From fitting ellipsoids to random points, to learning in large neural networks

Antoine Maillard

ETH zürich

- arXiv:2310.01169 *(w. D. Kunisky)* – IEEE Trans. Inf. Theory '24
- arXiv:2310.05787 *(w. A. Bandeira)*
- arXiv:2408.03733 *(w. E. Troiani, S. Martin, F. Krzakala, L. Zdeborová)*

Roccella – September 4th 2024

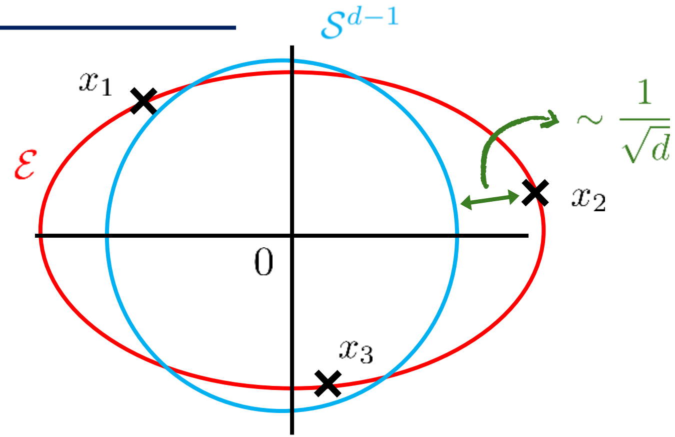
Part I: Fitting ellipsoids to random points

Fitting ellipsoids to random points

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

$$n, d \to \infty$$

Does $\mathcal{E}$ exist ?



Ellipsoid Fitting Property

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

Principal axes of $\mathcal{E}$ $\Longleftrightarrow$ Eigenspaces of S

$$r_i(\mathcal{E}) = \lambda_i(S)^{-1/2}$$

Some motivations

- Low-rank matrix decomposition
  Saunderson & al '12 ; '13 ; '13
- Independent Components Analysis
  Podosinnikova & al '19
- Discrepancy of random matrices
  Potechin & al '22
- Neural networks with quadratic activations
  More on that later !

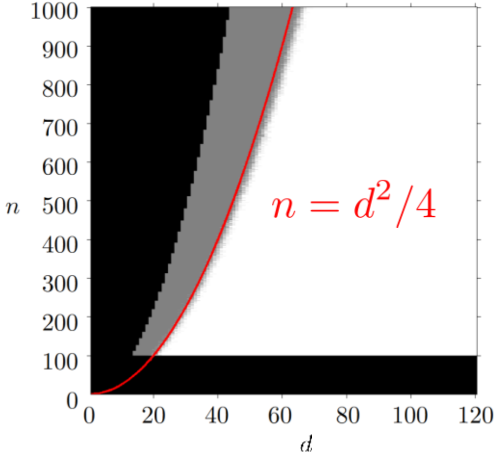
The ellipsoid fitting conjecture

Ellipsoid fitting is a **semidefinite program**

⇩

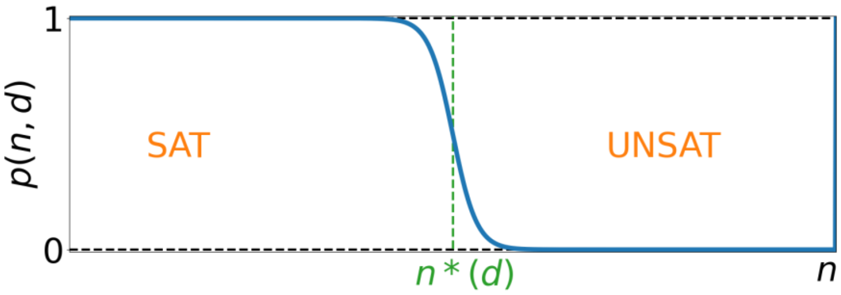
Convex problem + efficient solvers

■ : No simulation ▨ : No solutions □ : Solutions exist



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

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

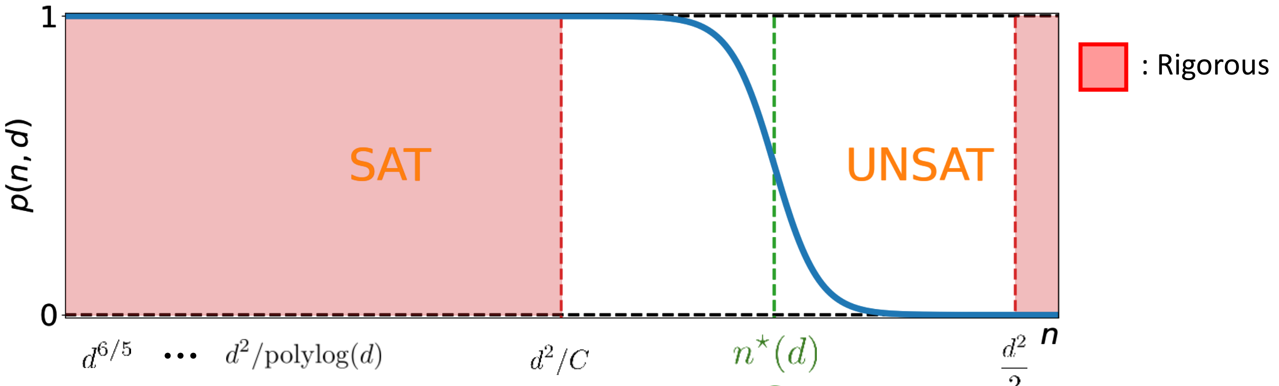


Open conjecture

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

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

?

Dimension counting
$\dim(\{S = S^\top\}) \simeq d^2/2$

This talk

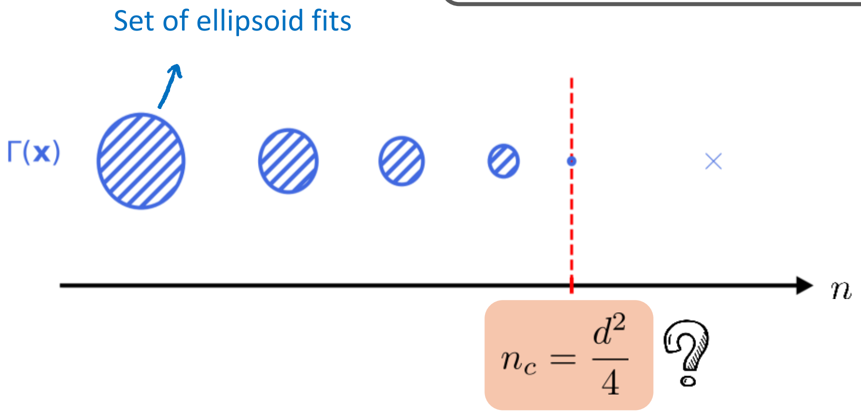
We see EFP as a **Random Constraint Satisfaction Problem**

$$\begin{cases} S \succeq 0 & \rightarrow \text{“spectral” constraint} \\ \{x_i^\top S \underbrace{x_i}_{\text{“disordered” model}} = 1\}_{i=1}^n \end{cases}$$

Statistical physics of ellipsoid fitting

Ellipsoid Fitting Property

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



$$n_c = \frac{d^2}{4} \ ?$$

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 calculation of $\mathbb{E} \log \mathcal{Z}$

- **Analytical derivation** of the threshold $n_c = d^2/4$
- **Shape** (spectrum) of typical ellipsoid fits
- Extensions to **non–Gaussian** vectors

[**M.** & Kunisky '23]

Connections to "HCIZ" integrals in random matrix theory [Matytsin '94 ; Guionnet&al'02]

Mathematical physics for ellipsoid fitting [M. & Bandeira '23]

Two-steps proof

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

[Goldt & al '22, Montanari & Saeed '22, Hu & Lu '22, …]

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

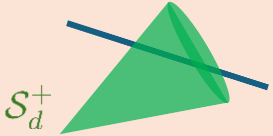
CRITICAL

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

II: • **Random convex geometry** tools for $\mathcal{Z}_G$

Extensions of Gordon's **min-max theorem**

[Gordon '88, Amelunxen & al'14]



$\{\mathrm{Tr}(SG_i) = 1,\ \forall i \in [n]\}$

uniformly randomly oriented

Theorem: The problem associated to $\mathcal{Z}_G$ is
- SAT (whp) if $n \leq (1-\varepsilon)\omega(\mathcal{S}_d^+)^2$
- UNSAT (whp) if $n \geq (1+\varepsilon)\omega(\mathcal{S}_d^+)^2$ ←

Gaussian width

$$\omega(\mathcal{S}_d^+) := \mathbb{E} \max_{\substack{S \succeq 0 \\ \|S\|_F = 1}} \mathrm{Tr}[GS]$$

$$\omega(\mathcal{S}_d^+) \sim_{d\to\infty} \frac{d}{2} \Rightarrow \boxed{n^\star(\mathcal{Z}_G) \sim \frac{d^2}{4}}$$

Mathematical physics for ellipsoid fitting [M. & Bandeira ’23]

I: “**Gaussian universality**” lemma ✚ II: **Random convex geometry** tools

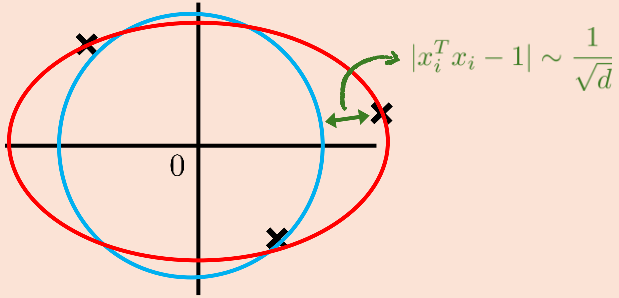
Theorem

EFP$_{\varepsilon,\, M}$ **:** $\exists S \in \mathbb{R}^{d\times d} \,:\, \mathrm{Sp}(S) \subseteq [0, M] \text{ and } \frac{1}{n}\sum_{i=1}^{n} |x_i^T S x_i - 1| \le \frac{\varepsilon}{\sqrt{d}}$

$S \succeq 0 \text{ and } |x_i^T S x_i - 1| \text{ for all } i \in [n]$

EFP $=$ **EFP**$_{0,\,\infty}$

$$n/d^2 \to \alpha \begin{cases} \alpha < 1/4 \quad \exists M_\alpha \,:\, \forall \varepsilon > 0,\ \mathbb{P}[\mathbf{EFP}_{\varepsilon, M_\alpha}] \to_{d\to\infty} 1 \\ \alpha > 1/4 \quad \exists \varepsilon_\alpha \,:\, \forall M > 0,\ \mathbb{P}[\mathbf{EFP}_{\varepsilon_\alpha, M}] \to_{d\to\infty} 0 \end{cases}$$



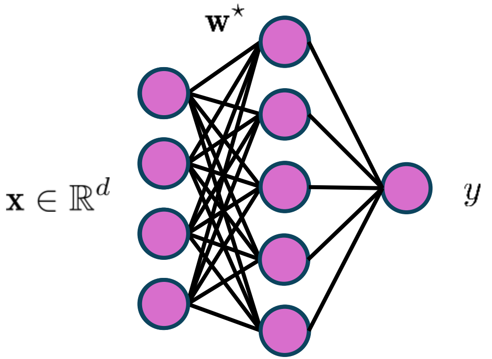
- Strengthen proof to **exact** ellipsoid fitting ?
- Extension to **other high-dimensional semidefinite programs** ?
- What does it have to do with **learning in neural networks** ??

Part II : Learning in neural networks

A two-layer neural network with quadratic activation

[**M.**, Troiani, Martin, Krzakala, Zdeborová '24]

Teacher network



$$y_i = f_{\mathbf{W}^*}(\mathbf{x}_i) := \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T \cdot \mathbf{x}_i\right]^2$$

$\sim \mathcal{N}(0, \mathrm{I}_d)$

$\mathbf{w}_k^\star \sim \mathcal{N}(0, \mathrm{I}_d)$

High-dimensional limit

$$d \to \infty \, ; \, m = \Theta(d)$$

Learning from data

$$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \Rightarrow \boxed{\mathbf{W}^\star} \; ?$$

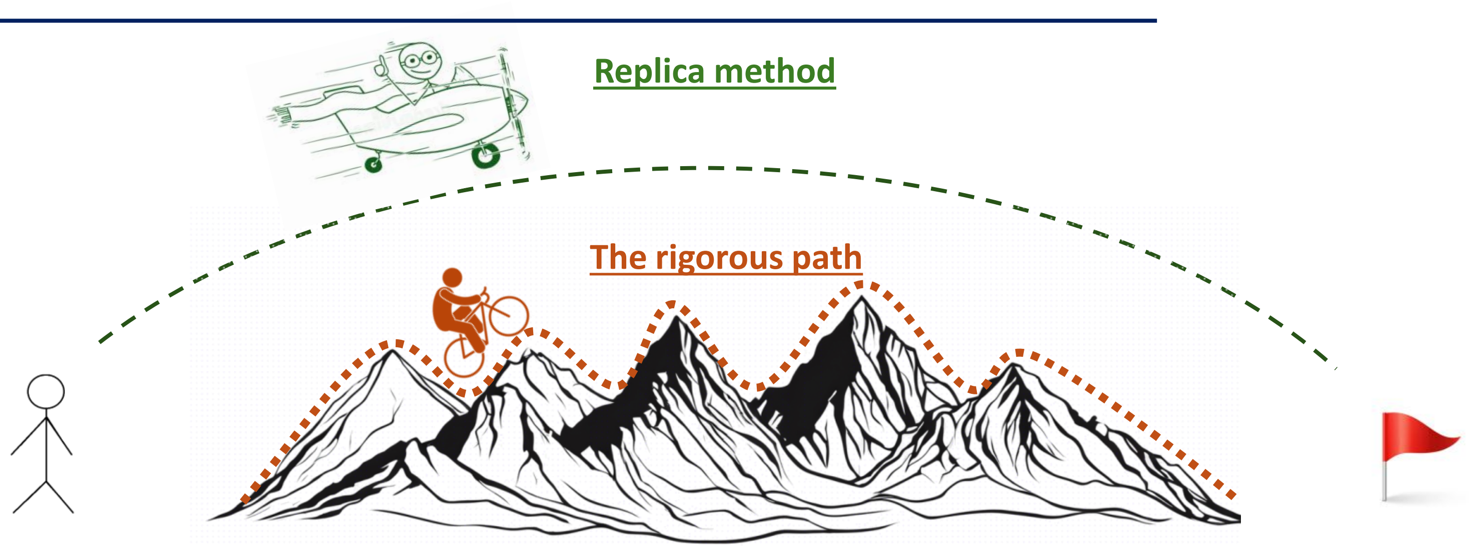
Bayes-optimal generalization error

$$\mathcal{E}_{\text{gen.}} := \mathbb{E}_{\mathbf{W}^\star, \{\mathbf{x}_i\}} \min_{\hat{y}(\{y_i, \mathbf{x}_i\})} \mathbb{E}_{\mathbf{x}_{\text{test}}}[(\hat{y}(\mathbf{x}_{\text{test}}) - f_{\mathbf{W}^\star}(\mathbf{x}_{\text{test}}))^2]$$

- If $n = \mathcal{O}(d)$, the optimal error can be reached by **linear regression**… Cui&al '23
- But there are $\Theta(d^2)$ weights to learn…

What happens for $n = \Theta(d^2)$?

All roads lead to Rome



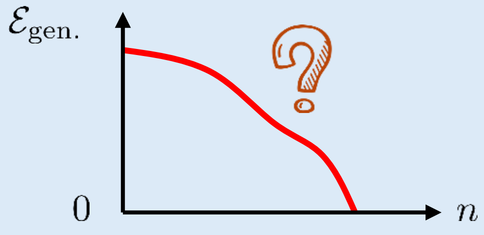
Replica method

The rigorous path

$$m = \Theta(d) \qquad n = \Theta(d^2)$$

$$\left\{ y_i = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T \cdot \mathbf{x}_i\right]^2 \right\}_{i=1}^{n}$$

$$\lim_{d\to\infty} \mathcal{E}_{\text{gen.}} = f\left(\frac{n}{d^2}, \frac{m}{d}\right)$$



Taking the long road

Step 0: $$y = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T\cdot\mathbf{x}\right]^2 = \frac{1}{d}\mathbf{x}^\top\mathbf{S}^\star\mathbf{x} = \mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}]$$

$$\mathbf{S}^\star := \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^\star(\mathbf{w}_k^\star)^\top \sim \mathcal{W}_{m,d} \qquad \mathbf{\Phi} := \frac{1}{d}\mathbf{x}\mathbf{x}^\top$$

Can be generalized to **noisy pre-activations**

$$\mathbf{w}_k^\star\cdot\mathbf{x} \to \mathbf{w}_k^\star\cdot\mathbf{x} + \sqrt{\Delta}\xi_k$$

$$y \sim P_{\mathrm{out}}\left(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}]\right)$$

Goal: $\{y_i, \mathbf{x}_i\}_{i=1}^n \Rightarrow \hat{\mathbf{S}}_{\mathrm{opt.}} = \arg\min \mathcal{E}_{\mathrm{gen.}}(\hat{\mathbf{w}}_k) = \arg\min \|\hat{\mathbf{S}} - \mathbf{S}^\star\|_F^2$

$\simeq$ **planted "ellipsoid fitting-like" problem**

Step 1 : "Gaussian universality"

$n = \Theta(d^2)$

⚠ Same scaling regime as ellipsoid fitting !

Universality of Bayes-optimal generalization error

$$\min \mathcal{E}_{\mathrm{gen.}}(\hat{\mathbf{w}}_k) = \min \|\hat{\mathbf{S}} - \mathbf{S}^\star\|_F^2 \quad = \quad \min \widetilde{\mathcal{E}}_{\mathrm{gen.}}(\hat{\mathbf{S}}) = \min \|\hat{\mathbf{S}} - \mathbf{S}^\star\|_F^2 \times(1+o(1))$$

from $\{y_i \sim P_{\mathrm{out}}\left(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}_i]\right)\}_{i=1}^n$ from $\{\tilde{y}_i \sim P_{\mathrm{out}}(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{G}_i])\}_{i=1}^n$

Gaussian matrix

Leverages our ellipsoid fitting analysis [**M.** & Bandeira '23]

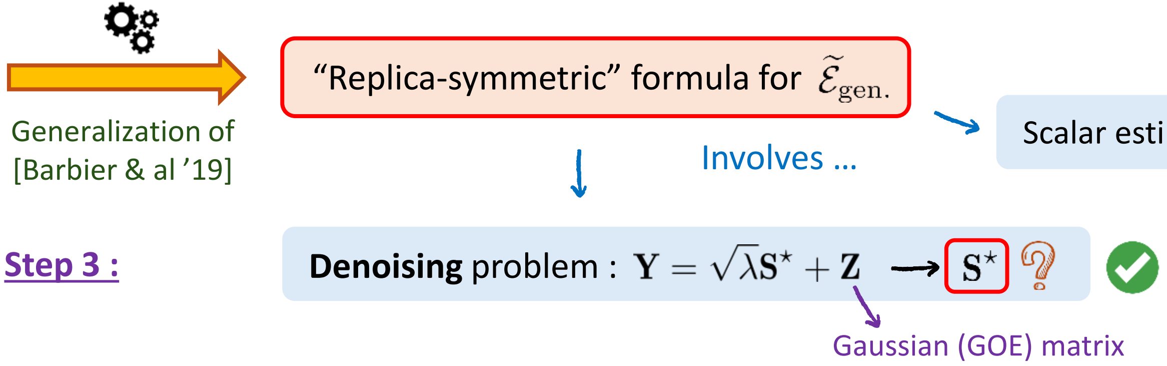
CRITICAL

Taking the long road

Step 2 : $\{\tilde{y}_i \sim P_{\text{out}}(\cdot|\mathrm{Tr}[\mathbf{S}^\star \mathbf{G}_i])\}_{i=1}^n$

Just a (generalized) **linear model on $\mathbf{S}^\star$**, with…

- Gaussian data $\mathbf{G} := \begin{pmatrix} \mathrm{flatt}(\mathbf{G}_1) \\ \vdots \\ \mathrm{flatt}(\mathbf{G}_n) \end{pmatrix}$ ✚ Wishart prior $\mathbf{S}^\star \sim \mathcal{W}_{m,d}$



Generalization of [Barbier & al '19]

"Replica-symmetric" formula for $\widetilde{\mathcal{E}}_{\text{gen.}}$

Involves …

Scalar estimation problem involving P_{out} ✓

Step 3 : **Denoising** problem : $\mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \boxed{\mathbf{S}^\star}$? ✓

Gaussian (GOE) matrix

[Bun & al '16 ; **M.**, Krzakala & al '22 ; Pourkamali & al '23 ; Semerjian '24 ; …]

❑ The optimal estimator is **spectral :** $\mathbf{Y} = \mathbf{O}\mathbf{D}\mathbf{O}^\top \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O} f_{\text{opt.}}(\mathbf{D})\mathbf{O}^\top$

❑ Analytical expressions for $f_{\text{opt.}}$ and the **asymptotic MMSE** $\lim_{d\to\infty} \|\hat{\mathbf{S}}(\mathbf{Y}) - \mathbf{S}^\star\|_F^2$

Taking the long road

$$\left\{ y_i = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T\cdot\mathbf{x}_i + \sqrt{\Delta}\xi_k\right]^2\right\}_{i=1}^{n}$$

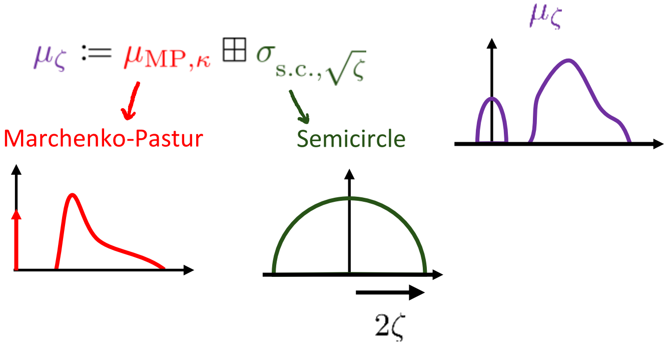
$$d \to \infty \qquad \begin{aligned} m &= \kappa d \\ n &= \alpha d^2 \end{aligned}$$

Combining all steps...

$$\boxed{\lim_{d\to\infty} \mathcal{E}_{\text{gen.}} = 2\kappa\alpha\zeta - \Delta(2+\Delta)}$$

ζ solves the self-consistent equation

$$(1-2\alpha) + \frac{\Delta(2+\Delta)}{\kappa\zeta} = \frac{4\pi^2\zeta}{3}\int \mu_\zeta(y)^3\,\mathrm{d}y$$



$$\mu_\zeta := \mu_{\mathrm{MP},\kappa} \boxplus \sigma_{\mathrm{s.c.},\sqrt{\zeta}}$$

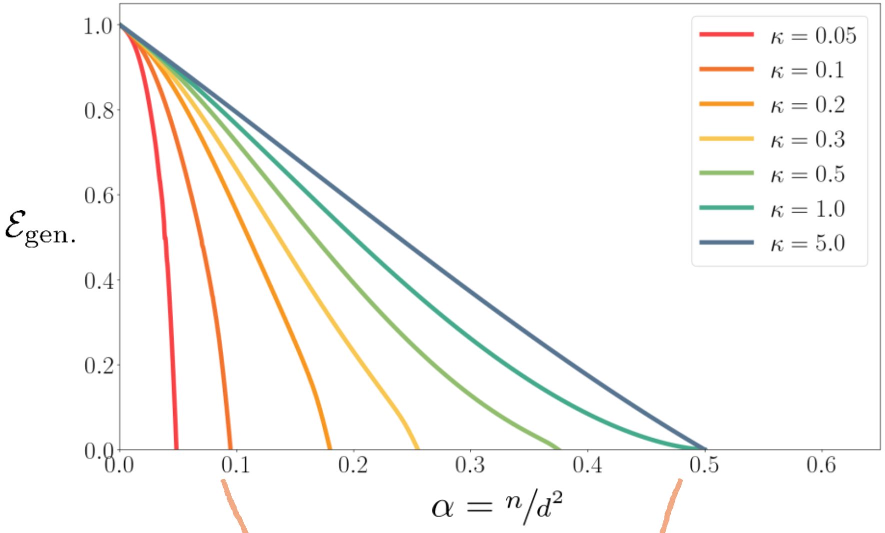
Marchenko-Pastur

Semicircle

- ➢ **Easy-to-evaluate formula** for the Bayes-optimal generalization error
- ➢ Not a fully rigorous theorem, some mathematical subtleties in **Steps 1 and 2**

Optimal generalization error

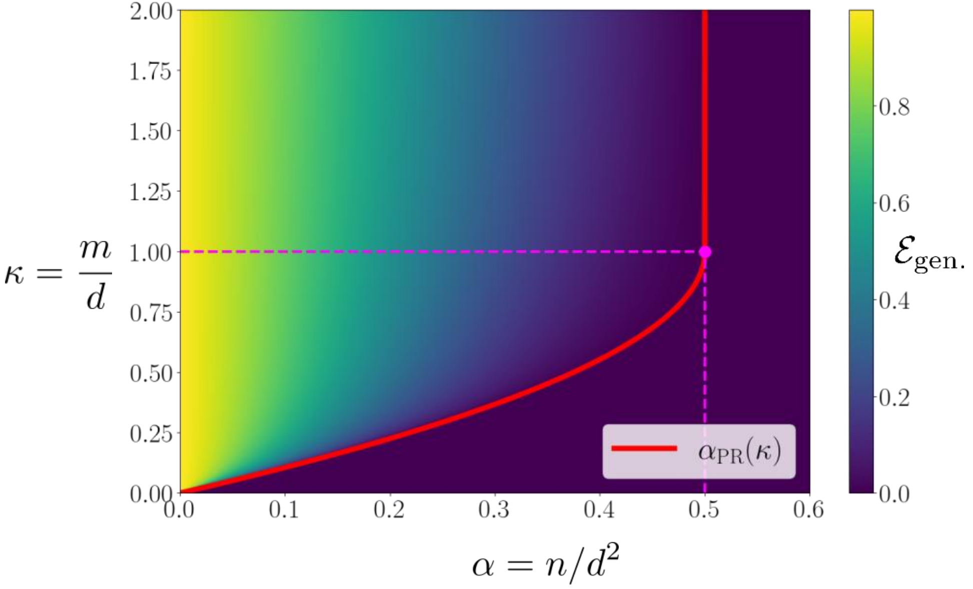
Intensive width $\kappa = {}^m/_d$; Sample complexity $\alpha = {}^n/_{d^2}$



$$\alpha_{\text{PR}}(\kappa) = \min\left(\kappa - \frac{\kappa^2}{2}, \frac{1}{2}\right)$$

Perfect recovery threshold

Noiseless setting : $\Delta = 0$



Matches a naïve "counting argument" $\text{DOF}[\{\mathbf{S} : \mathbf{S} = \mathbf{S}^\top \text{ and } \text{rk}(\mathbf{S}) \leq \kappa d\}] \simeq \alpha_{\text{PR}}(\kappa) d^2$

The GAMP-RIE algorithm

$y_i \sim P_{\mathrm{out}}\left(\cdot|\mathrm{Tr}[\mathbf{S}^\star \mathbf{\Phi}_i]\right)$

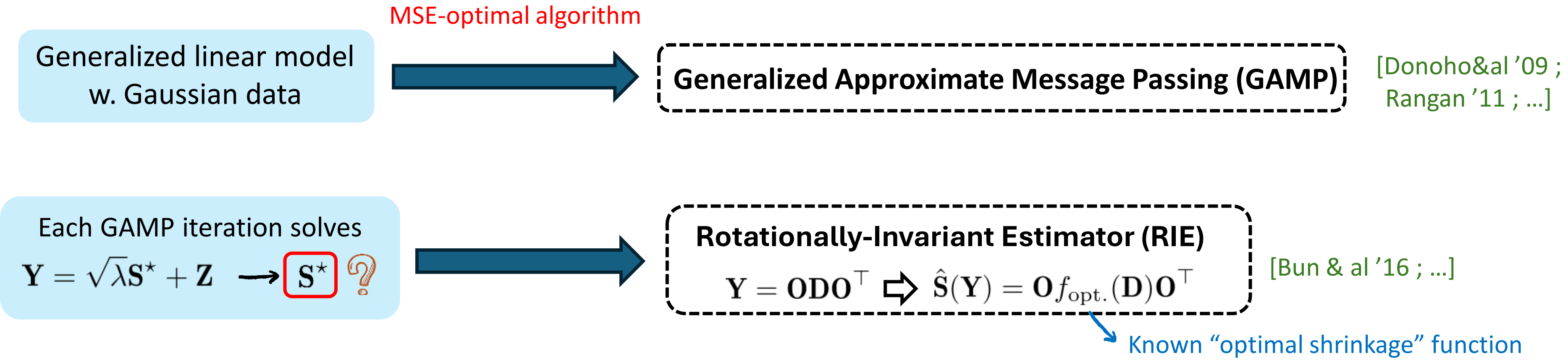
$\mathbf{\Phi}_i := (\mathbf{x}_i\mathbf{x}_i^\top - \mathrm{I}_d)/\sqrt{d}$

$\mathbf{S}^\star \sim \mathcal{W}_{m,d}$ (Wishart)

P_{out} : Noise channel

Informal hypothesis

Universality $\mathbf{\Phi}_i \Rightarrow \mathbf{G}_i$ also holds "at the level of algorithms"



Generalized linear model w. Gaussian data

MSE-optimal algorithm

Generalized Approximate Message Passing (GAMP)

[Donoho&al '09 ; Rangan '11 ; …]

Each GAMP iteration solves

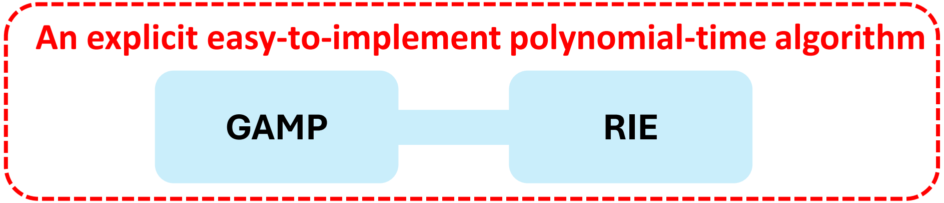
$\mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \boxed{\mathbf{S}^\star}$

Rotationally-Invariant Estimator (RIE)

$\mathbf{Y} = \mathbf{O}\mathbf{D}\mathbf{O}^\top \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O} f_{\mathrm{opt.}}(\mathbf{D})\mathbf{O}^\top$

[Bun & al '16 ; …]

Known "optimal shrinkage" function



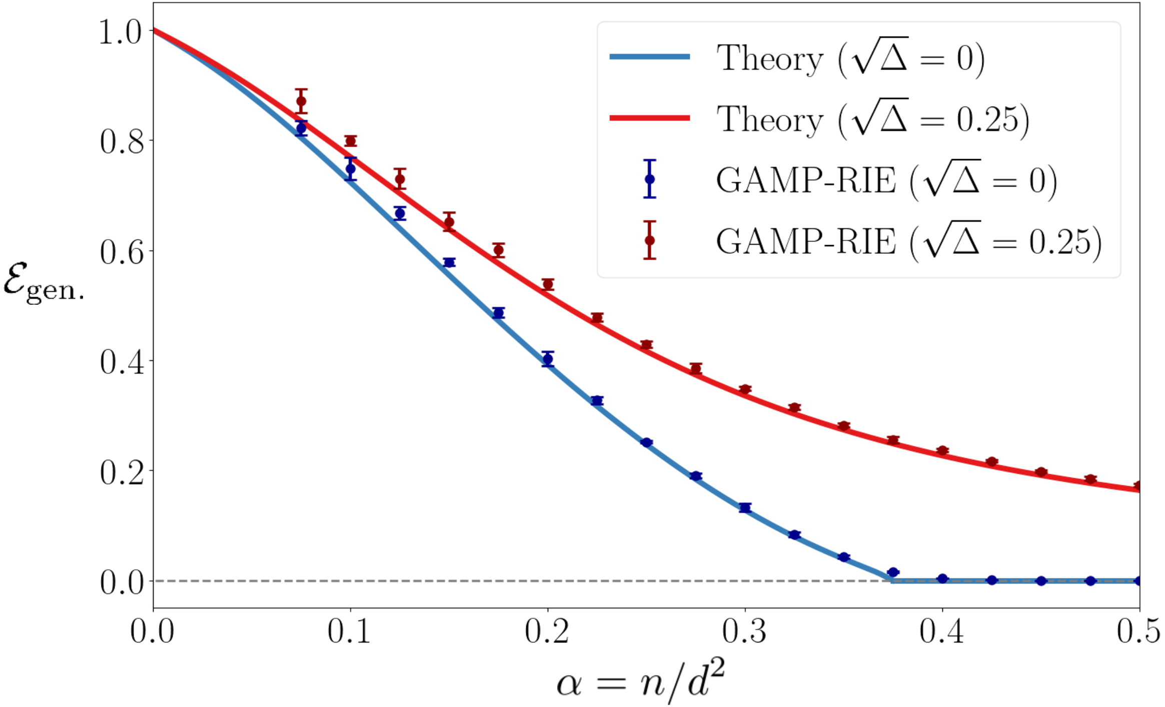
An explicit easy-to-implement polynomial-time algorithm

Absence of hard phase

Intensive width $\kappa = {}^{m}/_{d}$; Sample complexity $\alpha = {}^{n}/_{d^2}$

$\kappa = 0.5$
$d = 200$

$$y_i = f_{\mathbf{W}^*}(\mathbf{x}_i) := \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T \cdot \mathbf{x}_i + \sqrt{\Delta}\xi_{k,i}\right]^2$$



No computational-to-statistical gap / hard phase

For $m = \mathcal{O}(1)$ there is a hard phase !

cf. [Barbier & al '19] for $m = 1$

Gradient descent : noiseless case

Intensive width $\kappa = {}^{m}/_{d}$; Sample complexity $\alpha = {}^{n}/_{d^2}$

$$\mathcal{L}(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{f}_{\mathbf{W}}(\mathbf{x}_i)\right)^2 \text{, where } \tilde{f}_{\mathbf{W}}(\mathbf{x}) := \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k)^T \cdot \mathbf{x}\right]^2$$

For **any** κ, AGD seems to reach the Bayes-optimal MMSE

- For $\kappa \geq 1$ ($m \geq d$), the problem is **convex** over $\mathbf{S} := ({}^{1}/_{m})\sum_{k=1}^{m}\mathbf{w}_k\mathbf{w}_k^\top$

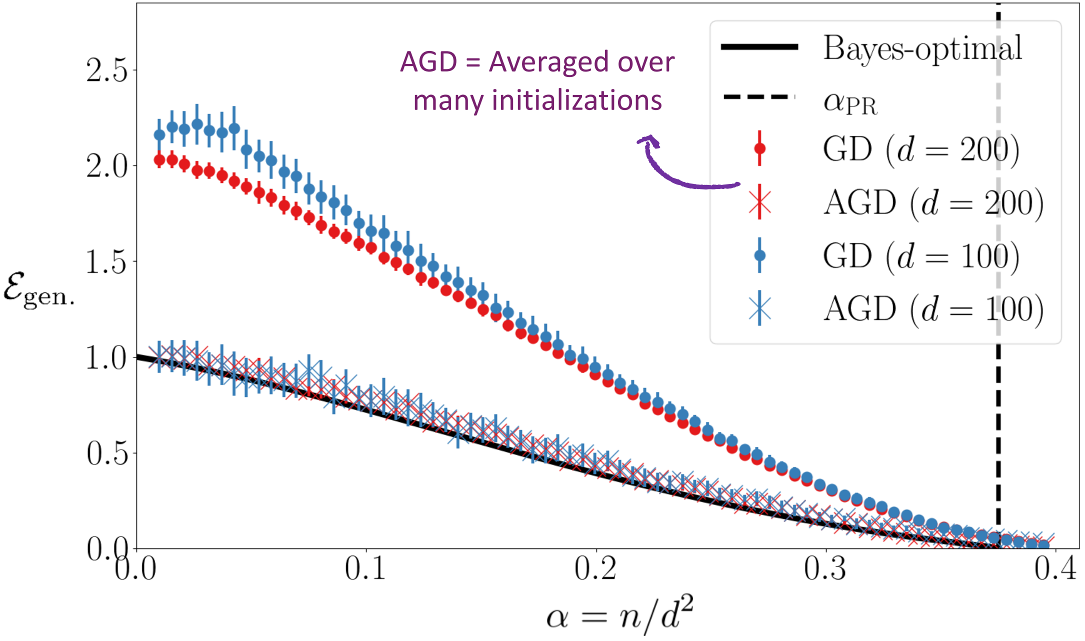
The landscape of $\mathcal{L}(\mathbf{W})$ trivializes in this regime

[Du & Lee '18 ; Soltanolkotabi & al '18 ; Venturi & al '19]

- For $\kappa < 1$, **non-convex problem**. Still, naïve GD reaches optimal error !

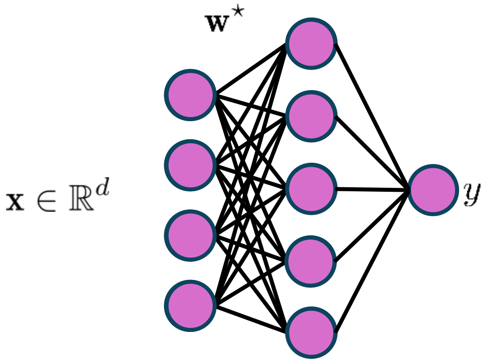
No longer true for noisy pre-activations.

$\kappa = {}^{m}/_{d} = {}^{1}/_{2}$



AGD = Averaged over many initializations

Summary



1. Analytical formula for the **Bayes-optimal generalization error.**

2. Optimal algorithm (GAMP-RIE), **no computational-statistical gap.**

3. (Averaged) **Gradient descent seems to sample from the posterior for noiseless pre-activations**, even in the non-convex regime $\kappa < 1$!

   Not true for noisy case

$$\left\{ y_i = f_{\mathbf{W}^*}(\mathbf{x}_i) := \frac{1}{m} \sum_{k=1}^{m} \left[\frac{1}{\sqrt{d}} (\mathbf{w}_k^*)^T \cdot \mathbf{x}_i + \sqrt{\Delta} \xi_{k,i} \right]^2 \right\}_{i=1}^{n}$$

$\sim \mathcal{N}(0, \mathrm{I}_d)$

$\mathbf{w}_k^\star \sim \mathcal{N}(0, \mathrm{I}_d)$

$$n = \alpha d^2 \;\; ; \;\; m = \kappa d$$

- **Other activations ?** (beyond quadratic)
  Other architectures ?

- Theoretical analysis of GD properties ?

- …

THANK YOU !