Constructive regularization of the random matrix norm.

Liza Rebrova

University of California Los Angeles

GeorgiaTech, March 2019

Non-asymptotic random matrix theory framework

$A = (A_{ij})_{n \times m}$. A_{ij} are taken from some distribution.
By definition,

$$\|A\| := \sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{u,v \in S^{n-1}} |\langle Au, v \rangle| = s_1(A)$$

$$\text{Norm of the inverse } 1/\|A^{-1}\| = \inf_{\|x\|_2=1} \|Ax\|_2 = s_n(A)$$

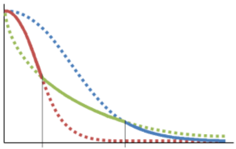
Singular values – real spectrum of the matrix

$$s(A) = \sqrt{eig(A^T A)}, \quad s_1 \geq s_2 \geq \ldots \geq s_n \geq 0.$$

What is optimal norm order?

Let $A = (A_{ij})_{n \times n}$ be a square random matrix with i.i.d. entries.

	Gaussian	Subgaussian
$s_1(A)$	for any $t \geq 0$ $s_1 \leq 2\sqrt{n} + t$ with prob $1 - 2e^{-t^2/2}$ from Gordon's theorem	for any $t \geq C_0$ $s_1 \leq t\sqrt{n}$ with prob $1 - e^{-ct^2 n}$ from Bernstein's inequality



Blue - gaussian, Red - subgaussian, Green - heavy-tailed

Def.: A_{ij} are subgaussian if $\mathbb{P}\{|A_{ij}| > t\} \leq C_1 e^{-c_2 t^2}$ for any $t > 0$

(Picture is taken from D.Mixon blog "Short, fat matrices")

Not an optimal order

Light tails ((sub)gaussian, 4 finite moments): with high probability,

$$\|A\| = s_{max}(A) \sim \sqrt{n} \quad \text{and} \quad s_{min}(A) \sim 1/\sqrt{n}.$$

Heavy tails (2 finite moments): with high probability,

$$\|A\| = s_{max}(A) \nsim \sqrt{n} \quad \text{and} \quad s_{min}(A) \sim 1/\sqrt{n}.$$

Example ($\|A\| \sim n \gg \sqrt{n}$)

- Litvak-Spector: Constructive example of $\|A\| \sim O(n^{1-\beta})$ for any $\beta \geq 0$ with probability at least $1/2$.
- Bai-Silverstein-Yin: 4 moments are needed for $\|A\| \to \sqrt{n}$.
- $$\sup_{\|x\|=1} \|Ax\| \geq \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} n^{-1/2} \\ n^{-1/2} \\ n^{-1/2} \\ n^{-1/2} \end{bmatrix} = n$$

Local norm regularization

Questions:

1. Can we regularize the norm correcting just a small fraction of the entries of A?
2. What in the structure of a heavy-tailed matrix causes norm to blow up from the "ideal" order $O(\sqrt{n})$?

Local regularization: $A \mapsto \bar{A}$, such that

- $\bar{A}$ differs from A in a small $\varepsilon n \times \varepsilon n$ sub-matrix
- $\|\bar{A}\| \lesssim \sqrt{n}$

Theorem (with R. Vershynin, informal statement)

Let A be a large enough random square matrix with i.i.d. elements. Local regularization is possible with high probability $\iff$

$$\mathbb{E}A_{ij} = 0 \text{ and } \mathbb{E}A_{ij}^2 \text{ is bounded.}$$

Local norm regularization

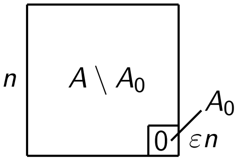
Theorem (Part 1: local obstructions)

Let $A = (A_{ij})_{n \times n}$ *has i.i.d. entries, such that* $\mathbb{E}A_{ij} = 0$, $\mathbb{E}A_{ij}^2 = 1$. *For any* $\varepsilon \in (0, 1/6]$,

with probability $\geq 1 - 11e^{-\varepsilon n/12}$

there exists an $\varepsilon n \times \varepsilon n$ *sub-matrix* $A_0 \subset A$:

$$\|A \setminus A_0\| \leq C_\varepsilon \sqrt{n}, \quad C_\varepsilon = C \cdot \frac{\ln(\varepsilon^{-1})}{\sqrt{\varepsilon}}$$



$A \setminus A_0$ = zero out all entries in A_0

- log-optimal dependence of on size ε
- can consider any $\varepsilon < 1$ in trade of larger constants
- inconstructive: does not identify A_0

Proof idea

"Ideal" norm relation?

$$\|A\| \lesssim \frac{\|A\|_{\infty\to 2}}{\sqrt{n}} \lesssim \|A\|_{2\to\infty} \lesssim \sqrt{n}$$

Definition

$\|A\|_{\infty\to 2} := \|A : l_\infty \to l_2\| = \max_{x\in\{-1,1\}^n} \|Ax\|_2$ (Cut norm)
$\|A\|_{2\to\infty}" = \|A : l_2 \to l_\infty\| = \max_i \|row(A)_i\|_2$ (Max row norm)

Example (True for gaussian matrices)

For gaussian matrix (i.i.d. N(0,1) entries) we have:

$$\|A\|_{2\to\infty} \sim \sqrt{n}, \quad \|A\|_{\infty\to 2} \sim n, \quad \|A\| \sim \sqrt{n}$$

Proof idea

"Ideal" norm relation?

$$\cancel{\|A\| \lesssim \frac{\|A\|_{\infty\to 2}}{\sqrt{n}} \lesssim \|A\|_{2\to\infty} \lesssim \sqrt{n}}$$

Not true for heavy-tailed :) Instead, we prove

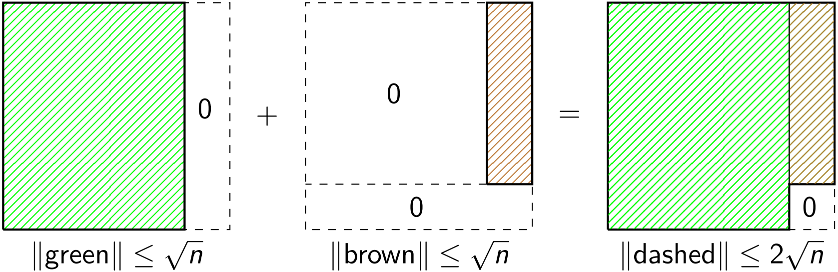
$$\|A_{J_3^c}\| \lesssim \frac{\|A_{J_2^c}\|_{\infty\to 2}}{\sqrt{n}} \lesssim \|A_{J_1^c}\|_{2\to\infty} \lesssim \sqrt{n},$$

where $J_1 \subset J_2 \subset J_3$ ($|J_i| \leq \varepsilon n$) are small column subsets to remove

- (*) εn column cut $\sim \varepsilon n \times \varepsilon n$ sub-matrix cut
- (**) last step: Grothendieck-Pietsch factorization for matrices (inconstructive!)

(*) εn columns cut is as good

It is enough to show that εn - columns cut regularizes the norm:



(**) Grothendieck-Pietsch factorization

Standard estimate: $\frac{1}{\sqrt{n}}\|B\|_{\infty\to 2} \leq \|B\| \leq \|B\|_{\infty\to 2}$

We want: $\|A_{J_3^c}\| \lesssim \frac{1}{\sqrt{n}}\|A_{J_2^c}\|_{\infty\to 2}$ with high probability

Theorem (Grothendieck-Pietsch, sub-matrix version)

Let B be a $n \times n_1$ real matrix and $\delta > 0$. Then there exists $J \subset [n_1]$ with $|J| \geq (1-\varepsilon)n_1$ such that

$$\|B_{[n]\times J}\| \leq \frac{2\|B\|_{\infty\to 2}}{\sqrt{\varepsilon n_1}}.$$

We apply it to $B = A_{J_2^c}$ with $n_1 = (1-\frac{\varepsilon}{2})n$ to find $|J| \geq (1-\varepsilon)n$.

Constructive regularization

Question:

1. Can we regularize the norm correcting just a small fraction of the entries of A?

Yes, iff $\mathbb{E}A_{ij} = 0$, $\mathbb{E}A_{ij}^2 = 1$.

2. What in the structure of a heavy-tailed matrix causes norm to blow up from the "ideal" order $O(\sqrt{n})$?

Or, how to perform local regularization constructively?

Individual entries correction

Theorem (more than 2 moments, any $\delta > 0$)

Let $A = (A_{ij})_{n \times n}$ has i.i.d. entries, s.t. $\mathbb{E} A_{ij} = 0$, $\mathbb{E}|A_{ij}|^{2+\delta} \le 1$. With high probability, zeroing $n^{1-\delta/9}$ largest entries of A leads to

$$\|\tilde{A}\| \le 8\sqrt{n}.$$

Proof based of Bandeira-Van Handel theorem: for any $\gamma > 1$

$$\|A\| \le \gamma \cdot \sigma + t \quad \text{with prob. } 1 - n \exp\left(-\frac{t^2}{c_\gamma \sigma_*^2}\right),$$

where

- σ is max expected row/col norm; $\sigma_{row}^2 = \max_i \|row(\mathbb{E} A_{ij}^2)\|_2^2$
- σ_* is max entry; $\sigma_*^2 = \max_{ij} \|A_{ij}\|_\infty$

We have $t, \sigma \sim \sqrt{n}$, and $\sigma_* \ll \sqrt{n}$.

If we have just finite 2*nd* moment...

Matrix Bernstein inequality: zeroing a few entries $\therefore \|\tilde{A}\| \lesssim \ln n\sqrt{n}$.

Example (failure of individual corrections approach)

Consider scaled Bernoulli matrix $A_{ij} \sim \sqrt{n} \cdot \mathrm{Ber}(\frac{1}{n})$.

- There is a row with at least $(\ln n / \ln \ln n)$ non-zero elements w.h.p. So, norm regularization is needed, as

$$\|A\| \geq \|A_i\|_2 \gg \sqrt{n}$$

- Entries are $\{0, \sqrt{n}\}$, so looking at the size only, we can only delete all or nothing.
- There are too many non-zero entries to fit in $\varepsilon n \times \varepsilon n$ sub-matrix

Need to use some information about entries locations with respect to each other (in given realization)

Constructive norm regularization

Theorem (Main)

Let $A = (A_{ij})_{n \times n}$ has i.i.d. symmetrically distributed entries, such that $\mathbb{E}A_{ij}^2 = 1$. For any $\varepsilon \in (0, 1/6]$ and $r > 1$,

$$\textit{with probability } \geq 1 - n^{0.1-r}$$

zeroing out εn rows and εn columns with the largest L_2-norms leads to the matrix $\tilde{A}$:

$$\|\tilde{A}\| \leq C\sqrt{c_\varepsilon \ln\ln n \cdot n}, \qquad \textit{where } c_\varepsilon = \ln(\varepsilon^{-1})/\varepsilon$$

- simple & constructive way to regularize the norm
- better description of the obstructions (to the good norm)
- extra $\ln\ln n$ term and symmetry assumprion

Constructive norm regularization

Theorem (Main, equivalent version)

Let $A = (A_{ij})_{n\times n}$ has i.i.d. symmetrically distributed entries, such that $\mathbb{E}A_{ij}^2 = 1$. For any $\varepsilon \in (0, 1/6]$ and $r > 1$,

$$\textit{with probability } \geq 1 - n^{0.1-r}$$

zeroing out any product subset of the entries such that on the rest all rows and columns have $\|row_i(A)\|_2, \|col_i(A)\|_2 \leq C\sqrt{c_\varepsilon n}$ produces $\tilde{A}$:

$$\|\tilde{A}\| \leq C\sqrt{c_\varepsilon \ln\ln n \cdot n}, \qquad \textit{where } c_\varepsilon = \ln(\varepsilon^{-1})/\varepsilon$$

- simple & constructive way to regularize the norm
- better description of the obstructions (to the good norm)
- extra $\ln\ln n$ term and symmetry assumprion

Proof background: Bernoulli matrices

B is $n \times n$ matrix with 0-1 entries, $\mathbb{E}B_{ij} = p$.

$$\mathbb{E}(B_{ij} - \mathbb{E}B_{ij})^2 \sim p \quad \therefore \text{ optimal norm } \|B - \mathbb{E}B\| \sim \sqrt{np}.$$

This is known:

- (Feige-Ofek) if $p \gtrsim \sqrt{\ln n}$, then $\|B - \mathbb{E}B\| \sim \sqrt{np}$ w.h.p.
- (Krievelevich-Sudakov) if $p \ll \sqrt{\ln n}$, No, counterexample

Note: $p = 1/n$ means exactly 2 finite (constant) moments
Regularization for the sparse case:

- (Feige-Ofek) zero out all rows and columns of A with more than $10d$ non-zero elements,
- (Le-Levina-Vershynin) reweight or zero out some elements s.t. sum of elements in every row and column is at most $10d$,

where $d := \mathbb{E}\{$ number of non-zero elements in a row/column $\}$.

Proof background: Bernoulli matrices

B is $n \times n$ matrix with 0-1 entries, $\mathbb{E}B_{ij} = p$.

$$\mathbb{E}(B_{ij} - \mathbb{E}B_{ij})^2 \sim p \quad \therefore \text{ optimal norm } \|B - \mathbb{E}B\| \sim \sqrt{np}.$$

This is known:

- (Feige-Ofek) if $p \gtrsim \sqrt{\ln n}$, then $\|B - \mathbb{E}B\| \sim \sqrt{np}$ w.h.p.
- (Krievelevich-Sudakov) if $p \ll \sqrt{\ln n}$, No, counterexample

Note: $p = 1/n$ means exactly 2 finite (constant) moments

Regularization for the sparse case:

- (Feige-Ofek) zero out all rows and columns of A with more than $10d$ non-zero elements,
- (Le-Levina-Vershynin) reweight or zero out some elements s.t. sum of elements in every row and column is at most $10d$,

where $d := \mathbb{E}\{L_2\text{-norm of a row/column }\}$.

Constructive regularization: proof ideas

High-level idea: split

$$|A_{ij}| \sim \sum_k 2^k \mathbb{1}_{\{|A_{ij}| \in (2^{k-1}, 2^k]\}} = \sum 2^k B^k$$

and apply Bernoulli results at each "level".

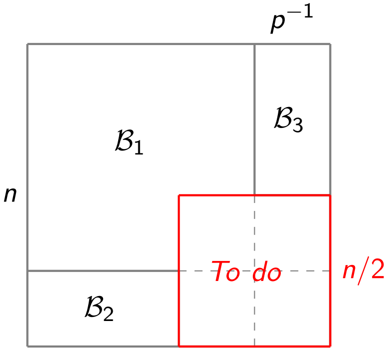
Some challenges:

1. Even though rows/columns of $\tilde{A}$ have bounded L_2-norms, some levels can be too heavy (compensated by other light)
2. Pass to absolute value (we cannot directly approximate $|||A|||$, it is too large - mean zero is needed for local regularization)
3. Consider as few levels as possible

1. Bernoulli matrix decomposition

With probability $1 - 3n^{-r}$ all entries of $B = \mathcal{B}_1 \sqcup \mathcal{B}_2 \sqcup \mathcal{B}_3$:

- $\#(row_i(\mathcal{B}_1)) \lesssim rnp$, $\#(col_i(\mathcal{B}_1)) \lesssim rnp$ - bounded rows&cols
- $\#(row_i(\mathcal{B}_2)) \lesssim r$ - very sparse rows
- $\#(col_i(\mathcal{B}_3)) \lesssim r$ - very sparse columns



Lemma (1, $pn > 4$)

In every $n2^{-k} \times n2^{-k}$ submatrix of B there are at most $2^{-k}/p$ columns with $> C_1 rnp$ non-zeros

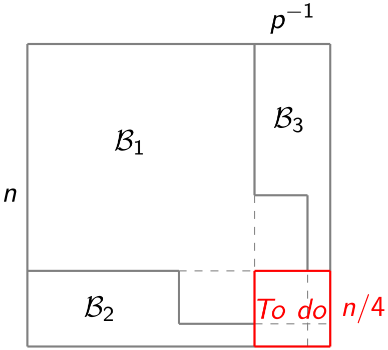
Lemma (2, $pn > 4$)

In every $n2^{-k} \times 2^{-k}/p$ submatrix of B there are at most $2^{-k}n/4$ columns with $> C_2 r$ non-zeros

1. Bernoulli matrix decomposition

With probability $1 - 3n^{-r}$ all entries of $B = \mathcal{B}_1 \sqcup \mathcal{B}_2 \sqcup \mathcal{B}_3$:

- $\#(row_i(\mathcal{B}_1)) \lesssim rnp$, $\#(col_i(\mathcal{B}_1)) \lesssim rnp$ - bounded rows&cols
- $\#(row_i(\mathcal{B}_2)) \lesssim r$ - very sparse rows
- $\#(col_i(\mathcal{B}_3)) \lesssim r$ - very sparse columns



Lemma (1, $pn > 4$)

In every $n2^{-k} \times n2^{-k}$ submatrix of B there are at most $2^{-k}/p$ columns with $> C_1 rnp$ non-zeros

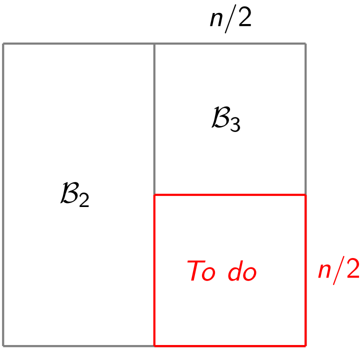
Lemma (2, $pn > 4$)

In every $n2^{-k} \times 2^{-k}/p$ submatrix of B there are at most $2^{-k}n/4$ columns with $> C_2 r$ non-zeros

1. Bernoulli matrix decomposition

With probability $1 - 3n^{-r}$ all entries of $B = \mathcal{B}_1 \sqcup \mathcal{B}_2 \sqcup \mathcal{B}_3$:

- $\#(row_i(\mathcal{B}_1)) \lesssim rnp$, $\#(col_i(\mathcal{B}_1)) \lesssim rnp$ - bounded rows&cols
- $\#(row_i(\mathcal{B}_2)) \lesssim r$ - very sparse rows
- $\#(col_i(\mathcal{B}_3)) \lesssim r$ - very sparse columns



Lemma (1, $pn \leq 4$)

In every $n2^{-k} \times n2^{-k}$ submatrix of B there are at most $n2^{-k-1}$ columns with $> C_1 r$ non-zeros

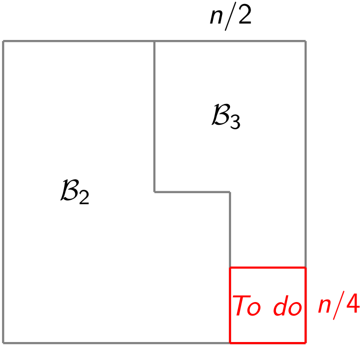
Lemma (2, $pn \leq 4$)

In every $n2^{-k} \times 2^{-k-1}$ submatrix of B there are at most $2^{-k-1}n$ columns with $> C_2 r$ non-zeros

1. Bernoulli matrix decomposition

With probability $1 - 3n^{-r}$ all entries of $B = \mathcal{B}_1 \sqcup \mathcal{B}_2 \sqcup \mathcal{B}_3$:

- $\#(row_i(\mathcal{B}_1)) \lesssim rnp$, $\#(col_i(\mathcal{B}_1)) \lesssim rnp$ - bounded rows&cols
- $\#(row_i(\mathcal{B}_2)) \lesssim r$ - very sparse rows
- $\#(col_i(\mathcal{B}_3)) \lesssim r$ - very sparse columns



Lemma (1, $pn \leq 4$)

In every $n2^{-k} \times n2^{-k}$ submatrix of B there are at most $n2^{-k-1}$ columns with $> C_1 r$ non-zeros

Lemma (2, $pn \leq 4$)

In every $n2^{-k} \times 2^{-k-1}$ submatrix of B there are at most $2^{-k-1}n$ columns with $> C_2 r$ non-zeros

1. Bernoulli matrices: after decomposition

Recall:

$$|A_{ij}| \sim \sum_k 2^k \mathbb{1}_{\{|A_{ij}| \in (2^{k-1}, 2^k]\}} = \sum 2^k B^k$$

- For $A_{part1} = \sum B^k_{\mathcal{B}_2} \cup B^k_{\mathcal{B}_3}$ use

Lemma (Norm of sparse matrices)

For any matrix Q and vectors $u, v \in S^{n-1}$, we have

$$\|Q\| \leq \underbrace{\max_j \|col_j(Q)\|_2}_{\sqrt{n}} \cdot \underbrace{\sqrt{\max_i \#(row_i(Q))}}_{\sqrt{const \cdot \#(terms)}}.$$

- For each $B^k_{ij} \in \mathcal{B}_1$ all rows and columns are bounded by $O(np_k) \implies$ we can use results for Bernoulli matrices

2. Heavy and light indices: Bernoulli

Using definition $\|B\| = \sup_{u,v \in S^{n-1}} |\sum_{ij} B_{ij} u_i v_j|$.
Light indices $:= \{(i,j) : |u_i v_j| \leq \sqrt{p/n}\}$ for every u, v.
Split the sum

$$|\sum_{ij}(B_{ij} - \mathbb{E}B_{ij})u_i v_j| \leq$$

$$|\sum_{light}(B_{ij} - \mathbb{E}B_{ij})u_i v_j| + |\sum_{heavy} \mathbb{E}B_{ij} u_i v_j| + |\sum_{heavy} B_{ij} u_i v_j|$$

- Light part - bounded members - Bernstein's concentration
- Expectation part - $\#(\text{heavy indices}) \leq n/p$ - Cauchy-Swartz
- Heavy part - Feige-Ofek theorem (bound follows from tail estimate for $e(S,T) =$ number of non-zero entries in some $S \times T$ sub-block)

2. Heavy and light indices: general case

Light indices $:= \{(i,j) : |u_i v_j A_{ij}| \le \sqrt{4/n}\}$ for every u, v.
Split the sum

$$|\sum_{ij} A_{ij} u_i v_j| \le |\sum_{light} A_{ij} u_i v_j| + \sum_{heavy} |A_{ij}| u_i v_j$$

$\mathbb{E}|A_{ij}| \ne 0$, but we do not care, split into Bernoulli levels and use Feige-Ofek theorem at each level!

$$\sum_{heavy} |A_{ij}| u_i v_j \le \sum_{ij} \sum_k 2^k B^k_{ij} u_i v_j \le \sum_k 2^k \sqrt{np_k}$$

$$\le \sqrt{n} \sum_k 2^{2k} p_k \cdot \sqrt{\#(\text{levels})}.$$

From second moment condition $1 \ge \mathbb{E}A^2_{ij} \ge 0.25 \sum_k 2^{2k} p_k$.
Number of levels is an extra term - minimize it.

3. Only average levels matter

- Large entries ($\gtrsim \sqrt{nc_\varepsilon}$) are zeroed (they produce heavy rows)
- Small entries ($\lesssim \sqrt{n/\ln n}$) are bounded separately by Bandeira-van Handel theorem

Number of levels is at most

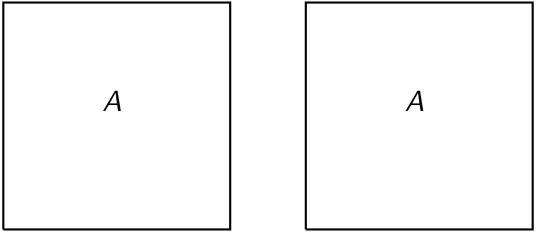
$$\log_2(Cc_\varepsilon n) - \log_2\left(\frac{cn}{\ln n}\right) \leq \log_2 \frac{Cc_\varepsilon n \cdot \ln n}{c_1 n} \sim \log\log n.$$

Note: symmetry is needed only to keep zero mean in various truncations.

Q.E.D.

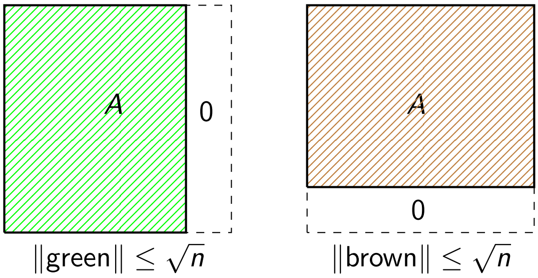
What is we want to zero out $\varepsilon n \times \varepsilon n$ block only?

Need to find the most "dense" part of the matrix.
Enough to find exceptional εn subset of columns (only), exceptional εn subset of rows (only) and take an intersection.



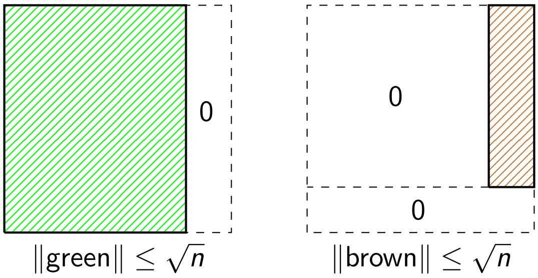
What is we want to zero out $\varepsilon n \times \varepsilon n$ block only?

Need to find the most "dense" part of the matrix.
Enough to find exceptional εn subset of columns (only), exceptional εn subset of rows (only) and take an intersection.



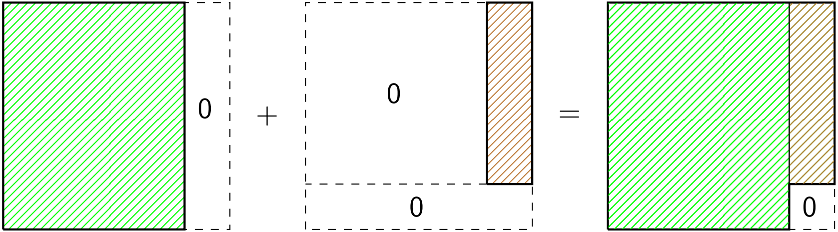
What is we want to zero out $\varepsilon n \times \varepsilon n$ block only?

Need to find the most "dense" part of the matrix.
Enough to find exceptional εn subset of columns (only),
exceptional εn subset of rows (only) and take an intersection.



What is we want to zero out $\varepsilon n \times \varepsilon n$ block only?

Need to find the most "dense" part of the matrix.
Enough to find exceptional εn subset of columns (only), exceptional εn subset of rows (only) and take an intersection.



$\|\text{green}\| \leq \sqrt{n}$ $\|\text{brown}\| \leq \sqrt{n}$ $\|\text{dashed}\| \leq 2\sqrt{n}$

Algorithm idea

Idea: find εn columns to replace with zeros, such that all rows and columns have bounded L_2-norms + apply Main Theorem.

Lemma (with K.Tikhomirov)

B is $n \times n$ matrix with 0-1 entries, $\mathbb{E}B_{ij} = p_k$. Then for any $L \geq 10$ with probability $1 - \exp(-n\exp(-Lp_k n))$ there are at most np_k columns to be deleted to achieve

$$\|row_i(\bar{A})\|_2^2, \|col_i(\bar{A})\|_2^2 \lesssim Ln$$

for every $i = 1, \ldots, n$.

This lemma will be applied for $p_k = 2^k\varepsilon/n$ for $k = 1, \ldots$.

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & & & & \\ & \delta_1 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & \delta_1 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1-st row: damping with the weight $0 < \delta_1 < 1$

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & & & & \\ & \delta_1 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & \delta_1 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2-nd row: all good

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & & & & \\ & \delta_1^2 & & & \\ & & \delta_1 & & \\ & & & 0 & \\ & & & & \delta_1 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_1 & \delta_1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

3-rd row: damping with the weight $0 < \delta_1 < 1$

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_2 & & & & \\ & \delta_1^2\delta_2 & & & \\ & & \delta_1 & & \\ & & & 0 & \\ & & & & \delta_1\delta_2 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_1 & \delta_1 & 0 & 0 \\ \delta_2 & \delta_2 & 0 & 0 & \delta_2 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4-th row: damping with the weight $0 < \delta_2 < \delta_1 < 1$

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ \color{red}{1} & \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} \end{bmatrix} \begin{bmatrix} \delta_2 & & & & \\ & \delta_1^2\delta_2 & & & \\ & & \delta_1 & & \\ & & & 0 & \\ & & & & \delta_1\delta_2 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_1 & \delta_1 & 0 & 0 \\ \delta_2 & \delta_2 & 0 & 0 & \delta_2 \\ \color{red}{1} & \color{red}{0} & \color{red}{0} & \color{red}{0} & \color{red}{0} \end{bmatrix}$$

5-th row: all good

Lemma is constructive:

Idea: we construct a diagonal matrix of weights that regularizes each row

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_2 & & & & \\ & \delta_1^2\delta_2 & & & \\ & & \delta_1 & & \\ & & & 0 & \\ & & & & \delta_1\delta_2 \end{bmatrix} = \begin{bmatrix} 0 & \delta_1 & 0 & 0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_1 & \delta_1 & 0 & 0 \\ \delta_2 & \delta_2 & 0 & 0 & \delta_2 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2-nd column has small weight: to be deleted

From Bernoulli to general matrices

Split

$$A_{ij}^2 \le \sum_k q_k \mathbb{1}_{\{A_{ij}^2 \in (q_{k-1}, q_k]\}}, \quad I_k := (q_{k-1}, q_k].$$

To pass from Bernoulli to general case now we need p_k to be in control: not too small (probability estimate), not too large (cardinality estimate). Indeed, for $p_k = P(\text{be inside a level } I_k)$, we want $p_k \sim 2^k$ for convegence.

One can take

Definition (2^{-k} quantiles)

$q_k := \min\{t : \mathbb{P}\{A_{ij}^2 > q_k\} = 2^{-k}\}$

However, the knowledge of quantiles (distribution) is undesirable.

Quantiles and order statistics

Note: quantiles q_k can be approximated by order statistics of A_{ij} (it is a free set of samples from the distribution!) So, the algorithm is distribution-oblivious.

Lemma

Let $A_{(1)} \geq A_{(2)} \geq \ldots \geq A_{(n^2)}$ be the order statistics of the elements A_{ij}. With probability at least $1 - 4\exp(-n^2 2^{-k-2}))$ for all $k = 1, \ldots, k_1$

$$q_{k-2} \leq A^2_{(\lceil n^2 2^{1-k} \rceil)} < q_k.$$

Proof idea: Chernoff's inequality
$\nu_1 :=$ { number of elements $A^2_{ij} > q$ }, then $\mathbb{E}\nu_1 = 2^{-k} n^{-2}$.

$$\mathbb{P}\{A^2_{(\lceil n^2 2^{1-k} \rceil)} \geq q_k\} = \mathbb{P}\{\nu_1 > 2^{1-k} n^2\} \text{ is small.}$$

Notations for the algorithm

Order statistics:

$$A_{(1)} \geq A_{(2)} \geq \ldots \geq A_{(n^2)}$$

Due to lemma, we can approximate "levels" with high probability as

$$\mathcal{A}_1 = A_{(n\varepsilon/2)} \ldots A_{(n\varepsilon)}$$

$$\mathcal{A}_2 = A_{(n\varepsilon+1)} \ldots A_{(2n\varepsilon)}$$

$$\vdots$$

Damping weights are defined as

$$W_{ij}^k := \begin{cases} 1 & \text{if } |\{i : A_{ij} \in \mathcal{A}_k\}| \leq C_\varepsilon p_k n; \\ \frac{C_\varepsilon p_k n}{|\{i : A_{ij} \in \mathcal{A}_k\}|} & \text{otherwise.} \end{cases}$$

$$V_j^k := \prod_{i=1}^{n} W_{ij}^k \leq 0.1.$$

Submatrix norm regularization algorithm

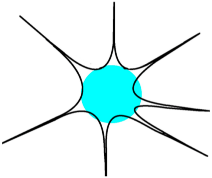
1. delete $n\varepsilon/2$ largest entries
2. small entries are fine without regularization
3. for each average k construct weights for $\mathcal{A}_k$: W_{ij}^k and V_j^k to find an exceptional subset of columns J_k:

$$|\cup_k J_k| \leq \varepsilon n/2 \text{ with high probability}$$

4. $J = \hat{J} \cup (\cup_k J_k)$, where $\hat{J}$ is a subset of $\varepsilon n/2$ columns with largest norms
5. repeat the process for A^T to find an exceptional row subset I
6. intersection of I and J gives a $\varepsilon n \times \varepsilon n$ exceptional matrix A_0

$$\implies \|\tilde{A}\| = \|A \setminus A_0\| \sim \sqrt{n \ln \ln n} \quad \text{by Main Theorem.}$$

THANKS FOR YOUR ATTENTION!



Feige-Ofek theorem

Part 1 - a good tail bound for all submatrices of Bernoulli matrices (number of edges in a sub-graph):

$$e(S,T) := \sum_{S\times T} B_{ij} \text{ for index subsets } S, T \subset [n]$$

Theorem (Part 1)

Let B be $n \times n$ $p-$Bernoulli matrix, $r \geq 1$. Suppose $\mathcal{B} \subset B$ such that all rows and columns have at most $C_0 np$ ones in $\mathcal{B}$. Then with probability at least $1 - n^{-r}$ one of the following holds:

(A) $e(S,T) \leq C|S||T|p$, *where* $e(S,T) := \sum_{S\times T\cap\mathcal{B}} B_{ij}$,

(B) $e(S,T)\log(\frac{e(S,T)}{|S||T|p}) \leq C_2|T|\log(\frac{n}{|T|})$

Feige-Ofek theorem

Part 2 - a non-random corolllary, based on partitioning coordinates u_i, v_j into 2^k-order "levels" and convergence of geometric series:

Theorem (Part 2)

Let B be a matrix with 0-1 entries, $p > 0$ and every row and column of B contains at most $C_0 np$ ones. If for all $S, T \subset [n]$ either (A) or (B) holds, then

$$\sum_{|u_i v_j| \geq \sqrt{p/n}} B_{ij} |u_i v_j| \leq C\sqrt{pn}.$$

Proof sketch for Part 1

- if $|T| \geq n/e$ (wlog $|T| \geq |S|$), then

$$e(S,T) \leq C_0 np|S| \leq C_0 ep|S||T|.$$

- otherwise, by Chernoff's inequality,

$$\mathbb{P}(e(S,T) > Kp|S||T|) \leq \exp\left(-\frac{K \ln Kp|S||T|}{3}\right).$$

Choose the smallest K that guarantees a good probability, including union bound over all S, T:

$$\exp\left(-\frac{K \ln Kp|S||T|}{3}\right)\binom{n}{|S|}\binom{n}{|T|} \leq \frac{1}{n^r}$$

Enough to take

$K \ln K \geq \dfrac{21|T|}{p|S||T|} \ln \dfrac{n}{|T|}$ to have $\dfrac{e(S,T)}{p|S||T|} \lesssim K$ with high prob.