Regularization of the random matrix norm: local and global obstructions

Liza Rebrova

joint with Roman Vershynin

University of Michigan

Center for Data Science, NYU, March, 2017

Global vs local obstructions

Setting: Object lacking some good property
Question: can we gain this property by a local change of an object?

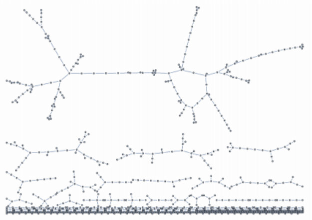
We can ask this for various structures/objects and various properties.

Example

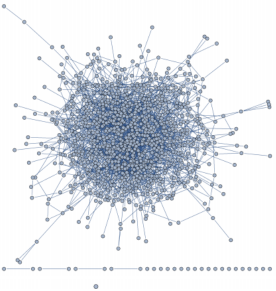
Object - Erdös-Rényi random graph $G(n, p)$;
property - connectivity; local change - in $o(n)$ edges
When $p \sim \frac{1}{n}$ structure properties change: a giant component appear:

1. before threshold - lots of connected components
2. after threshold - a giant component + a few other components

Global vs local obstructions - Example



$p < \frac{1}{n}$



$p > \frac{1}{n}$

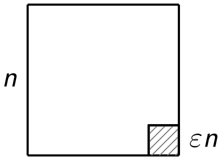
There are $O(n)$ small components, we cannot connect them all by $o(n)$ edges – obstructions to connectivity are "global"

A giant component and $log(n)$ other components, we can connect everything by a short cycle – obstructions are "local"

(Pictures are taken from A. Novozhilov's course in Mathematics of Networks, NDSU)

Key example - norm regularization problem

Object: $n \times n$ random matrix A with i.i.d. (independent identically distributed) entries
Property: $\|A\| \lesssim C\sqrt{n}$ w/high probability
Local change: in a small $\varepsilon n \times \varepsilon n$ submatrix



Notations:
"With high probability" – for all large matrices ($n > N_0$), property holds with probability 1 - o(1) (ideally, $1 - e^{-cn}$)

$$\|A\| := \sup_{\|x\|_2 = 1} \|Ax\|_2 \text{ – operator (spectral) norm}$$

It is equal to the maximum singular value of A

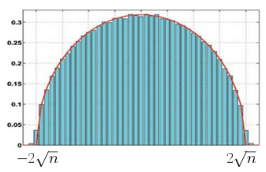
$$\|A\| = s_1(A) := \max_{\lambda} \sqrt{\lambda(A^T A)},$$

where $\lambda(X)$ denotes eigenvalue of X.

How large is $\|A\|$?

Example

- If $a_{ij} \sim N(0,1)$, then $\|A\| \simeq 2\sqrt{n}$ (Wigner semicircular law)
- If a_{ij} are subgaussian, then also $\|A\| \simeq C\sqrt{n}$ with probability $1 - e^{-cn}$ (Bernstein concentration inequality)



A random variable ξ is called subgaussian, if for any $t > 0$

$$\mathbb{P}\{|\xi| > t\} \leq C\exp(-ct^2)$$

Example

But if just $\mathbb{E}a_{ij}^2 = 1$, then there are examples $\|A\| \sim O(n^{2/\alpha})$ for any $\alpha \geq 2$ with probability at least $1/2$ (A.Litvak, S.Spector)

Norm regularization problem

Question: If $\|A\| \gg \sqrt{n}$ with substantial probability, is it a global or local obstruction?

Example

If A_{ij} are not mean zero: $\mathbb{E}A_{ij} \sim 1$, then $\|A\| \geq O(n)$, and the problem is global.

So, we assume $\mathbb{E}A_{ij} = 0$. Can we improve the norm of a random matrix by deletion of its small sub-matrix?

Theorem (L.R-R.Vershynin, informal statement)

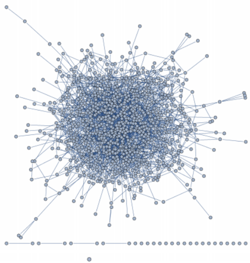
A is a random square matrix with i.i.d. centered elements A_{ij}.

- *if A_{ij} have finite variance $\therefore$ local obstructions*
- *if not $\therefore$ global obstructions*

Application to the random graphs

Consider an inhomogeneous Erdös-Rényi random graph $G(n, p_{ij})$ with *expected degrees* $np_{ij} \sim d$

$$A = \frac{1}{\sqrt{p}} \cdot \text{ Adjacency matrix}$$



$$p := \max p_{ij}$$
$$A_{ij} = \tfrac{1}{\sqrt{p}}\mathrm{Ber}(p), \text{ hence}$$
$$\mathbb{E}A_{ij}^2 = 1 \text{ and } \|\mathbb{E}A\| \sim \sqrt{p}n$$

Lemma

Dense graphs concentrate around their mean: if $d \geq \log n$, *then*

$$\|A - \mathbb{E}A\| \lesssim \sqrt{n},$$

while $\|\mathbb{E}A\| \geq \sqrt{n \log n}$

Lemma

Sparse graphs do not concentrate: if expected degree $d < \log n$, especially if $d \lesssim const$, then

$$\|A - \mathbb{E}A\| \gg \sqrt{n},$$

while $\|\mathbb{E}A\| \sim \sqrt{d}\sqrt{n}$.

Why do we care?

Spectral methods for, e.g. community detection problem, are based on idea:

- eigenstructure$(A) \sim$ eigenstructure$(\mathbb{E}A)$
- let's study the structure of $\mathbb{E}A$ instead

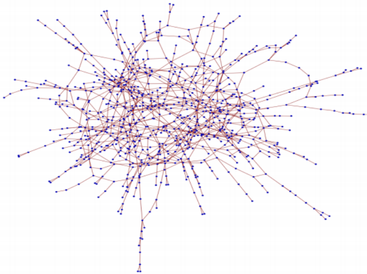
And it fails without concentration. Idea: preprocess our sparse graph to make it concentrate.

Application - random graphs

We want to change the graph, so that for new adjacency matrix

$$\|A' - \mathbb{E}A\| \lesssim \sqrt{n}.$$

1. When this change can be made on small fraction of vertices only? (local or global obstructions?)
2. What are the obstructions for such regularization?



Obstructions for random graphs

1. Is it local or global obstructions?

   Obstructions are local (known, Feige-Ofek)

2. What causes the obstructions (in terms of graph)?

   Idea: obstructions are in high-degree vertices.
   For the regularization it is enough to
   - U.Feige-E.Ofek: delete all high-degree vertices ($>10\cdot$expected degree)
   - C.Le-R.Vershynin: reweight or delete some of the edges adjacent to high-degree vertices (to make all the degrees bounded)
   - L.R-R.Vershynin (Bernoulli case corollary): delete a small $\varepsilon n \times \varepsilon n$ sub-graph

Finite $2 + \varepsilon$ moment

Theorem (L.R-R.Vershynin, informal statement)

A is a random square matrix with i.i.d. mean 0 elements A_{ij}.

- *if A_{ij} have finite 2nd moment $\therefore$ local obstructions*
- *if not $\therefore$ global obstructions*

Proposition (if we have more than 2nd moment)

Let A as before and $\mathbb{E}A_{ij} = 0$ and $\mathbb{E}A_{ij}^{2+\varepsilon} = 1$ for some $\varepsilon > 0$. Then with probability at least $1 - n^{-c}$ the norm of A can be regularized to the order $O(\sqrt{n})$ by correcting a few $o(n)$ individual entries.

This can be concluded from Bandeira-van Handel, or Seginer, or Auffinger results.

Plan of the proof:

- Let us zero out all the entries from the set

$$\mathcal{X} := \{A_{ij} : |A_{ij}| > c\frac{\sqrt{n}}{\sqrt{\log n}}\}$$

  The cardinality $|\mathcal{X}| \leq n^{1-\varepsilon/8}$ with probability at least $1 - e^{-n^{1-\varepsilon/8}}$ (Markov + Chernoff's inequalities).
- With probability at least $1 - \frac{1}{n}$ Euclidean norms of all the rows in $\bar{A} := A \setminus \mathcal{X}$ are at most $\sqrt{5n}$ (Bernstein's inequality)
- By Bandeira-van Handel's result:

$$\mathbb{P}\{\|\bar{A}\| \geq 3\sigma + t\} \leq n \exp(-t^2/C\sigma_*^2),$$

  where

$$\sigma_* := \max |\bar{A}_{ij}| \lesssim \frac{\sqrt{n}}{\sqrt{\log n}}; \quad \sigma := \max_i \sqrt{\sum_j A_{ij}^2} \leq \sqrt{5n}.$$

  Take $t = \sqrt{n}$ to see that $\|\bar{A}\| \lesssim \sqrt{n}$ with probability $1 - n^{-c}$.

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

...individual entries correction would not work for regularization!

Example

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

- There will be a row with at least $\log n / \log \log n$ non-zero elements. So, the norm is large:

$$\|A\| \geq \frac{\log n}{\log \log n} \sqrt{n} >> \sqrt{n}$$

- Entries are 0-1, so looking at them individually, we can only delete all non-zeros (but there are $O(n)$ non-zero entries)
- Or use some information about their locations with respect to each other (in given realization), such as heavy rows/columns etc. And this works!

Theorem (Local obstructions)

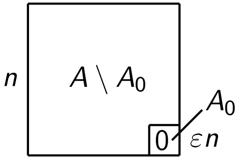
Let A be a random $n \times n$ matrix with i.i.d. elements, $\mathbb{E}A_{ij} = 0$, $\mathbb{E}A_{ij}^2 = 1$. Then for any $\varepsilon \in (0, \frac{1}{2}]$ with probability

$$1 - 11\exp(-\varepsilon n/6)$$

there exists an $\varepsilon n \times \varepsilon n$ sub-matrix $A_0 \subset A$, such that

$$\|A \setminus A_0\| \le C_\varepsilon \sqrt{n}$$

Here, $A \setminus A_0$ is a matrix we obtain by zeroing out all elements of A, that belong to sub-matrix A_0:



Dependence on ε

Optimal dependence would be $C_\varepsilon = O(\frac{1}{\sqrt{\varepsilon}})$.
To see this, consider

- $\varepsilon \lesssim \frac{1}{n}$ and $\|A\| = O(n) = O(\sqrt{n}/\sqrt{\varepsilon})$ with probability 1/2, or
- any $\varepsilon \leq 1/2$ and Bernoulli matrix A with rare $\frac{\sqrt{n}}{\sqrt{\varepsilon}}$ spikes

Our argument gives log-optimal dependence $C_\varepsilon = O(\frac{\ln(\varepsilon^{-1})}{\sqrt{\varepsilon}})$

Example (Bernoulli case)

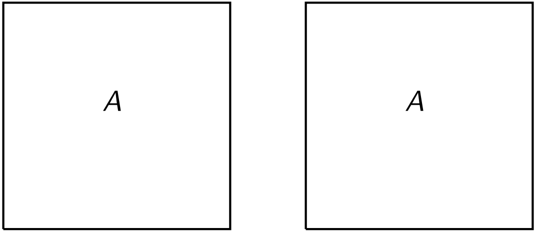
However, in "good Bernoulli case" dependence is better. Let A be a square matrix with i.i.d. elements distributed like

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{p}} & \text{with probability } p \\ 0 & \text{otherwise} \end{cases}, \quad p \cdot n = d = O(1) \geq 4$$

Then the theorem is hold with $C_\varepsilon = O(\ln(\varepsilon^{-1}))$.

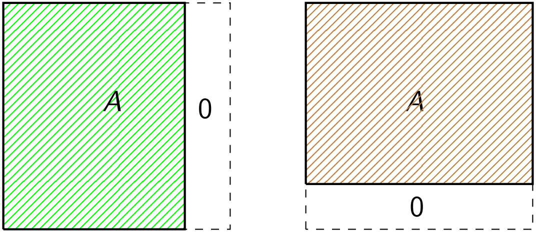
Observation 0: εn columns cut

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



Observation 0: εn columns cut

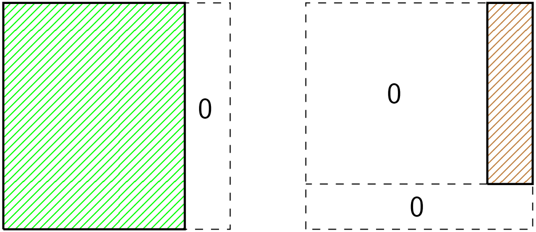
It is enough to show that εn -columns cut regularizes the norm:



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

Observation 0: εn columns cut

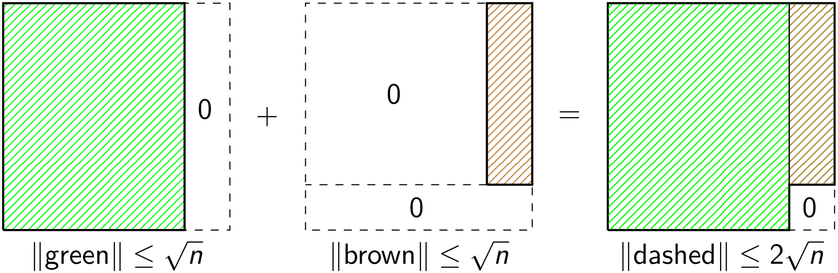
It is enough to show that εn -columns cut regularizes the norm:



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

Observation 0: εn columns cut

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



Three norms

Definition

- Operator norm

$$\|A\| = \|A : l_2 \to l_2\| = \sup_{\|x\|_2=1} \|Ax\|_2$$

- Infinity to 2 (cut norm)

$$\|A\|_{\infty\to 2} = \|A : l_\infty \to l_2\| = \max_{x\in\{-1,1\}^n} \|Ax\|_2$$

- 2 to infinity (maximum row norm)

$$\|A\|_{2\to\infty} = \|A : l_2 \to l_\infty\| = \max_i \|A_i\|_2,$$

where A_i, $i = 1, \ldots, n$ denote rows of matrix A.

Example

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

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

"Ideal" norm relation?

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

Example

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

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

"Ideal" norm relation?

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

Not true :) Instead,

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

where J_1, J_2, J_3 are small subsets of columns that we zero out ($J_1 \subset J_2 \subset J_3$ with cardinalities $|J_i| \leq \varepsilon n$)

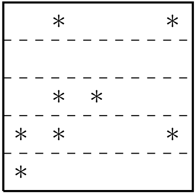
The $2 \to \infty$ norm: damping

Lemma

Consider an $n \times n$ random matrix A with i.i.d. entries A_{ij} which have mean zero, unit variance and $|A_{ij}| \le \frac{\sqrt{n}}{2}$ a.s. Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - e^{-\varepsilon n}$, there exists a subset $J_1 \in [n]$ with cardinality $|J_1| \le \varepsilon n$ such that

$$\|A_{J_1^c}\|_{2\to\infty} \le C\sqrt{\ln \varepsilon^{-1}} \cdot \sqrt{n}.$$

Warning: we cannot just cut columns with large elements!



Damping: Bernoulli example

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 $0 < \delta_1 < 1$

Damping: Bernoulli example

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

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

2-nd row: all good

Damping: Bernoulli example

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 $0 < \delta_1 < 1$

Damping: Bernoulli example

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 $0 < \delta_2 < \delta_1 < 1$

Damping: Bernoulli example

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

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

5-th row: all good

Damping: Bernoulli example

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

Damping: Bernoulli example

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

Proposition (L.R-K.Tikhomirov)

Let $\varepsilon \in (0,1]$ and A is our martix. Then with high probability there exists a diagonal weight matrix$D = (d_{ii})_{i=1}^n$, $d_i \in (0,1)$, such that

(1) $\|AD\|_{2\to\infty} \le C\sqrt{\ln \varepsilon^{-1}}\sqrt{n}$

(2) $\mathbb{E}(d_{11} \cdot d_{22} \cdot \ldots \cdot d_{nn}) \ge \exp(-\varepsilon n)$

- Condition (2) implies that there all but εn columns have weights d_{ii}'s such that: $d_{ii} > e^{-2}$. We can cut the rest!

Damping for each row

So, enough to show that for every row A_i exists $D^i = (d^i_1, \dots, d^i_n)$:

- $\sum_j d^i_{jj} \cdot A^2_{ij} \le C_\varepsilon n$
- $\mathbb{E}(d^i_{11} \cdot d^i_{22} \cdot \ldots \cdot d^i_{nn}) \ge e^{-\varepsilon}$

For Bernoulli matrix:

$$\begin{cases} d^i_{jj} := 0, & \text{if } A_{ij} = 0 \\ d^i_{jj} := 1, & \text{if } \|A^2_i\|_1 \le Cn \\ d^i_{jj} := \frac{Cn}{\|A^2_i\|_1}, & \text{otherwise} \end{cases}$$

where $A^2_i := (A^2_{i1}, \dots, A^2_{in})$.

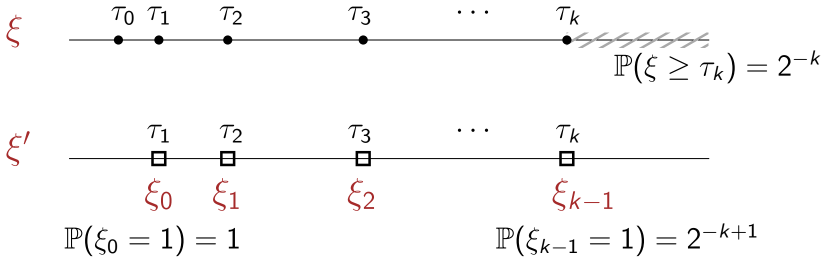
For general case:

Naive regularization ($d^i_{jj} := \frac{\text{expected norm}}{\text{real norm}}$) would not work

Damping: general disctibution case

Main idea: any random variable ξ (for us $\xi = A_{ij}^2$) can be almost surely approximated above by the sum of Bernoulli random variables ξ_i, such that $\mathbb{P}(\xi_i = 1) = 2^{-i}$,

$$\xi' := \sum_{i=}^{\infty} \tau_i \xi_i \geq \xi, \quad \text{and} \quad \mathbb{E}\xi' \leq 2\mathbb{E}\xi.$$



Step 2: $\|.\|_{\infty\to 2}$ norm - playing with the signs

Reminder: we are proving

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

Lemma

Let A be an $n \times n$ random matrix whose entries are independent, symmetric random variables. Then

$$\|A\|_{\infty\to 2} \le C\sqrt{n}\|A\|_{2\to\infty}$$

with probability at least $1 - e^{-n}$.

Rough idea: condition on $|A_{ij}|$, and consider linear combination of Rademacher random variables ($\gamma := \pm 1$ with probability $1/2$)

Corollary

If $J_1 \subset [n]$ be a random subset, which is independent of the signs of the entries of A, then with the same high probability

$$\|A_{J_1^c}\|_{\infty\to 2} \leq C\sqrt{n}\|A_{J_1^c}\|_{2\to\infty}$$

So, $J_1 = J_2$, there are no loss on Step 2.

Removing symmetry assumption:

- Note that for Lemma basic anti-symmetrization inequality will do (norm is a convex function)

$$\mathbb{E}\varphi(\|\sum_i X_i\|) \leq \mathbb{E}\varphi(2\|\sum_i \gamma_i X_i\|) \quad \text{(from Ledoux-Talagrand)}$$

- However, for Corollary (columns deletion makes it non-convex) more delicate argument is needed.

Proof sketch

Lemma

Let A be an $n \times n$ random matrix whose entries are independent, symmetric random variables. Then

$$\|A\|_{\infty\to 2} \le C\sqrt{n}\|A\|_{2\to\infty}$$

with probability at least $1 - e^{-n}$.

We want to show:

$$\max_{\{-1,1\}^n} \|Ax\|_2^2 \le Cn \max_{rows} \|A_i\|_2^2 \quad \text{w/high probability}$$

Enough to show: for each $x \in \max_{\{-1,1\}^n}$

$$\|Ax\|_2^2 \le Cn \max_{rows} \|A_i\|_2^2 \quad + \text{union bound}$$

$\|Ax\|_2^2 \leq Cn \max \|A_i\|_2^2$ - ?

Left hand side $\|Ax\|_2^2 = \sum \xi_i^2$, where

$$\xi_i = \langle A_i, x \rangle = \sum_j A_{ij} x_j = \sum_j A_{ij} \gamma_{ij} x_j = \sum_j A_{ij} \gamma_{ij}.$$

Linear combination of ± 1 symmetric random variables γ_{ij} - they are subgaussian. Bernstein for subgaussians: ξ_i is also subgaussian with $\|\xi_i\|_{\psi_2}^2 = \sum_j A_{ij}^2 = \|A_i\|_2^2$.

$$\xi_i - \text{subgaussian} \quad \therefore \xi_i^2 - \text{subexponential}$$

Concentration for sum of subexponentials:

$$\|Ax\|_2^2 = \sum \xi_i^2 \leq C \cdot n \|\xi_i\|_{\psi_2} \leq Cn \|A_i\|_2^2.$$

Done!

Step 3: $\|.\|$ 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 use it with $n_1 = (1-\varepsilon)n$ to find $|J| \geq (1-2\varepsilon)n$, such that

$$\|A_{[n]\times J}\| \leq \frac{2\|A \setminus A'\|_{\infty\to 2}}{\sqrt{\varepsilon n}} \leq \frac{C_\varepsilon n}{\sqrt{\varepsilon n}} \leq \frac{C_\varepsilon}{\sqrt{\varepsilon}}\sqrt{n}.$$

Fighting for a good C_ε

Solution (for bounded entries): consider only "small" entries of the matrix $|a_{ij}| \lesssim \sqrt{n}$, then on Step 1 $\|A \setminus A'\|_{2\to\infty} \leq C\sqrt{\ln(\varepsilon^{-1})}n$.

Hence, for a matrix A such that $\mathbb{E}A_{ij} = 0, \mathbb{E}{A_{ij}}^2 \leq 1$, $|a_{ij}| \leq \frac{\sqrt{n}}{2}$ a.s.:

$$\|A \setminus A'\| \leq C\frac{\sqrt{\ln(\varepsilon^{-1})}}{\sqrt{\varepsilon}}\sqrt{n}.$$

General case:

$$A = A \cdot \mathbb{1}_{\{|A_{ij}| \lesssim \sqrt{n}\}} + \underbrace{A \cdot \mathbb{1}_{\{\sqrt{n} \lesssim |A_{ij}| \lesssim \frac{\sqrt{n}}{\sqrt{\varepsilon}}\}}}_{\substack{\downarrow \\ \text{sparsity and size} \\ \text{(most non-zero elements} \\ \text{belong to sparse rows)}}} + \underbrace{A \cdot \mathbb{1}_{\{\frac{\sqrt{n}}{\sqrt{\varepsilon}} \lesssim |A_{ij}|\}}}_{\substack{\downarrow \\ \text{very sparse} \\ (\varepsilon n \text{ non-zero} \\ \text{elements)}}}$$

Theorem (Part 2: global obstructions)

Let A is an $n \times n$ matrix with i.i.d. entries, such that

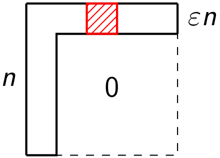
- $\mathbb{E}A_{ij}^2 \geq M$,
- $|A_{ij}| \leq \sqrt{n}$ *almost surely.*

If $M = M(C, \varepsilon)$ is a large enough constant, then any $\varepsilon n \times \varepsilon n$ sub-matrix A_0 has large norm

$$\|A_0\| \geq C\sqrt{n},$$

with probability at least $1 - \exp(-\varepsilon n)$.

So, if we were to cut some part for regularization, we need to cut almost everything! No $\varepsilon n \times \varepsilon n$ sub-matrix can survive.



Proof idea

$$\text{Frobenius norm } \|A_0\|_F^2 := \sum_{i=1}^{n} s_i^2 \leq n \cdot \max s_i^2 = n \cdot \|A_0\|^2$$

- it's enough to show that Frobenius norm is large

$$\|A_0\|_F \geq Cn \; - ?$$

- split elements onto levels "by size":

$$\|A_0\|_F^2 = \sum_{A_{ij} \in A_0} A_{ij}^2 = \sum_{k=0}^{\infty} \sum_{A_{ij} \in A_0} a_{ij}^2 \mathbb{1}_{\{2^k \leq a_{ij}^2 < 2^{k+1}\}}$$

- argue that the majority of the levels in any $\varepsilon n \times \varepsilon n$ sub-block contain many non-zero elements (use Chernoff's inequality).

Done!

Theorem (informal statement)

A is a random square matrix with i.i.d. centered elements a_{ij},

- *if $\mathbb{E}A_{ij}^2$ bounded $\therefore$ there are local obstructions*
- *if not, and entries are $\sqrt{n}$-bounded $\therefore$ there are global obstructions*

for the regularization of the operator norm $\|A\|$.

Thanks for your attention! :)

Appendix

What else can be done with similar techniques...

Theorem (Rudelson, Vershynin)

Let $n \geq n_0$ and let $A = (A_{ij})$ be an $n \times n$ random matrix with i.i.d mean zero subgaussian *entries. Then for any $\varepsilon > 0$ we have*

$$\mathbb{P}\{s_n(A) \leq \varepsilon n^{-1/2}\} \leq L\varepsilon + u^n,$$

where $L > 0$ and $u \in (0,1)$ depend only on the distribution of A_{ij}.

Corollary: i.i.d. matrices with subgaussian entries are well-invertible, as

$$\|A^{-1}\| = s_{\max}(A^{-1}) = 1/s_n(A) \sim \sqrt{n}$$

Appendix

What else can be done with similar techniques...

Theorem (R, Tikhomirov)

Let $n \geq n_0$ and let $A = (A_{ij})$ be an $n \times n$ random matrix with i.i.d mean zero $\mathbb{E}A_{ij}^2 = 1$ entries. Then for any $\varepsilon > 0$ we have

$$\mathbb{P}\{s_n(A) \leq \varepsilon n^{-1/2}\} \leq L\varepsilon + u^n,$$

where $L > 0$ and $u \in (0, 1)$ depend only on the distribution of A_{ij}.

Corollary: i.i.d. matrices with heavy-tailed entries are also well-invertible, as

$$\|A^{-1}\| = s_{\max}(A^{-1}) = 1/s_n(A) \sim \sqrt{n}$$

Subgaussian case, idea of the proof

$$s_n(A) := s_{\min}(A) = \min_{x \in S^{n-1}} \|Ax\|$$

Approximation by the ε−net $\mathcal{N} \subset S^{n-1}$.
For any $x \in S^{n-1}$ find the closest $y \in \mathcal{N}$:

$$\|Ax\| \geq \|Ay\| - \|A(x-y)\| \geq \|Ay\| - \|A\|\|x-y\| \geq \inf_{y \in \mathcal{N}} \|Ay\| - \sqrt{n} \cdot \varepsilon$$

Lemma

For A be a $n \times n$ random matrix with ii.d. subgaussian entries

$$\mathbb{P}\{\|A\| \geq t\sqrt{n}\| \leq \exp(-c_0 t^2 n) \quad \textit{for } t \geq C_0.$$

Challenge: find an ε-net with the sufficiently low cardinality

$$\mathcal{N} \sim \left(\frac{c}{\varepsilon\sqrt{n}}\right)^n$$

Heavy-tailed case, idea of the proof

$$s_n(A) := s_{\min}(A) = \min_{x \in S^{n-1}} \|Ax\|$$

Approximation by the $\varepsilon-$net $\mathcal{N} \subset S^{n-1}$.
For any $x \in S^{n-1}$ find the closest $y \in \mathcal{N}$:

$$\|Ax\| \geq \|Ay\| - \|A(x-y)\| \geq \|Ay\| - \|A\|\|x-y\| \geq ???$$

Norm is too large:

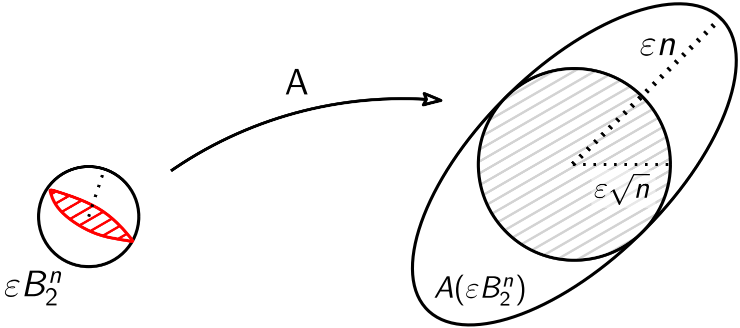
$$\|A\| \sim n \gg \sqrt{n}$$

New challenge: obtain an estimate $\|A(x-y)\| \geq \sqrt{n}\varepsilon$, where x, y are in the same ε-net element.

So, for any x, y taken from one net element we would like to have

$$\|A(x-y)\| \leq \sqrt{n}\varepsilon$$

New net is random (depends on realization of A):



And the net should be refined without blowing up cardinality $|\mathcal{N}|$.
It is possible, as A cannot have too many large directions!
Damping, discretization, . . .