Stein's Method for Matrix Concentration

Lester Mackey[†]

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

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

December 10, 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 completion from sparse random measurements

  (Gross, 2011; Recht, 2011; Negahban and Wainwright, 2010; 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)

Motivation: Matrix Completion

Goal: Recover a matrix $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ from a subset of its entries

$$\begin{bmatrix} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 & 1 & \dots & 4 \\ 3 & 4 & 5 & \dots & 1 \\ 2 & 5 & 3 & \dots & 5 \end{bmatrix}$$

Examples

- Collaborative filtering: How will user i rate movie j?
- Ranking on the web: Is URL j relevant to user i?
- Link prediction: Is user i friends with user j?

Motivation: Matrix Completion

Goal: Recover a matrix $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ from a subset of its entries

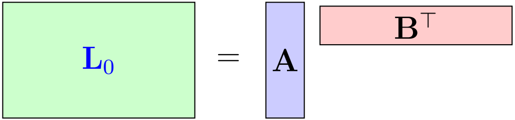
$$\begin{bmatrix} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 & 1 & \dots & 4 \\ 3 & 4 & 5 & \dots & 1 \\ 2 & 5 & 3 & \dots & 5 \end{bmatrix}$$

Bad News: Impossible to recover a generic matrix

- Too many degrees of freedom, too few observations

Good News:

- Small number of latent factors determine preferences
  - Movie ratings cluster by genre and director



- These **low-rank** matrices are easier to complete

How to Complete a Low-rank Matrix

Suppose Ω is the set of observed entry locations.

First attempt:

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \operatorname{rank} \mathbf{A} \\ \text{subject to} & \mathbf{A}_{ij} = \mathbf{L}_{0ij} \quad (i,j) \in \Omega \end{array}$$

Problem: NP-hard $\Rightarrow$ computationally intractable!

Solution: Solve **convex** relaxation (?)

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \|\mathbf{A}\|_* \\ \text{subject to} & \mathbf{A}_{ij} = \mathbf{L}_{0ij} \quad (i,j) \in \Omega \end{array}$$

where $\|\mathbf{A}\|_* = \sum_k \sigma_k(\mathbf{A})$ is the trace/nuclear norm of $\mathbf{A}$.

Can Convex Optimization Recover $\mathbf{L}_0$?

Yes, with high probability.

Theorem (Recht, 2011)

If $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ has rank r and $s \gtrsim \beta r n \log^2(n)$ entries are observed uniformly at random, then (under some technical conditions) convex optimization **recovers** $\mathbf{L}_0$ **exactly** with probability at least $1 - n^{-\beta}$.

- See also Gross (2011); Mackey, Talwalkar, and Jordan (2011)
- Past results (Candès and Recht, 2009; Candès and Tao, 2009) required stronger assumptions and more intensive analysis
- Streamlined approach reposes on a matrix variant of a classical Bernstein inequality (1946)

Scalar Bernstein Inequality

Theorem (Bernstein, 1946)

Let $(Y_k)_{k\geq 1}$ be independent random variables in $\mathbb{R}$ satisfying

$$\mathbb{E}\, Y_k = 0 \quad \text{and} \quad |Y_k| \leq R \quad \text{for each index } k.$$

Define the variance parameter

$$\sigma^2 := \sum\nolimits_k \mathbb{E}\, Y_k^2.$$

Then, for all $t \geq 0$,

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

- Gaussian decay controlled by variance when t is small
- Exponential decay controlled by uniform bound for large t

Matrix Bernstein Inequality

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

Let $(\boldsymbol{Y}_k)_{k\geq 1}$ be independent matrices in $\mathbb{R}^{m\times n}$ satisfying

$$\mathbb{E}\,\boldsymbol{Y}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{Y}_k\| \leq R \quad \text{for each index } k.$$

Define the variance parameter

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

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \boldsymbol{Y}_k\right\| \geq t\right\} \leq (m+n)\cdot\exp\left\{\frac{-t^2}{3\sigma^2 + 2Rt}\right\}$$

- See also Tropp (2011); Oliveira (2009); Recht (2011)
- Gaussian tail when t is small; exponential tail for large t

Matrix Bernstein Inequality

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

For all $t \geq 0$,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \boldsymbol{Y}_k\right\| \geq t\right\} \leq (m+n) \cdot \exp\left\{\frac{-t^2}{3\sigma^2 + 2Rt}\right\}$$

Consequences for matrix completion

- Recht (2011) showed that uniform sampling of entries captures most of the information in incoherent low-rank matrices
- Negahban and Wainwright (2010) showed that i.i.d. sampling of entries captures most of the information in non-spiky (near) low-rank matrices
- Foygel and Srebro (2011) characterized the generalization error of convex MC through Rademacher complexity

Concentration Inequalities

Matrix concentration

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

Difficulty: Matrix multiplication is not commutative

$$\Rightarrow \mathrm{e}^{\boldsymbol{X}+\boldsymbol{Y}} \neq \mathrm{e}^{\boldsymbol{X}}\mathrm{e}^{\boldsymbol{Y}}$$

Past approaches (Ahlswede and Winter, 2002; Oliveira, 2009; Tropp, 2011)

- Rely on deep results from matrix analysis
- Apply to sums of independent matrices and matrix martingales

This work

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

Roadmap

1. Motivation
2. Stein's Method Background and Notation
3. Exponential Tail Inequalities
4. Polynomial Moment Inequalities
5. Dependent Sequences
6. Extensions

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 $\boldsymbol{B}$

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 exchangeable pair, and let $\boldsymbol{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function. Define the random matrices

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

$(\boldsymbol{X}, \boldsymbol{X}')$ 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, $\mathbb{E}\,\boldsymbol{X}^2$

Take-home Message

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, exponential tail for large t
- When $d = 1$, improves scalar result of Chatterjee (2007)
- The dimensional factor d cannot be removed

Matrix Hoeffding Inequality

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

Let $\boldsymbol{X} = \sum_k \boldsymbol{Y}_k$ for 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

$$\sigma^2 := \left\|\sum\nolimits_k \boldsymbol{A}_k^2\right\|.$$

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
- Can replace variance parameter with $\sigma^2 = \frac{1}{2}\left\|\sum_k \left(\boldsymbol{A}_k^2 + \mathbb{E}\,\boldsymbol{Y}_k^2\right)\right\|$
  - 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 $m(\theta) := \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{X}}$

- Problem: $\mathrm{e}^{\boldsymbol{X}+\boldsymbol{Y}} \neq \mathrm{e}^{\boldsymbol{X}}\mathrm{e}^{\boldsymbol{Y}}$ when $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{H}^d$

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

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

  - 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

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

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 $g : \mathbb{R} \to \mathbb{R}$ and

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

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

Exponential Concentration: Proof Sketch

3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

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

Exponential Concentration: Proof Sketch

3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

$$m'(\theta) \le \theta \cdot \mathbb{E}\,\mathrm{tr}\left[\boldsymbol{\Delta}_{\boldsymbol{X}}\,\mathrm{e}^{\theta \boldsymbol{X}}\right].$$

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

- Yields differential inequality

$$\begin{aligned} m'(\theta) &\le c\theta\,\mathbb{E}\,\mathrm{tr}\left[\boldsymbol{X}\,\mathrm{e}^{\theta \boldsymbol{X}}\right] + v\theta\,\mathbb{E}\,\mathrm{tr}\left[\mathrm{e}^{\theta \boldsymbol{X}}\right] \\ &= c\theta \cdot m'(\theta) + v\theta \cdot m(\theta). \end{aligned}$$

- Solve to bound $m(\theta)$ and thereby bound

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \ge t\} \le \inf_{\theta > 0} \mathrm{e}^{-\theta t} \cdot m(\theta) \le d \cdot \exp\left\{\frac{-t^2}{2v + 2ct}\right\}.$$

Refined Exponential Concentration

Relaxing the constraint $\mathbf{\Delta}_{\boldsymbol{X}} \preccurlyeq c\boldsymbol{X} + v$

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

Let $(\boldsymbol{X}, \boldsymbol{X}')$ be a bounded matrix Stein pair with $\boldsymbol{X} \in \mathbb{H}^d$. Define the function

$$r(\psi) := \frac{1}{\psi} \log \mathbb{E} \operatorname{tr}(\mathrm{e}^{\psi \mathbf{\Delta}_{\boldsymbol{X}}}/d) \quad \text{for each } \psi > 0.$$

Then, for all $t \geq 0$ and all $\psi > 0$,

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

- $r(\psi)$ measures typical magnitude of conditional variance
  - $\mathbb{E}\, \lambda_{\max}(\mathbf{\Delta}_{\boldsymbol{X}}) \leq \inf_{\psi > 0}\left[r(\psi) + \frac{\log d}{\psi}\right]$
- When $d = 1$, improves scalar result of Chatterjee (2008)
- Proof extends to unbounded random matrices

Matrix Bernstein Inequality

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

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

$$\mathbb{E}\,\boldsymbol{Y}_k = \boldsymbol{0} \quad \text{and} \quad \|\boldsymbol{Y}_k\| \leq R \quad \text{for each index } k.$$

Define the variance parameter

$$\sigma^2 := \left\| \sum\nolimits_k \mathbb{E}\,\boldsymbol{Y}_k^2 \right\|.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \boldsymbol{Y}_k \right) \geq t \right\} \leq d \cdot \exp\left\{ \frac{-t^2}{3\sigma^2 + 2Rt} \right\}$$

- Gaussian tail controlled by improved variance when t is small
- **Key proof idea:** Apply refined concentration, and bound $r(\psi) = \frac{1}{\psi} \log \mathbb{E}\,\mathrm{tr}(\mathrm{e}^{\psi \boldsymbol{\Delta}_{\boldsymbol{X}}}/d)$ using unrefined concentration
- Constants better than Oliveira (2009), worse than Tropp (2011)

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 $\mathbb{E}\operatorname{tr}|\boldsymbol{X}|^{2p} < \infty$. Then

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

- **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

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 $p \geq 1.5$,

$$\mathbb{E}\operatorname{tr}\Big(\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\Big)^{2p} \leq (2p-1)^p \cdot \operatorname{tr}\Big(\sum\nolimits_k \boldsymbol{A}_k^2\Big)^p.$$

- 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

Adding Dependence

1. Motivation
   - Matrix Completion
   - Matrix Concentration
2. Stein's Method Background and Notation
3. Exponential Tail Inequalities
4. Polynomial Moment Inequalities
5. Dependent Sequences
   - Sums of Conditionally Zero-mean Matrices
   - Combinatorial Sums
6. Extensions

Sums of Conditionally Zero-mean Matrices

Definition (Sum of Conditionally Zero-Mean Matrices)

Given a sequence of Hermitian matrices $(\boldsymbol{Y}_k)_{k=1}^n$ satisfying the

Conditional zero mean property $\quad \mathbb{E}[\boldsymbol{Y}_k \,|\, (\boldsymbol{Y}_j)_{j \neq k}] = \mathbf{0}$

for all k, define the random sum $\boldsymbol{X} := \sum_{k=1}^n \boldsymbol{Y}_k$.

- **Note:** $(\boldsymbol{Y}_k)_{k \geq 1}$ is a martingale difference sequence

Examples

- Sums of independent centered random matrices
- Many sums of conditionally independent random matrices:

$$\boldsymbol{Y}_k \perp\!\!\!\perp (\boldsymbol{Y}_j)_{j \neq k} \,|\, Z \quad \text{and} \quad \mathbb{E}[\boldsymbol{Y}_k \,|\, Z] = \mathbf{0}$$

  - Rademacher series with random matrix coefficients

$$\boldsymbol{X} = \sum_k \varepsilon_k \boldsymbol{W}_k$$

  - $(\boldsymbol{W}_k)_{k \geq 1}$ Hermitian, $(\varepsilon_k)_{k \geq 1}$ independent Rademacher

Sums of Conditionally Zero-mean Matrices

Definition (Conditional Zero Mean Property)

$$\mathbb{E}[\boldsymbol{Y}_k \mid (\boldsymbol{Y}_j)_{j \neq k}] = \mathbf{0}$$

Matrix Stein Pair for $\boldsymbol{X} := \sum_{k=1}^n \boldsymbol{Y}_k$

- Let $\boldsymbol{Y}_k'$ and $\boldsymbol{Y}_k$ be conditionally i.i.d. given $(\boldsymbol{Y}_j)_{j \neq k}$
- Draw index K uniformly from $\{1, \ldots, n\}$
- Define $\quad \boldsymbol{X}' := \boldsymbol{X} + \boldsymbol{Y}_K' - \boldsymbol{Y}_K$
- Check Stein pair condition

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}' \mid (\boldsymbol{Y}_j)_{j \geq 1}] &= \mathbb{E}[\boldsymbol{Y}_K - \boldsymbol{Y}_K' \mid (\boldsymbol{Y}_j)_{j \geq 1}] \\
&= \frac{1}{n} \sum_{k=1}^n \left(\boldsymbol{Y}_k - \mathbb{E}[\boldsymbol{Y}_k' \mid (\boldsymbol{Y}_j)_{j \neq k}] \right) \\
&= \frac{1}{n} \sum_{k=1}^n \boldsymbol{Y}_k = \frac{1}{n} \boldsymbol{X}
\end{aligned}$$

Sums of Conditionally Zero-mean Matrices

Definition (Conditional Zero Mean Property)

$$\mathbb{E}[\boldsymbol{Y}_k \mid (\boldsymbol{Y}_j)_{j\neq k}] = \mathbf{0}$$

Conditional Variance for $\boldsymbol{X} := \boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}$

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{X}} &= \frac{n}{2}\cdot\mathbb{E}\left[(\boldsymbol{X}-\boldsymbol{X}')^2 \mid (\boldsymbol{Y}_j)_{j\geq 1}\right] \\ &= \frac{n}{2}\cdot\mathbb{E}\left[(\boldsymbol{Y}_K-\boldsymbol{Y}_K')^2 \mid (\boldsymbol{Y}_j)_{j\geq 1}\right] \\ &= \frac{1}{2}\sum_{k=1}^{n}\left(\boldsymbol{Y}_k^2 + \mathbb{E}[\boldsymbol{Y}_k^2 \mid (\boldsymbol{Y}_j)_{j\neq k}]\right).\end{aligned}$$

$\Rightarrow$ Conditional variance controlled when summands are bounded

$\Rightarrow$ Dependent analogues of concentration and moment inequalities

Combinatorial Sums of Matrices

Definition (Combinatorial Matrix Statistic)

Given a deterministic array $(\boldsymbol{A}_{jk})_{j,k=1}^{n}$ of Hermitian matrices and a uniformly random permutation π on $\{1, \ldots, n\}$, define the *combinatorial matrix statistic*

$$\boldsymbol{Y} := \sum\nolimits_{j=1}^{n} \boldsymbol{A}_{j\pi(j)} \quad \text{with mean} \quad \mathbb{E}\,\boldsymbol{Y} = \frac{1}{n} \sum\nolimits_{j,k=1}^{n} \boldsymbol{A}_{jk}.$$

- Generalizes the scalar statistics studied by Hoeffding (1951)

Example

- Sampling without replacement from $\{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n\}$

$$\boldsymbol{W} := \sum\nolimits_{j=1}^{s} \boldsymbol{B}_{\pi(j)}$$

Combinatorial Sums of Matrices

Definition (Combinatorial Matrix Statistic)

$$\boldsymbol{Y} := \sum\nolimits_{j=1}^{n} \boldsymbol{A}_{j\pi(j)} \quad \text{with mean} \quad \mathbb{E}\,\boldsymbol{Y} = \frac{1}{n}\sum\nolimits_{j,k=1}^{n} \boldsymbol{A}_{jk}.$$

Matrix Stein Pair for $\boldsymbol{X} := \boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}$

- Draw indices (J, K) uniformly from $\{1, \dots, n\}^2$
- Define $\quad \pi' := \pi \circ (J, K) \quad$ and $\quad \boldsymbol{X}' := \sum_{j=1}^{n} \boldsymbol{A}_{j\pi'(j)} - \mathbb{E}\,\boldsymbol{Y}$
- Check Stein pair condition

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}' \,|\, \pi] &= \mathbb{E}\left[\boldsymbol{A}_{J\pi(J)} + \boldsymbol{A}_{K\pi(K)} - \boldsymbol{A}_{J\pi(K)} - \boldsymbol{A}_{K\pi(J)} \,\middle|\, \pi\right] \\
&= \frac{1}{n^2}\sum\nolimits_{j,k=1}^{n} \boldsymbol{A}_{j\pi(j)} + \boldsymbol{A}_{k\pi(k)} - \boldsymbol{A}_{j\pi(k)} - \boldsymbol{A}_{k\pi(j)} \\
&= \frac{2}{n}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) = \frac{2}{n}\boldsymbol{X}
\end{aligned}$$

Combinatorial Sums of Matrices

Definition (Combinatorial Matrix Statistic)

$$\boldsymbol{Y} := \sum_{j=1}^{n} \boldsymbol{A}_{j\pi(j)} \quad \text{with mean} \quad \mathbb{E}\,\boldsymbol{Y} = \frac{1}{n}\sum_{j,k=1}^{n} \boldsymbol{A}_{jk}.$$

Conditional Variance for $\boldsymbol{X} := \boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}$

$$\begin{aligned}
\boldsymbol{\Delta}_{\boldsymbol{X}}(\pi) &= \frac{n}{4}\,\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{X}')^2 \,|\, \pi\right] \\
&= \frac{1}{4n}\sum_{j,k=1}^{n}\left[\boldsymbol{A}_{j\pi(j)} + \boldsymbol{A}_{k\pi(k)} - \boldsymbol{A}_{j\pi(k)} - \boldsymbol{A}_{k\pi(j)}\right]^2 \\
&\preccurlyeq \frac{1}{n}\sum_{j,k=1}^{n}\left[\boldsymbol{A}_{j\pi(j)}^2 + \boldsymbol{A}_{k\pi(k)}^2 + \boldsymbol{A}_{j\pi(k)}^2 + \boldsymbol{A}_{k\pi(j)}^2\right]
\end{aligned}$$

$\Rightarrow$ Conditional variance controlled when summands are bounded

$\Rightarrow$ Dependent analogues of concentration and moment inequalities

Extensions

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: $\lambda_{\max}(\mathscr{D}(\boldsymbol{B})) = \|\boldsymbol{B}\|$

Beyond Sums

- Matrix-valued functions satisfying a self-reproducing property
  - e.g., Matrix second-order Rademacher chaos: $\sum_{j,k} \varepsilon_j \varepsilon_k \boldsymbol{A}_{jk}$
  - Yields a dependent bounded differences inequality for matrices

Generalized Matrix Stein Pairs

- Satisfy $\mathbb{E}[g(\boldsymbol{X}) - g(\boldsymbol{X}') \,|\, Z] = \alpha \boldsymbol{X}$ almost surely for $g : \mathbb{R} \to \mathbb{R}$ weakly increasing.

References I

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

Bernstein, S. The theory of probabilities. *Gastehizdat Publishing House*, 1946.

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

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009.

Candès, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Info. Theory*, 2009. URL `arXiv:0903.1476`. To appear. Available at `arXiv:0903.1476`.

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

Chatterjee, S. *Concentration inequalities with exchangeable pairs*. PhD thesis, Stanford University, Palo Alto, Feb. 2008. URL `arxiv:math/0507526v1`.

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.optimization-online.org/DB_FILE/2011/01/2898.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.

Foygel, R. and Srebro, N. Concentration-based guarantees for low-rank matrix reconstruction. *Journal of Machine Learning Research - Proceedings Track*, 19:315–340, 2011.

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

Hoeffding, W. A combinatorial central limit theorem. *Ann. Math. Statist.*, 22:558–566, 1951.

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

References II

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 Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 1134–1142. 2011.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. Matrix concentration inequalities via the method of exchangeable pairs. URL `http://arxiv.org/abs/1201.6002`, 2012.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. `arXiv:1009.2118v2[cs.IT]`, 2010.

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. Simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

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.