Learning

Given $\{(x_i, y_i)\}_{i=1}^n$ "Training Data"

$x_i \in \mathbb{X}$: "input" features, representations, covariates, var. ind.

$y_i \in \mathbb{Y}$: "output" targets, labels, dependent variable.

Goal: Find a map $f: \mathbb{X} \to \mathbb{Y}$

{if $\mathbb{X} := \{\text{null}\}$ "unsupervised"}

Generalization Error (Lower is better) $= \int_{\mathbb{X}\times\mathbb{Y}} \Psi(\hat{f}(x), y)\, d\mu(x,y)$ — μ: Test distribution; $\Psi: \mathbb{Y}\times\mathbb{Y} \to \mathbb{R}$ (error metric)

High Dimensional Analysis

$\mathbb{X} = \mathbb{R}^{d_n}$ $\mathbb{Y} = \mathbb{R}^{c_n}$

$d_n, c_n \nearrow n$ e.g. $d_n = \beta n$ "proportional regime" $\beta \in (0, \infty)$

Operator Norm

$$\|M\|_{op} := \sup_{\vec{v}\neq 0} \frac{\|M\vec{v}\|_2}{\|\vec{v}\|_2}$$

$= \max_i |\lambda_i(M)|$ if M is symmetric.

Problem

$y_1, y_2, \cdots, y_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma_{d_n})$ $\vec{y}_i \in \mathbb{R}^{d_n}$ Jt. Gaussian

Estimate Σ_{d_n}

$\hat{\Sigma}_n = \langle y_i y_i^T \rangle \{= \frac{1}{n}\sum_{i=1}^n y_i y_i^T\}$ — estimator of Σ_{d_n}

Classical Regime: $d_n = d_0$ fixed. $\hat{\Sigma}_n \to \Sigma_{d_0}$ a.s.

if $d_n = \beta n$ $\hat{\Sigma}_n$ is random

Special Case if $\Sigma_{d_n} = I_{d_n}$

then $\lim_{n\to\infty} \|\hat{\Sigma}_n - \Sigma_{d_n}\|_{op} = \beta + 2\sqrt{\beta}$ a.s. for $\beta \in (0,1)$

For a symmetric positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$

Spectral Measure

$$F_A(t) = \frac{1}{d}\sum_{i=1}^d 1\{\lambda_i(A) \le t\}$$

Marchenko Pastur Theorem

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix} \in \mathbb{R}^{n\times d} \quad (\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n) \sim \mathcal{N}(0, I_{d\times d}), \quad d = \beta n$$

Then $F_{\hat{\Sigma}} \to F$

PDF of location of EV of $\hat{\Sigma}$ as $n\to\infty$

$$\frac{dF(s)}{ds} = \frac{\sqrt{(b-s)(s-a)}}{2\pi\beta t}, \quad b = (1+\sqrt{\beta})^2, \quad a = (1-\sqrt{\beta})^2, \quad 0 < \beta < 1$$

Pre/Post multiplying by diagonal matrix scales rows/columns

Regression (1D)

$\{(x_i, y_i)\}_{i=1}^n$ $f(x) = \beta x + c$

1D OLS (ordinary least squares) $(\hat{\beta}, \hat{c}) = \underset{(\beta,c)}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta x_i - c)^2$

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x_i, y_i)}{\text{Cov}(x_i)}, \quad \hat{c} = \bar{y} - \hat{\beta}\bar{x}$$

High Dimensions $f(x) = \beta^T x + c$ $\beta \in \mathbb{R}^d$

$$\hat{c} = \bar{y} - \hat{\beta}\bar{x} \qquad \bar{x} = \frac{\sum_{i=1}^n \vec{x}_i}{n}$$

$$\hat{\beta}_* = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \vec{y}$$

$$\bar{X}_{n\times d} = X_{n\times d} - \underset{\text{broadcast}}{\bar{x}_{1\times d}} = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_n - \bar{x})^T \end{bmatrix}_{n\times d}$$

can also be written as $(\hat{\beta}_*, \hat{c}) = \operatorname{argmin} \|\bar{X}\beta - \bar{y}\|_2^2$

$\hat{\beta}_*$ is not the only sol^n but least norm

$\hat{\beta} = \hat{\beta}_* + \alpha\beta_{null}$, $\beta_{null} \in \text{Null}(\bar{X})$

SVD $X_{n\times d} = U_{n\times n} S_{n\times d} V_{d\times d}^T$ — U, V: unitary; S: diagonal with non-ve values

Pseudo Inverse $X^+ = V S^+ U^T$ $\quad (S^+)_{ii} = \begin{cases} \frac{1}{s_{ii}} & \text{if } s_{ii} > 0 \\ 0 & s_{ii} = 0 \end{cases}$

Probability Measure: countable additive map from a σ-field of subsets of Ω to real numbers which assigns value 0 to the empty set.

Notation $P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\})$

'Probability that X is in A'

$X: (\Omega, \mathcal{F}) \mapsto (S, \mathcal{B}(S))$, $X^{-1}(A) \in \mathcal{F}$ $\forall A \in \mathcal{B}(S)$

is called a R.V. ($\approx$ a measurable map $(\Omega, \mathcal{F}) \mapsto (S, \underset{\sigma(S)}{\mathcal{S}})$)

$X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}$

$P(X \in A) := P(X^{-1}(A))$

1D CDF $F_X(t) = P(X \le t)$

a) Non-Decreasing b) $\lim_{t\to-\infty} F_X(t) = 0$ & $\lim_{t\to+\infty} F_X(t) = 1$ c) right continuous

However. Left limit exists: $P(X = a) = F_X(a) - \lim_{t\to a^-} F_X(t)$

Right $\lim_{t\to a^+} F_X(t) - F_X(a) = P(X \in (t, t+\varepsilon]) \overset{t\to a}{=} 0$

However in higher dimensions, it must also satisfy:

For every finite hypercube $A = (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d]$

$$P(A) = \sum_{v \in V(A)} \text{sgn}(v)\, F_X(v) \ge 0 \quad \text{where}$$

$V(A) = \{a_1, b_1\} \times \{a_2, b_2\} \times \cdots \times \{a_d, b_d\}$ are vertices of A.

$\text{sgn}(v) = (-1)^{\#\text{ of } a\text{'s in } v}$

For every $F: \mathbb{R}^d \to \mathbb{R}$ that satisfies these 4 conditions, $\exists$ a unique R.V. Furthermore, that R.V. satisfies

$$P(X \in A) = \sum_{v \in V(A)} \text{sgn}(v) F_X(v)$$

Median (m) $P(X \le m) \ge 1/2$ & $P(X \ge m) \ge \frac{1}{2}$, $\lim_{x \to m^-} F(x) \le \frac{1}{2} \le F(x)$

$F_X(t)$ may not be ONTO

Quantile Function $F_X^{-1}: [0,1] \to \mathbb{R}$ $F_X^{-1}(q) = \inf_{t \in \mathbb{R}} \{t : F_X(t) > q\}$

$F_X^{-1}(1/4) = 1^{st}$ quartile, $F_X^{-1}(3/4) = 3^{rd}$ quartile.

$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ for support(y): $\{y : f_Y(y) > 0\}$ $\int f_{X|Y}(x|y)dx = \begin{cases} 1 \\ 0 \end{cases}$ $\forall y \in \text{Supp}(Y)$

Expectation of R.V. $\mathbb{E}[X] = \int x dF_X(x) = \int x f_X(x) dx$ (if density exists)

Jensen's Inequality If ϕ is convex: $\phi(\alpha x_1 + (1-\alpha)x_2) \le \alpha\phi(x_1) + (1-\alpha)\phi(x_2)$

$\mathbb{E}[\phi(X)] \ge f(\mathbb{E}[X])$

If $p > q \ge 1$ ~~If~~ $\|X\|_p \ge \|X\|_q$ Proof uses Holder. $\|X\|_p = (\mathbb{E}[|X|^p])^{\frac{1}{p}}$

Holder if $\frac{1}{p} + \frac{1}{q} = 1$ $|\mathbb{E}[XY]| \le \|X\|_p \|Y\|_q$

k^{th} moment exists $\mathbb{E}|X|^k < \infty$ means smaller moments also exist

$\text{Var}(X) = \min_{c \in \mathbb{R}} \mathbb{E}((X-c)^2) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$

if $X \in \mathbb{R}^d$, $\mathbb{E}[X] \in \mathbb{R}^d$

Covariance Matrix / Variance Covar. Matrix $\mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}^T[X] = \mathbb{E}(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T$

Markov If $X \ge 0$ & $\mathbb{E}X < \infty$ $P(X \ge t) \le \frac{\mathbb{E}[X]}{t}$

Stronger Version Ramdas et al $U \perp\!\!\!\perp X$ & U uniform in $[0,1]$ then

$P(X \ge tU) \le \mathbb{E}[X]/t$

MGF does not always exist $M_X(t) = \mathbb{E}(e^{tX})$ $M_X^{(k)}(0) = \mathbb{E}[X^k]$

Characteristic F^n always exists $\phi_X(t) = \mathbb{E}\, e^{itX}$ $\phi_X(0) = 1$, $|\phi_X(t)| \le 1$ $\phi_X'(t) = j\,\mathbb{E}[Xe^{jXt}]$

Chebychev if $\mathbb{E}|X|^2 < \infty$ $P(|X - \mathbb{E}[X]| > t) \le \frac{\text{Var}(X)}{t}$

Chernoff $P(X \ge t) = P(uX \ge ut)$ if $u > 0$

$= P(e^{uX} \ge e^{ut}) \le \mathbb{E}[e^{uX}]/e^{ut}$

$\le \inf_{u>0} e^{-ut} M_X(u)$

Convergence of sequence of R.V. $X_1, X_2, \cdots, X_n$ dealing with joint $(\lim_{n\to\infty} X_n, X)$

① $X_n \to X$ a.s. if $\forall \varepsilon > 0$ $P(\lim_{n\to\infty} |X_n - X| > \varepsilon) = 0$

② $X_n \to X$ i.p. $X_n \xrightarrow{p} X$

if $\forall \varepsilon > 0$ $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$ joint (X_n, X)

a.s. $\Rightarrow$ i.p.

In statistics $\hat{\theta}_n$: estimate from n samples of θ^*

if $\hat{\theta}_n \xrightarrow{p} \theta^*$ "consistency"

↳ allowed to be Random.

WLLN $\{X_i\}_{i=1}^{\infty}$ i.i.d $F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^n F_{X_i}(x_i)$ independence

$= \prod_{i=1}^n F_X(x_i)$ identically distributed

with $\mathbb{E}[X_i] = \mu$

↳ $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mu$ (Random → const)

↳ Stronger version: for pairwise independence also holds.

$Y = X \oplus Z$, $X, Z \sim \text{Ber}\frac{1}{2}$. Pairwise but not jtly independent

SLLN $\{X_n\}$ pairwise independence. Let $X_1, X_2, \cdots, X_n$ be i.i.d. R.V. & assume $\mathbb{E}[|X|] < \infty$. $\mu = \mathbb{E}[X]$

$\sum_{i=1}^n X_i/n \to \mu$ a.s.

(Weak*) Convergence in distribution/Law $\{X_i\} \xrightarrow{d} X$

if $F_{X_i}(t) \to F_X(t) \Leftrightarrow \phi_{X_i}(t) \to \phi_X(t) \Leftrightarrow \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$

for all continuity pts of F_X — Pointwise convergence

CLT $\{X_i\}_{i=1}^{\infty}$ i.i.d. $\mathbb{E}|X_i|^2 < \infty$ $\mu = \mathbb{E}X_i$ $\sigma^2 = \text{Var}(X_1)$

$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0,1)$

↳ if it was $\sqrt{n}$ it would converge to 0 by SLLN

The standard inner product on $\mathbb{R}^{m\times n}$ is given by

$$\langle X, Y\rangle = \operatorname{tr}(X^T Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}$$

inner product of associated vectors in $\mathbb{R}^{mn}$

classmate Date Page

Statistics

data: $y_1, y_2, \cdots, y_n$ i.i.d. observations of R.V. $Y \sim P$ distribution

P_θ: Parametrised family of distribution

$\hat{\theta}_n$: "estimator of θ" $\hat{\theta}_n : \mathcal{Y}^n \to \Theta$,

$\theta \in \Theta$ "parameter space"

$\hat{\theta}_n = \hat{\theta}_n(Y_1, \cdots, Y_n)$ R.V, $E[\hat{\theta}_n]$: constant

Bias of $\hat{\theta}_n$ $= E[\hat{\theta}_n] - \theta^*$

Consider $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, $\hat{\sigma}_n^2 = \frac{1}{n}(Y_i - \hat{\mu}_n)^2$

For all distributions whose 1st & 2nd moment exists

$E[Y_i] = \mu \quad E[Y_i^2] = \sigma^2 + \mu^2$

Then $\hat{\mu}_n$ is unbiased estimator of μ

$\text{Bias}\left(\hat{\sigma}_n^2/n\right) = -\frac{\sigma^2}{n}$, $\text{Bias}\left(\frac{n}{n-1}\hat{\sigma}_n^2\right) = 0$

MSE($\hat{\theta}_n$) (not random) $= E[|\hat{\theta}_n - \theta^*|^2] = E[|\hat{\theta}_n - E\hat{\theta}_n + E\hat{\theta}_n - \theta^*|^2]$

$= E[|\hat{\theta}_n - E\hat{\theta}_n|^2] + E[|\theta^* - E\hat{\theta}_n|^2]$ (can drop E)

$= \text{Variance}(\hat{\theta}_n) + \text{Bias}^2(\hat{\theta}_n)$

Theorem If bias $\to 0$ & $\text{Var}(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n$ is consistent

$\because$ Conv. in Mean Square Sense $\Rightarrow$ Convergence i.p.

Exponential Family Distributions

$$f_Y(y) = e^{\langle \theta, T(y)\rangle - A(\theta)} h(y) = \frac{e^{\langle \theta, T(y)\rangle} h(y)}{Z(\theta)}$$

θ: natural param, $h(y)$: Base measure/ind. of θ

"s-param family" $\langle \theta, T(y)\rangle = \sum_{i=1}^s \theta_i T_i(y)$ $T_i : \mathcal{Y} \to \mathbb{R}$, $T : \mathcal{Y} \to \mathbb{R}^s$

T: sufficient statistic f^n

Z: Partition F^n / A: log Partition F^n

$\int f_Y(y)\,dy = 1 \Rightarrow \boxed{Z(\theta) = \int h(y) e^{\langle \theta, T(y)\rangle} dy}$

$\Theta = \{\theta : Z(\theta) < \infty\}$ Natural param space.

In $\mathcal{N}(\mu, \Sigma)$ Σ: CoVar Matrix, Σ^{-1}: Precision Matrix

$y^T \Sigma^{-1} y = \langle \text{flatten}(\Sigma^{-1}), \text{flatten}(yy^T)\rangle$

Properties $A(\theta) = \ln Z(\theta)$

1. $\nabla_\theta A(\theta) = E_\theta T(y)$
2. $\nabla_\theta^2 A(\theta)$ = Covariance Matrix of sufficient statistics

$$(\nabla_\theta^2 A(\theta))_{ij} = \frac{\partial}{\partial \theta_i}\left(\frac{\partial A(\theta)}{\partial \theta_j}\right) = \text{Cov}(T_i(Y), T_j(Y))$$

→ log Partition Function

$A(\theta)$ is a convex f^n of θ

Proof 1 | $\nabla_\theta^2 A(\theta) \geq 0$ Proof 2 | Using Hölder's Inequality

$$\int f(y)g(y)\,dy \leq \left(\int |f(y)|^p dy\right)^{1/p}\left(\int |g(y)|^q dy\right)^{1/q}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, $p \geq 1$

Likelihood Function $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$

Eg. Exponential Family i.i.d. data $Y_1, Y_2, \cdots, Y_n \overset{iid}{\sim} f_{Y,\theta}(y)$

$$\mathcal{L}(\theta) = \prod_{i=1}^n e^{\langle \theta, T(y_i)\rangle - A(\theta)} h(y_i)$$

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta\in\Theta} e^{\langle \theta, \sum_i T(y_i)\rangle - nA(\theta)}$$

$$= \operatorname{argmin}_{\theta\in\Theta} A(\theta) - \left\langle \theta, \frac{1}{n}\sum_{i=1}^n T(y_i)\right\rangle$$

Convex + Affine = Convex

Convex opt problem when Θ is convex set.

$\nabla A(\theta) - \frac{1}{n}\sum_{i=1}^n T(y_i) = 0 \Leftrightarrow E_{Y\sim f_{Y,\theta}} T(y) = \frac{1}{n}\sum_{i=1}^n T(y_i)$

↳ Unique Sol^n if $\nabla^2 A(\theta) > 0$ ($\Leftrightarrow T_1(Y), \cdots, T_s(Y)$ are linearly independent

If sufficient statistics are L.I.

Exponential Families are called minimal.

Conjugate of a Function:

$$f^*(y) = \sup_{x\in \text{dom} f} (y^T x - f(x))$$

↳ Convex $\forall$ if f not convex

$(f^*)^*$ = Biconjugate of f = f if f is convex, lower semi continuous

Dual Norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, Associated dual norm $\|\cdot\|_*$ is

$$\|z\|_* = \sup\{z^T x \mid \|x\| \leq 1\}$$

$z^T x \leq \|x\|\,\|z\|_*$

Dual of ℓ_p norm is ℓ_q norm $\frac{1}{p} + \frac{1}{q} = 1$

$$f(x) = \frac{\|x\|_p^p}{p} \qquad f^*(y) = \frac{\|y\|_q^q}{q}$$

$$f(x) = \|x\| \qquad f^*(y) = \begin{cases} 0 & \|y\|_* \leq 1 \\ \infty & \text{otherwise.} \end{cases}$$

$\{\lambda_i\}, \{\vec{v}_i\}$ EV of $A \in \mathbb{R}^{n\times n}$

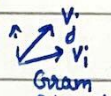
Properties of Symmetric Matrix

① $A = A^T$ ② All Eigen Values Real

③ $\lambda_i, \vec{v}_i$ & $\lambda_j, \vec{v}_j$ EV, EV pair

a) $\lambda_i \neq \lambda_j \Rightarrow \vec{v}_i^T \vec{v}_j = 0$

b) $\lambda_i = \lambda_j$, Eigenspace of λ_i = span(v_i, v_j)



WLOG $\{\vec{v}_i\}_{i=1}^n$ orthonormal $\quad v_i^T v_j = \mathbb{1}\{i=j\}$

$\forall\ \vec{x} \in \mathbb{R}^n,\ \vec{x} = \sum_{i=1}^n \alpha_i \vec{v}_i$ where $\alpha_i = v_i^T x$

$A\vec{x} = \sum_{i=1}^n \lambda_i \vec{x}\vec{v}_i^T \vec{v}_i = \left(\sum_{i=1}^n \lambda_i \vec{v}_i \vec{v}_i^T\right) x$

Eigen Decomposition of A

$A = \sum_{i=1}^n \lambda_i \vec{v}_i \vec{v}_i^T = \underset{n\times n}{V} \underset{n\times n}{\Lambda} \underset{n\times n}{V^T}$

$V = [\vec{v}_1\ \vec{v}_2\ \cdots\ \vec{v}_n]$ $\quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ convention λ_i descending order of modulii

$r = \text{rank}(A)$ $\quad \dim(\text{null}(A)) = n - r$

$V_{n\times n} = \left[\underbrace{V_r}_{n\times r}\ \underbrace{V_{n-r}}_{n\times(n-r)}\right]$ $\quad \{\vec{v}_i : \lambda_i \neq 0\}$ $\{\vec{v}_i \mid \lambda_i = 0\}$

$\Lambda = \begin{bmatrix} \Lambda_r & \\ & 0_{n-r} \end{bmatrix}$

$B = V\Lambda V^T = \underset{n\times r}{V_r} \underset{r\times r}{\Lambda_r} \underset{r\times n}{V_r^T} = \sum_{i=1}^r \lambda_i v_i v_i^T$ ← Non-zero EV

Every Symmetric Matrix has an Eigen Decomposition ✓

$(A^TA)_{n\times n}$ a) Symmetric

b) all EV are non-negative

c) If $(\lambda_i, \vec{v}_i)$ is Eigen Pair of A^TA, the (λ_i, Av_i) is EP of AA^T

$(AA^T)Av_i = A(A^TAv_i) = \lambda_i Av_i$, $\quad \vec{u}_i = \frac{A\vec{v}_i}{\sqrt{\lambda_i}}$ $\quad \|\vec{u}_i\| = 1$

$\|Av_i\| = \sqrt{\lambda_i}$

WLOG $\|v_i\|^2 = 1$

$A^TA\vec{v}_i = \lambda_i \vec{v}_i$

$\lambda_i = \|Av_i\|^2 / \|v_i\|^2$

$A^TA = V\Lambda_n V^T$ $\quad \Lambda_n = \begin{bmatrix} \Lambda_r & \\ & 0_{n-r} \end{bmatrix} \in \mathbb{R}^{n\times n}$

$AA^T = U\Lambda_m U^T$ $\quad \Lambda_m = \begin{bmatrix} \Lambda_r & \\ & 0_{m-r} \end{bmatrix} \in \mathbb{R}^{m\times m}$

$A^TA = V_r\Lambda_r V_r^T$ $\quad AA^T = U_r\Lambda_r U_r^T$

span(V_r) = range(A^TA) = range(A^T)

span(V_{n-r}) = null(A)

span(U_r) = range(A)

span(U_{m-r}) = null(A^T)

$x \in \mathbb{R}^n$: $X_{n(A)^\perp} + X_{n(A)} = \sum_{i=1}^r \alpha_i \vec{v}_i + \sum_{i=r+1}^n \alpha_i \vec{v}_i$

$Ax = \sum_{i=1}^r \alpha_i A\vec{v}_i = \sum_{i=1}^r \alpha_i \vec{u}_i \sqrt{\lambda_i}$

$AV = A[\vec{v}_1\ \vec{v}_2 \cdots \vec{v}_n]$

$= [\vec{u}_1\sqrt{\lambda_1},\ \vec{u}_2\sqrt{\lambda_2} \cdots \vec{u}_n\sqrt{\lambda_n}] = U_n \Lambda_n^{\frac{1}{2}}$

$A = U_n \Lambda_n^{\frac{1}{2}} V^T$ assume $m \geq n$ $\quad U_n$: 1st n columns of U

$A = \sum_{i=1}^r \sqrt{\lambda_i}\, \vec{u}_i \vec{v}_i^T = U_r \Lambda_r^{1/2} V_r^T$ "Economy version of SVD"

$\sqrt{\lambda_i}$: singular values of A ≥ 0

- Diagonal from right (left) scales columns (rows)

if $m < n$ work with A^T $\quad A^T = V_m \Lambda_m^{1/2} U^T$

? Projection onto span(A^T) $\quad V_r V_r^T = A^+A$

$\dot{x}(t) = -Ax(t)$ $\quad x(t) \in \mathbb{R}^{n\times 1}$ $\quad A \in \mathbb{R}^{n\times n}$, $A \in S_+^n$

$x(t) = e^{-At}x(0)$, $\quad A = V\Lambda V^T$

$\dot{x} = -V\Lambda V^T x \Rightarrow \dot{u} = -\Lambda u$ $\quad u := V^T x$

$u_i(t) = e^{-\lambda_i t} u_i(0)$ $\quad$ wlog $\lambda_1 > \lambda_2 > \cdots > \lambda_n$

slowest decaying $u_n(t) = e^{-\lambda_n t} u_n(0)$

$u^Tu = x^TVV^Tx = x^Tx$ $\quad \|x(t)\|_2 = \|u(t)\|_2 \leq e^{-\lambda_n t}\|u_n(0)\|$

$e^A = I + A + \frac{A^2}{2!} + \cdots = Ve^{\Lambda}V^T$ $\quad x(t) = e^{-At}x(0)$

$x(t) = Ve^{-\Lambda t}V^Tx(0)$

0 has multiplicity $n-r$, $\lambda_{r+1}, \cdots, \lambda_n$ $\quad u = V^Tx = \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} x$

$\lim_{t\to\infty} \|x(t)\|^2 = \sum_{k=r+1}^n u_k^2(0) = \|V_{n-r}^T x(0)\|^2$

↳ squared length of projection into $N(A)$

$x(t) = Ve^{-\Lambda t}V^Tx(0) = [V_r\ V_{n-r}] \begin{bmatrix} e^{-\Lambda_r t} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} x(0)$

$= V_r e^{-\Lambda_r t} V_r^T x(0) + V_{n-r} I V_{n-r}^T x(0)$

$I = VV^T = [V_r\ V_{n-r}] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} = V_rV_r^T + V_{n-r}V_{n-r}^T$

Digression Maximal Inequalities.

Let $X_1, \dots, X_N$ be N R.V. such that $X_i \sim \text{subG}(\sigma^2)$

Then $E[\max_{1\le i\le N} X_i] \le \sigma\sqrt{2\log N}$ $E[\max_{1\le i\le N} |X_i|] \le \sigma\sqrt{2\log 2N}$

$P(\max_{1\le i\le N} X_i > t) \le N e^{-t^2/2\sigma^2}$

Note that Random Variables need not be independent.

Proof: $E[\max_i X_i] = \frac{1}{s}\mathbb{E}[\log e^{s\max X_i}] \le \frac{1}{s}\log E[e^{s\max X_i}]$ Jensen

$\le \frac{1}{s}\log\sum_i E[e^{sX_i}] \le \frac{1}{s}\log N e^{\frac{\sigma^2 s^2}{2}}$ $\quad s = \sqrt{\frac{2\log N}{\sigma^2}}$

$\max_{1\le i\le N}|X_i| = \max_{1\le i\le 2N} X_i \quad X_{N+i} = -X_i$ for $i = 1, \dots, N$

Forward Euler Discretization $\quad \Delta t$ stepsize/learning rate/resolution

$\frac{X_{t+\Delta t} - X_t}{\Delta t} = -A X_t \qquad t = k\Delta t$

$X_{k+1} = (\underbrace{I - \Delta t\, A}_{B}) X_k$

$B = W\Sigma W^T, \quad u_k := W^T x_k$

$\vec{u}_{k+1} = \Sigma \vec{u}_k \qquad \Sigma = \text{diag}(\sigma_i) \quad \sigma_i := 1 - \Delta t\lambda_i$

$|\sigma_i| < 1 \Rightarrow \quad 0 < \Delta t < \frac{2}{\lambda_1} \qquad \lambda_1 > \lambda_2 > \cdots$

Max step size for forward Euler eqn to converge ODE $\quad$ track/

$x_k = (VV^T - \Delta t\, V\Lambda V^T)^k x_0 = V(I - \Delta t\,\Lambda)^k V^T x_0$

$= \begin{bmatrix} V_r & V_{n-r} \end{bmatrix} \begin{bmatrix} (I - \Delta t\,\Lambda_r)^k & 0 \\ 0 & (I_{n-r}) \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} x_0$

$x_k = V_r(I - \Delta t\,\Lambda_r)^k V_r^T x_0 + V_{n-r} V_{n-r}^T x_0$

$\| V_r(I - \Delta t\,\Lambda_r)^k V_r^T x_0 \|_2 \le \varepsilon$

$\Rightarrow \quad \max(|1 - \Delta t\lambda_1|, |1 - \Delta t\lambda_r|)^k \|u_0^r\| \le \varepsilon$

Mean Estimation in High Dimensions

$Y_i \overset{iid}{\sim} N(\theta^*, I_{d_n}) \qquad Y_i \in \mathbb{R}^{d_n}$ dropped

Estimator for θ^* $\quad \bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$ (Random) $\qquad \bar{Y}_n \sim N(\theta^*, \frac{I_{d_n}}{n})$ MAF

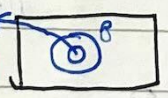
avg in i^{th} component

$E\|\bar{Y}_n - \theta^*\|^2 = E\sum_{i=1}^{d_n}(\bar{Y}_{ni} - \theta_i^*)^2 = \sum_{i=1}^{d_n} \text{Var}(\bar{Y}_{ni}) = \frac{d_n}{n}$

When does $\bar{Y}_n \xrightarrow{L^2} \theta^*$ if $\quad$ when $d_n = o(n)$

$A \Rightarrow B \Leftrightarrow B^c \Rightarrow A^c$ then $P(A) \le P(B)$

$\bar{Y}_n \xrightarrow{L^2} \theta^* \Rightarrow \bar{Y}_n \xrightarrow{P} \theta^*$



$n\|\bar{Y}_n - \theta^*\|^2 \sim \chi^2(d)$; $\chi^2(d) = \sum_{i=1}^{d} Z_i^2$ $Z_i \sim iid\, N(0,1)$ $\quad$ dof

$P(\|\bar{Y}_n - \theta^*\| > t) = P(n\|\bar{Y}_n - \theta^*\|^2 > nt^2) = 1 - F_{\chi^2(d)}(nt^2)$

$= Q(\frac{d}{2}, \frac{nt^2}{2})$

$d_n = \alpha n$

$\lim_{n\to\infty} Q(\frac{\alpha n}{2}, \frac{\alpha n}{2}) = \frac{1}{2}$

$\lim_{n\to\infty} P(\|Y_n - \theta^*\| > \sqrt{\alpha}) = \frac{1}{2}$

Def^n $\quad$ Convergence in Probability in High Dimensions

$X_n \in \mathbb{R}^d \quad X \in \mathbb{R}^d \quad X_n \xrightarrow{P} X$

$\lim_{n\to\infty} P(\sup_i |X_{ni} - X_i| > t) = 0$

$\Leftrightarrow \lim_{n\to\infty} P(\|X_n - X\|_\infty > t) = 0$

Add^n Structural Assumption : Sparsity. $\quad$ deterministic

$\bar{Y}_n \sim N(\theta^*, \frac{I_d}{n})$ $\quad \theta^*$ 1-sparse wlog $\theta_1^* \neq 0$

i^{th} component $\bar{Y}_{ni} \sim N(\theta_i^*, \frac{1}{n})$

$i^* = \text{argmax}_{i\in[d]} |\bar{Y}_{ni}|$, $\quad \hat{\theta} = \begin{cases} \hat{\theta}_{i^*} = \bar{Y}_{ni^*} \\ \theta_i^* = 0 \quad \forall i \neq i^* \end{cases}$

Claim $\quad E\|\hat{\theta} - \theta^*\|^2 \le \frac{C\log d}{n}$, $\quad d_n = o(e^n)$ we can now still have & have $\hat{\theta} \xrightarrow{L^2} \theta^*$

S1] Show $P(\|\hat{\theta} - \theta^*\|_2 > t) \le 2d\, e^{-\frac{nt^2}{4}}$

S2] Use for $Z \ge 0 \quad E[Z] = \int_0^\infty P(Z > t)\, dt$

S1] $\|\hat{\theta} - \theta^*\| = \sqrt{(\theta_1^* - \hat{\theta}_1 \mathbb{1}\{i^* = 1\})^2 + \sum_{i=2}^{d} (0 - \hat{\theta}_i \mathbb{1}\{i^* = i\})^2}$

$E\|\hat{\theta} - \theta^*\| \ge t \qquad \hat{\theta}_i \mathbb{1}\{i^* = i\} = Y_{ni}\mathbb{1}\{i^* = i\}$

$Q \Rightarrow \max(|\theta_1^* - \bar{Y}_{n1}|^2, \bar{Y}_{n2}^2, \dots, \bar{Y}_{nd}^2) \ge t^2/2 \qquad E \Rightarrow Q$

$\therefore P(E) \le P(Q) \le P(|\theta_1^* - \bar{Y}_{n1}| > \frac{t}{\sqrt{2}}) + \sum_{i=2}^{d} P(|\bar{Y}_{ni}| > \frac{t}{\sqrt{2}})$

$\le 1 \cdot 2e^{-(\ )} + (d-1)\, 2e^{-(\ )}$

$= 2d\, e^{-nt^2/4}$

S2] Having a Tail Bound Helps to get a bound in expectation

$E[\|\hat{\theta} - \theta^*\|^2] = \int_0^\infty P(\|\hat{\theta} - \theta^*\|^2 \ge t)\, dt = \int_0^\infty 2d\, e^{-nt/4} dt = \frac{8d}{n}$

if a factor of d comes up split the integral.

$\le \int_0^u P(\ )\, dt + \int_u^\infty 2d\, e^{-nt/4} dt$

$\le u + \frac{8d}{n} e^{-nu/4}$ $\quad$ minimized at $u = \frac{4}{n}\ln\frac{d}{2}$

(Approximately) Sparse Mean Estimation.

$Y_i \sim N(\theta^*, I_d)$ θ^* is approximately s-sparse.

$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ (can be other estimator also)

S/S^*: top s magnitude coordinates of $\hat{\theta}/\theta^*$

Deterministic bound. **S1** $\|\hat{\theta}-\theta^*\| \geq 3\sqrt{\frac{2s}{n}\log\frac{2d}{\delta}} + \sqrt{2}\|\theta^*_{-S^*}\|_1$

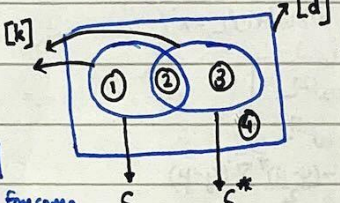
$\Rightarrow \exists i \in [d]$ s.t. $|\bar{Y}_{ni} - \theta^*_i| > \sqrt{\frac{2}{n}\log\frac{2d}{\delta}}$

$\Leftrightarrow \forall i \in [d]$ $|\bar{Y}_{ni}-\theta^*_i| \leq \sqrt{\frac{2}{n}\log\frac{2d}{\delta}}$ (ct) $\Rightarrow \|\hat{\theta}-\theta^*\| \leq t$

S2 $P(\|\hat{\theta}-\theta^*\| \geq \text{(1)}) \leq P(\exists i \in [d] \text{ s.t. } |\bar{Y}_{ni}-\theta^*_i| > ct)$

$3\sqrt{s}ct + \sqrt{2}\|\theta^*_{-S^*}\|_1 \leq 2de^{-\frac{nc^2t^2}{2}} \leq \delta$ $\quad tc = \sqrt{\frac{2}{n}\log\frac{2d}{\delta}}$

begin with this



$i \in$ then

① ② $|\hat{\theta}_i - \theta^*_i| = |\bar{Y}_{ni} - \theta^*_i| \leq ct$

③ $|\hat{\theta}_i - \theta^*_i| = |\hat{\theta}_i| \leq |\theta^*_i - \bar{Y}_{ni}| + |\bar{Y}_{ni}|$
$\leq ct + |\bar{Y}_{nj}|$ for some $j \in$ ①
$\leq ct + |\bar{Y}_{nj} - \theta^*_j| + |\theta^*_j|$
$\leq 2ct + |\theta^*_j|$

④ $|\hat{\theta}_i - \theta^*_i| = (\theta^*_i)$

$\|\hat{\theta}-\theta^*\|^2 = \sum_{i\in①\cup②} |\hat{\theta}_i-\theta^*_i|^2 \leq 8c^2t^2$ $\quad (x+y)^2 \leq 2(x^2+y^2)$

$+ \sum_{i\in③} |\hat{\theta}_i-\theta^*_i|^2 \left(\leq \sum_{j\in①}(2ct+|\theta^*_j|)^2 \leq \sum_{j\in①} 2(8c^2t^2+|\theta^*_j|^2)\right)$

$+ \sum_{i\in④} |\hat{\theta}_i-\theta^*_i|^2 \left(= \sum_{i\in④}|\theta^*_i|^2\right)$

$\leq 9sc^2t^2 + 2\|\theta^*_{-S^*}\|^2 \left(= \sum_{j\in①\cup④}|\theta^*_j|^2\right)$

Consider

(A) $Y_i \overset{iid}{\sim} N(\theta^*, \Sigma)$ Σ known θ^* sparse & unknown

(B) $Y_i \overset{iid}{\sim} N(A\theta^*, I_m)$ A known $m \times d$, " ↓

(A) can be reduced to (B) if Σ full rank

$y_i = A\theta^* + \xi_i$ $\quad \xi_i \sim N(0, I_m)$

Sample Average $\bar{Y}_n = A\theta^* + \xi$ $\quad \xi \sim N(0, \frac{I_m}{n})$

Consider noiseless case $n \to \infty$ so $\xi = 0$

Opt problem $\min_{\theta\in\mathbb{R}^d} \|\theta\|_0$ such that $X\vec{\theta} = \vec{y}$ — Cost f^n non-differentiable & non-convex.

⇓ Convex Relaxation

Basis Pursuit Linear Program $\min_{\theta\in\mathbb{R}^d} \|\theta\|_1$ such that $X\theta = y$

Exact Recovery & Restricted Nullspace.

When is solving BP equivalent to solving l_0 problem?

Tangent Cone $T(\theta^*) = \{\Delta \in \mathbb{R}^d \mid \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}$

Unfavourable case $\theta^* + \text{null}(X)$ passes through Tangent cone

$\text{null}(X) := \{\Delta \in \mathbb{R}^d \mid X\Delta = 0\}$

$C(S) := \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$

- The matrix X satisfies the Restricted Nullspace Property wrt S if $C(S) \cap \text{null}(X) = \{0\}$

$C(S)$ captures behaviour of tangent cone $T(\theta^*)$ ind. of θ^*

For any S-sparse vector θ^*, tangent cone $T(\theta^*)$ is contained within $C(S)$ & conversely $C(S)$ is contained in union of such tangent cones.

The following two properties are equivalent

a) For any vector $\theta^* \in \mathbb{R}^d$ with support S, BP applied with $y = X\theta^*$ has unique sol^n $\hat{\theta} = \theta^*$

b) Matrix X satisfies restricted nullspace property wrt S.

Restricted Isometry Property (RIP) For given $s \in \{1, \cdots, d\}$ we say that $X \in \mathbb{R}^{n\times d}$ satisfies RIP of order s with const $\delta_s(X) > 0$ if

$\left|\left|\left| \frac{X_S^T X_S}{n} - I_S \right|\right|\right|_2 \leq \delta_s(X)$ for all subsets S of size atmost s.

if A satisfies RIP $(1-\delta_s)\|\theta\|^2 \leq \frac{\|A\theta\|^2}{n} \leq (1+\delta_s)\|\theta\|^2$ $\quad \theta$: s sparse

Prediction f^n $f(x)=\theta^T x$

$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_i (y_i - x_i^T\theta)^2 + \lambda\|\theta\|^2 = (X^TX+\lambda\vec{I})^{-1}X^Ty$

"Robust Regression" $\hat{\theta}_{Robust} = \underset{}{\operatorname{argmin}} \sum_i \underbrace{h(x_i^T\theta - y_i)}_{cvx} + \lambda\|\theta\|^2$

Logistic Regression $P(y_i=+1|x_i) = p_{x_i} = \dfrac{e^{x_i^T\theta_i^*}}{1+e^{x_i^T\theta_i^*}}$

$P(y_i|x_i) = \dfrac{1}{1+e^{-y_i x_i^T\theta_i^*}}$

$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^n P(y_i|x_i) = \underset{\theta\in R^d}{\operatorname{argmin}} \sum \log(1+e^{-y_i x_i^T\theta})$

Gradient of loss lies in span of data.

Convexity $\not\Rightarrow$ GD converges

$\operatorname{prox}_f(x) = \underset{y\in R^d}{\operatorname{argmin}} \left(f(y) + \frac{1}{2}\|y-x\|^2\right)$

$\nabla \cdot \operatorname{prox}_f = \sum_i \frac{\partial}{\partial x_i} \operatorname{prox}_f$; $\nabla\cdot F = \operatorname{div} F = \sum_{i=1}^d \frac{\partial F}{\partial x_i}$

$R = X + N(0,\sigma^2)$

observed

$P_X = f_X(t) \propto e^{-s(t)}$ $\quad -\frac{(R-x)^2}{2\sigma^2}$

$P(R|x)$ likelihood $\propto e$

$P(x)$: prior, $P(x|R) = \frac{P(x,R)}{P(R)} = \frac{P(x)P(R|x)}{P(R)}$

Posterior $P(x|R) \propto e^{-\left(\frac{(R-x)^2}{2\sigma^2} + s(x)\right)}$

MAP $\Rightarrow \hat{x}(R) = \underset{x}{\operatorname{argmax}} P(x|R) = \operatorname{prox}_{\sigma^2 g}(R)$

$\hat{x}(R) = E[x|R] = \dfrac{\int t\, e^{-\left(\frac{(R-t)^2}{2\sigma^2}+s(t)\right)} dt}{e^{-\left(\frac{(R-t)^2}{2\sigma^2}+s(t)\right)} dt}$

$E[x|R] = \underset{f}{\operatorname{argmin}} \mathbb{E}\left(x - \underbrace{f(R)}_{-E[x]+E[x]}\right)^2 = E\left[E[x-f(R)]^2|R\right]$

$\min_w \ell(w) = \sum_i \log(1+e^{-y_i x_i^T w})$

Better formulation

SVM $\min \|w\|$ $y_i x_i^T w \geq 1$

Soft Margin SVM $\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \sum_i \xi_i$

$y_i x_i^T w \geq 1 - \xi_i$

$J(w,\xi,\alpha) = \frac{\lambda}{2}\|w\|^2 + 1^T\xi + \alpha^T(1-\xi-\tilde{X}w)$ $\quad \alpha\geq 0$ $\quad \tilde{X} = \begin{bmatrix} y_1 x_1^T \\ \vdots \\ y_n x_n^T \end{bmatrix}$

$g(\alpha) = \min_{\substack{w,\xi \\ \xi\geq 0}} J(w,\xi,\alpha)$

$g(\alpha) = 1^T\alpha + \min_w \frac{\lambda}{2}\|w\|^2 - \alpha^T\tilde{X}w + \min_{\xi\geq 0} \xi^T(1-\alpha)$

$g(\alpha) = -\frac{1}{2\lambda}\alpha^T\tilde{X}\tilde{X}^T\alpha - \delta_{(-\infty,1]^d}(\alpha) + 1^T\alpha$

$\max_{\alpha\geq 0} g(\alpha) = \max_{0\leq\alpha\leq 1} 1^T\alpha - \frac{\alpha^T\tilde{X}\tilde{X}^T\alpha}{2\lambda}$

Projected GD $\max_{a_n\leq\alpha\leq a_n} -\frac{\alpha^T\tilde{X}\tilde{X}^T\alpha}{2\lambda} + 1^T(\alpha)$

$\alpha_{t+1} =$

$\alpha_{t+1} = \operatorname{proj}_{[0,1]^n}\left(\alpha_t + \eta_t\left(-\frac{\tilde{X}\tilde{X}^T}{\lambda}\alpha_t + 1\right)\right)$

$\hat{\theta}_{rr} = \underset{\theta\in R^d}{\operatorname{argmin}} \frac{1}{2}\|y-A\theta\|^2 + \frac{\lambda}{2}\|\theta\|^2$

$\hat{\theta}_{rr} = (A^TA+\lambda I_d)^{-1}A^Ty$, $y = A\theta^* + \xi$

$E\|\hat{\theta}_{rr} - \theta^*\|^2 = \|(A^TA+\lambda I_d)^{-1}A^TA\theta^* - \theta^*\|^2 + E\|(A^TA+\lambda I_d)^{-1}A^T\xi\|^2$

use trace trick + SVD + assumption $E[\xi_i\xi_j] = 1(i=j)$

$Var = \sigma^2 \sum_{i=1}^d \frac{s_i^2}{s_i^2+\lambda^2}$

$Bias^2 = \sum_{i=1}^d \frac{\beta_i^2\lambda^2}{(s_i^2+\lambda)^2}$

$\lim_{\lambda\to 0^+} Bias^2 = \sum_{i=r+1}^d \beta_i^2 = \|V_{d-r}^T\theta^*\|^2$

$M = U\Sigma V^T$

$\sum_{i=1}^d s_i^2 = \|M\|_F^2 = tr(MM^T) = tr(\Sigma\Sigma^T)$

Sum of squares of singular values

$\min_w \|Y - Xw^T\|^2 + \|w^T\|_L^2$

$\|w^T\|_L^2 = \sum_{j=1}^d (w^j)^T L w^j$ $\quad L\in R^{t\times t}$

Sylvester eqn $X^TXw^T + w^TL = X^TY$

$w = \begin{bmatrix} -w^{(1)}- \\ -w^{(2)}- \\ -w^{(d)}- \end{bmatrix}$

$f_Y(y) = e^{-\frac{(y-\mu)^T\Sigma^{-1}(y-\mu)}{2}}$

$y^T\Sigma^{-1}y = \langle yy^T, \Sigma^{-1}\rangle = \langle vec(yy^T), vec(\Sigma^{-1})\rangle$

$\min_\theta \frac{1}{2}\|y-A\theta\|^2 + \lambda f(\theta)$

$A = USV^T$ $\quad \tilde{y} = U^Ty$

$\min_{\theta,\beta} \frac{1}{2}\|\tilde{y} - S\beta\|^2 + \lambda f(\theta)$

s.t. $V^T\theta = \beta$

init r_0, s_0

$\theta_t = \operatorname{prox}_{\frac{h}{s_t}}(r_t)$ $\quad \beta_t = \operatorname{prox}_{\frac{h}{\delta_t}}(q_t)$

$\eta_t = \frac{\rho\delta_t}{\nabla\cdot\operatorname{prox}_{\frac{h}{s_t}}(r_t)}$ $\quad \gamma_t = \frac{\rho\delta_t}{\nabla\cdot\operatorname{prox}_{\frac{h}{\delta_t}}}$

$\delta_t = \eta_t - s_t$ $\quad s_{t+1} = \gamma_t - \delta_t$

$q_t = V^T\left(\frac{\eta_t\theta_t - s_t r_t}{\delta_t}\right)$ $\quad r_{t+1} = V\left(\frac{\gamma_t\beta_t - \delta_t q_t}{s_{t+1}}\right)$

Mean Estimation in High Dimensions

$Y_i \overset{iid}{\sim} \mathcal{N}(\theta^*, I_{d_n})$ $\quad \bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$

$E[\|\bar{Y}_n - \theta^*\|^2] = E\sum_{i=1}^{d_n}(\bar{Y}_{ni} - \theta_i^*)^2 = \sum_{i=1}^{d_n} \mathrm{Var}(\bar{Y}_{ni}) \doteq d_n/n$

$\bar{Y}_n \xrightarrow{L^2} \theta^* \Rightarrow \bar{Y}_n \xrightarrow{P} \theta^*$

$\bar{Y}_n \xrightarrow{P} \theta^* \not\Rightarrow \bar{Y}_n \xrightarrow{L^2} \theta^*$

$A \Rightarrow B \Leftrightarrow B^c \Rightarrow A^c$ and $P(A) \le P(B)$

$n\|\bar{Y}_n - \theta^*\|^2 \sim \chi^2(d)$, $\chi^2(d) = \sum_{i=1}^d Z_i^2$ $\quad Z_i \sim \mathcal{N}(0,1)$ iid.

$P(\|\bar{Y}_n - \theta^*\| > t) = P(n\|\bar{Y}_n - \theta^*\|^2 > nt^2) = 1 - F_{\chi^2(d)}(nt^2)$

$= Q\left(\frac{d_n}{2}, \frac{nt^2}{2}\right)$

$\lim_{n\to\infty} Q\left(\frac{d_n}{2}, \frac{nt^2}{2}\right) = \frac{1}{2}$

Convergence in Probability in High Dimensions

$X_n \in \mathbb{R}^d$ $\quad X \in \mathbb{R}^d$ $\quad X_n \xrightarrow{p} X$ if

$\lim_{n\to\infty} P(\sup_i |X_{ni} - X_i| > t) = 0$ $\quad \left(\lim_{n\to\infty} P(\|X_n - X\|_\infty > t) = 0\right)$

Maximal Inequality

Let $X_1, \dots, X_N$ be N R.V. s.t. $X_i \sim \mathrm{SubG}(\sigma^2)$

Then $E\left[\max_{1\le i\le N} X_i\right] \le \sigma\sqrt{2\log N}$.

Conc^n inequality for SubGaussian R.V.

Y with μ, if $E[e^{t(Y-\mu)}] \le e^{\frac{\sigma^2 t^2}{2}}$ $\quad Y \sim \mathrm{SubG}(\sigma^2)$

$P(|Y-\mu| > \varepsilon) \le 2\exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right)$

Sparse Mean Estimation

$Y_i \sim \mathcal{N}(\theta^*, I_d)$ $\quad \theta^*$ is s-approximately sparse $\quad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$

$\hat{\theta}$: $\hat{\theta}(i) = \bar{Y}_n(i)\, 1(i \in$ Top s coordinates of $\bar{Y}_n)$

S1 Begin with $|\bar{Y}_{ni} - \theta_i^*| \le ct$ $\forall i$ & find upper bound on $|\hat{\theta} - \theta^*| \le u$

S2 $\|\hat{\theta} - \theta^*\| > 3\sqrt{\frac{2s}{n}\log\frac{2d}{\delta}} + \|\theta^*_{-S^*}\|_1\sqrt{2}$

$\Rightarrow \exists\, i : |\bar{Y}_{ni} - \theta_i^*| \ge ct$

$A \Rightarrow B$

$P(A) \le P(B)$

$\le \sum_i P(|\bar{Y}_{ni} - \theta_i^*| \ge ct) = 2de^{\frac{-nc^2t^2}{2}} \quad (=\delta)$

$Y_i \overset{iid}{\sim} \mathcal{N}(A\theta^*, I_m)$

Basis Pursuit (BP)

$\min_{\theta\in\mathbb{R}^d} \|\theta\|_1$ such that $X\theta = y$

$T(\theta^*) = \{\Delta \in \mathbb{R}^d \mid \|\theta^* + t\Delta\|_1 \le \|\theta^*\|_1, t > 0\}$

$C(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1\}$

Def: The matrix X satisfies the Restricted Nullspace Property wrt S if $C(S) \cap \mathrm{null}(X) = \{0\}$

The following are equivalent

a) For any vector $\theta^* \in \mathbb{R}^d$ with support S, BP applied with $y = X\theta^*$ has unique sol^n $\hat{\theta} = \theta^*$

b) The matrix X satisfies Restricted Nullspace property wrt X.

$\hat{\Delta} = \hat{\theta} - \theta^*$

b$\Rightarrow$a $\|\theta^*_S\|_1 = \|\theta^*\|_1 \ge \|\theta^* + \hat{\Delta}\|_1$

$\ge \|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1$

Conclude $\hat{\Delta} \in C(S)$ but also $\hat{\Delta} \in \mathrm{null}(X)$

Thus $\hat{\Delta} = 0$ & $\hat{\theta} = \theta^*$

a$\Rightarrow$b it suffices to show that if BP relaxation succeeds $\forall s$-sparse vectors then $\mathrm{null}(X)\setminus\{0\} \cap C(S) = \phi$. Choose $\theta^* \in \mathrm{null}(X)$

$\min \|\beta\|_1$ s.t. $X\beta = X\begin{bmatrix}\theta^*_S \\ 0\end{bmatrix}$ by ass. unique optimal sol^n $\hat{\beta} = [\theta^*_S\ 0]^T$

$[0\ -\theta^*_{S^c}]^T$ also sol^n $\|\theta^*_S\|_1 < \|\theta^*_{S^c}\|_1$ Thus $\theta^* \notin C(S)$

RIP

$A \in \mathbb{R}^{m\times d}$ satisfies RIP of order s with const $\delta_s(A)$ if

$\left\|\frac{A_S^T A_S}{n} - I_S\right\|_2 \le \delta_s(A)$

for all subsets S of size atmost s.

Let $\theta \in \mathbb{R}^d$ have support S.

$-\delta_s\|\theta_S\|^2 \le \theta_S^T\left(\frac{A_S^T A_S}{n} - I_S\right)\theta_S \le \delta_s\|\theta_S\|^2$

$(1-\delta_s)\|\theta\|^2 \le \frac{1}{n}\|A\theta\|^2 \le (1+\delta_s)\|\theta\|^2$

$\frac{\frac{1}{n}\|A\theta\|^2}{1+\delta_s} \le \|\theta\|^2 \le \frac{\frac{1}{n}\|A\theta\|^2}{1-\delta_s}$

- If RIP const of order $2s$ is bounded as $\delta_{2s}(X) < 1/3$, then uniform restricted nullspace property holds for any subset S of cardinality $|S| \le s$

$\theta \in \mathrm{Null}(X)$ it suffices to show $\|\theta_S\|_1 \le \|\theta_{S^c}\|_1$ $\quad \tilde{\theta}_j = \begin{cases}\theta_j & j \in A \\ 0\end{cases}$

S: largest s in θ $\quad S^c = \cup_{j\ge1} S_j$

$\theta = \tilde{\theta}_S + \sum_{k\ge1}\tilde{\theta}_{S_k}$

$\|\tilde{\theta}_S\|^2 \le \frac{1}{1-\delta_{2s}}\left\|\frac{1}{\sqrt{n}}X\tilde{\theta}_S\right\|^2$

$\le \frac{1}{1-\delta_{2s}}\left|\sum_{j\ge1}\left\langle X\tilde{\theta}_{S_0}, X\tilde{\theta}_{S_j}\right\rangle\right|$

$= \frac{1}{1-\delta_{2s}}\left|\sum_{j\ge1}\tilde{\theta}_{S_0}\left[\frac{X^TX}{n} - I_d\right]\tilde{\theta}_{S_j}\right|$

$\|\tilde{\theta}_{S_0}\|_2 \le \frac{\delta_{2s}}{1-\delta_{2s}}\sum_{j\ge1}\|\tilde{\theta}_{S_j}\|_2$

$\|\tilde{\theta}_{S_j}\|_\infty \le \frac{1}{s}\|\tilde{\theta}_{S_{j-1}}\|_1 \Rightarrow \|\tilde{\theta}_{S_j}\|_2 \le \frac{1}{\sqrt{s}}\|\tilde{\theta}_{S_{j-1}}\|_2$

$\|\tilde{\theta}_{S_0}\|_1 \le \sqrt{s}\|\tilde{\theta}_{S_0}\|_2 \le \frac{\delta_{2s}}{1-\delta_{2s}}\left\{\|\tilde{\theta}_{S_0}\|_1 + \sum_{j\ge1}\|\tilde{\theta}_{S_j}\|_1\right\}$