Concentration of measure: fundamentals and tools

Tyrone Vincent, Luis Tenorio and Michael Wakin

Contents

1 Notation

Notation

- X - random variable or vector
- $\mathbb{E}[X]$ - expected value of random variable X.
- $\mathrm{Var}(X)$ - Variance of random variable X.
- $\mathbb{I}_{\mathbb{A}}$ - Indicator function for the event $\mathbb{A} \subset \Omega$.
- $\Pr[\mathbb{A}]$ - Probability of the event $\mathbb{A}$.
- $\|\cdot\|_2$, $\|\cdot\|_1$ - Euclidean norm, Absolute sum norm.
- $\mathcal{N}(\mu, P)$ - Gaussian distribution with mean μ and covariance P.
- $(x)_+$ - $\max\{0, x\}$.
- $\mathrm{tr}(M)$ - trace of the square matrix M.

2 Introduction

Abstract: This talk provides a tutorial covering basic material on concentration inequalities of functions of independent random variables around their mean. We will start with the inequalities of Markov, Chernoff and Hoeffding and end with the logarithmic Sobolev inequalities of Ledoux. We will also discuss other inequalities that apply to Gaussian processes. The focus will be on inequalities that play a role in applications to signal processing and compressive sensing and thus we will provide examples that show the practical use of the results.

2.1 References

The follow lecture notes, available on-line, are an excellent introduction to the subject, and covers all the results given here (Markov, Chernoff, Hoeffding, Efron-Stein and Sobolev inequalities) in good detail, along with many examples from statistical learning theory.

> Gábor Lugosi, *Concentration-of-measure Inequalities*, Lecture Notes.
> online: http://www.econ.upf.edu/∼lugosi/anu.pdf
>
> Alexander Barvinok, Lecture notes, University of Michigan.
> online: http://www.math.lsa.umich.edu/∼barvinok/total710.pdf

Ladoux and Talagrand developed many of the techniques for obtaining exponential concentration bounds. These books contain further material that extends the basic results.

> Michel Ladoux, *The Concentration of Measure Phenomenon*, American Mathematical Society, 2001
>
> Michel Talagrand, *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*, Springer-Verlag, 2005

The following are the original references where concentration of measure is applied towards proving the Johnson-Lindenstrauss lemma.

> P. Frankl and H. Maehara, “The Johnson-Lindenstrauss lemma at the sphericity of some graphs,” *Journal of Combinatorial Theory, Ser. B,* vol. 44, no. 3, pp.355-362, 1988
>
> P. Indyk and R. Motwani, *Approximate nearest neighbors: towards removing the curse of dimensionality*, 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, pp. 604-613, 1998.
>
> S. Dasgupta and A. Gupta, *An elementary proof of the Johnson-Lindenstrauss lemma*, Technical Report 99-006, UC Berkeley, March 1999.

2.2 Motivation

Concentration of Measure: What is it?

- Recall: the Weak Law of Large Numbers
  - X_i are independent random variables with common mean μ and uniformly bounded variance.
  - $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - Result:

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \Pr\left[\left|\bar{X}_n - \mu\right| < \epsilon\right] = 1$$

- This is a statement about a particular function of independent random variables being concentrated about its mean

$$\bar{X}_n = f\left(X_1, X_2, \cdots, X_n\right)$$

Concentration of Measure: The behavior of functions of independent random variables

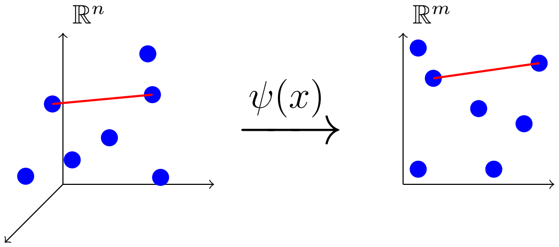
- Other functions are of interest, especially the norm of a linear mapping

$$f(X_1, X_2, \cdots, X_n) = \|\Phi X\|_2$$

- Possible mappings Φ
  - Projection Operator
  - Convolution Operator
  - Dictionary
- Concentration probabilities for finite n are useful
- Rates of decay can be important (want tight bounds)

2.3 Examples

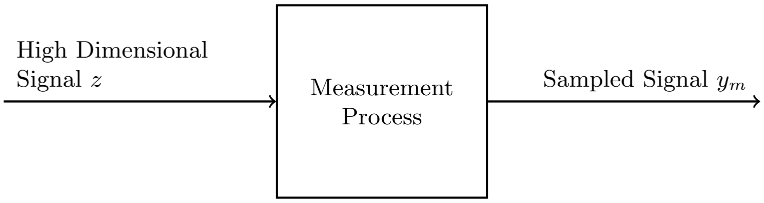
Example 1: Stable Embeddings



- Map set of N data points into lower dimensional space while preserving pair-wise distances.
  - Possible applications: search for nearest neighbors, compact data representations, clustering
- Questions:
  - For a given N and n, what is the required m to meet a specific distortion bound? (Johnson and Lindenstrauss)
  - How do we find the mapping ψ?

Example 2: Signal Recovery

- Basic signal processing question: How many measurements needed to represent a signal?



Example 2: Signal Recovery: Spectral Recovery

- Answer depends on signal model ($s \in \mathbb{S}$) and measurement model ($y_m = \phi_m(s)$).
- Signal model: Signal has spectral representation (in Fourier basis)

$$s(t) = \sum_k \alpha_k e^{j\omega_0 k t}$$

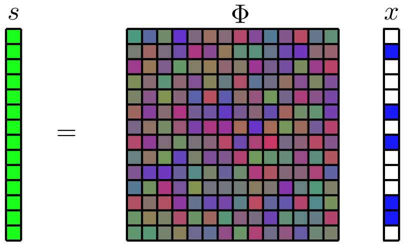
- Measurement model: Sampling

$$y_m = s(m\Delta t)$$

- Nyquist theorem: Original signal s can be recovered from samples y_m (over one period) if the sampling rate is twice the signal bandwidth.

Example 2: Signal Recovery: Compressive Sensing

- Compressive Sensing has different signal and measurement models.
- Signal model: Signal has sparse representation on some basis



- Measurement model: Linear mapping
- Questions (Answered next lecture):
  - What are the conditions on the measurement process that guarantee that all signals s of given sparsity can be recovered?
  - How can we design a good measurement process?

Example 3: Trace Estimate of a Matrix

- In large scale problems, the matrix multiplication Mx may be feasible, but $\text{tr}(M)$ may not be.
  - M may not fit in memory, and may be defined via other operations
- Estimate of trace for symmetric $M \in \mathbb{R}^{n \times n}$:
  - Select $x \sim \mathcal{N}(0, I)$.
  - Calculate $r = x'(Mx)$.
- $\mathbb{E}[r] = \text{tr}M$.
- Does this estimate concentrate around its mean? How does the concentration probability depend on the properties of M?

3 Basic Results

3.1 Markov and Chebyshev inequalities

The Statement of Markov's Inequality

Theorem 1 (Markov's Inequality)**.** *For any nonnegative random variable X with finite mean and $t > 0$,*

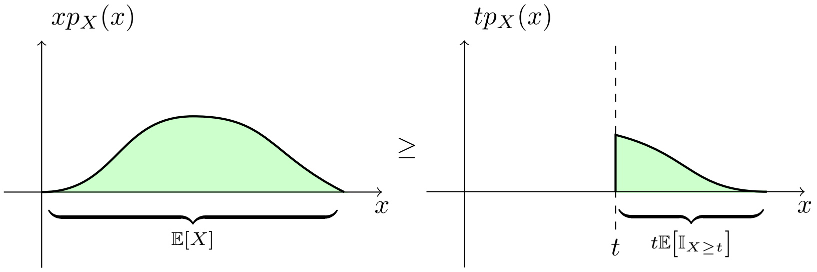
$$\Pr\left[X \geq t\right] \leq \frac{\mathbb{E}\left[X\right]}{t}$$

Remark 1. *Markov's inequality follows directly from the following:*

$$\begin{aligned}\mathbb{E}\left[X\right] &= \mathbb{E}\left[X\mathbb{I}_{X\geq t}\right] + \mathbb{E}\left[X\mathbb{I}_{X<t}\right]\\ &\geq t\mathbb{E}\left[\mathbb{I}_{X\geq t}\right]\\ &= t\Pr\left[X \geq t\right].\end{aligned}$$

This is illustrated below for a random variable with pdf $p_X(x)$.

Proof of Markov's Inequality



$$\mathbb{E}\left[X\right] \geq t\Pr\left[X \geq t\right]$$

Application of Markov's Inequality: Chebyshev's Inequality

Theorem 2 (Chebyshev's Inequality)**.** *For random variable X with finite variance σ^2,*

$$\Pr\left[\left|X - \mathbb{E}\left[X\right]\right| \geq t\right] \leq \frac{\sigma^2}{t^2} \quad \forall t > 0$$

Proof of Chebyshev's Inequality

- Note that $\Pr\left[\left|X - \mathbb{E}\left[X\right]\right| \geq t\right] = \Pr\left[\left|X - \mathbb{E}\left[X\right]\right|^2 \geq t^2\right]$
- Apply Markov's Inequality to the random variable

$$\phi = \left|X - \mathbb{E}\left[X\right]\right|^2.$$

- $\mathbb{E}\left[\phi\right] = \mathrm{Var}\left(X\right)$

$$\Pr\left[\phi \geq t^2\right] \leq \frac{\mathbb{E}\left[\phi\right]}{t^2}$$

$$\Pr\left[\left|X - \mathbb{E}\left[X\right]\right|^2 \geq t^2\right] \leq \frac{\mathrm{Var}\left(X\right)}{t^2}$$

$$\Pr\left[\left|X - \mathbb{E}\left[X\right]\right| \geq t\right] \leq \frac{\mathrm{Var}\left(X\right)}{t^2}$$

Application of Chebyshev's Inequality: The Weak Law of Large Numbers

- X_i are independent random variables with common mean μ and uniform variance bound σ^2_{sup}
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\begin{aligned}
\mathbb{E}\left[\bar{X}_n\right] &= \mu \\
\operatorname{Var}\left(\bar{X}_n\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}\left(X_i\right) \\
&\leq \frac{1}{n}\sup_i \operatorname{Var}\left(X_i\right) =: \frac{\sigma^2_{sup}}{n}
\end{aligned}$$

- Chebyshev's Inequality

$$\Pr\left[\left|\bar{X}_n - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2_{\text{sup}}}{n\epsilon^2}$$

$$\lim_{n\to\infty}\Pr\left[\left|\bar{X}_n - \mu\right| \geq \epsilon\right] = 0$$

How Tight is Chebyshev's Inequality?

- Chebyshev bound

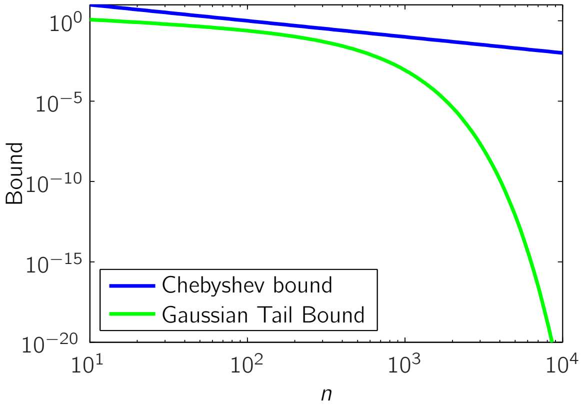
$$\Pr\left[\left|\bar{X}_n - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2_{\text{sup}}}{n\epsilon^2}$$

- Suppose X_i are Gaussian, $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- Then $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ (would approach Gaussian regardless by CLT)
- From tail bound on Gaussian distribution,

$$\Pr\left[\left|\bar{X}_n - \mu\right| \geq \epsilon\right] \leq \frac{\sigma}{\epsilon\sqrt{2\pi n}} e^{-n\epsilon^2/(2\sigma^2)}$$

- Chebyshev's bound decreases as $1/n$. The actual probability decreases exponentially in n.

Comparison of bounds



- Exponential dependence implies *critical* n. If probability of failure is small for $n = n_0$, it is *really* small for $n = 10n_0$.

3.2 Chernoff's bounding method

Idea of Chernoff's bounding method

- For Chebyshev's bound, we applied the second moment function $\phi(x) = x^2$ before applying Markov's inequality.
- Some moments may be better than others.
- Idea: choose
$$\phi(x, s) = e^{sx},$$
(which includes all moments,) then optimize over s.

Process for Chernoff's bounding method

- Given: random variable X.
- By monotonicity of e^{sx} for $s > 0$,
$$\Pr\left[X \geq t\right] = \Pr\left[e^{sX} \geq e^{st}\right]$$
- Apply Markov's inequality to right hand side
$$\Pr\left[X \geq t\right] \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}}$$
- $\mathbb{E}\left[e^{sX}\right]$ is moment generating function for X (when finite around $s = 0$)

Chernoff's bounding method summary

Theorem 3 (Chernoff's bounding method)**.** *For any random variable X and $t > 0$,*

$$\Pr\left[X \geq t\right] \leq \min_{s>0} \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}}$$
$$\Pr\left[X \leq t\right] \leq \min_{s>0} \frac{\mathbb{E}\left[e^{-sX}\right]}{e^{-st}}$$

when RHS exists.

Application: Norm of a Random Vector

- Let
$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$
be a Gaussian random vector with mean 0 and covariance matrix P.
- Does $\|X\|_2^2$ concentrate around its mean?

Application: Norm of a Random Vector Step 1: Moment Generating Function

- Moment Generating Function for $\|X\|_2^2$:

$$\mathbb{E}\left[e^{\pm s\|X\|_2^2}\right] = \frac{1}{\sqrt{\det\left(I \mp 2sP\right)}}$$

when $s \leq \lambda_{max}(P)_2$.

- Proof: Completion of squares

$$\begin{aligned}\mathbb{E}\left[e^{\pm s\|X\|_2^2}\right] &= \int \frac{1}{\left(2\pi \det\left(P\right)\right)^{\frac{1}{2}}} e^{\pm sX'X} e^{-\frac{1}{2}X'P^{-1}X} dX \\ &= \int \frac{1}{\left(2\pi \det\left(P\right)\right)^{\frac{1}{2}}} e^{-\frac{1}{2}X'\left(P^{-1} \mp 2sI\right)X} dX \\ &= \frac{\det^{\frac{1}{2}}\left(\left(P^{-1} \mp 2sI\right)^{-1}\right)}{\det^{\frac{1}{2}}\left(P\right)} \\ &= \frac{1}{\left(\det\left(P^{-1} \mp 2sI\right)\det P\right)^{\frac{1}{2}}} \\ &= \frac{1}{\sqrt{\det\left(I \mp 2sP\right)}}.\end{aligned}$$

- Special case: $P = I$ ($\|X\|_2^2 \sim \chi_n^2$)

$$\mathbb{E}\left[e^{s\|X\|_2^2}\right] = (1-2s)^{-\frac{n}{2}}$$

Application: Norm of a Random Vector Step 2: Use Chernoff's Method

- Concentration of norm of $X \sim \mathcal{N}(0, \sigma^2 I)$ around mean.
- Expected Norm

$$\mathbb{E}\left[\|X\|_2^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] = n\mathrm{Var}\left(X_1\right) = n\sigma^2$$

- Chernoff's bound, $\epsilon > 0$:

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq \min_{s>0} \frac{\mathbb{E}\left[e^{s\|X\|_2^2}\right]}{e^{s(1+\epsilon)n\sigma^2}}$$

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq \min_{s>0} \left(1-2s\sigma^2\right)^{-\frac{n}{2}} e^{-s(1+\epsilon)n\sigma^2}$$

Application: Norm of a Random Vector Step 3: Optimize over s

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq \min_{s>0} \left(1-2s\sigma^2\right)^{-\frac{n}{2}} e^{-s(1+\epsilon)n\sigma^2}$$

- optimal $s = \frac{\epsilon}{2(1+\epsilon)\sigma^2}$

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq \left((1+\epsilon)e^{-\epsilon}\right)^{\frac{n}{2}}$$

Note that

$$(1+\epsilon)e^{-\epsilon} = e^{-\epsilon+\log(1+\epsilon)}.$$

It is easy to verify that $\log(1+\epsilon) \leq \epsilon - \epsilon^2/2 + \epsilon^3/3$. Thus

$$(1+\epsilon)e^{-\epsilon} \leq e^{-\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)}.$$

Substituting into the probability bound,

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq \left(e^{-\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)}\right)^{\frac{n}{2}}$$

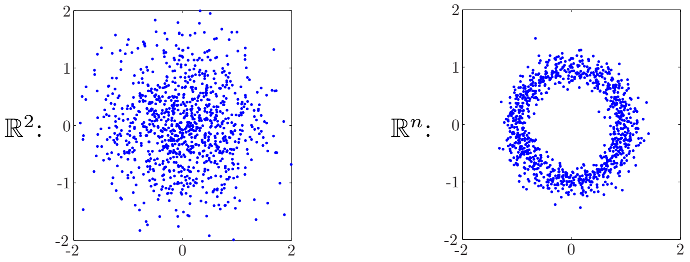
A second bound comes from noting that $\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} \geq \frac{\epsilon^2}{3}$ for $0 < \epsilon < 1/2$. Thus,

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq e^{-\epsilon^2 n/6} \quad 0 < \epsilon < 1/2$$

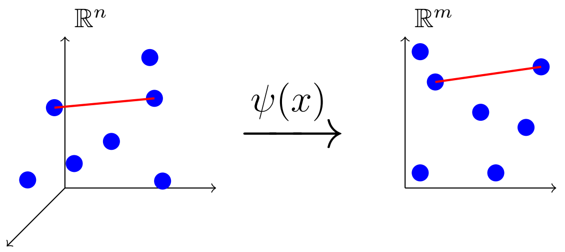
Application: Norm of a Random Vector: Result

$$\Pr\left[\|X\|_2^2 \geq (1+\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq e^{-\epsilon^2 n/6}$$
$$\Pr\left[\|X\|_2^2 \leq (1-\epsilon)\mathbb{E}\left[\|X\|_2^2\right]\right] \leq e^{-\epsilon^2 n/4}$$

- In high dimensions, $X \sim \mathcal{N}(0, \frac{1}{n}I)$ is concentrated near the unit sphere



Application: Stable Embedding



Theorem 4 (Johnson-Lindenstrauss)**.** *Given $\epsilon > 0$ and integer N, let m be a positive integer such that*

$$m \geq m_0 = O\left(\frac{\log N}{\epsilon^2}\right).$$

For every set $\mathbb{P}$ of N points in $\mathbb{R}^n$, there exists $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that for all $u, v \in \mathbb{P}$,

$$(1-\epsilon)\|u-v\|^2 \leq \|\psi(u)-\psi(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

Application: Stable Embedding

- Original proof utilized geometric approximation theory
- Simplified and *tightened* by Frankl and Maehara, Indyk and Motwani, Dasgupta and Gupta, using random mappings/concentration of measure

Application: Stable Embedding: Proof of J-L theorem

- Choose mapping

$$\psi(x) := \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{n2} & \cdots & a_{mn} \end{bmatrix} x = Ax$$

  where $a_{ij} \sim \mathcal{N}\left(0, \frac{1}{m}\right)$, indepdenent.

- Given set $\mathbb{P}$ of N points, there are $\binom{N}{2}$ vectors $x = u - v$, $u, v \in \mathbb{P}$.

Application: Stable Embedding: Proof of J-L theorem, step 1

- For fixed x consider $y = Ax$.
- By properties of Gaussian variables, $y_i \sim \mathcal{N}\left(0, \frac{\|x\|_2^2}{m}\right)$, independent.
- $\mathbb{E}\left[\|Ax\|_2^2\right] = \mathbb{E}\left[\|y\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^m y_i^2\right] = \|x\|_2^2$
- By "Norm of a Random Vector" result, for $0 < \epsilon < 0.5$,

$$\Pr\left[(1-\epsilon)\|x\|_2^2 \geq \|Ax\|_2^2 \geq (1+\epsilon)\|x\|_2^2\right] \leq 2e^{-\frac{\epsilon^2 m}{6}}$$

Application: Stable Embedding: Proof of J-L theorem, step 2

- Now consider $\binom{N}{2}$ vectors x.
- Using union bound $P(A \cup B) < P(A) + P(B)$,

$$\begin{aligned} \Pr\left[(1-\epsilon)\|x\|_2^2 \geq \|Ax\|_2^2 \geq (1+\epsilon)\|x\|_2^2\right] &\leq 2\binom{N}{2} e^{-\frac{\epsilon^2 m}{6}} \\ &\leq 2\left(eN/2\right)^2 e^{-\frac{\epsilon^2 m}{6}} \\ &= \frac{1}{2} e^2 e^{-\frac{\epsilon^2 m}{6} + 2\log N} \end{aligned}$$

- Probability of *not* achieving JL-embedding small if $m > O\left(\frac{\log N}{\min(\epsilon, 0.5)^2}\right)$

Application: Stable Embedding: Proof of J-L theorem, step 3

- Once the probability of failure drops below 1, a mapping exists.
- A *linear* mapping that is generated *randomly* will work with high probability for $m > m_0 = O\left(\frac{\log N}{\epsilon^2}\right)$.
- Probability of success depends exponentially on m.

Application: Trace Estimate: Problem Statement

- Estimate of trace for symmetric $M \in \mathbb{R}^{n \times n}$:
  - Select $x \sim \mathcal{N}(0, I)$.
  - Calculate $r = x'(Mx)$.
- $\mathbb{E}\left[r\right] = \mathrm{tr}M$.
- Using eigenvalue/eigenvector decomposition of $M = UDU'$,

  $$r = x'UDU'x = z'Dz = \sum_{i=1}^{n} \lambda_i z_i^2$$

  where $z_i \sim \mathcal{N}(0, I)$, λ_i: eigenvalues of M.

Application: Trace Estimate: Apply Chernoff Bound

- Chernoff bound ($0 < \epsilon < 1$):

  $$\Pr\left[r \leq (1-\epsilon)\mathrm{tr}M\right] \leq e^{s(1-\epsilon)\mathrm{tr}M}\mathbb{E}\left[e^{-s\sum \lambda_i z_i^2}\right]$$

- We found

  $$\mathbb{E}\left[e^{-s\lambda_i z_i^2}\right] = \frac{1}{\sqrt{1+2s\lambda_i}}$$

- Thus

  $$\begin{aligned}\Pr\left[r \leq (1-\epsilon)\mathrm{tr}M\right] &\leq \frac{e^{s(1-\epsilon)\mathrm{tr}M}}{\prod_i \sqrt{1+2s\lambda_i}} \\ &\leq e^{-\epsilon s(\mathrm{tr}M)} e^{s^2 \sum_i \lambda_i^2}\end{aligned}$$

  where we used $1/\sqrt{1+x} = e^{-0.5\log(1+x)}$ and $\log(1+x) \geq x - \frac{x^2}{2}$ for $x > 0$.

Application: Trace Estimate: Result

- Bound so far

  $$\Pr\left[r \leq (1-\epsilon)\mathrm{tr}M\right] \leq e^{-\epsilon s(\mathrm{tr}M)} e^{s^2 \sum_i \lambda_i^2}$$

- Optimal $s = \frac{\epsilon(\mathrm{tr}M)}{2\sum_i \lambda_i^2}$

  $$\Pr\left[r \leq (1-\epsilon)\mathrm{tr}M\right] \leq e^{-\epsilon^2/4\gamma(M)}$$

  where $\gamma(M) = \frac{\sum_i \lambda_i^2}{\mathrm{tr}M^2} = \frac{\sum_i \lambda_i^2}{(\sum_i \lambda_i)^2}$

- $\gamma(M)$ is related to the "spread" of eigenvalues
  - M orthonormal, $\gamma(M) = \frac{1}{n}$.

3.3 Hoeffding's Inequality

The Statement of Hoeffding's Inequality

- Problem: the moment generating function is not always easy to find, (any may not exist.)

Theorem 5 (Hoeffding's Inequality)**.** *Let X be a bounded random variable with mean 0 and $a \leq X \leq b$. Then for $s > 0$*

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}$$

- Proof: Use convexity of the exponential function: for $s \in [a, b]$,

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

Hoeffding's Tail Inequality

- Plugging into Chernoff's bound:

Theorem 6. *Let X_i be independent bounded random variables and $a_i \leq X_i \leq b_i$. Let $S_n = \sum_{i=1}^n X_i$. Then for all $\epsilon > 0$*

$$\Pr\left[S_n \geq \mathbb{E}\left[S_n\right] + \epsilon\right] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$
$$\Pr\left[S_n \leq \mathbb{E}\left[S_n\right] - \epsilon\right] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Application: Inner-Product of Sequence with Rademacher Distribution

- Suppose X is a length n random vector with elements drawn independently from $\{-, 1, 1\}$ with equal probability
- Let w be a length n vector with deterministic entries
- Consider inner product

$$S_n = \langle w, X \rangle = \sum_{i=1}^n w_i X_i$$

- Note that $w_i X_i$ is a random variable bounded between $-w_i$ and w_i, and $\mathbb{E}\left[S_n\right] = 0$.
- Using Hoeffding's Tail Inequality:

$$\Pr\left[|S_n| \geq \epsilon\right] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (2w_i)^2}\right)$$

$$\Pr\left[|S_n| \geq \epsilon\right] \leq \exp\left(\frac{-t^2}{2\|w\|_2^2}\right)$$

4 Logarithmic Sobolev inequalities

What comes next?

- So far, we have looked at inequalities for the 2-norm and inner products (which is still sums of random variables)
- In what follows, we will look at some inequalities that are useful for general functions of independent (but not necessarily identically distributed) random variables, which are not necessarily bounded

$$Z := g(X_1, \cdots, X_n)$$

4.1 Efron-Stein Inequality

Prediction

- Prediction plays an important role in signal processing
- Basic problem: Given measurement of Y, estimate X.
  - Y: radar return, X: airplane location
  - Y: reflectance measurement, X film thickness
  - $\cdots$

Theorem 7 (Minimum Mean Square Estimate)**.** *Given random variables X and Y, the (measureable) function $g(Y)$ that minimizes*

$$\mathbb{E}\left[(X - g(Y))^2\right]$$

is the conditional mean

$$\widehat{g}(Y) = \mathbb{E}\left[X|Y\right]$$

Efron-Stein Inequality, conditional mean version

Definition 8. Given (independent) random variables $X_1, \cdots, X_n$ and measurable function $Z = g(X_1, \cdots, X_n)$, define

$$\mathbb{E}\left[Z|X_{-i}\right] := \mathbb{E}\left[Z|X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\right]$$

Theorem 9 (Efron-Stein Inequality, conditional mean version)**.**

$$\mathrm{Var}\left(Z\right) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z|X_{-i}\right]\right)^2\right]$$

- Proof: See, e.g. Lugosi. Uses simple properties of conditional expectation.
- Note: If Z is sum of X_i, then $\mathbb{E}\left[\left(Z - \mathbb{E}\left[Z|X_{-i}\right]\right)^2\right] = \mathrm{Var}\left(X_i\right)$ and equality is achieved.

Efron-Stein Inequality: Modification of conditional mean

Definition 10. Given random variables $X_1, \cdots, X_n$ and measurable function $Z = g(X_1, \cdots, X_n)$, let $\tilde{X}_i$ be independent and identically distributed as X_i and define

$$Z_i := g(X_1, \cdots, X_{i-1}, \tilde{X}_i, X_{i+1}, \cdots, X_n)$$

- For any iid random variables X, Y

$$\mathrm{Var}\left(X\right) = \frac{1}{2}\mathbb{E}\left[(X - Y)^2\right] = \mathbb{E}\left[(X - Y)^2 \mathbb{I}_{X>Y}\right]$$

- Note that Z_i and $\mathbb{E}\left[Z|X_{-i}\right]$ are iid, conditioned on X_{-i}.

Efron-Stein Inequality: Theorem Statement

Theorem 11 (Efron-Stein Inequality)**.**

$$\mathrm{Var}\left(Z\right) \leq \frac{1}{2}\sum_{i}^{n}\mathbb{E}\left[\left(Z - Z_i\right)^2\right] = \sum_{i}^{n}\mathbb{E}\left[\left(Z - Z_i\right)^2\mathbb{I}_{Z>Z_i}\right]$$

- Can be used with Chebyshev inequality, but doesn't give exponential bounds.

Application: Largest Eigenvalue of a Random Matrix: Problem Statement

- Let $A \in \mathbb{R}^{n\times n}$ be a symmetric real matrix with elements $[A]_{ij}$, $1 \leq i \leq j \leq n$ independent random variables with magnitude bounded by 1.
- Let λ_i be the (real) eigenvalues of A, and define

$$Z = \max_i \lambda_i$$

- is Z concentrated around its mean?

Application: Largest Eigenvalue of a Random Matrix: Characterization of Max Eigenvalue

- Max gain property of largest eigenvalue of a symmetric matrix.

$$Z = \max_{\|u\|=1} u'Au$$

- The unit eigenvector v associated with the max eigenvalue attains the max gain.

Application: Largest Eigenvalue of a Random Matrix: Find Bound on Perturbed Value

- Let $\tilde{A}$ be matrix obtained by replacing $[A]_{ij}$ with an iid copy, and Z_{ij} be the max eigenvalue of this matrix. Then

$$\begin{aligned}(Z - Z_{ij})\mathbb{I}_{Z>Z_{ij}} &\leq (v'Av - v'\tilde{A}v)\mathbb{I}_{Z>Z_{ij}}\\ &\leq \left(v_i([A]_{ij} - [\tilde{A}]_{ij})v_j\right)_+\end{aligned}$$

- Since $[A]_{ij}$ and $-[\tilde{A}]_{ij}$ are bounded by 1,

$$(Z - Z_{ij})\mathbb{I}_{Z>Z_{ij}} \leq 2|v_i v_j|$$

Application: Largest Eigenvalue of a Random Matrix: Result

- Result:

$$\mathrm{Var}\left(Z\right) \leq \sum_{1\leq i\leq j\leq n} 4|v_i v_j|^2 \leq 4\|v\|^2 = 4$$

- Using Chebyshev's Inequality,

$$\Pr\left[|Z - \mathbb{E}\left[Z\right]| \geq \epsilon\right] \leq \frac{4}{\epsilon^2}$$

4.2 Entropy Method - Logarithmic Sobolev Inequality

Towards Exponential Bounds: Preliminaries

- Let $M(s) = \mathbb{E}\left[e^{sZ}\right]$ be the moment generating function of Z. If it exists,

$$\mathbb{E}[Z] = M'(s)|_{s=0} = \left.\frac{M'(s)}{M(s)}\right|_{s=0}$$

- Suppose there exists $C > 0$ such that the following bound holds:

$$F'(s) < C$$

  Then clearly for $s > 0$, $F(s) < F(0) + sC$.

Towards Exponential Bounds: What if...

- Suppose

$$\frac{M'(s)}{sM(s)} - \frac{\log M(s)}{s^2} \le C$$

- Then with $F(s) = \frac{\log M(s)}{s}$,

$$F'(s) \le C$$

- Thus, for $s > 0$,

$$\begin{aligned} \frac{\log M(s)}{s} &< \lim_{s \to 0} \frac{\log M(s)}{s} + sC \\ &= \left.\frac{M'(s)}{M(s)}\right|_{s=0} + sC \\ &= \mathbb{E}[Z] + sC \end{aligned}$$

- Implying

$$M(s) < e^{s\mathbb{E}[Z] + s^2 C}$$

Towards Exponential Bounds: Recap

- Inequality

$$sM'(s) - M(s)\log M(s) \le s^2 C M(s)$$

  implies the bound on moment generating function

$$M(s) < e^{s\mathbb{E}[Z] + s^2 C}.$$

- This can be used with Chebyshev's bounding method to show, e.g.

$$\Pr[Z - \mathbb{E}[Z] \ge \epsilon] \le e^{-\epsilon^2/4C}$$

Entropy Method

- Note that since
$$\mathrm{Var}\left(Z\right) = \mathbb{E}\left[Z^2\right] - \left(\mathbb{E}\left[Z\right]\right)^2$$
the conditional mean version of the Efron-Stein Inequality can be re-written as
$$\mathbb{E}\left[\phi(Z)\right] - \phi\left(\mathbb{E}\left[Z\right]\right) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\phi(Z)|X_{-i}\right] - \phi\left(\mathbb{E}\left[Z|X_{-i}\right]\right)\right]$$
where $\phi(z) = z^2$.
- Idea: Prove this is true with for $\phi(z) = z\log(z)$, and use $Z \leftarrow e^{sZ}$, since in this case
$$\mathbb{E}\left[\phi(Z)\right] = sM'(s), \quad \phi\left(\mathbb{E}\left[Z\right]\right) = M(s)\log M(s)$$

Why is this called Entropy Method?

Definition 12. Given two probability distributions P and Q with densities $p(x)$ and $q(x)$, define the *relative entropy* (or Kullback-Leibler divergence) of P from Q to be
$$D(P||Q) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

- Given an optimal coding of Q, the relative entropy is the expected extra number of bits needed to transmit samples from P using this code.

Entropy interpretation

- Given distribution P of X_i with density $p(x)$, Let Q be the distribution with density $q(X) = g(X)p(X)$.
- Interpretation: Let $\mathbb{E}\left[Z\right] = 1$. Then
$$\begin{aligned}\mathbb{E}\left[\phi(Z)\right] - \phi\left(\mathbb{E}\left[Z\right]\right) &= \mathbb{E}\left[Z\log(Z)\right] - \mathbb{E}\left[Z\right]\log(\mathbb{E}\left[Z\right]) \\ &= \mathbb{E}\left[Z\log(Z)\right] \\ &= \int g(x)\log(g(x))p(x)dx \\ &= \int q(x)\log\frac{q(x)}{p(x)}dx \\ &= D(P||Q)\end{aligned}$$

Tensorization inequality of the entropy

Theorem 13. *Let $\phi(x) = x\log(x)$ for $x > 0$. Let $X_1, \cdots, X_n$ be independent random variables, and let g be a positive-valued function of these variables, with $Z = g(X_1, \cdots, X_n)$. Then for $\phi(z) = z\log(z)$,*
$$\mathbb{E}\left[\phi(Z)\right] - \phi\left(\mathbb{E}\left[Z\right]\right) \leq \frac{1}{2}\sum_{i}^{n}\mathbb{E}\left[\mathbb{E}\left[\phi(Z)|X_{-i}\right] - \phi\left(\mathbb{E}\left[Z|X_{-i}\right]\right)\right]$$

- Proof: Lugosi, Ledoux.

A Logarithmic Sobolev Inequality...

Theorem 14. *Suppose there exists a positive constant C such that (a.s.)*

$$\sum_{i=1}^{n}(Z - Z_i)^2 \mathbb{I}_{Z>Z_i} \leq C.$$

Let $M(s) = \mathbb{E}\left[e^{sZ}\right]$ be the moment generating function of Z. Then

$$sM'(s) - M(s)\log M(s) \leq s^2 C M(s)$$

- This is exactly the kind of bound we are looking for!
- Proof sketch: bound right hand side using

$$\mathbb{E}\left[\phi(e^{sZ})|X_{-i}\right] - \phi\left(\mathbb{E}\left[e^{sZ}|X_{-i}\right]\right) \leq \mathbb{E}\left[s^2 e^{sZ}(Z - Z_i)^2 \mathbb{I}_{Z>Z_i}|X_{-i}\right]$$

... Gives a Concentration of Measure Inequality

Corollary 15. *Suppose there exists a positive constant C such that*

$$\sum_{i=1}^{n}(Z - Z_i)^2 \mathbb{I}_{Z>Z_i} \leq C.$$

Then for all $t > 0$,

$$\Pr\left[Z - \mathbb{E}\left[Z\right] \geq \epsilon\right] \leq e^{-\epsilon^2/4C}$$

Application: Largest Eigenvalue of a Random Matrix, again

Theorem 16. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric real matrix with elements $[A]_{ij}$, $1 \leq i \leq j \leq n$ independent random variables with magnitude bounded by 1. Let Z be the max eigenvalue of A. Then*

$$\Pr\left[Z - \mathbb{E}\left[Z\right] \geq \epsilon\right] \leq e^{-\epsilon^2/16}$$

Conclusion

- Everything starts with Markov's inequality
- For exponential bounds, we needed
  - Chernoff's bounding method
  - Logarithmic Sobolev Inequality
- Next lecture: Concentration of Measure applied to Compressive Sensing