Instructor: James R. Lee

1 Von Neumann's trace inequality

Last lecture, we studied the Frobenius inner product: $(A, B) \mapsto \mathrm{Tr}(A^*B)$. We used this to define the absolute value of a matrix $A \in \mathbb{M}_n(\mathbb{C})$:

$$|A| := (A^*A)^{1/2},$$

(which is always PSD), and the Schatten p-norms:

$$\|A\|_p := (|A|^p)^{1/p}.$$

And we observed that if $\sigma_1, \sigma_2, \ldots, \sigma_n \geqslant 0$ are the singular values of A, then

$$\|A\|_p = \|(\sigma_1, \ldots, \sigma_n)\|_p. \tag{1.1}$$

This follows readily from the fact that $|A|$ is a positive semidefinite matrix whose eigenvalues are the singular values of A.

While (1.1) is quite pleasant, it is not particularly surprising. Indeed, the Schatten p-norms are unitarily invariant: $\|UAV\|_p = \|A\|_p$ for all unitaries U, V, hence we can diagonalize. So the following inequality due to von Neumann is somewhat deeper, as it involves pairs of matrices.

Lemma 1.1. *If $A, B \in \mathbb{M}_n(\mathbb{C})$ have singular values $\sigma_1(A) \geqslant \cdots \geqslant \sigma_n(A)$ and $\sigma_1(B) \geqslant \cdots \geqslant \sigma_n(B)$, then*

$$|\mathrm{Tr}(A^*B)| \leqslant \sum_{i=1}^n \sigma_i(A)\sigma_i(B).$$

Note that both $|A|$ and $|A^*|$ are PSD matrices whose eigenvalues are the singular values of A, hence the following result will allow us to assume that A and B are PSD in Lemma 1.1.

Lemma 1.2. *For any $A, B \in \mathbb{M}_n(\mathbb{C})$, it holds that*

$$|\mathrm{Tr}(A^*B)| \leqslant \left(\mathrm{Tr}(|A||B|)\mathrm{Tr}(|A^*||B^*|)\right)^{1/2}. \tag{1.2}$$

Proof. Write the singular value decompositions $A = \sum_i \sigma_i u_i v_i^*$ and $B = \sum_i \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^*$, with $\sigma_i, \tilde{\sigma}_i \geqslant 0$ for all i, and $\{u_i\}, \{v_i\}, \{\tilde{u}_i\}, \{\tilde{v}_i\}$ orthonormal. Then by Cauchy-Schwarz,

$$|\mathrm{Tr}(A^*B)| = \left|\sum_{i,j} \sigma_i \tilde{\sigma}_j \langle u_i, \tilde{u}_j\rangle \langle v_i, \tilde{v}_j\rangle\right| \leqslant \left(\sum_{i,j} \sigma_i \tilde{\sigma}_j |\langle u_i, \tilde{u}_j\rangle|^2\right)^{1/2} \left(\sum_{i,j} \sigma_i \tilde{\sigma}_j |\langle v_i, \tilde{v}_j\rangle|^2\right)^{1/2}.$$

Now observe that the factors on the right are equal to $\mathrm{Tr}(|A||B|)^{1/2}$ and $\mathrm{Tr}(|A^*||B^*|)^{1/2}$, respectively:

$$|A| = (A^*A)^{1/2} = \left(\sum_i \sigma_i^2 v_i v_i^*\right)^{1/2} = \sum_i \sigma_i v_i v_i^*,$$

$$|B| = (B^*B)^{1/2} = \left(\sum_i \tilde{\sigma}_i^2 \tilde{v}_i \tilde{v}_i^*\right)^{1/2} = \sum_i \tilde{\sigma}_i \tilde{v}_i \tilde{v}_i^*,$$

$$|A^*| = (AA^*)^{1/2} = \left(\sum_i \sigma_i^2 u_i u_i^*\right)^{1/2} = \sum_i \sigma_i u_i u_i^*,$$

$$|B^*| = (BB^*)^{1/2} = \left(\sum_i \tilde{\sigma}_i^2 \tilde{u}_i \tilde{u}_i^*\right)^{1/2} = \sum_i \tilde{\sigma}_i \tilde{u}_i \tilde{u}_i^*. \qquad \square$$

Let us now prove von Neumann's inequality.

Proof of Lemma 1.1. Write $A = \sum_i a_i u_i u_i^*$ and $B = \sum_i b_i v_i v_i^*$, where $\{u_i\}$ and $\{v_i\}$ are orthogonal bases and $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$ and $b_1 \geqslant b_2 \geqslant \cdots \geqslant b_n$. Then:

$$\mathrm{Tr}(A^*B) = \mathrm{Tr}(AB) = \sum_{i,j} a_i b_j |\langle u_i, v_j\rangle|^2.$$

If we define the matrix $M_{ij} := |\langle u_i, v_j\rangle|^2$, we see that M is doubly-stochastic. Beacuse each $\{u_i\}$ and $\{v_i\}$ is an orthornomal basis, all the rows and columns sum to 1.

In other words, M is a fractional matching from $[n]$ to $[n]$, and hence (by the Birkhoff-von Neumann Theorem), M is a convex combination of permutation matrices. In other words, we can write

$$\mathrm{Tr}(A^*B) = \mathrm{Tr}(AB) = \sum_\pi w_\pi \sum_{i=1}^n a_i b_{\pi(i)},$$

where the sum is over all permutations of $[n]$, $\{w_\pi \geqslant 0\}$, and $\sum_\pi w_\pi = 1$. Hence

$$\mathrm{Tr}(A^*B) \leqslant \max_\pi \sum_{i=1}^n a_i b_{\pi(i)} = \sum_{i=1}^n a_i b_i. \qquad \square$$

Sketch of Birkhoff-von Neumann. Let $\mathcal{M}$ denote the fractional matching polytope on a complete bipartite graph (equivalently, the set of doubly stochastic matrices). Every polytope is equal to the convex hull of its extreme points (this is an exercise in convex geometry, and for more general compact convex sets, is the content of the Krein-Milman Theorem). Hence it suffices to show that all the extreme points of $\mathcal{M}$ are integral.

Consider a fractional perfect matching $M \in \mathcal{M}$. Let $E \subseteq [n] \times [n]$ denote the set of edges on which M is fractional. Then E cannot be a forest, as a leaf of E would have only a single fractional edge incident to it. Therefore E has a cycle, and since the graph is bipartite, this cycle has even length.

Let C be such a cycle and define $\varepsilon := \min_{(i,j)\in C} \min(M_{ij}, 1 - M_{ij})$. By adding ε to the even edges of C and subtracting ε from the odd edges, we obtain a new fractional matching M'. By subtracting ε from the even edgs and adding ε to the odd edges, we obtain another fractional matching M''. Moreover, we have $M = \frac{1}{2}(M' + M'')$, hence M is not an extreme point.

1.1 Symmetric gauge functions and the noncommutative Hölder inequality

Say that $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a *symmetric gauge function* if Φ is a norm that is invariant under permutations of the coordinates and under flipping signs of the coordinates. Let Φ^* denote the dual norm:

$$\Phi^*(u) := \max\{\langle v, u\rangle : \Phi(v) \leqslant 1\}.$$

Then we have the inequality:

$$|\langle u, v\rangle| \leqslant \Phi(u)\Phi^*(v). \tag{1.3}$$

When $\Phi(u) = \|u\|_2 = \Phi^*(u)$, this is the Cauchy-Schwarz inequality, and when $\Phi(u) = \|u\|_p$, we have $\Phi^*(u) = \|u\|_q$ where $1/p + 1/q = 1$, in which case this is Hölder's inequality. Of course, (1.3) holds without the assumption that Φ is symmetric.

But using the symmetry property and Lemma 1.1, we immediately obtain

$$|\mathrm{Tr}(A^*B)| \leqslant \Phi(\sigma(A))\Phi^*(\sigma(B)),$$

where $\sigma(A)$ denotes the vector of singular values of A. In particular, this gives the noncommutative Hölder inequalities:

Lemma 1.3. *For all $A, B \in \mathbb{M}_d(\mathbb{C})$, it holds that*

$$|\mathrm{Tr}(A^*B)| \leqslant \|A\|_p \|B\|_q \tag{1.4}$$

for $p, q \geqslant 1$ and $1/p + 1/q = 1$.

2 Majorization and unitarily invariant norms

Say that a norm $|||\cdot|||$ on $\mathbb{M}_n(\mathbb{C})$ is *unitarily invariant* if it holds that $|||UAV||| = |||A|||$ for all unitaries U, V and $A \in \mathbb{M}_n(\mathbb{C})$. If $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a symmetric gauge function, then

$$|||A||| := \Phi(\sigma(A))$$

is a unitarily invariant norm, as $\sigma(A)$—interpreted as a multiset—is unitarily invariant. In fact, this is a characterization: Every unitarily invariant norm is of this form (see [§I.4, BhatiaMA]).

It is straightforward that if A, B are PSD and $A \succeq B$, then $|||A||| \geqslant |||B|||$ for any unitarily invariant norm. Indeed, if we order the eigenvalues of A and B so that $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \lambda_n(A) \geqslant 0$ and $\lambda_1(B) \geqslant \lambda_2(B) \geqslant \cdots \geqslant \lambda_n(B) \geqslant 0$, then we have $\lambda_j(A) \geqslant \lambda_j(B)$ for all $j = 1, \ldots, n$. This can be seen using the variational characterization:

$$\lambda_k(A) = \max_{\dim(S)=k} \min\{\langle x, Ax\rangle : x \in S\},$$

where the maximum is over k-dimensional subspaces of $\mathbb{C}^n$. If we take S to be the span of the k eigenvectors of B corresponding to $\lambda_1(B), \ldots, \lambda_k(B)$, then we obtain

$$\lambda_k(A) \geqslant \min\{\langle x, Ax\rangle : x \in S\} \geqslant \{\langle x, Bx\rangle : x \in S\} \geqslant \lambda_k(B).$$

One might similarly ask how strong this consequence is: What does it mean that $|||A||| \geqslant |||B|||$ for every unitarily invariant norm?

2.1 Weak majorization

For a vector $x \in \mathbb{R}^n$, write $x^{\downarrow}$ for the vector that results from x by sorting the coordinates to be non-increasing:

$$x_1^{\downarrow} \geqslant x_2^{\downarrow} \geqslant \cdots \geqslant x_n^{\downarrow}.$$

For two vectors $x, y \in \mathbb{R}_+^n$, write $x \prec_w y$ if it holds that

$$x_1^{\downarrow} + x_2^{\downarrow} + \cdots + x_k^{\downarrow} \leqslant y_1^{\downarrow} + y_2^{\downarrow} + \cdots + y_k^{\downarrow}, \qquad \forall 1 \leqslant k \leqslant n.$$

We extend this to matrices $A, B \in \mathbb{M}_n(\mathbb{C})$, writing $A \prec_w B$ if $\sigma(A) \prec_w \sigma(B)$. We have the following characterization.

Theorem 2.1. *For every $x, y \in \mathbb{R}^n_+$, it holds that $\Phi(x) \leqslant \Phi(y)$ for every symmetric gauge function if and only if $x \prec_w y$. Equivalently, for every $A, B \in \mathbb{M}_n(\mathbb{C})$, it holds that $|||A||| \leqslant |||B|||$ for every unitarily invariant norm $|||\cdot|||$ if and only if $A \prec_w B$.*

One direction is straightforward because $\Phi_{(k)}(x) := |x_1^{\downarrow}| + \cdots + |x_k^{\downarrow}|$ is a symmetric gauge function for every $1 \leqslant k \leqslant n$. For the other direction, the next fact helps. Write $x \prec y$ if $x \prec_w y$ and, in addition, $x_1 + \cdots + x_n = y_1 + \cdots + y_n$.

Lemma 2.2. *For any $x, y \in \mathbb{R}^n$, it holds that $x \prec y$ if and only if x lies in the convex hull of permutations of y, i.e., $x = Py$, where P is a doubly-stochastic matrix. If holds that $x \prec_w y$ if and only if $x \leqslant u$ (pointwise) and $u \prec y$.*

Therefore for $x, y \in \mathbb{R}^n_+$, we have

$$x \prec_w y \implies \exists u \text{ s.t. } x \leqslant u, u = Py \implies \Phi(x) \leqslant \Phi(u) = \Phi(Py) \leqslant \Phi(y).$$

where the first inequality follows from the monotonicity of Φ, and the second from convexity of norms, along with symmetry of Φ.

2.2 Schur-convex functions

While $A \prec_w B$ is obviously a much weaker condition than $A \preceq B$, one of the advantages is that $\prec_w$ is monotone with respect to some interesting operations. For instance, the function $A \mapsto e^A$ is *not* operator monotone, i.e., there are Hermitian matrices A, B such that $A \preceq B$, but $e^A \npreceq e^B$. On the other hand, it does hold that $A \prec_w B \implies e^A \prec_w e^B$.

Lemma 2.3. *If $\varphi : \mathbb{R} \to \mathbb{R}$ is monotone increasing and convex, then $A \prec_w B \implies \varphi(A) \prec_w \varphi(B)$ for all $A, B \in \mathbb{M}_n(\mathbb{C})$.*

A consequence of this fact and Theorem 2.1 is that, while for a *particular* unitarily invariant norm, it might be that $|||A||| \leqslant |||B||| \not\Longrightarrow |||e^A||| \leqslant |||e^B|||$, the following does hold:

$$\begin{aligned} |||A||| \leqslant |||B||| \ \forall \text{ unitarily invariant } |||\cdot||| &\implies A \prec_w B \\ &\implies e^A \prec_w e^B \\ &\implies |||e^A||| \leqslant |||e^B||| \ \forall \text{ unitarily invariant } |||\cdot||| \end{aligned} \tag{2.1}$$

Generalized Golden-Thompson. Later, we will see the following generalization of the Golden-Thompson inequality: For all A, B Hermitian and unitarily invariant norms $|||\cdot|||$, it holds that $|||e^{A+B}||| \leqslant |||e^{B/2}e^A e^{B/2}|||$. In fact, we will prove the stronger inequality $|||A+B||| \leqslant |||\log(e^{B/2}e^A e^{B/2})|||$ and then use (2.1).

Hadamard's inequality and Schur's majorization. If $A \in \mathbb{M}_n(\mathbb{C})$ is PSD, then Hadamard's determinant inequality asserts that $\det(A) \leqslant \prod_{i=1}^n A_{ii}$. In attempting to understand this, Schur proved the following.

Theorem 2.4. *If $A \in \mathbb{M}_n(\mathbb{C})$ is Hermitian with eigenvalues $\lambda_1, \ldots, \lambda_n$, then we have the majorization*

$$(A_{11}, \ldots, A_{nn}) \prec (\lambda_1, \ldots, \lambda_n).$$

Proof. By Lemma 2.2, it suffices to prove that $(A_{11}, \ldots, A_{nn}) = P(\lambda_1, \ldots, \lambda_n)$ for some doubly-stochastic P. Observe that if $A = \sum_{i=1}^n \lambda_i u_i u_i^*$ and $\{u_i\}$ is an orthonormal basis, then

$$A_{jj} = \langle e_j, Ae_j \rangle = \sum_{i=1}^n \lambda_i \langle e_j, u_i u_i^* e_j \rangle = \sum_{i=1}^n \lambda_i |\langle u_i, e_j \rangle|^2 = \sum_{i=1}^n \lambda_i P_{ji},$$

where $P_{ji} = |\langle u_i, e_j \rangle|^2$. Since this P is doubly-stochastic, the proof is complete. □

To obtain Hadamard's inequality, note that since $(A_{11}, \ldots, A_{nn}) = P(\lambda_1, \ldots, \lambda_n)$, it holds that for $A \geq 0$,

$$\log(A_{jj}) = \log\left(\sum_{i=1}^n P_{ji}\lambda_i\right) \geqslant \sum_{i=1}^n P_{ji} \log(\lambda_i),$$

using concavity of the logarithm. Hence,

$$\log\left(\prod_{j=1}^n A_{jj}\right) = \sum_{j=1}^n \log(A_{jj}) \geqslant \sum_{i,j=1}^n P_{ji} \log(\lambda_i) = \sum_{j=1}^n \log(\lambda_j) = \log\det(A).$$