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

Antoine Maillard

ETH zürich

- arXiv:2310.01169 *(w. D. Kunisky)*
- arXiv:2310.05787 *(w. A. Bandeira)*
- arXiv:2406.???? *(w. E. Troiani, S. Martin, F. Krzakala, L. Zdeborová)*

LemanTh – May 29th 2024

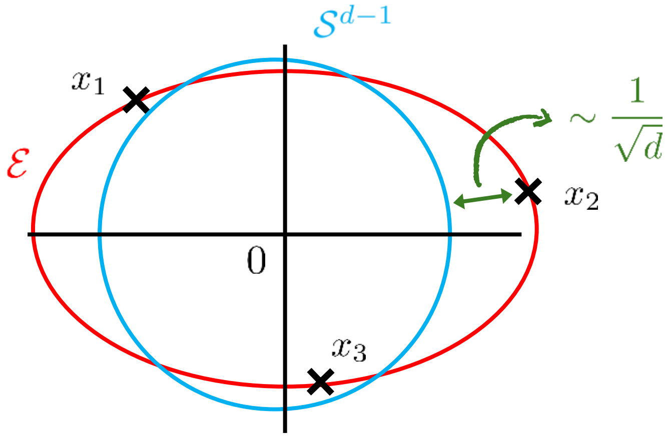
Part I: Fitting ellipsoids to random points

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

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

Fitting ellipsoids to random points

Ellipsoid Fitting Property

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

Some motivations

Potechin & al '22

- **Low-rank matrix decomposition** — Saunderson & al '12 ; '13 ; '13

  Recommendation systems, community detection, ...

  $$X = \underbrace{D^\star}_{\text{Diagonal}} + \underbrace{L^\star}_{\succeq 0 \text{ + low-rank}} \in \mathbb{R}^{n\times n}$$

  $$\mathrm{MTFA} := \min_{\substack{D,L \,:\, X = D + L \\ L \succeq 0}} \mathrm{Tr}(L)$$

  $$\mathrm{col}(L^\star) \sim \mathrm{Unif}[r-\text{dim subspaces}] \implies \mathbb{P}[\text{MTFA recovers } (L^\star, D^\star)] = p(n, n-r)$$

- **Independent Components Analysis** — Podosinnikova & al '19

  Signal processing

- **Discrepancy of random matrices** — Potechin & al '22

  SDP lower bounds certification

- **Neural networks with quadratic activations** — **More on that later !**

  Optimization, machine learning,...

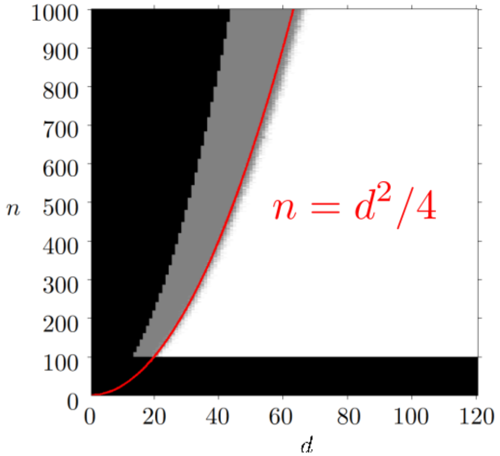
The ellipsoid fitting conjecture

Ellipsoid fitting is a **semidefinite program**

⇩

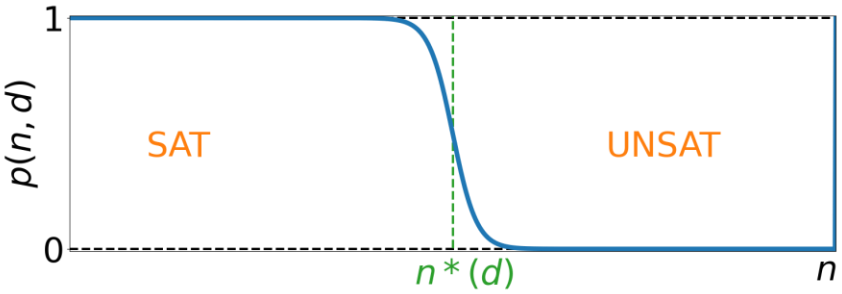
Convex problem + efficient solvers

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



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

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

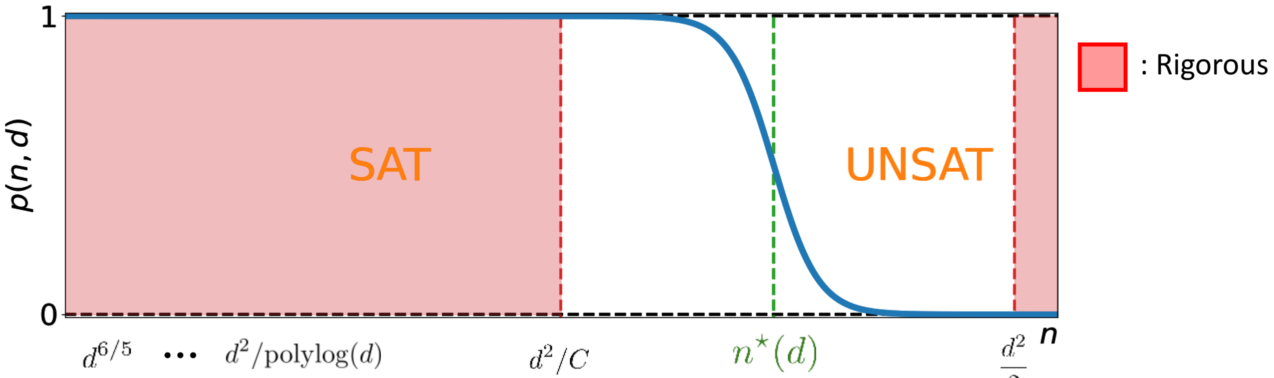


Open conjecture

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

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**

$$\begin{cases} S \succeq 0 & \longrightarrow \text{"spectral" constraint} \\ \{x_i^\top S x_i = 1\}_{i=1}^n & \end{cases}$$

"disordered" model

Statistical physics tools for ellipsoid fitting [M. & Kunisky '23]

Ellipsoid Fitting Property

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

Set of ellipsoid fits

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

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

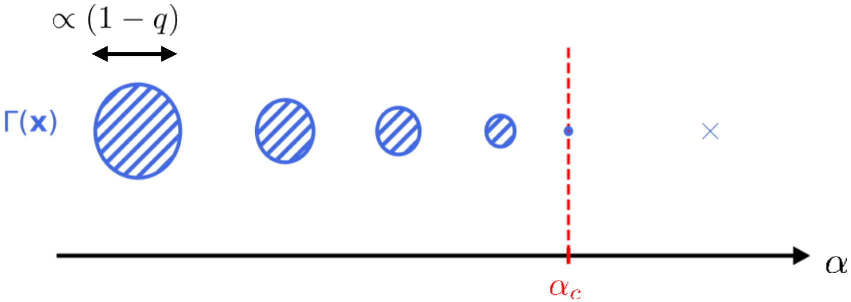
$\mathrm{supp}(P_0) \subseteq \mathcal{S}_d^+$

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

Statistical physics tools for ellipsoid fitting [M. & Kunisky '23]

$$^{n}/_{d^2} \to \alpha > 0$$

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



Replica method + convexity
("replica symmetry")

$$\frac{1}{d^2}\mathbb{E}\log\mathcal{Z} \to \sup_{q\in[0,1]} \sup_{\mu\in\mathcal{M}_1^+(\mathbb{R})} \left[F(\alpha,q,\mu) + I_{\mathrm{HCIZ}}\left(\frac{1}{\sqrt{1-q}}, \mu, \sigma_{\mathrm{s.c.}}\right)\right]$$

"Overlap"

Typical spectrum of solutions (ellipsoid shape)

$$I_{\mathrm{HCIZ}}(\theta, A, B) := \lim_{d\to\infty} \frac{1}{d^2} \log \int_{\mathcal{O}(d)} \mathcal{D}O \exp\{\theta \mathrm{Tr}[OAO^\top B]\}$$

Hard asymptotic expressions via PDEs [Matytsin '94 ; Guionnet&al'02]

$\alpha \to \alpha_c$
$q \to 1$

+

"Dilute" expansion ($\theta \to \infty$) of $I_{\mathrm{HCIZ}}(\theta, A, B)$ [Bun & al '16]

$$\boxed{\alpha_c = \frac{1}{4}}$$

+

- Computation of typical μ
- Extensions to non-Gaussian x_i
- …

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, …]

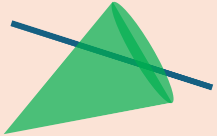
$x_i^\top S x_i \longrightarrow \mathrm{Tr}(S G_i)$ ← Gaussian matrix

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

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

Extensions of Gordon's **min-max theorem**
[Gordon '88, Amelunxen & al'14]

$\mathcal{S}_d^+$



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

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

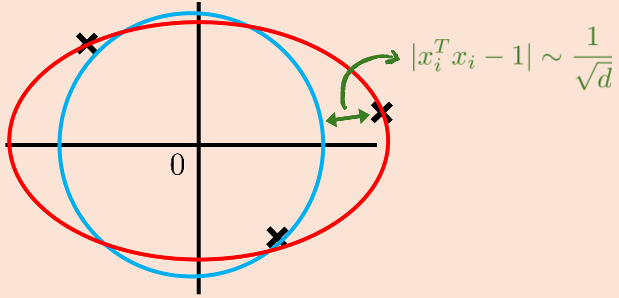
Theorem

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

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

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

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



Ellipsoid fitting: summary

1. Best-known **lower bound** $n^\star(d) \geq \frac{d^2}{C}$ — Bandeira, **M.**, Mendelson & Paquette '23

2. Refinement and extension of the conjecture to **non-Gaussian points.** — **M.** & Kunisky '23, to appear in IEEE Trans. Inf. Theory

3. Theorem: $\boxed{n^\star(d) = \frac{d^2}{4}}$ in **approximate ellipsoid fitting**. — **M.** & Bandeira '23

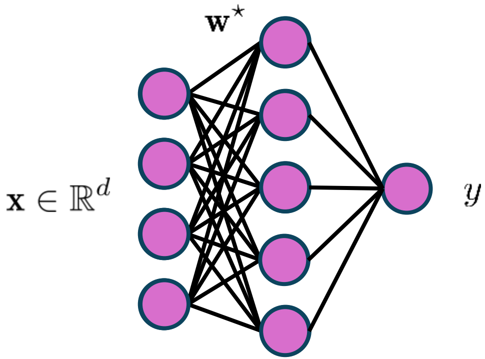
   First rigorous characterization of the transition

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

Part II : Learning in neural networks

Learning in large neural networks [**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

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

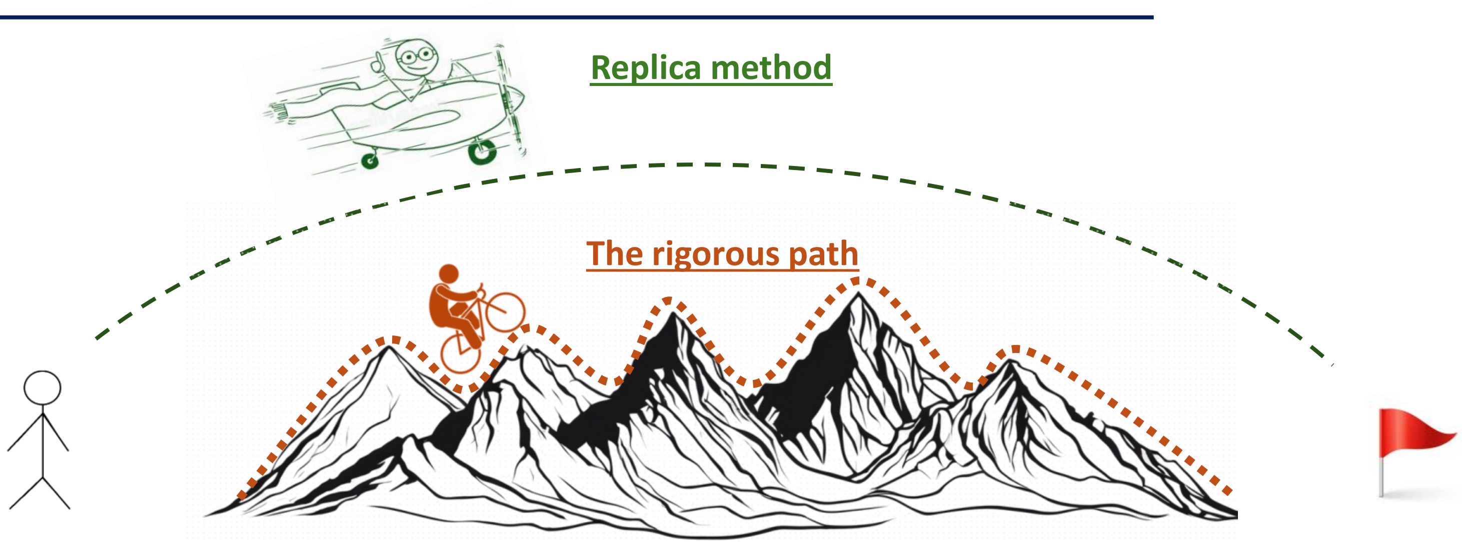
Bayes-optimal generalization error

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

- 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 $n = \Theta(d^2)$?

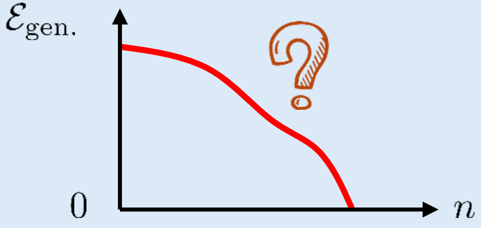
All roads lead to Rome



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

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

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



Taking the long road

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

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

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" $\quad n = \Theta(d^2)$ $\quad$ 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 $\{y_i \sim P_{\mathrm{out}}\left(\cdot|\mathrm{Tr}[\mathbf{S}^\star \mathbf{\Phi}_i]\right)\}_{i=1}^n$ $\quad$ 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]

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 :** $\mathbf{Y} = \mathbf{ODO}^\top \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O} f_{\text{opt.}}(\mathbf{D})\mathbf{O}^\top$
- 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

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

$$d \to \infty \qquad m = \kappa d \qquad n = \alpha d^2$$

Combining all steps...

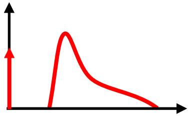
$$\boxed{\lim_{d \to \infty} \mathcal{E}_{\text{gen.}} = 2\kappa\alpha\zeta - \Delta(2 + \Delta)}$$

ζ solves the self-consistent equation

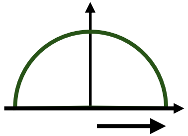
$$(1 - 2\alpha) + \frac{\Delta(2+\Delta)}{\kappa\zeta} = \frac{4\pi^2\zeta}{3} \int \mu_\zeta(y)^3 \, \mathrm{d}y$$

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

Marchenko-Pastur

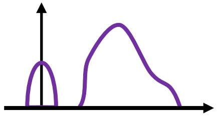


Semicircle



2ζ

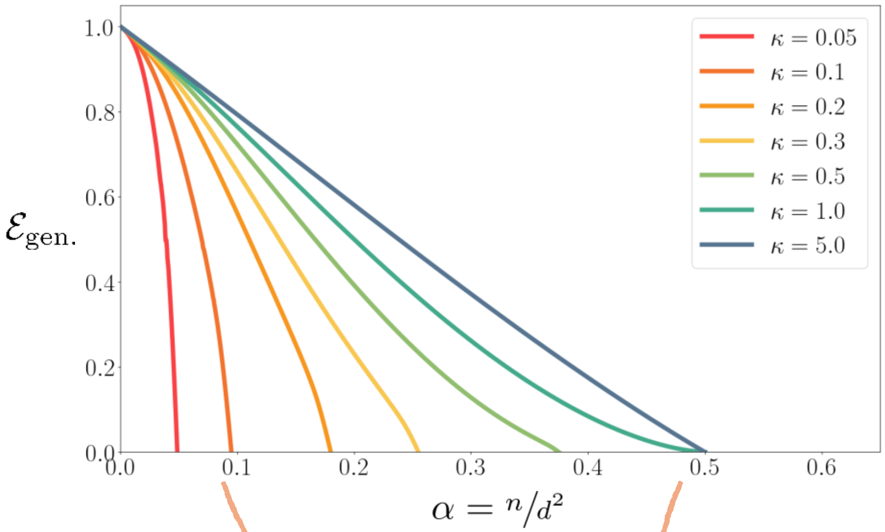
μ_ζ



- **Easy-to-evaluate formula** for the Bayes-optimal generalization error
- Not a fully rigorous theorem yet, work in progress in **Steps 1 and 2**

Optimal generalization error

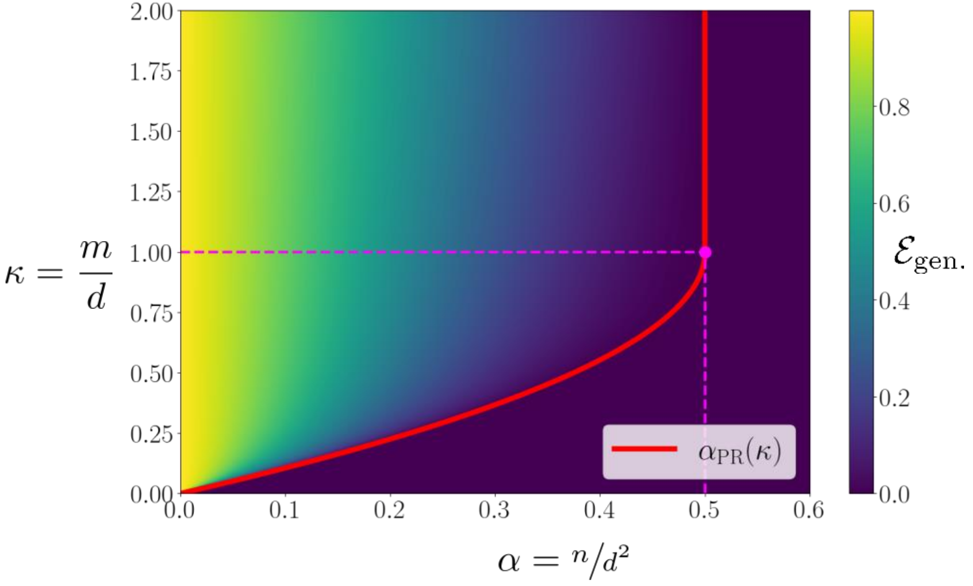
Intensive width $\kappa = {}^{m}/_{d}$; Sample complexity $\alpha = {}^{n}/_{d^2}$



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

Perfect recovery threshold

Noiseless setting : $\Delta = 0$



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

Gradient descent

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

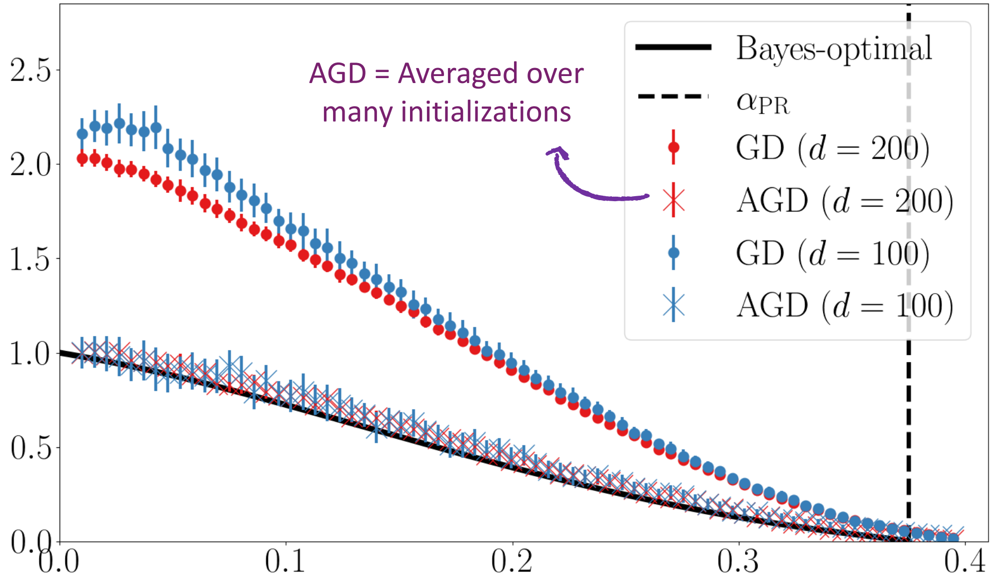
For **any** κ, AGD seems to reach the Bayes-optimal MMSE

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

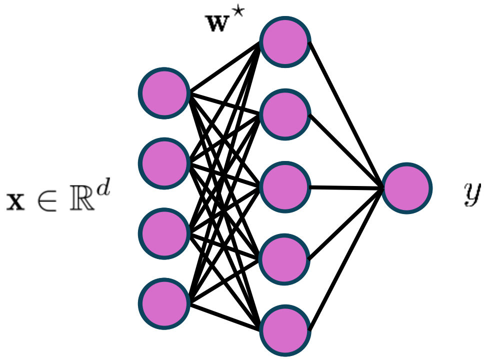
The landscape of $\mathcal{L}(\mathbf{W})$ trivializes in this regime
[Du & Lee '18 ; Soltanolkotabi & al '18 ; Venturi & al '19]

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

$\kappa = m/d = 1/2$



Summary



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

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

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

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

1. Analytical formula for the **Bayes-optimal generalization error.**
2. (Averaged) **Gradient descent seems to sample from the posterior**, even in the non-convex regime $\kappa < 1$!
3. Analysis (th. + exp.) is extended to **noisy pre-activations.**

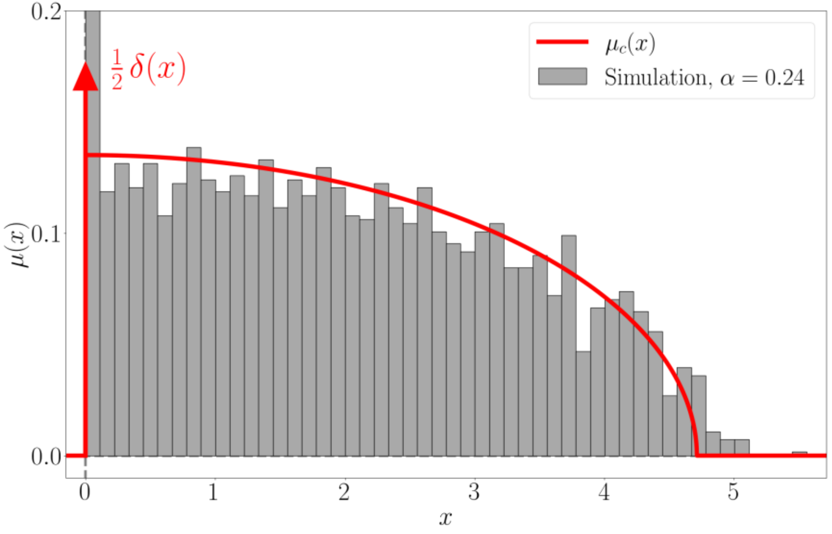
- What about **other activations ?** (beyond quadratic)
- Algorithms **provably reaching the MMSE** ?
- Theoretical analysis of GD properties ?
- …

THANK YOU !

Statistical physics tools for ellipsoid fitting [M. & Kunisky '23]

Spectrum of solutions / Shape of ellipsoids

Near the transition $\alpha \uparrow 1/4$



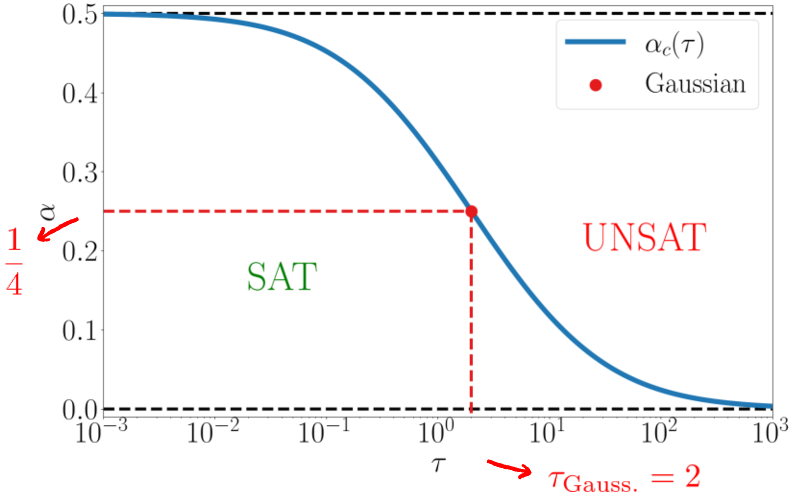
- Truncated semicircular distribution
- As $\alpha \uparrow 1/4$, ellipsoid fits are "cylinders" in $d/2$ directions !

Generalization to non-Gaussian random vectors

$x_i = \sqrt{r_i}\omega_i$

$\omega_i \sim \mathrm{Unif}(\mathcal{S}^{d-1})$

$\mathbb{E}[r_i] = 1$ + $\mathrm{Var}(r_i) = \frac{\tau}{d}$



Larger norm fluctuations $\Rightarrow$ Ellipsoid fits harder to find