Clustering a mixture of Gaussians with unknown covariance*

Mateo Díaz

Computing and Mathematical Sciences

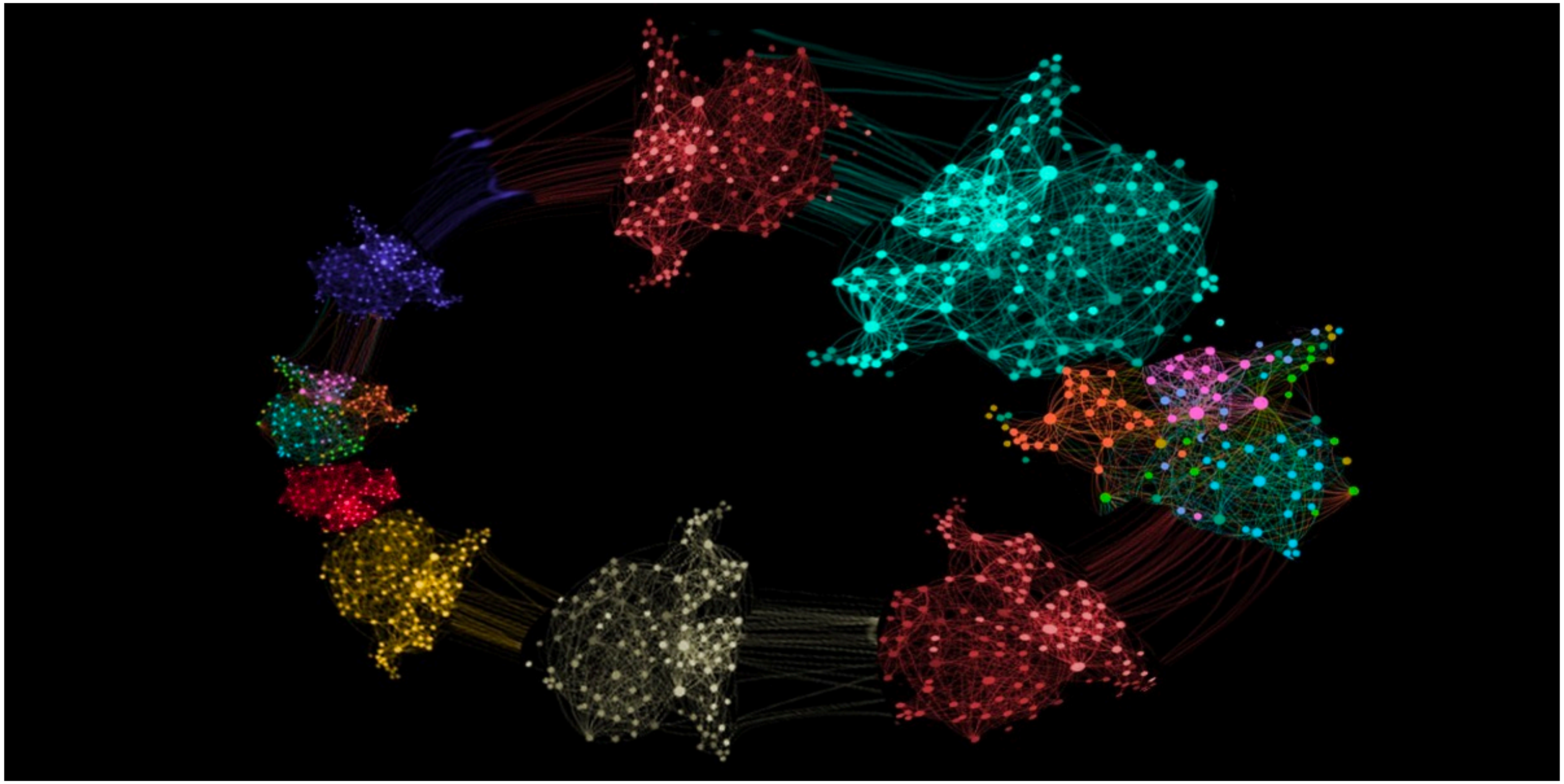
Robust and Resilience Workshop - Erice

May 2022

*Joint work with Damek Davis and Kaizheng Wang.

Clustering

Clustering if the process of partitioning a heterogeneous, unlabeled dataset into **groups of similar samples**.



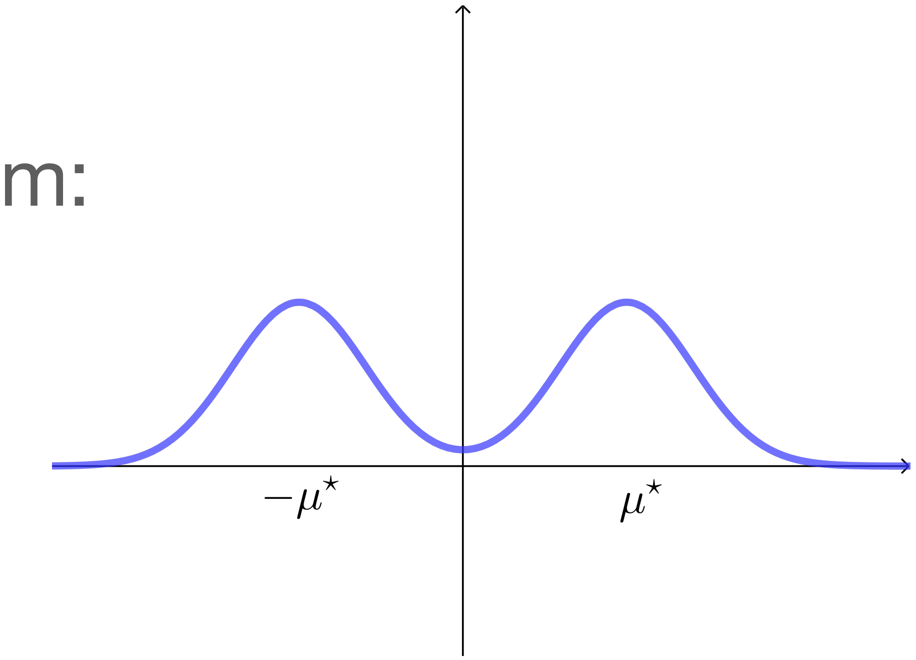
Ubiquitous task in ML and statistics with **applications** to computer vision, data analysis, network analysis, genomics, among others.

The problem today

Model

We consider a **Gaussian mixture** of the form:

$$\boldsymbol{X}_i \sim \frac{1}{2}N(-\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star) + \frac{1}{2}N(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$$



Equivalently,

$$\boldsymbol{X}_i = y_i \boldsymbol{\mu} + \boldsymbol{z}_i \quad \text{with} \quad \begin{cases} \boldsymbol{z} \sim N(0, \boldsymbol{\Sigma}), \\ \mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2. \end{cases}$$

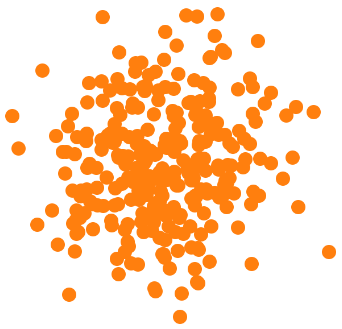
Goal

Given a sample $\{\boldsymbol{X}_i\}_{i=1}^n$, we want to recover the labels $\{y_i\}_{i=1}^n$.

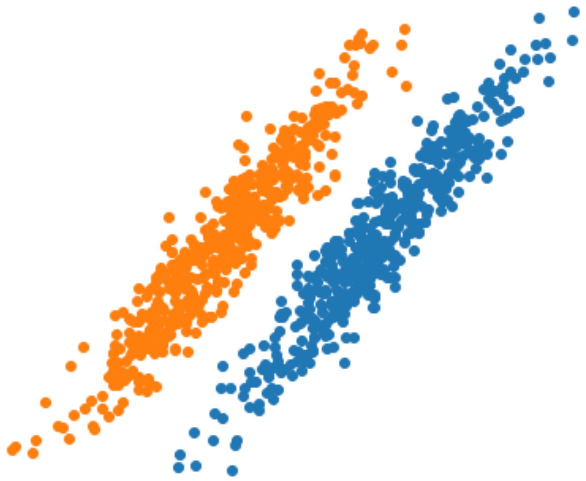
Challenges

Stretched mixtures

We are interested in the case when **the covariance is ill-conditioned.**



Spherical



Ill-conditioned

The **de facto solutions (PCA and k-means) struggle** in this setting.

Efficiency

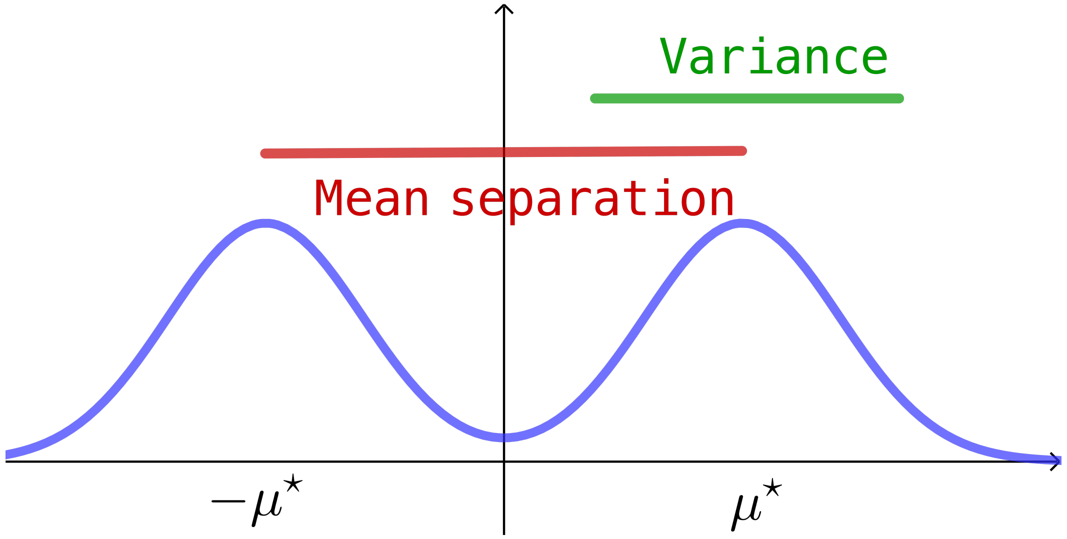
Make **efficient use of statistical and computational resources.**

How to measure separation?

Baby steps

In **one dimension** a natural way to measure the signal-strength is

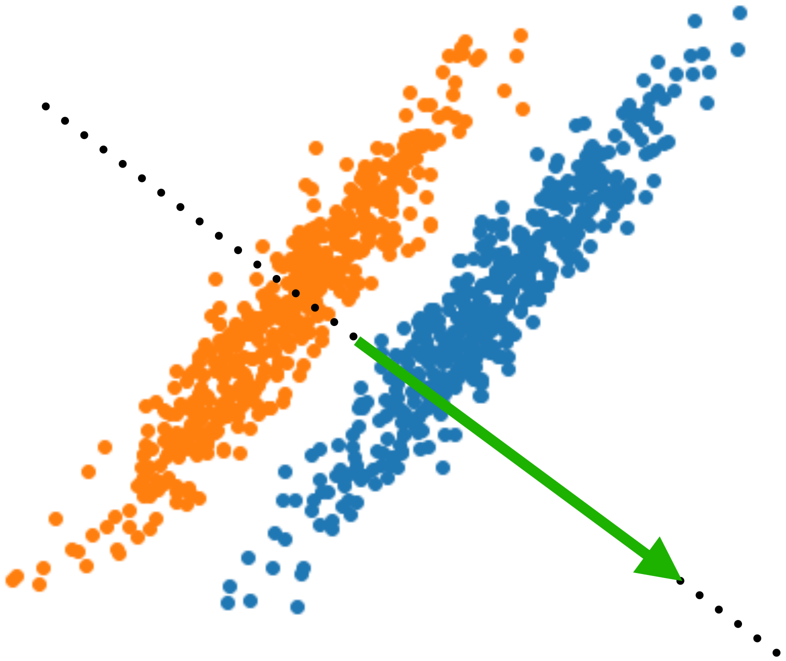
$$\mathrm{SNR} = \left(\frac{\mu^\star}{\sigma^\star}\right)^2$$



Signal-to-noise ratio

Generalizing this to higher dimensions:

$$\mathrm{SNR} = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$



Statistical metrics

Misclassification error

$$\mathcal{R}(\widehat{\boldsymbol{y}}, \boldsymbol{y}^\star) = n^{-1} \min_{s=\pm 1} \left|\{i \in [n] : \; s\widehat{y}_i \neq y_i^\star\}\right|.$$

Baseline

The **Bayes-optimal error**

$$\widehat{y}_i = \mathrm{sign}(\langle \boldsymbol{\Sigma}^{\star -1} \boldsymbol{\mu}^\star, \boldsymbol{X}_i \rangle)$$

$$\mathbb{E}(\mathcal{R}(\widehat{y}, y^\star)) = \exp(-\Omega(\mathrm{SNR}))$$

Sample complexity

Minimum number of samples necessary to achieve **Bayes-optimal error.**

Baseline

If the labels are given the **sample complexity** is $n = \Omega(d)$.

Questions

Statistical question

*When the **labels, mean, and covariance are unknown**, is it possible to achieve the **Bayes-optimal rate** with (near) **linear sample complexity**?*

Yes, using an optimization problem over the discrete hypercube $\{\pm 1\}^n$.

Computational question

*If so, is there a **computationally efficient estimator**?*

Doesn't seem possible unless more samples are provided.

Related work

Previous works do not tackle our questions because they either:

1. Need to **know the covariance matrix**.

Anandkumar, Ge, Hsu, Kakade, Telgarsky (2014),
Balakrishnan, Wainwright, Yu (2017),
Chen, Yang (2021),
Daskalakis, Tzamos, Zampetakis (2017),
Dwivedi, Ho, Khamaru, Wainwright, Jordan, Yu (2020),
…

Jin, Ke, Wang (2017),
Kwon, Caramanis (2020),
Löffler, Zhang, Zhou (2019),
Ndaoud (2018)
Vempala, Wang (2004),

2. Have **high sample complexity**.

Bakshi, Diakonikolas, Jia, Kane, Kothari (2020),
Bakshi, Kothari (2020),
Belkin, Sinha (2010),
Brubaker, Vempala (2008),
…

Cai, Ma, Zhang (2019)
Ge, Huang, Kakade (2015)
Moitra, Valiant (2010)
Tan, Vershynin (2018),

3. Exhibit **suboptimal dependency on the signal-to-noise ratio**.

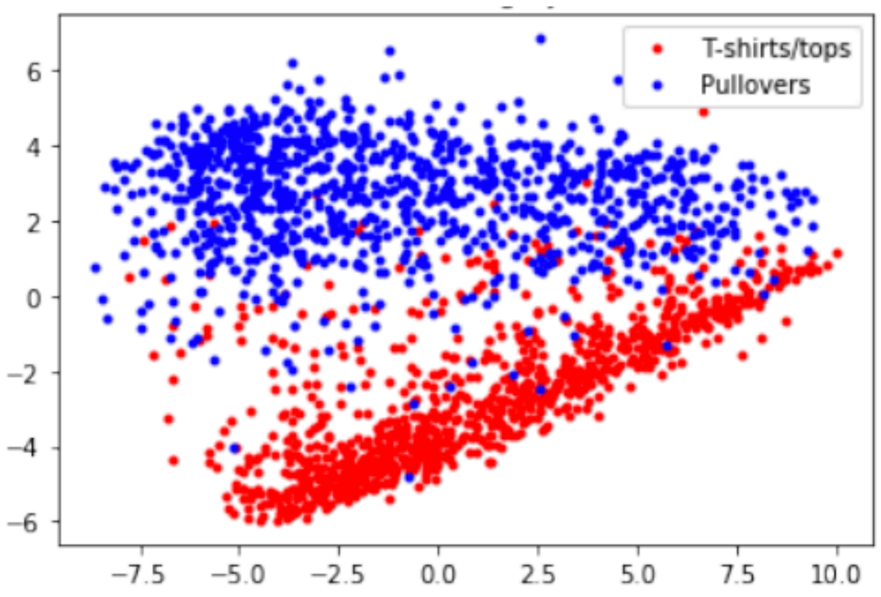
Abbe, Fan, Wang (2020),
Chen, Yang (2021),
Fei, Chen (2018),
Giraud, Verzelen (2019),

Lu, Zhou (2016),
Mixon, Villar, Ward (2017),
Royer (2017),
…

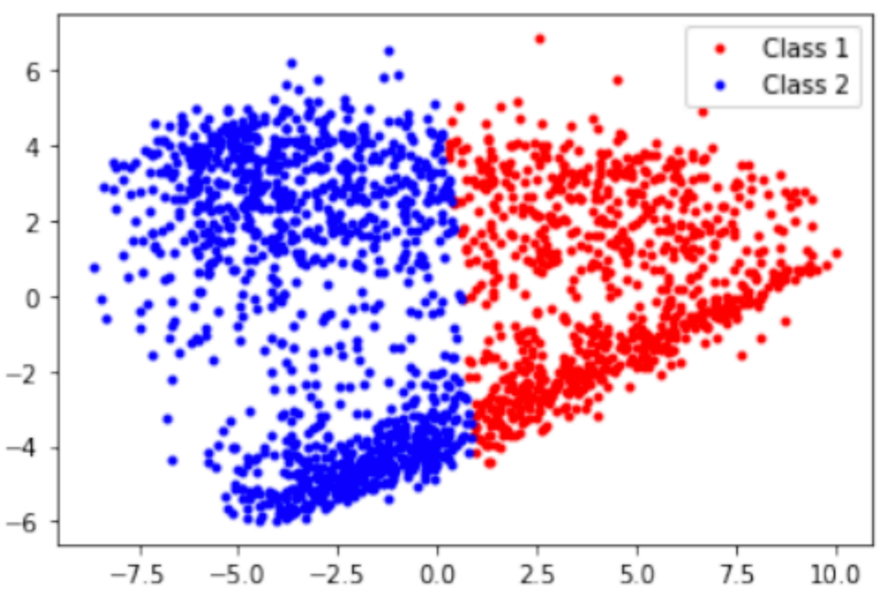
Numerical illustration: FashionMNIST

We sampled 1000 T-shirts/tops and Pullovers from **FashionMNIST** and classified them with **k-means** and our **new algorithm**.

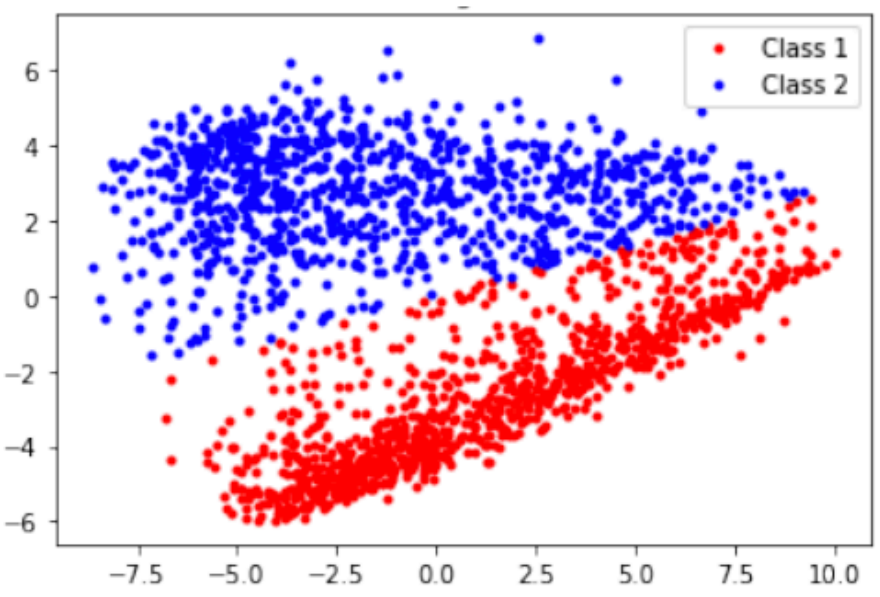
True labels



k-means



New algorithm



Misclassification rate: 44.7% and **7.1%**.

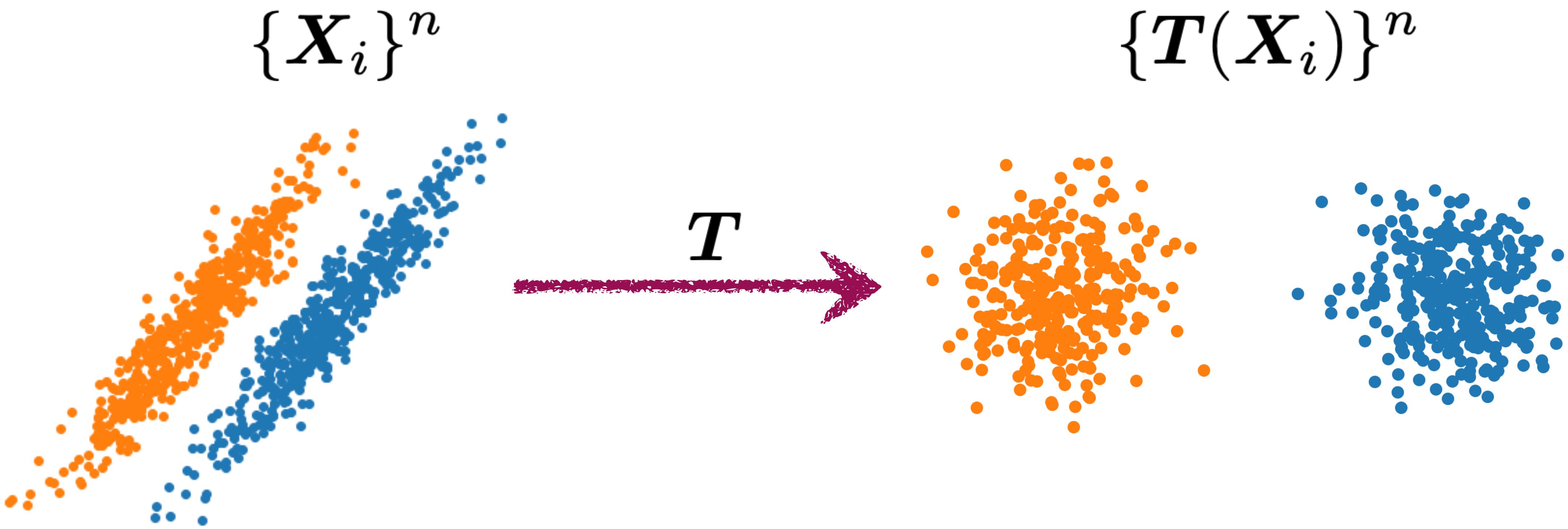
Agenda

- Introduction
- MAX-CUT formulation
- Efficient spectral algorithm
- Potential statistical-computational gap

Max-Cut formulation

Insight: Invariance

Let $\boldsymbol{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary nonsingular linear transformation



Then, we have

$$\mathrm{SNR}(\{\boldsymbol{X}_i\}^n) = \mathrm{SNR}(\{\boldsymbol{T}\boldsymbol{X}_i\}^n)$$

Key insight

Any estimator that depends on SNR **has to be invariant**.

Canonical form

Lemma (Canonical form)

There is a map $\boldsymbol{T}\colon \mathbb{R}^d \to \mathbb{R}^d$ such that:

$$\sigma = \frac{1}{\sqrt{\mathrm{SNR}+1}}$$

$$\begin{pmatrix} — & (\boldsymbol{T}\boldsymbol{X}_1)^\top & — \\ & \vdots & \\ — & (\boldsymbol{T}\boldsymbol{X}_n)^\top & — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \sqrt{1-\sigma^2}\boldsymbol{y}^\star + \sigma\mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_d \\ | & | & & | \end{pmatrix}.$$

where $\mathbf{g}_i \sim N(0, \boldsymbol{I})$ are iid and independent from labels.

Intuition

Every datapoint has its **label in the first component** and noise everywhere else.

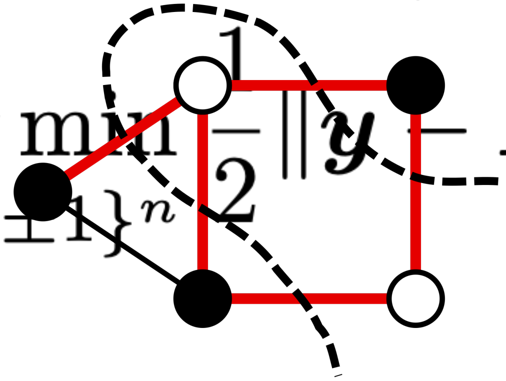
Maximum likelihood estimator

Define $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ to be the projection onto the range of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$.

Lemma (From MLE to Max-Cut)

The **MLE** is given by the following problem over the hypercube:

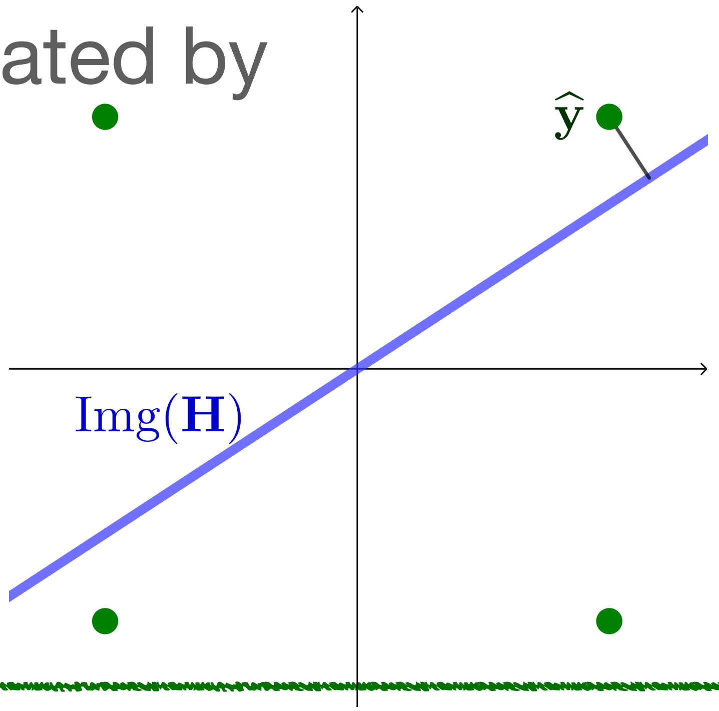
$$\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\{\pm 1\}^n} \boldsymbol{y}^\top \boldsymbol{H}\boldsymbol{y} = \arg\min_{\boldsymbol{y}\in\{\pm 1\}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|^2$$



Intuition

Minimize the **distance to the subpspace** generated by

$$\begin{pmatrix} | & | & & | \\ \sqrt{1-\sigma^2}\boldsymbol{y}^\star + \sigma \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_d \\ | & | & & | \end{pmatrix}$$



Optimality of Max-Cut

Theorem (Max-Cut)

Assume that $\mathrm{SNR} \to \infty$ and $n/(d \log n) \to \infty$. **The solution of the Max-Cut problem** satisfies:

1. (**Expected error**)

$$\mathrm{SNR} \le C \log n \implies \mathbb{E}\mathcal{R}(\widehat{\boldsymbol{y}}, \boldsymbol{y}^\star) \le e^{-\mathrm{SNR}/[2+o(1)]}.$$

2. (**Probability of zero error**)

$$\mathrm{SNR} \ge (2+\varepsilon) \log n \implies \mathbb{P}[\mathcal{R}(\widehat{\boldsymbol{y}}, \boldsymbol{y}^\star) = 0] = 1 - o(1).$$

Remarks.

1. Expected error matches Bayes-optimal error (up to the little-o term).
2. Separation holds with high probability if and only if $\mathrm{SNR} > 2 \log n$.
3. Best known algorithm to solve Max-Cut takes exponential time.

Why is this Max-Cut instance hard?

Natural ideas to solve the Max-Cut:

Spectral relaxation

Relax the constrained set to a **sphere**:

$$\arg\max_{\|\boldsymbol{y}\|^2=n} \boldsymbol{y}^\top \boldsymbol{H} \boldsymbol{y}.$$

Issue: the **leading eigenvector is not unique**.

Semidefinite relaxation

Relax to a spectrahedron a la **Goemans-Williamson** (1995):

$$\max_{\boldsymbol{Y}\in\mathbb{R}^{n\times n}} \langle \boldsymbol{H}, \boldsymbol{Y} \rangle \qquad \text{s.t.} \quad \boldsymbol{Y} \succeq 0, \quad \operatorname{diag}(\boldsymbol{Y}) = \mathbf{1}.$$

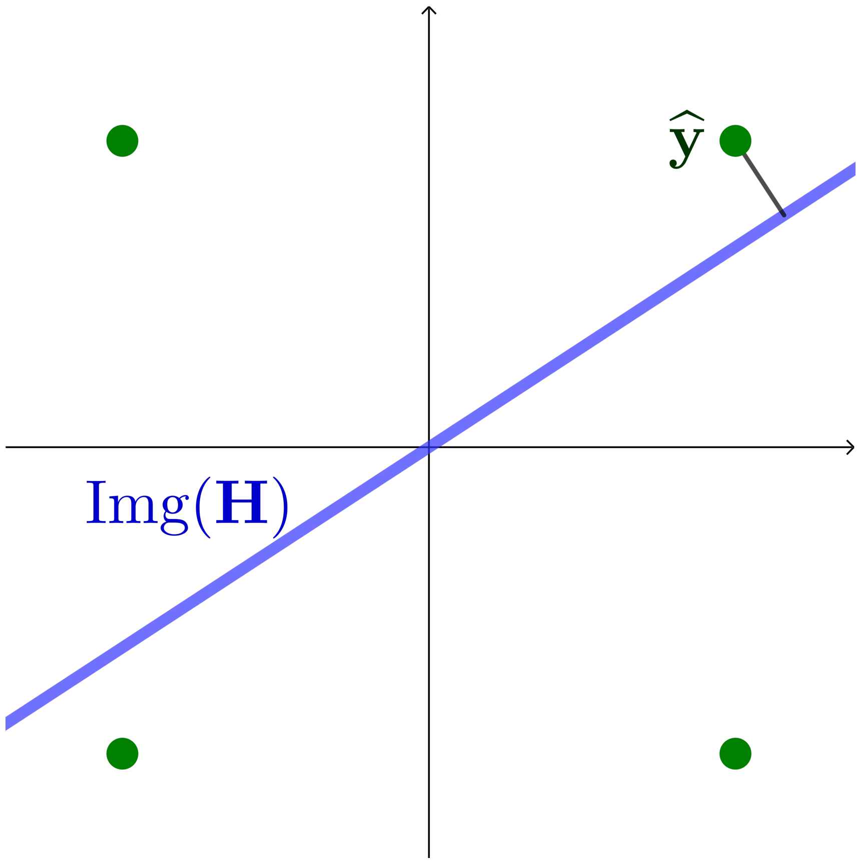
Issue: Guarantees do not apply since $\boldsymbol{H}$ might have **negative entries**.

Efficient algorithm

Two stage algorithm

We want to solve the **nonconvex problem:**

$$\operatorname*{arg\,min}_{\boldsymbol{y}\in\{\pm 1\}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|^2$$



Strategy split the algorithm into two:

- **Stage 1: Initialization (spectral method)**
  Finds a point **close to the solution.**

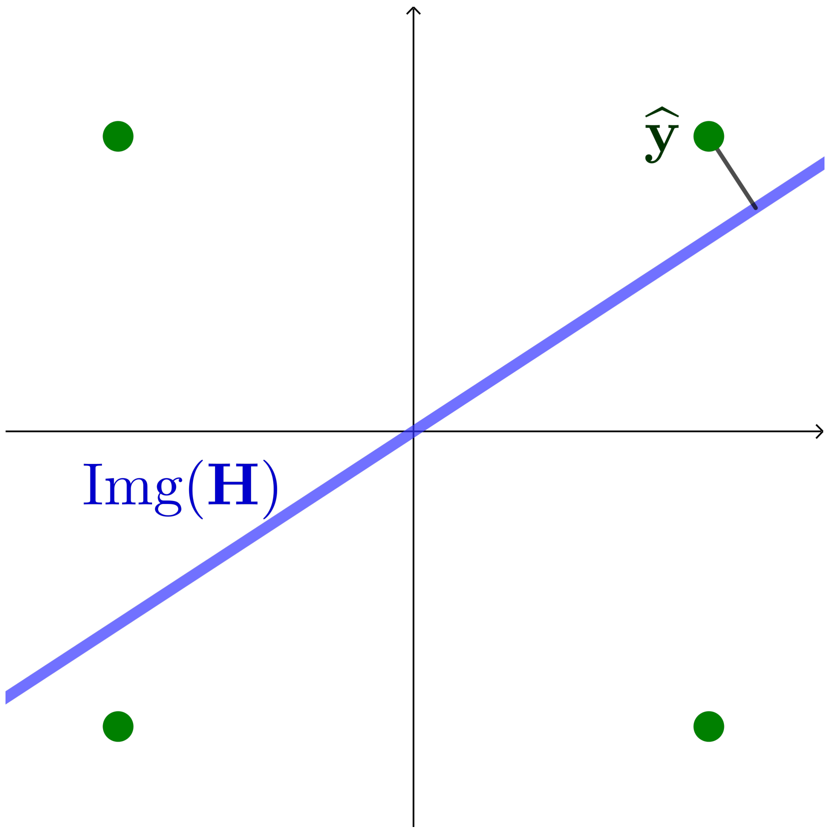
- **Stage 2: Local refinement (optimization)**
  Iterative algorithm that solves the optimization problem.

Projected power iteration

Algorithm 1 Projected power iteration

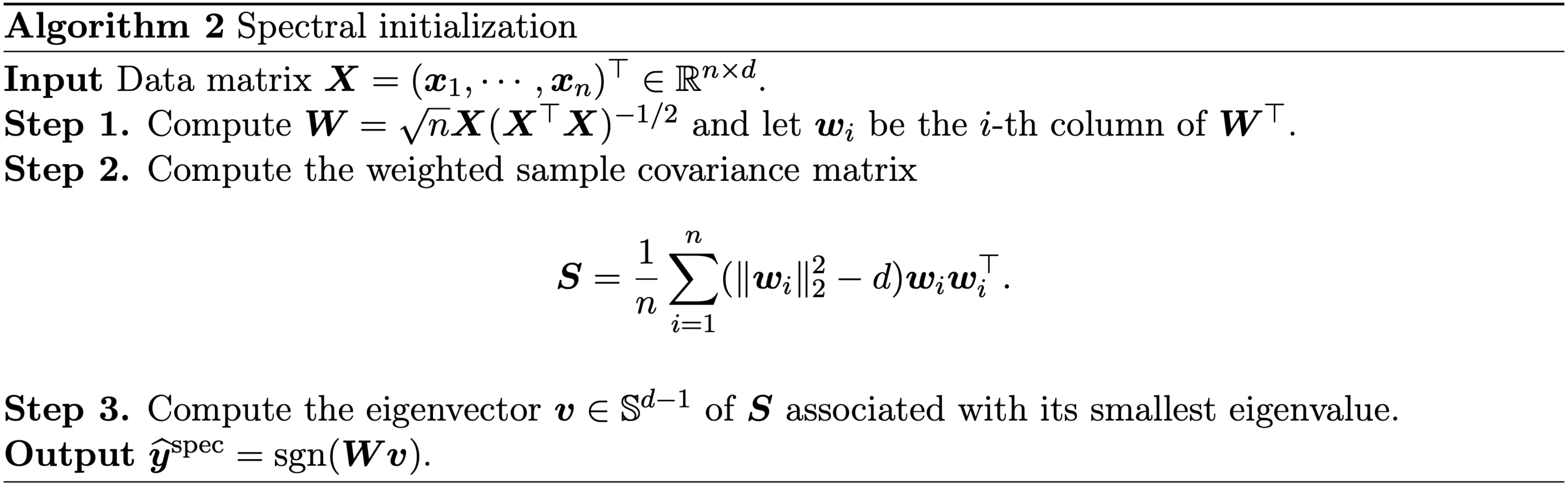
```
**Input** data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, initial guess $\boldsymbol{y}^0 \in \{\pm 1\}^n$.
**Compute** $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and set $T = 4\lceil \log_2 n \rceil + 4$.
**For** $t = 0, 1, \ldots, T-1$
    $\boldsymbol{y}^{t+1} = \mathrm{sgn}(\boldsymbol{H}\boldsymbol{y}^t)$        // *applied in an entry-wise manner*
**Return** $\widehat{\boldsymbol{y}}^{\mathrm{PPI}} = \boldsymbol{y}^T$.
```

This algorithm is **simply alternating projections!**



At each iteration we **project onto a subspace** and then **onto the discrete hypercube**.

Spectral algorithm



```
Algorithm 2 Spectral initialization
Input Data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d}$.
Step 1. Compute $\boldsymbol{W} = \sqrt{n}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}$ and let $\boldsymbol{w}_i$ be the $i$-th column of $\boldsymbol{W}^\top$.
Step 2. Compute the weighted sample covariance matrix
    $$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\|\boldsymbol{w}_i\|_2^2 - d)\boldsymbol{w}_i\boldsymbol{w}_i^\top.$$
Step 3. Compute the eigenvector $\boldsymbol{v} \in \mathbb{S}^{d-1}$ of $\boldsymbol{S}$ associated with its smallest eigenvalue.
Output $\widehat{\boldsymbol{y}}^{\text{spec}} = \operatorname{sgn}(\boldsymbol{W}\boldsymbol{v})$.
```

Key insight

Once $n = \tilde{\Omega}(d^2)$, the matrix concentrates around:

$$\boldsymbol{S} \approx 2\left(\boldsymbol{I} - c\frac{\boldsymbol{v}^\star\boldsymbol{v}^{\star\top}}{\|\boldsymbol{v}^\star\|^2}\right)$$

where the vector $\boldsymbol{v}^\star$ gives the optimal classifier for $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$.

Global convergence guarantee

Let $\widehat{\boldsymbol{y}}$ be the output of combining the spectral method and alternating projections algorithm.

Theorem (Global convergence)

Assume that $\mathrm{SNR} \to \infty$, $n/(d^2 \log^3 n) \to \infty$. Then:

1. (**Expected error**)

$$\mathrm{SNR} \le C \log n \quad \Longrightarrow \quad \mathbb{E}\mathcal{R}(\widehat{\boldsymbol{y}}, \boldsymbol{y}^\star) \le e^{-\mathrm{SNR}/[2+o(1)]}.$$

2. (**Probability of zero error**)

$$\mathrm{SNR} \ge (2+\varepsilon) \log n \quad \Longrightarrow \quad \mathbb{P}[\mathcal{R}(\widehat{\boldsymbol{y}}, \boldsymbol{y}^\star) = 0] = 1 - o(1).$$

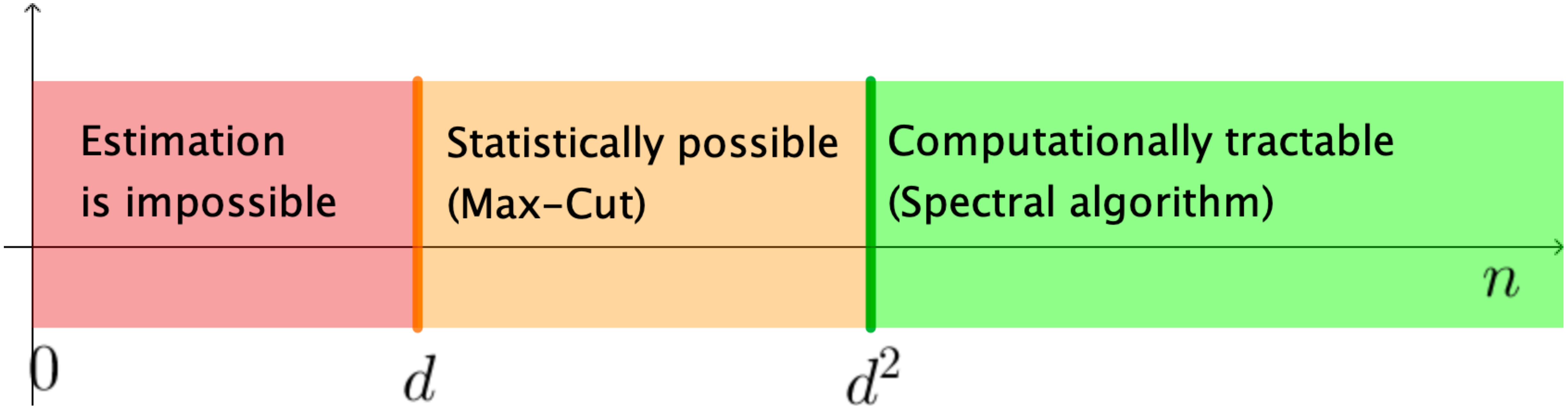
Remark

Exactly the same guarantees as Max-Cut with **quadratic sample complexity.**

A potential statistical-computational gap

A statistical-computational gap?

So far we have established that when $S \lesssim 1 \ll \mathrm{SNR}$, then



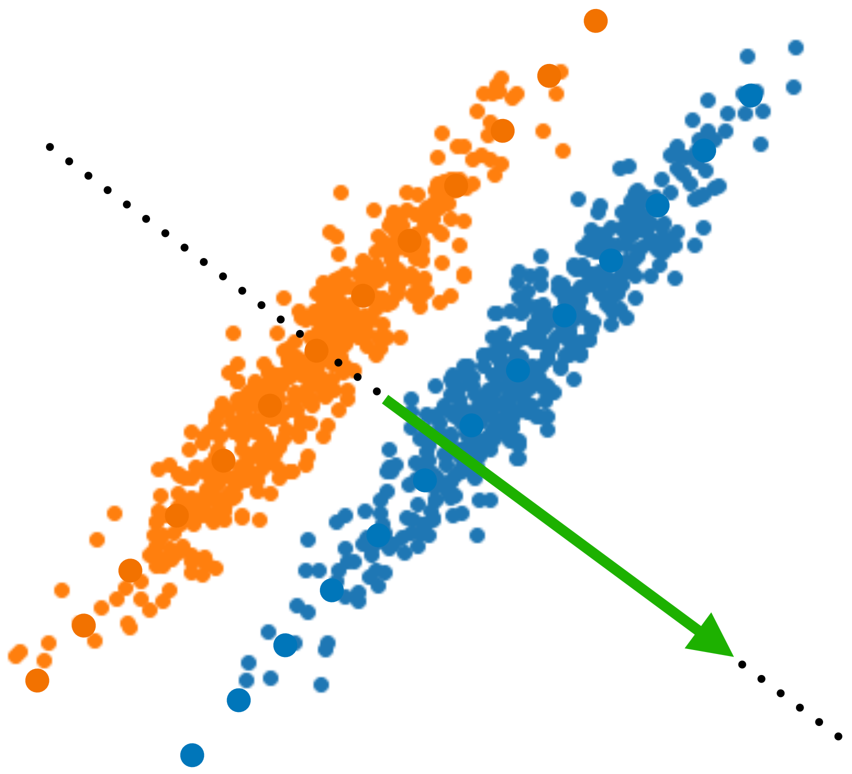
We conjecture that **no polynomial time algorithm performs better than a random guess** in the regime:

$$d \lesssim n \ll d^2.$$

Note that there is no statistical-computational gap when $S \gg 1$.

Caveat: Conjecture requires noise!

Recent papers by **Zadik, Song, Wein, and Bruna (2021)** and **Diakonikolas and Kane (2022)** disproved the conjecture when there is **no noise**, e.g., infinite SNR.



Their result is based on **Lenstra–Lenstra–Lovász lattice basis reduction** and only require $d + 1$ samples.

A statistical-computational gap?

We collect three pieces of evidence:

Numerical evidence

Popular polynomial time methods seem to need $n = \Omega(d^2)$.

Reduction from a hard testing problem

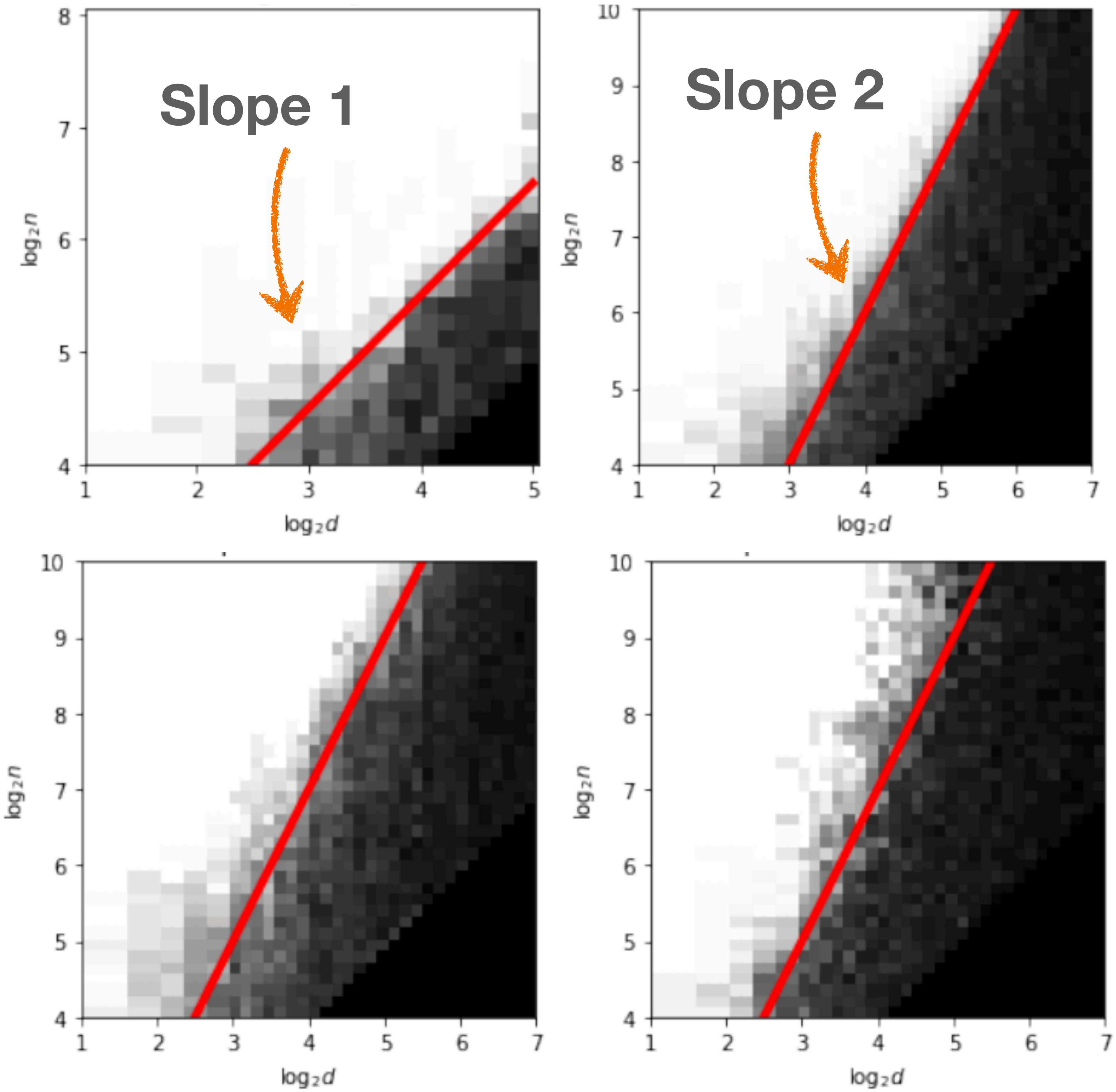
We reduce the problem from a **hypothesis testing problem** believed to be hard in the regime $n \ll d^2$.

Lower bound for sum-of-squares relaxations

We prove that **SoS relaxations** of the Max-Cut formulations **fail** when $n \ll d^{3/2}$.

Numerical evidence

Recovery frequency with $\mathrm{SNR} = 3\log n.$



Max-Cut

Semidefinite relaxation

Spectral + alternating projections

Expectation-Maximization algorithm

A hard testing problem

Strategy: reduce from a problem that we believe is hard.

Testing problem

We observe $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. We want a **test** $\psi : \mathbb{R}^{n \times d} \to \{H_0, H_1\}$ that can **decide between two hypothesis:**

Null hypothesis:

$$\boldsymbol{X} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_d) \quad \text{with i.i.d.} \quad \boldsymbol{g}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$$

Alternative hypothesis:

$$\boldsymbol{X} = \tilde{\boldsymbol{X}} \boldsymbol{Q} \quad \text{with} \quad \begin{cases} \tilde{\boldsymbol{X}} = (\boldsymbol{y}^\star, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_d) \\ \boldsymbol{Q} \text{ unknown rotation.} \end{cases}$$

We want a test with small **Type I and Type II errors**.

$$\mathbb{P}\Big(\psi(\boldsymbol{X}) = H_1 \Big| H_0\Big) + \mathbb{P}\Big(\psi(\boldsymbol{X}) = H_0 \Big| H_1\Big)$$

Spectral methods lower bound

A family of spectral tests

For a fixed degree $\boldsymbol{p}$. Let $\boldsymbol{M} : \mathbb{R}^{n \times d} \to \mathcal{S}^{n^p}$ with polynomial entries of degree at most $\boldsymbol{p}$. Then, consider the test

$$\psi(\boldsymbol{X}) = \begin{cases} H_0 & \text{if } \|\boldsymbol{M}(\boldsymbol{X})\| < t, \\ H_1 & \text{otherwise.} \end{cases}$$

Informal theorem (Mao and Wein, 2021)

Assume that $n \ll d^2$. Any test $\psi : \mathbb{R}^{n \times d} \to \{H_0, H_1\}$ coming from the above family has to satisfy

$$\mathbb{P}(\psi = H_1 \mid H_0) + \mathbb{P}(\psi = H_0 \mid H_1) \geq n^{-C+o(1)}.$$

A reduction from testing

Theorem (Reduction)

Assume that $n = \Omega(d)$ and $\mathrm{SNR} = c \log n$. Then if there is an **estimator for the clustering** problem such that

$$\mathbb{E}\mathcal{R}[\varphi(\boldsymbol{X}), \boldsymbol{y}^\star] = e^{-\mathrm{SNR}/[2+o(1)]}$$

Then, one can construct a **simple test** $\psi : \mathbb{R}^{n \times d} \to \{H_0, H_1\}$ that achieves:

$$\mathbb{P}(\psi = H_1 \mid H_0) + \mathbb{P}(\psi = H_0 \mid H_1) \leq n^{-c+o(1)}.$$

Conclusion

It is unlikely that there is a polynomial time estimator.

Max-Cut Semidefinite relaxation

Observation

We can rewrite the Max-Cut as a linear program:

Cut polytope

$$\max_{\boldsymbol{Y}\in\mathcal{C}} \langle \boldsymbol{H}, \boldsymbol{Y} \rangle \quad \text{where} \quad \mathcal{C} = \operatorname{conv}(\{\boldsymbol{y}\boldsymbol{y}^\top : \ \boldsymbol{y} \in \{\pm 1\}^n\}),$$

and LPs achieve their **maximum at vertices**.

Obstruction: Optimizing over the Cut polytope is **NP-hard.**

Idea (Goemans-Williamson 1995)

Substitute $\mathcal{C}$ by a set with a tractable SDP representation

$$\mathcal{S}_2 = \{\boldsymbol{Y} \in \mathbb{R}^{n\times n} : \ \boldsymbol{Y} \succeq 0, \ \operatorname{diag}(\boldsymbol{Y}) = \mathbf{1}\}$$

Intuitively, we match all the second-moment information we have.

Sum-of-squares relaxations

Sum-of-squares (Parrilo 2000, Laserre 2001)

We could take a hierarchy of tractable sets such that

$$\mathcal{S}_2 \supseteq \mathcal{S}_4 \supseteq \cdots \supseteq \mathcal{S}_m = \mathcal{C}$$

where each set has a **tractable SDP representation** and intuitively

$$\mathcal{S}_{2k} = \text{“Matches the moment information up to degree } 2k.\text{”}$$

For each fixed level we can solve the following in **polynomial time**

$$\max_{\boldsymbol{Y} \in \mathcal{S}_k} \langle \boldsymbol{H}, \boldsymbol{Y} \rangle$$

This **strategy has been very successful** for several problems in machine learning, and theoretical computer science.

Sum-of-squares lower bound

Theorem (lower-bound SoS) based on an obstruction by **Ghosh et al. (2020)**

Assume that $n \leq d^{3/2-\epsilon}$ and let $k \leq n^{c\epsilon}$. Then, with high probability there **exists a solution**

$$\widehat{\boldsymbol{Y}} \in \arg\max_{\boldsymbol{Y} \in \mathcal{S}_k} \langle \boldsymbol{H}, \boldsymbol{Y} \rangle,$$

that is **statistically independent from the true labels $\boldsymbol{y}^\star$**.

Intuition

When $n = O(d^{3/2})$, any relaxation of small degree k has a solution that performs as a **random guess**.

Summary

Statistically optimal procedure

The **Max-Cut** formulation

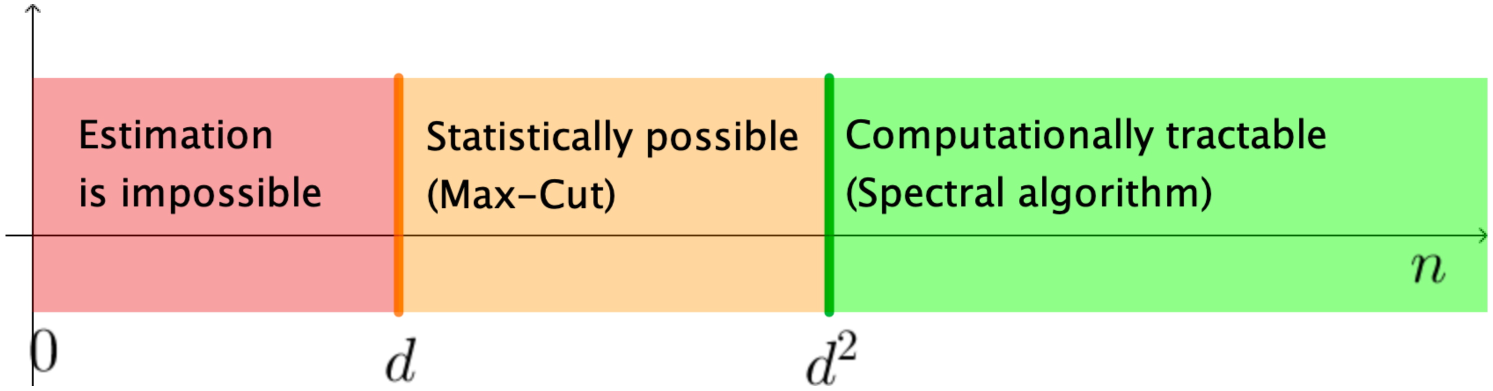
$$\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\{\pm 1\}^n} \boldsymbol{y}^\top \boldsymbol{H} \boldsymbol{y}$$

gives **Bayes-optimal error** with (near) **linear sample size**.

Computationally tractable procedure

Spectral method + alternating projections yield **optimal error** provided **quadratic sample size**.

Conjecture: statistical-computational gap



Estimation is impossible
Statistically possible (Max−Cut)
Computationally tractable (Spectral algorithm)
n
0
d
d^2

Thank you