Phase Transitions in Semidefinite Relaxations

Andrea Montanari
[with Adel Javanmard, Federico Ricci-Tersenghi, Subhabrata Sen]

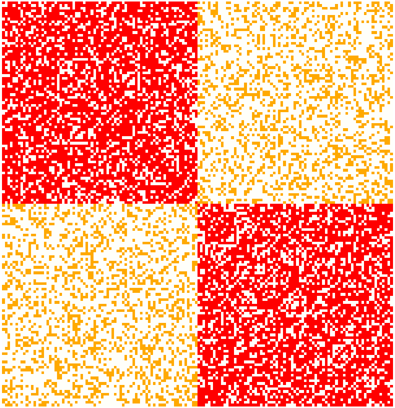
Stanford University

April 5, 2016

What is this talk about?

SDP for Matrix/Graph estimation

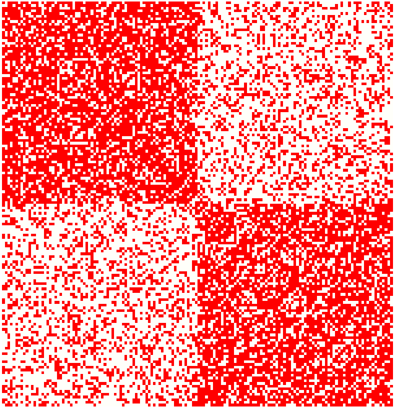
The hidden partition model



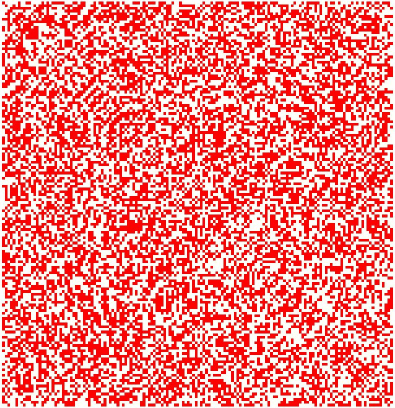
Vertices V, $|V| = n$, $V = V_+ \cup V_-$, $|V_+| = |V_-| = n/2$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} p & \text{if } \{i,j\} \subseteq V_+ \text{ or } \{i,j\} \subseteq V_-, \\ q < p & \text{otherwise.} \end{cases}$$

Of course entries are not colored. . .



...and rows/columns are not ordered



Problem: Detect/estimate the partition

What is this talk about?

SDP for Matrix/Graph estimation

Exact phase transition(?)

Outline

[arXiv:1603.1603.04064]
[arXiv:1511.08769, PNAS]
[arXiv:1504.05910, STOC]

Background

Estimation

$\boldsymbol{x_0} \in \{+1,-1\}^n$, $\langle \boldsymbol{x_0}, \mathbf{1} \rangle = 0$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} p & \text{if } x_{0,i} = x_{0,j}, \\ q & \text{otherwise.} \end{cases}$$

Estimation ($p = a/n$, $q = b/n$)

$\boldsymbol{x_0} \in \{+1, -1\}^n$, $\langle \boldsymbol{x_0}, \mathbf{1} \rangle = 0$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} a/n & \text{if } x_{0,i} = x_{0,j}, \\ b/n & \text{otherwise.} \end{cases}$$

Estimation threshold

$\boldsymbol{x_0} \in \{+1,-1\}^n$, $\langle \boldsymbol{x_0}, \mathbf{1}\rangle = 0$

$$\mathbb{P}\{(i,j)\in E\} = \begin{cases} a/n & \text{if } x_{0,i} = x_{0,j}, \\ b/n & \text{otherwise.} \end{cases}$$

$$\widehat{\mathrm{x}} : \mathcal{G} \to \{+1,-1\}^n ,$$
$$\mathrm{Overlap}_n(\widehat{\mathrm{x}}) = \frac{1}{n}\mathbb{E}\{\big|\langle \widehat{\mathrm{x}}(G), \boldsymbol{x_0}\rangle\big|\} .$$

Fow which a, b, $\liminf_{n\to\infty} \mathrm{Overlap}_n(\widehat{\mathrm{x}}) \geq \delta > 0$?

Estimation threshold

$x_0 \in \{+1,-1\}^n$, $\langle x_0, \mathbf{1}\rangle = 0$

$$\mathbb{P}\{(i,j)\in E\} = \begin{cases} a/n & \text{if } x_{0,i} = x_{0,j}\,, \\ b/n & \text{otherwise.} \end{cases}$$

$$\hat{\mathrm{x}} : \mathcal{G} \to \{+1,-1\}^n\,,$$
$$\mathrm{Overlap}_n(\hat{\mathrm{x}}) = \frac{1}{n}\mathbb{E}\{\big|\langle \hat{\mathrm{x}}(G), x_0\rangle\big|\}\,.$$

Fow which a, b, $\liminf_{n\to\infty} \mathrm{Overlap}_n(\hat{\mathrm{x}}) \geq \delta > 0$?

Estimation threshold

$x_0 \in \{+1,-1\}^n$, $\langle x_0, \mathbf{1}\rangle = 0$

$$\mathbb{P}\{(i,j)\in E\} = \begin{cases} a/n & \text{if } x_{0,i} = x_{0,j}, \\ b/n & \text{otherwise.} \end{cases}$$

$$\widehat{\mathrm{x}} : \mathcal{G} \to \{+1,-1\}^n ,$$
$$\mathrm{Overlap}_n(\widehat{\mathrm{x}}) = \frac{1}{n}\mathbb{E}\{\big|\langle \widehat{\mathrm{x}}(G), x_0\rangle\big|\} .$$

Fow which a, b, $\liminf_{n\to\infty} \mathrm{Overlap}_n(\widehat{\mathrm{x}}) \geq \delta > 0$?

Information theory threshold

Theorem (Mossel, Neeman, Sly, 2012)

There is an estimator that is better than random if and only if

$$\frac{a-b}{\sqrt{2(a+b)}} > 1\,.$$

[Proves conjecture by Decelle, Krzakala, Moore, Zdeborova, 2011]

Information theory threshold

Theorem (Mossel, Neeman, Sly, 2012)

There is an estimator that is better than random if and only if

$$\frac{a-b}{\sqrt{2(a+b)}} > 1\,.$$

[Proves conjecture by Decelle, Krzakala, Moore, Zdeborova, 2011]

Computational threshold

- Dyer, Frieze 1989 — $p = na > q = nb$ fixed.
- Condon, Karp 2001 — $a - b \gg n^{1/2}$
- McSherry 2001 — $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 — $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a-b}{\sqrt{2(a+b)}} > 1$$

Very ingenious spectral methods!

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed.
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a-b}{\sqrt{2(a+b)}} > 1$$

Very ingenious spectral methods!

Computational threshold

- Dyer, Frieze 1989 — $p = na > q = nb$ fixed.
- Condon, Karp 2001 — $a - b \gg n^{1/2}$
- McSherry 2001 — $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 — $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a - b}{\sqrt{2(a + b)}} > 1$$

Very ingenious spectral methods!

What if I am not ingenious?

Maximum Likelihood

Adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{A} = (A_{ij})_{1 \le i,j \le n}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j \,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0 \,, \\ & \sigma_i \in \{+1, -1\} \,. \end{aligned}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j \,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0 \,, \\ & \sigma_i \in \{+1, -1\} \,. \end{aligned}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j \,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0 \,, \\ & \sigma_i \in \{+1, -1\} \,. \end{aligned}$$

Centered adjacency matrix

$$A^{\text{cen}}_{ij} = \begin{cases} 1 - (d/n) & \text{if } (i,j) \in E, \\ -(d/n) & \text{otherwise.} \end{cases}$$

$$\boldsymbol{A}^{\text{cen}} = \boldsymbol{A} - \frac{d}{n}\mathbf{1}\,\mathbf{1}^{\mathsf{T}}$$

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

- NP-hard
- SDP($\boldsymbol{A}^{\text{cen}}$) is a very natural convex relaxation

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle\,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n\,. \end{aligned}$$

- NP-hard
- $\mathsf{SDP}(\boldsymbol{A}^{\text{cen}})$ is a very natural convex relaxation

Relaxation

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

SDP($\boldsymbol{A}^{\text{cen}}$):

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{X} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n \times n}, \; \boldsymbol{X} \succeq 0 \,, \\ & X_{ii} = 1 \,. \end{aligned}$$

- This is really **off-the-shelf**

- How well does it work?

Near-optimality of SDP

Before we pass to SDP

- What's the problem with sparse graphs?
- What's the problem vanilla PCA?

Why PCA?

Ground truth

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

Data = RankOne + Wigner

$$\frac{1}{\sqrt{d}} \boldsymbol{A}^{\text{cen}} = \frac{\lambda}{n} x_0 x_0^{\mathsf{T}} + \boldsymbol{W}, \qquad \lambda \equiv \frac{a-b}{\sqrt{2(a+b)}}$$

$$\mathbb{E}\{W_{ij}\} = 0, \qquad \mathbb{E}\{W_{ij}^2\} \in \left\{\frac{a}{dn}, \frac{b}{dn}\right\} \approx \frac{1}{n}.$$

Why PCA?

Ground truth

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

Data = RankOne + Wigner

$$\frac{1}{\sqrt{d}} \boldsymbol{A}^{\text{cen}} = \frac{\lambda}{n} \boldsymbol{x_0} \boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \lambda \equiv \frac{a-b}{\sqrt{2(a+b)}}$$

$$\mathbb{E}\{W_{ij}\} = 0\,, \qquad \mathbb{E}\{W_{ij}^2\} \in \left\{ \frac{a}{dn}\ ,\ \frac{b}{dn} \right\} \approx \frac{1}{n}\,.$$

The *right* parametrization

$$d = \frac{a+b}{2}, \qquad \lambda = \frac{a-b}{\sqrt{2(a+b)}}$$

Naive PCA

$$\widehat{\mathbf{x}}^{\rm PCA}(\boldsymbol{A}^{\rm cen}) = {\rm sign}(\boldsymbol{v}_1(\boldsymbol{A}^{\rm cen}))\,.$$

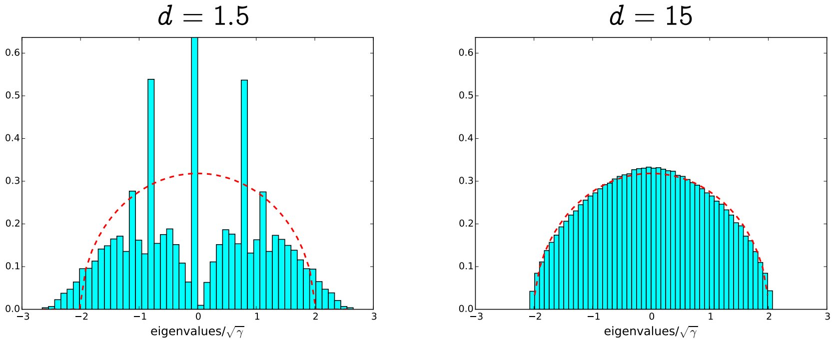
Does it work?

No!

Does it work?

No!

Spectral relaxation bad in the sparse regime!

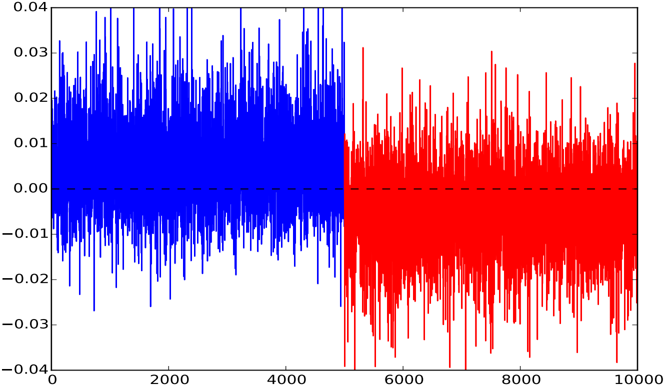


Theorem (Krivelevich, Sudakov 2003)

With high probability, for $d = O(1)$, $\lambda = 0$

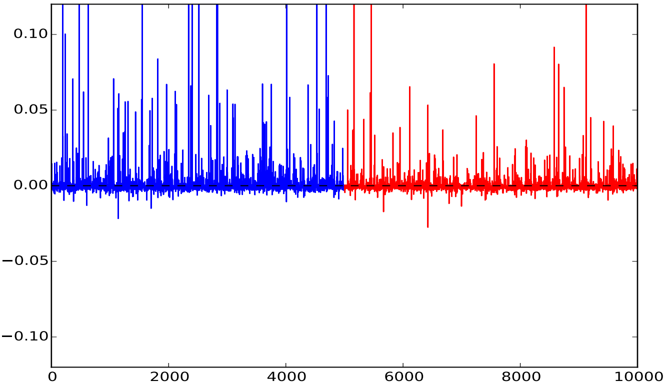
$$\lambda_{\max}(A^{cen}/\sqrt{d}) = C\sqrt{\log n/(\log\log n)}(1 + o(1)) \, .$$

Example: $d = 20$, $\lambda = 1.2$, $n = 10^4$



$v_1(A^{\text{cen}})$

Example: $d = 3$, $\lambda = 1.2$, $n = 10^4$



$v_1(A^{\text{cen}})$

Why should SDP work better?

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\rm cen}, \boldsymbol{X} \rangle\,, \\ \text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n\times n},\ \boldsymbol{X} \succeq 0\,, \\ & X_{ii} = 1\,. \end{aligned}$$

Recall the info-theoretic limit

$$d = \frac{a+b}{2} > 1\,, \qquad \lambda = \frac{a-b}{\sqrt{2(a+b)}}$$

Theorem (Mossel, Neeman, Sly, 2012)

There is an estimator that is better than random if and only if

$$\lambda > 1\,.$$

SDP has nearly optimal threshold

Theorem (Montanari, Sen 2015)

The SDP estimator is better than random if[a]

$$\lambda \geq 1 + o_d(1)\,.$$

[a] Can replace $o_d(1)$ by $C_0\,d^{-c_1}$.

SDP estimator

Ideally

- Let $\boldsymbol{X}_* \in \mathbb{R}^{n\times n}$ be an optimizer
- Compute the principal eigen-vector $v_1(\boldsymbol{X}_*)$.
- Return $\hat{\boldsymbol{x}}^{\rm SDP}(G) = {\rm sign}(v_1(\boldsymbol{X}_*))$

In reality, a few tricks

SDP estimator

Ideally

- Let $\boldsymbol{X}_* \in \mathbb{R}^{n\times n}$ be an optimizer
- Compute the principal eigen-vector $v_1(\boldsymbol{X}_*)$.
- Return $\hat{\boldsymbol{x}}^{\rm SDP}(G) = {\rm sign}(v_1(\boldsymbol{X}_*))$

In reality, a few tricks

Key technical result: SDP value

Theorem (Montanari, Sen 2015)

Assume $G \sim \mathsf{G}(n, d, \lambda)$.
If $\lambda \le 1$, *then, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + o_d(1)\,.$$

If $\lambda > 1$, *then there exists* $\Delta(\lambda) > 0$ *such that, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + \Delta(\lambda) + o_d(1)\,.$$

Earlier/related work

Optimal spectral methods

- Massoulie 2013
- Mossel, Neeman, Sly, 2013
- Bordenave, Lelarge, Massoulie, 2015

SDP, $d = \Theta(\log n)$

- Abbe, Bandeira, Hall 2014
- Hajek, Wu, Xu 2015

SDP, detection

- Guédon, Vershynin, 2015 (requires $\lambda \geq 10^4$, very different proof)

How does SDP work ‘in practice’?

Thresholds

- $\lambda_c^{\rm opt}(d) \equiv$ Threshold for optimal estimator

- $\lambda_c^{\rm SDP}(d) \equiv$ Threshold for SDP-based estimator

What we know

- $\lambda_c^{\text{opt}}(d) = 1$ [Mossel, Neeman, Sly, 2013]
- $\lambda_c^{\text{SDP}}(d) = 1 + o_d(1)$ [Montanari, Sen, 2015]

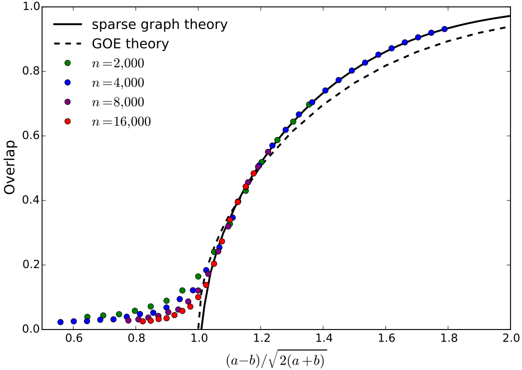
How big is the $o_d(1)$ gap?

What we know

- $\lambda_c^{\rm opt}(d) = 1$ [Mossel, Neeman, Sly, 2013]
- $\lambda_c^{\rm SDP}(d) = 1 + o_d(1)$ [Montanari, Sen, 2015]

How big is the $o_d(1)$ gap?

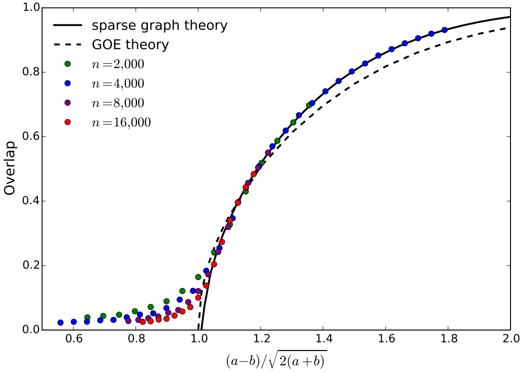
Simulations: $d = 5$, $N_{\text{sample}} = 500$ (with Javanmard and Ricci)



SDP estimator $\hat{x}^{\text{SDP}} \in \{+1, -1\}^n$

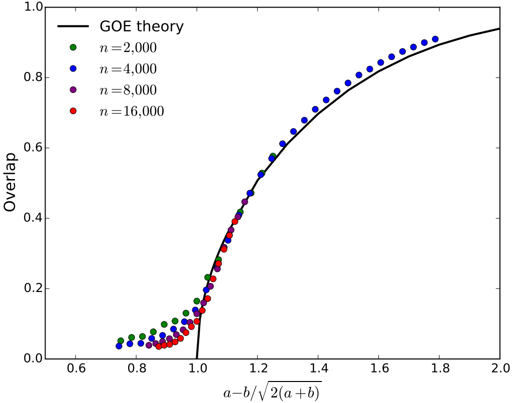
$$\text{Overlap}_n(\hat{\mathbf{x}}) = \frac{1}{n}\mathbb{E}\{\left|\langle \hat{x}^{\text{SDP}}(G), x_0 \rangle\right|\}.$$

Simulations: $d = 5$, $N_{\text{sample}} = 500$



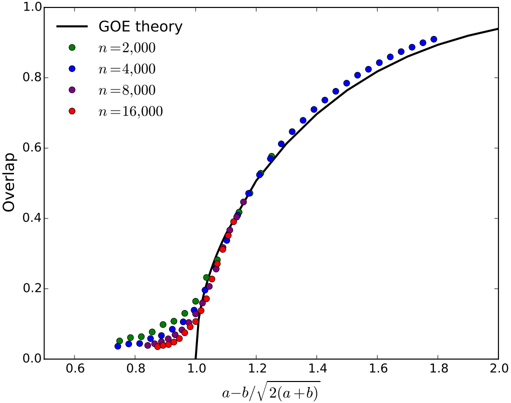
$$\lambda_c^{\text{SDP}}(d = 5) \approx 1\,.$$

Simulations: $d = 10$, $N_{\text{sample}} = 500$



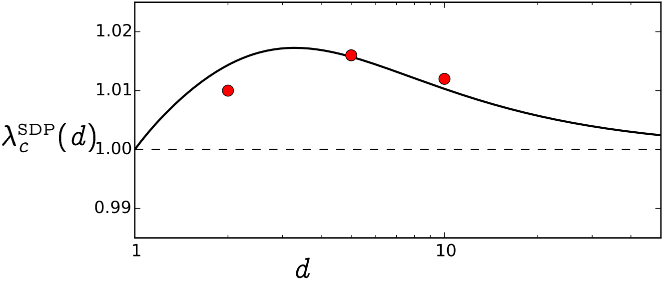
$$\lambda_c^{\text{SDP}}(d = 10) \approx 1\,.$$

Simulations: $d = 10$, $N_{\rm sample} = 500$



Estimate $\lambda_c^{SDP}(d)$ from data... *(a long story)*

$\lambda_c^{\rm SDP}(d)$



- Dots: Numerical estimates
- Line: Non-rigorous analytical approximation (using statistical physics)
- **At most 2% sub-optimal!**
- (Moitra, Perry, Wein 2016) suggests $\lambda_c^{\rm SDP}(d) > 1$ for d small
- (Fan, Montanari, 2016) $\lambda_c^{\rm SDP}(d) < c(d) \leq 2$ for all d

How do you actually solve the SDP?

How do you *actually* solve the SDP?

- Can do $n = 10^5$ in minutes.
- Code: http://web.stanford.edu/∼montanar/SDPgraph
- Burer-Monteiro trick: solve rank-constrained problem

Rank-constrained SDP (non-convex!!!)

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{M}, \boldsymbol{X} \rangle\,, \\ \text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n\times n},\ \boldsymbol{X} \succeq 0\,, \\ & X_{ii} = 1\,, \\ & \text{rank}(\boldsymbol{X}) \le k \end{aligned}$$

Equivalently:

$$\begin{aligned} \text{maximize} \quad & F_M(\boldsymbol{\sigma}) \equiv \sum_{i,j=1}^{n} M_{ij} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle\,, \\ \text{subject to} \quad & \boldsymbol{\sigma}_i \in \mathbb{R}^k,\ \|\boldsymbol{\sigma}_i\|_2 = 1\,. \end{aligned}$$

Rank-constrained SDP (non-convex!!!)

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{M}, \boldsymbol{X} \rangle\,, \\ \text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n\times n},\ \boldsymbol{X} \succeq 0\,, \\ & X_{ii} = 1\,, \\ & \text{rank}(\boldsymbol{X}) \le k \end{aligned}$$

Equivalently:

$$\begin{aligned} \text{maximize} \quad & F_{\boldsymbol{M}}(\boldsymbol{\sigma}) \equiv \sum_{i,j=1}^{n} M_{ij} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle\,, \\ \text{subject to} \quad & \boldsymbol{\sigma}_i \in \mathbb{R}^k,\ \ \|\boldsymbol{\sigma}_i\|_2 = 1\,. \end{aligned}$$

A Grothendieck-type inequality for local maxima

Theorem (Montanari, 2016)

Let $\boldsymbol{\sigma}$ be a local maximum of F_M over $(S^{k-1})^n$. Then

$$F_M(\boldsymbol{\sigma}) \geq \mathsf{SDP}(M) - \frac{8}{\sqrt{k}}\, n\|M\|_2\,.$$

Typically:

- $\|M\|_2 = O(1)$, $\mathsf{SDP}(M) = \Theta(n)$.
- **Sufficient to take $k = O(1)$!!!!**

Related work

- Bandeira, Boumal, Voroninski, 2016:
  $\mathbb{Z}_2$ synchronization, $k = 2, \lambda > C$

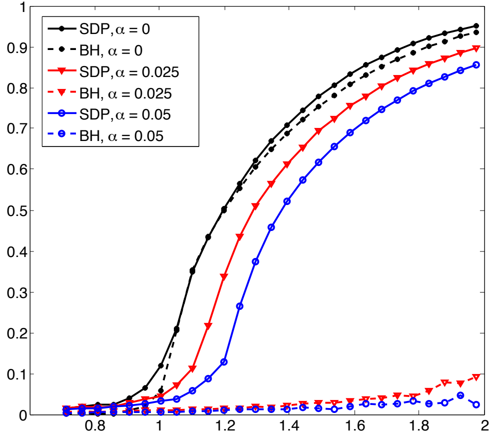
One last question

Is this approach robust to model miss-specifications?

An experiment

- Select $S \subseteq V$ uniformly at random. with $|S| = n\alpha$.
- For each $i \in S$, connect all of its neighbors.

An experiment



- Solid line:SDP
- Dashed line: Spectral
  (Non-backtracking walk [Krzakala, Moore, Mossel, Neeman, Sly, Zdeborova, Zhang, 2013])

Proof ideas

What we want to prove

Theorem (Montanari, Sen 2015)

If $\lambda \leq 1$*, then, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + o_d(1)\,.$$

If $\lambda > 1$*, then there exists* $\Delta(\lambda) > 0$ *such that, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + \Delta(\lambda) + o_d(1)\,.$$

Strategy

I. Prove equivalence to Gaussian model

II. Analyze Gaussian model

Strategy

I. Prove equivalence to Gaussian model

II. Analyze Gaussian model

Gaussian model: $\mathbb{Z}_2$ synchronization

$$B(\lambda) \equiv \frac{\lambda}{n} x_0 x_0^{\mathsf{T}} + W\,.$$

$x_0 \in \{+1, -1\}^n$, $W \sim \mathsf{GOE}(n)$

- $(W_{ij})_{i<j} \sim_{iid} \mathsf{N}(0, 1/n)$
- $W = W^{\mathsf{T}}$

- *A lot is known about spectral properties of B*

Need to characterize the SDP value with Gaussian data

Gaussian model: $\mathbb{Z}_2$ synchronization

$$\boldsymbol{B}(\lambda) \equiv \frac{\lambda}{n}\,\boldsymbol{x_0}\boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}\,.$$

$\boldsymbol{x_0} \in \{+1,-1\}^n,\ \boldsymbol{W} \sim \mathsf{GOE}(n)$

- $(W_{ij})_{i<j} \sim_{iid} \mathrm{N}(0, 1/n)$
- $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$

- *A lot is known about spectral properties of* $\boldsymbol{B}$

Need to characterize the SDP value with Gaussian data

Gaussian model: $\mathbb{Z}_2$ synchronization

$$\boldsymbol{B}(\lambda) \equiv \frac{\lambda}{n}\, \boldsymbol{x_0} \boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}\, .$$

$\boldsymbol{x_0} \in \{+1, -1\}^n,\ \boldsymbol{W} \sim \mathsf{GOE}(n)$

- $(W_{ij})_{i<j} \sim_{iid} \mathrm{N}(0, 1/n)$
- $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$

- *A lot is known about spectral properties of* $\boldsymbol{B}$

Need to characterize the SDP value with Gaussian data

Notation

$$\bar{\mathsf{s}}(\lambda) \equiv \lim\sup_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{B}(\lambda))\,,$$

$$\underline{\mathsf{s}}(\lambda) \equiv \lim\inf_{n\to\infty} \frac{1}{n}\mathsf{SDP}(\boldsymbol{B}(\lambda))\,.$$

$$\mathsf{SDP}(\boldsymbol{B}) \equiv \max\{\langle \boldsymbol{B}, \boldsymbol{X}\rangle : \ \boldsymbol{X} \succeq 0, X_{ii} = 1\,\forall i\}$$

Phase transition at $\lambda = 1$!

Theorem (Montanari, Sen, 2015)

The following holds almost surely

$$\lambda \in [0,1] \Rightarrow \qquad \underline{s}(\lambda) = \overline{s}(\lambda) = 2\,,$$

$$\lambda \in (1,\infty) \Rightarrow \qquad \underline{s}(\lambda) > 2 \quad \textit{(strictly)}\,.$$

For explicit probability bounds, see the paper

Proof of Gaussian phase transition

Simple facts:

- $$\overline{\mathsf{s}}(\lambda) \le \lim_{n\to\infty} \sigma_{\max}(\boldsymbol{B}(\lambda)) = \begin{cases} 2 & \text{if } \lambda \in [0,1], \\ \lambda + \lambda^{-1} & \text{if } \lambda \in (1,\infty). \end{cases}$$

  [Baik, Ben Arous, Peche, 2005]

- $$\underline{\mathsf{s}}(\lambda) \ge \lim_{n\to\infty} \frac{1}{n}\langle \mathbf{1}, \boldsymbol{B}(\lambda)\mathbf{1}\rangle = \lambda$$

- $\underline{\mathsf{s}}(\lambda)$, $\overline{\mathsf{s}}(\lambda)$ are non-random, non-decreasing

Proof of Gaussian phase transition

Simple facts:

- $$\bar{\mathsf{s}}(\lambda) \le \lim_{n\to\infty} \sigma_{\max}(\boldsymbol{B}(\lambda)) = \begin{cases} 2 & \text{if } \lambda \in [0,1], \\ \lambda + \lambda^{-1} & \text{if } \lambda \in (1,\infty). \end{cases}$$

  [Baik, Ben Arous, Peche, 2005]

- $$\underline{\mathsf{s}}(\lambda) \ge \lim_{n\to\infty} \frac{1}{n}\langle \mathbf{1}, \boldsymbol{B}(\lambda)\mathbf{1}\rangle = \lambda$$

- $\underline{\mathsf{s}}(\lambda)$, $\bar{\mathsf{s}}(\lambda)$ are non-random, non-decreasing

Proof of Gaussian phase transition

Simple facts:

- $$\bar{\mathsf{s}}(\lambda) \le \lim_{n\to\infty} \sigma_{\max}(\boldsymbol{B}(\lambda)) = \begin{cases} 2 & \text{if } \lambda \in [0,1], \\ \lambda + \lambda^{-1} & \text{if } \lambda \in (1,\infty). \end{cases}$$

  [Baik, Ben Arous, Peche, 2005]

- $$\underline{\mathsf{s}}(\lambda) \ge \lim_{n\to\infty} \frac{1}{n}\langle \mathbf{1}, \boldsymbol{B}(\lambda)\mathbf{1}\rangle = \lambda$$

- $\underline{\mathsf{s}}(\lambda)$, $\bar{\mathsf{s}}(\lambda)$ are non-random, non-decreasing

Proof of Gaussian phase transition

Simple facts:

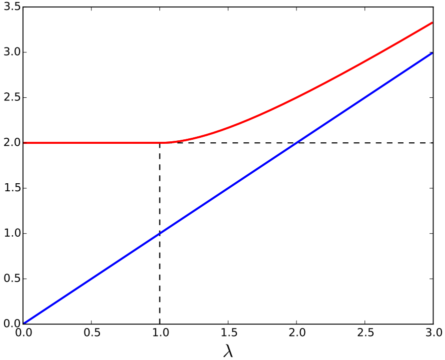
- $$\overline{\mathsf{s}}(\lambda) \le \lim_{n\to\infty} \sigma_{\max}(\boldsymbol{B}(\lambda)) = \begin{cases} 2 & \text{if } \lambda \in [0,1], \\ \lambda + \lambda^{-1} & \text{if } \lambda \in (1,\infty). \end{cases}$$

  [Baik, Ben Arous, Peche, 2005]

- $$\underline{\mathsf{s}}(\lambda) \ge \lim_{n\to\infty} \frac{1}{n}\langle \mathbf{1}, \boldsymbol{B}(\lambda)\mathbf{1}\rangle = \lambda$$

- $\underline{\mathsf{s}}(\lambda)$, $\overline{\mathsf{s}}(\lambda)$ are non-random, non-decreasing

Summarizing



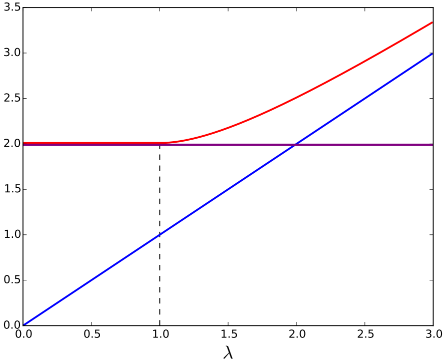
- Red: Upper bound
- Blue: Lower bound

Proof of Gaussian phase transition

Part 1:

Prove that $\underline{\mathsf{s}}(\lambda = 0) \geq 2$

Hence



- Red: Upper bound
- Blue: Lower bound
- Purple: Non-trivial lower bound

Proof of Gaussian phase transition

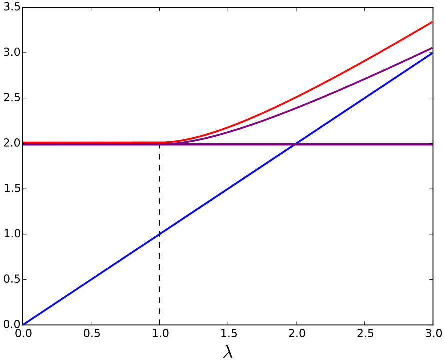
Part 1:

Prove that $\underline{\mathsf{s}}(\lambda = 0) \geq 2$

Part 2:

Prove that $\underline{\mathsf{s}}(\lambda = 1 + \varepsilon) > 2$

Hence



- Red: Upper bound
- Blue: Lower bound
- Purple: Non-trivial lower bound

Proof of Gaussian phase transition

Part 1:

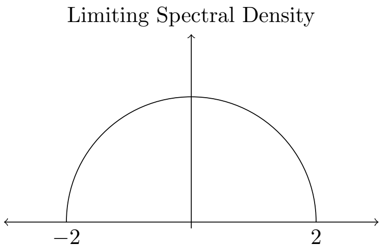
Prove that $\underline{\mathsf{s}}(\lambda = 0) \geq 2$

Part 2:

Prove that $\underline{\mathsf{s}}(\lambda = 1 + \varepsilon) > 2$

Technique: Construct feasible $\boldsymbol{X}$, such that $\langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq \dots$.

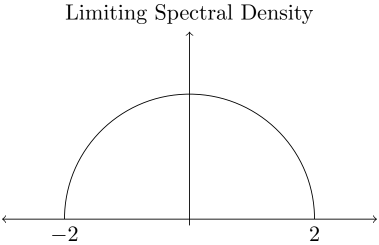
Part 1: $\lambda = 0$



First idea:

- $v_1 = v_1(B) \equiv$ principal eginvector of B
- Take $X = n\, v_1 v_1^{\mathsf{T}}$
- **Wrong:** $X_{ii} \approx \mathrm{N}(0,1)^2 \neq 1$

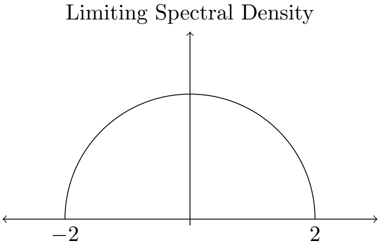
Part 1: $\lambda = 0$



First idea:

- $v_1 = v_1(B) \equiv$ principal eginvector of B
- Take $X = n\, v_1 v_1^{\mathsf{T}}$
- **Wrong:** $X_{ii} \approx \mathrm{N}(0,1)^2 \neq 1$

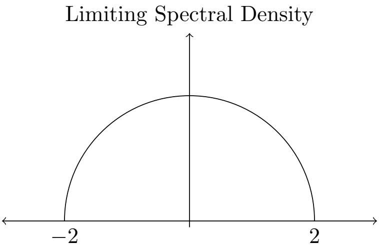
Part 1: $\lambda = 0$



First idea:

- $\boldsymbol{v}_1 = \boldsymbol{v}_1(\boldsymbol{B}) \equiv$ principal eginvector of $\boldsymbol{B}$
- Take $\boldsymbol{X} = n\, \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}}$
- **Wrong:** $X_{ii} \approx \mathrm{N}(0, 1)^2 \neq 1$

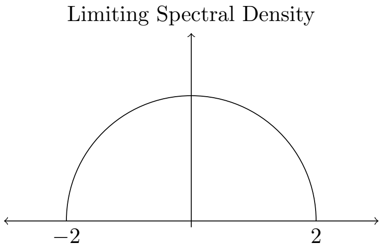
Part 1: $\lambda = 0$



Good idea:

- Let $U = [v_1|v_2|\cdots|v_{n\delta}] \in \mathbb{R}^{n\times n\delta}$, δ small.
- $D \equiv \mathrm{Diag}(U\,U^{\mathsf{T}}) \in \mathbb{R}^{n\times n}$. Claim $D \approx \delta\,\mathbf{I}$ ($D_{ii} \sim n^{-1}\chi_{n\delta}$)
- Set $X = D^{-1/2}(U\,U^{\mathsf{T}})D^{-1/2}$

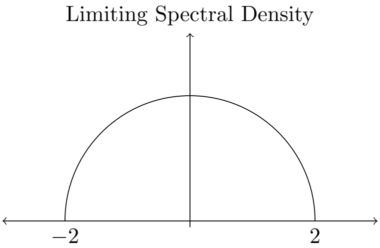
Part 1: $\lambda = 0$



Good idea:

- Let $\boldsymbol{U} = [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_{n\delta}] \in \mathbb{R}^{n\times n\delta}$, δ small.
- $\boldsymbol{D} \equiv \text{Diag}(\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}) \in \mathbb{R}^{n\times n}$. Claim $\boldsymbol{D} \approx \delta\,\mathbf{I}$ $(D_{ii} \sim n^{-1}\chi_{n\delta})$
- Set $\boldsymbol{X} = \boldsymbol{D}^{-1/2}(\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}})\boldsymbol{D}^{-1/2}$

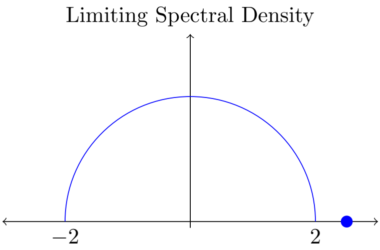
Part 1: $\lambda = 0$



Good idea:

- Let $\boldsymbol{U} = [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_{n\delta}] \in \mathbb{R}^{n\times n\delta}$, δ small.
- $\boldsymbol{D} \equiv \text{Diag}(\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}) \in \mathbb{R}^{n\times n}$. Claim $\boldsymbol{D} \approx \delta\,\mathbf{I}$ $(D_{ii} \sim n^{-1}\chi_{n\delta})$
- Set $\boldsymbol{X} = \boldsymbol{D}^{-1/2}(\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}})\boldsymbol{D}^{-1/2}$

Part 2: $\lambda = 1 + \varepsilon$



Construction:

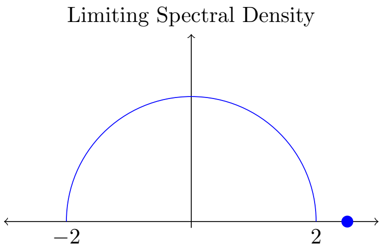
- $T(x) \equiv \max(\min(x, +1), -1)$, $\varphi \in \mathbb{R}^n$

$$\varphi_i \equiv T(\varepsilon\sqrt{n}\, v_{1,i})\,.$$

- $\boldsymbol{U} = [\boldsymbol{v}_2|\boldsymbol{v}_3|\cdots|\boldsymbol{v}_{n\delta+1}] \in \mathbb{R}^{n\times n\delta}$
- $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ diagonal $D_{ii} \equiv \sqrt{1-\varphi_i^2}/\|\boldsymbol{U}\mathbf{e}_i\|_2$.
-

$$\boldsymbol{X} \equiv \varphi\varphi^{\mathsf{T}} + \boldsymbol{D}\,\boldsymbol{U}\,\boldsymbol{U}^{\mathsf{T}}\boldsymbol{D}$$

Part 2: $\lambda = 1 + \varepsilon$



Construction:

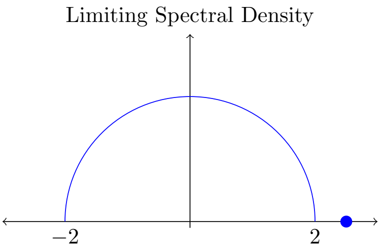
- $T(x) \equiv \max(\min(x, +1), -1)$, $\varphi \in \mathbb{R}^n$

$$\varphi_i \equiv T(\varepsilon\sqrt{n}\, v_{1,i})\,.$$

- $U = [v_2 | v_3 | \cdots | v_{n\delta+1}] \in \mathbb{R}^{n \times n\delta}$
- $D \in \mathbb{R}^{n \times n}$ diagonal $D_{ii} \equiv \sqrt{1 - \varphi_i^2} / \| U e_i \|_2$.
-

$$X \equiv \varphi\varphi^{\mathsf{T}} + D\,U\,U^{\mathsf{T}} D$$

Part 2: $\lambda = 1 + \varepsilon$



Construction:

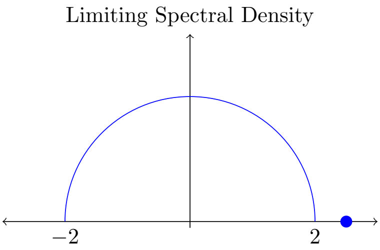
- $T(x) \equiv \max(\min(x, +1), -1)$, $\varphi \in \mathbb{R}^n$

$$\varphi_i \equiv T(\varepsilon\sqrt{n}\, v_{1,i})\,.$$

- $\boldsymbol{U} = [\boldsymbol{v}_2 | \boldsymbol{v}_3 | \cdots | \boldsymbol{v}_{n\delta+1}] \in \mathbb{R}^{n \times n\delta}$
- $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ diagonal $D_{ii} \equiv \sqrt{1 - \varphi_i^2} / \|\boldsymbol{U} \mathrm{e}_i\|_2$.
-

$$\boldsymbol{X} \equiv \varphi\varphi^{\mathsf{T}} + \boldsymbol{D}\,\boldsymbol{U}\,\boldsymbol{U}^{\mathsf{T}}\boldsymbol{D}$$

Part 2: $\lambda = 1 + \varepsilon$



Construction:

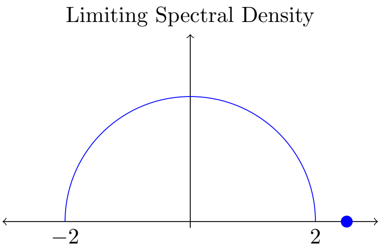
- $T(x) \equiv \max(\min(x, +1), -1)$, $\boldsymbol{\varphi} \in \mathbb{R}^n$

$$\varphi_i \equiv T(\varepsilon\sqrt{n}\, v_{1,i})\,.$$

- $\boldsymbol{U} = [\boldsymbol{v}_2|\boldsymbol{v}_3|\cdots|\boldsymbol{v}_{n\delta+1}] \in \mathbb{R}^{n\times n\delta}$
- $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ diagonal $D_{ii} \equiv \sqrt{1-\varphi_i^2}/\|\boldsymbol{U}\mathbf{e}_i\|_2$.
-

$$\boldsymbol{X} \equiv \boldsymbol{\varphi}\boldsymbol{\varphi}^{\mathsf{T}} + \boldsymbol{D}\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{D}$$

Part 2: $\lambda = 1 + \varepsilon$



Construction:

- $T(x) \equiv \max(\min(x, +1), -1)$, $\boldsymbol{\varphi} \in \mathbb{R}^n$

$$\varphi_i \equiv T(\varepsilon\sqrt{n}\, v_{1,i})\,.$$

- $\boldsymbol{U} = [\boldsymbol{v}_2|\boldsymbol{v}_3|\cdots|\boldsymbol{v}_{n\delta+1}] \in \mathbb{R}^{n\times n\delta}$
- $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ diagonal $D_{ii} \equiv \sqrt{1-\varphi_i^2}/\|\boldsymbol{U}\mathbf{e}_i\|_2$.
- $$\boldsymbol{X} \equiv \boldsymbol{\varphi}\boldsymbol{\varphi}^{\mathsf{T}} + \boldsymbol{D}\,\boldsymbol{U}\,\boldsymbol{U}^{\mathsf{T}}\boldsymbol{D}$$

Strategy

I. Prove equivalence to Gaussian model

II. Analyze Gaussian model

We want to prove

$$\frac{1}{\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}^{\rm cen}) \approx \mathsf{SDP}(\boldsymbol{B}(\lambda))$$

Lindeberg method: Replace the entries one-by-one

We want to prove

$$\frac{1}{\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}^{\rm cen}) \approx \mathsf{SDP}(\boldsymbol{B}(\lambda))$$

Lindeberg method: Replace the entries one-by-one

A simple Lindeberg lemma

- $X_1, X_2, \dots X_M$ iid

$$X_i = \begin{cases} \frac{1}{\sqrt{d}}\Big(1-\frac{d}{n}\Big) & \text{with probability } \frac{d}{n}\,, \\ -\frac{\sqrt{d}}{n} & \text{with probability } \Big(1-\frac{d}{n}\Big)\,. \end{cases}$$

$\mathbb{E}\{X_i\} = 0$, $\mathbb{E}\{X_i^2\} = (1/n) - (d/n^2)$

- $Z_1, Z_2, \dots Z_M \sim_{i.i.d.} \mathrm{N}(0, 1/n)$

Lemma

Assume $M = n(n-1)/2$, $F \in C_3(\mathbb{R}^M)$, $d \le n^{2/3}/10$. *Then*

$$\Big|\mathbb{E}F(\boldsymbol{X}) - \mathbb{E}F(\boldsymbol{Z})\Big| \le \frac{n}{3\sqrt{d}} \max_{i\in[M]} \Big(\|\partial_i^2 F\|_\infty \vee \|\partial_i^3 F\|_\infty\Big)\,.$$

where $\partial_i^\ell F(x) \equiv \frac{\partial^\ell F}{\partial x_i^\ell}$, *and* $\|\partial_i^\ell F\|_\infty \equiv \sup_{x\in\mathbb{R}^M} |\partial_i^\ell F(x)|$.

Problem: $F(\,\cdot\,) = \mathsf{SDP}(\,\cdot\,) \notin C_3(\mathbb{R}^M)$

A simple Lindeberg lemma

- $X_1, X_2, \dots X_M$ iid

$$X_i = \begin{cases} \frac{1}{\sqrt{d}}\left(1 - \frac{d}{n}\right) & \text{with probability } \frac{d}{n}\,, \\ -\frac{\sqrt{d}}{n} & \text{with probability } \left(1 - \frac{d}{n}\right). \end{cases}$$

$\mathbb{E}\{X_i\} = 0$, $\mathbb{E}\{X_i^2\} = (1/n) - (d/n^2)$

- $Z_1, Z_2, \dots Z_M \sim_{i.i.d.} \mathrm{N}(0, 1/n)$

Lemma

Assume $M = n(n-1)/2$, $F \in C_3(\mathbb{R}^M)$, $d \le n^{2/3}/10$. *Then*

$$\Big|\mathbb{E}F(\boldsymbol{X}) - \mathbb{E}F(\boldsymbol{Z})\Big| \le \frac{n}{3\sqrt{d}} \max_{i \in [M]} \left(\|\partial_i^2 F\|_\infty \vee \|\partial_i^3 F\|_\infty \right).$$

where $\partial_i^\ell F(x) \equiv \frac{\partial^\ell F}{\partial x_i^\ell}$, *and* $\|\partial_i^\ell F\|_\infty \equiv \sup_{\boldsymbol{x} \in \mathbb{R}^M} |\partial_i^\ell F(\boldsymbol{x})|$.

Problem: $F(\,\cdot\,) = \mathrm{SDP}(\,\cdot\,) \notin C_3(\mathbb{R}^M)$

A simple Lindeberg lemma

- $X_1, X_2, \dots X_M$ iid

$$X_i = \begin{cases} \frac{1}{\sqrt{d}}\left(1 - \frac{d}{n}\right) & \text{with probability } \frac{d}{n}, \\ -\frac{\sqrt{d}}{n} & \text{with probability } \left(1 - \frac{d}{n}\right). \end{cases}$$

  $\mathbb{E}\{X_i\} = 0$, $\mathbb{E}\{X_i^2\} = (1/n) - (d/n^2)$
- $Z_1, Z_2, \dots Z_M \sim_{i.i.d.} \mathrm{N}(0, 1/n)$

Lemma

Assume $M = n(n-1)/2$, $F \in C_3(\mathbb{R}^M)$, $d \le n^{2/3}/10$. *Then*

$$\left|\mathbb{E}F(\boldsymbol{X}) - \mathbb{E}F(\boldsymbol{Z})\right| \le \frac{n}{3\sqrt{d}} \max_{i\in[M]} \left(\|\partial_i^2 F\|_\infty \vee \|\partial_i^3 F\|_\infty\right).$$

where $\partial_i^\ell F(\boldsymbol{x}) \equiv \frac{\partial^\ell F}{\partial x_i^\ell}$, *and* $\|\partial_i^\ell F\|_\infty \equiv \sup_{\boldsymbol{x}\in\mathbb{R}^M} |\partial_i^\ell F(\boldsymbol{x})|$.

Problem: $F(\,\cdot\,) = \mathsf{SDP}(\,\cdot\,) \notin C_3(\mathbb{R}^M)$

Smoothing

SDP($\boldsymbol{A}$):

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \dots, \boldsymbol{\sigma}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times n} \,, \\ & \boldsymbol{\sigma}_i \in \mathbb{R}^n \,, \quad \|\boldsymbol{\sigma}_i\|_2 = 1 \,. \end{aligned}$$

Free energy

$$\Phi_k(\beta, k; A) \equiv \frac{1}{\beta} \log \left\{ \int \exp\left(\beta \sum_{i,j=1}^{n} A_{i,j} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle\right) \nu_{0,k}(\mathrm{d}\boldsymbol{\sigma}) \right\}$$

- $\nu_{0,k}(\mathrm{d}\boldsymbol{\sigma}) \equiv$ uniform measure on $S^{k-1} \times \cdots \times S^{k-1}$
- Control $\beta \to \infty$, $k \to \infty$

Smoothing

SDP($\boldsymbol{A}$):

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \dots, \boldsymbol{\sigma}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times n} \,, \\ & \boldsymbol{\sigma}_i \in \mathbb{R}^n \,, \quad \|\boldsymbol{\sigma}_i\|_2 = 1 \,. \end{aligned}$$

Free energy

$$\Phi_k(\beta, k; \boldsymbol{A}) \equiv \frac{1}{\beta} \log \left\{ \int \exp\left(\beta \sum_{i,j=1}^{n} A_{i,j} \langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle \right) \nu_{0,k}(\mathrm{d}\boldsymbol{\sigma}) \right\}$$

- $\nu_{0,k}(\mathrm{d}\boldsymbol{\sigma}) \equiv$ uniform measure on $S^{k-1} \times \cdots \times S^{k-1}$
- Control $\beta \to \infty$, $k \to \infty$

Conclusion

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, . . .]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!