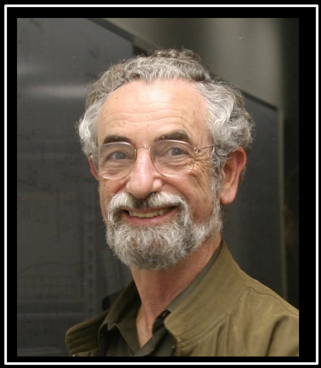
Lectures 3 and 4

(e) Lieb's theorem:

Let $H = H^*$ be fixed.
Then the function

$$A \mapsto \operatorname{tr} \exp(H + \log A)$$

is concave on a set of p.d. matrices.



Elliott H. Lieb

Theorem (Matrix Bernstein's inequality: Ahlswede-Winter '02, Oliveira '10, Tropp '11)

Let $X_1, \dots, X_n$ be independent self-adjoint random matrices such that $\mathbb{E} X_j = 0 \ \forall j$, and that $\|X_j\| \le L$ with prob. 1.

Let $\sigma^2 = \left\| \sum_{j=1}^{n} \mathbb{E} X_j^2 \right\|$ ("matrix variance").

Then (i) $$P\left(\left\| \sum_{j=1}^{n} X_j \right\| \ge t \right) \le 2d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + \frac{tL}{3}} \right),$$ and

(ii) $$\mathbb{E} \left\| \sum_{j=1}^{n} X_j \right\| \le \sqrt{2\sigma \cdot \log(2d)} + \frac{1}{3} L \log(2d).$$

Remark: (a) often, a more useful form of the bound is

$$P\left(\left\|\sum_{j=1}^{n} X_j\right\| \geq \max\left[2\sigma\sqrt{t}, \frac{4}{3} L t\right]\right) \leq 2d\, e^{-t}.$$

Indeed, consider two cases: (i) $\sigma^2 \geq \frac{tL}{3}$

(ii) $\sigma^2 < \frac{tL}{3}$

Then $P\left(\left\|\sum_1^n X_j\right\| \geq t\right) \leq 2d \max\left(e^{-\frac{t^2}{4\sigma^2}}, e^{-\frac{3t}{4L}}\right)$

Replace t by $\max\left(2\sigma\sqrt{s}, \frac{4}{3} L s\right)$ to get the result.

(b) Similar result holds if instead of the assumption $\|X_j\| \leq L$ w.p. 1, we require that $\mathbb{E} X_j^p \preceq \frac{p!}{2} L^{p-2} A_j^2$ for some $A_1, \dots, A_n$. Then $P\left(\left\|\sum_1^n X_j\right\| \geq t\right) \leq 2d \exp\left(\frac{-t^2/2}{\sigma^2 + Lt}\right)$, where $\sigma^2 = \left\|\sum_1^n A_j^2\right\|$.

Proof: (i) For any $M = M^*$, $\|M\| = \max(\lambda_{max}(M), -\lambda_{min}(M))$

$$P(\|Z\| \geq t) \leq P(\lambda_{max}(Z) \geq t) + P(-\lambda_{min}(Z) \geq t)$$

But $-\lambda_{min}(Z) = \lambda_{max}(-Z)$

$\Rightarrow P(-\lambda_{min}(Z) \geq t) = P(\lambda_{max}(-Z) \geq t)$

$\Rightarrow$ it suffices to bound $P\left(\lambda_{max}\left(\sum_{j=1}^n X_j\right) \geq t\right)$.

$$P\left(\lambda_{max}\left(\sum_{j=1}^n X_j\right) \geq t\right) = P\left(\lambda_{max}\left(\theta \sum_1^n X_j\right) \geq \theta t\right)$$

$$= P\left(\lambda_{max}\left(e^{\theta \sum_1^n X_j}\right) \geq e^{\theta t}\right)$$

$$\leq \mathbb{E}\,\lambda_{max}\left(e^{\theta \sum_1^n X_j}\right) \cdot e^{-\theta t}$$

$$\leq \mathbb{E}\,\mathrm{tr}\, e^{\theta \sum_1^n X_j} \cdot e^{-\theta t}$$

<u>Problem:</u> $\mathrm{tr}\, e^{\sum_j A_j} \neq \mathrm{tr}\, e^{A_1} \cdot \ldots \cdot e^{A_n}$

However, $\mathrm{tr}\, e^{A+B} \leq \mathrm{tr}(e^A \cdot e^B)$ – Golden - Thompson inequality

Instead, use Lieb's concavity theorem + Jensen's inequality:

$$\mathrm{tr}\, e^{\theta \sum_1^n X_j} = \mathrm{tr}\, e^{\theta \sum_{j=1}^{n-1} X_j + \log e^{X_n}}$$

$$\mathbb{E}\,\mathrm{tr}\, e^{\theta \sum_1^n X_j} = \mathbb{E}_{X_1,\ldots,X_{n-1}} \mathbb{E}\left[\mathrm{tr}\, e^{\overbrace{\theta \sum_1^{n-1} X_j}^{H} + \log \overbrace{e^{X_n}}^{A}} \,\middle|\, X_1 \ldots X_{n-1}\right]$$

$$\leq \mathbb{E}_{X_1 \ldots X_{n-1}} \mathrm{tr}\, e^{\theta \sum_1^{n-1} X_j + \log \mathbb{E} e^{X_n}}$$

$$\leq \ldots \leq \mathrm{tr}\, e^{\sum_1^n \log \mathbb{E} e^{X_j}}$$

$$\mathbb{E} e^{\theta X} = \mathbb{E}\left(I + \theta X + \frac{\theta^2 X^2}{2} + \ldots\right) =$$

$$= I + \frac{\theta^2}{2}\mathbb{E} X^2 + \frac{\theta^3}{3!}\mathbb{E} X^3 + \ldots$$

$$\mathbb{E} X^k = \mathbb{E}\, X \cdot X^{k-2} \cdot X \preceq L^{k-2}\, \mathbb{E} X^2$$

$$\Rightarrow \mathbb{E} e^{\theta X} \preceq I + \mathbb{E} X^2 \left(\frac{\theta^2}{2!} + \frac{L\theta^3}{3!} + \dots + L^{k-2} \frac{\theta^k}{k!} \right)$$

$$= I + \mathbb{E} X^2 \frac{\left(e^{L\theta} - L\theta - 1 \right)}{L^2}$$

$$\preceq e^{\mathbb{E} X^2 g(\theta, L)}, \text{ where } g(\theta, L) = \frac{e^{L\theta} - L\theta - 1}{L^2}$$

$$\Rightarrow \mathbb{E} \operatorname{tr} e^{\theta \sum_1^n X_i} \le \mathbb{E} \operatorname{tr} e^{g(\theta, L) \cdot \sum_1^n \mathbb{E} X_j^2}$$

$$\le d\, \mathbb{E}\, e^{g(\theta, L) \underbrace{\| \sum_1^n \mathbb{E} X_j \|^2}_{\sigma^2}}$$

Finally, it remains to minimize

$-\theta t + g(\theta, L) \sigma^2$ over $\theta > 0$: assume that $L = 1$ for simplicity (otherwise, rescale $X_j \to X_j / L$)

Then $\theta_* = \log\left(1 + \frac{t}{\sigma^2} \right)$

$$-\theta_* t + g(\theta, L) \sigma^2 = -\sigma^2 h\left(\frac{t}{\sigma^2} \right),$$

$$h(u) = (1+u) \log(1+u) - u$$

$h(u) \geq \dfrac{u^2/2}{1 + u/3}$ (check!) yields the desired bound.

(ii) $$e^{\theta \mathbb{E} \lambda_{max}\left(\sum_{j=1}^n X_j \right)} \le \mathbb{E}\, e^{\theta \lambda_{max}\left(\sum_1^n X_i \right)}$$

$$= \mathbb{E} \lambda_{max} \left(e^{\theta \sum_1^n X_i} \right) \le \mathbb{E} \operatorname{tr} e^{\theta \sum_1^n X_i}$$

$$\Rightarrow \mathbb{E} \lambda_{max} \left(\sum_1^n X_i \right) \le \inf_{\theta > 0} \frac{\log \mathbb{E} \operatorname{tr} e^{\theta \sum_1^n X_i}}{\theta}.$$

$$\le \inf \frac{1}{\theta} \left(\log d + \sigma^2 \left(e^{\theta} - \theta - 1 \right) \right)$$

$$e^{\theta} - \theta - 1 = \sum_{j=2}^{\infty} \frac{\theta^j}{j!} \leq \frac{\theta^2}{2} \sum_{j=2}^{\infty} \frac{\theta^{j-2}}{3^{j-2}} = \frac{\theta^2}{2} \frac{1}{1-\theta/3}$$

$$\Rightarrow \mathbb{E}\, \lambda_{max}\left(\sum_1^n X_j\right) \leq \inf_{\theta > 0} \frac{1}{\theta}\left(\log d + \sigma^2 \frac{\theta^2}{2} \frac{1}{1-\theta/3}\right)$$

$$\theta_{opt} = \frac{6L \log d + 9\sqrt{2\sigma^2 \log d}}{2L^2 t + 9\sigma^2 + 6L\sqrt{2\sigma^2 \log d}}$$

- What about rectangular matrices?

Let $X_1, \dots, X_n \in \mathbb{C}^{d_1 \times d_2}$ be independent, $\mathbb{E} X_j = 0_{d_1 \times d_2}$ $\forall j$

Then $$P\left(\left\|\sum_{j=1}^n X_j\right\| \geq t\right) \leq 2(d_1 + d_2) \exp\left(\frac{-t^2/2}{\sigma^2 + \frac{Lt}{3}}\right)$$

where $\sigma^2 = \max\left(\left\|\sum_1^n \mathbb{E} X_j X_j^*\right\|, \left\|\sum_1^n \mathbb{E} X_j^* X_j\right\|\right)$.

<u>Trick:</u> "self-adjoint dilation":

$$\mathbb{C}^{d_1 \times d_2} \ni X \mapsto \mathcal{H}(X) = \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix} \in \mathbb{C}^{(d_1 \times d_2) \times (d_1 + d_2)}$$

Note that (a) $\mathcal{H}(X)$ is self-adjoint

(b) $$(\mathcal{H}(X))^2 = \begin{pmatrix} XX^* & 0 \\ 0 & X^*X \end{pmatrix} \in \mathbb{C}^{(d_1 + d_2) \times (d_1 + d_2)}$$

$$\Rightarrow \|\mathcal{H}(X)\| = \|X\|$$

<u>Exercise</u> (a) Let $Y_1, \dots, Y_d \in \mathbb{R}^d$ be independent random vectors such that $\|Y_j\| \leq L$ with probability 1.

Show that $P\left(\|\sum_1^n (Y_j - \mathbb{E}Y_j)\|_2 \geq t\right) \leq 2d\, e^{-\frac{t^2/2}{\sigma^2 + tL/3}}$,

where $\sigma^2 = \sum_{j=1}^n \mathbb{E}\|Y_j - \mathbb{E}Y_j\|_2^2 = \sum_{j=1}^n tr\, \Sigma_j$

(b) Let $X_1, \dots, X_n \in \mathbb{R}^{d_1 \times d_2}$ be independent random matrices.

Find an upper bound for

$$P\left(\|\sum_{j=1}^n (X_j - \mathbb{E}X_j)\|_F \geq t\right) \text{ where } \|X\|_F^2 = tr(XX^T)$$

$\left[\sigma^2 = \sum_1^n \mathbb{E}\|X_j\|_F^2 \text{ in this case}\right]$ is the Frobenius norm.

- Let's discuss the dimensional factor d:

$$P\left(\|\sum_{j=1}^n (X_j - \mathbb{E}X_j)\| \geq t\right) \leq 2d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + \frac{tL}{3}}\right)$$

Is it necessary? In general, yes: consider a sequence $\{\varepsilon_{i,j}\}_{i,j \in \mathbb{N}}$ of iid random signs, and let

$$X_j = \frac{1}{\sqrt{n}}\begin{pmatrix} \varepsilon_{j,1} & & 0 \\ & \ddots & \\ 0 & & \varepsilon_{j,d} \end{pmatrix}. \quad \text{Then (i) } \|X_j\| = \frac{1}{\sqrt{n}}$$

$$\text{(ii) } \mathbb{E}X_j^2 = \frac{1}{n} I_d$$

$$\Rightarrow P\left(\|\sum_1^n X_j\| \geq t\right) \leq 2d \exp\left(\frac{-t^2/2}{1 + \frac{t}{3\sqrt{n}}}\right).$$

On the other hand, $\sum_1^n X_j = \begin{pmatrix} \frac{1}{\sqrt{n}}\sum_1^n \varepsilon_{j,1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{n}}\sum_1^n \varepsilon_{j,d} \end{pmatrix}$

$$\Rightarrow \|\sum_1^n X_j\| = \max_{i=1..d}\left(\left|\frac{1}{\sqrt{n}}\sum_1^n \varepsilon_{j,1}\right|, \dots, \left|\frac{1}{\sqrt{n}}\sum_1^n \varepsilon_{j,d}\right|\right) > c\sqrt{\log(d)}$$

for some $c > 0$.

with high probability by the central limit theorem, for large n.

- In some cases, the resulting bounds are suboptimal: consider $X_{i,j} = g_{i,j}(e_i e_j^T + e_j e_i^T)$, $g_{i,j} \sim N(0,1)$, $i,j = 1, \dots d$

  Then $X_{i,j}^2 = g_{i,j}^2 (e_i e_i^T + e_j e_j^T)$

  $$(X_{i,j})^p = g_{i,j}^p (e_i e_i^T + e_j e_j^T) \text{ for } p \text{ even}$$

  $$(X_{i,j})^p = g_{i,j}^p (e_i e_j^T + e_j e_i^T), \; p \text{ odd}$$

  $$\Rightarrow \mathbb{E}(X_{i,j})^p = \mathbb{E}(g_{i,j}^p) \cdot A_{i,j}^2 \rightarrow e_i e_i^T + e_j e_j^T, \; p \text{ even}$$

  $$= (p-1)!! = (p-1)(p-3)\dots 3 \cdot 1 \; A_{i,j}^2$$

  $$\leq \frac{p!}{2} \cdot 1 \cdot A_{i,j}^2 \quad [L = 1]$$

  $$\sigma^2 = \left\|\sum_{i,j=1}^{d} A_{i,j}^2\right\| = \left\|\begin{pmatrix} d & & 0 \\ & \ddots & \\ 0 & & d \end{pmatrix}\right\| \Rightarrow \sigma^2 = d$$

  $$\Rightarrow \mathbb{E}\left\|\sum_{i,j} X_{i,j}\right\| \lesssim \sqrt{d \log d} \text{ by matrix Bernstein}$$

  However, it is well known that $\mathbb{E}\|\sum_{i,j} X_{i,j}\| \leq 2\sqrt{d}$.

- The factor d can be improved as follows: given $M \succeq 0$, define the <u>effective rank</u> of M via $r(M) := \frac{\operatorname{tr} M}{\|\Sigma\|} = \frac{\sum_1^d \lambda_j(M)}{\lambda_{max}(M)}$. Clearly,

  $$1 \leq r(M) \leq \operatorname{rank}(M) \leq d.$$

Then the matrix Bernstein holds with d replaced by $r\left(\sum_{j=1}^{n} \mathbb{E} X_j^2\right)$, modulo minor technical details.

Example: recall the vector concentration inequality, and show that $r = 2$ in this case.

Proof idea: replace e^t by $e^t - t - 1$ in Markov's inequality

see (a) arXiv: 1112.5448 [S. Minsker '11]

(b) J. Tropp: An Introduction to Matrix Concentration inequalities (Chapter 7)

- Finally, let's discuss the "matrix variance" $\sigma^2 = \|\sum_1^n \mathbb{E} X_j^2\|$: on the one hand, it controls $\mathbb{E}\|\sum_1^n X_j\|$, on the other hand, it also controls the deviations:

(i) $$P\left(\|\sum_{j=1}^{n} X_j\| \geq t\right) \leq 2d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + \frac{tL}{3}}\right),$$

(ii) $$\mathbb{E}\|\sum_{j=1}^{n} X_j\| \leq \sqrt{2\sigma \cdot \log(2d)} + \frac{1}{3} L \log(2d).$$

The "scalar" concentration inequalities on the other hand have the form $P\left(\left|\|\sum_1 X_j\| - \mathbb{E}\|\sum_1 X_j\|\right| > t\right) \leq \psi(t, \sigma_*^2)$ where $\sigma_*^2 = \sup_{\|v\|_2 = 1} \mathbb{E}\langle Xv, v\rangle^2$ - the "weak" variance.

E.g., if Z is a random matrix $Z = \sum_j g_j A_j$ for some fixed self-adjoint A_j's, then
$$P\left(\|Z\| \geq \mathbb{E}\|Z\| + t\right) \leq e^{-t^2/2\sigma_*^2}$$

[Talagrand's inequality also involves σ_*^2]

__Lemma__ $\sigma_*^2 \le \sigma^2 = \|\mathbb{E} X^2\|$.

__Proof__ : note that $\langle Xv, v\rangle^2 = v^T X v \, v^T X v$

$$= (Xv)^T v v^T (Xv)$$

Since $vv^T \lesssim I_d$, $\langle Xv, v\rangle^2 \lesssim (Xv)^T(Xv) = v^T X^2 v$.
Take the expectation to get the result.

__Exercise__ $\sigma^2 \le d\,\sigma_*^2$, and the inequality is sharp
(consider random matrices of the form
$Z = \sum_j g_j A_j$ where g_j are iid $N(0,1)$ and
A_j's are appropriately chosen fixed matrices).

- What about random matrices that do not admit a sup norm bound or the moment growth condition?

The following matrix version of Rosenthal's moment inequality is due to R. Chen, A. Gittens & J. Tropp ('18):

Let $Y_1, \dots, Y_n \in \mathbb{C}^{d\times d}$ be a sequence of independent self-adjoint matrices, $\mathbb{E} Y_j = 0$. Then for all $q \ge 2$

$$\left(\mathbb{E}\left\|\sum_1^n Y_j\right\|^q\right)^{1/q} \le \sqrt{e \max(q, \log(2d))}\,\left\|\left[\sum_1^n \mathbb{E}\, Y_j^2\right]^{1/2}\right\|$$

$$+ 2e \max(q, \log(2d)) \left(\mathbb{E} \max_j \|Y_j\|^q\right)^{1/q}$$

__Exercise__ Show, using matrix Bernstein's inequality, that

$$\left(\mathbb{E}\left\|\sum_1^n Y_j\right\|^q\right)^{1/q} \le C\sqrt{\max(q, \log(2d))}\,\left\|\left[\sum_1^n \mathbb{E}\, Y_j^2\right]^{1/2}\right\|$$

$$+ \max(q, \log(2d)) \cdot L,$$

where $\|Y_j\| \leq L$ almost surely [this moment bound is weaker].

The proof is based on the "matrix Khintchine" inequality due to Lust-Picquard and Pisier: let $r \geq 2$, and consider a sequence of fixed self-adjoint matrices $A_1, \dots, A_n$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent random signs. Then

$$\left[\mathbb{E} \left\| \sum_{j=1}^{n} \varepsilon_j A_j \right\|_r^r \right]^{1/r} \leq \sqrt{r} \left\| \left(\sum_{j=1}^{n} A_j^2 \right)^{1/2} \right\|_r .$$

Here, $\|A\|_r^r = \left(\sum_1^d (\lambda_j(A))^r \right)$ is the Schatten norm of A.

<u>Exercise</u> Show that

$$\left(\mathbb{E} \left\| \sum_{j=1}^{n} \varepsilon_j A_j \right\|^2 \right)^{1/2} \leq \sqrt{e(1+2\log(d))} \left\| \sum_1^n A_j^2 \right\|^{1/2}$$ by

using the inequality between the Schatten p norm and the operator norm.