Matrix Completion and Matrix Concentration

Lester Mackey[†]

Collaborators:

Ameet Talwalkar[‡], Michael I. Jordan[††],
Richard Y. Chen[*], Brendan Farrell[*], Joel A. Tropp[*], and Daniel Paulin[**]

[†]Stanford University [‡]UCLA [††]UC Berkeley
[*]California Institute of Technology [**]National University of Singapore

February 9, 2016

Part I

Divide-Factor-Combine

Motivation: Large-scale Matrix Completion

Goal: Estimate a matrix $\mathbf{L}_0 \in \mathbb{R}^{m\times n}$ given 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?
  - Netflix: 40 million users, 200K movies and television shows
- Ranking on the web: Is URL j relevant to user i?
  - Google News: millions of articles, 1 billion users
- Link prediction: Is user i friends with user j?
  - Facebook: 1.5 billion users

Motivation: Large-scale Matrix Completion

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

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

State of the art MC algorithms

- Strong estimation guarantees
- Plagued by expensive subroutines (e.g., truncated SVD)

This talk

- Present divide and conquer approaches for scaling up any MC algorithm while maintaining strong estimation guarantees

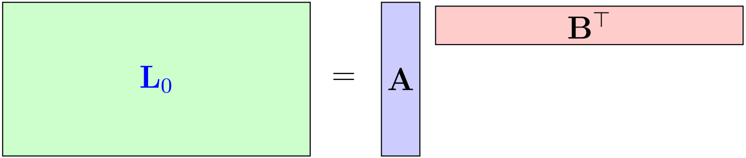
Exact Matrix Completion

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

Noisy Matrix Completion

Goal: Given entries from a matrix $\mathbf{M} = \mathbf{L}_0 + \mathbf{Z} \in \mathbb{R}^{m \times n}$ where $\mathbf{Z}$ is entrywise noise **and $\mathbf{L}_0$ has rank $\mathbf{r}$** $\ll m, n$, estimate $\mathbf{L}_0$

- Good news: $\mathbf{L}_0$ has $\sim (m+n)r \ll mn$ degrees of freedom



  - Factored form: $\mathbf{A}\mathbf{B}^\top$ for $\mathbf{A} \in \mathbb{R}^{m \times r}$ and $\mathbf{B} \in \mathbb{R}^{n \times r}$
- Bad news: Not all low-rank matrices can be recovered

Question: What can go wrong?

What can go wrong?

Entire column missing

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

- No hope of recovery!

Standard solution: Uniform observation model

Assume that the set of s observed entries Ω is drawn uniformly at random:

$$\Omega \sim \text{Unif}(m, n, s)$$

- Can be relaxed to non-uniform row and column sampling (Negahban and Wainwright, 2010)

What can go wrong?

Bad spread of information

$$\mathbf{L} = \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

- Can only recover $\mathbf{L}$ if $\mathbf{L}_{11}$ is observed

Standard solution: Incoherence with standard basis (Candès and Recht, 2009)

A matrix $\mathbf{L} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(\mathbf{L}) = r$ is *incoherent* if

Singular vectors are not too skewed: $\begin{cases} \max_i \|\mathbf{U}\mathbf{U}^\top \mathbf{e}_i\|^2 \leq \mu r/m \\ \max_i \|\mathbf{V}\mathbf{V}^\top \mathbf{e}_i\|^2 \leq \mu r/n \end{cases}$

and not too cross-correlated: $\|\mathbf{U}\mathbf{V}^\top\|_\infty \leq \sqrt{\frac{\mu r}{mn}}$

(In this literature, it's good to be incoherent)

How do we estimate $\mathbf{L}_0$?

First attempt:

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \operatorname{rank}(\mathbf{A}) \\ \text{subject to} & \sum_{(i,j)\in\Omega}(\mathbf{A}_{ij} - \mathbf{M}_{ij})^2 \leq \Delta^2. \end{array}$$

Problem: Computationally intractable!

Solution: Solve **convex** relaxation (Fazel, Hindi, and Boyd, 2001; Candès and Plan, 2010)

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \|\mathbf{A}\|_* \\ \text{subject to} & \sum_{(i,j)\in\Omega}(\mathbf{A}_{ij} - \mathbf{M}_{ij})^2 \leq \Delta^2 \end{array}$$

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

Questions:

- Will the nuclear norm heuristic successfully recover $\mathbf{L}_0$?
- Can nuclear norm minimization scale to large MC problems?

Noisy Nuclear Norm Heuristic: Does it work?

Yes, with high probability.

Typical Theorem

If $\mathbf{L}_0$ with rank r is incoherent, $s \gtrsim rn\log^2(n)$ entries of $\mathbf{M} \in \mathbb{R}^{m\times n}$ are observed uniformly at random, and $\hat{\mathbf{L}}$ solves the noisy nuclear norm heuristic, then

$$\|\hat{\mathbf{L}} - \mathbf{L}_0\|_F \leq f(m,n)\Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$.

- See Candès and Plan (2010); Mackey, Talwalkar, and Jordan (2011); Keshavan, Montanari, and Oh (2010); Negahban and Wainwright (2010)
- Implies **exact** recovery in the noiseless setting ($\Delta = 0$)

Noisy Nuclear Norm Heuristic: Does it scale?

Not quite...

- Standard interior point methods (Candès and Recht, 2009): $O(|\Omega|(m+n)^3 + |\Omega|^2(m+n)^2 + |\Omega|^3)$
- More efficient, tailored algorithms:
  - Singular Value Thresholding (SVT) (Cai, Candès, and Shen, 2010)
  - Augmented Lagrange Multiplier (ALM) (Lin, Chen, Wu, and Ma, 2009)
  - Accelerated Proximal Gradient (APG) (Toh and Yun, 2010)
  - All require rank-k truncated SVD on **every** iteration

Take away: These provably accurate MC algorithms are too expensive for large-scale or real-time matrix completion

Question: How can we scale up a given matrix completion algorithm and still retain estimation guarantees?

Divide-Factor-Combine (DFC)

Our Solution: Divide and conquer

1. Divide $\mathbf{M}$ into submatrices.
2. Complete each submatrix **in parallel**.
3. Combine submatrix estimates, using techniques from randomized low-rank approximation.

Advantages

- Completing a submatrix often much cheaper than completing $\mathbf{M}$
- Multiple submatrix completions can be carried out in parallel
- DFC works with **any** base MC algorithm
- The right choices of division and recombination yield estimation guarantees comparable to those of the base algorithm

DFC-Proj: Partition and Project

1. Randomly partition $\mathbf{M}$ into t column submatrices $\mathbf{M} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_t \end{bmatrix}$ where each $\mathbf{C}_i \in \mathbb{R}^{m \times l}$
2. Complete the submatrices **in parallel** to obtain
$$\begin{bmatrix} \hat{\mathbf{C}}_1 & \hat{\mathbf{C}}_2 & \cdots & \hat{\mathbf{C}}_t \end{bmatrix}$$
   - Reduced cost: Expect t-fold speed-up per iteration
   - Parallel computation: Pay cost of one cheaper MC
3. Project submatrices onto a single low-dimensional column space
   - Estimate column space of $\mathbf{L}_0$ with column space of $\hat{\mathbf{C}}_1$
$$\hat{\mathbf{L}}^{proj} = \hat{\mathbf{C}}_1 \hat{\mathbf{C}}_1^+ \begin{bmatrix} \hat{\mathbf{C}}_1 & \hat{\mathbf{C}}_2 & \cdots & \hat{\mathbf{C}}_t \end{bmatrix}$$
   - Common technique for randomized low-rank approximation (Frieze, Kannan, and Vempala, 1998)
   - Minimal cost: $O(mk^2 + lk^2)$ where $k = \operatorname{rank}(\hat{\mathbf{L}}^{proj})$
4. **Ensemble:** Project onto column space of each $\hat{\mathbf{C}}_j$ and average

DFC: Does it work?

Yes, with high probability.

Theorem (Mackey, Talwalkar, and Jordan, 2014b)

If $\mathbf{L}_0$ with rank r is incoherent and $s = \omega(r^2 n \log^2(n)/\epsilon^2)$ entries of $\mathbf{M} \in \mathbb{R}^{m \times n}$ are observed uniformly at random, then $l = o(n)$ random columns suffice to have

$$\|\hat{\mathbf{L}}^{proj} - \mathbf{L}_0\|_F \leq (2 + \epsilon) f(m, n) \Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$ and the noisy nuclear norm heuristic is used as a base algorithm.

- Can sample vanishingly small fraction of columns ($l/n \to 0$)
- Implies exact recovery for noiseless ($\Delta = 0$) setting
- Analysis streamlined by matrix Bernstein inequality

DFC: Does it work?

Yes, with high probability.

Proof Ideas:

1. If $\mathbf{L}_0$ is incoherent (has good spread of information), its partitioned submatrices are incoherent w.h.p.
2. Each submatrix has sufficiently many observed entries w.h.p.

$\Rightarrow$ Submatrix completion succeeds

3. Random submatrix captures the full column space of $\mathbf{L}_0$ w.h.p.
   - Analysis builds on randomized ℓ_2 regression work of Drineas, Mahoney, and Muthukrishnan (2008)

$\Rightarrow$ Column projection succeeds

DFC Noisy Recovery Error

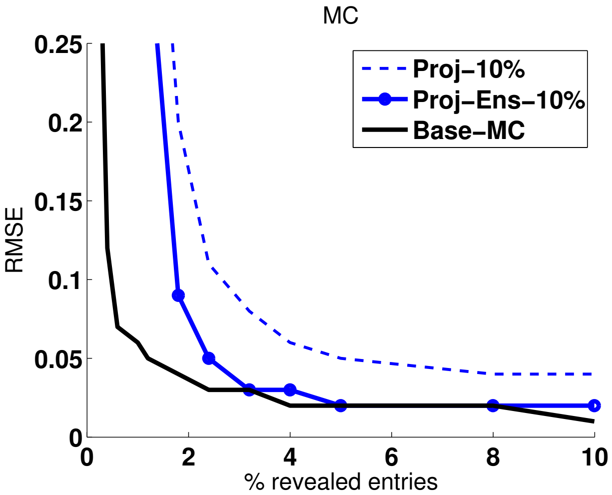


Figure : Recovery error of DFC relative to base algorithm (APG) with $m = 10K$ and $r = 10$.

DFC Speed-up

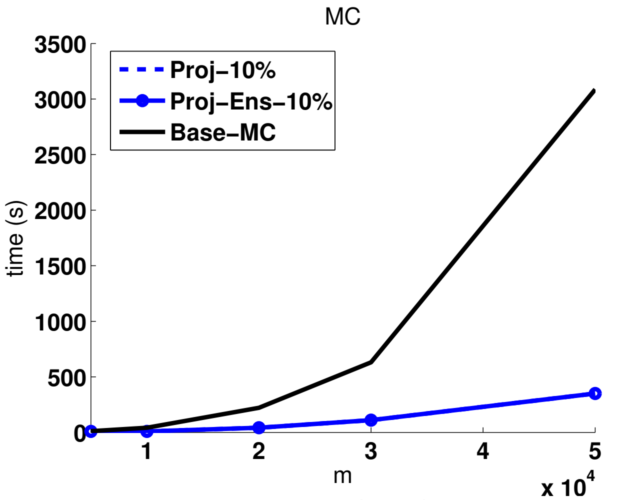


Figure : Speed-up over base algorithm (APG) for random matrices with $r = 0.001m$ and 4% of entries revealed.

Application: Collaborative filtering

Task: Given a sparsely observed matrix of user-item ratings, predict the unobserved ratings

Issues

- Full-rank rating matrix
- Noisy, non-uniform observations

The Data

- **Netflix Prize Dataset**[1]
  - 100 million ratings in $\{1, \ldots, 5\}$
  - 17,770 movies, 480,189 users

[1] http://www.netflixprize.com/

Application: Collaborative filtering

Task: Predict unobserved user-item ratings

Method	Netflix	
	RMSE	Time
Base method (APG)	0.8433	2653.1s
DFC-Proj-25%	0.8436	689.5s
DFC-Proj-10%	0.8484	289.7s
DFC-Proj-Ens-25%	0.8411	689.5s
DFC-Proj-Ens-10%	0.8433	289.7s

Future Directions

New Applications and Datasets

- Practical structured recovery problems with large-scale or real-time requirements
- Video background modeling via robust matrix factorization
  (Mackey, Talwalkar, and Jordan, 2014b)
- Image tagging / video event detection via subspace segmentation
  (Talwalkar, Mackey, Mu, Chang, and Jordan, 2013)

New Divide-and-Conquer Strategies

- Other ways to reduce computation while preserving accuracy
- More extensive use of ensembling

DFC-NYS: Generalized Nyström Decomposition

1. Choose a random column submatrix $\mathbf{C} \in \mathbb{R}^{m \times l}$ and a random row submatrix $\mathbf{R} \in \mathbb{R}^{d \times n}$ from $\mathbf{M}$. Call their intersection $\mathbf{W}$.

$$\mathbf{M} = \begin{bmatrix} \mathbf{W} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{M}_{21} \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} \mathbf{W} & \mathbf{M}_{12} \end{bmatrix}$$

2. Recover the low rank components of $\mathbf{C}$ and $\mathbf{R}$ **in parallel** to obtain $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$
3. Recover $\mathbf{L}_0$ from $\hat{\mathbf{C}}$, $\hat{\mathbf{R}}$, and their intersection $\hat{\mathbf{W}}$

$$\hat{\mathbf{L}}^{nys} = \hat{\mathbf{C}}\hat{\mathbf{W}}^{+}\hat{\mathbf{R}}$$

   - Generalized Nyström method (Goreinov, Tyrtyshnikov, and Zamarashkin, 1997)
   - Minimal cost: $O(mk^2 + lk^2 + dk^2)$ where $k = \mathrm{rank}(\hat{\mathbf{L}}^{nys})$
4. **Ensemble:** Run p times in parallel and average estimates

Future Directions

New Applications and Datasets

- Practical structured recovery problems with large-scale or real-time requirements

New Divide-and-Conquer Strategies

- Other ways to reduce computation while preserving accuracy
- More extensive use of ensembling

New Theory

- Analyze statistical implications of divide and conquer algorithms
  - Trade-off between statistical and computational efficiency
  - Impact of ensembling
- Developing suite of matrix concentration inequalities to aid in the analysis of randomized algorithms with matrix data

Part II

Stein's Method for Matrix Concentration

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, 2014b)
- Randomized matrix multiplication and factorization

  (Drineas, Mahoney, and Muthukrishnan, 2008; Hsu, Kakade, and Zhang, 2011)
- 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

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

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

Our work (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a; Paulin, Mackey, and Tropp, 2016)

- Stein's method of exchangeable pairs (1972), as advanced by Chatterjee (2007) for scalar concentration
  - $\Rightarrow$ Improved exponential tail inequalities (Hoeffding, Bernstein, Bounded differences)
  - $\Rightarrow$ 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 $\boldsymbol{B}$

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 $\mathbf{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ 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

Method of Exchangeable Pairs

Why Matrix Stein pairs?

- Furnish more convenient expressions for moments of $\boldsymbol{X}$

Lemma (Method of Exchangeable Pairs)

Let $(\boldsymbol{X}, \boldsymbol{X}')$ be a matrix Stein pair with scale factor α and $\boldsymbol{F} : \mathbb{H}^d \to \mathbb{H}^d$ a measurable function with $\mathbb{E}\|(\boldsymbol{X} - \boldsymbol{X}')\boldsymbol{F}(\boldsymbol{X})\| < \infty$. Then

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

Intuition

- Expressions like $\mathbb{E}\left[\boldsymbol{X}\mathrm{e}^{\theta \boldsymbol{X}}\right]$ and $\mathbb{E}[\boldsymbol{X}^p]$ arise naturally in concentration settings
- Eq. 1 allows us to bound these integrals using the smoothness properties of $\boldsymbol{F}$ and the discrepancy $\boldsymbol{X} - \boldsymbol{X}'$

The Conditional Variance

Why Matrix Stein pairs?

- Give rise to a measure of spread of the distribution of $\boldsymbol{X}$

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, $\mathbb{E}\,\boldsymbol{X}^2 = \frac{1}{2\alpha}\mathbb{E}[(\boldsymbol{X} - \boldsymbol{X}')^2] = \mathbb{E}\,\boldsymbol{\Delta}_{\boldsymbol{X}}$

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, 2014a)

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 $d = 1$, reduces to scalar result of Chatterjee (2007)
- The dimensional factor d cannot be removed

Matrix Hoeffding Inequality

Corollary (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

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 scale 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 scale 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 scalar Hoeffding inequality (1963)

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

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

  - Perfectly suited for rewriting using exchangeable pairs!

Exponential Concentration: Proof Sketch

2. Method of Exchangeable Pairs

- Rewrite the derivative of the trace mgf

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

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, 2014a)

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a weakly increasing function and that $h : \mathbb{R} \to \mathbb{R}$ is a function with convex derivative h'. 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

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

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

Exponential Concentration: Proof Sketch

3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

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

Exponential Concentration: Proof Sketch

3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

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

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

- 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

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

Polynomial Moments for Random Matrices

Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

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, 2014a)

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

- Combinatorial matrix statistics (e.g., sampling w/o replacement)
- Dependent bounded differences inequality for matrices

General Exchangeable Matrix Pairs (Paulin, Mackey, and Tropp, 2016)

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.

Cai, J. F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4), 2010.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9 (6):717–772, 2009.

Candès, E.J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, 2010.

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.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.

Fazel, M., Hindi, H., and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *In Proceedings of the 2001 American Control Conference*, pp. 4734–4739, 2001.

Frieze, A., Kannan, R., and Vempala, S. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Foundations of Computer Science*, 1998.

Goreinov, S. A., Tyrtyshnikov, E. E., and Zamarashkin, N. L. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1-3):1 – 21, 1997.

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

References II

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`, 2011.

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.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 99: 2057–2078, 2010.

Lin, Z., Chen, M., Wu, L., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report UILU-ENG-09-2215, 2009.

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. *The Annals of Probability*, 42(3):906–945, 2014a.

Mackey, L., Talwalkar, A., and Jordan, M. I. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 2014b. In press.

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`.

References III

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.

Paulin, D., Mackey, L., and Tropp, J. A. Efron-Stein Inequalities for Random Matrices. *The Annals of Probability*, to appear 2016.

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.

Talwalkar, Ameet, Mackey, Lester, Mu, Yadong, Chang, Shih-Fu, and Jordan, Michael I. Distributed low-rank subspace segmentation. December 2013.

Toh, K. and Yun, S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.

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