Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-3 — Week 14

Instructor: Stas Minsker *Scribe: S. Minsker*

1.1 Matrix Recovery Problems and Sparse Linear Regression

Our final goal for this course is to understand the matrix recovery problems with Gaussian and non-Gaussian designs. We will study the general "matrix regression" framework, and results for the usual "vector" regression will be obtained as corollaries.
Let $A_1, A_2 \in \mathbb{R}^{m_1 \times m_2}$, and define the inner product

$$\langle A_1, A_2 \rangle := \mathrm{tr}(A_1^T A_2) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (A_1)_{ij} (A_2)_{ij},$$

where for $A \in \mathbb{R}^{m \times m}$, $\mathrm{tr}(A) = \sum_{i=1}^m A_{ii}$ is the trace of A. The corresponding norm is the Frobenius norm $\|A\|_F^2 := \langle A, A \rangle$. Assume that (X, Y) is a random couple where $X \in \mathbb{R}^{m_1 \times m_2}$ and $Y \in \mathbb{R}$, and that

$$\mathbb{E}(Y|X) = \langle A_0, X \rangle$$

for some fixed matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$. Moreover, let

$$\xi = Y - \langle A_0, X \rangle$$

be the "noise" variable, so that $Y = \langle A_0, X \rangle + \xi$ that we will refer to as the *trace regression model.* We will assume that $\mathrm{Var}(\xi) = \sigma^2 < \infty$ (also, $\mathbb{E}\xi = 0$ by its definition). Assume that we observe n i.i.d. copies (X_j, Y_j), $j = 1, \ldots, n$ of (X, Y) so that

$$Y_j = \langle A_0, X_j \rangle + \xi_j, \ j = 1, \ldots, n. \tag{1.1}$$

We will also assume that the distribution of the random matrix X is known (this is the case in problems of interest to us). The goal is to estimate A_0 based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Example 1. Assume that $X_j = \mathrm{diag}(x_{j,1}, \ldots, x_{j,m})$ are diagonal matrices, $A_0 = \mathrm{diag}(a_1^0, \ldots, a_m^0)$. Then

$$Y_j = \sum_{i=1}^m x_{j,i} a_i^0 + \xi_j$$

is the usual linear regression model.

Example 2 (Matrix completion)**.** Let $e_j(m) = (0, \ldots, \overset{\overset{j}{\uparrow}}{1}, \ldots, 0)^T, \ j = 1, \ldots, m$ be the standard Euclidean basis, and

$$\mathcal{X} = \{e_i(m_1) e_j(m_2)^T, \ i = 1, \ldots, m_1, \ j = 1, \ldots, m_2\}$$

be the basis of $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$. Assume that X_j are i.i.d. with uniform distribution on $\mathcal{X}$. In this case, Y_j is the noisy version of a randomly selected entry of A_0, and the resulting problem is known as the matrix completion problem.

Example 3 (Matrix compressed sensing)**.** Assume that $x_{i,j} = e_i^T X e_j$ are independent $N(0,1)$ random variables and that $\xi = 0$. Then the model we consider is

$$Y_j = \langle A_0, X_j \rangle, \ j = 1, \ldots, n$$

that resembles the compressed sensing problem for sparse vectors that we investigated before.

What is the natural definition of "sparsity" for the matrix? The most obvious choice, entry-wise sparsity, is non invariant with respect to the change of basis in which the matrix A_0 is represented, so it is often not a good choice. A more fruitful notion of "sparsity" is the rank of the matrix (which does not depend on the choice of the basis). It is a generalizatin of vector sparsity in a sense that a low-rank diagonal matrix corresponds to a sparse "diagonal" vector.

Assumption. A_0 has small rank or can be well-approximated by a matrix of small rank.

Note that for diagonal matrices, the rank is equal to the degree of sparsity of the diagonal.

Let us consider the framework of example 3 where X_j's are matrices with i.i.d. $N(0,1)$ entries. A natural problem to study would be the "rank minimization problem"

$$\begin{aligned} &\operatorname{rank}(A) \to \min_{A \in \mathbb{R}^{m_1 \times m_2}} \\ &\text{s.t. } \langle X_j, A \rangle = Y_j, \ j = 1, \ldots, n. \end{aligned} \tag{1.2}$$

This problem is non-convex and hard to solve directly. Following the approach explained last week, previously, we should try to replace the set of low-rank matrices with something convex. For the case of sparse vectors, we used the $\|\cdot\|_1$-norm unit ball which is the convex hull of 1-sparse vectors. Hence, a natural analogue would be the set of rank-1 matrices of Frobenius norm 1. Any such rank-1 matrix in $\mathbb{R}^{m_1 \times m_2}$ can be written as $uv^T, u \in \mathbb{R}^{m_1}, v \in \mathbb{R}^{m_2}$. Consider the set

$$K = \{uv^T, u \in \mathbb{R}^{m_1}, v \in \mathbb{R}^{m_2}\}.$$

We will prove that the convex hull of K is the unit ball in the nuclear norm.

Definition 1. The nuclear norm (trace norm, Schatten-1 norm) is defined as

$$\|A\|_* = \sum_{j=1}^{\operatorname{rank}(A)} \sigma_j(A),$$

where $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_r(A)$ are the singular values of A and $r = \operatorname{rank}(A)$.

Lemma 1. $\operatorname{co}(K) = \{A \in \mathbb{R}^{m_1 \times m_2}, \|A\|_* \leq 1\}$.

Proof.
(a) Assume that $\sum_j \sigma_j(A) \le 1$, then

$$A = \sum_j \sigma_j(A) u_j v_j^T \in \mathrm{co}(K),$$

since $\sigma_j(A) \ge 0$ by defintion.
(b) Next, assume that $A \in \mathrm{co}(K)$ so that it can be written as

$$A = \sum_{j=1}^{l} \alpha_j u_j v_j^T$$

for some $l \ge 1$ and $u_j \in \mathbb{R}^{m_1},\ v_j \in \mathbb{R}^{m_2},\ j = 1, \ldots, l$. We will need the following fact: for any k between 1 and $r = \mathrm{rank}(A)$,

$$A \mapsto \sigma_1(A) + \ldots + \sigma_k(A) := |||A|||_l\,,$$

the sum of k largest singular values of A, defines a norm, the so-called Ky-Fan norm (to prove it, use the equality $\sigma_1(A) + \ldots + \sigma_k(A) = \sup_{u_j, v_j} \sum_{j=1}^k u_j^T A v_j$). Then

$$\begin{aligned}\sum_{j=1}^{r} \sigma_j\left(\sum_{i=1}^{l} \alpha_i u_i v_i^T\right) &= \left|\left|\left|\sum_{i=1}^{l} \alpha_i u_i v_i^T\right|\right|\right|_r \\ &\le \sum_{i=1}^{l} \left|\left|\left| u_i v_i^T \right|\right|\right|_r = \sum_{i=1}^{l}\sum_{j=1}^{r} \sigma_j\left(\alpha_i u_i v_i^T\right) = \sum_{i=1}^{l} \alpha_i = 1.\end{aligned}$$

■

Hence, it is natural to replace (1.2) by the following convex relaxation:

$$\begin{aligned}&\|A\|_* \to \min_{A \in \mathbb{R}^{m_1 \times m_2}} \\ &\text{s.t. } \langle X_j, A\rangle = Y_j,\ j = 1, \ldots, n.\end{aligned} \tag{1.3}$$

As shown in the previous lecture, to understand the expected error $\mathbb{E}\|\widehat{A} - A_0\|_F$ where $\widehat{A}_0$ is the solution to problem (1.4), we need to bound the Gaussian mean width of the unit ball with respect to the nuclear norm

$$B_* := B_*(0,1) = \{A \in \mathbb{R}^{m_1 \times m_2} : \|A\|_* \le 1\}.$$

To this end, let us recall the notion of the dual norm, and prove that the spectral norm is the dual of the nuclear norm.

Definition 2. Let $\|\cdot\|'$ be the norm in $\mathbb{R}^d$. The dual norm of $\|\cdot\|'$ is defined via

$$\|x\|_*' = \sup_{\|x\|' \le 1} \langle y, x\rangle$$

Lemma 2. The dual of the nuclear norm of the nuclear norm is the operator (spectral) norm.

Reminder. The operator norm of $A \in \mathbb{R}^{m_1 \times m_2}$ is

$$\|A\| = \max_{j=1,\dots,\mathrm{rank}(A)} \sigma_j(A).$$

Proof. Let $A = U\Sigma V^T$ be the SVD of A.
(a) First, we will prove that $\sup_{\|Q\|\le 1}\langle Q, A\rangle \ge \|A\|_*$. Take $Q = UV^T$, then

$$\langle Q, A\rangle = \mathrm{tr}(VU^TU\Sigma V^T) = \mathrm{tr}(\Sigma V^TV) = \mathrm{tr}(\Sigma) = \|A\|_*,$$

where we used the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
(b) It remains to show that $\sup_{\|Q\|\le 1}\langle Q, A\rangle \le \|A\|_*$. Indeed,

$$\begin{aligned}\sup_{\|Q\|\le 1}\langle Q, A\rangle &= \sup_{\|Q\|\le 1}\mathrm{tr}(Q^TU\Sigma V^T) = \sup_{\|Q\|\le 1}\mathrm{tr}(V^TQ^TU\Sigma) \\ &= \sup_{\|Q\|\le 1}\langle \Sigma, U^TQV\rangle = \sup_{\|Q\|\le 1}\sum_{j=1}^{\mathrm{rank}(A)}\sigma_j(A)(U^TQV)_{jj} \\ &= \sup_{\|Q\|\le 1}\sum_{j=1}^{\mathrm{rank}(A)}\sigma_j(A)\underbrace{\langle Qv_j, u_j\rangle}_{\le \sigma_1(Q)=\|Q\|} \le \sum_{j=1}^{\mathrm{rank}(A)}\sigma_j(A) = \|A\|_*.\end{aligned}$$

■

Theorem 1. The Gaussian mean width of a B_* satisfies

$$w(B_*) \le 2\left(\sqrt{m_1} + \sqrt{m_2}\right).$$

Proof. Let $X \in \mathbb{R}^{m_1\times m_2}$ have independent $N(0,1)$ entries and note that

$$w(B_*) = 2\mathbb{E}\sup_{A\in B_*}\langle A, X\rangle.$$

In view of duality between the spectral and nuclear norms, $\langle A, X\rangle \le \|A\|_*\|X\| = \|X\|$. Next, the celebrated theorem by Yehoram Gordon states that $\mathbb{E}\|X\| \le \sqrt{m_1} + \sqrt{m_2}$, and the claim follows. ■

Corollary 1. The solution of the problem (1.4) satisfies

$$\mathbb{E}\|\widehat{A} - A_0\|_F \le C\|A_0\|_*\sqrt{\frac{m_1+m_2}{n}} \le C\|A_0\|_F\sqrt{\frac{r(A_0)(m_1+m_2)}{n}}.$$

Here, $r(A_0)$ is the rank of A_0 and the second inequality follows from the fact that for any matrix A,

$$\|A\|_F \le \sqrt{\mathrm{rank}(A)}\|A\|_F \text{ (why?)}$$

Exercise. Try proving a weaker version of Gordon's theorem, namely, that $\mathbb{E}\|X\| \le C(\sqrt{m_1} + \sqrt{m_2})$ for some absolute constant $C > 0$. You may first show that

$$\mathbb{E}\|X\| = \mathbb{E}\sup_{\|v\|_2=1}\|Xv\|_2 \le C(\varepsilon)\mathbb{E}\max_{v\in S_\varepsilon}\|Xv\|_2$$

where S_ε is the ε-net of the unit sphere $\{x : \|x\|_2 = 1\}$. Then fix the value of ε (say, $\varepsilon = 1/3$), estimate the cardinality of S_ε and use the properties of sub-Gaussian random variables to estimate the resulting expression.

1.2 Additional material not covered during the lectures.

When the measurements are noisy (that is, $P(\xi \neq 0) > 0$) we will consider the following version of problem (1.3):

$$\|A\|_* \to \min_{A \in \mathbb{R}^{m_1 \times m_2}} \tag{1.4}$$
$$\text{s.t. } \frac{1}{n}\sum_{j=1}^{n} (Y_j - \langle X_j, A\rangle)^2 \le t.$$

A closely related problem (the "Lagrangian form" of (1.4)) is the penalized minimization problem

$$\tilde{A}_\tau = \operatorname*{argmin}_{A \in \mathbb{R}^{m_1 \times m_2}} \left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \langle X_i, A\rangle)^2 + \tau \|A\|_* \right]$$

where $\tau > 0$ (the exists 1-1 correspondence between t and τ so that solutions to the constrained and penalized forms of the problem coincide). Observe that

$$\frac{1}{n}\sum_{j=1}^{n} (Y_j - \langle A, X_j\rangle)^2 = \cancelto{\text{does not depend on } A}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} + \frac{1}{n}\sum_{j=1}^{n} \langle A, X_j\rangle^2 - \left\langle \frac{2}{n}\sum_{j=1}^{n} Y_j X_j, A \right\rangle.$$

Since the distribution Π of X_j's is assumed to be known, it is natural to replace the empirical average $\frac{1}{n}\sum_{j=1}^{n}\langle A, X_j\rangle^2$ by its expected value

$$\|A\|_{L_2(\Pi)}^2 := \mathbb{E}\langle A, X_1\rangle^2.$$

Even when X_j's are deterministic, we will set

$$\|A\|_{L_2(\Pi)}^2 := \frac{1}{n}\sum_{j=1}^{n}\langle A, X_j\rangle^2.$$

Finally, we get the following problem:

$$\hat{A}_\tau = \operatorname*{argmin}_{A \in \mathbb{R}^{m_1 \times m_2}} \left[\|A\|_{L_2(\Pi)}^2 - \left\langle \frac{2}{n}\sum_{j=1}^{n} Y_j X_j, A \right\rangle + \tau \|A\|_* \right]. \tag{1.5}$$

Example (Matrix completion)**.** Recall that in the matrix completion example, $X_j \sim U(\mathcal{X})$, hence

$$\mathbb{E}\langle A, X_1\rangle^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{1}{m_1 m_2} A_{ij}^2 = \frac{1}{m_1 m_2}\|A\|_F^2.$$

Example (Matrix compressed sensing)**.** If $(X_1)_{ij}$ are i.i.d. standard normal, then

$$\mathbb{E}\langle A, X_1\rangle^2 = \|A\|_F^2.$$

Example (LASSO)**.** When the design matrices are diagonal, problem (1.5) is known as LASSO – "Least Absolute Shrinkage & Selection Operator". In this case, A and $X_j,\ j = 1, \dots, n$ can be identified as vectors in $\mathbb{R}^p$, and

$$\hat{A}_\tau = \underset{A\in\mathbb{R}^p}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{j=1}^n \langle A, X_j\rangle^2 - \left\langle \frac{2}{n}\sum_{j=1}^n Y_j X_j, A \right\rangle + \tau\|A\|_1 \right].$$

Note that the nuclear norm of a diagonal matrix is simply the $\|\cdot\|_1$ norm of its diagonal.

1.2.1 Subdifferential of the Nuclear Norm

Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex. The subdifferential of f at the point x is defined as

$$\partial f(x) = \{y \in \mathbb{R}^p,\ f(x) \geq f(z) + \langle y, z - x\rangle \text{ for all } z \in \mathbb{R}^p\}.$$

We will need the following fact (try to prove it as an exercise): the subdifferental of any norm $\|\cdot\|'$ with dual norm $\|\cdot\|'_*$ is

$$\partial\|x\|' = \begin{cases} y :\ \|y\|'_* = 1,\ \|x\|' = \langle y, x\rangle, \text{ when } x \neq 0, \\ y :\ \|y\|'_* \leq 1 \text{ when } x = 0. \end{cases}$$

Applied to the nuclear norm, this yields

$$\partial\|A\|_* = \begin{cases} W \in \mathbb{R}^{m_1\times m_2} :\ \|W\| = 1,\ \|A\|_* = \langle y, x\rangle, \text{ when } A \neq 0, \\ W \in \mathbb{R}^{m_1\times m_2} :\ \|W\| \leq 1 \text{ when } A = 0. \end{cases}$$

1.2.2 Analysis of the estimator (1.5)

Given a matrix A of rank r with the singular value decomposition $A = \sum_{j=1}^r \sigma_j(A) u_j v_j^T$, define the subspaces S_1 and S_2 as

$$S_1 = \text{lin.span}\,(u_1, \dots, u_r)\,,$$
$$S_2 = \text{lin.span}\,(v_1, \dots, v_r)\,.$$

We will call (S_1, S_2) the *support* of A. Given a linear subspace L, let P_L stand for the orthogonal projector onto L.

For any $A \in \mathbb{R}^{m_1\times m_2}$, we define the linear operators $\mathcal{P}_A,\ \mathcal{P}_A^\perp : \mathbb{R}^{m_1\times m_2} \mapsto \mathbb{R}^{m_1\times m_2}$ via

$$\mathcal{P}_A^\perp(B) = P_{S_1^\perp} B P_{S_2^\perp},$$
$$\mathcal{P}_A(B) = B - P_{S_1^\perp} B P_{S_2^\perp} = B - \mathcal{P}_A^\perp(B).$$

We will need to consider the norm of $\mathcal{P}_A(B)$ restricted to a cone

$$\mathcal{C}(A, c_0) := \left\{B \in \mathbb{R}^{m_1\times m_2} :\ \left\|\mathcal{P}_A^\perp(B)\right\|_* \leq c_0 \left\|\mathcal{P}_A(B)\right\|_*\right\},$$

where $c_0 > 0$ is a constant (we will take $c_0 = 5$ in what follows). The "restricted norm" is defined as

$$\mu_{c_0}(A) = \inf\left\{\kappa > 0 :\ \left\|\mathcal{P}_A(B)\right\|_{\mathrm{F}} \leq \kappa \left\|B\right\|_{L_2(\Pi)} \text{ for all } B \in \mathcal{C}(A, c_0)\right\}.$$

Remark 1 (Matrix completion)**.** In the matrix completion case, we have

$$\|\mathcal{P}_A(B)\|_{\mathrm{F}} \le \|B\|_{\mathrm{F}} = m_1 m_2 \|B\|_{L_2(\Pi)},$$

hence $\mu_{c_0}(A) \le m_1 m_2$ for any $c_0 > 0$.

Remark 2 (Linear regression)**.** We will identify the unknown matrix A_0 with a vector λ_0 (to keep notation consistent with previous lectures), and consider the model

$$Y_j = \langle \lambda_0, X_j \rangle + \xi_j$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. and ξ_j's are independent from X_j's. In this case, vectors X_j, $j = 1, \ldots, n$ are identified with diagonal matrices, and $S_1(\lambda) = S_2(\lambda)$ is the subspace spanned by the basis vectors indexed by the "support" $J(\lambda)$ of λ. The operator $\mathcal{P}_\lambda(\lambda')$ acts by restricting λ' to $J(\lambda)$. The nuclear norm becomes the $\|\cdot\|_1$-norm, and the cone $\mathcal{C}(A, c_0)$ becomes the "cone of dominant coordinates" that we have considered before. Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \hline \vdots \\ \hline X_n \end{pmatrix}$$

to be the design matrix. Then

$$\|\lambda'\|^2_{L_2(\Pi)} = \frac{1}{n}\sum_{j=1}^{n} \langle \lambda', X_j \rangle^2 = \frac{1}{n}\left\|\mathbf{X}\lambda'\right\|_2^2,$$

and it is easy to see that $\mu_{c_0}(\lambda) = \frac{1}{\theta(c_0,\lambda)}$, where

$$\theta(c_0, \lambda) = \sup\left\{\tau > 0 : \frac{1}{n}\left\|\mathbf{X}\lambda'\right\|_2 \ge \tau \|\lambda'_{J(\lambda)}\|_2 \text{ for all } \lambda' \in \mathbb{R}^p, \right\}$$

where $\lambda'_{J(\lambda)}$ is the restriction of λ' onto the support of λ. The quantity

$$\min_{\lambda \in \mathbb{R}^p : \mathrm{card}(J(\lambda)) \le s} \theta(c_0, \lambda)$$

is known as the *restricted eigenvalue* of $\mathbf{X}$ corresponding to the sparsity level s. It follows from the result below that λ_0 can be estimated accurately if the restricted eigenvalue corresponding to $s = \mathrm{card}\,(\mathrm{supp}(\lambda_0))$ is not too small.

Our main result is the following theorem for the trace regression model (1.1) and the estimator defined in (1.5):

Theorem 2. Define the random matrix $M := \frac{1}{n}\sum_{j=1}^{n} (Y_j X_j - \mathbb{E}\,(Y_j X_j))$. The following inequality holds on the event $\mathcal{E} = \{\tau \ge 3\|M\|\}$:

$$\left\|\hat{A}_\tau - A_0\right\|^2_{L_2(\Pi)} \le \inf_{S \in \mathbb{R}^{m_1 \times m_2}} \left[\|S - A_0\|_{L_2(\Pi)} + \tau^2 \mu_5^2(S)\mathrm{rank}(S).\right]$$

The proof of this theorem will not be reproduced in the notes but can be found in the paper [**?**] (see Theorem 2 in that paper; you will also need to read the proof of Theorem 1). We will however prove the *Matrix Bernstein inequality*, one of the key technical tools that is required to estimate the size of $\|M\|$ and to find τ such that the event $\mathcal{E} = \{\tau \ge 3\|M\|\}$ holds with high probability. For the details on the application of the Matrix Bernstein inequality to trace regression model (1.1) in general, and to matrix completion and LASSO in particular, see section 6 in the paper [**?**] (this material has not been covered in class).

1.3 Matrix Bernstein inequality

The notes in this section were prepared by L. Goldstein based on the notes by Joel Tropp [?].
We will first recall several useful facts that we will rely on in the proofs.

1. Complex matrices B in $\mathbb{C}^{d_1 \times d_2}$. Trace $\mathrm{tr}(B) = \sum_{i=1}^{\min(d_1,d_2)} b_{ii}$, the sum of the diagonal elements of B. The matrix B is square when $d_1 = d_2$.

2. The matrix H is Hermitian when $H^* = H$ where H^* is the conjugate transpose of H. The set of all $d \times d$ Hermitian matrices is denoted $\mathbb{H}_d$. If H is Hermitian, it admits a diagonalization via a (unitary) imatrix Q satisfying $QQ^* = I$,

   $$H = Q\mathrm{diag}(\lambda_1, \cdots, \lambda_d)Q^*.$$

   The eigenvalues $\lambda_1, \ldots, \lambda_d$ of H are real, and the maximum eigenvalue is denoted $\lambda_{\max}(H)$, or just $\lambda_{\max}$, and the minimum one similarly.

   As the trace is rotationally invariant, that is $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, one consequence of this representation is that for $H \in \mathbb{H}_d$,

   $$\mathrm{tr}(H) = \sum_{j=1}^{d} \lambda_j. \tag{1.6}$$

   We say that $H \in \mathbb{H}^d$ is positive semi-definite when

   $$u^* H u \geq 0 \quad \text{for all } u \in \mathbb{C}^d.$$

   The matrix $H \in \mathbb{H}^d$ is positive semi-definite if and only if all its eigenvalues of are non-negative. In this case, by (1.6), we have

   $$\lambda_{\max}(H) \leq \mathrm{tr}(H).$$

3. A general matrix B has a singular value decomposition

   $$B = Q_1 \Sigma Q_2^*$$

   where Q_1 and Q_2 are unitary matrices, and Σ is a non-negative diagonal matrix. The spectral norm $\|B\|$ is the largest singular value of B, or equivalently $\|B\| = \max_i \sqrt{\lambda_i(B^*B)}$. For Hermitian matrices one has $\|A\| = \max\{\lambda_{\max}(A), -\lambda_{\min}(A)\}$.

4. For $d \times d$ Hermitian matrices A and H, we say A is less than H in the positive semidefinite order, and write $A \preceq H$, if

   $$u^* A u \leq u^* H u \quad \text{for all } u \in \mathbb{C}^d.$$

   Equivalently, $A \preceq H$ if $H - A$ is a positive semi-definite matrix. The positive semi definite order is preserved by conjugation:

   $$\text{if} \quad A \preceq H \quad \text{then} \quad B^*AB \preceq B^*HB.$$

5. Standard matrix functions. If $f : I \to R$ and the eigenvalues of a Hermitian matrix A lie in I, then let $f(A)$ be the matrix

$$f(A) = Q\text{diag}(f(\lambda_1), ...f(\lambda_d))Q^*.$$

Note that this definition agrees with matrix powers for $f(x) = x^n$, hence also agrees when f is given by a power series.

The Spectral mapping Theorem is an immediate consequence: If λ is an eigenvalue of A, then $f(\lambda)$ is an eigenvalue of $f(A)$.

6. Transfer rule: Let f and g be real valued functions on I, where the eigenvalues of the Hermitian matrix A lie. Then $f(a) \leq g(a)$ for all $a \in I$ implies $f(A) \preceq g(A)$.

7. With $\lambda_i(\cdot)$ denoting the i^{th} largest eigenvalue of a Hermitian matrix, when $A \preceq B$, the Courant-Fisher theorem yields that $\lambda_i(A) \leq \lambda_i(B)$. In some detail, by the characterization of the ordered eigenvalues that this result provides, we have

$$\lambda_i(A) = \max_{\dim L = i} \min_{u \in L} \frac{u^*Au}{u^*u} \leq \max_{\dim L = i} \min_{u \in L} \frac{u^*Bu}{u^*u} = \lambda_i(B),$$

where the maximum ranges over all i-dimensional linear subspaces L in the domain of the matrix, and where we apply the convention that $0/0 = 0$.

In particular, if f is a (weakly) increasing function on an interval that contains the eigenvalues of the Hermitian matrices A and B, then

$$A \preceq B \quad \text{implies} \quad \text{tr}f(A) \leq \text{tr}f(B).$$

8. The matrix exponential $\exp(A)$ is given as a standard matrix function, or equivalently as a power series. As e^x is an increasing function, by point 7 the trace exponential

$$A \to \text{tr}\exp(A),$$

has the property that

$$A \preceq H \quad \text{implies} \quad \text{tr}\exp(A) \leq \text{tr}\exp(H).$$

9. The matrix log is given as a standard matrix function, or as a power series. It is the inverse of the exponential function, so $\log\exp(A) = A$.

The matrix log is operator monotone, that is,

$$A \preceq H \quad \text{implies} \quad \log A \preceq \log H.$$

This same property does not hold for the matrix exponential.

10. Dilation. One can map any matrix $B \in \mathbb{C}^{d_1 \times d_2}$ to $H_{d_1+d_2}$ by

$$\mathcal{H}(B) = \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix},$$

for which

$$\lambda_{\max}(\mathcal{H}(B)) = \|\mathcal{H}(B)\| = \|B\|.$$

Hence information on the spectral norm of general matrices may be obtained via information on Hermitian matrices.

11. The expectation of a random matrix is given by taking component wise expectation, and it preserves the positive semi-definite order for Hermitian matrices. The variance of a Hermitian matrix Y is $\mathbb{E}(Y - \mathbb{E}Y)^2$. As for scalar valued random variables, the variance of the sum of independent Hermitian matrices is the sum of the variances.

    Define the variance parameter

    $$v(Y) = \|\mathrm{Var}(Y)\|$$

12. Matrix Laplace transform. If Y is a random Hermitian matrix, then for all $t \in \mathbb{R}$,

    $$\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) \leq \inf_{\theta>0} e^{-\theta t}\mathbb{E}\mathrm{tr}\, e^{\theta Y} \quad \text{and} \quad \mathbb{E}\lambda_{\max}(Y) \leq \inf_{\theta>0} \frac{1}{\theta} \log \mathbb{E}\mathrm{tr}\, e^{\theta Y}.$$

    The first inequality follows by Markov's inequality and the homogeneity of $\lambda_{\max}$, as for $\theta > 0$ we have

    $$\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) = \mathbb{P}\left(e^{\theta\lambda_{\max}(Y)} \geq \theta t\right) \leq e^{-\theta t}\mathbb{E}e^{\theta\lambda_{\max}(Y)} = e^{-\theta t}\mathbb{E}e^{\lambda_{\max}(\theta Y)},$$

    and that

    $$e^{\lambda_{\max}(\theta Y)} = \lambda_{\max}(e^{\theta Y}) \leq \mathrm{tr}\, e^{\theta Y},$$

    using the Spectral Mapping Theorem for the first identity and that the exponential function is increasing, that the exponential of a Hermitian matrix is positive semi-defnite, and that the trace of a positive semi-definte matrix is the sum of all its (non-negative) eigenvalues, which therefore dominates the maximum.

    For the second inequality, for $\theta > 0$ we have

    $$\mathbb{E}\lambda_{\max}(Y) = \frac{1}{\theta}\mathbb{E}\log e^{\lambda_{\max}(\theta Y)} \leq \frac{1}{\theta}\log \mathbb{E}e^{\lambda_{\max}(\theta Y)} = \frac{1}{\theta}\log \mathbb{E}\lambda_{\max}(e^{\theta Y}) \leq \frac{1}{\theta}\log \mathbb{E}\mathrm{tr}\, e^{\theta Y},$$

    where we have used Jensen's inequality, the Spectral Mapping Theorem, and that the trace dominates the eigenvalues of any positive semi-definite matrix.

13. Lieb's inequality: For a fixed Hermitian matrix $H \in \mathbb{H}_d$, the map

    $$A \to \mathrm{tr}\exp\left(H + \log A\right)$$

    is concave on the cone of positive definite matrices A in $\mathbb{H}_d$.

    Consequence, Subadditivity of the Matrix trace exponential: If $X_k, k = 1, \ldots, n$ are independent random Hermitian matrices of the same dimension, then

    $$\mathbb{E}\mathrm{tr}\exp\left(\sum_{k=1}^{n} \theta X_k\right) \leq \mathrm{tr}\exp\left(\sum_{k=1}^{n} \log \mathbb{E}e^{\theta X_k}\right)$$

14. Master inequalities for the sum of independent random Hermitian matrices $X_1, \dots, X_n$.

$$\mathbb{E}\lambda_{\max}\left(\sum_{k=1}^n X_k\right) \le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{k=1}^n \log \mathbb{E} e^{\theta X_k}\right)$$

and

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^n X_k\right) \ge t\right) \le \inf_{\theta>0} e^{-\theta t} \operatorname{tr} \exp\left(\sum_{k=1}^n \log \mathbb{E} e^{\theta X_k}\right)$$

1.3.1 Matrix Bernstein's inequality

Theorem 3. Let $X_1, \dots, X_n$ be independent random matrices in $\mathbb{H}_d$. Assume that there exists $L \ge 0$ such that for $k = 1, \dots, n$,

$$\mathbb{E}X_k = 0 \quad \text{and} \quad \lambda_{\max}(X_k) \le L.$$

Let

$$Y = \sum_{k=1}^n X_k \quad \text{and} \quad v(Y) = \|\mathbb{E}Y^2\| = \|\sum_{k=1}^n X_k^2\|.$$

Then

$$\mathbb{E}\lambda_{\max}(Y) \le \sqrt{2v(Y)\log d} + \frac{1}{3} L \log d,$$

and for all $t \ge 0$,

$$\mathbb{P}(\lambda_{\max}(Y) \ge t) \le d \exp\left(\frac{-t^2/2}{v(Y) + Lt/3}\right).$$

We first prove a moment and cumulant generating function bound for a random Hermitian matrix X with mean zero.

Theorem 4. Let X be a random Hermitian matrix such that for some L satisfies

$$\mathbb{E}X = 0 \quad \text{and} \quad \lambda_{\max}(X) \le L.$$

Then for all $0 < \theta < 3/L$,

$$\mathbb{E}e^{\theta X} \preceq \exp\left(\frac{\theta^2/2}{1-\theta L/3}\mathbb{E}X^2\right) \quad \text{and} \quad \log \mathbb{E}e^{\theta X} \preceq \frac{\theta^2/2}{1-\theta L/3}\mathbb{E}X^2.$$

Proof: Let $\theta > 0$ and write

$$e^{\theta X} = I + \theta X + (e^{\theta X} - \theta X - I) = I + \theta X + X f(X) X,$$

where

$$f(x) = \frac{e^{\theta x} - \theta x - 1}{x^2}, x \neq 0 \quad \text{and} \quad f(0) = \frac{\theta^2}{2}.$$

By taking derivatives, one may verify that the function f is increasing, hence $f(x) \le f(L)$ for all $x \le L$. As the eigenvalues of X are bounded above by L, writing $X = Q\text{diag}(\lambda_1, \ldots, \lambda_d)Q^*$ we have $\text{diag}(f(\lambda_1), \ldots, f(\lambda_d)) \le f(L)I$, and as the positive semi-definite order is preserved under conjugation, we have

$$f(X) \preceq f(L)I \quad \text{and} \quad Xf(X)X \preceq X(f(L)I)X,$$

and hence

$$e^{\theta X} \preceq I + \theta X + X(f(L)I)X = 1 + \theta X + f(L)X^2.$$

Now we develop a bound for $f(L)$ using Taylor series,

$$f(L) = \frac{e^{\theta L} - \theta L - 1}{L^2} = \frac{1}{L^2}\sum_{q=2}^{\infty}\frac{(\theta L)^q}{q!} \le \frac{1}{2\theta^2}\sum_{q=2}^{\infty}\frac{(\theta L)^{q-2}}{3^{q-2}} = \frac{\theta^2/2}{1 - \theta L/3},$$

where we have applied the bound $q! \ge 23^{q-2}$, valid for $q \ge 2$, and summed a geometric series, using that $0 < \theta < 3/L$. Hence we obtain, using that X^2 is positive semi-definite,

$$e^{\theta X} \preceq I + \theta X + \frac{\theta^2/2}{1 - \theta L/3}X^2.$$

As expectation preserves the positive semi-definite order, and using also that the mean of X is zero, we obtain

$$\mathbb{E}e^{\theta X} \preceq I + \frac{\theta^2/2}{1 - \theta L/3}\mathbb{E}X^2 \preceq \exp\left(\frac{\theta^2/2}{1 - \theta L/3}\mathbb{E}X^2\right),$$

where for the second order relation above we have applied the transfer rule for the inequality $1 + a \le e^a$, holding for all $a \in \mathbb{R}$. Hence the first claim of the theorem is shown, and the second follows using that the log is operator monotone. □

Proof of Theorem 3: Theorem 4 yields

$$\log \mathbb{E}e^{\theta X_k} \preceq g(\theta)\mathbb{E}X_k^2 \quad \text{where} \quad g(\theta) = \frac{\theta^2/2}{1 - \theta L/3}, 0 < \theta < 3/L.$$

To obtain a bound on $\mathbb{E}\lambda_{\max}(Y)$, by the master inequality we have

$$\begin{aligned}\mathbb{E}\lambda_{\max}(Y) &\le \inf_{\theta > 0}\frac{1}{\theta}\log \operatorname{tr}\exp\left(\sum_{k=1}^{n}\log \mathbb{E}e^{\theta X_k}\right)\\ &\le \inf_{0<\theta<3/L}\frac{1}{\theta}\log \operatorname{tr}\exp\left(g(\theta)\sum_{k=1}^{n}\mathbb{E}X_k^2\right)\\ &= \inf_{0<\theta<3/L}\frac{1}{\theta}\log \operatorname{tr}\exp\left(g(\theta)\mathbb{E}Y^2\right).\end{aligned}$$

where we have used that the trace exponential is monotone, and the additivity of the variance. Now

$$\begin{aligned}\mathbb{E}\lambda_{\max}(Y) &\leq \inf_{0<\theta<3/L} \frac{1}{\theta}\log\left[d\lambda_{\max}(\exp\left(g(\theta)\mathbb{E}Y^2\right))\right]\\ &\leq \inf_{0<\theta<3/L} \frac{1}{\theta}\log\left[d\exp\left(g(\theta)v(Y)\right)\right]\\ &= \inf_{0<\theta<3/L}\left[\frac{\log d}{\theta} + \frac{\theta/2}{1-\theta L/3}v(Y)\right],\end{aligned}$$

bounding the trace of the exponential by the dimension d times the maximum eigenvalue, then applying the Spectral Mapping Theorem. The optimal value of θ is given by

$$\theta = \frac{6L\log d + 9\sqrt{2v(Y)\log d}}{2L^2t + 9v(Y) + 6L\sqrt{2v(Y)\log d}}.$$

Next, for the tail bound, by the master inequality,

$$\begin{aligned}\mathbb{P}(\lambda_{\max} \geq t) &\leq \inf_{\theta>0} e^{-\theta t}\mathrm{tr}\exp\left(\sum_{k=1}^{n}\log\mathbb{E}e^{\theta X_k}\right)\\ &\leq \inf_{\theta>0} e^{-\theta t}\mathrm{tr}\exp\left(g(\theta)\sum_{k=1}^{n}\mathbb{E}X_k^2\right)\\ &= \inf_{\theta>0} e^{-\theta t}\mathrm{tr}\exp\left(g(\theta)\mathbb{E}Y^2\right)\\ &\leq \inf_{\theta>0} de^{-\theta t}\exp\left(g(\theta)v(Y)\right).\end{aligned}$$

Letting $\theta = t/(v(Y) + Lt/3)$ yields the result. □