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 $L_0 \in \mathbb{R}^{m \times n}$ given a subset of its entries

\[
\begin{bmatrix}
? & ? & 1 & \ldots & 4 \\
3 & ? & ? & \ldots & ? \\
? & 5 & ? & \ldots & 5 \\
\end{bmatrix} \rightarrow \begin{bmatrix}
2 & 3 & 1 & \ldots & 4 \\
3 & 4 & 5 & \ldots & 1 \\
2 & 5 & 3 & \ldots & 5 \\
\end{bmatrix}
\]

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 $L_0 \in \mathbb{R}^{m \times n}$ given a subset of its entries

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
Goal: Estimate a matrix $L_0 \in \mathbb{R}^{m \times n}$ given a subset of its entries.
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

- 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

\[
\begin{bmatrix}
1 & 2 & ? & 3 & \ldots & 4 \\
3 & 5 & ? & 4 & \ldots & 1 \\
2 & 5 & ? & 2 & \ldots & 5 \\
\end{bmatrix}
\]

- No hope of recovery!

Standard solution: Uniform observation model

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

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

- Can be relaxed to non-uniform row and column sampling
  (Negahban and Wainwright, 2010)
What can go wrong?

**Bad spread of information**

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

- Can only recover \(L\) if \(L_{11}\) is observed

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

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

Singular vectors are not too skewed:

\[
\left\{ \begin{array}{l}
\max_i \|UU^\top e_i\|^2 \leq \frac{\mu r}{m} \\
\max_i \|VV^\top e_i\|^2 \leq \frac{\mu r}{n}
\end{array} \right.
\]

and not too cross-correlated:

\[
\|UV^\top\|_\infty \leq \sqrt{\frac{\mu r}{mn}}
\]

(In this literature, it’s good to be incoherent)
How do we estimate $L_0$?

First attempt:

$$\begin{align*}
\text{minimize}_A & \quad \text{rank}(A) \\
\text{subject to} & \quad \sum_{(i,j) \in \Omega} (A_{ij} - M_{ij})^2 \leq \Delta^2.
\end{align*}$$

Problem: Computationally intractable!

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

$$\begin{align*}
\text{minimize}_A & \quad \|A\|_* \\
\text{subject to} & \quad \sum_{(i,j) \in \Omega} (A_{ij} - M_{ij})^2 \leq \Delta^2
\end{align*}$$

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

Questions:

- Will the nuclear norm heuristic successfully recover $L_0$?
- Can nuclear norm minimization scale to large MC problems?
Yes, with high probability.

Typical Theorem

If $L_0$ with rank $r$ is incoherent, $s \gtrsim rn \log^2(n)$ entries of $M \in \mathbb{R}^{m \times n}$ are observed uniformly at random, and $\hat{L}$ solves the noisy nuclear norm heuristic, then

$$\|\hat{L} - L_0\|_F \leq f(m,n)\Delta$$

with high probability when $\|M - L_0\|_F \leq \Delta$.

- 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):
  \[ O(|Ω|(m + n)^3 + |Ω|^2(m + n)^2 + |Ω|^3) \]

- 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 $M$ 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 $M$
- 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 $\hat{\mathbf{C}}_1$
     
     $\hat{\mathbf{L}}^{proj} = \hat{\mathbf{C}}_1\hat{\mathbf{C}}_1^+[\hat{\mathbf{C}}_1 \hat{\mathbf{C}}_2 \cdots \hat{\mathbf{C}}_t]$}

   - Common technique for randomized low-rank approximation
     (Frieze, Kannan, and Vempala, 1998)

   - 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, 2014b)

If $L_0$ with rank $r$ is incoherent and $s = \omega(\frac{r^2 n \log^2(n)}{\epsilon^2})$ entries of $M \in \mathbb{R}^{m \times n}$ are observed uniformly at random, then $l = o(n)$ random columns suffice to have

$$\| \hat{L}^{\text{proj}} - L_0 \|_F \leq (2 + \epsilon) f(m, n) \Delta$$

with high probability when $\| M - L_0 \|_F \leq \Delta$ 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 $L_0$ 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 $L_0$ w.h.p.
   - Analysis builds on randomized $\ell_2$ regression work of Drineas, Mahoney, and Muthukrishnan (2008)

$\Rightarrow$ Column projection succeeds
**Figure**: Recovery error of DFC relative to base algorithm (APG) with $m = 10K$ and $r = 10$. 

Matrix Completion and Concentration  
February 9, 2016 16 / 43
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/
**Task:** Predict unobserved user-item ratings

<table>
<thead>
<tr>
<th>Method</th>
<th>Netflix</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>Time</td>
</tr>
<tr>
<td>Base method (APG)</td>
<td>0.8433</td>
<td>2653.1s</td>
</tr>
<tr>
<td>DFC-Proj-25%</td>
<td>0.8436</td>
<td>689.5s</td>
</tr>
<tr>
<td>DFC-Proj-10%</td>
<td>0.8484</td>
<td>289.7s</td>
</tr>
<tr>
<td>DFC-Proj-Ens-25%</td>
<td>0.8411</td>
<td>689.5s</td>
</tr>
<tr>
<td>DFC-Proj-Ens-10%</td>
<td>0.8433</td>
<td>289.7s</td>
</tr>
</tbody>
</table>
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
Choose a random column submatrix $C \in \mathbb{R}^{m \times l}$ and a random row submatrix $R \in \mathbb{R}^{d \times n}$ from $M$. Call their intersection $W$.

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

Recover the low rank components of $C$ and $R$ in parallel to obtain $\hat{C}$ and $\hat{R}$.

Recover $L_0$ from $\hat{C}$, $\hat{R}$, and their intersection $\hat{W}$

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

- Generalized Nyström method (Goreinov, Tyrtyshnikov, and Zamarashkin, 1997)
- **Minimal cost:** $O(mk^2 + lk^2 + dk^2)$ where $k = \text{rank}(\hat{L}^{\text{nys}})$

**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

\[ P\{\|X - \mathbb{E}X\| \geq t\} \leq \delta \]
\[ P\{\lambda_{\text{max}}(X - \mathbb{E}X) \geq t\} \leq \delta \]

- 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)
Motivation

Concentration Inequalities

Matrix concentration
\[ \mathbb{P}\{\lambda_{\max}(X - \mathbb{E}X) \geq t\} \leq \delta \]

Difficulty: Matrix multiplication is not commutative
\[ e^{X+Y} \neq e^X e^Y \neq e^Y e^X \]

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

4 Motivation

5 Stein’s Method Background and Notation

6 Exponential Tail Inequalities

7 Polynomial Moment Inequalities

8 Extensions
Notation

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

- *All matrices in this talk are Hermitian.*

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

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

**Spectral norm:** $\| B \|$, the maximum singular value of $B$
Matrix Stein Pair

Definition (Exchangeable Pair)

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

Definition (Matrix Stein Pair)

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

\[
X := \Psi(Z) \quad \text{and} \quad X' := \Psi(Z').
\]

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

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

- 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 $X$

Lemma (Method of Exchangeable Pairs)

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

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

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

Why Matrix Stein pairs?
- Give rise to a measure of spread of the distribution of $X$

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

$$\Delta_X := \Delta_X(Z) := \frac{1}{2\alpha} \mathbb{E} \left[ (X - X')^2 \mid Z \right].$$

- $\Delta_X$ is a stochastic estimate for the variance,
  $$\mathbb{E} X^2 = \frac{1}{2\alpha} \mathbb{E}[(X - X')^2] = \mathbb{E} \Delta_X$$

Take-home Message
Control over $\Delta_X$ yields control over $\lambda_{\max}(X)$
Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

Let \((X, X')\) be a matrix Stein pair with \(X \in \mathbb{H}^d\). Suppose that
\[
\Delta_X \preceq cX + v I
\]
almost surely for \(c, v \geq 0\).

Then, for all \(t \geq 0\),
\[
P\{\lambda_{\text{max}}(X) \geq t\} \leq d \cdot \exp\left\{\frac{-t^2}{2v + 2ct}\right\}.
\]

- Control over the conditional variance \(\Delta_X\) yields
  - Gaussian tail for \(\lambda_{\text{max}}(X)\) 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 \( X = \sum_k Y_k \) for independent matrices in \( \mathbb{H}^d \) satisfying

\[
\mathbb{E} Y_k = 0 \quad \text{and} \quad Y_k^2 \preceq A_k^2
\]

for deterministic matrices \((A_k)_{k \geq 1}\). Define the scale parameter

\[
\sigma^2 := \left\| \sum_k A_k^2 \right\|.
\]

Then, for all \( t \geq 0 \),

\[
P \left\{ \lambda_{\max} \left( \sum_k Y_k \right) \geq t \right\} \leq d \cdot e^{-t^2/2\sigma^2}.
\]

- Improves upon the matrix Hoeffding inequality of Tropp (2011)
  - Optimal constant 1/2 in the exponent
- Can replace scale parameter with \( \sigma^2 = \frac{1}{2} \left\| \sum_k \left( A_k^2 + \mathbb{E} Y_k^2 \right) \right\| \)
  - 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 $X$

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

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

**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
    $$m'(\theta) = \mathbb{E} \text{tr} X e^{\theta X} \quad \text{for} \ \theta \in \mathbb{R}.$$  
    
  - Perfectly suited for rewriting using exchangeable pairs!
2. Method of Exchangeable Pairs

- Rewrite the derivative of the trace mgf

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

**Goal:** Use the smoothness of \( F(X) = e^{\theta X} \) to bound the derivative
Mean Value Trace Inequality

**Lemma** (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

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

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

- **Standard matrix functions:** If \( g : \mathbb{R} \rightarrow \mathbb{R} \) and

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

- For exponential concentration we let \( g(A) = A \) and \( h(B) = e^{\theta B} \)
- Inequality does not hold without the trace
3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf

\[
m'(\theta) = \frac{1}{2\alpha} \mathbb{E} \text{tr} \left[ (X - X') (e^{\theta X} - e^{\theta X'}) \right] \\
\leq \frac{\theta}{4\alpha} \mathbb{E} \text{tr} \left[ (X - X')^2 \cdot (e^{\theta X} + e^{\theta X'}) \right] \\
= \frac{\theta}{2\alpha} \mathbb{E} \text{tr} \left[ (X - X')^2 \cdot e^{\theta X} \right] \\
= \theta \cdot \mathbb{E} \text{tr} \left[ \frac{1}{2\alpha} \mathbb{E} \left[ (X - X')^2 \mid Z \right] \cdot e^{\theta X} \right] \\
= \theta \cdot \mathbb{E} \text{tr} \left[ \Delta X e^{\theta X} \right].
\]
3. Mean Value Trace Inequality

- Bound the derivative of the trace mgf
  \[ m'(\theta) \leq \theta \cdot \mathbb{E} \text{tr} \left[ \Delta_X e^{\theta X} \right]. \]

4. Conditional Variance Bound: \( \Delta_X \preceq cX + vI \)

- Yields differential inequality
  \[ m'(\theta) \leq c\theta \mathbb{E} \text{tr} \left[ X e^{\theta X} \right] + v\theta \mathbb{E} \text{tr} \left[ e^{\theta X} \right] \]
  \[ = c\theta \cdot m'(\theta) + v\theta \cdot m(\theta). \]

- Solve to bound \( m(\theta) \) and thereby bound
  \[ \mathbb{P}\{\lambda_{\max}(X) \geq t\} \leq \inf_{\theta > 0} e^{-\theta t} \cdot m(\theta) \leq d \cdot \exp \left\{ \frac{-t^2}{2v + 2ct} \right\}. \]
Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

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

\[
(\mathbb{E} \|X\|_{2p}^{2p})^{1/2p} \leq \sqrt{2p - 1} \cdot (\mathbb{E} \|\Delta X\|_p^p)^{1/2p}.
\]

- **Moral:** The conditional variance controls the moments of \( X \)
- 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
Corollary (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a)

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

\[
\left( \mathbb{E} \left\| \sum_k \varepsilon_k A_k \right\|_{2p}^{2p} \right)^{1/2p} \leq \sqrt{2p - 1} \cdot \left\| \left( \sum_k A_k^2 \right)^{1/2} \right\|_{2p}.
\]

- 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 $X$
- Yields matrix generalization of classical Bernstein inequality
- Offers tool for unbounded random matrices

General Complex Matrices
- Map any matrix $B \in \mathbb{C}^{d_1 \times d_2}$ to a Hermitian matrix via dilation
  \[ \mathcal{D}(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \in \mathbb{H}^{d_1+d_2}. \]
- Preserves spectral information: $\lambda_{\text{max}}(\mathcal{D}(B)) = \|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


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