LECTURE 5

- PROOF OF CONVERGENCE FOR IHT.

REMEMBER THAT:

$$x_{t+1} = H_k\left(x_t - \eta \cdot \nabla f(x_t)\right)$$

FOR LINEAR REGRESSION: $\nabla f(x_t) = -A^T(y - Ax_t)$. THUS:

$$x_{t+1} = H_k\left(x_t + \eta A^T(y - Ax_t)\right)$$

WE ASSUME WE KNOW $k = \|x^*\|_0$. ALSO, FOR THE MOMENT ASSUME $\eta = 1$. THUS:

$$x_{t+1} = H_k\left(x_t + A^T(y - Ax_t)\right)$$

QUESTION: CAN'T WE JUST USE THE RESULT FROM CONVEX PROJ. GRADIENT DESCENT

ANSWER: NO - WE USED THE FACT THAT:

$$\|H_k(x) - H_k(y)\|_2 \leq \|x - y\|_2$$

WHICH IS NOT TRUE. COUNTEREXAMPLE:

$$x = \begin{bmatrix} 10 \\ 1 \end{bmatrix} \qquad \|x - y\|_2 = \sqrt{(10-1)^2 + (1-10)^2} = 9 \cdot \sqrt{2}$$

$$\longrightarrow$$

$$y = \begin{bmatrix} 1 \\ 10 \end{bmatrix} \qquad \|H_1(x) - H_1(y)\|_2 = \sqrt{(10-0)^2 + (0-10)^2} = 10 \cdot \sqrt{2}$$

THUS: $\|H_1(x) - H_1(y)\|_2 \geq \|x - y\|_2$

WHAT CAN WE SAY ABOUT OUR PROJECTION?

DENOTE $\tilde{x}_t = x_t + A^T(y - Ax_t)$. THEN:

$$\|x_{t+1} - \tilde{x}_t\|_2^2 \leq \|x^* - \tilde{x}_t\|_2^2 \Rightarrow \qquad \text{(BY DEFINITION OF } H_k(\cdot))$$

$$\|(x_{t+1} - x^*) + (x^* - \tilde{x}_t)\|_2^2 \leq \|x^* - \tilde{x}_t\|_2^2 \Rightarrow$$

$$\|x_{t+1} - x^*\|_2^2 + \|x^* - \tilde{x}_t\|_2^2 + 2\langle x_{t+1} - x^*, x^* - \tilde{x}_t\rangle \leq \|x^* - \tilde{x}_t\|_2^2 \Rightarrow$$

$$\|x_{t+1} - x^*\|_2^2 \leq 2\langle x_{t+1} - x^*, \tilde{x}_t - x^*\rangle$$

SINCE
$$\tilde{x}_t = x_t + A^T(y - Ax_t) = x_t + A^T(Ax^* + w - Ax_t) = x_t + A^TA(x^* - x_t) + A^Tw$$

DEFINE
$$\mathcal{U} := \text{supp}(x_t) \cup \text{supp}(x_{t+1}) \cup \text{supp}(x^*)$$

THEN:
$$\|x_{t+1} - x^*\|_2^2 \leq 2\langle x_{t+1} - x^*, x_t + A^TA(x^* - x_t) + A^Tw - x^*\rangle$$
$$= 2\langle x_{t+1} - x^*, (I - A_u^TA_u)(x_t - x^*)\rangle + 2\langle x_{t+1} - x^*, A_u^Tw\rangle$$

KEY PROPERTY OF INNER PRODUCT: $\langle x, A^Ty\rangle = x^TA^Ty = (Ax)^Ty = \langle Ax, y\rangle$

THUS:

i) $$\langle x_{t+1} - x^*, A_u^Tw\rangle = \langle A_u(x_{t+1} - x^*), w\rangle \leq \|A_u(x_{t+1} - x^*)\|_2 \cdot \|w\|_2 \leq \sqrt{1+\delta} \cdot \|x_{t+1} - x^*\|_2 \cdot \|w\|_2$$

ii) $$\langle x_{t+1} - x^*, (I - A_u^TA_u)(x_t - x^*)\rangle \leq \|x_{t+1} - x^*\|_2 \cdot \|(I - A_u^TA_u)(x_t - x^*)\|_2 \leq \|x_{t+1} - x^*\|_2 \cdot \underbrace{\|I - A_u^TA_u\|_2}_{\text{RESTRICTED ON } 3k\text{-SPARSE SETS}} \cdot \|x_t - x^*\|_2$$

WHERE ONE CAN SHOW THAT: $\|I - A_u^TA_u\|_2 \leq \max\{(1+\delta) - 1, 1 - (1-\delta)\}$

COMBINING:
$$\|x_{t+1} - x^*\|_2^2 \leq 2\delta \cdot \|x_{t+1} - x^*\|_2 \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta} \cdot \|x_{t+1} - x^*\|_2 \cdot \|w\|_2$$

$$\Rightarrow \|x_{t+1} - x^*\|_2 \leq 2\delta \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta} \cdot \|w\|_2$$

ASSUMING $\delta < 1/2$, $\rho = 2\delta < 1$, AND $\|w\|_2 \leq \theta$

$$\|x_{t+1} - x^*\|_2 \leq \rho \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta$$
$$\leq \rho^t \cdot \|x_0 - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta \cdot \sum_{i=0}^{t} \rho^i$$
$$= \rho^t \cdot \|x_0 - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta \cdot \frac{1 - \rho^{t+1}}{1 - \rho} \leq \rho^t \|x_0 - x^*\|_2 + \frac{\sqrt{1+\delta} \cdot \theta}{1 - \rho}$$

THUS, TO OBTAIN $\| x_{t+1} - x^* \|_2 \leq \varepsilon$, WE NEED: $O\left(\log \frac{\| x_0 - x^* \|_2}{\varepsilon}\right)$

ITERATIONS ■

- A DIFFERENT STEP SIZE, BASED ON RIP.

WE WILL USE THE FACT THAT:

i) IN CONVEX OPTIMIZATION, $\eta = \frac{1}{L}$ WORKS

ii) WE WILL COMPUTE L IN OUR SCENARIO

BY DEFINITION OF $f(\cdot)$: FOR x_1, x_2 k-SPARSE.

$$\| \nabla f(x_1) - \nabla f(x_2) \|_2 = \| -A^T(y - Ax_1) + A^T(y - Ax_2) \|_2$$

$$= \| A^T A (x_1 - x_2) \|_2$$

$$\leq \max_{S: |S| \leq 2k} \| (A^T A)_S \|_2 \cdot \| x_1 - x_2 \|_2$$

$$\leq \underbrace{(1 + \delta)}_{L} \| x_1 - x_2 \|_2 \quad \text{BY DEFINITION OF RIP}$$

THUS, POTENTIALLY, $\eta = \frac{1}{L}$ COULD WORK (AND FOR $\delta > 0$, IT IS $\eta < 1$.) ■

- ADAPTIVE STEP SIZE SELECTION.

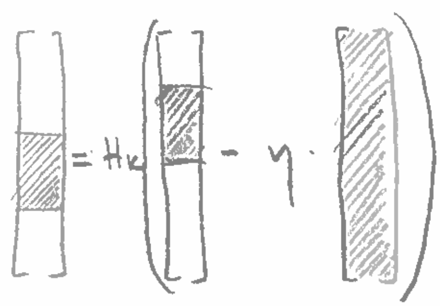
WE WANT TO COMPUTE η IN $x_{t+1} = H_k(x_t - \eta \cdot \nabla f(x_t))$

SOME OBSERVATIONS:

i) x_t IS k-SPARSE

ii) x_{t+1} IS k-SPARSE

iii) SCHEMATICALLY:



x_{t+1} HAS SUPPORT FROM x_t, $H_k(-\nabla f(x_t))$ OUTSIDE OF $\text{supp}(x_t)$ OR COMBINATION OF BOTH.

DEFINE: $S_t = \text{supp}(x_t)$

$S_{t+1} = \text{supp}(x_{t+1})$

$Q_t = S_t \cup \text{supp}\left(H_K\left(\nabla_{S_t^c} f(x_t)\right)\right)$

THEN: $$H_K\left(x_t - \eta \nabla f(x_t)\right) = H_K\left(x_t - \eta \cdot \nabla_{Q_t} f(x_t)\right)$$

GIVEN THE ABOVE, WE PERFORM LINE SEARCH FOR η AS:

$$\eta = \underset{\eta}{\text{argmin}} \left\| y - A\left(x_t - \eta \cdot \nabla_{Q_t} f(x_t)\right)\right\|_2^2$$

$\longrightarrow$ FINDS STEP SIZE THAT MINIMIZES $f(\cdot)$

TO FIND SUCH η, DEFINE $g(\eta) = \left\| y - A\left(x_t - \eta \nabla_{Q_t} f(x_t)\right)\right\|_2^2$

TAKING DERIVATIVE AND SETTING TO ZERO:

$$\nabla g(\eta) = 0 \Rightarrow 2\langle A\nabla_{Q_t} f(x_t), y - Ax_t \rangle + 2\eta \left\| A\nabla_{Q_t} f(x_t)\right\|_2^2 = 0$$

$$\Rightarrow \eta = \frac{-\langle A\nabla_{Q_t} f(x_t), y - Ax_t \rangle}{\left\| A\nabla_{Q_t} f(x_t)\right\|_2^2} = \frac{\left\| \nabla_{Q_t} f(x_t)\right\|_2^2}{\left\| A\nabla_{Q_t} f(x_t)\right\|_2^2}$$

SINCE $|\text{supp}(Q_t)| \leq 2k$,

$$\frac{1}{1+\delta} \leq \eta \leq \frac{1}{1-\delta} \qquad \blacksquare$$

- PROOF OF ADAPTIVE STEP SIZE IN IHT.

FOLLOWING THE SAME PROCEDURE AS IN $\eta = 1$, WE HAVE:

$$\|x_{t+1} - x^*\|_2 \leq 2 \cdot \|I - \eta A_u^T A_u\|_2 \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta} \cdot \eta \cdot \|w\|_2$$

BY RIP: $$\|I - \eta A_u^T A_u\|_2 \leq \max\left\{\eta(1+\delta) - 1, 1 - \eta(1-\delta)\right\}$$

$$\leq \max\left\{\frac{(1+\delta)}{(1-\delta)} - 1, 1 - \frac{(1-\delta)}{1+\delta}\right\}$$

BY THE PROPERTY $\frac{1}{\ldots} \leq \eta \leq \frac{1}{\ldots}$

THUS:

$$\|x_{t+1} - x^*\|_2 \leq 2 \frac{2\delta}{1-\delta} \cdot \|x_t - x^*\|_2 + \frac{2\sqrt{1+\delta}}{1-\delta} \|w\|_2$$

$$= \frac{4\delta}{1-\delta} \|x_t - x^*\|_2 + \frac{2\sqrt{1+\delta}}{1-\delta} \|w\|_2$$

ASSUMING $\delta < 1/5 \Longrightarrow \frac{4\delta}{1-\delta} := \rho < 1$.

COMPARE WITH PLAIN IHT. ∎

- GRAPHICAL MODEL SELECTION

LET $x \sim \mathcal{N}(\mu, \Sigma)$. THEN ITS PROBABILITY DENSITY SATISFIES:

$$f(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

DEFINE $\Theta = \Sigma^{-1}$ (INVERSE COVARIANCE MATRIX OR PRECISION MATRIX)

THEN:

$$f(x) = \frac{\det(\Theta)^{1/2}}{(2\pi)^{p/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot \Theta \cdot (x-\mu)\right\}$$

PROBLEM DEFINITION: ASSUME WE DO NOT KNOW (μ, Σ), BUT WE HAVE SAMPLES $\{x_i\}_{i=1}^n$, $x_i \sim \mathcal{N}(\mu, \Sigma)$. LET'S SEE WHAT WE CAN DO WITH THESE SAMPLES

ASSUME INDEPENDENCE BETWEEN x_i'S. THE LOG-LIKELIHOOD FUNCTION IS:

$$\ell(\mu, \Theta) = \sum_{i=1}^n \log f(x_i)$$

$$\propto \sum_{i=1}^n \log \det(\Theta)^{1/2} - \sum_{i=1}^n \frac{1}{2}(x_i - \mu)^T \Theta (x_i - \mu)$$

$$= \frac{n}{2} \log \det(\Theta) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \cdot \Theta (x_i - \mu)$$

OBSERVE THAT: $-\operatorname{tr}(\Theta \cdot \hat{\Sigma}) - (\mu - \hat{\mu})^T \Theta (\mu - \hat{\mu})$

$$= -\operatorname{tr}\left(\Theta \cdot \frac{1}{n} \sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i=1}^n x_i\right)\left(x_i - \frac{1}{n}\sum_{i=1}^n x_i\right)^T\right)$$

$$- \left(\mu - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^T \Theta \left(\mu - \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

(WHERE WE USED:
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$
)

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^T \Theta \left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$
$$- \left(\mu - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^T \Theta \left(\mu - \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^T \Theta (x_i - \mu)$$

THUS OUR $\ell(\cdot,\cdot)$ TRANFORMS INTO:

$$\ell(\mu, \Theta) = \frac{n}{2}\left(\log\det(\Theta) - tr(\Theta \cdot \hat{\Sigma}) - (\mu - \hat{\mu})^T \Theta (\mu - \hat{\mu})\right)$$

MAXIMUM LIKELIHOOD ESTIMATION OF (μ, Σ) LEADS TO:

$$\min_{\mu, \Theta \succ 0} -\log\det(\Theta) + tr(\Theta \cdot \hat{\Sigma}) + \underbrace{(\mu - \hat{\mu})^T \Theta (\mu - \hat{\mu})}_{\text{ONLY THIS TERM CONTAINS } \mu;\ \text{SINCE } \Theta \succ 0,\ \mu^* = \hat{\mu}}$$

$$\Big\downarrow \ \mu^* = \hat{\mu}$$

$$\min_{\substack{\Theta \succ 0 \\ \Theta \in \mathbb{R}^{p \times p}}} -\log\det(\Theta) + tr(\Theta \cdot \hat{\Sigma}) = -\log\det(\Theta) + \langle \Theta, \hat{\Sigma} \rangle$$

// THE DETERMINANT OF A SQUARED MATRIX IS (RELATIVELY) NOT AN EASY OBJECT/OPERATION TO DESCRIBE. THE GEOMETRIC WAY OF THINKING IT IS AS IF WE HAD A UNIT CUBE IN p DIMENSIONS; THEN $\det(\Theta)$ MEASURES THE VOLUME OF THE CUBE, AFTER APPLYING THE ROWS/COLUMNS OF Θ ON THAT CUBE. ANOTHER WAY TO SEE IT IS

$$\det(\Theta) = \prod_{i=1}^{p} \lambda_i(\Theta),$$ WHERE $\lambda_i(\Theta)$ IS THE i-TH EIGENVALUE OF Θ.

//

WHY DO WE CARE ABOUT ALL THIS?

THERE IS A VERY NICE THEORY CONNECTING UNDIRECTED GRAPHS UNDER GAUSSIAN ASSUMPTIONS AND COVARIANCE SELECTION.

"VARIABLES $x(i), x(j)$ FROM $x \sim \mathcal{N}(\mu, \Sigma)$ ARE CONDITIONALLY INDEPENDENT IFF $\Theta^*_{ij} = 0$."

(SEE EXAMPLE IN SLIDES)

<u>QUESTION</u>: GIVEN SAMPLES $\{x_i\}_{i=1}^n$, CAN WE INFER THE UNDERLYING UNDIRECTED GRAPH STRUCTURE?

<u>ANSWER #1</u>: TAKE MANY^2 SAMPLES $\longrightarrow$ COMPUTE $\hat{\mu}, \hat{\Sigma} \longrightarrow \hat{\Sigma}^{-1}$

IF $p = 10^5 - 10^6 \longrightarrow$ OFTEN IMPOSSIBLE.

ANSWER #2: FIND THE MOST IMPORTANT PART OF THE GRAPH: ASSUME SPARSITY IN Σ^{-1}

$$\min_{\Theta \succ 0} \; -\log\det(\Theta) + tr(\Theta \cdot \hat{\Sigma})$$

$$\text{s.t.} \quad \|\Theta\|_0 \leq k \qquad \text{(ASSUMING WE OBEY SYMMETRY)}$$

$-\log\det(\Theta) + tr(\Theta \cdot \hat{\Sigma})$ IS <u>LOCALLY</u> LIPSCHITZ GRADIENT

- PROOF OF RIP FOR SUBGAUSSIAN MATRICES

A RANDOM VARIABLE x IS CALLED SUBGAUSSIAN IF $\exists \beta, \kappa > 0$ SUCH THAT:

$$\mathbb{P}(|x| \geq t) \leq \beta e^{-\kappa t^2}, \; \forall t > 0$$

— " — " — x — " — SUBEXPONENTIAL — " —

$$\mathbb{P}(|x| \geq t) \leq \beta \cdot e^{-\kappa t}, \; \forall t > 0.$$

A VECTOR $y \in \mathbb{R}^p$ IS CALLED ISOTROPIC IF $\mathbb{E}\left[|\langle y, x\rangle|^2\right] = \|x\|_2^2, \; \forall x \in \mathbb{R}^p$

STEP 1: LET $A \in \mathbb{R}^{n \times p}$ WITH INDEPENDENT, ISOTROPIC AND SUBGAUSSIAN ROWS. THEN, $\forall x \in \mathbb{R}^p$ AND $\forall t \in (0,1)$: (8)

$$\mathbb{P}\left(\left| \frac{1}{n} \|Ax\|_2^2 - \|x\|_2^2 \right| \geq t \cdot \|x\|_2^2 \right) \leq 2 \cdot e^{-ct^2 n}, \quad c \text{ CONSTANT}$$

PROOF: W.L.O.G., $\|x\|_2 = 1$. LET $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}^p$ BE ROWS OF A.

DEFINE: $z_i = |\langle \alpha_i, x \rangle|^2 - \|x\|_2^2$. SINCE α_i IS ISOTROPIC, $\mathbb{E}[z_i] = 0$. FURTHER, z_i IS SUBEXPONENTIAL, SINCE $\langle \alpha_i, x \rangle$ IS SUBGAUSSIAN; THIS MEANS:

$$\mathbb{P}(|z_i| \geq r) \leq \beta e^{-\kappa r}, \quad \forall r > 0$$

OBSERVE:

$$\frac{1}{n}\|Ax\|_2^2 - \|x\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(|\langle \alpha_i, x\rangle|^2 - \|x\|_2^2\right) = \frac{1}{n}\sum_{i=1}^{n} z_i$$

SINCE α_i's ARE INDEPENDENT, z_i's ARE INDEPENDENT.

WE WILL USE THE FOLLOWING BERNSTEIN INEQUALITY:

// LET $x_1, x_2, \ldots, x_M$ BE INDEPENDENT, ZERO-MEAN, SUBEXPONENTIAL R.V.S., WITH CONSTANTS β, κ. THEN:

$$\mathbb{P}\left(\left| \sum_{i=1}^{M} x_i \right| \geq t \right) \leq 2 e^{-\frac{(\kappa t)^2/2}{2\beta M + \kappa t}}$$ //

IN OUR CASE, THIS TRANSLATES INTO:

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} z_i \right| \geq t \right) = \mathbb{P}\left(\left| \sum_{i=1}^{n} z_i \right| \geq tn \right) \leq 2 e^{-\frac{\kappa^2 n^2 t^2 / 2}{2\beta n + \kappa n t}}$$

$$\leq 2 \cdot e^{-\frac{\kappa^2}{4\beta + 2\kappa} \cdot n t^2} \quad \text{FOR } t \in (0,1)$$

■

STEP 2: ASSUME STEP 1 HOLDS. FIX A SET $S \subset [p]$ WITH $|S| = k$ AND $\delta, \xi \in (0,1)$. IF

$$n \geq \frac{c}{\delta^2}\left(7k + 2\ln\left(\frac{2}{\xi}\right)\right), \quad c \text{ CONSTANT},$$

THEN W.P. AT $1-\xi$:

$$\| A_S^T A_S - I \|_2 < \delta.$$

PROOF: WE WILL USE THE CONSTRUCTION OF ε-NETS OVER UNIT BALLS. LET $B = \{x \in \mathbb{R}^p, \|x\|_2 \leq 1\}$. AN ε-NET OVER B IS A SET SUCH THAT, FOR EVERY POINT IN B, THERE IS A POINT IN THE ε-NET THAT ε-CLOSE BY SOME DISTANCE FUNTION (E.g. $\|x-y\|_2 \leq \varepsilon$).

THE NUMBER OF POINTS IN SUCH ε-NET CAN BE BOUNDED BY:

$$\mathcal{N}(B, \|\cdot\|_2, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^p$$

IN OUR CASE, WE GENERATE AN ε-NET ON $B = \{x \in \mathbb{R}^p, \text{supp}(x) \subset S, \|x\|_2 \leq$

THIS IS THE SET OF VECTORS FOR SET S.

IN THIS CASE:

$$\mathcal{N}(B, \|\cdot\|_2, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^k$$

THEN, FROM STEP 1:

$$\mathbb{P}\left(\left| \|Au\|_2^2 - \|u\|_2^2 \right| \geq t \cdot \|u\|_2^2, \text{ FOR SOME } u \text{ IN } \varepsilon\text{-NET}\right)$$

$$\leq \sum_{u \text{ IN } \varepsilon\text{-NET}} \mathbb{P}\left(\left| \|Au\|_2^2 - \|u\|_2^2 \right| \geq t \cdot \|u\|_2^2\right)$$

$$\leq 2 \cdot \left(1 + \frac{2}{\varepsilon}\right)^k e^{-c t^2 n}$$

DEFINE: $D = A_S^T A_S - I$. THEN:

$$\left| \|Au\|_2^2 - \|u\|_2 \right| = \left| \langle A_S^T A_S u, u \rangle - \langle u, u \rangle \right|$$

$$= \left| \langle (A_S^T A_S - I) u, u \rangle \right| = \left| \langle Du, u \rangle \right|$$

THEN, OUR GOAL IS PROVE $|\langle Dx, x \rangle| < t$ (FOR $x \in B$, AND PROPER t) VIA $|\langle Du, u \rangle| < t$ WHERE u IS IN ε-NET.

ASSUME $|\langle Du, u \rangle| < t$. THIS OCCURS W.P. $1 - 2\left(1 + \frac{2}{\varepsilon}\right)^k e^{-ct^2 n}$

THEN, FOR SOME $x \in B$, AND SOME u IN ε-NET SUCH THAT:

$$\|x-u\|_2 \le \varepsilon < 1/2, \text{ WE GET:}$$

$$\begin{aligned}|\langle Dx,x\rangle| &= |\langle Du,u\rangle + \langle D(x+u), x-u\rangle| \\ &\le |\langle Du,u\rangle| + |\langle D(x+u), x-u\rangle| \\ &\le t + \|D\|_2 \cdot \|x+u\|_2 \cdot \|x-u\|_2 \le t + 2\cdot\|D\|_2\cdot\varepsilon\end{aligned}$$

TAKING MAXIMUM OVER $x \in B$:

$$\|D\|_2 < t + 2\|D\|_2 \cdot \varepsilon \Rightarrow \|D\|_2 \le \frac{t}{1-2\varepsilon}$$

CHOOSE $t = (1-2\varepsilon)\cdot\delta \longrightarrow \|D\|_2 < \delta$. THIS MEANS:

$$\mathbb{P}\left(\|A_S^T A_S - I\|_2 \geq \delta\right) \le 2\left(1+\frac{2}{\varepsilon}\right)^k e^{-c(1-2\varepsilon)^2\delta^2 n}$$

CHOOSING $\varepsilon = 2/e^{7/2}-1$, WE GET THAT $\|A_S^T A_S - I\|_2 \le \delta$ WITH PR. $1-\xi$

PROVIDED.

$$n \geq \frac{c}{\delta^2}\left(7k + 2\ln\left(\frac{2}{\xi}\right)\right)$$

STEP 3: WE PROVED THAT $\|A_S^T \cancel{A_S} - I\|_2 < \delta$ FOR A SINGLE S. TAKING ALL $\binom{p}{k}$ SUBSETS $S \subset [p]$ WITH $|S| = k$, WE GET:

$$\begin{aligned}\mathbb{P}\left(\sup_{S:\dots} \|A_S^T A_S - I\|_2 \geq \delta\right) &\le \sum_S \mathbb{P}\left(\|A_S^T A_S - I\|_2 \geq \delta\right) \\ &\le 2\cdot\binom{p}{k}\left(1+\frac{2}{\varepsilon}\right)^k \cdot e^{-c(1-2\varepsilon)^2\cdot\delta^2\cdot n} \\ &\le 2\cdot\left(\frac{ep}{k}\right)^k\left(1+\frac{2}{\varepsilon}\right)^k e^{-c(1-2\varepsilon)^2\delta^2 n}\end{aligned}$$

FORCING THIS PROBABILITY BE LESS THAN ξ WE GET

$$n \geq O\left(k\ln\left(\frac{ep}{k}\right) + \frac{14}{3}k + \frac{4}{3}\ln\left(\frac{2}{\xi}\right)\right)$$

∎