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

\[ \mathbb{P}\{|\|X - \mathbb{E}X\| \geq t\} \leq \delta \]

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

- 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, 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 $L_0 \in \mathbb{R}^{m \times n}$ from 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 $j$?
- 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 $L_0 \in \mathbb{R}^{m \times n}$ from a subset of its entries

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

**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
Suppose $\Omega$ is the set of observed entry locations.

**First attempt:**

$$\begin{align*}
\text{minimize}_A & \quad \text{rank } A \\
\text{subject to} & \quad A_{ij} = L_{0ij} \quad (i, j) \in \Omega
\end{align*}$$

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

**Solution:** Solve convex relaxation (?

$$\begin{align*}
\text{minimize}_A & \quad \|A\|_* \\
\text{subject to} & \quad A_{ij} = L_{0ij} \quad (i, j) \in \Omega
\end{align*}$$

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

How to Complete a Low-rank Matrix
Can Convex Optimization Recover $L_0$?

Yes, with high probability.

**Theorem** (Recht, 2011)

If $L_0 \in \mathbb{R}^{m\times n}$ has rank $r$ and $s \gtrsim \beta rn \log^2(n)$ entries are observed uniformly at random, then (under some technical conditions) convex optimization **recovers** $L_0$ **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

\[
  \sigma^2 := \sum_k \mathbb{E} Y_k^2.
\]

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

\[
  \mathbb{P} \left\{ \left| \sum_k Y_k \right| \geq t \right\} \leq 2 \cdot \exp \left\{ \frac{-t^2}{2\sigma^2 + 2Rt/3} \right\}
\]

- Gaussian decay controlled by variance when \( t \) is small
- Exponential decay controlled by uniform bound for large \( t \)
Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

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

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

Define the variance parameter

\[ \sigma^2 := \max \left( \left\| \sum_k \mathbb{E} Y_k Y_k^\top \right\|, \left\| \sum_k \mathbb{E} Y_k^\top Y_k \right\| \right). \]

Then, for all $t \geq 0$,

\[ \mathbb{P}\left\{ \left\| \sum_k Y_k \right\| \geq t \right\} \leq (m + n) \cdot \exp \left\{ \frac{-t^2}{3\sigma^2 + 2Rt} \right\} \]

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

### Matrix Bernstein Inequality

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

For all \( t \geq 0 \),

\[
P\left\{ \left\| \sum_k Y_k \right\| \geq t \right\} \leq (m + n) \cdot \exp\left\{ \frac{-t^2}{3\sigma^2 + 2Rt} \right\}
\]

### 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**

**Motivation**

**Matrix concentration**

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

**Difficulty:** Matrix multiplication is not commutative

\[ \Rightarrow e^{X+Y} \neq e^X e^Y \]

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

- All matrices in this talk are Hermitian.

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

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

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

- Matrix Stein pairs are exchangeable pairs
- Matrix Stein pairs always have zero mean
The Conditional Variance

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, \(\mathbb{E} X^2\)

Take-home Message

Control over \(\Delta_X\) yields control over \(\lambda_{\text{max}}(X)\)
Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

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

\[
\Delta_X \preceq cX + v I \quad \text{almost surely for } \ c, v \geq 0.
\]

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

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

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

$\mathbb{E} Y_k = 0$ and $Y_k^2 \preceq A_k^2$

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

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

Then, for all $t \geq 0$,

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

- Improves upon the matrix Hoeffding inequality of Tropp (2011)
  - Optimal constant $1/2$ in the exponent
- Can replace variance parameter with $\sigma^2 = \frac{1}{2} \left\| \sum_k (A_k^2 + \mathbb{E} Y_k^2) \right\|$  
  - Tighter than classical Hoeffding inequality (1963) when $d = 1$
Exponential Concentration: Proof Sketch

1. **Matrix Laplace transform method** (Ahlswede & Winter, 2002)
   - Relate tail probability to the *trace* of the mgf of $X$
     \[
     \mathbb{P}\{\lambda_{\text{max}}(X) \geq t\} \leq \inf_{\theta > 0} e^{-\theta t} \cdot m(\theta)
     \]
     where $m(\theta) := \mathbb{E} \text{tr} e^{\theta X}$
   - **Problem:** $e^{X+Y} \neq e^X e^Y$ when $X, Y \in \mathbb{H}^d$

**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
     \[
     m'(\theta) = \mathbb{E} \text{tr} X e^{\theta X} \quad \text{for } \theta \in \mathbb{R}.
     \]
   - Rewrite using exchangeable pairs
Lemma

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

\[
\mathbb{E}\| (X - X') F(X) \| < \infty.
\]

Then

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

Intuition

- Can characterize the distribution of a random matrix by integrating it against a class of test functions \(F\)
- Eq. 1 allows us to estimate this integral using the smoothness properties of \(F\) and the discrepancy \(X - X'\)
2. Method of Exchangeable Pairs

- Rewrite the derivative of the trace mgf

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

**Goal:** Use the smoothness of \( F(X) = e^{\theta X} \) to bound the derivative
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 \( A, B \in \mathbb{H}^d \), 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 \( g : \mathbb{R} \to \mathbb{R} \) and

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

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

Bound the derivative of the trace mgf

\[
m'(\theta) = \frac{1}{2\alpha} \mathbb{E} \text{tr} \left[ (X - X') (e^{\theta X} - e^{\theta X'}) \right]
\leq \frac{\theta}{4\alpha} \mathbb{E} \text{tr} \left[ (X - X')^2 \cdot (e^{\theta X} + e^{\theta X'}) \right]
= \frac{\theta}{2\alpha} \mathbb{E} \text{tr} \left[ (X - X')^2 \cdot e^{\theta X} \right]
= \theta \cdot \mathbb{E} \text{tr} \left[ \frac{1}{2\alpha} \mathbb{E} \left[ (X - X')^2 \mid Z \right] \cdot e^{\theta X} \right]
= \theta \cdot \mathbb{E} \text{tr} \left[ \Delta X e^{\theta X} \right].
\]
Exponential Concentration: Proof Sketch

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: \( \Delta_X \lesssim cX + vI \)

- 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

\[ \mathbb{P}\{\lambda_{\max}(X) \geq t\} \leq \inf_{\theta > 0} e^{-\theta t} \cdot m(\theta) \leq d \cdot \exp\left\{ \frac{-t^2}{2v + 2ct} \right\}. \]
Relaxing the constraint $\Delta X \preceq cX + v$

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

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

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

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

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

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

Matrix Bernstein Inequality

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

Let \((Y_k)_{k \geq 1}\) be independent matrices in \(\mathbb{H}^d\) satisfying

\[
E Y_k = 0 \quad \text{and} \quad \|Y_k\| \leq R \quad \text{for each index } k.
\]

Define the variance parameter

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

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

\[
P\left\{ \lambda_{\text{max}} \left( \sum_k Y_k \right) \geq t \right\} \leq d \cdot \exp\left\{ \frac{-t^2}{3\sigma^2 + 2Rt} \right\}
\]

- Gaussian tail controlled by improved variance when \(t\) is small
- **Key proof idea:** Apply refined concentration, and bound
  \(r(\psi) = \frac{1}{\psi} \log E \text{tr}(e^{\psi \Delta X} / d)\) using unrefined concentration
- Constants better than Oliveira (2009), worse than Tropp (2011)
Theorem (Mackey, Jordan, Chen, Farrell, and Tropp, 2012)

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

$$
(\mathbb{E} \text{tr} |X|^{2p})^{1/2p} \leq \sqrt{2p - 1} \cdot (\mathbb{E} \text{tr} \Delta^p_X)^{1/2p}.
$$

- **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, 2012)

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 \(p \geq 1.5\),

\[
\mathbb{E} \operatorname{tr}\left( \sum_k \varepsilon_k A_k \right)^{2p} \leq (2p - 1)^p \cdot \operatorname{tr}\left( \sum_k A_k^2 \right)^p.
\]

- 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 \(\sqrt{e}\) 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 \((Y_k)_{k=1}^n\) satisfying the Conditional zero mean property

\[ \mathbb{E}[Y_k \mid (Y_j)_{j \neq k}] = 0 \]

for all \(k\), define the random sum

\[ X := \sum_{k=1}^n Y_k. \]

Note: \((Y_k)_{k \geq 1}\) is a martingale difference sequence

Examples

- Sums of independent centered random matrices
- Many sums of conditionally independent random matrices:
  \[ Y_k \perp \perp (Y_j)_{j \neq k} \mid Z \quad \text{and} \quad \mathbb{E}[Y_k \mid Z] = 0 \]
  - Rademacher series with random matrix coefficients
    \[ X = \sum_k \varepsilon_k W_k \]
  - \((W_k)_{k \geq 1}\) Hermitian, \((\varepsilon_k)_{k \geq 1}\) independent Rademacher
Sums of Conditionally Zero-mean Matrices

Definition (Conditional Zero Mean Property)

\[ \mathbb{E}[Y_k \mid (Y_j)_{j \neq k}] = 0 \]

Matrix Stein Pair for \( X := \sum_{k=1}^{n} Y_k \)

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

\[
\mathbb{E}[X - X' \mid (Y_j)_{j \geq 1}] = \mathbb{E}[Y_K - Y'_K \mid (Y_j)_{j \geq 1}]
\]

\[
= \frac{1}{n} \sum_{k=1}^{n} (Y_k - \mathbb{E}[Y'_k \mid (Y_j)_{j \neq k}])
\]

\[
= \frac{1}{n} \sum_{k=1}^{n} Y_k = \frac{1}{n}X
\]
**Definition (Conditional Zero Mean Property)**

\[ \mathbb{E}[Y_k \mid (Y_j)_{j \neq k}] = 0 \]

**Conditional Variance for** \( X := Y - \mathbb{E} Y \)

\[ \Delta_X = \frac{n}{2} \cdot \mathbb{E} \left[ (X - X')^2 \mid (Y_j)_{j \geq 1} \right] \]

\[ = \frac{n}{2} \cdot \mathbb{E} \left[ (Y_K - Y'_K)^2 \mid (Y_j)_{j \geq 1} \right] \]

\[ = \frac{1}{2} \sum_{k=1}^{n} \left( Y_k^2 + \mathbb{E}[Y_k^2 \mid (Y_j)_{j \neq k}] \right). \]

⇒ Conditional variance controlled when summands are bounded

⇒ Dependent analogues of concentration and moment inequalities
Dependent Sequences  
Combinatorial Sums  

Combinatorial Sums of Matrices

**Definition (Combinatorial Matrix Statistic)**

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

$$Y := \sum_{j=1}^n A_{j\pi(j)}$$

with mean

$$\mathbb{E} Y = \frac{1}{n} \sum_{j,k=1}^n A_{jk}.$$  

- Generalizes the scalar statistics studied by Hoeffding (1951)

**Example**

- Sampling without replacement from $\{B_1, \ldots, B_n\}$

$$W := \sum_{j=1}^s B_{\pi(j)}$$
### Combinatorial Sums of Matrices

#### Definition (Combinatorial Matrix Statistic)

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

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

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

$$\mathbb{E}[X - X'|\pi] = \mathbb{E}\left[ A_{J\pi(J)} + A_{K\pi(K)} - A_{J\pi(K)} - A_{K\pi(J)} \bigg| \pi \right]$$

$$= \frac{1}{n^2} \sum_{j,k=1}^{n} A_{j\pi(j)} + A_{k\pi(k)} - A_{j\pi(k)} - A_{k\pi(j)}$$

$$= \frac{2}{n} (Y - \mathbb{E} Y) = \frac{2}{n} X$$
Combinatorial Sums of Matrices

Definition (Combinatorial Matrix Statistic)

\[ Y := \sum_{j=1}^{n} A_{j\pi(j)} \] with mean \[ \mathbb{E} Y = \frac{1}{n} \sum_{j,k=1}^{n} A_{jk}. \]

Conditional Variance for \( X := Y - \mathbb{E} Y \)

\[ \Delta_X(\pi) = \frac{n}{4} \mathbb{E} \left[ (X - X')^2 \mid \pi \right] \]

\[ = \frac{1}{4n} \sum_{j,k=1}^{n} \left[ A_{j\pi(j)} + A_{k\pi(k)} - A_{j\pi(k)} - A_{k\pi(j)} \right]^2 \]

\[ \approx \frac{1}{n} \sum_{j,k=1}^{n} \left[ A_{j\pi(j)}^2 + A_{k\pi(k)}^2 + A_{j\pi(k)}^2 + A_{k\pi(j)}^2 \right] \]

⇒ Conditional variance controlled when summands are bounded
⇒ Dependent analogues of concentration and moment inequalities
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: $\lambda_{\text{max}}(\mathcal{D}(B)) = \|B\|$ 

Beyond Sums

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

Generalized Matrix Stein Pairs

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


References II


