Applied Random Matrix Theory

Joel A. Tropp
Steele Family Professor of
Applied & Computational Mathematics

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

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

Open for Business

SIAM Journal on Mathematics of Data Science (SIMODS)

https://simods.siam.org/

What is a Random Matrix?

Definition. A **random matrix** is a matrix whose entries are random variables, not necessarily independent.

A random matrix in captivity:

$$\begin{bmatrix} 0.0000 & -1.3077 & -1.3499 & 0.2050 & 0.0000 \\ 1.8339 & 0.0000 & -1.3077 & 0.0000 & 0.2050 \\ -2.2588 & 1.8339 & 0.0000 & -1.3077 & -1.3499 \\ 2.7694 & 0.0000 & 1.8339 & 0.0000 & -1.3077 \\ 0.0000 & 2.7694 & -2.2588 & 1.8339 & 0.0000 \end{bmatrix}$$

What do we want to understand?

- Eigenvalues
- Eigenvectors
- Singular values
- Singular vectors
- Operator norms
- ...

Sources: Muirhead 1982; Mehta 2004; Nica & Speicher 2006; Bai & Silverstein 2010; Vershynin 2010; Tao 2011; Kemp 2013; Tropp 2015; ...

Random Matrices in Statistics

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:

$$dp = \frac{\begin{vmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{vmatrix}^{\frac{N-1}{2}}}{(\sqrt{\pi})^{\frac{1}{2}n(n-1)}\,\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)\cdots\Gamma\left(\frac{N-n}{2}\right)}$$

$$\times\, e^{-A_{11}a_{11}-A_{22}a_{22}-\cdots-A_{nn}a_{nn}-2A_{12}a_{12}-2A_{13}a_{13}-\cdots-2A_{n-1\,n}a_{n-1\,n}}$$

$$\times \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}^{\frac{N-n-2}{2}} da_{11}\,da_{12}\ldots da_{nn} \quad \ldots\ldots\ldots\ldots(9),$$

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

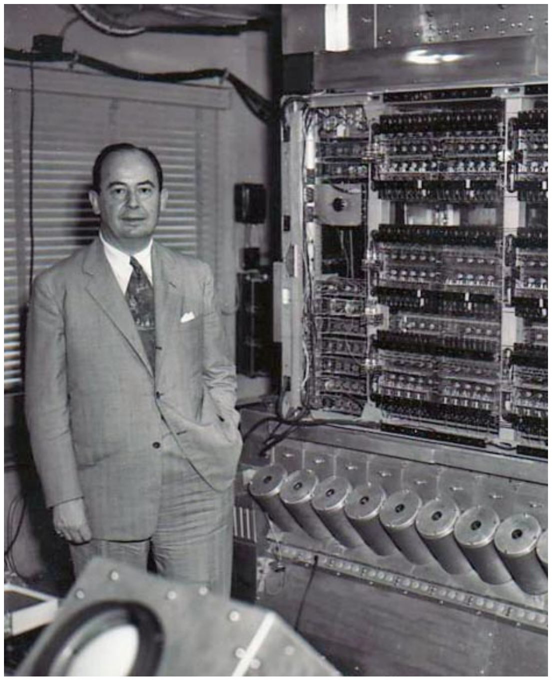
$$|\rho_{pq}|,\ p, q = 1, 2, 3, \ldots n,$$

and Δ_{pq} the minor of ρ_{pq} in Δ.

- Sample covariance matrix for the multivariate normal distribution

Sources: Wishart, *Biometrika* 1928. Photo from `apprendre-math.info`.

Random Matrices in Numerical Linear Algebra



John von Neumann

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

$$\text{Prob}\ (\lambda > 2\sigma^2 rn) < \frac{(rn)^{n-1/2}e^{-rn}\pi^{1/2}e^n \cdot 2^{n-2}}{\pi n^{n-1}(r-1)n} = \left(\frac{2r}{e^{r-1}}\right)^n \times \frac{1}{4(r-1)(r\pi n)^{1/2}}. \tag{8.8}$$

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 $r = 3.70$.

- Model for floating-point errors in LU decomposition

Sources: von Neumann & Goldstine, *Bull. AMS* 1947 and *Proc. AMS* 1951. Photo ©IAS Archive.

Random Matrices in Nuclear Physics

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.

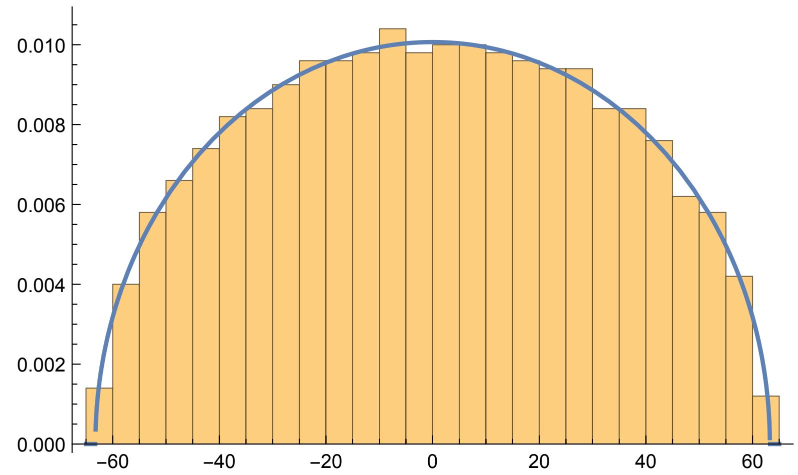
- Model for the Hamiltonian of a heavy atom in a slow nuclear reaction

Sources: Wigner, *Ann. Math.* 1955. Photo from Nobel Foundation.

Classical RMT

$$\begin{bmatrix} 0 & + & - & + & + & - & + \\ & 0 & + & - & - & - & + \\ & & 0 & + & - & + & + \\ & & & 0 & - & - & - \\ & & & & 0 & + & - \\ & * & & & & 0 & + \\ & & & & & & 0 \end{bmatrix}$$

Wigner ($d = 7$)



Distribution of eigenvalues ($d = 10^3$)

- Highly symmetric models
- Very precise results
- Strong resonances with other fields of mathematics

Contemporary Applications of RMT

- Numerical linear algebra
- Numerical analysis
- Uncertainty quantification
- High-dimensional statistics
- Econometrics
- Approximation theory
- Sampling theory
- Machine learning
- Learning theory
- Mathematical signal processing
- Optimization
- Computer graphics and vision
- Quantum information theory
- Theory of algorithms
- Combinatorics
- ...

Sources: (Drawn at random, nonuniformly) Halko et al. 2011; March & Biros 2014; Constantine & Gleich 2015; Koltchinskii 2011; Chen & Christensen 2013; Cohen et al. 2013; Bass & Groechenig 2013; Djolonga et al. 2013; Lopez-Paz et al. 2014; Fornasier et al. 2012; Morvant et al. 2012; Chen et al. 2014; Cheung et al. 2012; Chen et al. 2014; Holevo 2012; Harvey & Olver 2014; Cohen et al. 2014; Oliveira 2014.
Per Google Scholar, over $33,900$ papers with key "Random Matrix Theory."

Contemporary RMT

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$\downarrow$ (sample random columns) $\downarrow$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

- Wide range of examples, many data-driven
- Results may sacrifice precision for applicability
- Theory is still developing

Thesis Statement

Modern applications demand new random matrix models and new analytical tools

Matrix Concentration

- **Goal:** For a random matrix $\boldsymbol{Z}$, find probabilistic bounds for

$$\|\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z}\|$$

- An upper bound on this quantity ensures that
  - Singular values of $\boldsymbol{Z}$ and $\mathbb{E}\,\boldsymbol{Z}$ are close
  - Singular vectors of $\boldsymbol{Z}$ and $\mathbb{E}\,\boldsymbol{Z}$ are close (for isolated singular values)
  - Linear functionals of $\boldsymbol{Z}$ and $\mathbb{E}\,\boldsymbol{Z}$ are close
  - Spectral norm of $\boldsymbol{Z}$ is controlled: $\|\boldsymbol{Z}\| = \|\mathbb{E}\,\boldsymbol{Z}\| \pm \|\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z}\|$

$\|\cdot\|$ = spectral norm = largest singular value = ℓ_2 operator norm

The Independent Sum Model

$$\boldsymbol{Z} = \sum\nolimits_k \boldsymbol{S}_k$$

with $\boldsymbol{S}_k$ independent

Useful observation: $\mathbb{E}\,\boldsymbol{Z} = \sum_k \mathbb{E}\,\boldsymbol{S}_k$

Exercise: Express the sample covariance matrix in this model

Exercise: Express column sampling (with replacement) from a fixed matrix

The Bernstein Inequality

Fact 1 (Bernstein 1920s). **Suppose**

- ❧ $S_1, S_2, S_3, \dots$ *are independent real random variables*
- ❧ *Each one is centered:* $\mathbb{E}\, S_k = 0$
- ❧ *Each one is bounded:* $|S_k| \le L$

Then, *for* $t > 0$,

$$\mathbb{P}\left\{\left|\sum_k S_k\right| \ge t\right\} \quad \le \quad 2 \cdot \begin{cases} \mathrm{e}^{-\mathrm{c}t^2/v}, & t \le v/L \\ \mathrm{e}^{-\mathrm{c}t/L}, & t \ge v/L \end{cases} \qquad (\mathrm{c} = 3/8)$$

where the variance proxy is

$$v = \mathrm{Var}\left(\sum_k S_k\right) = \sum_k \mathbb{E}\, S_k^2$$

Sources: Bernstein 1927; Boucheron et al. 2013.

The Matrix Bernstein Inequality I

Theorem 2 (T 2011). **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are independent random matrices with dimension* $d_1 \times d_2$
- *Each one is centered:* $\mathbb{E}\, \boldsymbol{S}_k = \mathbf{0}$
- *Each one is bounded:* $\|\boldsymbol{S}_k\| \leq L$

Then, *for* $t > 0$,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \boldsymbol{S}_k\right\| \geq t\right\} \quad \leq \quad (d_1 + d_2) \cdot \begin{cases} \mathrm{e}^{-\mathrm{c}t^2/v}, & t \leq v/L \\ \mathrm{e}^{-\mathrm{c}t/L}, & t \geq v/L \end{cases} \qquad (\mathrm{c} = 3/8)$$

where the matrix variance proxy is

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

Sources: Tomczak-Jaegermann 1973; Lust-Piquard 1986; Pisier 1998; Rudelson 1999; Ahlswede & Winter 2002; Junge & Xu 2003, 2008; Rudelson & Vershynin 2005; Gross 2011; Recht 2011; Oliveira 2011; Tropp 2011–2015.

The Matrix Bernstein Inequality II

Theorem 3 (T 2011). **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are independent random matrices with dimension* $d_1 \times d_2$
- *Each one is centered:* $\mathbb{E}\, \boldsymbol{S}_k = \mathbf{0}$
- *Each one is bounded:* $\|\boldsymbol{S}_k\| \leq L$

Then

$$\mathbb{E} \left\| \sum_k \boldsymbol{S}_k \right\| \quad \leq \quad \sqrt{2v \cdot \log(d_1 + d_2)} + \frac{1}{3} L \cdot \log(d_1 + d_2)$$

where the matrix variance proxy is

$$v = \max \left\{ \left\| \sum_k \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \left\| \sum_k \mathbb{E}(\boldsymbol{S}_k^* \boldsymbol{S}_k) \right\| \right\}$$

Sources: Tomczak-Jaegermann 1973; Lust-Piquard 1986; Pisier 1998; Rudelson 1999; Ahlswede & Winter 2002; Junge & Xu 2003, 2008; Rudelson & Vershynin 2005; Gross 2011; Recht 2011; Oliveira 2011; Tropp 2011–2015.

Example: Matrix Sparsification

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix} \longrightarrow \hat{\boldsymbol{A}} = \begin{bmatrix} & 2 & & \\ & 4 & & 8 \\ 3 & 6 & 9 & 12 \\ & & 12 & 16 \end{bmatrix}$$

- **Goal:** Find a sparse matrix $\hat{\boldsymbol{A}}$ for which $\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|$ is small
- **Approach:** Non-uniform randomized sampling

Sources: Achlioptas & McSherry 2001, 2007; Arora et al. 2006; d'Asprémont 2008; Gittens & Tropp 2009; Nguyen et al. 2009; Drineas & Zouzias 2011; Achlioptas et al. 2013; Kundu & Drineas 2014; Tropp 2015.

Sparsification: Sampling Model

- Let $\boldsymbol{A}$ be a fixed $d_1 \times d_2$ matrix
- Construct a probability mass $\{p_{ij}\}$ on the matrix indices
- Define a 1-sparse random matrix $\boldsymbol{S}$ where

$$\boldsymbol{S} = \frac{a_{ij}}{p_{ij}} \mathbf{E}_{ij} \quad \text{with probability } p_{ij}$$

- The random matrix $\boldsymbol{S}$ is an unbiased estimator for $\boldsymbol{A}$

$$\mathbb{E}\, \boldsymbol{S} = \sum\nolimits_{ij} \frac{a_{ij}}{p_{ij}} \mathbf{E}_{ij} \cdot p_{ij} = \sum\nolimits_{ij} a_{ij} \mathbf{E}_{ij} = \boldsymbol{A}$$

- To reduce the variance, average r independent copies of $\boldsymbol{S}$

$$\hat{\boldsymbol{A}}_r = \frac{1}{r} \sum\nolimits_{k=1}^{r} \boldsymbol{S}_k \quad \text{where } \boldsymbol{S}_k \sim \boldsymbol{S}$$

- By construction, $\hat{\boldsymbol{A}}_r$ has at most r nonzero entries and approximates $\boldsymbol{A}$

Sparsification: Analysis

- **Recall:** $\boldsymbol{S} = (a_{ij}/p_{ij})\mathbf{E}_{ij}$ with probability p_{ij}
- Bound for spectral norm:

$$\|\boldsymbol{S} - \mathbb{E}\,\boldsymbol{S}\| \leq 2 \cdot \max_{ij} \frac{|a_{ij}|}{p_{ij}}$$

- Bound for variance:

$$\left\|\mathbb{E}(\boldsymbol{S} - \mathbb{E}\,\boldsymbol{S})(\boldsymbol{S} - \mathbb{E}\,\boldsymbol{S})^*\right\| \leq \left\|\mathbb{E}\,\boldsymbol{S}\boldsymbol{S}^*\right\| = \left\|\sum\nolimits_i \left(\sum\nolimits_j \frac{|a_{ij}|^2}{p_{ij}}\right)\mathbf{E}_{ii}\right\| = \max_i \sum\nolimits_j \frac{|a_{ij}|^2}{p_{ij}}$$

$$\left\|\mathbb{E}(\boldsymbol{S} - \mathbb{E}\,\boldsymbol{S})^*(\boldsymbol{S} - \mathbb{E}\,\boldsymbol{S})\right\| \leq \left\|\mathbb{E}\,\boldsymbol{S}^*\boldsymbol{S}\right\| = \left\|\sum\nolimits_j \left(\sum\nolimits_i \frac{|a_{ij}|^2}{p_{ij}}\right)\mathbf{E}_{jj}\right\| = \max_j \sum\nolimits_i \frac{|a_{ij}|^2}{p_{ij}}$$

- Construct probability mass $p_{ij} \propto |a_{ij}| + |a_{ij}|^2$ to control all terms

Sparsification: Result

Proposition 4 (Kundu & Drineas 2014; T 2015). **Suppose**

$$r \geq \varepsilon^{-2} \cdot \operatorname{srank}(\boldsymbol{A}) \cdot \max\{d_1, d_2\} \log(d_1 + d_2) \qquad (0 < \varepsilon \leq 1)$$

Then *the relative error in the* r*-sparse approximation* $\hat{\boldsymbol{A}}_r$ *satisfies*

$$\frac{\mathbb{E}\,\| \boldsymbol{A} - \hat{\boldsymbol{A}}_r \|}{\| \boldsymbol{A} \|} \leq 4\varepsilon$$

The stable rank

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

- The proof is an immediate consequence of matrix Bernstein

Application: Fast Laplacian Solvers

Theorem 5 (Kyng & Sachdeva 2016). **Suppose**

- *G is a weighted, undirected graph with n vertices and m edges*
- *$\boldsymbol{L}$ is the combinatorial Laplacian of the graph G*

Then, *with high probability, the* SPARSECHOLESKY *algorithm produces*

- *A lower-triangular matrix $\boldsymbol{C}$ with $\tilde{O}(m)$ nonzero entries that satisfies*

$$\frac{1}{2}\boldsymbol{L} \preccurlyeq \boldsymbol{C}\boldsymbol{C}^* \preccurlyeq \frac{3}{2}\boldsymbol{L}$$

- *The running time is $\tilde{O}(m)$*

In particular, we can solve $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{b}$ to machine precision in time $\tilde{O}(m)$

SPARSECHOLESKY for a Graph Laplacian

$$\boldsymbol{L} = \begin{bmatrix} a & \boldsymbol{u}^* \\ \boldsymbol{u} & \boldsymbol{L}_2 \end{bmatrix}_{n \times n} \quad \to \quad \boldsymbol{L}_2 - a^{-1} \begin{bmatrix} & & \\ & \boldsymbol{u}\boldsymbol{u}^* & \\ & & \end{bmatrix}_{(n-1)\times(n-1)}$$

Subtract rank-1

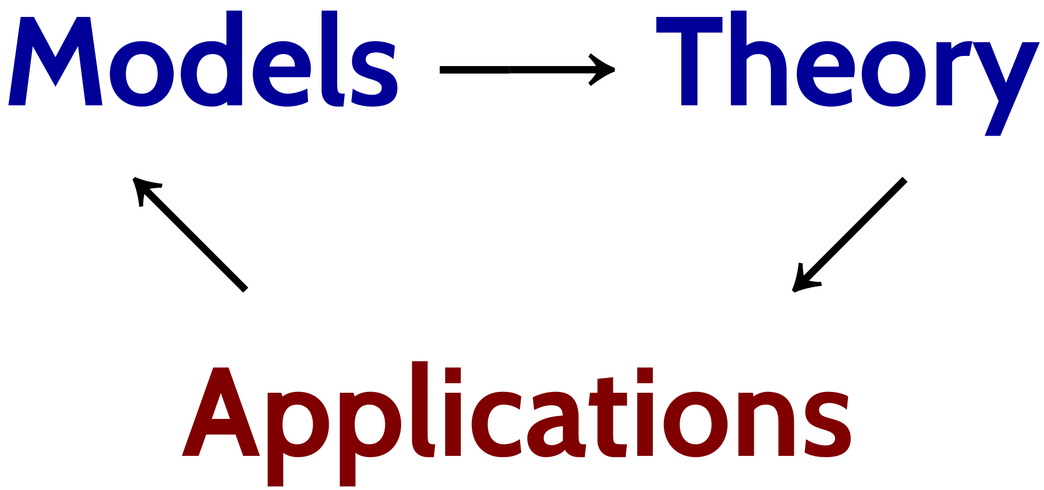
$$\to \quad \boldsymbol{L}_2 - a^{-1} \begin{bmatrix} & \times & \\ \times & & \times \\ & \times & \times \end{bmatrix}_{(n-1)\times(n-1)}$$

Sparsify rank-1

- Direct computation of Cholesky factorization requires $O(n^2)$ operations per step
- Randomized approximation in $\tilde{O}(m/n)$ operations per step (amortized)
- Sampling probabilities are computed using graph theory
- Analysis depends on Bernstein inequality for matrix martingales!

Sources: Pisier & Xu 1997; Junge & Xu 2003, 2008; Oliveira 2011; Tropp 2011; Kyng & Sachdeva 2016.

A Virtuous Cycle



Contact & Papers

email: `jtropp@cms.caltech.edu`
web: `http://users.cms.caltech.edu/~jtropp`

Monograph:

- *An Introduction to Matrix Concentration Inequalities*. *Found. Trends Mach. Learn.*, 2015. Preprint: `arXiv:1501.01571`

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
- "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
- "The masked sample covariance estimator" with R. Chen & A. Gittens. *I&I*, 2012
- "Tail bounds for all eigenvalues of a sum of random matrices" with A. Gittens. Caltech ACM Report 2014-02
- "Matrix concentration inequalities via the method of exchangeable pairs" with L. Mackey et al. *Ann. Probab.*, 2014
- "Subadditivity of matrix φ-entropy and concentration of random matrices" with R. Chen. *EJP*, 2014
- "Efron–Stein inequalities for random matrices" with D. Paulin & L. Mackey. *Ann. Probab.*, 2016
- "Second-order matrix concentration inequalities." *ACHA*, 2016
- "The expected norm of a sum of independent random matrices: An elementary approach." *HDP 7*, 2016