Lecture 7 Concerntration

IEMS 402 Statistical Learning

Northwestern

Asymptotic VS Non-Asymptotic

Drawback of Asymptotic Theory

Asymptotic : $\sqrt{n}(T_n - \theta^*) \xrightarrow{n \to \infty} T$ $\quad N(0, I_{\theta^*}^{-1})$

$\theta \in \mathbb{R}^d$: d is high. Total Variance $\propto d$ $\Rightarrow$ $\sqrt{\frac{d}{n}}$ ← Final Error

In most of the case. $d \propto n$, then $\sqrt{\frac{d}{n}}$ is $O(1)$

$O(1)$ distance to convergence.

Concerntration

First sense of Concerntration

inequalities of the form

#n

Randomly sampled data

$\phi(n,t)$

$$\boxed{\mathbb{P}}(X \geq t) \leq \underline{\phi(t)}$$

where ϕ goes to zero (quickly) as $t \to \infty$

Error/risk

First examples

Proposition (Markov's inequality)

If $X \geq 0$, *then* $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$ *for all* $t \geq 0$.

$$\mathbb{P}(X \geq t) = \mathbb{P}(X^2 \geq t^2) \leq \frac{\mathbb{E}[X^2]}{t^2}$$

(Markov's Inequality)

different convergence rate w.r.t t

First examples

Proposition (Markov's inequality)

If $X \geq 0$*, then* $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$ *for all* $t \geq 0$*.*

Proposition (Chebyshev's inequality)

For any $t \geq 0$, $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$

First examples

Proposition (Markov's inequality)

If $X \geq 0$, *then* $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{\boxed{t}}$ *for all* $t \geq 0$.

Proposition (Chebyshev's inequality)

For any $t \geq 0$, $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{\boxed{t^2}}$

$\frac{1}{t^3}$

Should be $O(e^{-t})$?

Moment Generating Function

moment generating function

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

A function

Argument of moment generating function

$$e^x = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \ldots$$

control the reweighting the polynomials.

Moment Generating Function

moment generating function

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

A function

Argument of moment generating function

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

$Y = X_1 + X_2$

$M_Y(t) = M_{X_1}(t)\, M_{X_2}(t)$

Sum of Independent Random Variables:

Suppose X_1, X_2, ..., X_n are n independent random variables, and the random variable Y is defined as

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then,

$$\begin{aligned} M_Y(s) &= E[e^{sY}] \\ &= E[e^{s(X_1+X_2+\cdots+X_n)}] \\ &= E[e^{sX_1} e^{sX_2} \cdots e^{sX_n}] \\ &= E[e^{sX_1}]E[e^{sX_2}] \cdots E[e^{sX_n}] \quad \text{(since the } X_i\text{'s are independent)} \\ &= M_{X_1}(s) M_{X_2}(s) \cdots M_{X_n}(s). \end{aligned}$$

Chernoff bound

$$P(X \geq a) \leq \inf_{t>0} M(t) e^{-ta}$$

$\Leftarrow$ $\mathbb{P}(X \geq a)$

$M(t) \downarrow \mathbb{E}[e^{tX}]$

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

$$= M(t) e^{-ta}$$

Chernoff bound

$$P(X \geq a) \leq \inf_{t>0} M(t)e^{-ta}$$

Reason 1 σ^2 is "Variance"

<u>sub-Gaussian random variable</u>

A mean-zero random variable X is σ^2-*sub-Gaussian* if

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Moment Generating function of a Gaussian Var is σ^2

reason 2

$X_1 + X_2 = Y$ is $\sigma_1^2 + \sigma_2^2$ sub-Gaussian

$M_{X_1}(t) M_{X_2}(t) = M_Y(t)$

$\wedge|$ $\quad$ $\wedge|$

$\exp\left(\frac{t^2\sigma_1^2}{2}\right)$ $\quad$ $\exp\left(\frac{t^2\sigma_2^2}{2}\right)$

$\shortparallel$

$\exp\left(\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right)$

Example

If $X \in [a, b]$, then

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

$\sigma = \frac{1}{4}(b-a)^2$

Exercise

Chernoff bound

$$P(X \geq a) \leq \inf_{t>0} M(t) e^{-ta}$$

<u>sub-Gaussian random variable</u>

A mean-zero random variable X is σ^2-*sub-Gaussian* if

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Hoeffding Inequality

$X_1, \dots\dots, X_n$ is $(\sum_{i=1}^n \sigma_i^2)$- sub gaussian.

Corollary (Hoeffding bounds)

If X_i are independent σ_i^2-sub-Gaussian random variables,

set probability to be one $O(1)$, $t = O(\frac{1}{\sqrt{n}})$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{nt^2}{\frac{2}{n}\sum_{i=1}^n \sigma_i^2}\right).$$

this is a constant

Should be $O(1/\sqrt{n})$?

$$\|$$

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geq nt\right) \leq \exp\left(-\frac{(nt)^2}{\sum_{i=1}^n \sigma_i^2}\right) = \exp\left(-\frac{nt^2}{\frac{2}{n}\sum_{i=1}^n \sigma_i^2}\right)$$

- usually stated as $X_i \in [a, b]$, so bound is $\exp(-\frac{2nt^2}{(b-a)^2})$

Moment Generating Function is Powerful

Not Required

Bernstein's Inequality

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \vee \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq -t\right) \leq \exp\left(-\frac{nt^2}{\boxed{2\sigma^2} + 2ct/3}\right),$$

different thing

σ^2 : variance $\quad |X_i| \leq c$

Special case: σ is 0

$$\exp\left(-\frac{nt^2}{ct}\right)$$

$$= \exp\left(-\frac{nt}{c}\right) \to t = O\left(\frac{1}{n}\right)$$

Homework 5, Question 3

Moment Generating Function is Powerful

Not Required

Proposition

Let $\{Z_i\}_{i=1}^N$ be σ^2-sub-Gaussian (not necessarily independent). Then

$$\mathbb{E}\left[\max_i Z_i\right] \leq \sqrt{2\sigma^2 \log N}.$$

$\Rightarrow$ max of n-random variables is log N !!

$\exp(\sqrt{2\sigma^2 \log N})$

$\Downarrow$

$$\exp\left(\mathbb{E}\left[\max_i Z_i\right]\right) \leq \mathbb{E}\left[\exp\left(t \max_i Z_i\right)\right]$$

$\Uparrow$

$\wedge\!|$

$$\mathbb{E}\left[\sum_{i=1}^{n} \exp(t Z_i)\right] = O(N)$$

Application

Johnson-Lindenstrauss Lemma

Lemma For any $0 < \epsilon < 1$ and any interger n let k be a possitive interger such that

error

ϵ is the error

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n \qquad (2)$$

log of max of #(i,j)-pairs r.v.

number of projection

then for any set A of n points $\in \Re^d$ there exists a map $f : \Re^d \to \Re^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2 \qquad (3)$$

prob p fails of a (i,j) pair

⇓

$\leq O(n^2 p)$ to fail the whole problem ⇒ set p to be $O(\frac{1}{n^2})$

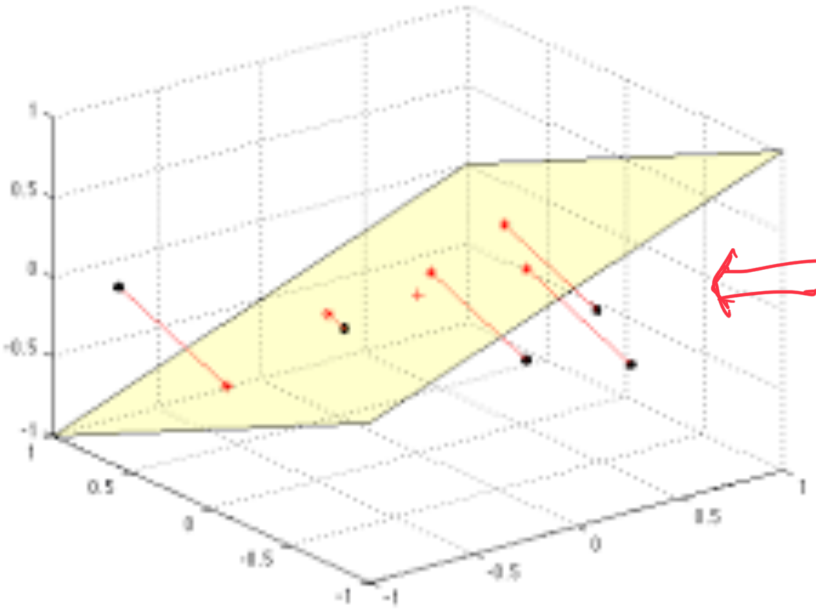
$\|R(x_i - x_j)\| \approx \|x_i - x_j\|$

How many (i,j) pairs? $O(n^2)$

https://cs.stanford.edu/people/mmahoney/cs369m/Lectures/lecture1.pdf

Why it's important

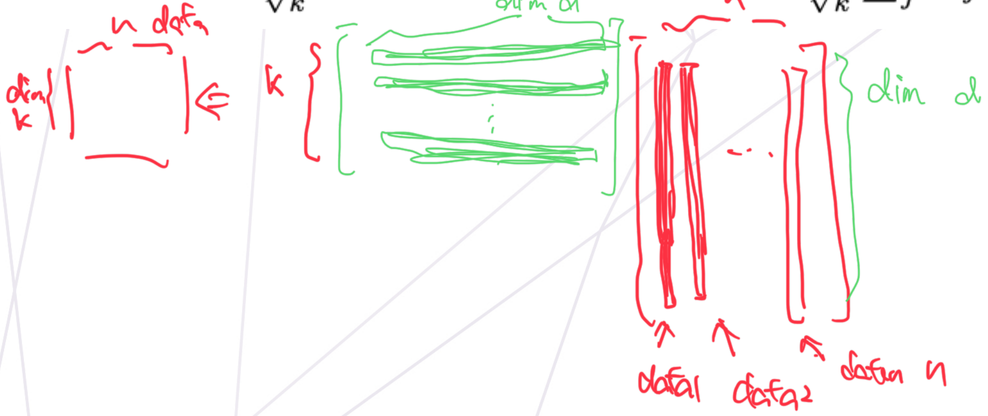
CIFAR 100: 6000 32x32images,



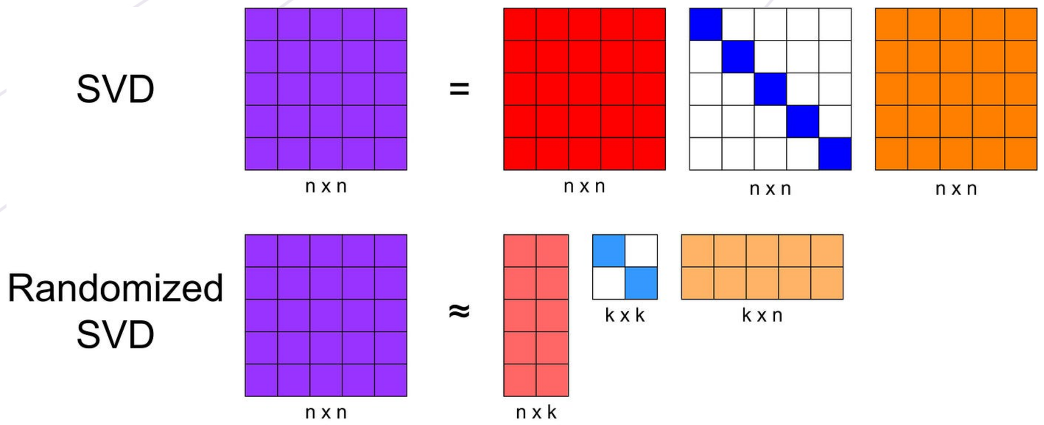
f- linear

Idea: random projection

Definition Let R be a random matrix of order $k \times d$ i.e $R_{ij} \overset{i.i.d}{\sim} N(0,1)$ and u be any fixed vector $\in \Re^d$. Define $v = \frac{1}{\sqrt{k}} R \cdot u$. Thus $v \in \Re^k$ and $v_i = \frac{1}{\sqrt{k}} \sum_j R_{ij} u_j$



Why it's important



Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." SIAM review

Idea: random projection

Definition Let R be a random matrix of order $k \times d$ i.e $R_{ij} \overset{i.i.d}{\sim} N(0,1)$ and u be any fixed vector $\in \Re^d$. Define $v = \frac{1}{\sqrt{k}} R \cdot u$. Thus $v \in \Re^k$ and $v_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} R_{ij} u_j$

normalization

Fact 1. $\mathbb{E}[\|v\|^2] = \|u\|^2$ $\rightarrow$ $u \in \mathbb{R}^d,\ r \in \mathbb{R}^d\ \ r \sim N(0, I)$ covariance

$\mathbb{E}[(u \cdot r)^2] = \mathbb{E}[u^T r r^T u] = u^T \mathbb{E}[r r^T] u = u^T I u = u^T u = \|u\|^2$

$\boxed{\left(\sum_{i=1}^{d} u_i \cdot r_i\right)^2} \Rightarrow N(0, (u_1^2 + u_2^2 + \cdots + u_d^2)) \Rightarrow \mathbb{E}[(u \cdot r)^2] = \|u\|^2$

$N(0,1)$

Question. $\mathbb{P}(\|v\|^2 \geq (1+\epsilon)\|u\|^2)$ Assume $\|u\| = 1$

Random projection

k of
1-dim
projection

Question. $\mathbb{P}(\|v\|^2 \geq (1+\epsilon)\|u\|^2)$

$$x_i = R_i^\top \cdot u$$

Means $\dfrac{\sum_{i=1}^{k} x_i^2}{k} \geq (1+\epsilon)$

Random projection

Question. $\mathbb{P}(\boxed{\|v\|^2 \geq (1+\epsilon)\|u\|^2})$

$$x_i = R_i^\top \cdot u$$

$x = \sum_{i=1}^{k} x_i^2$

Means $\dfrac{\sum_{i=1}^{k} x_i^2}{k} \geq (1+\epsilon) \rightarrow \boxed{e^{\lambda x}} \geq e^{\lambda(1+\epsilon)k}$

$$\mathbb{E}[e^{\lambda x}] = \prod_{i=1}^{k} \mathbb{E}[e^{\lambda x_i^2}] = \left(\mathbb{E}[e^{\lambda x_i^2}]\right)^k$$

Random projection

Question. $\mathbb{P}(\boxed{\|v\|^2 \geq (1+\epsilon)\|u\|^2})$

$x_i = R_i^\top \cdot u$

Means $\dfrac{\sum_{i=1}^k x_i^2}{k} \geq (1+\epsilon) \to \boxed{e^{\lambda x}} \geq e^{\lambda(1+\epsilon)k}$

$$\mathbb{E}[e^{\lambda x}] = \prod_{i=1}^{k} \mathbb{E}[e^{\lambda x_i^2}] = \left(\mathbb{E}[e^{\lambda x_i^2}]\right)^k$$

$\mathbb{E}[e^{\lambda x_i}]$ is the moment generating function of a χ^2.

Thus $\mathbb{P}[e^{\lambda(1+\epsilon)k}] \leq \left(\dfrac{1}{\sqrt{1-2\lambda}}\right)^k \cdot \dfrac{1}{e^{\lambda(1+\epsilon)k}}$

Random projection

Question. $\mathbb{P}(\|v\|^2 \geq (1+\epsilon)\|u\|^2)$ $\leq e^{-(\epsilon^2/2-\epsilon^3)k/2}$

$x_i = R_i^\top \cdot u$

Means $\frac{\sum_{i=1}^k x_i^2}{k} \geq (1+\epsilon) \to \boxed{e^{\lambda x}} \geq e^{\lambda(1+\epsilon)k}$

$$\mathbb{E}[e^{\lambda x}] = \prod_{i=1}^{k} \mathbb{E}[e^{\lambda x_i}] = \left(\mathbb{E}[e^{\lambda x_i}]\right)^k$$

Thus $\mathbb{P}[e^{\lambda(1+\epsilon)k}] \leq \left(\frac{1}{\sqrt{1-2\lambda}}\right)^k \cdot \frac{1}{e^{\lambda(1+\epsilon)k}}$ $\xrightarrow{\text{set } \lambda = \frac{\epsilon}{2(1+\epsilon)}}$ $\leq e^{-(\epsilon^2/2-\epsilon^3)k/2}$ $\boxed{\leq n^{-2}}$

Why?

Uniform bound!

Note

Not Required

another proof using epsilon-net: **Theorem 8.**
https://www.cs.princeton.edu/~smattw/Teaching/Fa19Lectures/lec9/lec9.pdf