User-Friendly Tools for Random Matrices

Joel A. Tropp

Computing + Mathematical Sciences
California Institute of Technology
jtropp@cms.caltech.edu

Research supported by ONR, AFOSR, NSF, DARPA, Sloan, and Moore.

Download the Notes:

tinyurl.com/bocrqhe

[URL] `http://users.cms.caltech.edu/~jtropp/notes/Tro12-User-Friendly-Tools-NIPS.pdf`

Random Matrices
in the Mist

Random Matrices in Statistics

❧ Covariance estimation for the multivariate normal distribution

John Wishart

3. *Multi-variate Distribution. Use of Quadratic co-ordinates.*

A comparison of equation (8) with the corresponding results (1) and (2) for uni-variate and bi-variate sampling, respectively, indicates the form the general result may be expected to take. In fact, we have for the simultaneous distribution in random samples of the n variances (squared standard deviations) and the $\frac{n(n-1)}{2}$ product moment coefficients the following expression:

$$dp = \frac{\begin{vmatrix} A_{11} & A_{12} \dots & A_{1n} \\ A_{21} & A_{22} \dots & A_{2n} \\ \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} \dots & A_{nn} \end{vmatrix}^{\frac{N-1}{2}}}{(\sqrt{\pi})^{\frac{1}{2}n(n-1)} \Gamma\left(\frac{N-1}{2}\right) \Gamma\left(\frac{N-2}{2}\right) \dots \Gamma\left(\frac{N-n}{2}\right)} \times e^{-A_{11}a_{11} - A_{22}a_{22} - \dots - A_{nn}a_{nn} - 2A_{12}a_{12} - 2A_{13}a_{13} - \dots - 2A_{n-1\,n}a_{n-1\,n}} \times \begin{vmatrix} a_{11} & a_{12} \dots & a_{1n} \\ a_{21} & a_{22} \dots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} \dots & a_{nn} \end{vmatrix}^{\frac{N-n-2}{2}} da_{11} da_{12} \dots da_{nn} \quad \dots\dots\dots\dots\dots (9),$$

where $a_{pq} = s_p s_q r_{pq}$, and $A_{pq} = \frac{N}{2\sigma_p \sigma_q} \cdot \frac{\Delta_{pq}}{\Delta}$, Δ being the determinant

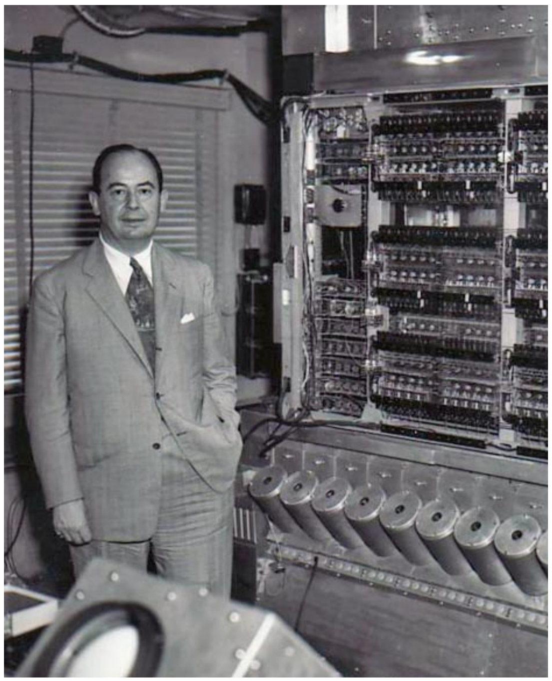
$$|\rho_{pq}|, p, q = 1, 2, 3, \dots n,$$

and Δ_{pq} the minor of ρ_{pq} in Δ.

[Refs] Wishart, *Biometrika* 1928. Photo from `apprendre-math.info`.

Random Matrices in Numerical Linear Algebra

- Model for floating-point errors in LU decomposition



John von Neumann

now combining (8.6) and (8.7) we obtain our desired result:

$$\text{Prob}\,(\lambda > 2\sigma^2 rn) < \frac{(rn)^{n-1/2} e^{-rn} \pi^{1/2} e^{n} \cdot 2^{n-2}}{\pi n^{n-1}(r-1)n} \tag{8.8}$$

$$= \left(\frac{2r}{e^{r-1}}\right)^{n} \times \frac{1}{4(r-1)(r\pi n)^{1/2}}.$$

We sum up in the following theorem:

(8.9) The probability that the upper bound $|A|$ of the matrix A of (8.1) exceeds $2.72\sigma n^{1/2}$ is less than $.027 \times 2^{-n} n^{-1/2}$, that is, with probability greater than 99% the upper bound of A is less than $2.72\sigma n^{1/2}$ for $n = 2, 3, \cdots$.

This follows at once by taking $r = 3.70$.

[Refs] von Neumann and Goldstine, *Bull. AMS* 1947 and *Proc. AMS* 1951.

Random Matrices in Nuclear Physics

❧ Model for the Hamiltonian of a heavy atom in a slow nuclear reaction

Eugene Wigner

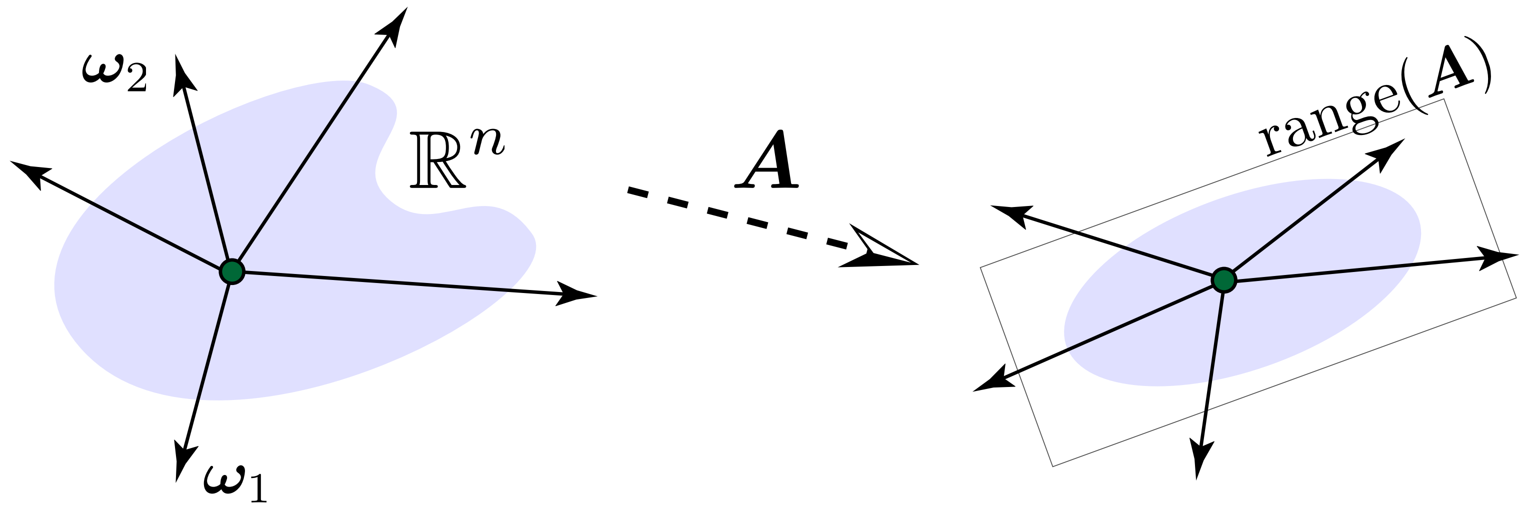
Random sign symmetric matrix

The matrices to be considered are $2N + 1$ dimensional real symmetric matrices; N is a very large number. The diagonal elements of these matrices are zero, the non diagonal elements $v_{ik} = v_{ki} = \pm v$ have all the same absolute value but random signs. There are $\mathfrak{N} = 2^{N(2N+1)}$ such matrices. We shall calculate, after an introductory remark, the averages of $(H^\nu)_{00}$ and hence the strength function $S'(x) = \sigma(x)$. This has, in the present case, a second interpretation: it also gives the density of the characteristic values of these matrices. This will be shown first.

[Refs] Wigner, *Ann. Math* 1955. Photo from Nobel Foundation.

Modern
Applications

Randomized Linear Algebra



Input: An $m \times n$ matrix $\boldsymbol{A}$, a target rank k, an oversampling parameter p

Output: An $m \times (k + p)$ matrix $\boldsymbol{Q}$ with orthonormal columns

1. Draw an $n \times (k + p)$ random matrix $\boldsymbol{\Omega}$
2. Form the matrix product $\boldsymbol{Y} = \boldsymbol{A\Omega}$
3. Construct an orthonormal basis $\boldsymbol{Q}$ for the range of $\boldsymbol{Y}$

[Ref] Halko–Martinsson–T, *SIAM Rev.* 2011.

Other Algorithmic Applications

- **Sparsification.** Accelerate spectral calculation by randomly zeroing entries in a matrix.

- **Subsampling.** Accelerate construction of kernels by randomly subsampling data.

- **Dimension Reduction.** Accelerate nearest neighbor calculations by random projection to a lower dimension.

- **Relaxation & Rounding.** Approximate solution of maximization problems with matrix variables.

[Refs] Achlioptas–McSherry 2001 and 2007, Spielman–Teng 2004; Williams–Seeger 2001, Drineas–Mahoney 2006, Gittens 2011; Indyk–Motwani 1998, Ailon–Chazelle 2006; Nemirovski 2007, So 2009...

Random Matrices as Models

- **High-Dimensional Data Analysis.** Random matrices are used to model multivariate data.

- **Wireless Communications.** Random matrices serve as models for wireless channels.

- **Demixing Signals.** Random model for incoherence when separating two structured signals.

[Refs] Bühlmann and van de Geer 2011, Koltchinskii 2011; Tulino–Verdú 2004; McCoy–T 2011.

Theoretical Applications

- **Algorithms.** Smoothed analysis of Gaussian elimination.
- **Combinatorics.** Random constructions of expander graphs.
- **High-Dimensional Geometry.** Structure of random slices of convex bodies.
- **Quantum Information Theory.** (Counter)examples to conjectures about quantum channel capacity.

[Refs] Sankar–Spielman–Teng 2006; Pinsker 1973; Gordon 1985; Hayden–Winter 2008, Hastings 2009.

Random Matrices:
My Way

The Conventional Wisdom

"Random Matrices are Tough!"

[Refs] `youtube.com/watch?v=NO0cvqT1tAE`, most monographs on RMT.

Principle A

“But...

In many applications, a random matrix can be decomposed as a sum of independent random matrices:

$$\boldsymbol{Z} = \sum_{k=1}^{n} \boldsymbol{S}_k$$

Principle B

and

There are exponential concentration inequalities for the spectral norm of a sum of independent random matrices:

$$\mathbb{P}\left\{ \|\boldsymbol{Z}\| \geq t \right\} \leq \exp(\quad \cdots \quad)$$

!!!"

Matrix
Gaussian Series

The Norm of a Matrix Gaussian Series

Theorem 1. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$, *and*
- $\gamma_1, \gamma_2, \gamma_3, \dots$ *are independent standard normal RVs.*

Define $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{ \left\| \sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \left\| \sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\}.$$

Then

$$\mathbb{P}\left\{ \left\| \sum\nolimits_k \gamma_k \boldsymbol{B}_k \right\| \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

The Norm of a Matrix Gaussian Series

Theorem 2. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$*, and*
- $\gamma_1, \gamma_2, \gamma_3, \dots$ *are independent standard normal RVs.*

Define $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \ \left\| \sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\}.$$

Then

$$\mathbb{E} \left\| \sum\nolimits_k \gamma_k \boldsymbol{B}_k \right\| \leq \sqrt{2\sigma^2 \log d}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

The Variance Parameter

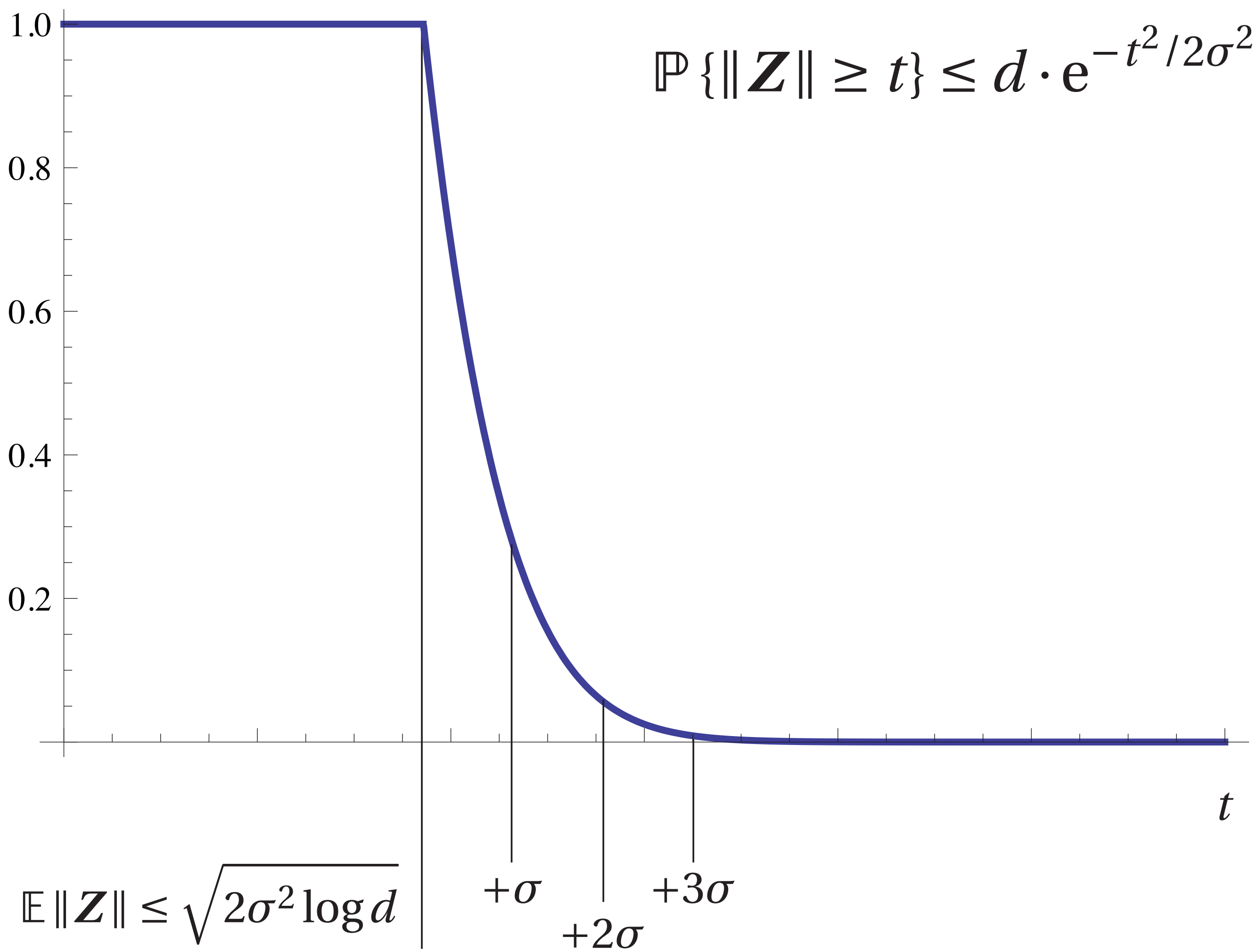
- Define the matrix Gaussian series $\boldsymbol{Z} = \sum_{k=1}^{n} \gamma_k \boldsymbol{B}_k$
- The variance parameter $\sigma^2(\boldsymbol{Z})$ derives from the "mean square of $\boldsymbol{Z}$"
- But a general matrix has *two* different squares!

$$\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{E}(\gamma_j \gamma_k) \boldsymbol{B}_j \boldsymbol{B}_k^* = \sum_{k=1}^{n} \boldsymbol{B}_k \boldsymbol{B}_k^*$$

$$\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbb{E}(\gamma_j \gamma_k) \boldsymbol{B}_j^* \boldsymbol{B}_k = \sum_{k=1}^{n} \boldsymbol{B}_k^* \boldsymbol{B}_k$$

- Variance parameter $\sigma^2(\boldsymbol{Z}) = \max\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|,\ \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\}$.

Schematic of Gaussian Series Tail Bound



Warmup: A Wigner Matrix

- Let $\{\gamma_{jk} : 1 \leq j < k \leq n\}$ be independent standard normal variables
- A Gaussian Wigner matrix:

$$\boldsymbol{W} = \begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} & \dots & \gamma_{1n} \\ \gamma_{12} & 0 & \gamma_{23} & \dots & \gamma_{2n} \\ \gamma_{13} & \gamma_{23} & 0 & & \gamma_{3n} \\ \vdots & \vdots & & \ddots & \vdots \\ \gamma_{1n} & \gamma_{2n} & \dots & \gamma_{n-1,n} & 0 \end{bmatrix}$$

- **Problem:** What is $\mathbb{E}\,\|\boldsymbol{W}\|$?

Notes: §4.4.1, page 35.

The Wigner Matrix, *qua* Gaussian Series

- Express the Wigner matrix as a Gaussian series:

$$\boldsymbol{W} = \sum_{1 \le j < k \le n} \gamma_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj})$$

- The symbol $\mathbf{E}_{jk}$ denotes the $n \times n$ matrix unit

$$\mathbf{E}_{jk} = \begin{bmatrix} & & \\ & & 1 \\ & & \end{bmatrix} \begin{matrix} \\ \leftarrow \; j \\ \\ \end{matrix}$$

$$\uparrow$$

$$k$$

Norm Bound for the Wigner Matrix

- Need to compute the variance parameter $\sigma^2(\boldsymbol{W})$
- Summands are symmetric, so both matrix squares are the same:

$$\sum_{1\le j<k\le n} (\mathbf{E}_{jk} + \mathbf{E}_{kj})^2 = \sum_{1\le j<k\le n} (\mathbf{E}_{jk}\mathbf{E}_{jk} + \mathbf{E}_{jk}\mathbf{E}_{kj} + \mathbf{E}_{kj}\mathbf{E}_{jk} + \mathbf{E}_{kj}\mathbf{E}_{kj})$$

$$= \sum_{1\le j<k\le n} (\mathbf{0} + \mathbf{E}_{jj} + \mathbf{E}_{kk} + \mathbf{0}) = (n-1)\,\mathbf{I}_n$$

- Thus, the variance $\sigma^2(\boldsymbol{W}) = \|(n-1)\,\mathbf{I}_n\| = n - 1.$

- **Conclusion:** $\mathbb{E}\,\|\boldsymbol{W}\| \le \sqrt{2(n-1)\log(2n)}$
- **Optimal:** $\mathbb{E}\,\|\boldsymbol{W}\| \sim 2\sqrt{n}$

[Refs] Wigner 1955, Davidson–Szarek 2002, Tao 2012.

Example: A Gaussian Toeplitz Matrix

- Let $\{\gamma_k\}$ be independent standard normal variables
- An unsymmetric Gaussian Toeplitz matrix:

$$\boldsymbol{T} = \begin{bmatrix} \gamma_0 & \gamma_1 & & \dots & & \gamma_{n-1} \\ \gamma_{-1} & \gamma_0 & \gamma_1 & & & \\ & \gamma_{-1} & \gamma_0 & \gamma_1 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \gamma_{-1} & \gamma_0 & \gamma_1 \\ \gamma_{-(n-1)} & & \dots & & \gamma_{-1} & \gamma_0 \end{bmatrix}$$

- **Problem:** What is $\mathbb{E} \|\boldsymbol{T}\|$?

Notes: §4.6, page 38.

The Toeplitz Matrix, *qua* Gaussian Series

❧ Express the unsymmetric Toeplitz matrix as a Gaussian series:

$$\boldsymbol{T} = \gamma_0 \, \mathbf{I} + \sum_{k=1}^{n-1} \gamma_k \boldsymbol{S}^k + \sum_{k=1}^{n-1} \gamma_{-k} (\boldsymbol{S}^k)^*$$

❧ The matrix $\boldsymbol{S}$ is the shift-up operator on n-dimensional column vectors:

$$\boldsymbol{S} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}.$$

Variance Calculation for the Toeplitz Matrix

- Note that

$$(\boldsymbol{S}^k)(\boldsymbol{S}^k)^* = \sum_{j=1}^{n-k} \mathbf{E}_{jj} \quad \text{and} \quad (\boldsymbol{S}^k)^*(\boldsymbol{S}^k) = \sum_{j=k+1}^{n} \mathbf{E}_{jj}.$$

- *Both* sums of squares take the form

$$\begin{aligned}
&\mathbf{I}^2 + \sum_{k=1}^{n-1} (\boldsymbol{S}^k)(\boldsymbol{S}^k)^* + \sum_{k=1}^{n-1} (\boldsymbol{S}^k)^*(\boldsymbol{S}^k) \\
&\quad = \mathbf{I} + \sum_{k=1}^{n-1} \left[\sum_{j=1}^{n-k} \mathbf{E}_{jj} + \sum_{j=k+1}^{n} \mathbf{E}_{jj} \right] = \sum_{j=1}^{n} \left[1 + \sum_{k=1}^{n-j} 1 + \sum_{k=1}^{j-1} 1 \right] \mathbf{E}_{jj} \\
&\quad = \sum_{j=1}^{n} (1 + (n-j) + (j-1))\, \mathbf{E}_{jj} = n\, \mathbf{I}_n.
\end{aligned}$$

Norm Bound for the Toeplitz Matrix

- The variance parameter $\sigma^2(\boldsymbol{T}) = \|n\,\mathbf{I}_n\| = n$
- **Conclusion:** $\mathbb{E}\,\|\boldsymbol{T}\| \leq \sqrt{2n\log(2n)}$
- **Optimal:** $\mathbb{E}\,\|\boldsymbol{T}\| \sim \mathrm{const}\cdot\sqrt{2n\log n}$
- The optimal constant is at least 0.8288...

[Refs] Bryc–Dembo–Jiang 2006, Meckes 2007, Sen–Virág 2011, T 2011.

Matrix
Rademacher Series

The Norm of a Matrix Rademacher Series

Theorem 3. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$*, and*
- $\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots$ *are independent Rademacher RVs.*

Then

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{B}_k\right\| \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{\left\|\sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^*\right\|, \; \left\|\sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k\right\|\right\}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

The Norm of a Matrix Rademacher Series

Theorem 4. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$*, and*
- $\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots$ *are independent Rademacher RVs.*

Then

$$\mathbb{E} \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{B}_k \right\| \leq \sqrt{2\sigma^2 \log d}$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \ \left\| \sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

Example: Modulation by Random Signs

Fixed matrix, in captivity:

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots \\ c_{21} & c_{22} & c_{23} & \dots \\ c_{31} & c_{32} & c_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{d_1 \times d_2}$$

Random matrix, formed by randomly flipping the signs of the entries:

$$\boldsymbol{Z} = \begin{bmatrix} \varepsilon_{11}\, c_{11} & \varepsilon_{12}\, c_{12} & \varepsilon_{13}\, c_{13} & \dots \\ \varepsilon_{21}\, c_{21} & \varepsilon_{22}\, c_{22} & \varepsilon_{23}\, c_{23} & \dots \\ \varepsilon_{31}\, c_{31} & \varepsilon_{32}\, c_{32} & \varepsilon_{33}\, c_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{d_1 \times d_2}$$

Problem: What is $\mathbb{E}\, \|\boldsymbol{Z}\|$?

Notes: §4.5, page 37.

The Random Matrix, *qua* Rademacher Series

❧ Express the random matrix as a Gaussian series:

$$\boldsymbol{Z} = \begin{bmatrix} \varepsilon_{11}\, c_{11} & \varepsilon_{12}\, c_{12} & \varepsilon_{13}\, c_{13} & \dots \\ \varepsilon_{21}\, c_{21} & \varepsilon_{22}\, c_{22} & \varepsilon_{23}\, c_{23} & \dots \\ \varepsilon_{31}\, c_{31} & \varepsilon_{32}\, c_{32} & \varepsilon_{33}\, c_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{d_1 \times d_2} = \sum\nolimits_{jk} \varepsilon_{jk}\, c_{jk}\, \mathbf{E}_{jk}$$

Variance of the Randomly Signed Matrix

The first term in the matrix variance σ^2 satisfies

$$\begin{aligned}\left\|\sum\nolimits_{jk}(c_{jk}\,\mathbf{E}_{jk})(c_{jk}\,\mathbf{E}_{jk})^*\right\| &= \left\|\sum\nolimits_{jk}|c_{jk}|^2\,\mathbf{E}_{jk}\mathbf{E}_{kj}\right\| \\ &= \left\|\sum\nolimits_{j}\left(\sum\nolimits_{k}|c_{jk}|^2\right)\mathbf{E}_{jj}\right\| \\ &= \left\|\begin{bmatrix}\sum_k |c_{1k}|^2 & & \\ & \sum_k |c_{2k}|^2 & \\ & & \ddots\end{bmatrix}\right\| \\ &= \max\nolimits_j \sum\nolimits_k |c_{jk}|^2\end{aligned}$$

The same argument applies to the second term. Thus,

$$\sigma^2 = \max\left\{\max\nolimits_j \sum\nolimits_k |c_{jk}|^2,\ \max\nolimits_k \sum\nolimits_j |c_{jk}|^2\right\}$$

Comparison with the Literature

Consider the randomly signed matrix $\boldsymbol{Z} = [\varepsilon_{jk}\, c_{jk}]$. Define

$$\sigma^2(\boldsymbol{Z}) = \max \left\{ \max_j \sum\nolimits_k |c_{jk}|^2,\ \max_k \sum\nolimits_j |c_{jk}|^2 \right\}$$

[T 2010], obtained via matrix Rademacher bound:

$$\mathbb{E}\, \|\boldsymbol{Z}\| \le \sqrt{2 \log d} \cdot \sigma$$

[Seginer 2000], obtained with path-counting arguments:

$$\mathbb{E}\, \|\boldsymbol{Z}\| \le \text{const} \cdot \sqrt[4]{\log d} \cdot \sigma$$

[Latała 2005], obtained with chaining arguments:

$$\mathbb{E}\, \|\boldsymbol{Z}\| \le \text{const} \cdot \left[\sigma + \sqrt[4]{\sum\nolimits_{jk} |c_{jk}|^4} \right]$$

Matrix
Chernoff Inequality

The Matrix Chernoff Bound

Theorem 5. [T 2010] **Suppose**

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \dots$ *are random* **psd** *matrices with dimension* d*, and*
- $\lambda_{\max}(\boldsymbol{X}_k) \leq R$ *for each* k*.*

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \leq (1-t)\cdot\mu_{\min}\right\} \leq d \cdot \left[\frac{\mathrm{e}^{-t}}{(1-t)^{1-t}}\right]^{\mu_{\min}/R}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq (1+t)\cdot\mu_{\max}\right\} \leq d \cdot \left[\frac{\mathrm{e}^{t}}{(1+t)^{1+t}}\right]^{\mu_{\max}/R}$$

where $\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$ *and* $\mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$*.*

[Refs] Ahlswede–Winter 2002, T 2011. **Notes: Thm. 5.1.1, page 48.**

The Matrix Chernoff Bound

Theorem 6. [T 2010] **Suppose**

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \dots$ *are random* **psd** *matrices with dimension* d*, and*
- $\lambda_{\max}(\boldsymbol{X}_k) \le R$ *for each* k*.*

Then

$$\mathbb{E}\,\lambda_{\min}\Big(\sum\nolimits_k \boldsymbol{X}_k\Big) \ge 0.6\,\mu_{\min} - R\log d$$

$$\mathbb{E}\,\lambda_{\max}\Big(\sum\nolimits_k \boldsymbol{X}_k\Big) \le 1.8\,\mu_{\max} + R\log d$$

where $\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$ *and* $\mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$*.*

[Refs] Ahlswede–Winter 2002, T 2011. **Notes: Thm. 5.1.1, page 48.**

Example: Random Submatrices

Fixed matrix, in captivity:

$$\boldsymbol{C} = \begin{bmatrix} | & | & | & | & & | \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \boldsymbol{c}_3 & \boldsymbol{c}_4 & \dots & \boldsymbol{c}_n \\ | & | & | & | & & | \end{bmatrix}_{d \times n}$$

Random matrix, formed by picking random columns:

$$\boldsymbol{Z} = \begin{bmatrix} & | & | & & & | \\ & \boldsymbol{c}_2 & \boldsymbol{c}_3 & & \dots & \boldsymbol{c}_n \\ & | & | & & & | \end{bmatrix}_{d \times n}$$

$$\quad\uparrow \quad \uparrow \quad\quad\quad \uparrow$$

Problem: What is the expectation of $\sigma_1(\boldsymbol{Z})$? What about $\sigma_d(\boldsymbol{Z})$?

Notes: §5.2.1, page 49.

Model for Random Submatrix

- Let $\boldsymbol{C}$ be a fixed $d \times n$ matrix with columns $\boldsymbol{c}_1, \dots, \boldsymbol{c}_n$
- Let $\delta_1, \dots, \delta_n$ be independent 0–1 random variables with mean s/n
- Define $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, \dots, \delta_n)$
- Form a random submatrix $\boldsymbol{Z}$ by turning off columns from $\boldsymbol{C}$

$$\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{\Delta} = \begin{bmatrix} | & | & & | \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_n \\ | & | & & | \end{bmatrix}_{d \times n} \begin{bmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_n \end{bmatrix}_{n \times n}$$

- Note that $\boldsymbol{Z}$ typically contains about s nonzero columns

The Random Submatrix, *qua* PSD Sum

❧ The largest and smallest singular values of $\boldsymbol{Z}$ satisfy

$$\sigma_1(\boldsymbol{Z})^2 = \lambda_{\max}(\boldsymbol{Z}\boldsymbol{Z}^*)$$
$$\sigma_d(\boldsymbol{Z})^2 = \lambda_{\min}(\boldsymbol{Z}\boldsymbol{Z}^*)$$

❧ Define the psd matrix $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^*$, and observe that

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^* = \boldsymbol{C}\boldsymbol{\Delta}^2\boldsymbol{C}^* = \boldsymbol{C}\boldsymbol{\Delta}\boldsymbol{C}^* = \sum\nolimits_{k=1}^{n} \delta_k \, \boldsymbol{c}_k\boldsymbol{c}_k^*$$

❧ We have expressed $\boldsymbol{Y}$ as a sum of independent psd random matrices

Preparing to Apply the Chernoff Bound

- Consider the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \delta_k \, \boldsymbol{c}_k \boldsymbol{c}_k^*$$

- The maximal eigenvalue of each summand is bounded as

$$R = \max_k \lambda_{\max}(\delta_k \, \boldsymbol{c}_k \boldsymbol{c}_k^*) \le \max_k \|\boldsymbol{c}_k\|^2$$

- The expectation of the random matrix $\boldsymbol{Y}$ is

$$\mathbb{E}(\boldsymbol{Y}) = \frac{s}{n} \sum\nolimits_{k=1}^{n} \boldsymbol{c}_k \boldsymbol{c}_k^* = \frac{s}{n} \boldsymbol{C}\boldsymbol{C}^*$$

- The mean parameters satisfy

$$\mu_{\max} = \lambda_{\max}(\mathbb{E}\,\boldsymbol{Y}) = \frac{s}{n} \sigma_1(\boldsymbol{C})^2 \quad \text{and} \quad \mu_{\min} = \lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}) = \frac{s}{n} \sigma_d(\boldsymbol{C})^2$$

What the Chernoff Bound Says

Applying the Chernoff bound, we reach

$$\mathbb{E}\left[\sigma_1(\boldsymbol{Z})^2\right] = \mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \leq 1.8 \cdot \frac{s}{n}\sigma_1(\boldsymbol{C})^2 + \max_k \|\boldsymbol{c}_k\|_2^2 \cdot \log d$$

$$\mathbb{E}\left[\sigma_d(\boldsymbol{Z})^2\right] = \mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \geq 0.6 \cdot \frac{s}{n}\sigma_d(\boldsymbol{C})^2 - \max_k \|\boldsymbol{c}_k\|_2^2 \cdot \log d$$

- Matrix $\boldsymbol{C}$ has n columns; the random submatrix $\boldsymbol{Z}$ includes about s
- The singular value $\sigma_i(\boldsymbol{Z})^2$ inherits an s/n share of $\sigma_i(\boldsymbol{C})^2$ for $i = 1, d$
- Additive correction reflects number d of rows of $\boldsymbol{C}$, max column norm
- **[Gittens–T 2011]** Remaining singular values have similar behavior

Key Example: Unit-Norm Tight Frame

- A $d \times n$ unit-norm tight frame $\boldsymbol{C}$ satisfies

$$\boldsymbol{C}\boldsymbol{C}^* = \frac{n}{d}\mathbf{I}_d \quad \text{and} \quad \|\boldsymbol{c}_k\|_2^2 = 1 \quad \text{for } k = 1, 2, \dots, n$$

- Specializing the inequalities from the previous slide...

$$\mathbb{E}\left[\sigma_1(\boldsymbol{Z})^2\right] \leq 1.8 \cdot \frac{s}{d} + \log d$$
$$\mathbb{E}\left[\sigma_d(\boldsymbol{Z})^2\right] \geq 0.6 \cdot \frac{s}{d} - \log d$$

- Choose $s \geq 1.67\, d \log d$ columns for a nontrivial lower bound
- Sharp condition $s > d \log d$ also follows from matrix Chernoff bound

[Refs] Rudelson 1999, Rudelson–Vershynin 2007, T 2008, Gittens–T 2011, T 2011, Chrétien–Darses 2012.

Matrix
Bernstein Inequality

The Matrix Bernstein Inequality

Theorem 7. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are indep. random matrices with dimension* $d_1 \times d_2$,
- $\mathbb{E}\, \boldsymbol{S}_k = \boldsymbol{0}$ *for each* k, *and*
- $\|\boldsymbol{S}_k\| \leq R$ *for each* k.

Then

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \boldsymbol{S}_k\right\| \geq t\right\} \leq d \cdot \exp\left\{\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{\left\|\sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k\boldsymbol{S}_k^*)\right\|,\ \left\|\sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k^*\boldsymbol{S}_k)\right\|\right\}$$

[Refs] Gross 2010, Recht 2011, Oliveira 2010, T 2011. **Notes: Cor. 6.2.1, page 64.**

The Matrix Bernstein Inequality

Theorem 8. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are indep. random matrices with dimension* $d_1 \times d_2$*,*
- $\mathbb{E}\, \boldsymbol{S}_k = \mathbf{0}$ *for each* k*, and*
- $\|\boldsymbol{S}_k\| \leq R$ *for each* k*.*

Then

$$\mathbb{E} \left\| \sum\nolimits_k \boldsymbol{S}_k \right\| \leq \sqrt{2\sigma^2 \log d} + \tfrac{1}{3} R \log d$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k^* \boldsymbol{S}_k) \right\| \right\}$$

[Refs] Gross 2010, Recht 2011, Oliveira 2010, T 2011. **Notes: Cor. 6.2.1, page 64.**

Example: Randomized Matrix Multiplication

Product of two matrices, in captivity:

$$\boldsymbol{B}\boldsymbol{C}^* = \begin{bmatrix} | & | & | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \boldsymbol{b}_3 & \boldsymbol{b}_4 & \dots & \boldsymbol{b}_n \\ | & | & | & | & & | \end{bmatrix}_{d_1 \times n} \begin{bmatrix} — & \boldsymbol{c}_1^* & — \\ — & \boldsymbol{c}_2^* & — \\ — & \boldsymbol{c}_3^* & — \\ — & \boldsymbol{c}_4^* & — \\ & \vdots & \\ — & \boldsymbol{c}_n^* & — \end{bmatrix}_{n \times d_2}$$

[Idea] Approximate multiplication by random sampling

[Refs] Drineas–Mahoney–Kannan 2004, Magen–Zouzias 2010, Magdon-Ismail 2010, Hsu–Kakade–Zhang 2011 and 2012.

A Sampling Model for Tutorial Purposes

- **Assume**

$$\|\boldsymbol{b}_j\|_2 = 1 \quad \text{and} \quad \|\boldsymbol{c}_j\|_2 = 1 \quad \text{for } j = 1, 2, \dots, n$$

- Construct a random variable $\boldsymbol{S}$ whose value is a $d_1 \times d_2$ matrix:
  - Draw $J \sim \textsc{uniform}\{1, 2, \dots, n\}$
  - Set $\boldsymbol{S} = n \cdot \boldsymbol{b}_J \boldsymbol{c}_J^*$
- The random matrix $\boldsymbol{S}$ is an unbiased estimator of the product $\boldsymbol{BC}^*$

$$\mathbb{E}\, \boldsymbol{S} = \sum\nolimits_{j=1}^{n} (n \cdot \boldsymbol{b}_j \boldsymbol{c}_j^*) \cdot \mathbb{P}\{J = j\} = \sum\nolimits_{j=1}^{n} \boldsymbol{b}_j \boldsymbol{c}_j^* = \boldsymbol{BC}^*$$

- Approximate $\boldsymbol{BC}^*$ by averaging m independent copies of $\boldsymbol{S}$

$$\boldsymbol{Z} = \frac{1}{m} \sum\nolimits_{k=1}^{m} \boldsymbol{S}_k \approx \boldsymbol{BC}^*$$

Notes: §6.4, page 67.

Preparing to Apply the Bernstein Bound I

- Let $\boldsymbol{S}_k$ be independent copies of $\boldsymbol{S}$, and consider the average

$$\boldsymbol{Z} = \frac{1}{m} \sum\nolimits_{k=1}^{m} \boldsymbol{S}_k$$

- We study the typical approximation error

$$\mathbb{E} \, \|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\| = \frac{1}{m} \cdot \mathbb{E} \left\| \sum\nolimits_{k=1}^{m} (\boldsymbol{S}_k - \boldsymbol{B}\boldsymbol{C}^*) \right\|$$

- The summands are independent and $\mathbb{E}\,\boldsymbol{S}_k = \boldsymbol{B}\boldsymbol{C}^*$, so we *symmetrize*:

$$\mathbb{E} \, \|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\| \leq \frac{2}{m} \cdot \mathbb{E} \left\| \sum\nolimits_{k=1}^{m} \varepsilon_k \boldsymbol{S}_k \right\|$$

where $\{\varepsilon_k\}$ are independent Rademacher RVs, independent from $\{\boldsymbol{S}_k\}$

Preparing to Apply the Bernstein Bound II

- The norm of each summand satisfies the uniform bound

$$R = \|\varepsilon \boldsymbol{S}\| = \|\boldsymbol{S}\| = \|n \cdot (\boldsymbol{b}_J \boldsymbol{c}_J^*)\| = n \, \|\boldsymbol{b}_J\|_2 \, \|\boldsymbol{c}_J\|_2 = n$$

- Compute the variance in two stages:

$$\begin{aligned}\mathbb{E}(\boldsymbol{S}\boldsymbol{S}^*) &= \sum\nolimits_{j=1}^{n} n^2 (\boldsymbol{b}_j \boldsymbol{c}_j^*)(\boldsymbol{b}_j \boldsymbol{c}_j^*)^* \, \mathbb{P}\{J = j\} = n \sum\nolimits_{j=1}^{n} \|\boldsymbol{c}_j\|_2^2 \, \boldsymbol{b}_j \boldsymbol{b}_j^* \\ &= n \, \boldsymbol{B}\boldsymbol{B}^* \\ \mathbb{E}(\boldsymbol{S}^*\boldsymbol{S}) &= n \, \boldsymbol{C}\boldsymbol{C}^*\end{aligned}$$

$$\begin{aligned}\sigma^2 &= \max\left\{ \left\| \sum\nolimits_{k=1}^{m} \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \ \left\| \sum\nolimits_{k=1}^{m} \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\| \right\} \\ &= \max\left\{ \|mn \cdot \boldsymbol{B}\boldsymbol{B}^*\|, \ \|mn \cdot \boldsymbol{C}\boldsymbol{C}^*\| \right\} \\ &= mn \cdot \max\{ \|\boldsymbol{B}\|^2, \ \|\boldsymbol{C}\|^2 \}\end{aligned}$$

What the Bernstein Bound Says

Applying the Bernstein bound, we reach

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}-\boldsymbol{B}\boldsymbol{C}^*\| &\le \frac{2}{m}\,\mathbb{E}\left\|\sum\nolimits_{k=1}^{m}\varepsilon_k\boldsymbol{S}_k\right\| \\ &\le \frac{2}{m}\left[\sigma\sqrt{2\log(d_1+d_2)}+\tfrac{1}{3}R\log(d_1+d_2)\right] \\ &= 2\sqrt{\frac{n\log(d_1+d_2)}{m}}\cdot\max\{\|\boldsymbol{B}\|,\|\boldsymbol{C}\|\}+\frac{2}{3}\cdot\frac{n\log(d_1+d_2)}{m}\end{aligned}$$

[Q] What can this possibly mean? Is this bound any good at all?

Detour: The Stable Rank

- The *stable rank* of a matrix is defined as

$$\operatorname{srank}(\boldsymbol{A}) := \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\|\boldsymbol{A}\|^2}$$

- In general, $1 \leq \operatorname{srank}(\boldsymbol{A}) \leq \operatorname{rank}(\boldsymbol{A})$
- When $\boldsymbol{A}$ has either n rows or n columns, $1 \leq \operatorname{srank}(\boldsymbol{A}) \leq n$
- **Assume** that $\boldsymbol{A}$ has n unit-norm columns, so that $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = n$
- When all columns of $\boldsymbol{A}$ are the same, $\|\boldsymbol{A}\|^2 = n$ and $\operatorname{srank}(\boldsymbol{A}) = 1$
- When all columns of $\boldsymbol{A}$ are orthogonal, $\|\boldsymbol{A}\|^2 = 1$ and $\operatorname{srank}(\boldsymbol{A}) = n$

Randomized Matrix Multiply, Relative Error

- Define the (geometric) *mean stable rank* of the factors to be

$$s := \sqrt{\operatorname{srank}(\boldsymbol{B}) \cdot \operatorname{srank}(\boldsymbol{C})}.$$

- Converting the error bound to a relative scale, we obtain

$$\frac{\mathbb{E}\,\|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\|}{\|\boldsymbol{B}\|\,\|\boldsymbol{C}\|} \le 2\sqrt{\frac{s \log(d_1 + d_2)}{m}} + \frac{2}{3} \cdot \frac{s \log(d_1 + d_2)}{m}$$

- For relative error $\varepsilon \in (0, 1)$, the number m of samples should be

$$m \ge \text{Const} \cdot \varepsilon^{-2} \cdot s \log(d_1 + d_2)$$

- **The number of samples is proportional to the mean stable rank!**
- We also pay weakly for the dimension $d_1 \times d_2$ of the product $\boldsymbol{B}\boldsymbol{C}^*$

More Things in Heaven & Earth

- **[More Bounds for Eigenvalues]** There are exponential tail bounds for maximum eigenvalues, minimum eigenvalues, and eigenvalues in between...
- **[More Exponential Bounds]** There is a matrix Hoeffding inequality and a matrix Bennett inequality, plus matrix Chernoff and Bernstein for unbounded matrices...
- **[Matrix Martingales]** There is a matrix Azuma inequality, a matrix bounded difference inequality, and a matrix Freedman inequality...
- **[Dependent Sums]** Exponential tail bounds hold for some random matrices based on dependent random variables...
- **[Polynomial Bounds]** There are matrix versions of the Rosenthal inequality, the Pinelis inequality, and the Burkholder–Davis–Gundy inequality...
- **[Intrinsic Dimension]** The dimensional dependence can sometimes be weakened...
- **[The Proofs!]** And the technical arguments are amazingly pretty...

[Refs] T 2011, Gittens–T 2011, Oliveira 2010, Mackey et al. 2012, ...

To learn more...

E-mail: `jtropp@cms.caltech.edu`
Web: `http://users.cms.caltech.edu/~jtropp`

Some papers:

- "User-friendly tail bounds for sums of random matrices," *FOCM*, 2011.
- "User-friendly tail bounds for matrix martingales." Caltech ACM Report 2011-01.
- "Freedman's inequality for matrix martingales," *ECP*, 2011.
- "A comparison principle for functions of a uniformly random subspace," *PTRF*, 2011.
- "From the joint convexity of relative entropy to a concavity theorem of Lieb," *PAMS*, 2012.

- "Improved analysis of the subsampled randomized Hadamard transform," *AADA*, 2011.
- "Tail bounds for all eigenvalues of a sum of random matrices" with A. Gittens. Submitted 2011.
- "The masked sample covariance estimator" with R. Chen and A. Gittens. *I&I*, 2012.
- "Matrix concentration inequalities..." with L. Mackey et al.. Submitted 2012.
- "User-Friendly Tools for Random Matrices: An Introduction." 2012.

See also...

- Ahlswede and Winter, "Strong converse for identification via quantum channels," *Trans. IT*, 2002.
- Oliveira, "Concentration of the adjacency matrix and of the Laplacian." Submitted 2010.
- Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2011.
- Minsker, "Some extensions of Bernstein's inequality for self-adjoint operators," 2011.