Lecture 10: Matrices Review

Isabella Zhu

11 March 2025

§1 Last Lecture Wrapup

We will wrap up the proof from lecture 9.

Theorem 1.1
Assume INC(k) with k equal to the sparsity of θ^* (i.e. $k = |\theta^*|_0$). Fix

$$2\tau = 8\sigma\sqrt{\log(2d)/n} + 8\sigma\sqrt{\log(1/\delta)/n}.$$

Then, the MSE of the lasso estimator is at most

$$\text{MSE}(\mathbb{X}\hat{\theta}^L) \le 32k\tau^2 \lesssim \frac{\sigma^2|\theta^*|_0}{n}\log(d/\delta)$$

Moreover,

$$|\hat{\theta} - \theta^*|_2^2 \le 2\text{MSE}(\mathbb{X}\hat{\theta}^L)$$

all happening with probability at least $1-\delta$.

Proof. For the five hundred millionth time, we start with the good ole basic inequality

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \le 2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}|_1$$

We bound

$$2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle \le 2|\mathbb{X}^T\epsilon|_\infty \cdot |\hat{\theta} - \theta^*|_1$$

We bound the highest column norm of $\mathbb{X}$. We have

$$|\mathbb{X}_j|_2^2 = (\mathbb{X}^T\mathbb{X})_{jj} \le n + \frac{n}{32k} \le 2n$$

by the incoherence property. Therefore, we get

$$2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle \le 2|\mathbb{X}^T\epsilon|_\infty \cdot |\hat{\theta} - \theta^*|_1 \le 2 \cdot 2n \cdot \frac{\tau}{4} \cdot |\hat{\theta} - \theta^*|_1 = n\tau|\hat{\theta} - \theta^*|_1$$

To summarize, we've proved so far that

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \le n\tau|\hat{\theta} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}|_1$$

We add $n\tau|\hat{\theta}-\theta^*|_1$ on both sides.

$$|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2^2+n\tau|\hat{\theta}-\theta^*|_1 \le 2n\tau|\hat{\theta}-\theta^*|_1+2n\tau|\theta^*|_1-2n\tau|\hat{\theta}|_1$$

Now we take the support S into account. We have

$$|\hat{\theta}|_1=|\hat{\theta}_S|_1+|\hat{\theta}_{S^c}|_1 \implies |\hat{\theta}-\theta^*|_1-|\hat{\theta}|_1=|\hat{\theta}_S-\theta^*|_1-|\hat{\theta}_S|_1.$$

Putting it together,

$$|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2^2+|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2^2 \le 2n\tau\left[|\hat{\theta}_S-\theta^*|_1+|\theta^*|_1-|\hat{\theta}|_S\right] \le 4n\tau|\hat{\theta}_S-\theta^*|_1$$

We have that

$$|\hat{\theta}-\theta^*|_1 \le 4|\hat{\theta}_S-\theta^*|_1 \Leftrightarrow |\hat{\theta}_{S^c}-\theta^*_{S^c}| \le 3|\hat{\theta}_S-\theta^*_S|$$

which is exactly the cone condition! Everything below this is kinda suspicious because I was playing squardle instead of paying attention. So for our lower bound, we get

$$\frac{2|\mathbb{X}(\hat{\theta}-\theta^*)|_2^2}{n} \ge |\hat{\theta}-\theta^*|_2^2$$

By Cauchy,

$$|\hat{\theta}_S-\theta^*|_1 \le \sqrt{k}|\hat{\theta}_s-\theta^*|_2 \le \sqrt{k}||\hat{\theta}-\theta^*|_2 \le \sqrt{\frac{2k}{n}}|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2$$

Therefore, we get

$$|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2^2 \le 4n\tau\sqrt{\frac{2k}{n}}|\mathbb{X}\hat{\theta}-\mathbb{X}\theta^*|_2$$

from which we divide and square to get the desired result.

§2 Matrix Estimation

We will go over some linear algebra "basics" which need to be known for later lectures. Apparently this lecture will be "boring to death" (not my words).

§2.1 SubGaussian Sequence Model

Our subGaussian sequence model is of the form $Y=\theta^*+\epsilon\in\mathbb{R}^d$. We can make this a matrix problem by just reshaping each vector into a matrix.

If θ^* is sparse, then we can just use $\hat{\theta}^{HARD}$, so we aren't utilizing matrix properties.

§2.2 An Aside: Netflix Prize 2006

Aka how Netflix got half the academic community to work for them for free. The problem is the following: consider matrix M, with n users and m movies, such that $M_{i,j}$ is how the ith person rated the jth movie.

Clearly, the matrix is very sparse. In fact, only 1% was filled. The goal was the fill the rest of the matrix.

§2.2.1 A Simple Model

Consider where M_{ij} only has two effects: user and movie. So,

$$M_{ij} = u_i \cdot v_j + \text{noise}.$$

For the simple model, we reduce the number of parameters from nm to $n + m$.

$$M = uv^T + \text{noise}$$

The rank of uv^T is 1. More generally, if the rank of M is r, we can write as

$$M = \sum_{j=1}^{r} u^{(j)} v^{(j)T}$$

§3 Matrix Redux

§3.1 Eigenvalues and Eigenvectors

Square matrix $A \in \mathbb{R}^{n \times n}$. Defines eigenvalue and eigenvector $Au = \lambda u$.

Fact 3.1. If A is symmetric, then all eigenvalues are real.

In this class, we will assume that all eigenvectors have norm 1.

Fact 3.2. If $u_1, \ldots u_n$ eigenvectors of symmetric A, they can form an orthogonal basis for column span of A. We will call this the **eigenbasis**.

§3.2 Singular Value Decomposition

Let $A \in \mathbb{R}^{m \times n}$. The **SVD** of A is A written as

$$A = UDV^T, \;\; U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{r \times n}, D \in \mathbb{R}^{r \times r}$$

where r is the rank of A, $U^T U = I_r$, $V^T V = I_r$, D is diagonal with positive entries.

This implies that $u_1, u_2, \ldots \in \text{colspan}(A)$ and $v_1^T, v_2^T, \ldots v_n^T \in \text{rowspan}(A)$.

The vector form of this is

$$A = \sum_{j=1}^{r} \lambda_j u_j v_j^T$$

> **Remark 3.3.** We have $AA^T u_j = \lambda_j^2 u_j$ and $A^T A v_j = \lambda_j^2 v_j$.

Consider the special case when A is positive semidefinite. The eigenvalues are positive and are equal to the singular values. U and V become the same matrix. In this case,

$$||A||_{op} = \max_{x \in B_2^m} |Ax|_2 = \lambda_{max}(A)$$

§3.3 Vector Norms and Inner Products

Let A and B be matrices. The **q-norm** is defined as

$$|A|_q = \left(\sum_{ij} |A_{ij}|^q\right)^{1/q}$$

> **Remark 3.4.** Note that $|A|_\infty = \max |A_{ij}|$ and $|A|_0$ is the number of nonzero entries. We also have $|A|_2 = \sqrt{Tr(A^T A)} = \sqrt{Tr(AA^T)} = ||A||_F$.

Then we can define the inner product

$$\langle A, B\rangle = Tr(A^T B) = Tr(AB^T)$$

§3.4 Spectral Norms

Let A have singular values $\lambda_1, \ldots, \lambda_r$. Consider vector $\lambda = (\lambda_1, \ldots, \lambda_r)$. The **Schatten q-norm** is defined as

$$||A||_q = |\lambda|_q$$

When $q = 2$, we have

$$||A||_2^2 = |\lambda|_2^2 = ||A||_F^2 = |A|_2^2$$

which can be derived trivially by plugging in SVD into $Tr(A^T A)$.

When $q = 1$, we call this the **nuclear/trace norm**.

$$||A||_1 = |\lambda|_1 = \sum \lambda_j = ||A||_A$$

§3.5 Matrix Inequalities

Let A and B be positive semidefinite. Order their eigenvalues in decreasing order.

Theorem 3.5

Weyl. We have

$$\max_j |\lambda_j(A) - \lambda_j(B)| \leq ||A - B||_{op}$$

Theorem 3.6

Hoffman-Wielaudt. We have

$$\sum_j |\lambda_j(A) - \lambda_j(B)|^2 \leq ||A - B||_F^2$$

Theorem 3.7

Holder. We have for $\frac{1}{p} + \frac{1}{q} = 1$,

$$\langle A, B\rangle \leq ||A||_p ||B||_q$$

§3.6 Eckert-Young

Also known as best rank-k approximation.

Lemma 3.8

Let matrix A be of rank r. Look at SVD $A = \sum_{j=1}^{r} \lambda_j u_j v_j^T$ and assume singular values are in decreasing order. For any $k \leq r$, define the truncated SVD

$$A = \sum_{j=1}^{k} \lambda_j u_j v_j^T$$

This matrix has rank k. Then, we have

$$||A - A_k||_F^2 = \inf_{\text{rank}(B) \leq k} ||A - B||_F^2 = \sum_{j=k+1}^{r} \lambda_j^2$$