High Dimensional Probability and Statistics **2nd Semester, 2023-2024**

Lecture 8: Random Matrices and Applications

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2024/05/05)*

Motivation: The study of random matrices is directly motivated by the estimation of covariance matrices. Let $X \in \mathbb{R}^n$ be a *mean zero* random vector. Then the covariance matrix corresponding to X is given by

$$\Sigma = \mathbb{E}\left[XX^T\right].$$

However, since we typically do not know the distribution of X but only have access to m i.i.d samples $\{X_k\}_{k=1}^m$ of X, a natural estimator[1] of Σ is

$$\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T.$$

Then we would like to know how close the random matrix Σ_m to its mean Σ, *particularly in terms of the matrix spectral norm.*

The approaches for studying the concentration of random matrices relies on the knowledge of the distribution of the elements. For example, if the random matrix has sub-Gaussian entries, we can establish the concentration results based on the concentration of random variables through the variational expression for the matrix spectral norm. When there is no explicit distributions associated with the elements of the random matrix, the matrix concentration bound can be developed by imitating the Chernoff method for random variables. That is, either we can use the concentration inequalities for the random variables directly, or we can extend the proof techniques for the random variable case to the random matrix case.

Before proceeding, it is worth noting that we will study matrix concentration in terms the spectral norm rather than the Frobenius norm. This is largely due to that the deviation of principle directions associated with the covariance matrix is typically of interest, and the bound based on spectral norm is sufficiently tighter than that based on the Frobenius norm (which is the sum of the errors in all directions). In addition, it is trivial that the matrix concentration bound in terms of Frobenius norm can be reduced to concentration result of random variables.

Agenda:

- Covariance matrix under sub-Gaussian assumption
- Application: Clustering based on PCA
- Matrix Bernstein inequality
- Application: Covariance matrix for general distributions
- Application: Sparse Recovery

[1] When the covariance matrix is known to have certain structure, a better estimator can be constructed based on that structure, see for example Chapter 6.5 of [1].

8.1 Covariance Matrix under sub-Gaussian Assumption

In this section we will consider the concentration of the covariance matrix Σ_m when X is a sub-Gaussian random vector, defined as follows.

Definition 8.1 (Sub-Gaussian random vector) *A mean zero random vector $X \in \mathbb{R}^n$ is sub-Gaussian with parameter σ^2 if for each $v \in \mathbb{S}^{n-1}$ (i.e., $\|v\|_2 = 1$), $\langle X, v\rangle$ is a sub-Gaussian random variable with parameter σ^2.*

Example 8.2 *Assume $X \in \mathbb{R}^n$ has i.i.d σ^2-sub-Gaussian entries. Then,*

$$\mathbb{E}\left[e^{\lambda\langle X,v\rangle}\right] = \mathbb{E}\left[\prod_{k=1}^{n} e^{\lambda v_k X_k}\right] \leq \prod_{k=1}^{n} e^{\frac{\lambda^2 v_k^2 \sigma^2}{2}} = e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{for } v \in \mathbb{S}^{n-1},$$

meaning $\langle X, v\rangle$ is σ^2-sub-Gaussian. Thus, X is a σ^2-sub-Gaussian random vector.

Example 8.3 *Let $X \sim \mathcal{N}(0, \Sigma)$. Then for any $v \in \mathbb{S}^{n-1}$, $v^T X \sim \mathcal{N}(0, v^T\Sigma v)$. Since $v^T\Sigma v \leq \|\Sigma\|_2$, we can conclude that X is a sub-Gaussian random vector with parameter at most $\|\Sigma\|_2$.*

The following lemma provides a characterization of the spectral norm of a symmetric matrix in terms of the ε-net. We have indeed seen this result for general matrices in Lecture 4.

Lemma 8.4 *Let $Z \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Assume $\varepsilon \in [0, 1/2)$ and let N be a ε-net of $\mathbb{S}^{n-1}$ under the $\|\cdot\|_2$ metric. Then*

$$\|Z\|_2 \leq \frac{1}{1-2\varepsilon} \sup_{v\in N} |\langle Zv, v\rangle| .$$

Proof: For any $x \in \mathbb{S}^{n-1}$, by the definition of ε-net, there exists a vector $\pi(x) \in N$ such that $\|x - \pi(x)\|_2 \leq \varepsilon$. It follows that

$$\langle Zx, x\rangle - \langle Z\pi(x), \pi(x)\rangle = \langle Z(x - \pi(x)), x\rangle + \langle Z\pi(x), x - \pi(x)\rangle ,$$

and hence

$$|\langle Zx, x\rangle - \langle Z\pi(x), \pi(x)\rangle| \leq 2\varepsilon\|Z\|_2.$$

Consequently,

$$\|Z\|_2 = \sup_{x\in\mathbb{S}^{n-1}} |\langle Zx, x\rangle| \leq \sup_{x\in\mathbb{S}^{n-1}} \left(|\langle Z\pi(x), \pi(x)\rangle| + 2\varepsilon\|Z\|_2\right).$$

Then the proof is complete after rearrangement. ∎

Theorem 8.5 *Let $X \in \mathbb{R}^n$ be a mean zero σ^2-sub-Gaussian random vector and $\Sigma = \mathbb{E}\left[XX^T\right]$ be its covariance matrix. Let $\{X_k\}_{k=1}^m$ be i.i.d samples and define $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$. Then,*

$$\mathbb{P}\left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \geq c_1\left\{\sqrt{\frac{n}{m}} + \frac{n}{m}\right\} + t\right] \leq c_2 \exp\left(-c_3 \min\{t, t^2\}m\right) \quad \text{for all } t \geq 0.$$

Here, $c_1, c_2, c_3 > 0$ are absolute numerical constants.

Proof: Let $Z = \Sigma_m - \Sigma$. Taking N to be a 1/4-net of $\mathbb{S}^{n-1}$, we have $|N| \le 9^n$ and

$$\|Z\|_2 \le 2 \sup_{v \in N} |\langle Zv, v \rangle| .$$

The overall strategy of the proof is to first consider a fixed $v \in N$ and then take a union bound.

For any fixed $v \in N$, we have

$$\langle Zv, v \rangle = \frac{1}{m} \sum_{k=1}^{m} \left(\left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \right).$$

Since $X_k^T v$ is σ^2-sub-Gaussian, we have

$$\begin{aligned} \left\| \left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \right\|_{L_p} &\le \left\| \left(X_k^T v \right)^2 \right\|_{L_p} + \mathbb{E} \left[\left(X_k^T v \right)^2 \right] \\ &\lesssim \sigma^2 p, \end{aligned}$$

implying that $\left(X_k^T v \right)^2 - \mathbb{E} \left[\left(X_k^T v \right)^2 \right]$ is $c_4 \cdot \sigma^4$-sub-exponential. Thus the application of the Bernstein inequality yields that

$$\mathbb{P} \left[|\langle Zv, v \rangle| \ge \frac{\delta}{2} \right] \lesssim \exp \left(-c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right).$$

Taking a union bound yields that

$$\begin{aligned} \mathbb{P} \left[\|Z\|_2 \ge \delta \right] &\le \mathbb{P} \left[\sup_{v \in N} |\langle Zv, v \rangle| \ge \frac{\delta}{2} \right] \\ &\lesssim 9^n \exp \left(-c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right) \\ &= \exp \left(n \log 9 - c_5 \min \left\{ \frac{\delta^2}{\sigma^4}, \frac{\delta}{\sigma^2} \right\} m \right) \end{aligned} \tag{8.1}$$

Let $\delta = \left(c_1 \left\{ \sqrt{\frac{n}{m}} + \frac{n}{m} \right\} + t \right) \sigma^2$. Then,

$$\delta \ge \left(c_1 \frac{n}{m} + t \right) \sigma^2 \quad \text{and} \quad \delta^2 \ge \left(c_1^2 \frac{n}{m} + t^2 \right) \sigma^4.$$

Substituting them into (8.1) yields that

$$\mathbb{P} \left[\|Z\|_2 \ge \delta \right] \lesssim \exp \left(n \log 9 - c_5 \min \left\{ c_1 \frac{n}{m} + t, c_1^2 \frac{n}{m} + t^2 \right\} m \right).$$

The proof is complete if we take c_1 to be sufficiently large. ■

Remark 8.6 *Given the tail bound, it is anticipated to obtain the moment bound, in particularly on* $\mathbb{E} \left[\|\Sigma_m - \Sigma\|_2 \right]$. *Since*

$$\mathbb{E} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \right] = \int_0^\infty \mathbb{P} \left[\frac{\|\Sigma_m - \Sigma\|_2}{\sigma^2} \ge x \right] dx$$

$$
\begin{aligned}
&= \int_0^{c_1\left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\}} \mathbb{P}\left[\frac{\|\Sigma_m-\Sigma\|_2}{\sigma^2} \geq x\right] dx + \int_{c_1\left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\}}^{\infty} \mathbb{P}\left[\frac{\|\Sigma_m-\Sigma\|_2}{\sigma^2} \geq x\right] dx \\
&\leq c_1\left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\} + \int_0^{\infty} \mathbb{P}\left[\frac{\|\Sigma_m-\Sigma\|_2}{\sigma^2} \geq c_1\left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\}+t\right] dt \\
&\leq c_1\left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\} + c_2\int_0^{\infty} \exp\left(-c_3\min\{t,t^2\}m\right) dt \\
&\lesssim \sqrt{\frac{n}{m}}+\frac{n}{m},
\end{aligned}
$$

it follows that

$$
\mathbb{E}\left[\|\Sigma_m-\Sigma\|_2\right] \lesssim \left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\}\sigma^2.
$$

Moreover, we have

$$
\mathbb{E}\left[\|\Sigma_m\|_2\right] \lesssim \|\Sigma\|_2 + \left\{\sqrt{\frac{n}{m}}+\frac{n}{m}\right\}\sigma^2.
$$

Thus, an upper bound for $\mathbb{E}\left[\|\Sigma_m\|_2\right]$ *can be derived from the concentration result under less stringent conditions. Note this bound cannot be obtained via methods discussed in the previous lectures since they only work for (sub)-Gaussian processes.*

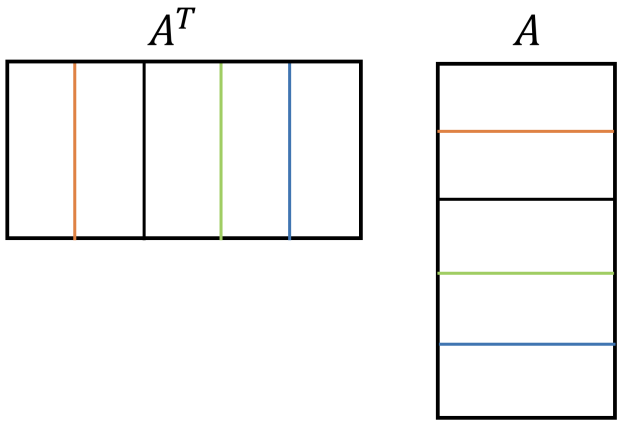

Figure 8.1: $\Sigma_m = \frac{1}{m}A^TA$.

Remark 8.7 *Assume* $\Sigma = I_n$ *and* X_k *is sub-Gaussian with parameter* $\sigma^2 = 1$*. Note that we can express* Σ_m *as* $\Sigma_m = \frac{1}{m}A^TA$*, where* $A^T = [X_1, \cdots, X_m]$ *(see Figure 8.3). Thus, Theorem 8.5 implies that, with high probability,*

$$
1 - c'\sqrt{\frac{n}{m}} \leq \frac{\sigma_{\min}(A)}{\sqrt{m}} \leq \frac{\sigma_{\max}(A)}{\sqrt{m}} \leq 1 + c'\sqrt{\frac{n}{m}}
$$

for some numerical constant $c' > 0$*, with the proviso that* $m \geq n$*. That is,* A *behaves more and more well-conditioned (like an orthogonal matrix) when* m/n *increases. This turns out to be a useful result itself.*

8.2 Application: Clustering Based on PCA

The PCA paradigm which first projects data onto a low dimensional subspace can be used for data clustering. For simplicity we consider the following Gaussian mixture model with two different means $\{-\mu, \mu\}$,

$$X = \varepsilon\mu + g, \tag{8.2}$$

where $\varepsilon \in \{1, -1\}$ is a Rademacher random variable, $\mu \in \mathbb{R}^n$ is deterministic and $g \in \mathcal{N}(0, I_n)$. In words, sampling from X will generate two clusters of data, obeying $\mathcal{N}(-\mu, I_n)$ and $\mathcal{N}(\mu, I_n)$ respectively, see Figure 8.2.

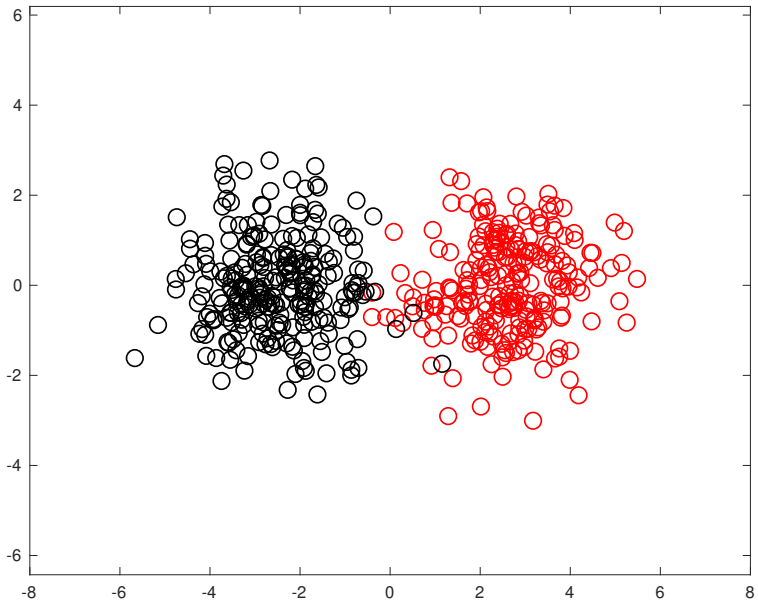

Figure 8.2: A simulation of points generated according to the Gaussian mixture model (8.2).

Suppose we are given a sample of m points $\{X_k\}_{k=1}^m$ drawn according to the Gaussian mixture model and want to identify which points belong to which cluster (i.e., determine they are generated from which mean). From the simulation, it is not hard to see that the data generated from X is stretch in the direction of μ, and the data points from different clusters have different inner product with μ. Assuming $\|\mu\|_2 > 1$, noting that

$$\langle \varepsilon\mu + g, \mu \rangle = \varepsilon\|\mu\|_2^2 + \langle g, \mu \rangle ,$$

where the size of $\langle g, \mu \rangle$ is about $\|\mu\|_2$, the sign of the inner product will coincide with ε, and hence can tell which mean the data point corresponds to. Indeed, if we define

$$Z_k = (\text{sign}(\langle \underbrace{\varepsilon_k\mu + g_k}_{X_k}, \mu/\|\mu\|_2 \rangle) \neq \varepsilon_k),$$

by the Hoeffding inequality, it can be shown that with high probability the number of misclassifications $\sum_{k=1}^m Z_k$ cannot exceed a fraction of m (**show this!**).

In the situation when we do not know μ but only have access to $\{X_k\}_{k=1}^m$, we can approximate μ by PCA since the principal direction of PCA captures the direction that the data points stretch the most. This gives the spectral algorithm for data clustering (here "spectral" refers to using the eigenvectors of a matrix for the task since the eigen-decomposition of a matrix is also known as spectral decomposition),

- Compute the covariance matrix $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$.
- Compute the principal eigenvector q (of unit norm) of Σ_m, i.e., eigenvector corresponding to the largest eigenvalue of Σ_m.
- Partition the data points into two clusters based on the sign of $\langle X_k, q\rangle$ (data points with the same sign of $\langle X_k, q\rangle$ will be put into the same cluster).

Next we are going to show that q can be close to μ. To this end, we need the Davis-Kahan theorem.

Theorem 8.8 (Davis-Kahan) *Let S and T be two symmetric matrices with the same dimension. Suppose the i-th largest eigenvalue of S is well separated from the rest of them:*

$$\min_{j\neq i} |\lambda_j(S) - \lambda_i(S)| > \delta.$$

Then the acute angle θ_i between the unit-norm eigenvectors $\mu_i(S)$ and $\mu_i(T)$ corresponding to the i-th largest eigenvalues satisfies

$$\sin\theta_i \leq \frac{2\|S-T\|_2}{\delta}.$$

In particular, there exists a $\theta \in \{1,-1\}$ such that $\|\mu_i(S) - \mu_i(T)\|_2 \leq 2^{3/2}\|S-T\|_2/\delta$.

Note that

$$\Sigma = \mathbb{E}\left[XX^T\right] = \mu\mu^T + I_n,$$

and the largest eigenvalue of Σ is $1+\|\mu\|_2^2$, with the corresponding normalized eigenvector $\mu/\|\mu\|_2$. Since X is a sub-Gaussian random vector with the parameter proportional to $\|\mu\|_2^2$ (**check this!**), By Theorem 8.5, we have

$$\|\Sigma_m - \Sigma\|_2 \leq \rho\|\mu\|_2^2, \tag{8.3}$$

for a sufficiently small $\rho > 0$ when $m \gtrsim n$ (the hidden constant relies on ρ). Noting the gap between the first and second largest eigenvalues of Σ is $\|\mu\|_2^2$, the Davis-Kahan theorem together with (8.3) implies that

$$\exists\ \theta \in \{1,-1\} \text{ such that } \|q - \theta(\mu/\|\mu\|_2)\|_2 \leq \rho',$$

where $\rho' > 0$ is also a sufficiently small number (a multiple of ρ).

8.3 Matrix Bernstein Inequality

In the last section, we have studied the covariance matrix concentration based on the distributional information of the matrix elements (e.g, certain sub-Gaussian rows). When there is no distribution assumption to use, we may develop matrix concentration inequalities via the matrix Chernoff method, which imitates the Chernoff method for random variables. Both the matrix Hoeffding inequality and the matrix Bernstein inequality can be developed this way. In this section we focus on the more widely used matrix Bernstein inequality.

8.3.1 Matrix Calculus

In this section we use $\mathbb{S}^{n\times n}$ to denote the set of $n \times n$ symmetric matrices and use $\mathbb{S}_+^{n\times n}$ to denote the set of $n \times n$ symmetric and positive definite matrices. In addition, we say $X \preceq Y$ or $Y \succeq X$ if $Y - X$ is positive semidefinite.

Definition 8.9 (Matrix Function) *Let $X \in \mathbb{S}^{n\times n}$ with the eigenvalue decomposition $X = Q\Lambda Q^T = \sum_{k=1}^n \lambda_k q_k q_k^T$. Given a function $f : \mathbb{R} \to \mathbb{R}$, we define $f(X)$ as*

$$f(X) = \sum_{k=1}^n f(\lambda_k) q_k q_k^T$$

In other words, we compute $f(X)$ by applying $f(\cdot)$ to each eigenvalue of X while the eigenvectors remain unchanged.

Example 8.10 *Let $f(x) = a_0 + a_1 x + \cdots + a_j x^j$. Then,*

$$f(X) = a_0 I + a_1 X + \cdots + a_j X^j.$$

Example 8.11 *Let $f(x) = e^x$. Then,*

$$f(X) = e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

Example 8.12 *Let $f(x) = \log x$. Then, for $X \in \mathbb{S}_+^{n\times n}$,*

$$e^{f(X)} = e^{\log X} = X.$$

Exercise 8.13 *Let X and Y be two matrices in $\mathbb{S}^{n\times n}$.*

1. *Show that if the matrices commute (i.e., $XY = YX$), then*

$$e^{X+Y} = e^X e^Y.$$

2. *Give an example of two matrices X and Y such that*

$$e^{X+Y} \neq e^X e^Y.$$

Note that the identity $e^{x+y} = e^x e^y$ plays a crucial role in the proof of the concentration of the sum of random variables. Indeed, this identity allows us to tensorize, i.e., to break the moment generating function of variable sum into the product of exponentials. Unfortunately, as we see in the above exercise, similar identity does not hold for matrices in general. Nevertheless, there are useful substitutes in terms of the matrix trace, which are stated below without proofs.

Lemma 8.14 (Golden-Thompson inequality) *For two matrices X and Y in $\mathbb{S}^{n\times n}$, we have*

$$\operatorname{trace}\left(e^{X+Y}\right) \leq \operatorname{trace}\left(e^X e^Y\right).$$

Lemma 8.15 (Lieb inequality) *Let $H \in \mathbb{S}^{n\times n}$. Define the function on the set $\mathbb{S}_+^{n\times n}$,*

$$f(X) = \operatorname{trace}\left(\exp\left(H + \log X\right)\right).$$

Then $f(X)$ is a concave function on $\mathbb{S}_+^{n\times n}$.

Remark 8.16 *The Jensen inequality still holds for random matrices since we can interpret $f(X)$ as a function of all the entries of X. Thus, letting X be a random matrix, we have*

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(H + \log X\right)\right)\right] \leq \operatorname{trace}\left(\exp\left(H + \log \mathbb{E}\left[X\right]\right)\right)$$

Letting $X = e^Z$, we have

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(H + Z\right)\right)\right] \leq \operatorname{trace}\left(\exp\left(H + \log \mathbb{E}\left[e^Z\right]\right)\right). \tag{8.4}$$

This inequality will be used in the proof of the matrix Bernstein inequality.

Both the Golden-Thompson inequality and the Lieb inequality can be used to establish the matrix Bernstein inequality. We will use the Lieb inequality next as it tensorizes better and thus yields better parameter dependence.

8.3.2 Matrix Bernstein Inequality

Theorem 8.17 (Matrix Bernstein inequality) *Let $X_1, \cdots, X_m$ be independent, mean zero, $n\times n$ symmetric random matrices. Assume $\|X_k\|_2 \leq B$ almost surely for all k. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] \leq 2n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right),$$

where $\sigma^2 = \left\|\sum_{k=1}^m \mathbb{E}\left[X_k^2\right]\right\|_2$.

Note that the matrix Bernstein is an exact analogue of the Bernstein inequality for random variables. Thus, the overall proof strategy is similar to that for the variable case. We start by establishing a matrix MGF inequality.

Lemma 8.18 (Moment generating function of random matrix) *Let $X \in \mathbb{S}^{n\times n}$ be a mean zero random matrix which satisfies $\|X\|_2 \leq B$ almost surely. Then,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \preceq \exp\left(g(\lambda)\mathbb{E}\left[X^2\right]\right) \quad \text{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - B|\lambda|/3}$$

provided that $|\lambda| < 3/B$.

Proof: First it can be shown that (**check this!**)

$$e^z \leq 1 + z + \frac{1}{1-|z|/3} \cdot \frac{z^2}{2} \quad \text{if } |z| < 3.$$

Thus, for $|x| \leq B$, if $|\lambda| < 3/B$, then

$$e^{\lambda x} \leq 1 + \lambda x + g(\lambda) x^2.$$

It follows that

$$\exp(\lambda X) \preceq I + \lambda X + g(\lambda) X^2,$$

provided $\|X\|_2 \leq B$ and $|\lambda| < 3/B$. Taking expectation on both sides yields that

$$\mathbb{E}[\exp(\lambda X)] \preceq I + g(\lambda)\mathbb{E}\left[X^2\right] \preceq \exp\left(g(\lambda)\mathbb{E}\left[X^2\right]\right),$$

as desired. ■

Proof: [Proof of Theorem 8.17] Noting that

$$\left\|\sum_{k=1}^m X_k\right\|_2 = \max\left\{\lambda_{\max}\left(\sum_{k=1}^m X_k\right), \lambda_{\max}\left(-\sum_{k=1}^m X_k\right)\right\},$$

it suffices to show that $\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^m X_k\right) \geq t\right] \leq n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right)$, and the bound for $\mathbb{P}\left[\lambda_{\max}\left(-\sum_{k=1}^m X_k\right) \geq t\right]$ can be established in the same manner. To this end, for fixed $\lambda \geq 0$ and the application of the Markov inequality gives

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^m X_k\right) \geq t\right] &= \mathbb{P}\left[\exp\left(\lambda \cdot \lambda_{\max}\left(\sum_{k=1}^m X_k\right)\right) \geq \exp(\lambda t)\right] \\
&\leq \exp(-\lambda t)\,\mathbb{E}\left[\exp\left(\lambda \cdot \lambda_{\max}\left(\sum_{k=1}^m X_k\right)\right)\right] \\
&= \exp(-\lambda t)\,\mathbb{E}\left[\lambda_{\max}\left(\exp\left(\lambda \cdot \sum_{k=1}^m X_k\right)\right)\right] \\
&\leq \exp(-\lambda t)\,\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^m X_k\right)\right)\right]. \qquad (8.5)
\end{aligned}$$

To apply the Lieb inequality (8.4), letting $H = \lambda \sum_{k=1}^{m-1} X_k$ and $Z = \lambda X_m$, we have

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^m X_k\right)\right)\right] \leq \mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \sum_{k=1}^{m-1} X_k + \log \mathbb{E}\left[e^{\lambda X_m}\right]\right)\right)\right]$$

Repeating this process yields that

$$\mathbb{E}\left[\operatorname{trace}\left(\exp\left(\lambda \cdot \sum_{k=1}^m X_k\right)\right)\right] \leq \operatorname{trace}\left(\exp\left(\sum_{k=1}^m \log \mathbb{E}\left[e^{\lambda X_k}\right]\right)\right)$$

$$
\begin{aligned}
&\leq \operatorname{trace}\left(\exp\left(\sum_{k=1}^{m}\log\exp\left(g(\lambda)\mathbb{E}\left[X_k^2\right]\right)\right)\right)\\
&= \operatorname{trace}\left(\exp\left(g(\lambda)\sum_{k=1}^{m}\mathbb{E}\left[X_k^2\right]\right)\right)\\
&\leq n\left\|\exp\left(g(\lambda)\sum_{k=1}^{m}\mathbb{E}\left[X_k^2\right]\right)\right\|_2\\
&= n\cdot\exp\left(g(\lambda)\left\|\sum_{k=1}^{m}\mathbb{E}\left[X_k^2\right]\right\|_2\right)\\
&= n\cdot\exp\left(g(\lambda)\sigma^2\right)
\end{aligned}
$$

provided $|\lambda| \leq 3/B$, where in the second line we have used Lemma 8.18 for every $\mathbb{E}\left[e^{\lambda X_k}\right]$, the last line follows from the definition of σ^2. Plugging this bound into (8.5) gives

$$
\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{m}X_k\right)\geq t\right]\leq n\cdot\exp\left(-\lambda t+g(\lambda)\sigma^2\right).
$$

Note that this bound holds for all $0<\lambda<3/B$, and thus we can minimize the right side over this interval. Indeed, the minimum is attained at $\lambda = t/(\sigma^2 + Bt/3)$, yielding

$$
\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{m}X_k\right)\geq t\right]\leq n\cdot\exp\left(-\frac{t^2/2}{\sigma^2+Bt/3}\right),
$$

which is the desirable bound. ■

From the tail bound on $\|\sum_{k=1}^{m}X_k\|_2$, we can obtain a bound on the expectation.

Theorem 8.19 (Matrix Bernstein in expectation) *Let $X_1,\cdots,X_m$ be independent, mean zero, $n\times n$ symmetric random matrices. Assume $\|X_k\|_2\leq B$ almost surely for all k and let $\sigma^2 = \left\|\sum_{k=1}^{m}\mathbb{E}\left[X_k^2\right]\right\|_2$. Then,*

$$
\mathbb{E}\left[\left\|\sum_{k=1}^{m}X_k\right\|_2\right]\lesssim\sigma\sqrt{\log n}+B\log n.
$$

Proof: By Theorem 8.17, it is not hard to show that (**check this!**) there exists an absolute numerical constant $c>0$ such that

$$
\mathbb{P}\left[\left\|\sum_{k=1}^{m}X_k\right\|_2\geq c\left(\sigma\sqrt{\log n+u}+B(\log n+u)\right)\right]\leq 2e^{-u}.
$$

Thus,

$$
\mathbb{E}\left[\left\|\sum_{k=1}^{m}X_k\right\|_2\right]=\int_0^\infty\mathbb{P}\left[\left\|\sum_{k=1}^{m}X_k\right\|_2\geq t\right]dt
$$

$$
\begin{aligned}
&= \int_0^{c\left(\sigma\sqrt{\log n}+B\log n\right)} \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] dt + \int_{c\left(\sigma\sqrt{\log n}+B\log n\right)}^{\infty} \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] dt \\
&\leq c\left(\sigma\sqrt{\log n}+B\log n\right) \\
&\quad + \int_0^{\infty} \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq c\left(\sigma\sqrt{\log n+u}+B(\log n+u)\right)\right]\left(\frac{c\sigma}{2\sqrt{\log n+u}}+cB\right) du \\
&\leq c\left(\sigma\sqrt{\log n}+B\log n\right) \\
&\quad + \left(\frac{c\sigma}{2\sqrt{\log n}}+cB\right)\int_0^{\infty} \mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq c\left(\sigma\sqrt{\log n+u}+B(\log n+u)\right)\right] du \\
&\leq c\left(\sigma\sqrt{\log n}+B\log n\right) + \left(\frac{c\sigma}{\sqrt{\log n}}+cB\right)\int_0^{\infty} e^{-u} du \\
&\lesssim \sigma\sqrt{\log n}+B\log n,
\end{aligned}
$$

which completes the proof. ■

The matrix Bernstein inequality can be extended to non-symmetric and non-square matrices.

Theorem 8.20 (Matrix Bernstein inequality for rectangular matrices) *Let $X_1, \cdots, X_m$ be independent, mean zero, $n_1 \times n_2$ matrices. Assume $\|X_k\|_2 \leq B$ almost surely for all k. Then, for any $t \geq 0$, we have*

$$
\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 \geq t\right] \leq 2(n_1+n_2)\exp\left(-\frac{t^2/2}{\sigma^2+Bt/3}\right),
$$

where

$$
\sigma^2 = \max\left(\left\|\sum_{k=1}^m \mathbb{E}\left[X_k X_k^T\right]\right\|_2, \left\|\sum_{k=1}^m \mathbb{E}\left[X_k^T X_k\right]\right\|_2\right).
$$

Proof: Apply Theorem 8.17 to the sum of $\begin{bmatrix} 0 & X_k^T \\ X_k & 0 \end{bmatrix}$. ■

8.4 Application: Covariance Matrix for General Distributions

In the first section we have considered the covariance matrix problem when the random vector is sub-Gaussian. In this section we remove the sub-gaussian requirement and consider the case when the random vector has bounded ℓ_2-norm. In this situation, the Bernstein inequality will yield better result than simply using Theorem 8.5 with a crude estimation of the sub-Gaussian parameter based on the ℓ_2-norm of the random vector.

Theorem 8.21 *Let $X_1, \cdots, X_m \in \mathbb{R}^n$ be i.i.d zero mean random vectors with covariance $\Sigma = \mathbb{E}\left[X_k X_k^T\right]$. Assume $\|X_k\|_2 \leq \sqrt{b}$ almost surely. Then for any $t > 0$, the sample covariance matrix $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$ satisfies*

$$
\mathbb{P}\left[\|\Sigma_m - \Sigma\|_2 \geq t\right] \leq 2n \cdot \exp\left(-\frac{mt^2/2}{b\|\Sigma\|_2 + 2bt/3}\right).
$$

In addition, we have

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|_2\right] \lesssim \sqrt{\frac{b\|\Sigma\|_2 \log n}{m}} + \frac{b \log n}{m}.$$

Proof: First note that if $\|X_k\|_2 \leq \sqrt{b}$, there holds (**check this!**)

$$\|\Sigma\|_2 = \left\|\mathbb{E}\left[X_k X_k^T\right]\right\|_2 \leq b.$$

Letting $Z_k = \frac{1}{m}\left(X_k X_k^T - \Sigma\right)$, it follows that

$$\|Z_k\|_2 \leq \frac{1}{m}\left\|X_k X_k^T\right\|_2 + \frac{1}{m}\|\Sigma\|_2 \leq \frac{2b}{m}.$$

Moreover, we have

$$\mathbb{E}\left[Z_k^2\right] = \frac{1}{m^2}\left(\mathbb{E}\left[(X_k X_k^T)^2\right] - \Sigma^2\right) \preceq \frac{1}{m^2}\mathbb{E}\left[\|X_k\|_2^2 X_k X_k^T\right] \preceq \frac{b}{m^2}\Sigma.$$

It follows that,

$$\sigma^2 = \left\|\sum_{k=1}^m \mathbb{E}\left[Z_k^2\right]\right\|_2 \leq \frac{b\|\Sigma\|_2}{m}.$$

Thus, applying Theorems 8.17 and 8.19 concludes the proof. ∎

Example 8.22 *Let $X_k = \sqrt{n} e_{k_j}$, where e_{k_j} is the k_j-th canonical vector in $\mathbb{R}^n$ with k_j being sampled uniformly at random from $\{1, \cdots, n\}$. Then*

$$\mathbb{E}\left[X_k X_k^T\right] = \sum_{j=1}^n e_j e_j^T = I_n \quad \textit{and} \quad \|X_k\|_2 \leq \sqrt{n}.$$

Thus, by Theorem 8.21, we have

$$\mathbb{E}\left[\|\Sigma_m - I_n\|_2\right] \lesssim \sqrt{\frac{n \log n}{m}} + \frac{n \log n}{m}.$$

8.5 Application: Sparse Recovery

Consider the following underdetermined linear system (see Figure 8.3 for a pictorial illustration):

$$y = Ax^* + w, \tag{8.6}$$

where $A \in \mathbb{R}^{m \times n}$ is a fat matrix with $m < n$, y denotes the observation, x^* denotes the parameter to be estimated or signal to be reconstructed, and w denotes the measurement noise. *The goal is to infer or reconstruct x^* from the observation y.*

The linear model (8.6) arises in many statistical and signal processing applications. In statistics, (8.6) models the regime where the number of responses is fewer than the number of predictors (or covariates). In signal processing, it describes the problem where the number of measurements is smaller than size of the signal. Since the number of unknowns is larger than the number of equations, (8.6) does not admit a unique solution, in contrast to the classical least squares problem. Therefore, additional structures on the unknown vector x^* is needed to reduce the feasible space. In this section we will focus on the sparse solution, namely x^* only has a few nonzero entries.

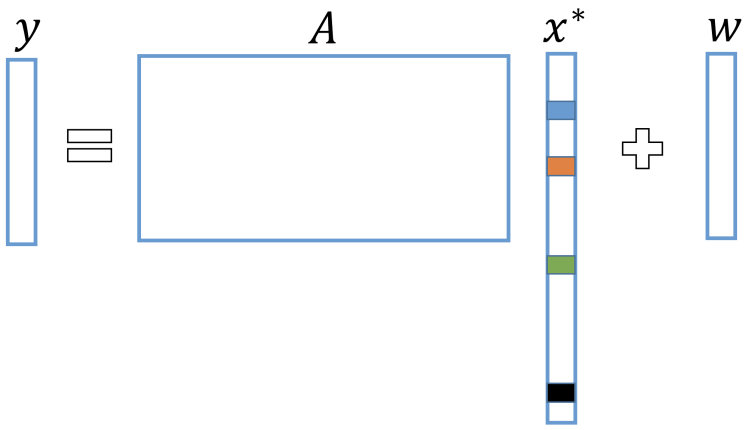

Figure 8.3: A pictorial illustration of (8.6).

.

Definition 8.23 (Sparse vector) *A vector $x \in \mathbb{R}^n$ is said to s-sparse if the number of nonzero entries in x is less than or equal to s. In other words, if we define*

$$\|x\|_0 = \#\{k \in \{1, \cdots, n\} : \ x_k \neq 0\}$$

which counts the number of nonzero entries in x, then x is s-sparse if $\|x\|_0 \leq s$.

In this lecture we will refer $\|\cdot\|_0$ as the ℓ_0-norm though it is technically not a norm. The notion of sparsity plays an important role in modern statistics, signal processing and machine learning, which characterizes a special type of low dimensional structure.

- In statistics, especially in the context of variable selection, it means only a number of covariates play an important role (a typical example is genome expression).
- In signal processing or machine learning, it means the signal of interest has the sparse structure itself or under certain linear transform.

A basic question to answer is how and when one can reconstruct the sparse vector x^* when there are fewer observations. There have been many methods for sparse parameter estimation or sparse signal reconstruction, including both the convex and nonconvex methods. In this lecture, we study the most widely studied methods based on the ℓ_1-norm. For simplicity, we only consider the noiseless case (i.e., $w = 0$). The noisy case can discussed in an overall similar way, see the references for details.

8.5.1 Exact Recovery in the Noiseless Setting

Since we know x^* is a sparse signal it is natural to reconstruct it by seeking the sparsest vector which is consistent with the measurement, namely by solving the following ℓ_0-minimization problem:

$$\min_{x \in \mathbb{R}^n} \|x\|_0 \quad \text{subject to} \quad Ax = y. \tag{8.7}$$

However, the ℓ_0 minimization problem is nonconvex and computationally intractable due to the combinatorial nature of ℓ_0-norm. In optimization, convex relaxation is a widely used technique

to handle nonconvex problems. Here, the nearest convex relaxation of the ℓ_0-norm is the ℓ_1-norm which sums up the magnitudes of all the entries of a vector (i.e., $\|x\|_1 = \sum_{k=1}^n |x_k|$). Replacing the ℓ_0-norm with the ℓ_1-norm in the objective leads to the following ℓ_1-minimization,

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad Ax = y. \tag{8.8}$$

The ℓ_1-minimization problem is also known as *basis pursuit* in the literature. It is a convex problem which can be rewritten as a linear programming. It can be solved by the first order or the second order methods. Indeed, the ℓ_1-minimization problem has spurred the significant development of the first order methods in optimization.

A central question in this section is when the ℓ_1-minimization is able to recover the target sparse solution x^*. To understand why the ℓ_1-minimization returns a sparse solution we first present the intuition and then give a rigorous analysis. Noting that (8.8) is trivially equivalent to

$$\min_{t\in\mathbb{R}} t \quad \text{subject to} \quad \|x\|_1 = t \text{ and } Ax = y.$$

That is, the solution to (8.8) can be found by gradually enlarge the ℓ_1-ball until the ball intersect with the solution set, see Figure 8.4. Since the ℓ_1-ball is pointy at its vertices (or the extreme sets in high dimension), the vertices will first touch the solution set. Noting the vertices have fewer nonzero entries, the ℓ_1-minimization tends to return a sparse solution.

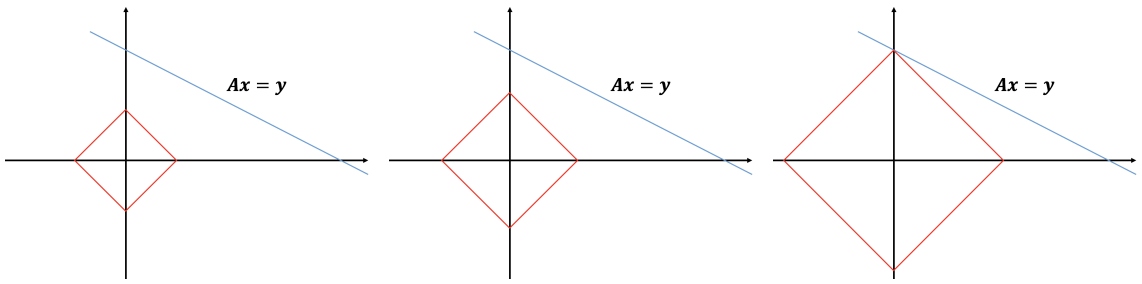

Figure 8.4: A pictorial illustration of ℓ_1-minimization.

.

There are several different conditions which have been developed for the guarantee analysis of the ℓ_1-minimization. In this lecture we will adopt the restricted isometry property proposed by Candes and Tao [2005].

Definition 8.24 (Restricted Isometry Property (RIP)) *Given an integer $s \in \{1, \cdots, n\}$, we say the matrix $A \in \mathbb{R}^{m\times n}$ $(m < n)$ satisfies the restricted isometry property with the constant δ_s if*

$$(1 - \delta_s)\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + \delta_s)\|x\|_2^2 \tag{8.9}$$

holds for all s-sparse vectors x such that $\|x\|_0 \le s$.

The restricted isometry property basically means that every s columns of A, denoted A_S with $|S| = s$, form a nearly orthogonal matrix when δ_s is small since it can be easily seen that (8.9) is equivalent to

$$\|A_S A_S^T - I_s\|_2 \le \delta_s \tag{8.10}$$

for any subset S of cardinality at most s, where A_S denotes the sub-matrix formed by the columns of A in S.

We are now in position to present a rigorous analysis about when the ℓ_1 minimization is able to exactly reconstruct the target solution x^* based on the restricted isometry property of the matrix.

Theorem 8.25 (Exact recovery) *Let $y = Ax^*$, where x^* is a s-sparse vector (i.e., $\|x^*\|_0 \leq s$). If the RIP constant of A of order $3s$ satisfies $\delta_{3s} < 1/3$, then the solution to (8.8) is x^*. That is, the ℓ_1 minimization is able to exactly recovery the sparse vector x^*.*

A careful reader may wonder when a matrix A satisfies the condition $\delta_{3s} < 1/3$. As can be seen in the last section, certain random matrix satisfies this condition with high probability when $m \gtrsim s \log n$.

Proof: [Proof of Theorem 8.25] Let S denote the support of x^* and S^c denote the complement of S in $\{1, \cdots, n\}$. We first show that for any $x = x^* + h \in \mathbb{R}^n$, if $\|x\|_1 \leq \|x^*\|_1$, then there must hold

$$\|h_{S^c}\|_1 \leq \|h_S\|_1. \tag{8.11}$$

This follows from

$$\|x^*\|_1 \geq \|x\|_1 = \|x^* + h\|_1 = \|x^*_S + h_S\|_1 + \|h_{S^c}\|_1 \geq \underbrace{\|x^*_S\|_1}_{=\|x^*\|_1} - \|h_S\|_1 + \|h_{S^c}\|_1.$$

Thus it suffices to show the following nullspace property[2]: for any h in the nullspace of A (i.e., $Ah = 0$), if h satisfies (8.11), then we must have $h = 0$.

Next we are going to show that if $\delta_{3s} < 1/3$, the nullspace property holds. To this end, let $S_0 = S$ be the support of x^*, let S_1 be the first $2s$ largest entries (in magnitude) of h_{S^c}, let S_2 be the second $2s$ largest entries (in magnitude) of h_{S^c},and so on. Let $h_{S_j} \in \mathbb{R}^n$ be the vector such $h_{S_j}(i) = h(i)$ when $i \in S_j$ and $h(S_j)(i) = 0$. With a slight abuse of notion, we also use h_{S_j} to denote the vector segment supported on S_j. Noting that

$$0 = Ah = Ah_{S_0 \cup S_1} + \sum_{j \geq 2} Ah_{S_j},$$

we have

$$\begin{aligned} 0 &\geq \|Ah_{S_0 \cup S_1}\|_2 - \|\sum_{j \geq 2} Ah_{S_j}\|_2 \\ &\geq \|Ah_{S_0 \cup S_1}\|_2 - \sum_{j \geq 2} \|Ah_{S_j}\|_2 \\ &\geq \sqrt{1 - \delta_{3s}} \|h_{S_0 \cup S_1}\|_2 - \sqrt{1 + \delta_{3s}} \sum_{j \geq 2} \|h_{S_j}\|_2. \end{aligned} \tag{8.12}$$

[2]The nullspace property for sparse recovery which basically means that the nullspace of A does not insects with the descent direction of the ℓ_1-norm at x^*. It is actually both sufficient and necessary for exact recovery of basis pursuit, see for example [1]. Theorem 8.25 gives a sufficient condition for this property to hold in terms of the RIP constant.

Moreover, a simple calculation yields that

$$\begin{aligned}\sum_{j\geq 2}\|h_{S_j}\|_2 &\leq \sum_{j\geq 2}\sqrt{2s}\|h_{S_j}\|_\infty \\ &\leq \sum_{j\geq 2}\frac{\|h_{S_{j-1}}\|_1}{\sqrt{2s}} \\ &\leq \frac{1}{\sqrt{2s}}\|h_{S^c}\|_1 \\ &\leq \frac{1}{\sqrt{2s}}\|h_S\|_1 \\ &\leq \frac{1}{\sqrt{2}}\|h_S\|_2 \\ &\leq \frac{1}{\sqrt{2}}\|h_{S_0\cup S_1}\|_2, \end{aligned} \tag{8.13}$$

where the fourth line follows from (8.11). Inserting this inequality into (8.12) gives

$$\left(\sqrt{1-\delta_{3s}} - \frac{\sqrt{1+\delta_{3s}}}{\sqrt{2}}\right)\|h_{S_0\cup S_1}\|_2 \leq 0.$$

Since $\sqrt{1-\delta_{3s}} - \frac{\sqrt{1+\delta_{3s}}}{\sqrt{2}} > 0$ due to the assumption $\delta_{3s} < 1/3$, $\|h_{S_0\cup S_1}\|_2 = 0$ and thus $\|h\|_2 = 0$. ■

8.5.2 Random Matrices Satisfying RIP

Theorem 8.26 *Let A be an $m\times n$ matrix whose rows A_i are independent, isotropic (i.e., $\mathbb{E}\left[A_i^T A_i\right] = I_n$), sub-Gaussian vectors with parameter $\sigma^2 = 1$. Then, if*

$$m \gtrsim \delta^{-2} s\log n,$$

the matrix $A/\sqrt{m}$ satisfies the RIP with a small constant $0 < \delta < 1$ with probability at least $1 - c_2\cdot\exp\left(-c_4\delta^2 m\right)$, where c_2 and c_4 are numerical constants.

Proof: Recall that, by (8.10), it is enough to show

$$\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \leq \delta$$

for all subsets S of cardinality s, where A_S denotes the sub-matrix constructed from the columns of A in S.

For a fixed subset S, first note that $A_i(S)$ is also σ^2-sub-Gaussian (**why?**) and it also satisfies $\mathbb{E}\left[A_i(S)^T A_i(S)\right] = I_s$. Thus, the application of Theorem 8.5 implies that

$$\mathbb{P}\left[\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \geq c_1\sqrt{\frac{s}{m}} + t\right] \leq c_2\exp\left(-c_3\min\{t, t^2\}m\right),$$

provided $m \geq s$. Let $t = \frac{\delta}{2}$. If $m \gtrsim c\cdot\delta^{-2}s\log n$ for a sufficiently large constant $c > 0$, then

$$\left\|\frac{1}{m}A_S^T A_S - I_s\right\|_2 \leq c_1\sqrt{\frac{s}{m}} + \frac{\delta}{2} \leq \delta$$

for all subsets S of cardinality s with probability at least

$$1 - \binom{n}{s} \cdot c_2 \exp\left(-c_3 \delta^2 m\right) \geq 1 - c_2 \cdot \exp\left(s \log n - c_3 \delta^2 m\right) \geq 1 - c_2 \cdot \exp\left(-c_4 \delta^2 m\right),$$

which completes the proof. ■

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 6.2, 6.3, 6.4, 7.1, 7.2, 7.3.

[2] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapters 4.6, 4.7, 5.4, 5.6, 10.5, 10.6.