# Computational Barriers to Estimation from Low-Degree Polynomials

Tselil Schramm (Stanford)

with Alex Wein (NYU)

# High-dimensional statistics


Can fast algorithms use noisy data to give solid conclusions?

information - computation tradeoff

# Simple planted models

## We (simply, but faithfully) model our data as signal + noise

e.g. the spiked Wigner matrix model

observe: $M = \lambda \cdot uu^\top + G \in \mathbb{R}^{d\times d}$, with i.i.d. $G_{ij} = G_{ji} \sim N\left(0, \frac{1}{d}\right)$

goal(s):

- **detection**: Is the signal $u$ present? (hypothesis testing with null hypothesis $M \sim N\left(0, \frac{1}{d}\right)^{d\times d}$)
- **estimation/recovery**: find $\hat{u}$ with $\|u - \hat{u}\|$ as small as possible

Several variations on a theme

- Sparse PCA: i.i.d. $u_i \sim \sqrt{\frac{1}{k}} \cdot \mathrm{Ber}\left(\frac{k}{d}\right) \cdot \mathrm{Rad}\left(\frac{1}{2}\right)$
- Planted Dense Submatrix: i.i.d. $u_i \sim \sqrt{\frac{1}{k}} \cdot \mathrm{Ber}\left(\frac{k}{d}\right)$

**These models are *famous*!**

[Johnstone-Lu'09, Baik-Ben-Arous-Peche'05, Brennan-Bresler'19, Chen-Xu'16, Ding-Kunisky-Wein-Bandeira'19, Deshpande-Montanari'14,Holtzman-Soffer-Vilenchik'20, Butucea-Ingster'13, Barbier-Macris-Rush'20, plus **dozens more**…]

# Information-computation gaps

When are detection and recovery possible (with fast algorithms)?

detection + recovery impossible

detection + recovery are possible

SPCA: both appear to be hard

PDS: detection has fast algs, recovery does not

Easy

$\tilde{o}\left(\frac{1}{\sqrt{d}}\right)$ $\tilde{O}(1)$

signal-to-noise parameter: $\frac{\sqrt{d}}{k}$

(|signal entry|/|noise entry|)

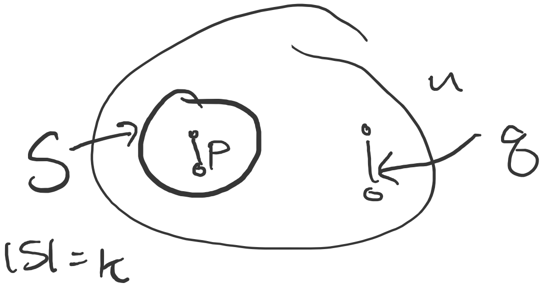
$$M = \frac{1}{100} u u^\top + G \in \mathbb{R}^{d \times d},\ G_{ij} = G_{ji} \sim N\left(0, \frac{1}{d}\right)$$

Sparse PCA: $u_i \sim \sqrt{\frac{1}{k}} \cdot \mathrm{Ber}\left(\frac{k}{d}\right) \cdot \mathrm{Rad}\left(\frac{1}{2}\right)$

Planted Dense Submatrix: $u_i \sim \sqrt{\frac{1}{k}} \cdot \mathrm{Ber}\left(\frac{k}{d}\right)$

[Johnstone-Lu'09],[Paul-Johnstone'12],[Cai-Ma-Wu'13],[Lesieur-Krazakala-Zdeborova'15],
[Kolar-Balakrishnan-Rinaldo-Singh'11],[Butucea-Ingster-Suslina'15],[Chen-Xu'16],...

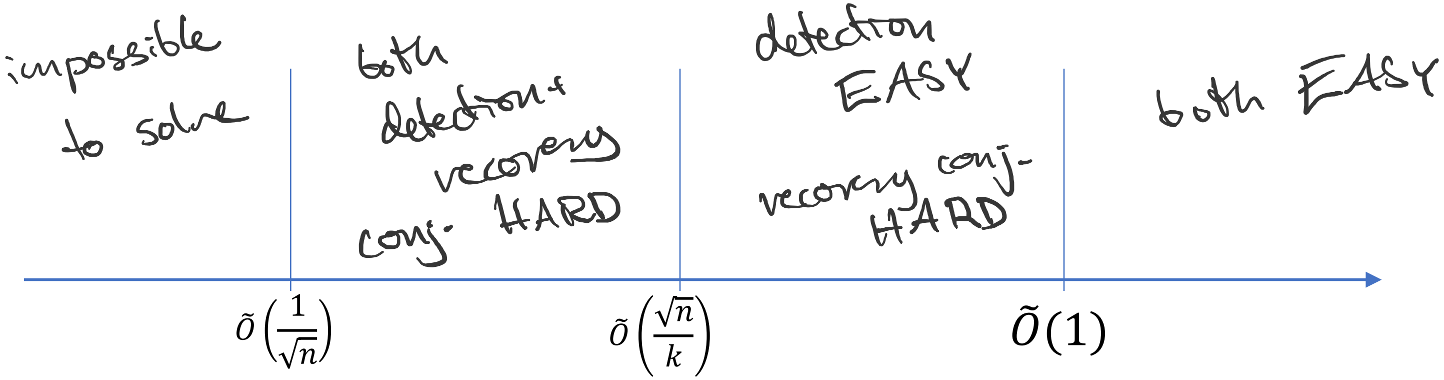
# Another simple model




**Planted dense subgraph:**

Observe a graph $G = (V, E)$ on $\cancel{d}$ $n$ vertices with a hidden subgraph on $S \subset V$, $|S| = k$ where

$$\Pr[(i,j) \in E(G)] = \begin{cases} p & \text{if } i,j \in S \\ q & \text{otherwise} \end{cases}$$




signal-to-noise parameter: $\frac{pk}{\sqrt{qn}}$

(avg degree in S / std. dev. of degree in G)

[Bhaskara-Charikar-Chlamtac-Feige-Vijayraghavan’10], [Ames’13], [Hajek-Wu-Xu’15], [ Verzelen-Arias-Castro’15], [Chen-Xu’16], …

# Explaining intractability

Planted Dense Subgraph

$\boxed{p=1,\ q=\frac{1}{2}} \not\to \boxed{p=1,\ q=1/4}$

Can we give rigorous evidence for computational barriers?

HARD PROBLEM $\to$ OUR PROBLEM

1. Reductions

   [Ma-Wu’15], [Chen-Xu’16], [Brennan-Bresler’19], [Brennan-Bresler-Huleihel’19], etc…

2. Lower bounds against restricted models of computation

   statistical query lower bounds, convex program (sum-of-squares) lower bounds, approximate message passing/belief propagation lower bounds, “energy barriers”, low-degree polynomials

# Low-degree polynomials

f : Graph adj matrix → YES/1 if graph has clique of size ≥ k
NO/0 o/w

Classical complexity theory: $f: \{0,1\}^n \to \{0,1\}$.

What is the degree of $f$ as a polynomial (over $\mathbb{R}$, $\mathbb{F}_2$)?

How well can a degree-$d$ polynomial approximate $f$?

degree-d function in time $n^d$

Complexity of statistics:

How well can a degree-$d$ polynomial detect/estimate?

f: data → conclusion

# Low-degree polynomials in high-dimensional statistics

$M$ ,  spectral gap  $M = \lambda u u^T + G$

$\lambda > (1+\varepsilon)\lambda_{max}(G)$

## Why low-degree polynomials?

$u \propto M^{\log n/\varepsilon} g \leftarrow$ random vector

- Many algorithms are (approximately) $\underbrace{\text{low-degree}}_{O(\log n)}$

  e.g. many spectral algorithms, message passing, [folklore]
  "reasonable" statistical query algorithms [Brennan-Bresler-Hopkins-Li-**S**'20],
  sum-of-squares semidefinite programs ? [Barak-Hopkins-Kelner-Kothari-Moitra-Potechin'16]

**Conclusion:** if we rule out low-degree polynomials, it is unlikely that other go-to algorithms work

- Accurately predict current computational thresholds for detection!

  degree $\omega(\log n)$ required above known computational threshold

# Computational barriers to *detection* from low-degree polynomials

planted: $\lambda uu^T + G$

null: $G$

Predictions consistent with the detection threshold for many problems:

Planted Clique [Barak-Hopkins-Kelner-Kothari-Moitra-Potechin'16], sparse PCA [Ding-Bandeira-Kunisky-Wein'19], tensor PCA [Bandeira-Kunisky-Wein'19], community detection in block models [Hopkins-Steurer'17], random CSPs…

Convenient closed form solution for detection when testing against null measure with product structure (e.g. sparse PCA, null = $N(0, \mathrm{I}_d)$)

**But** for some problems, we observe a detection-recovery gap

planted dense submatrix, planted dense subgraph, overcomplete tensor decomposition, graph matching in mildly correlated random graphs,…

# Our results: computational barriers to estimation from low-degree polynomials

$f(d) = \hat{u}$

$M = \lambda u u^T + G$

$f: \mathbb{R}^{d \times d} \to \mathbb{R}^d$

**Theorem** (Planted Submatrix):
If $\sqrt{d}/k \leq d^{-\delta}$, no degree-$O(d^{\varepsilon(\delta)})$ polynomial outperforms the trivial estimator
(which guesses every coordinate is equally likely to be in the support)

resolve open problems from
[Kolar-Balakrishnan-Rinaldo-Singh'11], [Ma-Wu'15], [Chen-Xu'16]

**Theorem** (Planted Dense Subgraph):
If $pk/\sqrt{qn} \leq n^{-\delta}$, no degree-$O(n^{\varepsilon(\delta)})$ polynomial outperforms the trivial estimator
(which guesses every coordinate is equally likely to be in the subgraph)

Both derive from more general result characterizing the minimum mean squared error of best degree-$\ell$ estimator in additive Gaussian or binary observation model.

# Outline

- Background: degree lower bounds for detection
- Degree lower bounds for estimation
  - Framework
  - Results for additive gaussian and binary observation models

# Degree lower bounds for detection (background)

Setting: "null" model $Q$ (e.g. $N(0, \mathrm{I}_n)$) and "planted" model $P$ (e.g. $\int N(\mu, \mathrm{I}_n) d\mu$)

Goal: find $f \in \mathbb{R}[x]^{\leq \ell}$ with $\left|\mathbb{E}_P f - \mathbb{E}_Q f\right| \gg \sqrt{\max(\mathbb{V}_Q f, \mathbb{V}_P f)}$

Convex program:

Strategy: upper bound

Lemma: the optimizer is $f^* \propto \left(\frac{dP}{dQ} - 1\right)^{\leq \ell}$, with objective value $\left\|\left(\frac{dP}{dQ} - 1\right)^{\leq \ell}\right\|_Q$.

$$\max_{f \in \mathbb{R}[x]^{\leq \ell}} \mathbb{E}_P f$$

$$\text{s.t. } \mathbb{E}_Q f = 0,$$

$$\mathbb{V}_Q f \leq 1,$$

$$\cancel{\mathbb{V}_P f \leq 1}.$$

Conclusion: if $\left\|\left(\frac{dP}{dQ} - 1\right)^{\leq \ell}\right\|_Q = o(1)$, then there is no degree-$\ell$ polynomial which meets our criteria

(no tests with difference in mean larger than the standard deviation)

# Degree lower bounds for detection

Strategy: compute $\left\| \left( \frac{dP}{dQ} - 1 \right)^{\leq \ell} \right\|_Q$, rule out degree-$\ell$ polynomials if $o(1)$.

This framework was used to give consistent detection lower bounds for many problems.

Planted clique [Barak-Hopkins-Kelner-Kothari-Moitra-Potechin'16], sparse PCA [Ding-Bandeira-Kunisky-Wein'19], tensor PCA [Bandeira-Kunisky-Wein'19], community detection in block models [Hopkins-Steurer'17], random CSPs…

Successful when $\mathbb{R}[x]^{\leq \ell}$ has a nice orthogonal (w.r.t $\langle \cdot, \cdot \rangle_Q$) basis (e.g. product measures).

Also, used to give new algorithms! (Evaluate $\left( \frac{dP}{dQ} - 1 \right)^{\leq \ell}$ and threshold)

Overlapping block models [Hopkins-Steurer'17], graph matching [Barak-Chou-Lei-**S**-Sheng'19]

# Outline

- Background: degree lower bounds for detection
- Degree lower bounds for estimation
  - Framework
  - Results for additive gaussian and binary observation models

# Degree lower bounds for estimation

$g(x) = g(x) \in \mathbb{R}^n$

Setting: "planted" model $P$ (e.g. $\int N(\mu, \mathrm{I}_n) d\mu$)

Goal: find $g \in \left(\mathbb{R}[x]^{\leq \ell}\right)^{\otimes n}$ with $\mathbb{E}_{(\mu,x)\sim P} \|g(x) - \mu\|^2 = o(\mathbb{E}_{\mu \sim P} \|\mu\|^2)$

degree-$\ell$ MMSE:

$$\min_{g \in \left(\mathbb{R}[x]^{\leq \ell}\right)^{\otimes n}} \mathbb{E}_P \|g(x) - \mu\|^2$$

scaling

$$\max_{h \in \left(\mathbb{R}[x]^{\leq \ell}\right)^{\otimes n}} \mathbb{E}_P \langle h(x), \mu \rangle$$

$$\text{s.t. } \mathbb{E}_P \|h(x)\|^2 \leq 1$$

symmetry

$$\boxed{\begin{aligned} \mathrm{OPT}_i &= \max_{h_i} \langle h_i, \mu_i^{\leq \ell} \rangle_P \\ \text{s.t. } &\|h_i\|_P \leq 1 \end{aligned}}$$

Lemma: $\mathrm{OPT}_\mathrm{i} = \left\| (\mu_i)^{\leq \ell} \right\|_P$.

Conclusion: if for all $i$, $\mathrm{OPT}_\mathrm{i} \leq \sqrt{\alpha \cdot \mathbb{E}_P \mu_i^2}$, error $\geq (1-\alpha) \cdot \mathbb{E}_P \|\mu\|^2$.

# A familiar story?

So… why not just compute $\left\|(\mu_i)^{\leq \ell}\right\|_P$ for each $i$? (like for detection)

For planted distributions of interest, $\mathbb{R}[x]^{\leq \ell}$ has a nasty orthogonal basis w.r.t $\langle \cdot,\cdot \rangle_P$!

Compare to hypothesis testing of $Q = N(0, \mathrm{I})$ vs. $P = \mathbb{E}_\mu N(\mu, \mathrm{I})$.

$$\mathrm{OPT}_i = \max_{h_i} \left\langle h_i, \mu_i^{\leq \ell} \right\rangle_P \quad \text{s.t. } \|h_i\|_P \leq 1$$

change-of-basis

$$\mathrm{OPT}_i = \max_{f_i} \left\langle f_i, \left(\frac{dP}{dQ}\mu_i\right)^{\leq \ell} \right\rangle_Q \quad \text{s.t. } \|B f_i\|_Q \leq 1$$

closed form

$$\mathrm{OPT}_i = \left\| (B^{\cancel{-1}})^\top \left(\frac{dP}{dQ}\mu_i\right)^{\leq \ell} \right\|_Q$$

# A solution for additive Gaussian models

Suppose $P = \mathbb{E}_\mu N(\mu, \mathrm{I})$.

For $x = \mu + G \sim P$, think of first sampling the signal $\mu$, then the noise $G \sim Q = N(0, \mathrm{I})$.

$$\forall i \in [n], \mathrm{OPT}_i = \max_{h_i \in \mathbb{R}[x]^{\leq \ell}} \mathbb{E}_{x\sim P} h_i(x) \mu_i$$
$$\text{s.t. } \mathbb{E}_{x\sim P} h_i(x)^2 \leq 1$$

relaxation, by Jensen

$$\forall i \in [n], \mathrm{rOPT}_i = \max_{h_i \in \mathbb{R}[x]^{\leq \ell}} \mathbb{E}_{\mu,G} h_i(\mu + G) \mu_i$$
$$\text{s.t. } \mathbb{E}_G \left[\mathbb{E}_\mu h_i(\mu + G)\right]^2 \leq 1$$

intuitively: not too lossy when recovery is impossible

Let $f_i(G) = \mathbb{E}_\mu h_i(\mu + G)$; in additive Gaussian models, we can write $f_i = A h_i$ for $A$ **upper triangular**.

$\forall i \in [n], \mathrm{rOPT}_i = \|(A^{-1})^T c_i\|_Q$ is tractable to compute, and we get a closed form.

# Estimation lower bounds in additive Gaussian + Binary observation models

Closed form for $\mathrm{OPT}_i$ for additive Gaussian planted models

in which we observe $x \sim P = \mathbb{E}_\mu N(\mu, \mathrm{I}_n)$

Also for *Binary observation* models

in which $\mu \sim D([0,1]^n)$, we observe $x_i \sim \mathrm{Ber}(\mu_i)$ $\forall i$ (e.g. planted dense subgraph)

Exact expression for degree-$\ell$ MMSE.

Special case: lower bounds for estimation in *planted submatrix* and *planted dense subgraph*.

# Conclusion

tl;dr : we extend methods for lower bounding the polynomial degree of hypothesis tests to lower bounding the polynomial degree of estimators.

We give the first(ish) rigorous evidence for hardness (against a restricted class of algorithms) of planted submatrix and planted dense subgraph below known algorithmic thresholds.

## Open:

$n^{d\text{-ish}}$

- Optimal degree-vs.-estimation tradeoff?
- Estimation lower bounds for other models with detection-recovery gaps?
  (favorite problem: overcomplete tensor decomposition)
- Extending consequences to other models
  (degree lower bounds imply SoS lower bounds?)

# Thank you!