

# Matrix Completion and Matrix Concentration

Michael Jordan<sup>†</sup>

Collaborators: Lester Mackey<sup>†</sup>, Ameet Talwalkar<sup>†</sup>,  
Richard Y. Chen<sup>\*</sup>, Brendan Farrell<sup>\*</sup>, and Joel A. Tropp<sup>\*</sup>

<sup>†</sup>University of California, Berkeley

<sup>\*</sup>California Institute of Technology

February 21, 2012

# 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: 10 million users, 100K DVD titles
- Ranking on the web: Is URL  $j$  relevant to user  $i$ ?
  - Google News: millions of articles, millions of users
- Link prediction: Is user  $i$  friends with user  $j$ ?
  - Facebook: 500 million 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  $r \ll m, n$ , estimate  $\mathbf{L}_0$

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

The diagram illustrates the factored form of a low-rank matrix  $\mathbf{L}_0$ . On the left is a large light green rectangle labeled  $\mathbf{L}_0$ . To its right is an equals sign. Further right is a light blue vertical rectangle labeled  $\mathbf{A}$ . To the right of  $\mathbf{A}$  is a light red horizontal rectangle labeled  $\mathbf{B}^\top$ . This visualizes the equation  $\mathbf{L}_0 = \mathbf{A}\mathbf{B}^\top$ .

- 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!

## Solution: Uniform observation model

Assume that the set of  $s$  observed entries  $\Omega$  is drawn uniformly at random:

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

# What can go wrong?

## Bad spread of information

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

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

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

A matrix  $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \in \mathbb{R}^{m \times n}$  with  $\text{rank}(\mathbf{L}) = r$  is  $(\mu, r)$ -coherent if

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

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



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

First attempt:

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

**Problem:** Intractable to solve!

**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  $\mathbf{L}_0$  is  $(\mu, r)$ -coherent,  $s = O(\mu r n \log^2(n))$  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:** Provably accurate MC algorithms are still **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  $M$  into submatrices.
- 2 Factor each submatrix **in parallel**.
- 3 Combine submatrix estimates to estimate  $L_0$ .

## Advantages

- Factoring a submatrix is often much cheaper than factoring  $M$
- Multiple submatrix factorizations can be carried out in parallel
- DFC works with **any** base MC algorithm
- With the right choice of division and recombination, yields estimation guarantees comparable to those of the base algorithm

# DFC-PROJ: Partition and Project

- 1 Randomly partition  $\mathbf{M}$  into  $n/l$  column submatrices  $\mathbf{M} = [\mathbf{C}_1 \ \mathbf{C}_2 \ \cdots \ \mathbf{C}_{n/l}]$  where each  $\mathbf{C}_i \in \mathbb{R}^{m \times l}$

- 2 Complete the submatrices **in parallel** to obtain

$$[\hat{\mathbf{C}}_1 \ \hat{\mathbf{C}}_2 \ \cdots \ \hat{\mathbf{C}}_{n/l}]$$

- **Reduced cost:** Expect  $\min(n/l, m/d)$  speed-up per iteration
- **Parallel computation:** Pay cost of one cheaper MC

- 3 Recover a single factorization for  $\mathbf{M}$  by projecting each submatrix onto the column space of  $\hat{\mathbf{C}}_1$

$$\hat{\mathbf{L}}^{proj} = \hat{\mathbf{C}}_1 \hat{\mathbf{C}}_1^+ [\hat{\mathbf{C}}_1 \ \hat{\mathbf{C}}_2 \ \cdots \ \hat{\mathbf{C}}_{n/l}]$$

- **Minimal cost:**  $O(mk^2 + lk^2)$  where  $k = \text{rank}(\hat{\mathbf{L}}^{proj})$

- 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, 2011)

If  $\mathbf{L}_0$  is  $(\mu, r)$ -coherent and  $s$  entries of  $\mathbf{M} \in \mathbb{R}^{m \times n}$  are observed uniformly at random, then

$$l = O\left(\frac{\mu^2 r^2 n^2 \log^2(n)}{s \epsilon^2}\right)$$

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 ( $l/n \rightarrow 0$ ) whenever  $s = \omega(n \log^2(n))$
- Implies exact recovery for noiseless ( $\Delta = 0$ ) setting

# DFC: Does it work?

Yes, with high probability.

## Proof Ideas:

- ① Uniform column/row sampling yields **submatrices with low coherence** (high spread of information) w.h.p.
  - ② Each submatrix has **sufficiently many observed entries** w.h.p.
- ⇒ Submatrix completion succeeds
- ③ Uniform sampling of columns/rows **captures the full column/row space** of  $\mathbf{L}_0$  w.h.p.
    - Noisy analysis builds on randomized  $\ell_2$  regression work of Drineas, Mahoney, and Muthukrishnan (2008)
- ⇒ Column projection succeeds

# DFC Noisy Recovery Error

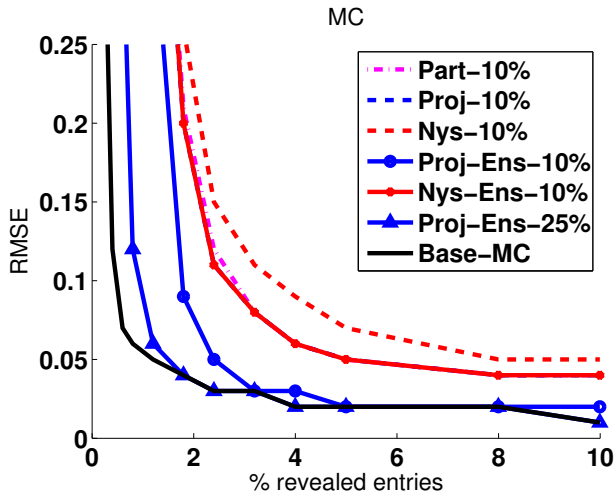


Figure: Recovery error of DFC relative to base algorithms with  $(m = 10K, r = 10)$ .



# DFC Speed-up

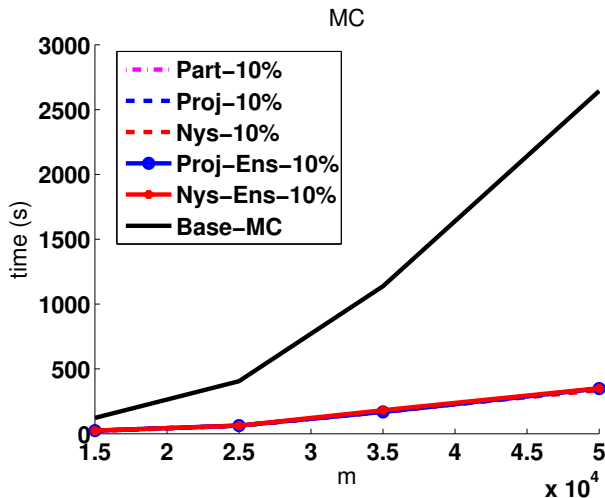


Figure: Speed-up over APG for random matrices with  $r = 0.001m$  and 4% of entries revealed.

# 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**<sup>1</sup>
  - 100 million ratings in  $\{1, \dots, 5\}$
  - 17,770 movies, 480,189 users

---

<sup>1</sup><http://www.netflixprize.com/>

# Application: Collaborative filtering

Method	Netflix	
	RMSE	Time
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

# Part II

## Stein's Method for Matrix Concentration Inequalities

# Concentration Inequalities

## Matrix concentration

$$\mathbb{P}\{\|\mathbf{X} - \mathbb{E} \mathbf{X}\| \geq t\} \leq \delta$$

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

- 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

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

**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
  - ⇒ Improved exponential tail inequalities (Hoeffding, Bernstein)
  - ⇒ Polynomial moment inequalities (Khintchine, Rosenthal)
  - ⇒ Dependent sums and more general matrix functionals

# Roadmap

- 3 Motivation
- 4 Stein's Method Background and Notation
- 5 Exponential Tail Inequalities
- 6 Polynomial Moment Inequalities
- 7 Extensions

# Notation

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

- *All matrices in this talk are Hermitian.*

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

**Trace:**  $\text{tr } \mathbf{B}$ , the sum of the diagonal entries of  $\mathbf{B}$

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

**Schatten  $p$ -norm:**  $\|\mathbf{B}\|_p := (\text{tr} |\mathbf{B}|^p)^{1/p}$  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  $\Psi : \mathcal{Z} \rightarrow \mathbb{H}^d$  be a measurable function. Define the random matrices

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

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

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

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

# The Conditional Variance

## Definition (Conditional Variance)

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

$$\Delta_{\mathbf{X}} := \Delta_{\mathbf{X}}(Z) := \frac{1}{2\alpha} \mathbb{E} [(\mathbf{X} - \mathbf{X}')^2 | Z].$$

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

# Exponential Concentration for Random Matrices

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

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

$$\Delta_{\mathbf{X}} \preceq c\mathbf{X} + v\mathbf{I} \quad \text{almost surely for } c, v \geq 0.$$

Then, for all  $t \geq 0$ ,

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

- Control over the conditional variance  $\Delta_{\mathbf{X}}$  yields
  - Gaussian tail for  $\lambda_{\max}(\mathbf{X})$  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  $(\mathbf{Y}_k)_{k \geq 1}$  be independent matrices in  $\mathbb{H}^d$  satisfying

$$\mathbb{E} \mathbf{Y}_k = \mathbf{0} \quad \text{and} \quad \mathbf{Y}_k^2 \preceq \mathbf{A}_k^2$$

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

$$\sigma^2 := \frac{1}{2} \left\| \sum_k (\mathbf{A}_k^2 + \mathbb{E} \mathbf{Y}_k^2) \right\|.$$

Then, for all  $t \geq 0$ ,

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

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

# Exponential Concentration: Proof Sketch

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

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

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

where  $m(\theta) := \mathbb{E} \operatorname{tr} e^{\theta \mathbf{X}}$

## 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

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

- Rewrite using exchangeable pairs

# Method of Exchangeable Pairs

## Lemma

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

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

Then

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

## Intuition

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

# Exponential Concentration: Proof Sketch

## 2. Method of Exchangeable Pairs

- Rewrite the derivative of the trace mgf

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

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

# Mean Value Trace Inequality

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

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

$$\begin{aligned} & \operatorname{tr}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (h(\mathbf{A}) - h(\mathbf{B}))] \leq \\ & \frac{1}{2} \operatorname{tr}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))]. \end{aligned}$$

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

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

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



# Exponential Concentration: Proof Sketch

## 3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

$$\begin{aligned} m'(\theta) &= \frac{1}{2\alpha} \mathbb{E} \operatorname{tr} [(\mathbf{X} - \mathbf{X}') (e^{\theta \mathbf{X}} - e^{\theta \mathbf{X}'})] \\ &\leq \frac{\theta}{4\alpha} \mathbb{E} \operatorname{tr} [(\mathbf{X} - \mathbf{X}')^2 \cdot (e^{\theta \mathbf{X}} + e^{\theta \mathbf{X}'})] \\ &= \theta \cdot \mathbb{E} \operatorname{tr} [\Delta_{\mathbf{X}} e^{\theta \mathbf{X}}]. \end{aligned}$$

## 4. Conditional Variance Bound: $\Delta_{\mathbf{X}} \preceq c\mathbf{X} + v\mathbf{I}$

- Yields differential inequality

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

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

# Polynomial Moments for Random Matrices

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

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

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

- **Moral:** The conditional variance controls the moments of  $\mathbf{X}$
- 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  $(\mathbf{A}_k)_{k \geq 1}$  be a deterministic sequence of Hermitian matrices. Then if  $p = 1$  or  $p \geq 1.5$ ,

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

- 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  $\sqrt{e}$  of optimal

# Extensions

## Refined Exponential Concentration

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

## General Complex Matrices

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

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

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

## 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

# References I

- Ahlsweide, 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.se.cuhk.edu.hk/~mancho/papers/cclmi\\_sta.pdf](http://www.se.cuhk.edu.hk/~mancho/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.
- Fazel, M., Hindi, H., and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, pp. 4734–4739, 2001.
- Goreinov, S. A., Tyrtshnikov, 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.
- 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.
- 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 *Advances in Neural Information Processing Systems 24*. 2011.
- 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.
- 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.

# References III

- 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.
- 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.