Statistical-to-Computational Gaps: The Low-Degree Method and Free-Energy Barriers

Afonso S. Bandeira
ETH Zürich

partly based on `arXiv:2205.09727[math.ST]` joint with A. El Alaoui (Cornell), S. B. Hopkins (MIT), T. Schramm (Stanford), A. S. Wein (Gatech), I. Zadik (MIT)

- *"The world's most valuable resource is no longer oil, but* **data**"

  – The Economist

- Are there **limits** to what we can learn?

- Are there **limits** to what we can learn?

- Which methods work? **Why?**

– XKCD

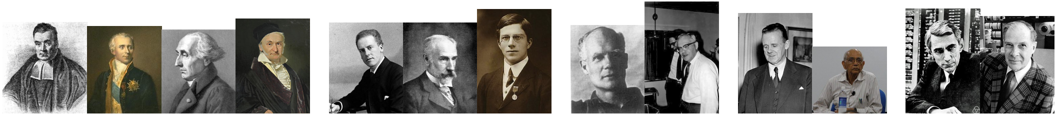
- Are there **limits** to what we can learn?
- Which methods work? **Why?**
- What are the **bottlenecks**?

– XKCD

Statistics & Information Theory — What are limits to learning?

- 1700's - **Bayesian Statistics**
- 1900-1920 - **Fisher Information**
  — How much information does a sample have about a parameter?
- 1933: **Neyman-Pearson Lemma**:
  — Limits on Hypothesis Testing
- 1940's: **Cramér-Rao Bound**:
  — Limits on Statistical Estimation
- late 1940's: **Information Theory**:
  – **Shannon Entropy:** # of bits "of information" needed to identify a draw of X
- 1950+ Minimax, Contiguity, ...

Is there enough information in the data?



Bayes 1760's | Laplace 1770's | Lagrange | Gauss 1800's | K.Pearson 1890's | Edgeworth 1900's | Fisher 1920's | E.Pearson Neyman 1930's | Cramér Rao 1940's | Shannon 1948 | Hamming 1950 | · · ·

Learning/Estimating is (also) optimization

Goal: Find parameter/signal/model that best “fits” the data

- Maximum likelihood estimation
- Training of Neural Networks
- ...

Are these computational tasks feasible/easy?

Learning/Estimating is (also) optimization

Goal: Find parameter/signal/model that best "fits" the data

- Maximum likelihood estimation
- Training of Neural Networks
- . . .

Are these computational tasks feasible/easy?

Many optimization/computational problems are **NP-hard** (e.g. Knapsack)

1956: Gödel's letter to von Neumann
(and John Nash's 1955)

1971-72: Cook and Karp's **NP-hardness**

Learning/Estimating is (also) optimization

Goal: Find parameter/signal/model that best "fits" the data

- Maximum likelihood estimation
- Training of Neural Networks
- ...

Are these computational tasks feasible/easy?

Many optimization/computational problems are **NP-hard** (e.g. Knapsack)

1956: Gödel's letter to von Neumann
(and John Nash's 1955)

1971-72: Cook and Karp's **NP-hardness**

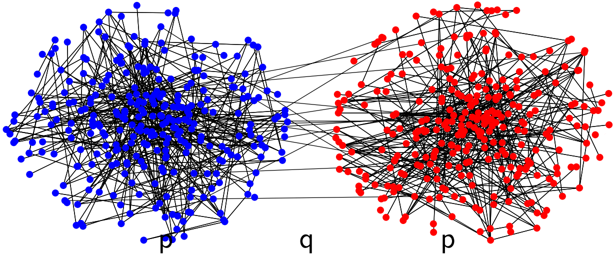
Should we design (statistical) models so that optimization is easy?

Linearity, Convexity, ...

An example: Communities in Social Networks

Given two disjoint sets of $m = \frac{n}{2}$ nodes each. Independently:

- pairs between clusters have an edge with probability p
- pairs across clusters have an edge with probability $q < p$



A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

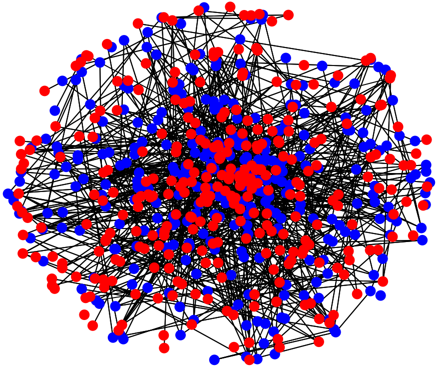
E. Mossel. J. Neeman, A. Sly, 2012, 2013.

L. Massoulie, 2013.

E. Abbe, A. S. Bandeira, G. Hall, 2014.

A. S. Bandeira, 2018.

An example: Communities in Social Networks



A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

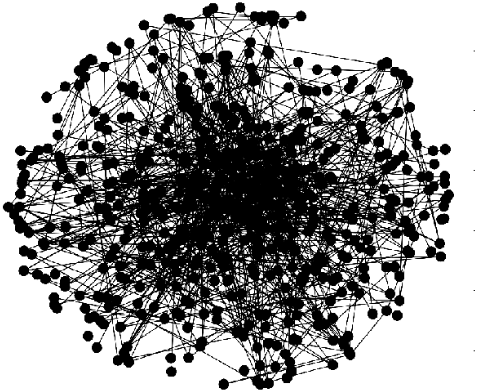
E. Mossel. J. Neeman, A. Sly, 2012, 2013.

L. Massoulie, 2013.

E. Abbe, A. S. Bandeira, G. Hall, 2014.

A. S. Bandeira, 2018.

An example: Communities in Social Networks



Can we recover the colors/partition?

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

E. Mossel. J. Neeman, A. Sly, 2012, 2013.

L. Massoulie, 2013.

E. Abbe, A. S. Bandeira, G. Hall, 2014.

A. S. Bandeira, 2018.

An example: continued

- **Theorem:** For $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$, If (iff)

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

Minimum Bisection coincides with the true communities with high probability.

E. Abbe, A. S. Bandeira, G. Hall, 2014.

E. Mossel, J. Neeman, and A. Sly, 2014

B. Hajek, Y. Wu, and J. Xu., 2014

A. S. Bandeira, 2015.

An example: continued

- **Theorem:** For $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$, If (iff)

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

  Minimum Bisection coincides with the true communities with high probability.
- **Theorem: Minimum Bisection is an NP-hard problem.**

E. Abbe, A. S. Bandeira, G. Hall, 2014.

E. Mossel, J. Neeman, and A. Sly, 2014

B. Hajek, Y. Wu, and J. Xu., 2014

A. S. Bandeira, 2015.

An example: continued

- **Theorem:** For $p = \alpha\frac{\log n}{n}$ and $q = \beta\frac{\log n}{n}$, If (iff)

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

  Minimum Bisection coincides with the true communities with high probability.

- **Theorem:** **Minimum Bisection is an NP-hard problem.**
- **Theorem:** If

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

  Minimum Bisection can be computed efficiently with high probability.

E. Abbe, A. S. Bandeira, G. Hall, 2014.

E. Mossel, J. Neeman, and A. Sly, 2014

B. Hajek, Y. Wu, and J. Xu., 2014

A. S. Bandeira, 2015.

An example: continued

- **Theorem:** For $p = \alpha\frac{\log n}{n}$ and $q = \beta\frac{\log n}{n}$, If (iff)

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

  Minimum Bisection coincides with the true communities with high probability.

- **Theorem: Minimum Bisection is an NP-hard problem.**
- **Theorem:** If

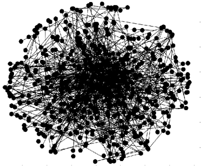
$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2},$$

  Minimum Bisection can be computed efficiently with high probability.

Does this always happen?

E. Abbe, A. S. Bandeira, G. Hall, 2014.
E. Mossel, J. Neeman, and A. Sly, 2014
B. Hajek, Y. Wu, and J. Xu., 2014
A. S. Bandeira, 2015.

Statistical-to-Computational Gaps

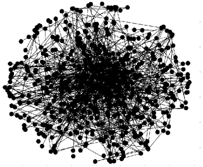


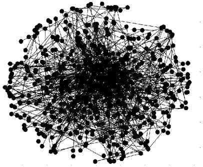
Hidden Clique Problem

- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

Statistical-to-Computational Gaps

Hidden Clique Problem





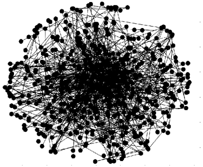
- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

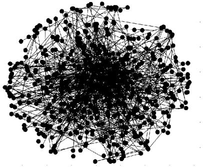
Vs

- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — **k**-clique added at random

Statistical-to-Computational Gaps

Hidden Clique Problem





- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

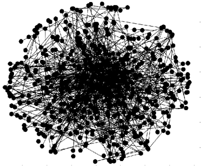
Vs

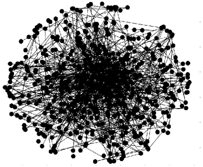
- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — **k**-clique added at random

$2 \log n$	**Largest Clique**	**k**

Statistical-to-Computational Gaps

Hidden Clique Problem





- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

Vs

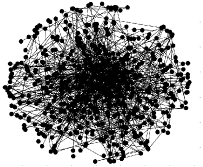
- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — $\mathbf{k}$-clique added at random

Largest Clique

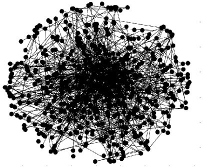
$2 \log n$ | $\mathbf{k}$

- *Alon-Krivelevich-Sudakov '98:* Efficient algorithm for $\mathbf{k} \gtrsim \sqrt{n}$ (as opposed to $\mathbf{k} > 2 \log n$)

Statistical-to-Computational Gaps



Hidden Clique Problem



- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

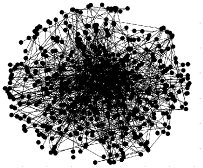
Vs

- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — **k**-clique added at random

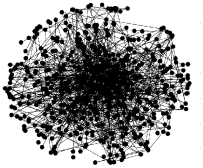
$2 \log n$ **Largest Clique** **k**

- *Alon-Krivelevich-Sudakov '98:* Efficient algorithm for $\mathbf{k} \gtrsim \sqrt{n}$ (as opposed to $\mathbf{k} > 2 \log n$)
- No improvement since; believed to be **hard** and used as reduction primitive (e.g. Berthet-Rigollet '12)

Statistical-to-Computational Gaps



Hidden Clique Problem



- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

Vs

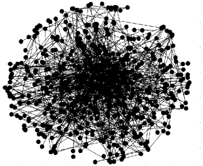
- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — **k**-clique added at random

	Largest Clique	
$2 \log n$		$\mathbf{k}$

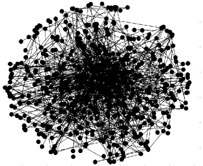
- *Alon-Krivelevich-Sudakov '98:* Efficient algorithm for $\mathbf{k} \gtrsim \sqrt{n}$ (as opposed to $\mathbf{k} > 2 \log n$)
- No improvement since; believed to be **hard** and used as reduction primitive (e.g. Berthet-Rigollet '12)

Statistical-to-Computational Gap "Hypothesis"

Statistical-to-Computational Gaps



Hidden Clique Problem



- A graph $G\left(n, \frac{1}{2}\right)$
  — each edge appears with probability $\frac{1}{2}$

Vs

- $G\left(n, \frac{1}{2}\right) + \mathbf{k}$-clique
  — **k**-clique added at random

$2 \log n$	**Largest Clique**	**k**

- *Alon-Krivelevich-Sudakov '98:* Efficient algorithm for $\mathbf{k} \gtrsim \sqrt{n}$ (as opposed to $\mathbf{k} > 2 \log n$)
- No improvement since; believed to be **hard** and used as reduction primitive (e.g. Berthet-Rigollet '12)

Statistical-to-Computational Gap "Hypothesis"

What Makes a Problem Hard?

$\mathbb{P}($ node colors $\mid$ SBM Graph $) \quad \leftrightarrow \quad$ Spin Glass (Physics)

Complexy of Posterior / **Geometry of Solutions** / **Free-Energy Overlap Barrier**

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

D. Gamarnik, M. Sudan, 2013

S. Hopkins, D. Steurer, 2017

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019

What Makes a Problem Hard?

$\mathbb{P}($ node colors $|$ SBM Graph $) \quad \leftrightarrow \quad$ Spin Glass (Physics)

Complexy of Posterior / Geometry of Solutions / Free-Energy Overlap Barrier

What if one can only use low-degree polynomials of the data? (Hopkins-Steurer '17)

Restricted class of algorithms / Low-Degree Method

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

D. Gamarnik, M. Sudan, 2013

S. Hopkins, D. Steurer, 2017

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019

What Makes a Problem Hard?

$\mathbb{P}(\text{ node colors} \mid \text{SBM Graph })$ $\leftrightarrow$ Spin Glass (Physics)

Complexy of Posterior / **Geometry of Solutions** / **Free-Energy Overlap Barrier**

What if one can only use low-degree polynomials of the data? (Hopkins-Steurer '17)

Restricted class of algorithms / Low-Degree Method

Goal: When are different approaches equivalent?

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, 2011

D. Gamarnik, M. Sudan, 2013

S. Hopkins, D. Steurer, 2017

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019

Planted Model (parameter $n \to \infty$)

- **Signal:** $x \sim \mu$ drawn from prior μ
- **Data:** $Y \sim \mathbb{P}_x := \mathbb{P}(Y|x)$

Planted Model (parameter $n \to \infty$)

- **Signal:** $x \sim \mu$ drawn from prior μ
- **Data:** $Y \sim \mathbb{P}_x := \mathbb{P}(Y|x)$
- **Goal:** Estimate x from Y

Planted Model (parameter $n \to \infty$)

- **Signal:** $x \sim \mu$ drawn from prior μ
- **Data:** $Y \sim \mathbb{P}_x := \mathbb{P}(Y|x)$
- **Goal:** Estimate x from Y

Hypothesis Testing

- Planted model (with "signal") $Y \sim \mathbb{P}_n$ $Y \sim \mathbb{P}_x$ with $x \sim \mu$
- Null model (just noise) $Y \sim \mathbb{Q}_n$ e.g. $G(n, 1/2)$

Planted Model (parameter $n \to \infty$)

- **Signal:** $x \sim \mu$ drawn from prior μ
- **Data:** $Y \sim \mathbb{P}_x := \mathbb{P}(Y|x)$
- **Goal:** Estimate x from Y

Hypothesis Testing

- Planted model (with "signal") $Y \sim \mathbb{P}_n$ $Y \sim \mathbb{P}_x$ with $x \sim \mu$
- Null model (just noise) $Y \sim \mathbb{Q}_n$ e.g. $G(n, 1/2)$

Le Cam Contiguity

If $\mathbb{E}_{\mathbb{Q}}\, L(Y)^2 = O(1)$, where $L(Y) := \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$, then H.T. with $o(1)$ error probability is impossible

Planted Model (parameter $n \to \infty$)

- **Signal:** $x \sim \mu$ drawn from prior μ
- **Data:** $Y \sim \mathbb{P}_x := \mathbb{P}(Y|x)$
- **Goal:** Estimate x from Y

Hypothesis Testing

- Planted model (with "signal") $\quad Y \sim \mathbb{P}_n \quad Y \sim \mathbb{P}_x$ with $x \sim \mu$
- Null model (just noise) $\quad Y \sim \mathbb{Q}_n \quad$ e.g. $G(n, 1/2)$

Le Cam Contiguity

If $\mathbb{E}_{\mathbb{Q}} L(Y)^2 = O(1)$, where $L(Y) := \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$, then H.T. with $o(1)$ error probability is impossible

Proof: Let A be the event where test says "Planted Model".

$$\mathbb{P}(A) = \mathbb{E}_{\mathbb{P}} 1_A(Y) = \mathbb{E}_{\mathbb{Q}} L(Y) 1_A(Y) \leq \sqrt{\mathbb{E}_{\mathbb{Q}} L(Y)^2} \sqrt{\mathbb{E}_{\mathbb{Q}} 1_A(Y)^2} = \sqrt{\mathbb{E}_{\mathbb{Q}} L(Y)^2} \sqrt{\mathbb{Q}(A)}$$

The Low-Degree Method (e.g. [Hopkins-Steurer '17])

Goal: *Hypothesis Testing* between two distributions with $o(1)$ error:

S. Hopkins, D. Steurer, 2017

S. Hopkins, 2018 (PhD thesis)

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

The Low-Degree Method (e.g. [Hopkins-Steurer '17])

Goal: *Hypothesis Testing* between two distributions with $o(1)$ error:

- Null model (just noise) $Y \sim \mathbb{Q}_n$ e.g. $G(n, 1/2)$
- Planted model (with "signal") $Y \sim \mathbb{P}_n$ e.g. $G(n, 1/2) \cup \{\text{random } k\text{-clique}\}$

S. Hopkins, D. Steurer, 2017

S. Hopkins, 2018 (PhD thesis)

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

The Low-Degree Method (e.g. [Hopkins-Steurer '17])

Goal: *Hypothesis Testing* between two distributions with $o(1)$ error:

- Null model (just noise) $Y \sim \mathbb{Q}_n$ e.g. $G(n, 1/2)$
- Planted model (with "signal") $Y \sim \mathbb{P}_n$ e.g. $G(n, 1/2) \cup \{\text{random } k\text{-clique}\}$

Is there a low degree polynomial $f(Y)$ that is big when $Y \sim \mathbb{P}$ and close to zero when $Y \sim \mathbb{Q}$?

S. Hopkins, D. Steurer, 2017

S. Hopkins, 2018 (PhD thesis)

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

The Low-Degree Method (e.g. [Hopkins-Steurer '17])

Goal: *Hypothesis Testing* between two distributions with $o(1)$ error:

- Null model (just noise) $Y \sim \mathbb{Q}_n$ e.g. $G(n, 1/2)$
- Planted model (with "signal") $Y \sim \mathbb{P}_n$ e.g. $G(n, 1/2) \cup \{\text{random } k\text{-clique}\}$

Is there a low degree polynomial $f(Y)$ that is big when $Y \sim \mathbb{P}$ and close to zero when $Y \sim \mathbb{Q}$?

IDEA: Compute $\max\limits_{f \text{ deg } D} \dfrac{\mathbb{E}_{Y \sim \mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)^2]}}$ $\qquad \dfrac{\text{mean in } \mathbb{P}}{\text{fluctuations in } \mathbb{Q}}$

S. Hopkins, D. Steurer, 2017

S. Hopkins, 2018 (PhD thesis)

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y \sim \mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)^2]}}$$

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{Q}}[L(Y)f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$

Likelihood ratio: $L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{Q}}[L(Y)f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$

Likelihood ratio: $L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$

$$\langle f, g\rangle = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)g(Y)]$$

$$\|f\| = \sqrt{\langle f, f\rangle}$$

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$
$$= \max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{Q}}[L(Y)f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$
$$= \max_{f \text{ deg } D} \frac{\langle L, f\rangle}{\|f\|}$$

Likelihood ratio: $L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$

$\langle f, g\rangle = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)g(Y)]$

$\|f\| = \sqrt{\langle f, f\rangle}$

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y \sim \mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)^2]}} = \max_{f \text{ deg } D} \frac{\mathbb{E}_{Y \sim \mathbb{Q}}[L(Y) f(Y)]}{\sqrt{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)^2]}} = \max_{f \text{ deg } D} \frac{\langle L, f \rangle}{\|f\|}$$

Likelihood ratio: $L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$

$\langle f, g \rangle = \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y) g(Y)]$

$\|f\| = \sqrt{\langle f, f \rangle}$

Maximizer: $f = L^{\leq D} :=$ projection of L onto degree-D subspace

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$

$$= \max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{Q}}[L(Y)f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}}$$

$$= \max_{f \text{ deg } D} \frac{\langle L, f\rangle}{\|f\|}$$

$$= \|\mathbf{L}^{\leq \mathbf{D}}\|$$

Likelihood ratio: $L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y)$

$$\langle f, g\rangle = \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)g(Y)]$$

$$\|f\| = \sqrt{\langle f, f\rangle}$$

Maximizer: $f = L^{\leq D}$:= projection of L onto degree-D subspace

Norm of low-degree likelihood ratio

The Low-Degree Method

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y \sim \mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)^2]}} = \|L^{\leq D}\|$$

The Low-Degree Method

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \|L^{\leq D}\|$$

Heuristically $\quad \|L^{\leq D}\| = \begin{cases} \xrightarrow[n]{} \infty & \text{degree-}D \text{ polynomial can distinguish } \mathbb{Q}, \mathbb{P} \\ O(1) & \text{degree-}D \text{ polynomials fail} \end{cases}$

The Low-Degree Method

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \|L^{\leq D}\|$$

Heuristically $\quad \|L^{\leq D}\| = \begin{cases} \xrightarrow[n]{} \infty & \text{degree-}D \text{ polynomial can distinguish } \mathbb{Q},\mathbb{P} \\ O(1) & \text{degree-}D \text{ polynomials fail} \end{cases}$

Conjecture (informal variant of [Hopkins '18])

For "nice" $\mathbb{Q},\mathbb{P}$, *if* $\|L^{\leq D}\| = O(1)$ *for some* $D \gg \log n$ *then no polynomial-time algorithm can distinguish* $\mathbb{Q},\mathbb{P}$ *with success probability* $1 - o(1)$.

The Low-Degree Method

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \|L^{\leq D}\|$$

Heuristically $\quad \|L^{\leq D}\| = \begin{cases} \underset{n}{\to} \infty & \text{degree-}D \text{ polynomial can distinguish } \mathbb{Q}, \mathbb{P} \\ O(1) & \text{degree-}D \text{ polynomials fail} \end{cases}$

Conjecture (informal variant of [Hopkins '18])

For "nice" $\mathbb{Q}, \mathbb{P}$, if $\|L^{\leq D}\| = O(1)$ for some $D \gg \log n$ then no polynomial-time algorithm can distinguish $\mathbb{Q}, \mathbb{P}$ with success probability $1 - o(1)$.

degree-D polynomials $\iff$ time-$n^{\tilde{\Theta}(D)}$ algorithms

The Low-Degree Method

$$\max_{f \text{ deg } D} \frac{\mathbb{E}_{Y\sim\mathbb{P}}[f(Y)]}{\sqrt{\mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)^2]}} = \|L^{\le D}\|$$

Heuristically $\quad \|L^{\le D}\| = \begin{cases} \xrightarrow[n]{} \infty & \text{degree-}D \text{ polynomial can distinguish } \mathbb{Q}, \mathbb{P} \\ O(1) & \text{degree-}D \text{ polynomials fail} \end{cases}$

Conjecture (informal variant of [Hopkins '18])

For "nice" $\mathbb{Q}, \mathbb{P}$, if $\|L^{\le D}\| = O(1)$ for some $D \gg \log n$ then no polynomial-time algorithm can distinguish $\mathbb{Q}, \mathbb{P}$ with success probability $1 - o(1)$.

degree-D polynomials $\iff$ time-$n^{\tilde{\Theta}(D)}$ algorithms

- If $\|L^{\le D}\| = O(1)$ for some $D \gg \log n$ then no spectral method can distinguish $\mathbb{Q}$ from $\mathbb{P}$ (in a particular sense) [Kunisky, Wein, B '19]
- Spectral methods are believed to be as powerful as sum-of-squares for average-case problems [HKPRSS '17]
  - e.g. Predicts exact sub-exponential computational cost of sparse PCA [Ding, Kunisky, Wein, B. '19]

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} L(Y)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} \left(\mathop{\mathbb{E}}_{x\sim\mu} L_x(Y) \right)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} \mathop{\mathbb{E}}_{x\sim\mu} \mathop{\mathbb{E}}_{x'\sim\mu} L_x(Y) L_{x'}(Y) = \mathop{\mathbb{E}}_{x\sim\mu} \mathop{\mathbb{E}}_{x'\sim\mu} \langle L_x, L_{x'} \rangle_{L^2(\mathbb{Q})}$$

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} L^{\leq D}(Y)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} \left(\mathop{\mathbb{E}}_{x\sim\mu} L_x^{\leq D}(Y) \right)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} \mathop{\mathbb{E}}_{x\sim\mu} \mathop{\mathbb{E}}_{x'\sim\mu} L_x^{\leq D}(Y) L_{x'}^{\leq D}(Y) = \mathop{\mathbb{E}}_{x\sim\mu} \mathop{\mathbb{E}}_{x'\sim\mu} \langle L_x^{\leq D}, L_{x'}^{\leq D} \rangle_{L^2(\mathbb{Q})}$$

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} L^{\leq D}(Y)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}\left(\mathop{\mathbb{E}}_{x\sim\mu} L_x^{\leq D}(Y)\right)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}\mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{x'\sim\mu} L_x^{\leq D}(Y) L_{x'}^{\leq D}(Y) = \mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{x'\sim\mu}\langle L_x^{\leq D}, L_{x'}^{\leq D}\rangle_{L^2(\mathbb{Q})}$$

Gaussian Additive Model: $\mathbb{P} : \mathbf{Y} = \lambda\mathbf{X} + \mathbf{Z}$ vs $\mathbb{Q} : \mathbf{Y} = \mathbf{Z}$

$\mathbf{X} \sim \mu$, distribution over $\mathbb{R}^N$ and $\mathbf{Z} \in \mathbb{R}^N$ i.i.d. $\mathcal{N}(0,1)$

$$\mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} L^{\leq D}(Y)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}} \left(\mathop{\mathbb{E}}_{x\sim\mu} L_x^{\leq D}(Y)\right)^2 = \mathop{\mathbb{E}}_{Y\sim\mathbb{Q}}\mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{x'\sim\mu} L_x^{\leq D}(Y) L_{x'}^{\leq D}(Y) = \mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{x'\sim\mu} \langle L_x^{\leq D}, L_{x'}^{\leq D}\rangle_{L^2(\mathbb{Q})}$$

Gaussian Additive Model: $\mathbb{P}: \mathbf{Y} = \lambda\mathbf{X} + \mathbf{Z}$ vs $\mathbb{Q}: \mathbf{Y} = \mathbf{Z}$

$\mathbf{X} \sim \mu$, distribution over $\mathbb{R}^N$ and $\mathbf{Z} \in \mathbb{R}^N$ i.i.d. $\mathcal{N}(0,1)$

$$L(Y) = \frac{d\mathbb{P}}{d\mathbb{Q}}(Y) = \frac{\mathbb{E}_X \exp(-\frac{1}{2}\|Y-\lambda X\|^2)}{\exp(-\frac{1}{2}\|Y\|^2)} = \mathop{\mathbb{E}}_X \exp\left(\lambda\langle Y, X\rangle - \frac{1}{2}\lambda^2\|X\|^2\right)$$

$L = \sum_\alpha c_\alpha h_\alpha$ where $\{h_\alpha\}$ are the *Hermite polynomials* *(orthonormal basis w.r.t. $\mathbb{Q}$)*

$\|L^{\leq D}\|^2 = \sum_{|\alpha|\leq D} c_\alpha^2$ where $c_\alpha = \langle L, h_\alpha\rangle = \mathbb{E}_{Y\sim\mathbb{Q}}[L(Y)h_\alpha(Y)] = \mathbb{E}_{Y\sim\mathbb{P}}[h_\alpha(Y)]$

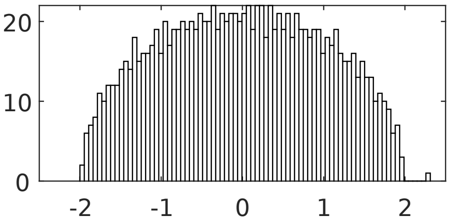
Result: $\|\mathbf{L}^{\leq \mathbf{D}}\|^2 = \sum_{d=0}^{D} \frac{1}{d!} \mathop{\mathbb{E}}_{X,X'}[\lambda^{2d}\langle X, X'\rangle^d] = \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}'} \exp^{\leq \mathbf{D}}(\lambda^2\langle \mathbf{X}, \mathbf{X}'\rangle)$

Example: **Wigner Spike Model**

$$Y = \lambda v v^T + W$$

W Wigner, $\|v\| = 1$.

$W_{ij} \sim \mathcal{N}(0, 1/n)$



Goal: Recover/Detect v from $\lambda v v^T + W$

Johnstone, AoS 2001.

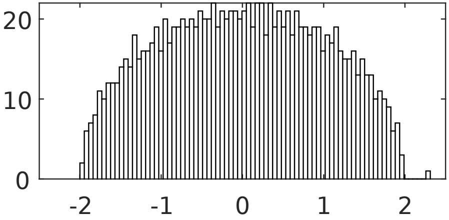
Baik, Ben-Arous, Peche, AoP 2005.

D. Feral, S. Peche, CMP 2006.

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

Example: **Wigner Spike Model**

$$Y = \lambda v v^T + W$$
W Wigner, $\|v\| = 1$.
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$



Goal: Recover/Detect v from $\lambda v v^T + W$

Visible on largest eigenvalue if

$$\lambda > \mathbf{1}$$

Johnstone, AoS 2001.

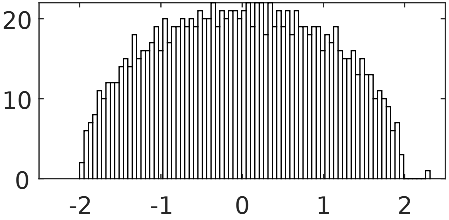
Baik, Ben-Arous, Peche, AoP 2005.

D. Feral, S. Peche, CMP 2006.

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

Example: **Wigner Spike Model**

$$Y = \lambda v v^T + W$$
$$W \text{ Wigner}, \ \|v\| = 1.$$
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$



Goal: Recover/Detect v from $\lambda v v^T + W$

Visible on largest eigenvalue if

$$\lambda > \mathbf{1}$$

For $\mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{\mathbf{n-1}})$, there is no gap

$$(\lambda^*_{STAT} = 1)$$

Johnstone, AoS 2001.

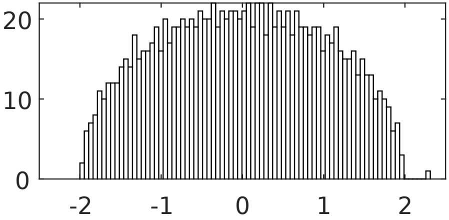
Baik, Ben-Arous, Peche, AoP 2005.

D. Feral, S. Peche, CMP 2006.

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

Example: **Wigner Spike Model**

$$Y = \lambda v v^T + W$$
$$W \text{ Wigner, } \|v\| = 1.$$
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$



Johnstone, AoS 2001.

Baik, Ben-Arous, Peche, AoP 2005.

D. Feral, S. Peche, CMP 2006.

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

Goal: Recover/Detect v from $\lambda v v^T + W$

Visible on largest eigenvalue if

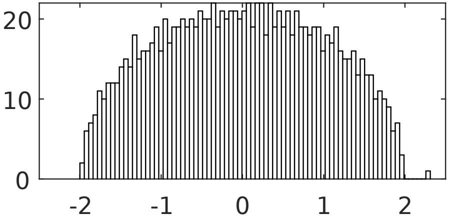
$$\lambda > \mathbf{1}$$

For $\mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{\mathbf{n-1}})$, there is no gap $(\lambda^*_{STAT} = 1)$

Sparse PCA: For other priors, such as
$\mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{\mathbf{n-1}} \cap \|\mathbf{v}\|_{\mathbf{0}} \leq \mathbf{n}/\mathbf{100})$
gaps appear $(\lambda^*_{STAT} < 1,\ \lambda^*_{COMP} = 1)$

Example: **Wigner Spike Model**

$$Y = \lambda v v^T + W$$
W Wigner, $\|v\| = 1$.
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$



Johnstone, AoS 2001.

Baik, Ben-Arous, Peche, AoP 2005.

D. Feral, S. Peche, CMP 2006.

A. S. Bandeira, D. Kunisky, A. S. Wein, 2019 (survey)

Goal: Recover/Detect v from $\lambda v v^T + W$

Visible on largest eigenvalue if

$$\lambda > \mathbf{1}$$

For $\mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{\mathbf{n-1}})$, there is no gap $(\lambda^*_{STAT} = 1)$

Sparse PCA: For other priors, such as
$\mathbf{v} \sim \mathrm{Unif}(\mathbb{S}^{\mathbf{n-1}} \cap \|\mathbf{v}\|_{\mathbf{0}} \leq \mathbf{n}/\mathbf{100})$
gaps appear $(\lambda^*_{STAT} < 1,\ \lambda^*_{COMP} = 1)$

Tensor PCA, ...

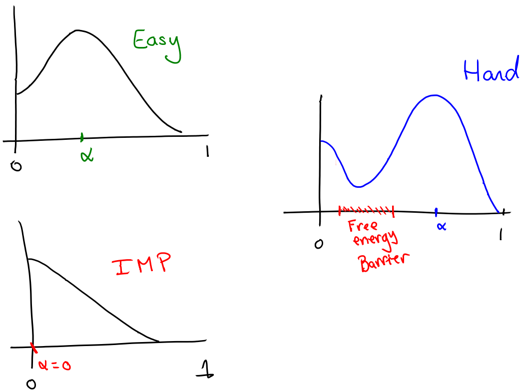
Statistical Physics and Free-Energy Barriers

- x^* ground truth
- x draw from posterior
  $x \sim \mathbb{P}(x \mid Y)$
- $q(x) \leftarrow$ measure of quality of draw
  $q(x^*) = 1$
  $q = 0 \longleftrightarrow$ random guessing
- What is the distribution of $q(x)$?

$q(x) \rightarrow \alpha > 0$ stat. possible

$q(x) \rightarrow 0$ stat impossible

Plot $\frac{1}{n} \log \mathbb{P}_Y \left(q(x) = s \right)$



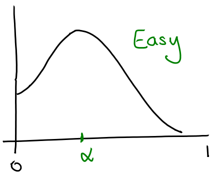
Statistical Physics and Free-Energy Barriers

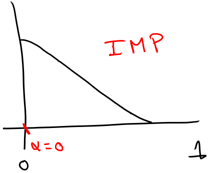
- x^* ground truth
- x draw from posterior
  $x \sim \mathbb{P}(x \mid Y)$
- $q(x) \leftarrow$ measure of quality of draw
  $q(x^*) = 1$
  $q = 0 \leftrightarrow$ random guessing
- What is the distribution of $q(x)$?

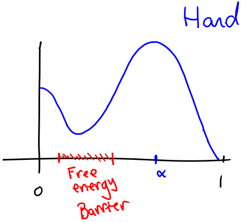
$q(x) \rightarrow \alpha > 0$ stat. possible

$q(x) \rightarrow 0$ stat impossible

Plot $\frac{1}{n} \log \mathbb{P}_Y \left(q(x) = s \right)$







- ▶ Free-energy barrier often created by energy/**likelihood** vs entropy/**volume trade-offs**

Statistical Physics and Free-Energy Barriers

- x^* ground truth
- x draw from posterior
  $x \sim \mathbb{P}(x \mid Y)$
- $q(x) \leftarrow$ measure of quality of draw

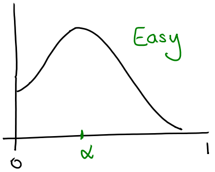
  $q(x^*) = 1$

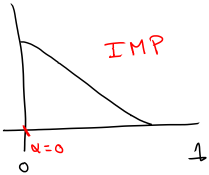
  $q = 0 \leftrightarrow$ random guessing
- What is the distribution of $q(x)$?

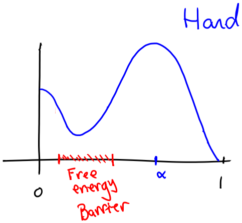
$q(x) \to \alpha > 0$ stat. possible

$q(x) \to 0$ stat impossible

Plot $\frac{1}{n} \log \mathbb{P}_Y \left(q(x) = s \right)$







- Free-energy barrier often created by energy/**likelihood** vs entropy/**volume trade-offs**
- Tightly connected to Franz-Parisi Potential [FP'95]. Also connected to Bethe Free-Energy and Overlap Gap Property.

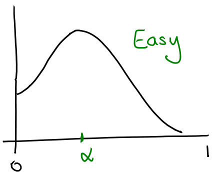
Statistical Physics and Free-Energy Barriers

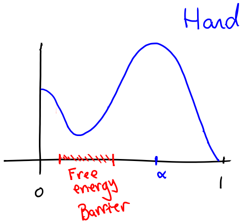
- x^* ground truth
- x draw from posterior
  $x \sim \mathbb{P}(x \mid Y)$
- $q(x) \leftarrow$ measure of quality of draw
  $q(x^*) = 1$
  $q = 0 \leftrightarrow$ random guessing
- What is the distribution of $q(x)$?

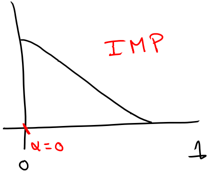
$q(x) \to \alpha > 0$ stat. possible

$q(x) \to 0$ stat impossible

Plot $\frac{1}{n} \log \mathbb{P}_Y \left(q(x) = s \right)$







- Free-energy barrier often created by energy/**likelihood** vs entropy/**volume trade-offs**
- Tightly connected to Franz-Parisi Potential [FP'95]. Also connected to Bethe Free-Energy and Overlap Gap Property.
- Possible to show Markov Chain Monte Carlo lower bounds

The Franz-Parisi Criterion

Low-Degree Method: Problem is hard if $LD(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{LD}(D) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x^{\leq D}, L_{x'}^{\leq D} \rangle$$

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

The Franz-Parisi Criterion

Low-Degree Method: Problem is hard if $LD(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{LD}(D) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x^{\leq D}, L_{x'}^{\leq D} \rangle$$

Franz-Parisi Criterion: Problem is hard if $FP(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{FP}(D) = \mathrm{LO}(\delta) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x, L_{x'} \rangle 1_{|\langle x,x' \rangle| \leq \delta} \qquad \delta \text{ s.t. } \mathrm{Prob}_{\mathbb{Q}}(|\langle x, x' \rangle| = \delta) \sim e^{-D}$$

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

The Franz-Parisi Criterion

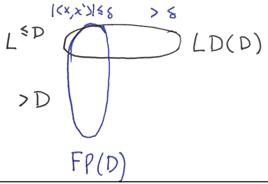
Low-Degree Method: Problem is hard if $LD(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{LD}(D) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x^{\leq D}, L_{x'}^{\leq D} \rangle$$

Franz-Parisi Criterion: Problem is hard if $FP(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{FP}(D) = \mathrm{LO}(\delta) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x, L_{x'} \rangle 1_{|\langle x,x' \rangle| \leq \delta} \qquad \delta \text{ s.t. } \mathrm{Prob}_{\mathbb{Q}}(|\langle x, x' \rangle| = \delta) \sim e^{-D}$$

Gaussian Additive Model: $Y = \lambda X + Z$ vs $Y = Z$



A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

The Franz-Parisi Criterion

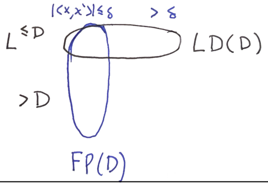
Low-Degree Method: Problem is hard if $LD(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{LD}(D) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x^{\leq D}, L_{x'}^{\leq D} \rangle$$

Franz-Parisi Criterion: Problem is hard if $FP(D)$ is bounded for some $D \gg \log(n)$

$$\mathrm{FP}(D) = \mathrm{LO}(\delta) = \mathop{\mathbb{E}}_{x,x' \sim \mu} \langle L_x L_{x'} \rangle 1_{|\langle x,x' \rangle| \leq \delta} \qquad \delta \text{ s.t. } \mathrm{Prob}_{\mathbb{Q}}(|\langle x, x' \rangle| = \delta) \sim e^{-D}$$

Gaussian Additive Model: $Y = \lambda X + Z$ vs $Y = Z$



$$\mathrm{LD}(D) = \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}'} \exp^{\leq \mathbf{D}}(\lambda^2 \langle \mathbf{X}, \mathbf{X}' \rangle)$$

$$\mathrm{FP}(D) = \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}'} \mathbf{1}_{|\langle \mathbf{X},\mathbf{X}' \rangle| \leq \delta} \exp(\lambda^2 \langle \mathbf{X}, \mathbf{X}' \rangle)$$

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

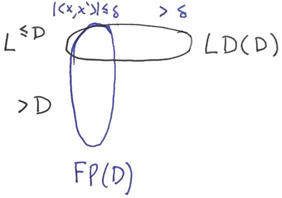
Main Results

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**

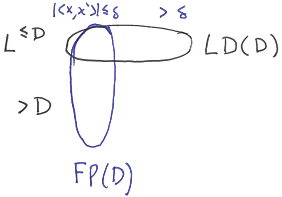


(proof by picture)

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**



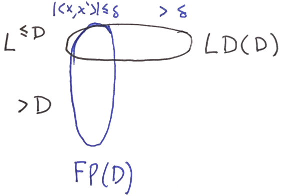
(proof by picture)

- Low-Degree Hard $\Rightarrow$ coldstarted local **Markov-Chain-Monte-Carlo Hard** (GAM)

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**



(proof by picture)

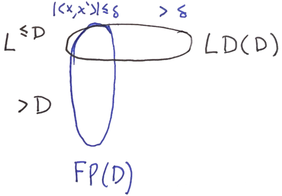
- Low-Degree Hard $\Rightarrow$ coldstarted local **Markov-Chain-Monte-Carlo Hard** (GAM)

- Franz-Parisi Hard $\Rightarrow$ Low-Degree Hard for certain planted sparse models
  gives new computational lower bound for **Sparse Linear Regression**

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**



(proof by picture)

- Low-Degree Hard $\Rightarrow$ coldstarted local **Markov-Chain-Monte-Carlo Hard** (GAM)

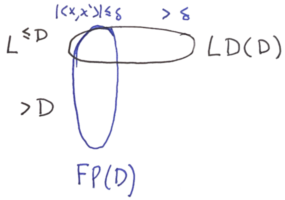
- Franz-Parisi Hard $\Rightarrow$ Low-Degree Hard for certain planted sparse models
  gives new computational lower bound for **Sparse Linear Regression**

- Unfortunately, criteria are not equivalent for all planted models

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results and questions

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**



(proof by picture)

- Low-Degree Hard $\Rightarrow$ coldstarted local **Markov-Chain-Monte-Carlo Hard** (GAM)

- Franz-Parisi Hard $\Rightarrow$ Low-Degree Hard for certain planted sparse models

  gives new computational lower bound for **Sparse Linear Regression**

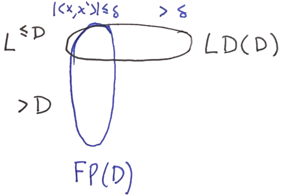
- Unfortunately, criteria are not equivalent for all planted models

  ★ A better free-energy-based criteria needed?

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Main Results and questions

- Low-Degree Method and Franz-Parisi Criterion are essentially equivalent for **Gaussian Additive Model**



(proof by picture)

- Low-Degree Hard $\Rightarrow$ coldstarted local **Markov-Chain-Monte-Carlo Hard** (GAM)

- Franz-Parisi Hard $\Rightarrow$ Low-Degree Hard for certain planted sparse models

  gives new computational lower bound for **Sparse Linear Regression**

- Unfortunately, criteria are not equivalent for all planted models

  ★ A better free-energy-based criteria needed?

  ★ Recovery vs Estimation?

A. S. Bandeira, A. El Alaoui, S. B. Hopkins, T. Schramm, A. S. Wein, I. Zadik, 2022

Thank You

Shameless plug I: For PhD & Postdoc positions visit: https://people.math.ethz.ch/~abandeira/positions.html

Shameless plug II: Draft available of a new book **Mathematics of Data Science**, and notes with Open Problems