Concentration of measure: fundamentals and tools

Tyrone Vincent, Luis Tenorio and Michael Wakin

Colorado School of Mines
1 Introduction
   - Motivation
   - Examples

2 Basic Results
   - Markov and Chebyshev inequalities
   - Chernoff’s bounding method
   - Hoeffding’s Inequality

3 Logarithmic Sobolev inequalities
   - Efron-Stein Inequality
   - Entropy Method - Logarithmic Sobolev Inequality
Concentration of Measure:
What is it?

- Recall: the Weak Law of Large Numbers
  - $X_i$ are independent random variables with common mean $\mu$ and uniformly bounded variance.
  - $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
  - Result:
    \[
    \forall \epsilon > 0 \quad \lim_{n \to \infty} \Pr \left[ |\bar{X}_n - \mu| < \epsilon \right] = 1
    \]

- This is a statement about a particular function of independent random variables being concentrated about its mean
  \[
  \bar{X}_n = f (X_1, X_2, \cdots, X_n)
  \]
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 \( \Phi \)
  - Projection Operator
  - Convolution Operator
  - Dictionary

- Concentration probabilities for finite \( n \) are useful
- Rates of decay can be important (want tight bounds)
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 $\psi$?
Example 2: Signal Recovery

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

![Diagram]

- High Dimensional Signal $z$
- Measurement Process
- Sampled Signal $y_m$
Example 2: Signal Recovery:
Spectral Recovery

- Answer depends on signal model \((s \in S)\) and measurement model \((y_m = \phi_m(s))\).
- Signal model: Signal has spectral representation (in Fourier basis)
  \[ s(t) = \sum_k \alpha_k e^{j\omega_0 kt} \]
- 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 $s$.

\[ \Phi \]

- 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 $M$?
The Statement of Markov’s Inequality

Theorem (Markov’s Inequality)

For any nonnegative random variable $X$ with finite mean and $t > 0$,

$$\Pr [X \geq t] \leq \frac{E[X]}{t}$$
Proof of Markov’s Inequality

\[ E[X] \geq t \Pr [X \geq t] \]
Theorem (Chebyshev's Inequality)

For random variable $X$ with finite variance $\sigma^2$, 

$$\Pr [ |X - \mathbb{E}[X]| \geq t ] \leq \frac{\sigma^2}{t^2} \quad \forall t > 0$$
Proof of Chebyshev’s Inequality

- Note that \( \Pr \left[ |X - \mathbb{E}[X]| \geq t \right] = \Pr \left[ |X - \mathbb{E}[X]|^2 \geq t^2 \right] \)

- Apply Markov’s Inequality to the random variable

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

- \( \mathbb{E}[\phi] = \text{Var}(X) \)

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

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

  \[ \Pr [|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}(X)}{t^2} \]
Application of Chebyshev’s Inequality: The Weak Law of Large Numbers

- \( X_i \) are independent random variables with common mean \( \mu \) and uniform variance bound \( \sigma_{sup}^2 \)
- \( \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \).

\[
\mathbb{E} [\bar{X}_n] = \mu
\]

\[
\text{Var} (\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var} (X_i)
\]

\[
\leq \frac{1}{n} \sup_{i} \text{Var} (X_i) =: \frac{\sigma_{sup}^2}{n}
\]

- Chebyshev’s Inequality

\[
\Pr \left[ \left| \bar{X}_n - \mu \right| \geq \epsilon \right] \leq \frac{\sigma_{sup}^2}{n\epsilon^2}
\]
How Tight is Chebyshev’s Inequality?

- Chebyshev bound

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

- 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 \textit{critical} \( n \). If probability of failure is small for \( n = n_0 \), it is \textit{really} small for \( n = 10n_0 \).
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 $s > 0$,
  \[ \Pr [X \geq t] = \Pr [e^{sX} \geq e^{st}] \]
- Apply Markov’s inequality to right hand side
  \[ \Pr [X \geq t] \leq \frac{\mathbb{E} [e^{sX}]}{e^{st}} \]
- $\mathbb{E} [e^{sX}]$ is moment generating function for $X$
Chernoff’s bounding method summary

Theorem (Chernoff’s bounding method)

For any random variable $X$ and $t > 0$,

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

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

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 $\|X\|^2_2$:**

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

- **Proof: Completion of squares**

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

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

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

- Chernoff’s bound, $\epsilon > 0$:

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

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

\[
\Pr \left[ \|X\|_2^2 \geq (1 + \epsilon) \mathbb{E} \left[ \|X\|_2^2 \right] \right] \leq e^{-\epsilon^2 n/6} \quad 0 < \epsilon < 1/2
\]
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 (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 \( \psi : \mathbb{R}^n \to \mathbb{R}^m \) 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 \( a_{ij} \sim \mathcal{N}(0, \frac{1}{m}) \), independent.

- Given set \( \mathcal{P} \) of \( N \) points, there are \( \binom{N}{2} \) vectors \( x = u - v \), \( u, v \in \mathcal{P} \).
Application: Stable Embedding:
Proof of J-L theorem, step 1

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

$$\Pr [(1 - \epsilon)\|x\|_2^2 \geq \|Ax\|_2^2 \geq (1 + \epsilon)\|x\|_2^2] \leq 2e^{-\frac{\epsilon^2 m}{6}}$$
Application: Stable Embedding:
Proof of J-L theorem, step 2

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

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

- 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 $r = x'(Mx)$.
- $\mathbb{E}[r] = \text{tr}M$.
- Using eigenvalue/eigenvector decomposition of $M = UDU'$,

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

where $z_i \sim \mathcal{N}(0, I)$, $\lambda_i$: eigenvalues of $M$. 
Application: Trace Estimate:

Apply Chernoff Bound

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

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

- We found

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

- Thus

\[
\Pr [ r \leq (1 - \epsilon)\text{tr}M ] \leq \frac{e^{s(1-\epsilon)\text{tr}M}}{\prod_i \sqrt{1 + 2s \lambda_i}} \leq e^{-\epsilon s (\text{tr}M)} e^{s^2 \sum_i \lambda_i^2}
\]
Application: Trace Estimate:

Result

- Bound so far

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

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

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

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

- \( \gamma(M) \) is related to the “spread” of eigenvalues
  - \( M \) orthonormal, \( \gamma(M) = \frac{1}{n} \).
The Statement of Hoeffding’s Inequality

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

Theorem (Hoeffding’s Inequality)

Let $X$ be a bounded random variable with mean 0 and $a \leq X \leq b$. Then for $s > 0$

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

- 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**

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 $\epsilon > 0$

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

\[
\Pr [S_n \leq \mathbb{E} [S_n] - \epsilon] \leq \exp \left( \frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)
\]
Application: Inner-Product of Sequence with Rademacher Distribution

- Suppose $X$ is a length $n$ random vector with elements drawn independently from $\{-1, 1, 1\}$ 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 $\mathbb{E}[S_n] = 0$.

- Using Hoeffding’s Tail Inequality:

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

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

\[ Z := g(X_1, \cdots, X_n) \]
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
  - ...

**Theorem (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*

$$
\hat{g}(Y) = \mathbb{E} [X | Y]
$$
**Efron-Stein Inequality, conditional mean version**

**Definition**

Given (independent) random variables $X_1, \ldots, X_n$ and measurable function $Z = g(X_1, \ldots, X_n)$, define

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

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

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

- **Proof:** See, e.g. Lugosi. Uses simple properties of conditional expectation.

- **Note:** If $Z$ is sum of $X_i$, then $\mathbb{E} \left[ (Z - \mathbb{E} [Z|X_{-i}])^2 \right] = \text{Var} (X_i)$ and equality is achieved.
Efron-Stein Inequality:
Modification of conditional mean

Definition
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 $X, Y$

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

- Note that $Z_i$ and $\mathbb{E} [Z|X_{-i}]$ are iid, conditioned on $X_{-i}$. 
Efron-Stein Inequality:
Theorem Statement

Theorem (Efron-Stein Inequality)

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

- 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 $\lambda_i$ 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.
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

\[
(Z - Z_{ij})I_{Z > Z_{ij}} \leq (v' Av - v' \tilde{A}v)I_{Z > Z_{ij}} \\
\leq \left( v_i ( [A]_{ij} - [\tilde{A}]_{ij} ) v_j \right)_+
\]

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

\[
(Z - Z_{ij})I_{Z > Z_{ij}} \leq 2|v_i v_j|
\]
Application: Largest Eigenvalue of a Random Matrix:

Result:

\[ \text{Var}(Z) \leq \sum_{1 \leq i \leq j \leq n} 4|v_i v_j|^2 \leq 4\|v\|^2 = 4 \]

Using Chebyshev’s Inequality,

\[ \Pr[|Z - \mathbb{E}[Z]| \geq \epsilon] \leq \frac{4}{\epsilon^2} \]
Towards Exponential Bounds:

Preliminaries

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

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

- 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

\[ \frac{M'(s)}{sM(s)} - \frac{\log M(s)}{s^2} \leq C \]

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

\[ F'(s) \leq C \]

- Thus, for \( s > 0 \),

\[ \frac{\log M(s)}{s} < \lim_{s \to 0} \frac{\log M(s)}{s} + sC \]

\[ = \frac{M'(s)}{M(s)} \bigg|_{s=0} + sC \]

\[ = \mathbb{E}[Z] + sC \]

- Implying

\[ M(s) < e^{s\mathbb{E}[Z]+s^2C} \]
Towards Exponential Bounds:

Recap

- Inequality

\[ sM'(s) - M(s) \log M(s) \leq s^2 CM(s) \]

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.

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

- Note that since
  \[ \text{Var} (Z) = \mathbb{E} [Z^2] - (\mathbb{E} [Z])^2 \]
  the conditional mean version of the Efron-Stein Inequality can be re-written as
  \[
  \mathbb{E} [\phi(Z)] - \phi (\mathbb{E} [Z]) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E} [\mathbb{E} [\phi(Z)|X_{-i}] - \phi (\mathbb{E} [Z|X_{-i}])]
  \]
  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
  \[
  \mathbb{E} [\phi(Z)] = sM'(s), \quad \phi (\mathbb{E} [Z]) = M(s) \log M(s)
  \]
Why is this called Entropy Method?

Definition

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 $\mathbb{E}[Z] = 1$. Then

\[
\mathbb{E}[\phi(Z)] - \phi(\mathbb{E}[Z]) = \mathbb{E}[Z \log(Z)] - \mathbb{E}[Z] \log(\mathbb{E}[Z]) \\
= \mathbb{E}[Z \log(Z)] \\
= \int g(x) \log(g(x))p(x)dx \\
= \int q(x) \log \frac{q(x)}{p(x)}dx \\
= D(P||Q)
\]
Tensorization inequality of the entropy

**Theorem**

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 \( \phi(z) = z \log(z) \),

\[
\mathbb{E} [\phi(Z)] - \phi (\mathbb{E} [Z]) \leq \frac{1}{2} \sum_{i} \mathbb{E} [ \mathbb{E} [\phi(Z)|X_{-i}] - \phi (\mathbb{E} [Z|X_{-i}])] 
\]

- **Proof:** Lugosi, Ledoux.
Theorem

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

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

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

$$\mathbb{E} [\phi(e^{sZ})|X_{-i}] - \phi (\mathbb{E} [e^{sZ}|X_{-i}]) \leq \mathbb{E} [s^2 e^{sZ} (Z - Z_i)^2 \mathbb{I}_{Z > Z_i}|X_{-i}]$$
... Gives a Concentration of Measure Inequality

**Corollary**

*Suppose there exists a positive constant $C$ such that*

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

*Then for all $t > 0$,*

$$
\Pr \left[ Z - \mathbb{E} [Z] \geq \epsilon \right] \leq e^{-\epsilon^2 / 4C}.
$$
Application: Largest Eigenvalue of a Random Matrix, again

Theorem

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

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