Concentration of Measure

Jarosław Błasiok

January 14, 2020

Concentration of Measure:

Some random variables are often close to their expectation.

Concentration of Measure:
Some random variables are often close to their expectation.

Example

Let $X_1, X_2, \ldots X_n \in \{0, 1\}$, be such that $\mathbb{E}X_i = p$ is bounded away from 0 and 1. Then

$$\mathbb{P}(|\frac{1}{n}\sum X_i - p| > \lambda) < \exp(-C\lambda^2 n)$$

Concentration of Measure:
Some random variables are often close to their expectation.

Example

Let $X_1, X_2, \ldots X_n \in \{0, 1\}$, be such that $\mathbb{E}X_i = p$ is bounded away from 0 and 1. Then

$$\mathbb{P}(|\frac{1}{n}\sum X_i - p| > \lambda) < \exp(-C\lambda^2 n)$$

Comparing with Central Limit Theorem?

The inequality above says that $\frac{1}{n}\sum X_i$ looks like gaussian with mean p and standard deviation $\frac{1}{\sqrt{n}}$.

If we tried to deduce a concentration inequality from CLT, we could only get that $\mathbb{P}(|\frac{1}{n}\sum X_i - p| > \lambda) \lesssim \frac{1}{\sqrt{n}}$.

Goal of this talk:

Johnson-Lindenstrauss Lemma

Theorem (JL)

For any set of points $x_1, \dots x_n \in \mathbb{R}^d$ and any $\varepsilon > 0$, there is $m = \mathcal{O}(\frac{\log n}{\varepsilon^2})$ and a linear map $T : \mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall i, j \qquad (1-\varepsilon)\|x_i - x_j\|_2 \leq \|Tx_i - Tx_j\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2.$$

Goal of this talk:
Johnson-Lindenstrauss Lemma

Theorem (JL)

For any set of points $x_1, \dots x_n \in \mathbb{R}^d$ and any $\varepsilon > 0$, there is $m = \mathcal{O}(\frac{\log n}{\varepsilon^2})$ and a linear map $T : \mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall i, j \qquad (1-\varepsilon)\|x_i - x_j\|_2 \leq \|Tx_i - Tx_j\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2.$$

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \qquad \mathbb{P}_{T \sim \mathcal{D}}\left(\left|\|Ty\|_2^2 - \|y\|_2^2\right| > \varepsilon\right) < \delta.$$

Goal of this talk:

Johnson-Lindenstrauss Lemma

Theorem (JL)

For any set of points $x_1, \dots x_n \in \mathbb{R}^d$ and any $\varepsilon > 0$, there is $m = \mathcal{O}(\frac{\log n}{\varepsilon^2})$ and a linear map $T : \mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall i, j \qquad (1-\varepsilon)\|x_i - x_j\|_2 \leq \|Tx_i - Tx_j\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2.$$

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \qquad \mathbb{P}_{T \sim \mathcal{D}}\left(\left|\|Ty\|_2^2 - \|y\|_2^2\right| > \varepsilon\right) < \delta.$$

DJL implies JL

Why? Take $\delta < \frac{1}{n^2}$ and union bound over all pairs x_i, x_j.

This is a concentration of measure kind of statement.

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \mathbb{P}\left(\left|\|Ty\|_2^2 - \|y\|_2^2\right| > \varepsilon\right) < \delta.$$

We will take T such that $\mathbb{E}\|Ty\|_2^2 = \|y\|_2^2$.

This is a concentration of measure kind of statement.

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \mathbb{P}\left(\left|\|Ty\|_2^2 - \|y\|_2^2\right| > \varepsilon\right) < \delta.$$

We will take T such that $\mathbb{E}\|Ty\|_2^2 = \|y\|_2^2$.

It is not too difficult

I will use this theorem as a motivating example to trick you into learning more general way of thinking about concentration inequalities.

Application of JL lemma: Sketch-and-solve paradigm

Example

Consider the k-means problem: given a high-dimensional dataset $x_1, x_2, \ldots x_n \in \mathbb{R}^d$, find clusters $C_1, C_2, \ldots C_k \subset [n]$ and cluster centers y_i minimizing $\sum_{i \leq k} \sum_{j \in C_i} \|x_j - y_k\|^2$.

Solution?

We can start by reducing dimension, then work on $\mathcal{O}(\log n)$ dimensional dataset.

Concentration of Measure

Markov Inequality

If $X \in \mathbb{R}$ is nonnegative random variable, then

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}X}{\lambda}$$

Proof.

$$\mathbb{E}X \geq \mathbb{E}\mathbf{1}(X \geq \lambda)X \geq \mathbb{E}\mathbf{1}(X \geq \lambda)\lambda = \mathbb{P}(X \geq \lambda)\lambda$$

□

Chebyshev Inequality

For any $p > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq \lambda) = \mathbb{P}(|X - \mathbb{E}X|^p \geq \lambda^p) \leq \frac{\|X - \mathbb{E}X\|_p^p}{\lambda^p},$$

where $\|Y\|_p := (\mathbb{E}|Y|^p)^{1/p}$. Equivalently

$$\mathbb{P}(|X - \mathbb{E}X| > \lambda \|X - \mathbb{E}X\|_p) \leq \frac{1}{\lambda^p}.$$

Chebyshev Inequality

For any $p > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq \lambda) = \mathbb{P}(|X - \mathbb{E}X|^p \geq \lambda^p) \leq \frac{\|X - \mathbb{E}X\|_p^p}{\lambda^p},$$

where $\|Y\|_p := (\mathbb{E}|Y|^p)^{1/p}$. Equivalently

$$\mathbb{P}(|X - \mathbb{E}X| > \lambda \|X - \mathbb{E}X\|_p) \leq \frac{1}{\lambda^p}.$$

Often we have $\|X - \mathbb{E}X\|_p \leq C_p \|X - \mathbb{E}X\|_2 = C_p \sigma$. In such a case we have

$$\mathbb{P}(|X - \mathbb{E}X| > \lambda\sigma) < \frac{C_p}{\lambda^p}.$$

Concentration of measure corresponds to moment inequalities.

If we can bound $\|X - \mathbb{E}X\|_p$, we get tail bound $\mathbb{P}(|X - \mathbb{E}X| > \lambda) < \ldots$

Concentration of measure corresponds to moment inequalities.

If we can bound $\|X - \mathbb{E}X\|_p$, we get tail bound $\mathbb{P}(|X - \mathbb{E}X| > \lambda) < \dots$

Properites of $\|\cdot\|_p$-norms

Often it is easier to work with moment bounds, and deduce the tail bound only at the end.

- Triangle inequality: $\forall p \geq 1, \|X + Y\|_p \leq \|X\|_p + \|Y\|_p$

Concentration of measure corresponds to moment inequalities.

If we can bound $\|X - \mathbb{E}X\|_p$, we get tail bound $\mathbb{P}(|X - \mathbb{E}X| > \lambda) < \dots$

Properites of $\|\cdot\|_p$-norms

Often it is easier to work with moment bounds, and deduce the tail bound only at the end.

- Triangle inequality: $\forall p \geq 1, \|X + Y\|_p \leq \|X\|_p + \|Y\|_p$
- Monotone: $\forall q < p, \|Y\|_q \leq \|Y\|_p$

Concentration of measure corresponds to moment inequalities.

If we can bound $\|X - \mathbb{E}X\|_p$, we get tail bound $\mathbb{P}(|X - \mathbb{E}X| > \lambda) < \dots$

Properites of $\|\cdot\|_p$-norms

Often it is easier to work with moment bounds, and deduce the tail bound only at the end.

- Triangle inequality: $\forall p \geq 1, \|X + Y\|_p \leq \|X\|_p + \|Y\|_p$
- Monotone: $\forall q < p, \|Y\|_q \leq \|Y\|_p$
- ℓ_∞-norm: $\lim_{p\to\infty} \|Y\|_p = \sup\{\lambda : \mathbb{P}(Y > \lambda) > 0\}$.

Weak converse to Markov inequality

By the way, this is almost equivalent. Let Y be non-negative random variable, such that for all λ we have

$$\mathbb{P}(|Y| > \tau\lambda) \leq \frac{1}{\lambda^p}$$

Then for any $q < p$ we have

$$\|Y\|_q \leq C(p,q)\tau$$

Weak converse to Markov inequality

By the way, this is almost equivalent. Let Y be non-negative random variable, such that for all λ we have

$$\mathbb{P}(|Y| > \tau\lambda) \leq \frac{1}{\lambda^p}$$

Then for any $q < p$ we have

$$\|Y\|_q \leq C(p,q)\tau$$

Proof.

W.l.o.g. $\tau = 1$ and $q = 1$. Then $p > 1$ and by integration by parts

$$\mathbb{E}|Y| = \int_0^\infty \mathbb{P}(|Y| > \lambda)\,\mathrm{d}\lambda \leq 1 + \int_1^\infty \frac{1}{\lambda^p}\,\mathrm{d}\lambda \lesssim 1$$

□

Take away message: moment bounds imply tail bounds, but tail bounds imply moment bounds!

Gaussian random variables

There is a distribution $\mathcal{N}(0,1)$ satisfying:

- $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 = 1$.

Gaussian random variables

There is a distribution $\mathcal{N}(0,1)$ satisfying:

- $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 = 1$.
- For $Z_1, Z_2, \ldots Z_n \sim \mathcal{N}(0,1)$ i.i.d., we have

$$\sum a_i Z_i \sim \|a\|_2 Z$$

Gaussian random variables

There is a distribution $\mathcal{N}(0,1)$ satisfying:

- $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 = 1$.
- For $Z_1, Z_2, \ldots Z_n \sim \mathcal{N}(0,1)$ i.i.d., we have

$$\sum a_i Z_i \sim \|a\|_2 Z$$

- For every $\lambda > 0$, we have $\mathbb{P}(Z > \lambda) < \exp(-\lambda^2/4)$.

Gaussian random variables

There is a distribution $\mathcal{N}(0,1)$ satisfying:

- $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 = 1$.
- For $Z_1, Z_2, \ldots Z_n \sim \mathcal{N}(0,1)$ i.i.d., we have

$$\sum a_i Z_i \sim \|a\|_2 Z$$

- For every $\lambda > 0$, we have $\mathbb{P}(Z > \lambda) < \exp(-\lambda^2/4)$.
- For every $p \geq 1$ we have $\|Z\|_p \simeq \sqrt{p}$.

Moment generating function

For a random variable X define $\Phi_X(\lambda) := \mathbb{E}e^{\lambda X}$, and $\phi_X(\lambda) = \ln \Phi_X(\lambda)$.

Moment generating function

For a random variable X define $\Phi_X(\lambda) := \mathbb{E}e^{\lambda X}$, and $\phi_X(\lambda) = \ln \Phi_X(\lambda)$.

Why is it useful?

If X, Y are independent random variables, then

$$\Phi_{X+Y}(\lambda) = \mathbb{E}e^{\lambda(X+Y)} = \mathbb{E}e^{\lambda X}\mathbb{E}e^{\lambda Y} = \Phi_X(\lambda)\Phi_Y(\lambda).$$

Therefore $\phi_{X+Y}(\lambda) = \phi_X(\lambda) + \phi_Y(\lambda)$.

Moment generating function

For a random variable X define $\Phi_X(\lambda) := \mathbb{E}e^{\lambda X}$, and $\phi_X(\lambda) = \ln \Phi_X(\lambda)$.

Why is it useful?

If X, Y are independent random variables, then

$$\Phi_{X+Y}(\lambda) = \mathbb{E}e^{\lambda(X+Y)} = \mathbb{E}e^{\lambda X}\mathbb{E}e^{\lambda Y} = \Phi_X(\lambda)\Phi_Y(\lambda).$$

Therefore $\phi_{X+Y}(\lambda) = \phi_X(\lambda) + \phi_Y(\lambda)$.

Why is it called Moment Generating Function?

$$\Phi_X(\lambda) = \sum \lambda^p \frac{\mathbb{E}X^p}{p!}$$

Moment generating function

For a random variable X define $\Phi_X(\lambda) := \mathbb{E}e^{\lambda X}$, and $\phi_X(\lambda) = \ln \Phi_X(\lambda)$.

Why is it useful?

If X, Y are independent random variables, then

$$\Phi_{X+Y}(\lambda) = \mathbb{E}e^{\lambda(X+Y)} = \mathbb{E}e^{\lambda X}\mathbb{E}e^{\lambda Y} = \Phi_X(\lambda)\Phi_Y(\lambda).$$

Therefore $\phi_{X+Y}(\lambda) = \phi_X(\lambda) + \phi_Y(\lambda)$.

Why is it called Moment Generating Function?

$$\Phi_X(\lambda) = \sum \lambda^p \frac{\mathbb{E}X^p}{p!}$$

MGF bounds imply moment bound!

If $\Phi_X(\lambda) < \tau$ then $\mathbb{E}X^p < \frac{p!}{\lambda^p}\tau$, therefore $\|X\|_p \leq \frac{p}{\lambda}\tau^{1/p}$.

Subgaussian random variable

For gaussian random variable $Z \sim \mathcal{N}(0,1)$ we have $\Phi_Z(\lambda) = e^{\lambda^2/2}$, or equivalently $\phi_Z(\lambda) = \lambda^2/2$.

Definition

For a mean zero random variable Y (i.e. $\mathbb{E}Y = 0$), we say that Y is σ-subgaussian if and only if $\forall\lambda$, $\phi_Y(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$.

Subgaussian random variable

For gaussian random variable $Z \sim \mathcal{N}(0,1)$ we have $\Phi_Z(\lambda) = e^{\lambda^2/2}$, or equivalently $\phi_Z(\lambda) = \lambda^2/2$.

Definition

For a mean zero random variable Y (i.e. $\mathbb{E}Y = 0$), we say that Y is σ-subgaussian if and only if $\forall\lambda$, $\phi_Y(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$.

Fact

If Y is σ-subgaussian, and c is constant, then cY is $c\sigma$-subgaussian.

Subgaussian random variable

For gaussian random variable $Z \sim \mathcal{N}(0,1)$ we have $\Phi_Z(\lambda) = e^{\lambda^2/2}$, or equivalently $\phi_Z(\lambda) = \lambda^2/2$.

Definition

For a mean zero random variable Y (i.e. $\mathbb{E}Y = 0$), we say that Y is σ-subgaussian if and only if $\forall\lambda$, $\phi_Y(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$.

Fact

If Y is σ-subgaussian, and c is constant, then cY is $c\sigma$-subgaussian.

Fact

If $X_1, \ldots X_n$ are σ_i-subgaussian, and independent then $\sum X_i$ is $\sqrt{\sum \sigma_i^2}$-subgaussian.

Proof.

$$\phi_{\sum X_i}(\lambda) = \sum \phi_{X_i}(\lambda) \leq \sum \sigma_i^2\lambda^2/2 = (\sum \sigma_i^2)\frac{\lambda^2}{2}$$

Subgaussian random variable

When $\mathbb{E}Y = 0$, the following are equivalent (up to constant factor scaling):

- Y is σ-subgaussian, i.e. $\phi_Y(\lambda) = \ln \mathbb{E}e^{\lambda Y} \leq \frac{\lambda^2}{2}$. *MGF bound*
- For all λ we have $\mathbb{P}(|Y| > \sigma\lambda) \lesssim \exp(-\lambda^2/4)$. *Tail bounds*
- For all $p \geq 1$ we have $\|Y\|_p \leq \sqrt{p}\sigma$. *Moment bounds*

Why they are equivalent?

- MGF bound imply moment bound:
  $\lambda^p \frac{\mathbb{E}X^p}{p!} \leq \sum_p \lambda^p \frac{\mathbb{E}X^p}{p!} = \Phi_X(\lambda) \leq \exp(\lambda^2)$.
  This yields $\|X\|_p \lesssim \frac{p}{\lambda} \exp(\lambda^2/p)$.
  Take $\lambda \simeq \sqrt{p}$ to get $\|X\|_p \lesssim \sqrt{p}$.

Why they are equivalent?

- MGF bound imply moment bound:
  $\lambda^p \frac{\mathbb{E}X^p}{p!} \le \sum_p \lambda^p \frac{\mathbb{E}X^p}{p!} = \Phi_X(\lambda) \le \exp(\lambda^2)$.
  This yields $\|X\|_p \lesssim \frac{p}{\lambda} \exp(\lambda^2/p)$.
  Take $\lambda \simeq \sqrt{p}$ to get $\|X\|_p \lesssim \sqrt{p}$.
- Moment bounds imply MGF bounds (assuming $\mathbb{E}X = 0$):
  $\Phi_X(\lambda) = \sum \frac{\lambda^p}{p!} \mathbb{E}X^p \le 1 + \sum_{p \ge 2} \left(\frac{\lambda^2}{p}\right)^{p/2} \le \exp(C\lambda^2)$.

Why they are equivalent?

- MGF bound imply moment bound:
  $\lambda^p \frac{\mathbb{E}X^p}{p!} \le \sum_p \lambda^p \frac{\mathbb{E}X^p}{p!} = \Phi_X(\lambda) \le \exp(\lambda^2)$.
  This yields $\|X\|_p \lesssim \frac{p}{\lambda} \exp(\lambda^2/p)$.
  Take $\lambda \simeq \sqrt{p}$ to get $\|X\|_p \lesssim \sqrt{p}$.
- Moment bounds imply MGF bounds (assuming $\mathbb{E}X = 0$):
  $\Phi_X(\lambda) = \sum \frac{\lambda^p}{p!} \mathbb{E}X^p \le 1 + \sum_{p \ge 2} \left(\frac{\lambda^2}{p}\right)^{p/2} \le \exp(C\lambda^2)$.
- Moment bounds imply tail bound: Chebyshev inequality:
  $\mathbb{P}(|X| > \lambda) < \frac{\mathbb{E}|X|^p}{\lambda^p} \le \left(\frac{\sqrt{p}}{\lambda}\right)^p$.

Why they are equivalent?

- MGF bound imply moment bound:
  $\lambda^p \frac{\mathbb{E}X^p}{p!} \le \sum_p \lambda^p \frac{\mathbb{E}X^p}{p!} = \Phi_X(\lambda) \le \exp(\lambda^2)$.
  This yields $\|X\|_p \lesssim \frac{p}{\lambda} \exp(\lambda^2/p)$.
  Take $\lambda \simeq \sqrt{p}$ to get $\|X\|_p \lesssim \sqrt{p}$.
- Moment bounds imply MGF bounds (assuming $\mathbb{E}X = 0$):
  $\Phi_X(\lambda) = \sum \frac{\lambda^p}{p!} \mathbb{E}X^p \le 1 + \sum_{p \ge 2} \left(\frac{\lambda^2}{p}\right)^{p/2} \le \exp(C\lambda^2)$.
- Moment bounds imply tail bound: Chebyshev inequality:
  $\mathbb{P}(|X| > \lambda) < \frac{\mathbb{E}|X|^p}{\lambda^p} \le \left(\frac{\sqrt{p}}{\lambda}\right)^p$.
  Pick $\lambda := \sqrt{p}/2$, so that $\left(\frac{\sqrt{p}}{\lambda}\right)^p = \left(\frac{1}{2}\right)^{4\lambda^2} \le \exp(-C\lambda^2)$

Why they are equivalent?

- MGF bound imply moment bound:
  $\lambda^p \frac{\mathbb{E}X^p}{p!} \le \sum_p \lambda^p \frac{\mathbb{E}X^p}{p!} = \Phi_X(\lambda) \le \exp(\lambda^2)$.
  This yields $\|X\|_p \lesssim \frac{p}{\lambda} \exp(\lambda^2/p)$.
  Take $\lambda \simeq \sqrt{p}$ to get $\|X\|_p \lesssim \sqrt{p}$.
- Moment bounds imply MGF bounds (assuming $\mathbb{E}X = 0$):
  $\Phi_X(\lambda) = \sum \frac{\lambda^p}{p!} \mathbb{E}X^p \le 1 + \sum_{p \ge 2} \left(\frac{\lambda^2}{p}\right)^{p/2} \le \exp(C\lambda^2)$.
- Moment bounds imply tail bound: Chebyshev inequality:
  $\mathbb{P}(|X| > \lambda) < \frac{\mathbb{E}|X|^p}{\lambda^p} \le \left(\frac{\sqrt{p}}{\lambda}\right)^p$.
  Pick $\lambda := \sqrt{p}/2$, so that $\left(\frac{\sqrt{p}}{\lambda}\right)^p = \left(\frac{1}{2}\right)^{4\lambda^2} \le \exp(-C\lambda^2)$
- Tail bounds imply moment bounds:
  $\mathbb{E}|X|^p \le \int_0^\infty \lambda^{p-1} \mathbb{P}(|X| > \lambda)\, \mathrm{d}\lambda \le \int_0^\infty \lambda^{p-1} \exp(-\lambda^2)$
  $\le (Cp)^{p/2}$.

Hoeffding inequality

Theorem

If X_i are bounded $|X_i| \leq a_i$, and independent, then $\mathbb{P}(|\sum X_i - \mathbb{E}\sum X_i| > \|a\|_2 \lambda) < e^{-C\lambda^2}$.

Proof.

Take $Y_i = X_i - \mathbb{E}X_i$. We have $\|Y_i\|_p \leq \|Y_i\|_\infty = a_i \leq \sqrt{p} a_i$, therefore each Y_i is a_i-subgaussian. This implies that $\sum Y_i$ is $\|a\|_2$-subgaussian, which yields the desired tail bounds. $\square$

Going back to Johnson-Lindenstrauss for a bit

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \qquad \mathbb{P}_{T \sim \mathcal{D}}\left(\left| \|Ty\|_2^2 - \|y\|_2^2 \right| > \varepsilon\right) < \delta.$$

Going back to Johnson-Lindenstrauss for a bit

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \qquad \mathbb{P}_{T \sim \mathcal{D}}\left(\left| \|Ty\|_2^2 - \|y\|_2^2 \right| > \varepsilon\right) < \delta.$$

Take matrix T such that $T_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, and without loss of generality $\|y\|_2 = 1$. We have:

$$\|Ty\|_2^2 = \sum_i \langle T_i, y \rangle^2$$

Since $T_{ij} = \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, we know that $\langle T_i, y \rangle$ has the same distribution as $\frac{1}{\sqrt{m}} Z_i$ for independent gaussian Z_i.

Going back to Johnson-Lindenstrauss for a bit

Theorem (Distributional JL)

For any $d, \varepsilon, \delta > 0$, there is $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over linear maps $\mathbb{R}^d \to \mathbb{R}^m$, such that

$$\forall y \in \mathbb{R}^d, \qquad \mathbb{P}_{T \sim \mathcal{D}}\left(\left| \|Ty\|_2^2 - \|y\|_2^2 \right| > \varepsilon\right) < \delta.$$

Take matrix T such that $T_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, and without loss of generality $\|y\|_2 = 1$. We have:

$$\|Ty\|_2^2 = \sum_i \langle T_i, y \rangle^2$$

Since $T_{ij} = \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, we know that $\langle T_i, y \rangle$ has the same distribution as $\frac{1}{\sqrt{m}} Z_i$ for independent gaussian Z_i.
Consider $Z_1, Z_2, \ldots Z_m$ independent gaussian random variables. What can we say about concentration of $\sum_{i \le m} \frac{1}{m} Z_i^2$ around its mean?

Square of a gaussian is not subgaussian.

Note that for large λ we have

$$\mathbb{P}(Z^2 - \mathbb{E}Z^2 > \lambda) \simeq \mathbb{P}(Z^2 > \lambda) = \mathbb{P}(|Z| > \sqrt{\lambda}) \simeq \exp(-\lambda),$$

this is much larger than $\exp(-\lambda^2)$.

Subgamma random variables

Definition

A mean zero random variable X is (σ, B)-subgamma if $\phi_X(\lambda) \leq \sigma^2\lambda^2$ for every $\lambda \leq \frac{1}{B}$.

Subgamma random variables

Definition

A mean zero random variable X is (σ, B)-subgamma if $\phi_X(\lambda) \leq \sigma^2\lambda^2$ for every $\lambda \leq \frac{1}{B}$.

Fact

If $X_1, X_2, \ldots X_n$ are independent and (σ_i, B_i)-subgamma, then $\sum X_i$ is $(\|\sigma\|, \max_i B_i)$-subgamma.

MGF-Moment-Tail equivalence for subgamma random variables

When $\mathbb{E}X = 0$,

- If X is (σ, B)-subgamma, then $\|X\|_p \leq \sigma\sqrt{p} + Bp$.

MGF-Moment-Tail equivalence for subgamma random variables

When $\mathbb{E}X = 0$,

- If X is (σ, B)-subgamma, then $\|X\|_p \leq \sigma\sqrt{p} + Bp$.
- If X is (σ, B)-subgamma, then $\mathbb{P}(|X| > \lambda) < \exp(-\lambda^2/\sigma^2) + \exp(-\lambda/B)$.

MGF-Moment-Tail equivalence for subgamma random variables

When $\mathbb{E}X = 0$,

- If X is (σ, B)-subgamma, then $\|X\|_p \leq \sigma\sqrt{p} + Bp$.
- If X is (σ, B)-subgamma, then $\mathbb{P}(|X| > \lambda) < \exp(-\lambda^2/\sigma^2) + \exp(-\lambda/B)$.
- If $\|X\|_p \leq \sigma\sqrt{p} + Bp$ for all p, then X is $(C(\sigma + B), CB)$-subgamma.

MGF-Moment-Tail equivalence for subgamma random variables

When $\mathbb{E}X = 0$,

- If X is (σ, B)-subgamma, then $\|X\|_p \leq \sigma\sqrt{p} + Bp$.
- If X is (σ, B)-subgamma, then $\mathbb{P}(|X| > \lambda) < \exp(-\lambda^2/\sigma^2) + \exp(-\lambda/B)$.
- If $\|X\|_p \leq \sigma\sqrt{p} + Bp$ for all p, then X is $(C(\sigma + B), CB)$-subgamma.
- If $\mathbb{P}(|X| > \lambda) \leq \exp(-\lambda^2/\sigma^2) + \exp(-\lambda/B)$ for all p, then $\|X\|_p \lesssim \sigma\sqrt{p} + Bp$ for all p.

Motivating example

If $\mathbb{E}X_i = 0$, $|X_i| < B_i$ and $\mathbb{E}X_i^2 = \sigma_i^2$, then X_i is (σ_i, B_i)-subgamma. (By MGF bound).

Theorem (Bernstein inequality)

If $\mathbb{E}X_i = 0$, $|X_i| < B$ with probability 1, and $\mathbb{E}X_i^2 \leq \sigma_i^2$, then

$$\mathbb{P}(|\sum X_i| > \lambda) < \exp(-\lambda^2/\|\sigma\|_2^2) + \exp(-\lambda/B).$$

Motivating example

If $\mathbb{E}X_i = 0$, $|X_i| < B_i$ and $\mathbb{E}X_i^2 = \sigma_i^2$, then X_i is (σ_i, B_i)-subgamma. (By MGF bound).

Theorem (Bernstein inequality)

If $\mathbb{E}X_i = 0$, $|X_i| < B$ with probability 1, and $\mathbb{E}X_i^2 \leq \sigma_i^2$, then

$$\mathbb{P}(|\sum X_i| > \lambda) < \exp(-\lambda^2/\|\sigma\|_2^2) + \exp(-\lambda/B).$$

Example

Consider $X_i = 1$ with probability $\frac{1}{n}$, and 0 otherwise.
Then all $\sigma_i = \mathcal{O}(\frac{1}{\sqrt{n}})$ and $B = 1$, we have $Var(\sum X_i) \simeq 1$, and $\mathbb{P}(|\sum X_i - 1| > \lambda) \lesssim \exp(-C\lambda)$.

Motivating example

If $\mathbb{E}X_i = 0$, $|X_i| < B_i$ and $\mathbb{E}X_i^2 = \sigma_i^2$, then X_i is (σ_i, B_i)-subgamma. (By MGF bound).

Theorem (Bernstein inequality)

If $\mathbb{E}X_i = 0$, $|X_i| < B$ with probability 1, and $\mathbb{E}X_i^2 \leq \sigma_i^2$, then

$$\mathbb{P}(|\sum X_i| > \lambda) < \exp(-\lambda^2/\|\sigma\|_2^2) + \exp(-\lambda/B).$$

Example

Consider $X_i = 1$ with probability $\frac{1}{n}$, and 0 otherwise.
Then all $\sigma_i = \mathcal{O}(\frac{1}{\sqrt{n}})$ and $B = 1$, we have $Var(\sum X_i) \simeq 1$, and $\mathbb{P}(|\sum X_i - 1| > \lambda) \lesssim \exp(-C\lambda)$.
If we only used Hoeffding inequality we could only deduce $\mathbb{P}(|\sum X_i - 1| > \lambda\sqrt{n}) < \exp(-\lambda^2)$.

Back to Johnson-Lindenstrauss

Where were we?

Take $Z_i \sim \mathcal{N}(0,1)$ independent, and $Y_i := Z_i^2 - \mathbb{E}Z_i^2$. We want to show the concentration result $\mathbb{P}(|\sum_{i \leq m} Y_i| > m\varepsilon) \leq \delta$, when $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$.

Back to Johnson-Lindenstrauss

Where were we?

Take $Z_i \sim \mathcal{N}(0,1)$ independent, and $Y_i := Z_i^2 - \mathbb{E}Z_i^2$. We want to show the concentration result $\mathbb{P}(|\sum_{i\le m} Y_i| > m\varepsilon) \le \delta$, when $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$.

Proof.

For $Z_i \sim \mathcal{N}(0,1)$, we have $Y_i := Z_i^2 - 1$ is $(1,1)$-subgamma (tail bound is easy to verify). This means that $\sum Y_i$ is $(\sqrt{m},1)$-subgamma, and therefore

$$\mathbb{P}(|\sum Y_i| > m\varepsilon) < \exp(-C\varepsilon^2 m) + \exp(-Cm\varepsilon) \lesssim \exp(-Cm\varepsilon^2).$$

Back to Johnson-Lindenstrauss

Where were we?

Take $Z_i \sim \mathcal{N}(0,1)$ independent, and $Y_i := Z_i^2 - \mathbb{E}Z_i^2$. We want to show the concentration result $\mathbb{P}(|\sum_{i\le m} Y_i| > m\varepsilon) \le \delta$, when $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$.

Proof.

For $Z_i \sim \mathcal{N}(0,1)$, we have $Y_i := Z_i^2 - 1$ is $(1,1)$-subgamma (tail bound is easy to verify). This means that $\sum Y_i$ is $(\sqrt{m},1)$-subgamma, and therefore

$$\mathbb{P}(|\sum Y_i| > m\varepsilon) < \exp(-C\varepsilon^2 m) + \exp(-Cm\varepsilon) \lesssim \exp(-Cm\varepsilon^2).$$

Now we can pick $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$ so that $Cm\varepsilon^2 = \ln \delta^{-1}$, and we have

$$\mathbb{P}(|\sum Y_i| > m\varepsilon) < \delta.$$

□

We proved something stronger: do not need gaussian entries!

Theorem (Distributional JL)

For a random matrix $T \in \mathbb{R}^{m \times d}$, such that entries T_ij are independent, $\mathbb{E}T_{ij} = 0$ and $\mathbb{E}T_{ij}^2 = 1$ and T_{ij} are $\mathcal{O}(1)$-subgaussian, and any vector $v \in \mathbb{R}^d$, we have

$$\mathbb{P}\left(\left|\frac{1}{m}\|Tv\|_2^2 - \|v\|_2^2\right| > \varepsilon\|v\|_2^2\right) < \delta,$$

where $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$.

We proved something stronger: do not need gaussian entries!

Theorem (Distributional JL)

For a random matrix $T \in \mathbb{R}^{m \times d}$, such that entries T_ij are independent, $\mathbb{E}T_{ij} = 0$ and $\mathbb{E}T_{ij}^2 = 1$ and T_{ij} are $\mathcal{O}(1)$-subgaussian, and any vector $v \in \mathbb{R}^d$, we have

$$\mathbb{P}\left(\left|\frac{1}{m}\|Tv\|_2^2 - \|v\|_2^2\right| > \varepsilon\|v\|_2^2\right) < \delta,$$

where $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$.

Example

We can take $T_{ij} = \pm 1$ independent coin flips. This matrix is much easier to compute with, than a gaussian matrix.