Stein's Method for Matrix Concentration

Lester Mackey[†]

Collaborators:
Michael I. Jordan[†], Richard Y. Chen[*], Brendan Farrell[*], and Joel A. Tropp[*]

[†]University of California, Berkeley [*]California Institute of Technology

BEARS 2012

Concentration Inequalities

Matrix concentration

$$\mathbb{P}\{\|\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}\| \geq t\} \leq \delta$$
$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \geq t\} \leq \delta$$

- Non-asymptotic control of random matrices with complex distributions

Applications

- Matrix estimation from sparse random measurements
  (Gross, 2011; Recht, 2009; Mackey, Talwalkar, and Jordan, 2011)
- Randomized matrix multiplication and factorization
  (Drineas, Mahoney, and Muthukrishnan, 2008; Hsu, Kakade, and Zhang, 2011b)
- Convex relaxation of robust or chance-constrained optimization
  (Nemirovski, 2007; So, 2011; Cheung, So, and Wang, 2011)
- Random graph analysis (Christofides and Markström, 2008; Oliveira, 2009)

Concentration Inequalities

Matrix concentration

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \geq t\} \leq \delta$$

Difficulty: Matrix multiplication is not commutative

Past approaches (Oliveira, 2009; Tropp, 2011; Hsu, Kakade, and Zhang, 2011a)

- Deep results from matrix analysis
- Sums of independent matrices and matrix martingales

This work

- Stein's method of exchangeable pairs (1972), as advanced by Chatterjee (2007) for scalar concentration
  - $\Rightarrow$ Improved exponential tail inequalities (Hoeffding, Bernstein)
  - $\Rightarrow$ Polynomial moment inequalities (Khintchine, Rosenthal)
  - $\Rightarrow$ Dependent sums and more general matrix functionals

Roadmap

Notation

Hermitian matrices: $\mathbb{H}^d = \{\boldsymbol{A} \in \mathbb{C}^{d\times d} : \boldsymbol{A} = \boldsymbol{A}^*\}$

- *All matrices in this talk are Hermitian.*

Maximum eigenvalue: $\lambda_{\max}(\cdot)$

Trace: $\operatorname{tr} \boldsymbol{B}$, the sum of the diagonal entries of $\boldsymbol{B}$

Spectral norm: $\|\boldsymbol{B}\|$, the maximum singular value of $\boldsymbol{B}$

Schatten p-norm: $\|\boldsymbol{B}\|_p := \left(\operatorname{tr}|\boldsymbol{B}|^p\right)^{1/p}$ for $p \geq 1$

Matrix Stein Pair

Definition (Exchangeable Pair)

(Z, Z') is an *exchangeable pair* if $(Z, Z') \stackrel{d}{=} (Z', Z)$.

Definition (Matrix Stein Pair)

Let (Z, Z') be an auxiliary exchangeable pair, and let $\mathbf{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function. Define the random matrices

$$\boldsymbol{X} := \mathbf{\Psi}(Z) \quad \text{and} \quad \boldsymbol{X}' := \mathbf{\Psi}(Z').$$

$(\boldsymbol{X}, \boldsymbol{X}')$ is a *matrix Stein pair* with scale factor $\alpha \in (0, 1]$ if

$$\mathbb{E}[\boldsymbol{X}' \mid Z] = (1 - \alpha)\boldsymbol{X}.$$

- Matrix Stein pairs are exchangeable pairs
- Matrix Stein pairs always have zero mean

The Conditional Variance

Definition (Conditional Variance)

Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a matrix Stein pair with scale factor α, constructed from the exchangeable pair (Z, Z'). The *conditional variance* is the random matrix

$$\boldsymbol{\Delta}_{\boldsymbol{X}} := \boldsymbol{\Delta}_{\boldsymbol{X}}(Z) := \frac{1}{2\alpha}\,\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{X}')^2 \,|\, Z\right].$$

- $\boldsymbol{\Delta}_{\boldsymbol{X}}$ is a stochastic estimate for the variance, $\mathbb{E}\,\boldsymbol{X}^2$
- Control over $\boldsymbol{\Delta}_{\boldsymbol{X}}$ yields control over $\lambda_{\max}(\boldsymbol{X})$

Exponential Concentration for Random Matrices

Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

Let $(\boldsymbol{X}, \boldsymbol{X}')$ be a matrix Stein pair with $\boldsymbol{X} \in \mathbb{H}^d$. Suppose that

$$\boldsymbol{\Delta}_{\boldsymbol{X}} \preccurlyeq c\boldsymbol{X} + v\,\mathbf{I} \quad \text{almost surely for} \quad c, v \geq 0.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \geq t\} \leq d \cdot \exp\left\{\frac{-t^2}{2v + 2ct}\right\}.$$

- Control over the conditional variance $\boldsymbol{\Delta}_{\boldsymbol{X}}$ yields
  - Gaussian tail for $\lambda_{\max}(\boldsymbol{X})$ for small t, Poisson tail for large t
- When $d = 1$, reduces to scalar result of Chatterjee (2007)
- The dimensional factor d cannot be removed

Application: Matrix Hoeffding Inequality

Corollary (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

Let $(\boldsymbol{Y}_k)_{k \geq 1}$ be independent matrices in $\mathbb{H}^d$ satisfying

$$\mathbb{E}\, \boldsymbol{Y}_k = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{Y}_k^2 \preccurlyeq \boldsymbol{A}_k^2$$

for deterministic matrices $(\boldsymbol{A}_k)_{k \geq 1}$. Define the variance parameter

$$\sigma^2 := \frac{1}{2} \left\| \sum\nolimits_k \left(\boldsymbol{A}_k^2 + \mathbb{E}\, \boldsymbol{Y}_k^2 \right) \right\|.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \boldsymbol{Y}_k \right) \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}.$$

- Improves upon the matrix Hoeffding inequality of Tropp (2011)
  - Optimal constant $1/2$ in the exponent
  - Variance parameter σ^2 smaller than the bound $\left\| \sum_k \boldsymbol{A}_k^2 \right\|$
- Tighter than classical Hoeffding inequality (1963) when $d = 1$

Exponential Concentration: Proof Sketch

1. Matrix Laplace transform method (Ahlswede & Winter, 2002)

- Relate tail probability to the *trace* of the mgf of $\boldsymbol{X}$

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \geq t\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot m(\theta)$$

where $m(\theta) := \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \boldsymbol{X}}$

How to bound the trace mgf?

- Past approaches: Golden-Thompson, Lieb's concavity theorem
- Chatterjee's strategy for scalar concentration
  - Control mgf growth by bounding derivative

$$m'(\theta) = \mathbb{E}\,\mathrm{tr}\,\boldsymbol{X}\mathrm{e}^{\theta \boldsymbol{X}} \quad \text{for } \theta \in \mathbb{R}.$$

  - Rewrite using exchangeable pairs

Method of Exchangeable Pairs

Lemma

Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a matrix Stein pair with scale factor α. Let $\boldsymbol{F} : \mathbb{H}^d \to \mathbb{H}^d$ be a measurable function satisfying

$$\mathbb{E}\|(\boldsymbol{X} - \boldsymbol{X}')\boldsymbol{F}(\boldsymbol{X})\| < \infty.$$

Then

$$\mathbb{E}[\boldsymbol{X}\ \boldsymbol{F}(\boldsymbol{X})] = \frac{1}{2\alpha}\mathbb{E}[(\boldsymbol{X} - \boldsymbol{X}')(\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{F}(\boldsymbol{X}'))]. \qquad (1)$$

Intuition

- Can characterize the distribution of a random matrix by integrating it against a class of test functions $\boldsymbol{F}$
- Eq. 1 allows us to estimate this integral using the smoothness properties of $\boldsymbol{F}$ and the discrepancy $\boldsymbol{X} - \boldsymbol{X}'$

Exponential Concentration: Proof Sketch

2. Method of Exchangeable Pairs

- Rewrite the derivative of the trace mgf

$$m'(\theta) = \mathbb{E} \operatorname{tr} \boldsymbol{X} \mathrm{e}^{\theta \boldsymbol{X}} = \frac{1}{2\alpha} \mathbb{E} \operatorname{tr} \left[(\boldsymbol{X} - \boldsymbol{X}') \left(\mathrm{e}^{\theta \boldsymbol{X}} - \mathrm{e}^{\theta \boldsymbol{X}'} \right) \right].$$

Goal: Use the smoothness of $\boldsymbol{F}(\boldsymbol{X}) = \mathrm{e}^{\theta \boldsymbol{X}}$ to bound the derivative

Mean Value Trace Inequality

Lemma (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a weakly increasing function and that $h : \mathbb{R} \to \mathbb{R}$ is a function whose derivative h' is convex. For all matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d$, it holds that

$$\operatorname{tr}[(g(\boldsymbol{A}) - g(\boldsymbol{B})) \cdot (h(\boldsymbol{A}) - h(\boldsymbol{B}))] \leq \frac{1}{2} \operatorname{tr}[(g(\boldsymbol{A}) - g(\boldsymbol{B})) \cdot (\boldsymbol{A} - \boldsymbol{B}) \cdot (h'(\boldsymbol{A}) + h'(\boldsymbol{B}))].$$

- *Standard matrix functions:* If $g : \mathbb{R} \to \mathbb{R}$, then

$$g(\boldsymbol{A}) := \boldsymbol{Q} \begin{bmatrix} g(\lambda_1) & & \\ & \ddots & \\ & & g(\lambda_d) \end{bmatrix} \boldsymbol{Q}^* \quad \text{when} \quad \boldsymbol{A} := \boldsymbol{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \boldsymbol{Q}^*$$

- Inequality does not hold without the trace
- For exponential concentration we let $g(\boldsymbol{A}) = \boldsymbol{A}$ and $h(\boldsymbol{B}) = \mathrm{e}^{\theta \boldsymbol{B}}$

Exponential Concentration: Proof Sketch

3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

$$\begin{aligned} m'(\theta) &= \frac{1}{2\alpha}\,\mathbb{E}\operatorname{tr}\left[(\boldsymbol{X}-\boldsymbol{X}')\left(\mathrm{e}^{\theta\boldsymbol{X}}-\mathrm{e}^{\theta\boldsymbol{X}'}\right)\right] \\ &\le \frac{\theta}{4\alpha}\,\mathbb{E}\operatorname{tr}\left[(\boldsymbol{X}-\boldsymbol{X}')^2\cdot\left(\mathrm{e}^{\theta\boldsymbol{X}}+\mathrm{e}^{\theta\boldsymbol{X}'}\right)\right] \\ &= \theta\cdot\mathbb{E}\operatorname{tr}\left[\boldsymbol{\Delta}_{\boldsymbol{X}}\,\mathrm{e}^{\theta\boldsymbol{X}}\right]. \end{aligned}$$

4. Conditional Variance Bound: $\boldsymbol{\Delta}_{\boldsymbol{X}} \preccurlyeq c\boldsymbol{X} + v\,\mathbf{I}$

- Yields differential inequality

$$m'(\theta) \le c\theta\cdot m'(\theta) + v\theta\cdot m(\theta).$$

- Solve to bound $m(\theta)$ and thereby bound $\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \ge t\}$

Polynomial Moments for Random Matrices

Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

Let $p = 1$ or $p \geq 1.5$. Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a matrix Stein pair where $\mathbb{E}\|\boldsymbol{X}\|_{2p}^{2p} < \infty$. Then

$$\left(\mathbb{E}\|\boldsymbol{X}\|_{2p}^{2p}\right)^{1/2p} \leq \sqrt{2p-1} \cdot \left(\mathbb{E}\|\boldsymbol{\Delta}_{\boldsymbol{X}}\|_{p}^{p}\right)^{1/2p}.$$

- **Moral:** The conditional variance controls the moments of $\boldsymbol{X}$
- Generalizes Chatterjee's version (2007) of the scalar Burkholder-Davis-Gundy inequality (Burkholder, 1973)
  - See also Pisier & Xu (1997); Junge & Xu (2003, 2008)
- Proof techniques mirror those for exponential concentration
- Also holds for infinite dimensional Schatten-class operators

Application: Matrix Khintchine Inequality

Corollary (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

Let $(\varepsilon_k)_{k\geq 1}$ be an independent sequence of Rademacher random variables and $(\boldsymbol{A}_k)_{k\geq 1}$ be a deterministic sequence of Hermitian matrices. Then if $p = 1$ or $p \geq 1.5$,

$$\left(\mathbb{E}\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right\|_{2p}^{2p}\right)^{1/2p} \leq \sqrt{2p-1} \cdot \left\|\left(\sum\nolimits_k \boldsymbol{A}_k^2\right)^{1/2}\right\|_{2p}.$$

- Noncommutative Khintchine inequality (Lust-Piquard, 1986; Lust-Piquard and Pisier, 1991) is a dominant tool in applied matrix analysis
  - e.g., Used in analysis of column sampling and projection for approximate SVD (Rudelson and Vershynin, 2007)
- Stein's method offers an unusually concise proof
- The constant $\sqrt{2p-1}$ is within $\sqrt{\mathrm{e}}$ of optimal

Extensions

Refined Exponential Concentration

- Relate trace mgf of conditional variance to trace mgf of $\boldsymbol{X}$
- Yields matrix generalization of classical Bernstein inequality
- Offers tool for unbounded random matrices

General Complex Matrices

- Map any matrix $\boldsymbol{B} \in \mathbb{C}^{d_1 \times d_2}$ to a Hermitian matrix via *dilation*

$$\mathscr{D}(\boldsymbol{B}) := \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix} \in \mathbb{H}^{d_1+d_2}.$$

- Preserves spectral information: $\lambda_{\max}(\mathscr{D}(\boldsymbol{B})) = \|\boldsymbol{B}\|$

Dependent Sequences

- Sums of conditionally zero-mean random matrices
- Combinatorial matrix statistics (e.g., sampling w/o replacement)
- Matrix-valued functions satisfying a self-reproducing property
  - Yields a dependent bounded differences inequality for matrices

The End

Thanks!

References I

Ahlswede, R. and Winter, A. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3): 569–579, Mar. 2002.

Burkholder, D. L. Distribution function inequalities for martingales. *Ann. Probab.*, 1:19–42, 1973. doi: 10.1214/aop/1176997023.

Chatterjee, S. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138:305–321, 2007.

Cheung, S.-S., So, A. Man-Cho, and Wang, K. Chance-constrained linear matrix inequalities with dependent perturbations: A safe tractable approximation approach. Available at `http://www.se.cuhk.edu.hk/~manchoso/papers/cclmi_sta.pdf`, 2011.

Christofides, D. and Markström, K. Expansion properties of random cayley graphs and vertex transitive graphs via matrix martingales. *Random Struct. Algorithms*, 32(1):88–100, 2008.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, Mar. 2011.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hsu, D., Kakade, S. M., and Zhang, T. Dimension-free tail inequalities for sums of random matrices. Available at `arXiv:1104.1672`, 2011a.

Hsu, D., Kakade, S. M., and Zhang, T. Dimension-free tail inequalities for sums of random matrices. `arXiv:1104.1672v3[math.PR]`, 2011b.

Junge, M. and Xu, Q. Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.*, 31(2):948–995, 2003.

Junge, M. and Xu, Q. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel J. Math.*, 167:227–282, 2008.

Lust-Piquard, F. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

Lust-Piquard, F. and Pisier, G. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

Mackey, L., Talwalkar, A., and Jordan, M. I. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems 24*. 2011.

References II

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. Matrix concentration inequalities via the method of exchangeable pairs. Available at arXiv, Jan. 2012.

Nemirovski, A. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Program.*, 109:283–317, January 2007. ISSN 0025-5610. doi: 10.1007/s10107-006-0033-0. URL http://dl.acm.org/citation.cfm?id=1229716.1229726.

Oliveira, R. I. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv:0911.0600, Nov. 2009.

Pisier, G. and Xu, Q. Non-commutative martingale inequalities. *Comm. Math. Phys.*, 189(3):667–698, 1997.

Recht, B. A simpler approach to matrix completion. arXiv:0910.0651v2[cs.IT], 2009.

Rudelson, M. and Vershynin, R. Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, 54(4):Article 21, 19 pp., Jul. 2007. (electronic).

So, A. Man-Cho. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Program.*, 130 (1):125–151, 2011.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, Berkeley, 1972. Univ. California Press.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, August 2011.