From disordered systems to the mathematics of data science

Antoine Maillard

ETH zürich

Argo team – February 12th 2024

The Mathematics of Data

Modern machine learning

Brown & al, **2020**

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

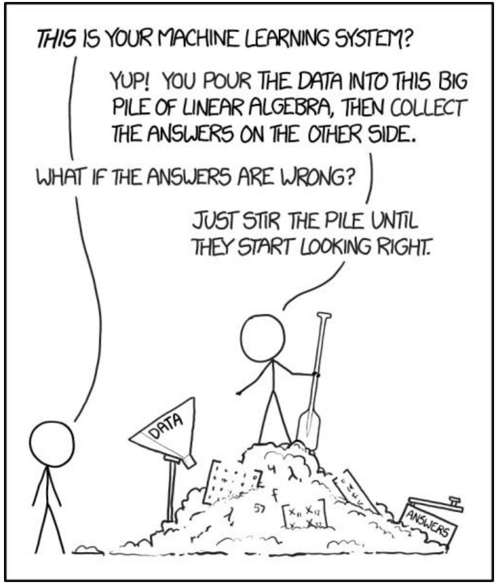
Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

High-dimensional statistics

Number of **parameters** $d \to \infty$

+

Size of **dataset** $n \to \infty$

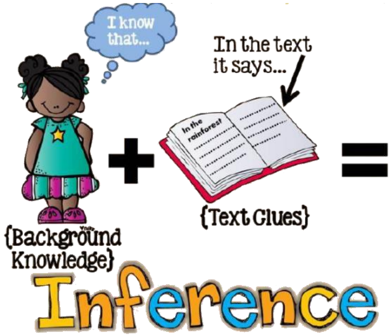


XKCD

Goal : fundamental limits in high-dimensional statistics

- When is learning/inference/optimisation **(im)possible** ?
- **Which algorithms** work ? Why ? Are there **bottlenecks ?**

A first example: from Bayes to Boltzmann



Inference : learn parameters from data

Model — Parameters $\mathbf{x}^\star \in \mathbb{R}^d$ $\Longrightarrow$ Measurements $y_1, \cdots, y_n$

Posterior distribution

High-dimensional — Random (e.g. noise)

$$\mathbb{P}[\mathbf{x}|\mathbf{y}] = \frac{1}{P(\mathbf{y})} P_0(\mathbf{x}) \prod_{i=1}^{n} P(y_i|\mathbf{x})$$

$\Longleftrightarrow$

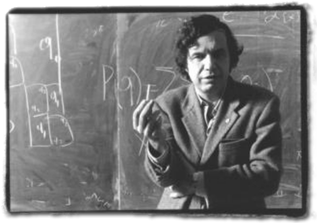
Disordered system - **Spin glass**

Random (noise, ...)

$$\mathbb{P}(\mathbf{x}) = \frac{e^{-\mathcal{H}(\mathbf{x})/T}}{\mathcal{Z}} \quad \begin{cases} T = 1 \\ \mathcal{H}(\mathbf{x}) = -\log P_0(x) - \sum_{i=1}^{n} \log P(y_i|\mathbf{x}) \end{cases}$$

Physicists and mathematicians have studied high-dimensional disordered systems / spin glasses for 40+ years !

2021



Giorgio Parisi

The physics of computation

A toolbox from disordered systems and high-dimensional probability for e.g.…

- Bayesian **inference** and **learning**

  Learning in two-layers neural networks ; Phase retrieval ; Matrix factorisation ; High-temperature expansions ; Generalised linear estimation ; Generative priors and data structure; Bottlenecks of MCMC algorithms ; …

$$\mathbb{P}[\mathbf{x}|\mathbf{y}] \propto P_0(\mathbf{x}) \prod_{i=1}^{n} P(y_i|\mathbf{x})$$

- Random **constraint satisfaction problems**

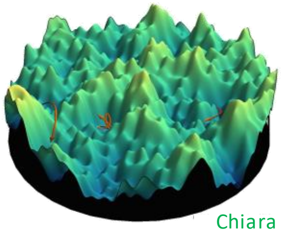
  Spherical perceptron ; random ellipsoid fitting ; …

graph colouring

$$\mathbb{P}[\mathbf{x}] \propto P_0(\mathbf{x}) \prod_{i=1}^{n} \mathcal{C}(y_i, \mathbf{x})$$

- High-dimensional **optimisation**

  Landscape complexity: Kac-Rice formula, large deviations for random matrices, …



Chiara Cammarota

$$\min_{\mathbf{x}\in\Sigma} \sum_{i=1}^{n} \mathcal{H}(y_i, \mathbf{x})$$

- …

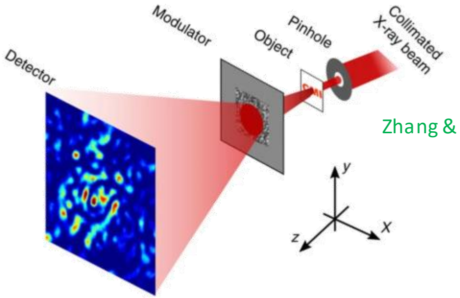
Example 1 : Phase retrieval

Example 1: Phase retrieval

Goal: Recover $x^\star \in \mathbb{C}^d$ from phaseless measurements $\{y_\mu = |\Phi_\mu \cdot x^\star|^2\}_{\mu=1}^n$

$n, d \to \infty$

known "sensing" matrix



Zhang & al '16

Random models for Φ
- Gaussian
- Subsampled DFT
- ...

imaging in complex media, ptychography, ...

Review: **Dong, Valzania, M., Pham, Gigan, Unser '23**

Our toolbox

- Analytical predictions using **replica/cavity method**
- **Efficient algorithms** (message-passing)
- **Rigorous proofs** of replica predictions

Parisi, Mézard, Virasoro '87 ; Krzakala & Zdeborová '16 ; ....

Mézard & Montanari '09 ; Donoho, Maleki & Montanari '09 ; ...

Guerra & Toninelli '02 ; Talagrand '06 '10 ;
Barbier, Krzakala, Macris, Miolane & Zdeborová '19 ; ....

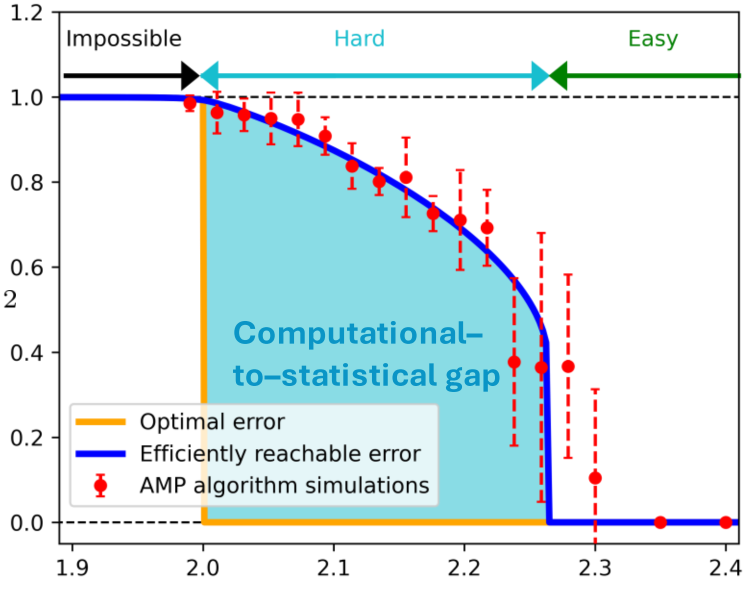
Computational-to-statistical gaps in phase retrieval

Theorem: (M., Loureiro, Krzakala & Zdeborová '20)

Φ : Haar-sampled column-unitary matrix

Mean Squared **Error**

$\min_{\hat{\mathbf{x}}(\mathbf{y})} \|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2$



$$\alpha = \frac{n}{d} = \frac{\#\text{measurements}}{\#\text{parameters}}$$

"Hard phase"

- $\alpha_{\text{IT}} = 2$: # of samples needed for perfect recovery for **any algorithm**
- $\alpha_{\text{AMP}} \simeq 2.27$: # of samples needed for perfect recovery for **approximate message passing**

conjectured optimal among **polynomial-time** ones (Gamarnik, Moore & Zdeborová '22)

Many extensions:

- Other models of Φ : a **zoology of hard phases**
- **Noisy versions, other activations…**

$$y_i \sim P_{\text{out}}(\cdot|\Phi_i \cdot x^\star)$$

Fast and optimal algorithms for phase retrieval

M., Lu, Krzakala & Zdeborová '22

- Semidefinite relaxations
- Non-convex optimisation
- Approximate Message-Passing

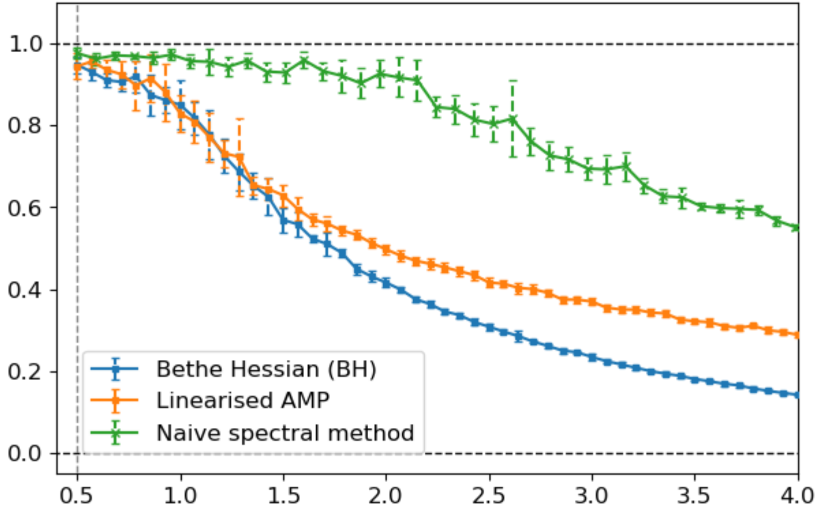
Computationally heavy / Need informed initialisation → **Spectral methods**

Two strategies inspired by **disordered systems**

- ➢ AMP **linearisation**
- ➢ **Bethe Hessian** $\simeq$ **non-backtracking matrix** in community detection [Saade & al '14]

Mean Squared **Error**

$\|\hat{\mathbf{x}} - \mathbf{x}^{\star}\|^2$



α_{WR}

$$\alpha = \frac{n}{d} = \frac{\#\text{measurements}}{\#\text{parameters}}$$

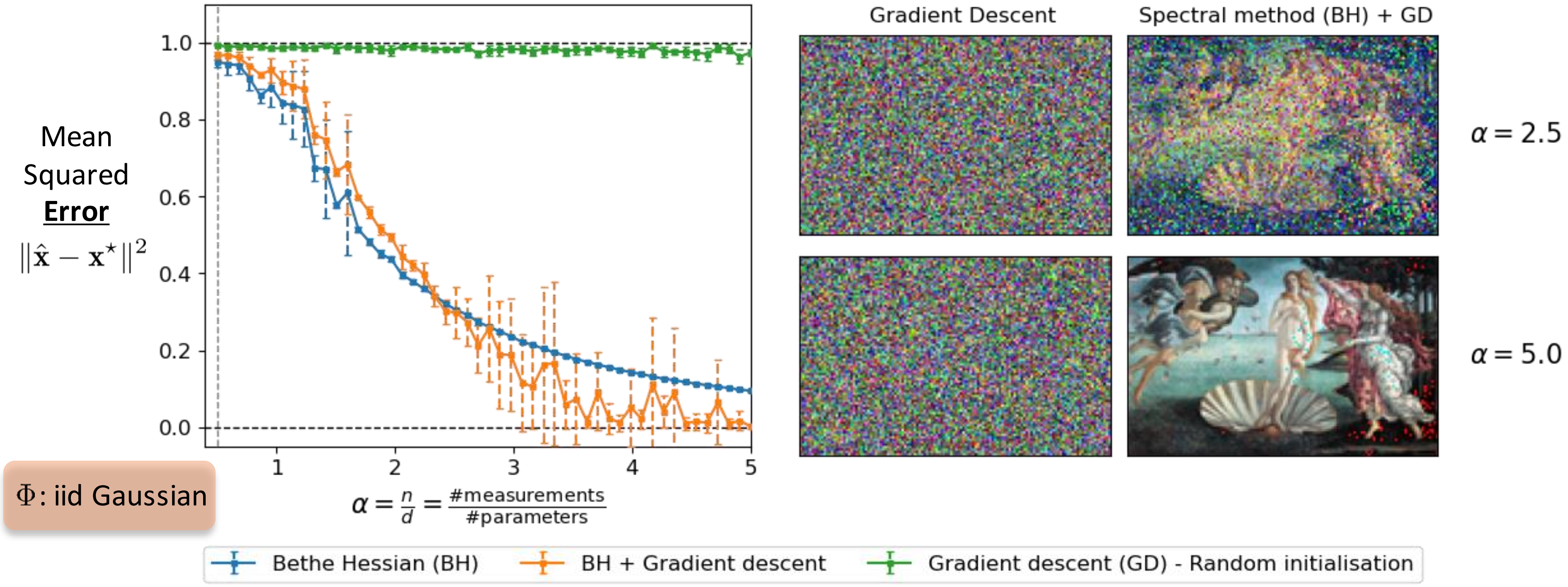
Φ : iid Gaussian

- ➢ Achieves **optimal weak recovery**
- ➢ **Bethe Hessian** is conjectured optimal

[M., Lu, Krzakala & Zdeborová '22]

Improving spectral methods with gradient descent

M., Lu, Krzakala & Zdeborová '22



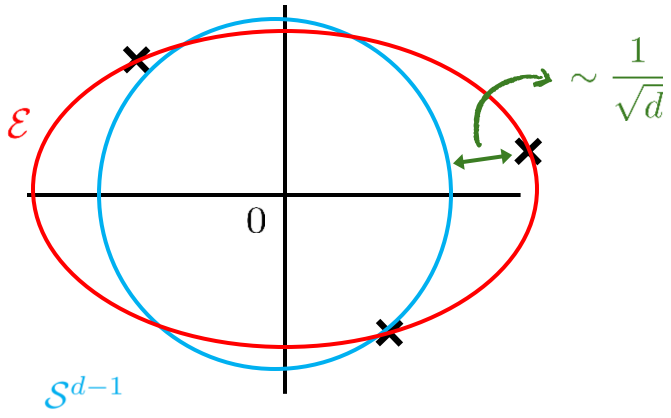
- Combined with gradient descent (on square loss): **efficient and cheap** !
- Derived for synthetic signals, but efficient on **real image recovery**

Example 2 :

Phase transition in a Matrix Constraint Satisfaction Problem

Example 2: Phase transition in a Matrix Constraint Satisfaction Problem

$$x_1, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d/d)$$
$$n, d \to \infty$$



Ellipsoid Fitting Property

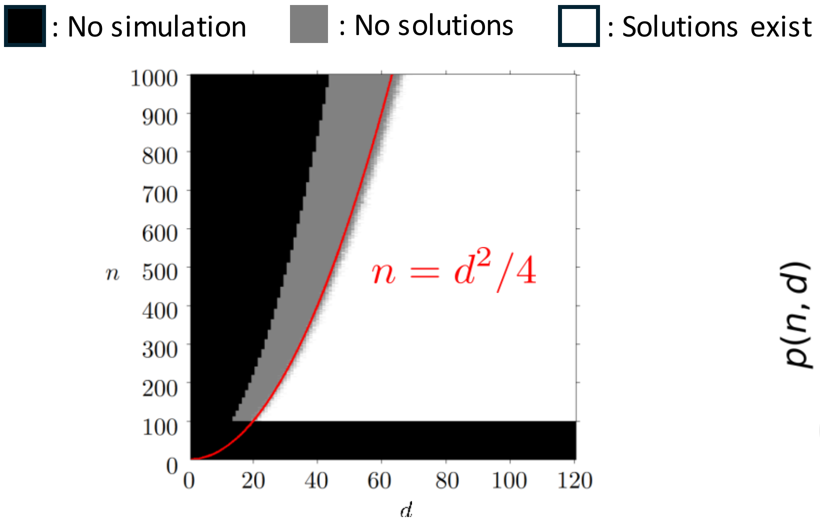
$$\boxed{\mathbb{P}[\exists S \in \mathbb{R}^{d\times d} \,:\, S \succeq 0 \text{ and } x_i^\top S x_i = 1 \text{ for all } i \in [n]] \text{ ?}}$$

A **semidefinite program (SDP)**
- Convex
- Efficient algorithms

Motivations

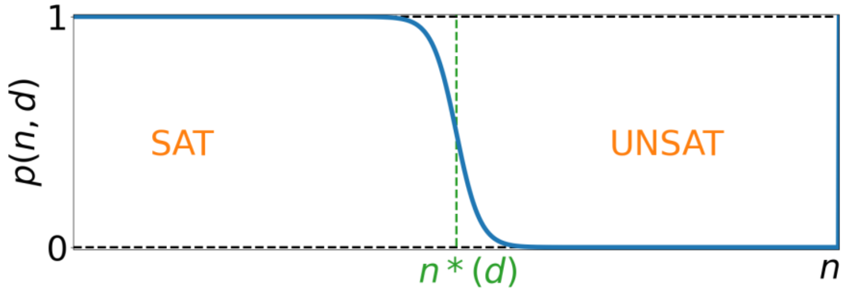
- Statistical inference
  - Minimum Trace Factor Analysis: Saunderson & al '12
  - Independent Components Analysis: Podosinnikova & al '19
- Theoretical computer science
  - Discrepancy of random matrices: Potechin & al '22
  - Average-case characterization of SDPs: Hopkins & al '17, Barak & al '19, …

The ellipsoid fitting conjecture



Saunderson, James, et al. *SIAM Journal on Matrix Analysis and Applications* 2012

$$p(n,d) = \mathbb{P}[\text{An ellipsoid fit exists}]$$

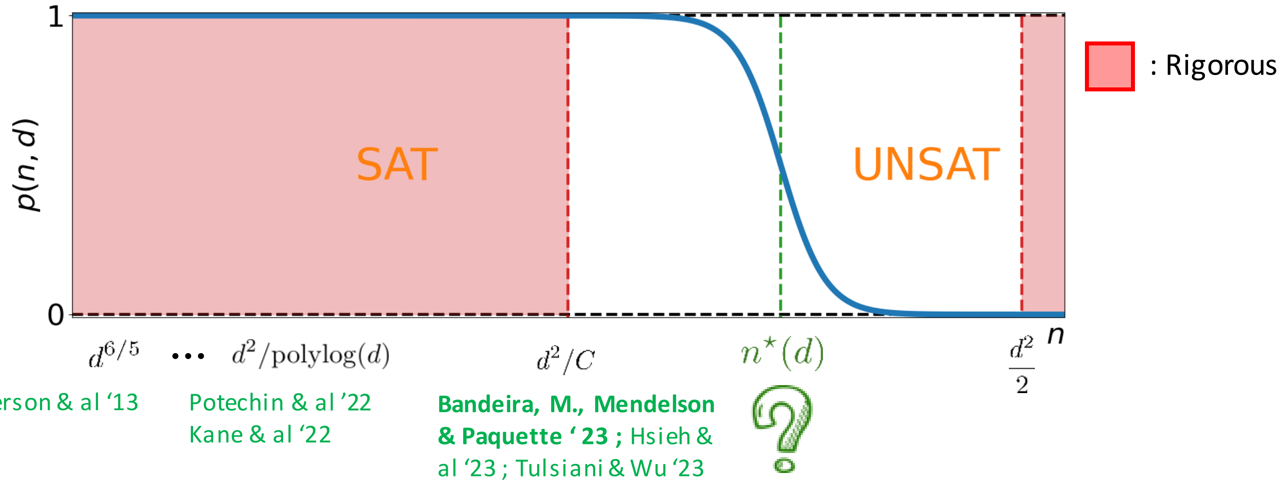


Open conjecture

$$\boxed{\lim_{d\to\infty} \frac{n^\star(d)}{d^2} = \frac{1}{4}}$$

The ellipsoid fitting conjecture: what is known

Conjecture $\boxed{\lim_{d\to\infty} \frac{n^\star(d)}{d^2} = \frac{1}{4}}$



Progress on lower bounds

Saunderson & al '13

Potechin & al '22
Kane & al '22

Bandeira, M., Mendelson & Paquette ' 23 ; Hsieh & al '23 ; Tulsiani & Wu '23

This talk

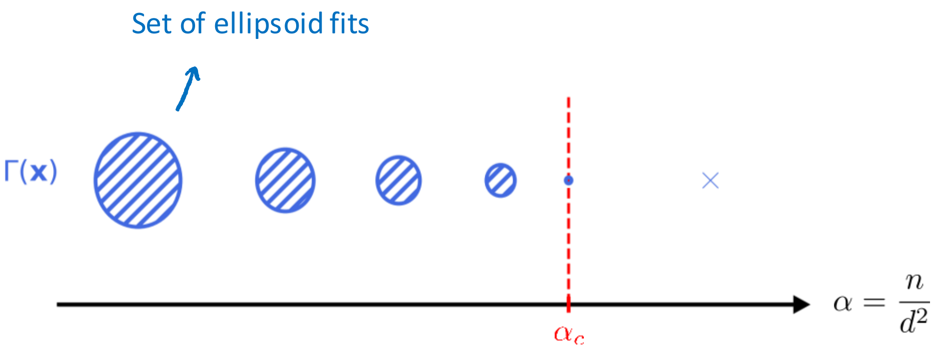
We see EFP as a **Random Constraint Satisfaction Problem**

$$\begin{cases} S \succeq 0 \\ \{x_i^\top S x_i = 1\}_{i=1}^n \end{cases}$$

→ "spectral" constraint

"disordered" model

Disordered systems tools for ellipsoid fitting



Volume of solutions / **"Partition function"**

$\mathrm{supp}(P_0) \subseteq \mathcal{S}_d^+$

$$\mathcal{Z} := \int P_0(\mathrm{d}S) \prod_{i=1}^{n} \delta(x_i^\top S x_i - 1)$$

Replica method and numerics: **M. & Kunisky '23** $\longrightarrow$

- Derivation of $\alpha_c = \frac{1}{4}$
- Analytical expressions for $\underset{d\to\infty}{\text{p-lim}} \frac{1}{n} \log \mathcal{Z}$
- ...

Non-rigorous analytical methods

Rigorous disordered systems tools for ellipsoid fitting (M. & Bandeira '23)

$$\mathcal{Z} := \int P_0(\mathrm{d}S) \prod_{i=1}^{n} \delta(x_i^\top S x_i - 1)$$

$$\mathcal{Z}_G := \int P_0(\mathrm{d}S) \prod_{i=1}^{n} \delta(\mathrm{Tr}(SG_i) - 1)$$

I: • "**Gaussian universality**" lemma : $\frac{1}{n} \log \mathcal{Z} \simeq \frac{1}{n} \log \mathcal{Z}_G$

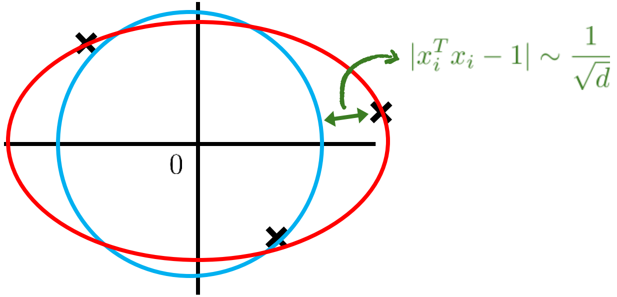
$x_i^\top S x_i \longrightarrow \mathrm{Tr}(SG_i)$ Gaussian matrix

II: • Random convex geometry tools for $\mathcal{Z}_G$ (Gordon '88 ; Amelunxen & al '14, ...)

Theorem (informal)

EFP' $\exists S \in \mathbb{R}^{d\times d} \, : \, S \succeq 0 \text{ and } |x_i^\top S x_i - 1| \ll \frac{1}{\sqrt{d}} \text{ for all } i \in [n]$

has a SAT / UNSAT transition at $n^\star(d) \simeq \frac{d^2}{4}$



First rigorous characterization of the SAT/UNSAT transition in (approximate) ellipsoid fitting