Fundamental limits of high-dimensional estimation

A stroll between statistical physics, probability, and random matrix theory

Antoine Maillard

Under the supervision of Florent Krzakala

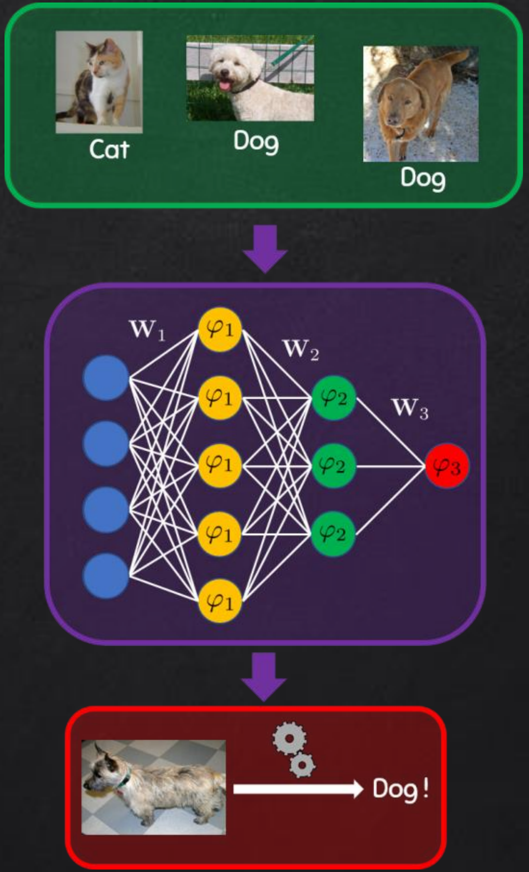
Département de Physique
École normale supérieure
ENS | PSL

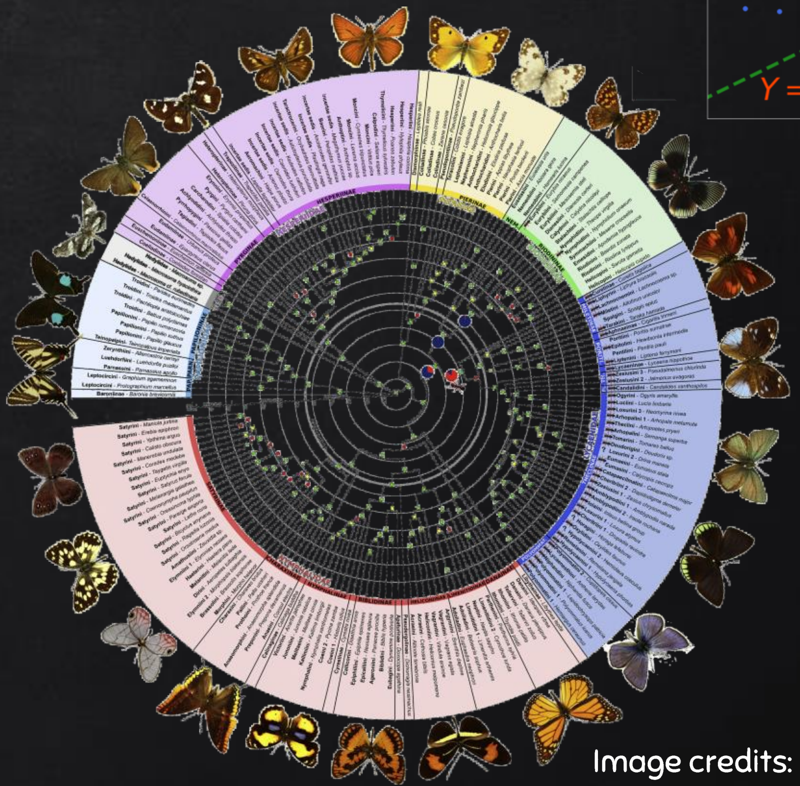
FONDATION CFM POUR LA RECHERCHE

PhD defense – August 30th 2021

What is statistical inference ?

- Supervised learning in "teacher-student" neural networks
- Signal processing
- Phase retrieval
- Matrix factorization
- Quantitative finance, particle physics, evolutionary biology…





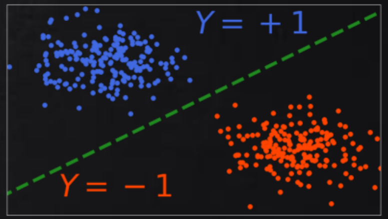


Image credits: [Espeland&al '18]

Statistics in high dimension

$$\left\{ \Phi_\mu \in \mathbb{R}^p \rightarrow \boxed{\text{Model } \mathbf{X}^\star \in \mathbb{R}^n} \rightarrow y_\mu \right\}_{\mu = 1, \cdots, m}$$

Data deluge

Gigantic databases and explosion of computing power.

↔

Theoretical revolution of the 2000s

High-dimensional statistics

[Donoho, AMS Lectures 2000: *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*]

"Modern machine learning": GoogLeNet [Szegedy&al '15]: $n \simeq 5 \times 10^6$ and $m \simeq 10^6$.

"High-dimensional" limit

Number of parameters $n \to \infty$
+
Number of data $m \to \infty$

In this presentation: $m/n \to \alpha > 0$ (sampling ratio).

Bayesian formalism

$\boldsymbol{\Phi}_\mu \rightarrow$ Model $\mathbf{X}^\star$ $\rightarrow y_\mu$

Posterior distribution

Observation channel, or likelihood

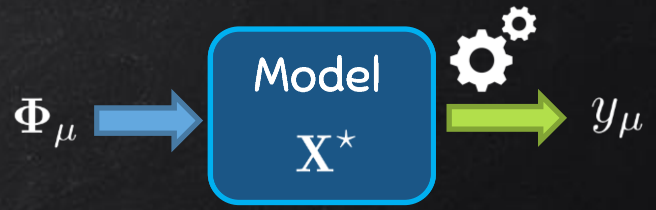
Prior distribution

$$\mathbb{P}(\mathbf{X}^\star = \mathbf{x}|\mathbf{Y}, \boldsymbol{\Phi}) = \frac{\mathbb{P}(\mathbf{Y}|\mathbf{X}^\star = \mathbf{x}, \boldsymbol{\Phi}) \times \mathbb{P}(\mathbf{X}^\star = \mathbf{x})}{\mathbb{P}(\mathbf{Y}|\boldsymbol{\Phi})}$$

Most of this talk: the prior and the observation channel are known to the statistician.

Bayes-optimal setting

Estimators



$\mathbb{P}(\mathbf{x}|\mathbf{Y}) \xrightarrow{\;?\;}$ Estimation of $\mathbf{X}^\star$

<u>Bayesian estimators:</u>

- *Maximum A Posteriori* $\hat{\mathbf{X}}_{\mathrm{MAP}} \equiv \arg\max_{\mathbf{x}} \mathbb{P}(\mathbf{x}|\mathbf{Y})$

- *Minimal Mean Squared Error* $\hat{\mathbf{X}}_{\mathrm{MMSE}} \equiv \arg\min_{\mathbf{x}} \left\{ \mathbb{E}_{\mathbf{Y}} \int \mathrm{d}\mathbf{x}' \mathbb{P}(\mathbf{x}'|\mathbf{Y}) \|\mathbf{x} - \mathbf{x}'\|^2 \right\}$

$$\boxed{\hat{\mathbf{X}}_{\mathrm{MMSE}} = \int \mathrm{d}\mathbf{x}\, \mathbb{P}(\mathbf{x}|\mathbf{Y})\, \mathbf{x} = \langle \mathbf{x} \rangle_{\mathbf{Y}}}$$

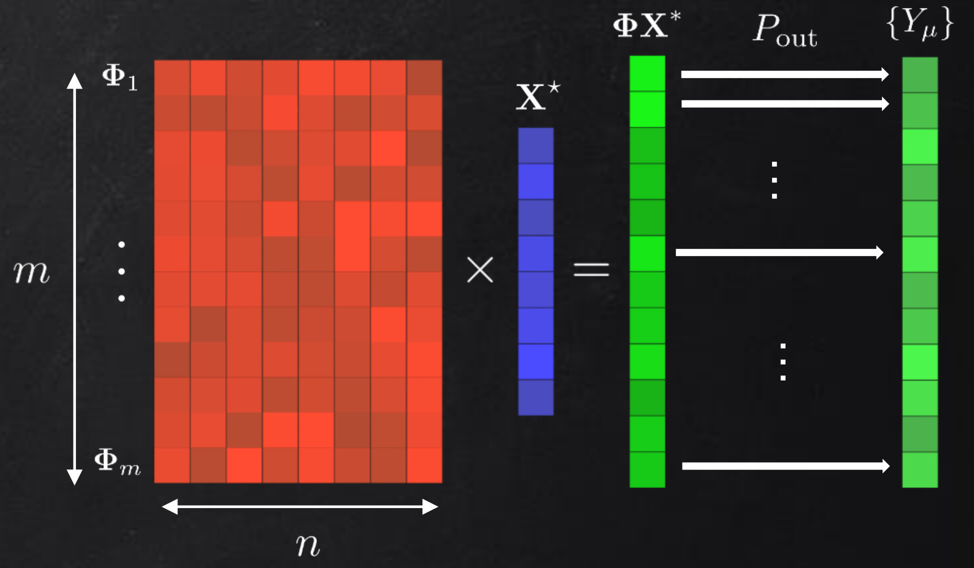
Many other types of estimators exist, such as the *Empirical Risk Minimizer* $\hat{\mathbf{X}} \equiv \arg\min_{\mathbf{x}} \sum_{\mu=1}^{m} L(x_\mu, y_\mu)$

"Loss" function

This presentation: we mainly focus on MMSE estimation and empirical risk minimization.

Important example – Generalized Linear Models

Goal : Recover $\mathbf{X}^\star \in \mathbb{R}^n$ from $\{\mathbf{\Phi}_\mu, Y_\mu\}_{\mu=1}^m$:



Observations $Y_\mu \in \mathbb{R}$

$$\boxed{Y_\mu \sim P_{\text{out}}\left(\cdot \,\middle|\, \frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi_{\mu i} X_i^\star\right) \quad \mu \in \{1, \cdots, m\}}$$

Probabilistic channel (noise)

Sensing matrix

$$Y_\mu = \frac{1}{n} |(\mathbf{\Phi X}^\star)_\mu|^2$$

Many examples: compressed sensing, perceptron learning, phase retrieval, ...

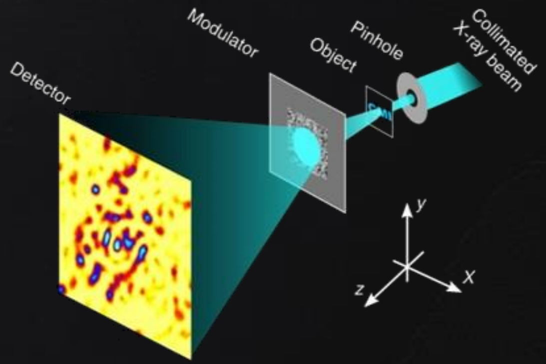


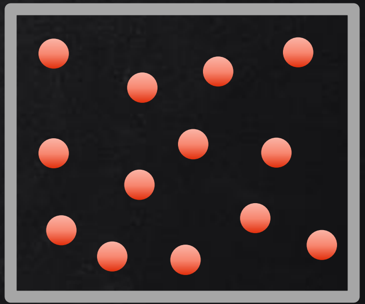
Image credits: [Zhang&al 16]

$$\boxed{\mathbb{P}(\mathbf{x}|\mathbf{Y}, \mathbf{\Phi}) = \frac{\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{Y}|\mathbf{x}, \mathbf{\Phi})}{\mathbb{P}(\mathbf{Y}|\mathbf{\Phi})} = \frac{1}{\mathcal{Z}_n(\mathbf{Y}, \mathbf{\Phi})} \prod_{i=1}^n P_X(x_i) \prod_{\mu=1}^m P_{\text{out}}\left[Y_\mu \middle| \frac{1}{\sqrt{n}} (\mathbf{\Phi x})_\mu\right].}$$

Prior knowledge on the signal

Where are the physics?

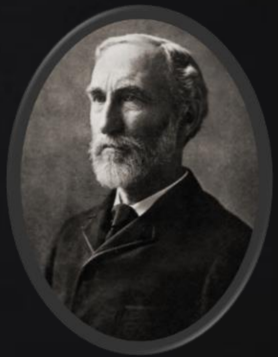
$$\mathcal{H} = \frac{m}{2}\sum_{i=1}^{n} v_i^2$$



"Statistical mechanics 101"

Consider n particles with position x_i, with distribution $P_X(x)$, interacting via the Hamiltonian $\mathcal{H}(\mathbf{x})$, at temperature $T = \eta^{-1}$.

Gibbs-Boltzmann probability: $\mathbb{P}_\eta(\mathbf{x}) = \frac{1}{\mathcal{Z}_n(\eta)} e^{-\eta \mathcal{H}(\mathbf{x})} \prod_{i=1}^{n} P_X(x_i)$



GLM: $\mathbb{P}(\mathbf{x}|\mathbf{Y},\mathbf{\Phi}) = \frac{1}{\mathcal{Z}_n(\mathbf{Y},\mathbf{\Phi})} \prod_{i=1}^{n} P_X(x_i) \prod_{\mu=1}^{m} P_{\text{out}}\left[Y_\mu \middle| \frac{1}{\sqrt{n}}(\mathbf{\Phi}\mathbf{x})_\mu\right].$

Statistical physics "disordered" model, with Hamiltonian $\mathcal{H}(\mathbf{x}) = -\sum_{\mu=1}^{m} \ln P_{\text{out}}\left[Y_\mu \middle| \frac{1}{\sqrt{n}}(\mathbf{\Phi}\mathbf{x})_\mu\right]$ $(T=1)$

Spin glasses

General connection for many statistical models

[Hopfield '82; Mézard&Parisi '85; Gardner&Derrida '89; Anderson '89; Mézard&Montanari '09; ...]

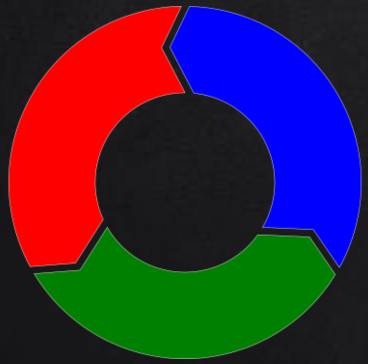
Where do we go from here?

Deep and detailed connection

- Bayesian estimation problems ⟷ Finite–temperature statistical physics
- Empirical risk minimization ⟷ (Zero–temperature) energy landscape minimization
- Posterior distribution ⟷ Gibbs–Boltzmann distribution
- High–dimensional limit ⟷ Thermodynamic limit
- Randomness of the observations (noise, …) ⟷ Disordered systems, "spin glasses"

<u>Theory of machine learning / inference</u>

Structure of the data



Architecture of the model

Algorithms

[Zdeborová '20]

Our "statistical physics–inspired" approach allows to study each of these pieces!

Main PhD projects

❖ Revisiting high-temperature expansions

- High-temperature expansions and message passing algorithms. *J.Stat.Mech. 2019.*
- Towards exact solution of extensive-rank matrix factorization. *In preparation.*

Approximation schemes and algorithms

II

✓ Exacts in high dimension

High-temperature expansions + Diagrammatics and random matrix theory → I + II

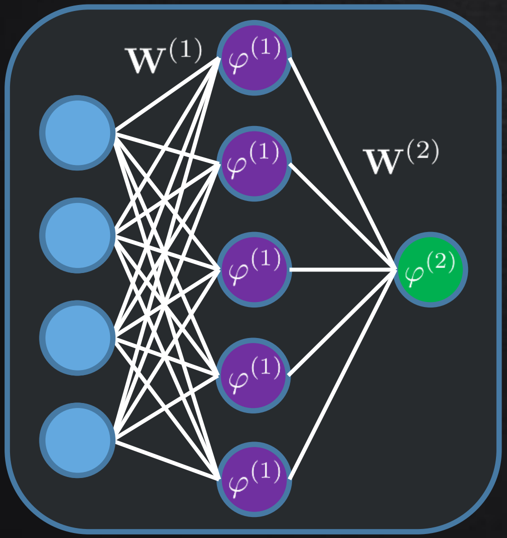
?

Extensive-rank matrix factorization $\mathbf{Y} = \mathbf{U}\mathbf{V}^\intercal + \mathbf{Z}$

Main PhD projects

❖ Optimal estimation in high-dimensional problems

- The mutual information in random linear estimation beyond iid matrices. *ISIT 2018.*
- Computational-to-statistical gaps in learning a two-layers neural network. *NeurIPS 2018 & J.Stat.Mech. 2019.*
- The spiked matrix model with generative priors. *IEEE Trans. Inf. Theory 2020 & NeurIPS 2019*
- Phase retrieval in high dimensions: statistical and computational phase transitions. *NeurIPS 2020.*
- Construction of optimal spectral methods in phase retrieval. *MSML 2021.*



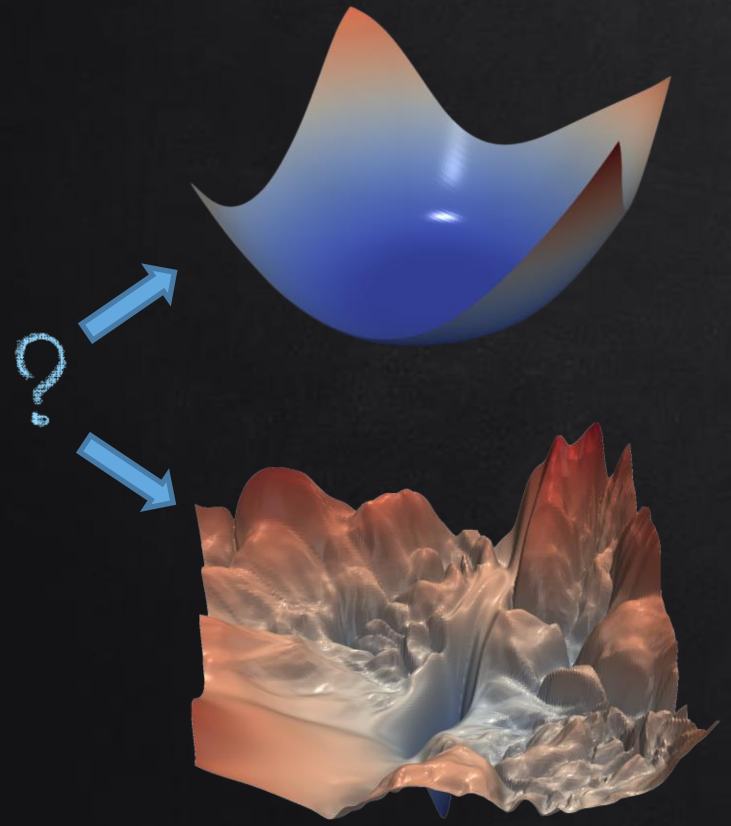
Committee machine

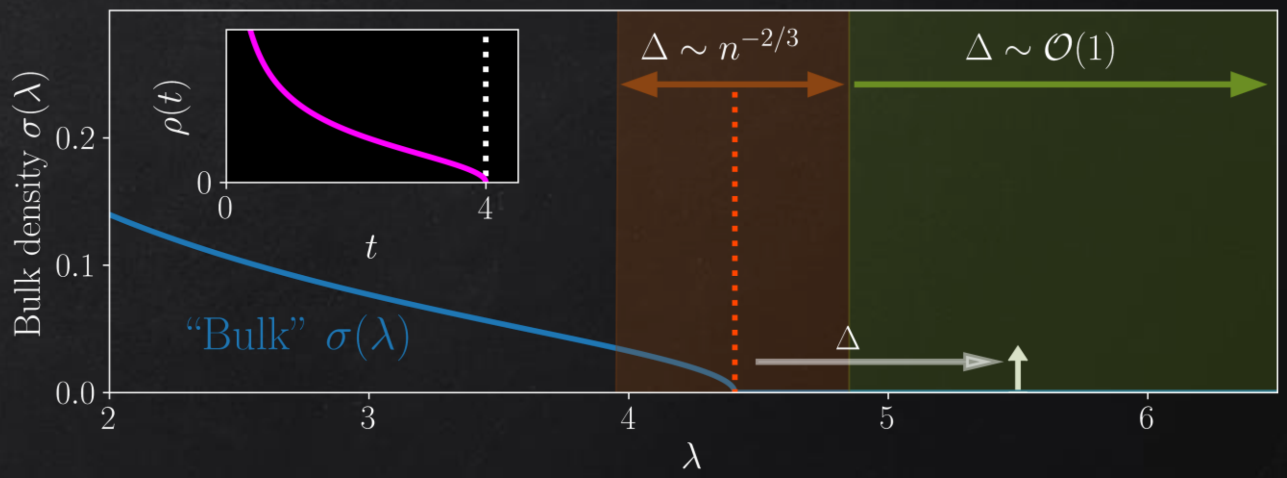
$$Y_\mu = \frac{1}{n}|(\mathbf{\Phi X}^\star)_\mu|^2$$

Main PhD projects

❖ Towards a topological approach to high-dimensional optimization

- Landscape complexity for the empirical risk of generalized linear models. *MSML 2020.*
- Large deviations of extreme eigenvalues of generalized sample covariance matrices. *EPL 2021.*





$$\mathbf{M} = \frac{1}{m}\sum_{\mu=1}^{m} \rho_\mu \mathbf{z}_\mu \mathbf{z}_\mu^\dagger$$

Main PhD projects

- Revisiting high-temperature expansions
  - High-temperature expansions and message passing algorithms. *J.Stat.Mech. 2019.*
  - Towards exact solution of extensive-rank matrix factorization. *In preparation.*

- Optimal estimation in high-dimensional problems
  - The mutual information in random linear estimation beyond iid matrices. *ISIT 2018.*
  - Computational-to-statistical gaps in learning a two-layers neural network. *NeurIPS 2018 & J.Stat.Mech. 2019.*
  - I { The spiked matrix model with generative priors. *IEEE Trans. Inf. Theory 2020 & NeurIPS 2019*
  - Phase retrieval in high dimensions: statistical and computational phase transitions. *NeurIPS 2020.*
  - Construction of optimal spectral methods in phase retrieval. *MSML 2021.*

- Towards a topological approach to high-dimensional optimization
  - II { Landscape complexity for the empirical risk of generalized linear models. *MSML 2020.*
  - Large deviations of extreme eigenvalues of generalized sample covariance matrices. *EPL 2021.*

I Exploiting data structure in spiked matrix estimation

Spiked Wigner model [Johnstone '01]

$$\boxed{\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{v}^\star(\mathbf{v}^\star)^\intercal + \sqrt{\Delta}\boldsymbol{\xi} \in \mathbb{R}^{p\times p}} \quad \begin{cases} \xi_{ij} & = \xi_{ji} \\ \xi_{ij} & \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1+\delta_{ij}) \end{cases}$$

$$\mathbf{v}^\star \sim P_v$$

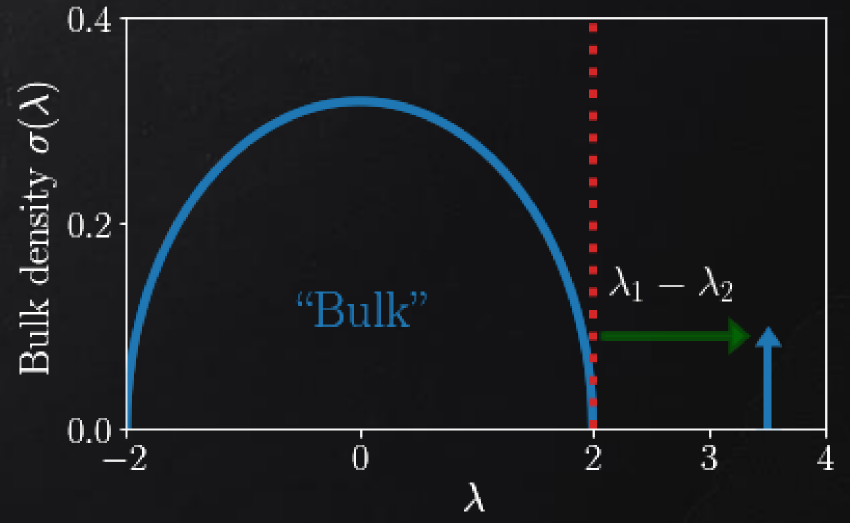
$$\rho_v \equiv \lim_{p\to\infty} \frac{1}{p}\mathbb{E}_{P_v}\|\mathbf{v}\|^2$$

Spiked Wishart model

$$\boxed{\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{u}^\star(\mathbf{v}^\star)^\intercal + \sqrt{\Delta}\boldsymbol{\xi} \in \mathbb{R}^{n\times p}}$$

$$\mathbf{u}^\star \sim P_u \qquad \mathbf{v}^\star \sim P_v \qquad \xi_{\mu i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

"BBP" transition



$$\Delta/\rho_v^2 \lessgtr 1$$

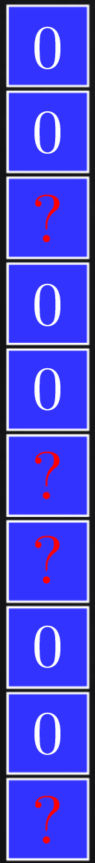
[Edwards&Jones '76; Baik, Ben Arous&Péché '04]

- PCA: the dominant eigenvector of $\mathbf{Y}$.

  Optimal for unstructured signal $P_v = \mathcal{N}(0,1)$.

- Leverage prior knowledge on the structure of the signal to improve recovery ? ➡ "Dimensionality reduction"


$$\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{v}^\star(\mathbf{v}^\star)^\intercal + \sqrt{\Delta}\xi \in \mathbb{R}^{p\times p}$$



Sparsity

- Natural representation, e.g.:

Images	$\Rightarrow$	Wavelet
Sound	$\Rightarrow$	Fourier

- Efficient algorithms: LASSO, compressed sensing,...

But...

- Large algorithmically hard phases
- Impossible to "beat" the BBP transition of PCA.

[Deshpande&al '14, Lesieur&al '15]

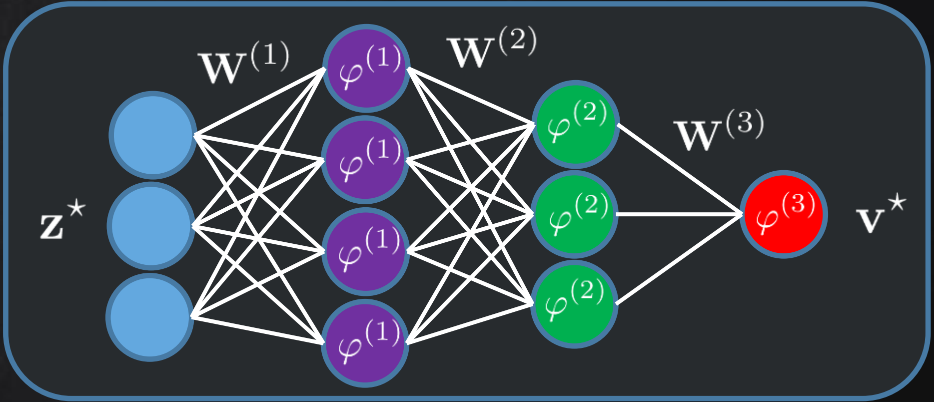
Generative prior

Random weights

Unstructured latent variable

$$\boxed{\mathbf{v}^\star = \varphi^{(L)}\left(\frac{1}{\sqrt{k_L}}\mathbf{W}^{(L)}\dots\varphi^{(1)}\left(\frac{1}{\sqrt{k}}\mathbf{W}^{(1)}\mathbf{z}^\star\right)\right)}$$

Structured signal



In the limit $p \to \infty$, for a given $\Delta > 0$, what is the optimal recovery:

- Information-theoretically (in exponential time) ?
- With which tractable (polynomial-time) algorithm ?
- Cheap (e.g. spectral) algorithms that outperform PCA ?

The replica-symmetric formula

$$\mathbb{P}(\mathbf{v}|\mathbf{Y}) = \frac{1}{\mathcal{Z}_p(\mathbf{Y})} P_v(\mathbf{v}) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - \frac{v_i v_j}{\sqrt{p}}\right)^2}$$

Theorem (informal)

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

$$\lim_{p\to\infty} -\frac{1}{p}\mathbb{E}_{\mathbf{Y}} \ln \mathcal{Z}_p(\mathbf{Y}) = \inf_{q\in(0,\rho_v)} f_{\mathrm{RS}}(\Delta, q)$$, with $$f_{\mathrm{RS}}(\Delta, q) = \frac{q(\rho_v - q)}{2\Delta} + \lim_{p\to\infty} \frac{1}{p}\mathrm{I}(\mathbf{v}; \mathbf{v} + \sqrt{\Delta/q_v}\boldsymbol{\xi})$$

Mutual information

Moreover: $$\mathrm{MMSE}_v(\Delta) = \rho_v - \arg\min_{q\in[0,\rho_v]} f_{\mathrm{RS}}(\Delta, q)$$

Information-theoretic MMSE.

How to:

- Derive the result using the non-rigorous replica method [Mézard, Parisi & Virasoro '87]...
- Prove the result (not the method!) using interpolation techniques [Guerra '03 ; Talagrand '06 ; Barbier&al '19]...

Similar results & strategy for: two-layers neural networks [Aubin, A.M&al '18], compressed sensing with non-i.i.d. matrices [Barbier, A.M&al'18], phase retrieval with rotationally-invariant matrices [A.M&al '20], ...

Application: single-layer generative prior

$$\boxed{\mathbf{v}^\star \sim \varphi\left(\frac{1}{\sqrt{k}}\mathbf{W}\mathbf{z}^\star\right)}$$

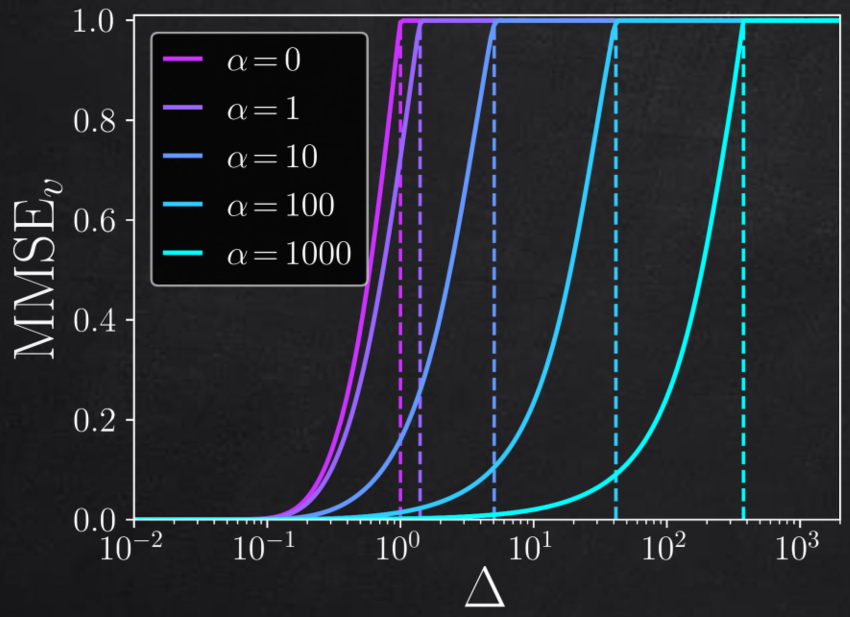
$$\begin{cases} W_{il} & \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ \mathbf{z}^\star & \sim P_z \\ \alpha & \equiv p/k \end{cases}$$

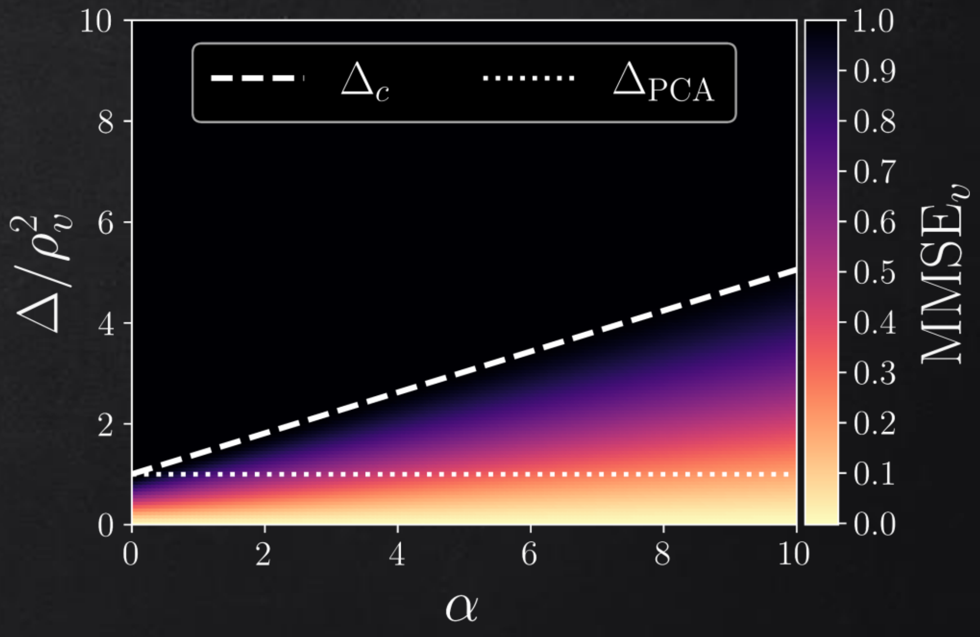
Unstructured (i.i.d.) prior

$$\boxed{f_{\mathrm{RS}}(\Delta, q) = \frac{q(\rho_v - q)}{2\Delta} + \frac{1}{\alpha}\min_{q_z,\hat{q}_z}\left[\frac{q_z\hat{q}_z}{2} - \Psi_z(\hat{q}_z) - \alpha\Psi_{\text{out}}\left(\frac{q}{\Delta}, q_z\right)\right]}$$

Tedious, but <u>completely scalar</u> potential!

$\varphi(x) = \text{sign}(x)$





$\boxed{\text{Weak-recovery: } \Delta_c \equiv \inf\{\Delta \,:\, \text{MMSE}_v(\Delta) = 1\}}$

- Sparse PCA: $\Delta_c = 1$
- 1-layer generative prior: $\Delta_c = 1 + \frac{4}{\pi^2}\alpha$

Algorithmic limits

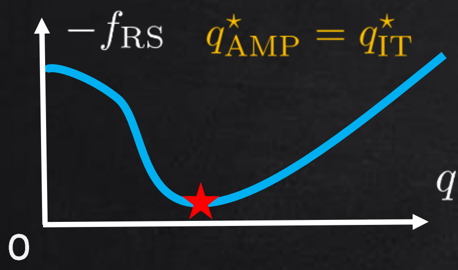
Can we algorithmically (i.e. in polynomial time) achieve the optimal MSE ?

Measure of algorithm reconstruction by the overlaps

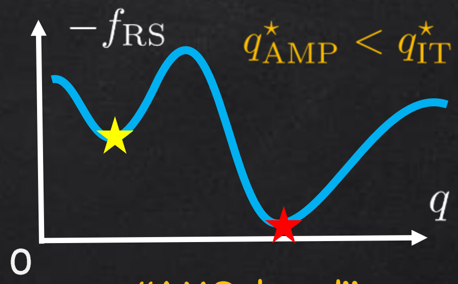
$$q \equiv \lim_{p\to\infty} \frac{1}{p}\mathbb{E}[\mathbf{v}^\intercal \mathbf{v}^\star] \qquad q_z \equiv \lim_{k\to\infty} \frac{1}{k}\mathbb{E}[\mathbf{z}^\intercal \mathbf{z}^\star]$$

$$q^{t+1} = 2\partial_q \Psi_{\text{out}}\left(\frac{q}{\Delta}, q_z\right) \; ; \; q_z^{t+1} = 2\partial_{\hat{q}_z}\Psi_z(\hat{q}_z^t) \; ; \; \hat{q}_z^{t+1} = 2\alpha\partial_{q_z}\Psi_{\text{out}}\left(\frac{q^t}{\Delta}, q_z^t\right)$$

State Evolution (SE) equations: "Fixed point algorithm" on f_{RS} !



"AMP-easy"



"AMP-hard"

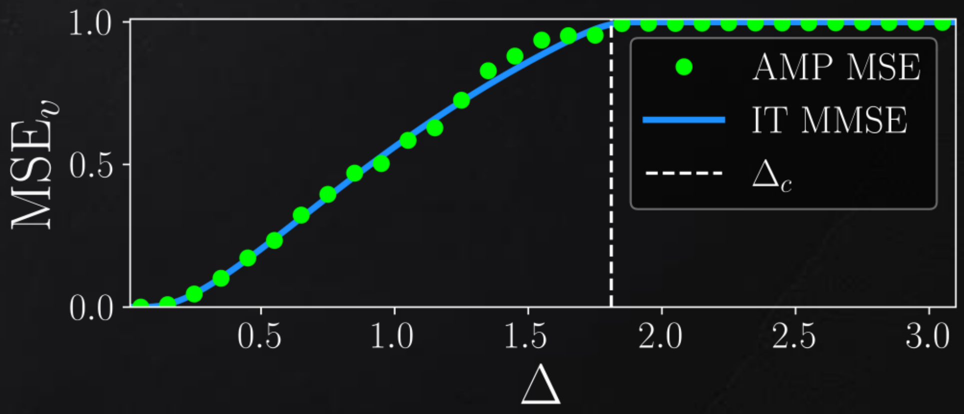
Tested settings: single/multi layer and $\varphi \in \{\text{linear, sign, ReLU}\}$.

No algorithmically hard phase: very different from sparse PCA !

Approximate Message Passing (AMP)

Input: $Y \in \mathbb{R}^{p\times p}$ and $W \in \mathbb{R}^{p\times k}$:
Initialize with: $\hat{\mathbf{v}}^{t=1} = \mathcal{N}(\mathbf{0}, \sigma^2 I_p)$, $\hat{\mathbf{z}}^{t=1} = \mathcal{N}(\mathbf{0}, \sigma^2 I_k)$, and $\hat{\mathbf{c}}_v^{t=1} = I_p$, $\hat{\mathbf{c}}_z^{t=1} = I_k$, $t = 1$.
repeat
  Spiked layer denoising:
  $\mathbf{B}_v^t = \frac{1}{\Delta}\frac{Y}{\sqrt{p}}\hat{\mathbf{v}}^t - \frac{1}{\Delta}\frac{\left(I_p^\intercal \hat{\mathbf{c}}_v^t\right)}{p}\hat{\mathbf{v}}^{t-1}$ and $A_v^t = \frac{1}{\Delta p}(\|\hat{\mathbf{v}}^t\|_2)^2 I_p$.
  Generative layer denoising:
  $V^t = \frac{1}{k}\left(I_k^\intercal \hat{\mathbf{c}}_z^t\right) I_p$, $\boldsymbol{\omega}^t = \frac{1}{\sqrt{k}} W\hat{\mathbf{z}}^t - V^t\mathbf{g}^{t-1}$
  $\mathbf{g}^t = f_{\text{out}}\left(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t\right)$
  $\Lambda^t = \frac{1}{k}\|\mathbf{g}^t\|_2^2 t I_k$ and $\boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}} W^\intercal \mathbf{g}^t + \Lambda^t \hat{\mathbf{z}}^t$.
  Marginals estimation:
  $\hat{\mathbf{v}}^{t+1} = f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t)$ and $\hat{\mathbf{c}}_v^{t+1} = \partial_B f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t)$,
  $\hat{\mathbf{z}}^{t+1} = f_z(\boldsymbol{\gamma}^t, \Lambda^t)$ and $\hat{\mathbf{c}}_z^{t+1} = \partial_\gamma f_z(\boldsymbol{\gamma}^t, \Lambda^t)$,
  $t = t + 1$.
until Convergence.
Output: $\hat{\mathbf{v}}, \hat{\mathbf{z}}$.

Iteration of the TAP equations of stat. phys.
Optimal among general first-order methods
[Celentano&al '20]

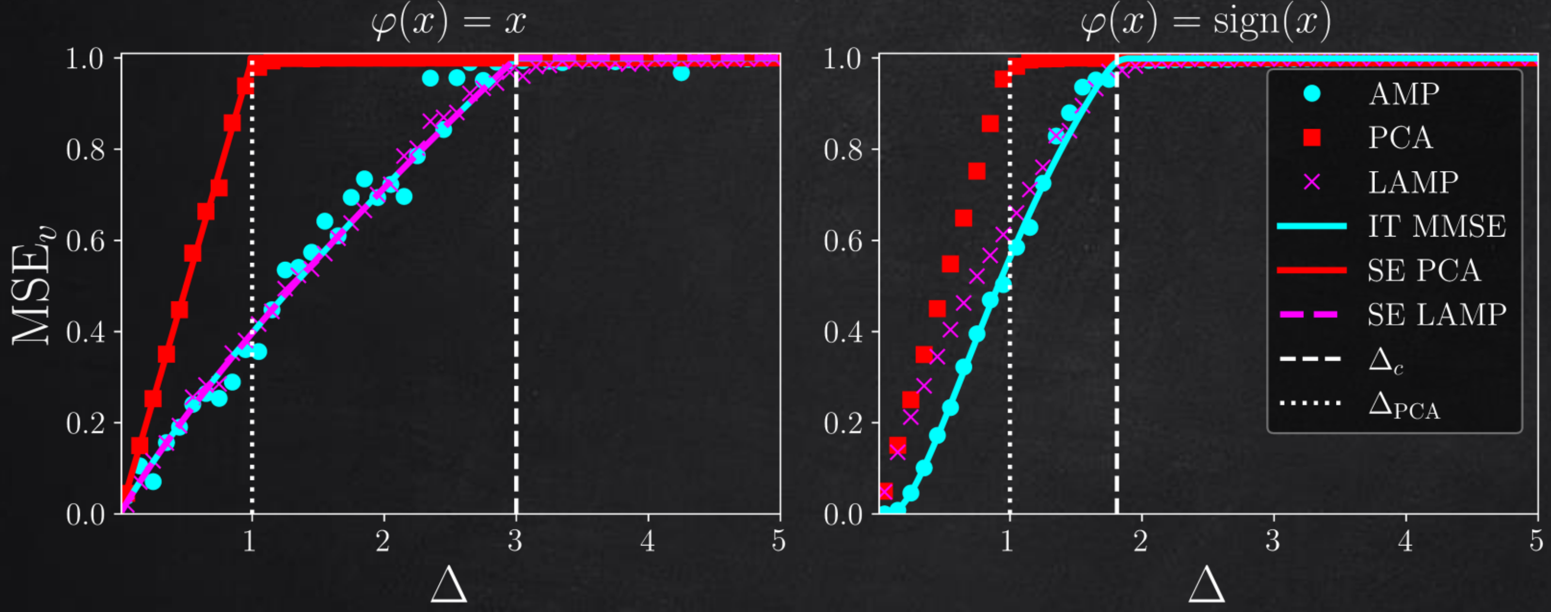


Spectral algorithms

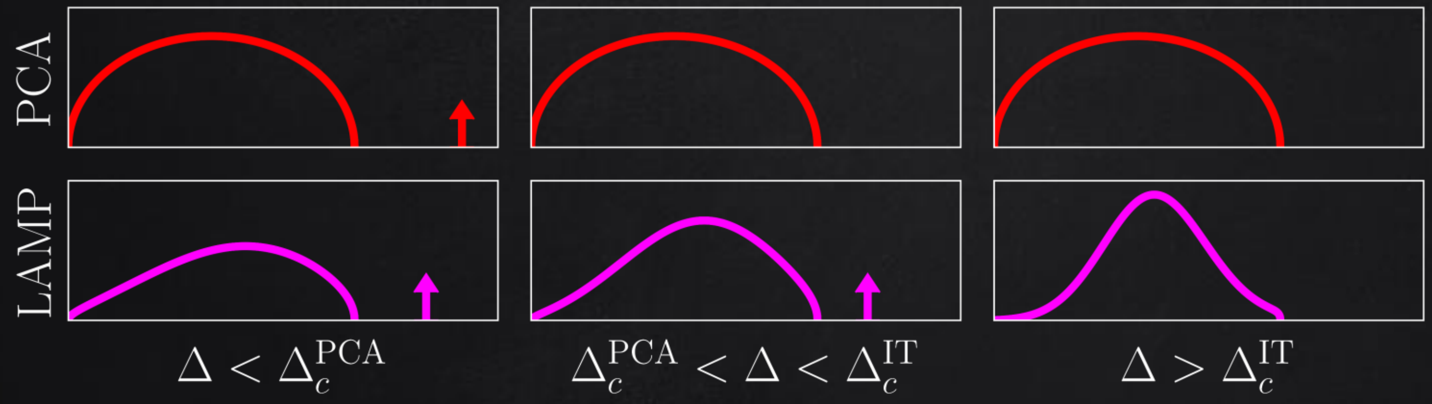
Linearized Approximate Message-Passing (LAMP)

Symmetries $\Rightarrow$ "Trivial" fixed point $\Rightarrow$

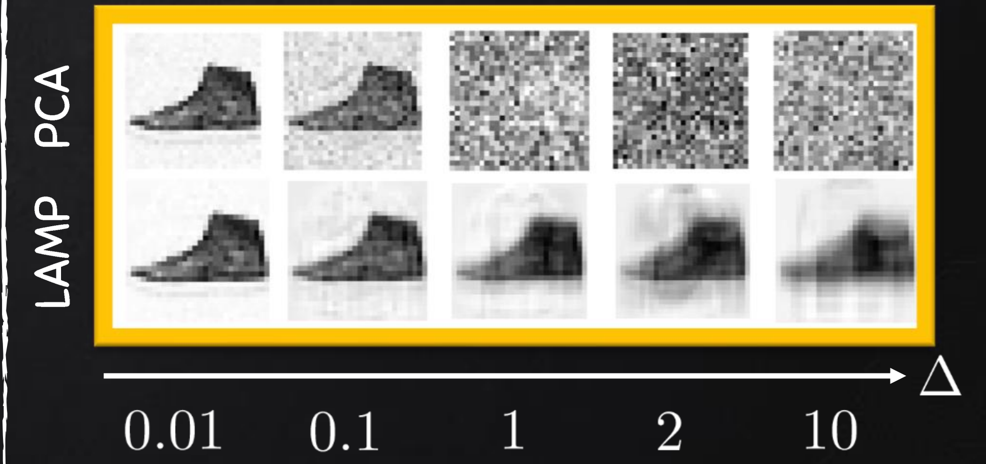
Leading eigenvector of $\boldsymbol{\Gamma}_p \equiv \frac{1}{\Delta}\mathbf{K}_p\left[\frac{\mathbf{Y}}{\sqrt{p}} - \mathrm{I}_p\right]$, with $\mathbf{K}_p \equiv \frac{1}{k}\mathbb{E}[\mathbf{v}\mathbf{v}^\intercal]$.



LAMP "beats" the BBP transition of PCA!



Realistic data $\quad \mathbf{K}_p \simeq \frac{1}{n}\sum_{\alpha=1}^{n}\mathbf{v}^\alpha(\mathbf{v}^\alpha)^\intercal$

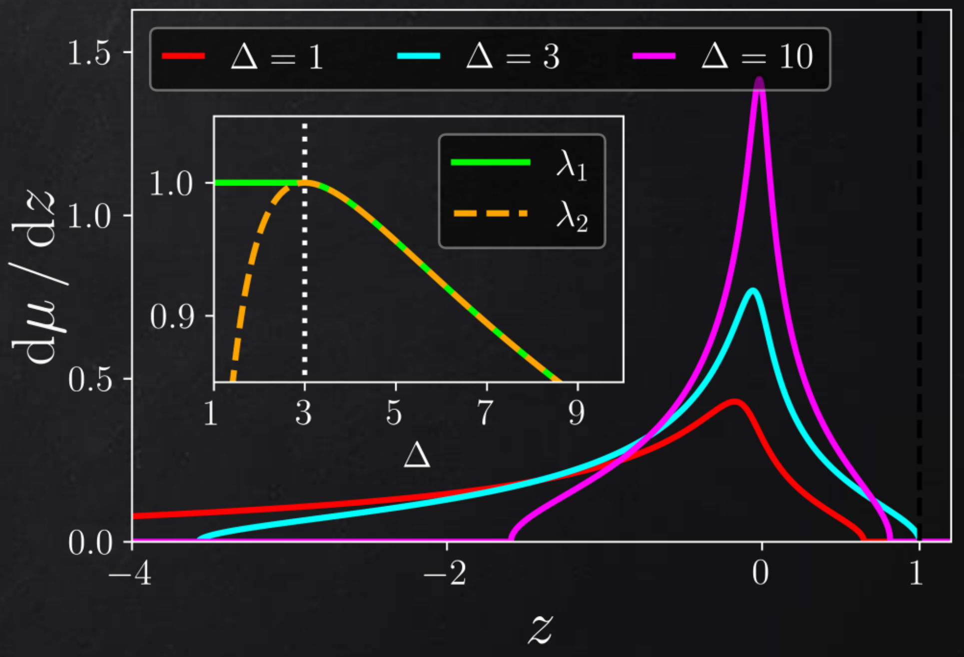


Random matrix analysis

Linear case $\varphi(x) = x$ $\Rightarrow$ $\boxed{\Gamma_p = \frac{1}{\Delta}\frac{\mathbf{W}\mathbf{W}^\intercal}{k}\left[\frac{\mathbf{Y}}{\sqrt{p}} - \mathrm{I}_p\right]}$

μ : asymptotic spectral density of Γ_p, with $\lambda_{\max}$ the right edge of its support.

$\alpha = 2 \Rightarrow \Delta_c = 1 + \alpha = 3$



Theorem: "BBP"-like transition

Let $\Delta_c(\alpha) \equiv 1 + \alpha$. We denote $\lambda_1 \geq \lambda_2$ the leading eigenvalues of Γ_p, with normalized eigenvectors $\mathbf{v}_1, \mathbf{v}_2$. Then:

- For $\Delta > \Delta_c(\alpha)$, $\lambda_1 \xrightarrow[p\to\infty]{\text{a.s.}} \lambda_{\max}$ and $\lambda_2 \xrightarrow[p\to\infty]{\text{a.s.}} \lambda_{\max}$, and $\lambda_{\max} < 1$.
- For $\Delta < \Delta_c(\alpha)$, $\lambda_1 \xrightarrow[p\to\infty]{\text{a.s.}} 1$ and $\lambda_2 \xrightarrow[p\to\infty]{\text{a.s.}} \lambda_{\max}$, and $\lambda_{\max} < 1$.

Moreover, if $\epsilon(\Delta) \equiv \lim_{p\to\infty} \frac{1}{p}|\mathbf{v}_1^\intercal \mathbf{v}^\star|$, then $\begin{cases} \epsilon(\Delta) & = 0 \text{ if } \Delta > \Delta_c(\alpha), \\ \epsilon(\Delta) & > 0 \text{ if } \Delta < \Delta_c(\alpha). \end{cases}$

- Main difficulty: correlation of $\mathbf{W}$ and $\mathbf{Y} = \frac{\mathbf{W}(\mathbf{z}\mathbf{z}^\intercal)\mathbf{W}^\intercal}{\sqrt{kp}} + \sqrt{\Delta}\xi$. We use a cavity computation, generalizing the classical arguments of [Baik, Ben Arous&Péché '04].
- Similar results in the spiked Wishart model.
- A RMT analysis of non-linear activations is still lacking !

Summary on the spiked matrix model

Sparse priors

- Large hard phases for sparse signals $\rho \ll 1$. [Deshpande&al '14, Lesieur&al '15]
- IT weak recovery: $\Delta_c^{\rm IT} > 1$. But no algorithm can beat the PCA threshold $\Delta_c^{\rm PCA} = 1$.

Generative priors

- No algorithmically hard phase, AMP achieves the IT MMSE.
- Spectral L-AMP outperforms PCA and achieves optimal weak-recovery at $\Delta_c^{\rm LAMP} = \Delta_c^{\rm AMP} > 1$.
- Rigorous RMT analysis of LAMP's performance in the linear case.

Generative priors lead to algorithmically better-behaved problems than sparsity!

Active line of research on the influence of the data structure

- Similar analysis followed in the group, e.g. [Aubin&al '20] for phase retrieval.
- [Goldt&al '19 ; Goldt&al '20] : "hidden manifold" model: theoretical and empirical evidence that many conclusions transfer to trained (non-random) generative priors.
- ....

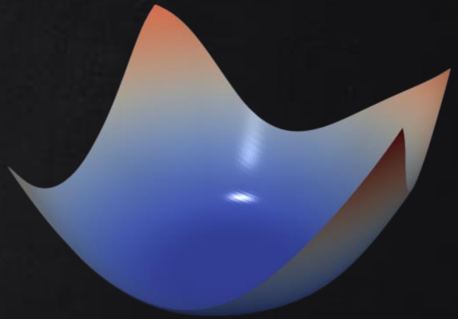
II Topology of high-dimensional landscapes

Squared loss of a noiseless GLM

$$L_2(\mathbf{x}) = \frac{1}{2m}\sum_{\mu=1}^{m}\left[\varphi(\mathbf{\Phi}_\mu \cdot \mathbf{x}) - \varphi(\mathbf{\Phi}_\mu \cdot \mathbf{X}^\star)\right]^2$$

$$\mathbf{X}^\star \in \mathbb{R}^n,\ \|\mathbf{X}^\star\|^2 = 1$$
$$\mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\|^2 = 1$$

i.i.d. Gaussian data



In many nonconvex problems: one can provably find regimes in which the optimization landscape is 'easy' (matrix decomposition, tensor factorization, neural nets…) [Soudry&al '16; Ge&al '16; Ge&Ma '17; …]

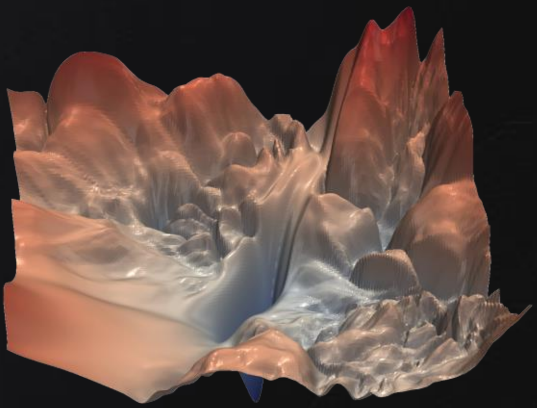
In practice, local optimization algorithms work far beyond these regimes !

WHY ?

→ The bounds on the simplicity of the landscape are not tight enough ?

→ Optimization algorithms work in the "hard" regime (i.e. many spurious minima) ?

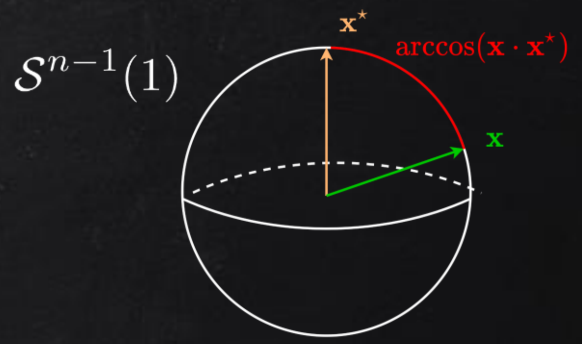
→ Analyze the topological transition, and characterize the 'hard' regime ?



[Li&al '17]

"Complex" landscapes

$$L_2(\mathbf{x}) = \frac{1}{2m}\sum_{\mu=1}^{m}\Big[\varphi(\mathbf{\Phi}_\mu\cdot\mathbf{x}) - \varphi(\mathbf{\Phi}_\mu\cdot\mathbf{X}^\star)\Big]^2$$



- High-dimensional limit $n, m \to \infty$ with $\alpha = m/n = \Theta(1)$.
- Count critical points with fixed "energy" $L_2(\mathbf{x})$ and overlap $q = \mathbf{X}^\star\cdot\mathbf{x}$?

$$\mathrm{Crit}_\star(B,Q) \equiv \sum_{\mathbf{x}:\,\mathrm{grad}\,L_2(\mathbf{x})=0} \mathbb{1}\{L_2(\mathbf{x})\in B,\ \mathbf{x}\cdot\mathbf{X}^\star\in Q\}$$

Random variable (randomness of the data)

- Typically of size $e^{\Theta(n)}$.
- Strongly fluctuating !

- Mean value : annealed complexity $\Sigma_\star^{(\mathrm{an.})}(B,Q) \equiv \lim_{n\to\infty}\frac{1}{n}\ln\mathbb{E}\,\mathrm{Crit}_\star(B,Q)$
- Typical value : quenched complexity $\Sigma_\star^{(\mathrm{qu.})}(B,Q) \equiv \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\ln\,\mathrm{Crit}_\star(B,Q)$

⚠ Quenched $\neq$ Annealed !
(Few exceptions [Subag '17])

Complexity at a fixed index $\mathrm{Crit}_k(B,Q) \equiv \sum_{\mathbf{x}:\,\mathrm{grad}\,L_2(\mathbf{x})=0} \mathbb{1}\{\mathrm{i}[\mathrm{Hess}\,L_2(\mathbf{x})]=k,\ L_2(\mathbf{x})\in B,\ \mathbf{x}\cdot\mathbf{X}^\star\in Q\}$

Our main tool: the Kac-Rice formula

Theorem (Kac-Rice)

e.g. [Adler&Taylor '09]

$f : \mathcal{S}^{n-1}(1) \to \mathbb{R}$ is a smooth random function that is a.s. Morse. Then:

$$\mathbb{E}\,\mathrm{Crit}_k(f) = \int_{\mathcal{S}^{n-1}(1)} \sigma(\mathrm{d}\mathbf{x})\, \varphi_{\mathrm{grad}\, f(\mathbf{x})}(0) \times \underline{\mathbb{E}\left[|\det \mathrm{Hess}\, f(\mathbf{x})| \middle| \mathrm{grad}\, f(\mathbf{x}) = 0; \mathrm{i}(\mathrm{Hess}\, f(\mathbf{x})) = k\right]}$$

Density of the (random) gradient taken in 0

- Random differential geometry ⟹ Random matrix theory.
- Many possible refinements: fix the value of $f(\mathbf{x})$, higher-order moments, ...

Takeaways:
- Distribution of $\{\mathrm{Hess}\, f(\mathbf{x}) | \mathrm{grad}\, f(\mathbf{x}) = 0\}$: intractable for "generic" functions !
- To compute $\mathbb{E}\mathrm{Crit}_k(f)$: we need the large deviations of the k-th largest eigenvalue of the Hessian.

⟶ Applications limited to Gaussian random functions

Pure p-spin and variants

$$f(\mathbf{x}) = \sum_{i_1,\cdots,i_p} \underbrace{J_{i_1,\cdots,i_p}}_{\mathcal{N}(0,1)} x_{i_1}\cdots x_{i_p}$$

[Bray&Moore '80, Crisanti&al '95, Fyodorov&al '07, Auffinger&al '13, Ros&al '19, ...] .

Main results

$$L_2(\mathbf{x}) = \frac{1}{2m}\sum_{\mu=1}^{m}\left[\varphi(\mathbf{\Phi}_\mu \cdot \mathbf{x}) - \varphi(\mathbf{\Phi}_\mu \cdot \mathbf{X}^\star)\right]^2 \xrightarrow{\text{simplification}} \boxed{L_1(\mathbf{x}) = \frac{1}{m}\sum_{\mu=1}^{m}\varphi(\mathbf{\Phi}_\mu \cdot \mathbf{x})}$$

Theorem & proof transfer to $L_2(\mathbf{x})$.

First exact high-dimensional result obtained with Kac-Rice for non-Gaussian functions !

Theorem

$$\Sigma_\star^{(\mathrm{an})}(B) = \frac{1+\ln\alpha}{2} + \sup_{\substack{\nu\in\mathcal{M}_1^+(\mathbb{R}) \\ \int \nu(\mathrm{d}t)\varphi(t)\in B}} \left[-\frac{1}{2}\ln\left\{\int \nu(\mathrm{d}t)\varphi'(t)^2\right\} - \underbrace{\alpha H\left(\nu|\mathcal{N}(0,1)\right)}_{\text{Relative entropy}} + \kappa_\alpha(\nu)\right]$$

Involved function : related to the logarithmic potential of the asymptotic spectral measure of $\mathbf{z}\mathbf{D}\mathbf{z}^\intercal/m$ if $\mathbf{z}\in\mathbb{R}^{n\times m}$ is a Gaussian i.i.d. matrix and $D_\mu = \varphi''(y_\mu)$ with $y_\mu \overset{\text{i.i.d.}}{\sim} \nu$.

- Term $\kappa_\alpha(\nu)$ $\Longrightarrow$ Analytically very hard variational problem!
- We derive a closed formula for the quenched complexity, using the heuristic replica method [Parisi&al '87, Ros&al '19].
- Generic result, applies to mixture of Gaussians, binary classification, ...

Sketch of proof

$$L_1(\mathbf{x}) = \frac{1}{m}\sum_{\mu=1}^m \varphi(\mathbf{\Phi}_\mu \cdot \mathbf{x})$$

Kac-Rice formula $\Longrightarrow$ Hessian conditioned by zero gradient.

Main idea : Condition everything by the i.i.d. Gaussian random variables $y_\mu \equiv \mathbf{\Phi}_\mu \cdot \mathbf{x}$

$\mathbb{E}_{\mathbf{\Phi}}[\cdots] = \mathbb{E}_{\mathbf{y}}\mathbb{E}[\cdots|\mathbf{y}]$ Under this conditional distribution, and under the gradient being zero:

$$\boxed{\mathrm{Hess}\, L_1(\mathbf{x}) \overset{\mathrm{d}}{=} \frac{1}{m}\sum_{\mu=1}^m \varphi''(y_\mu)\mathbf{z}_\mu\mathbf{z}_\mu^\intercal + t(\mathbf{y})\mathbb{1}_n (+\text{finite rank term})}$$

$\mathbf{z}_\mu$: i.i.d. standard Gaussian vectors

"Generalized" version of a sample covariance matrix

- We prove fast enough concentration of $\ln|\det \mathrm{Hess}|$ as a function of $\{y_\mu\}_{\mu=1}^m$: $\mathbb{E}|\det \mathrm{Hess}| \simeq e^{\mathbb{E}\ln|\det \mathrm{Hess}|}$
- The expectation only depends on the empirical distribution $\nu_{\mathbf{y}} \equiv (1/m)\sum_{\mu=1}^m \delta_{y_\mu}$: $\mathbb{E}\,\ln|\det \mathrm{Hess}| \simeq m\kappa_\alpha(\nu_{\mathbf{y}})$
- We use Sanov's theorem: the law of $\nu_{\mathbf{y}}$ satisfies large deviations with rate function : $I(\nu) = \alpha H(\nu|\mathcal{N}(0,1))$
- Kac-Rice formula and Varadhan's lemma $\Longrightarrow$ $\boxed{\Sigma_\star^{(\mathrm{an})} = \sup_{\nu\in\mathcal{M}_1^+(\mathbb{R})} [\kappa_\alpha(\nu) + \underline{G(\nu)} - \alpha H(\nu|\mathcal{N}(0,1))]}$ $\blacksquare$

(Gradient density term in Kac-Rice)

Summary & outlook

❖ First exact high-dimensional result obtained with Kac-Rice for non-Gaussian functions !

❖ Generalizes to other models: mixture of two Gaussians, binary linear classification… → Neural networks ?

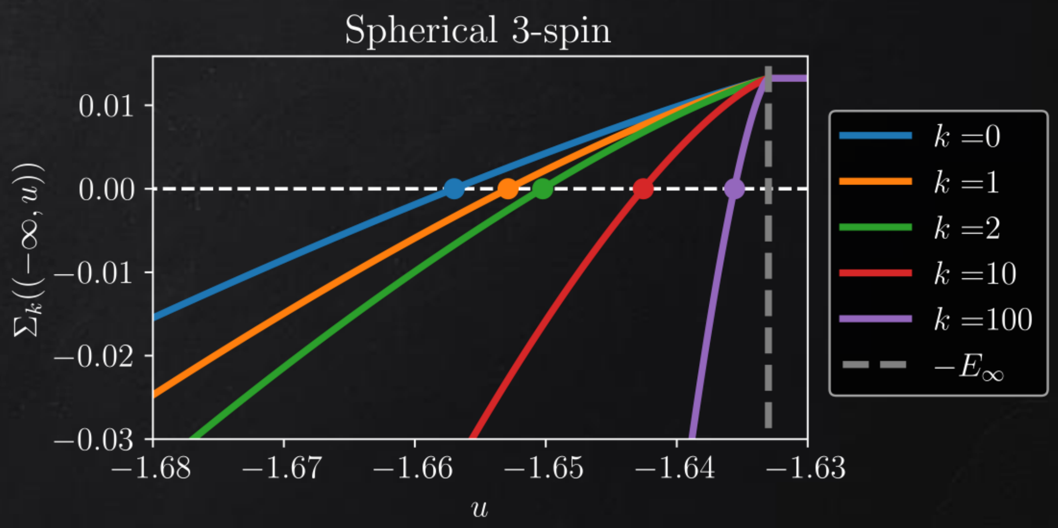
Physical discussion is lacking. Many problems ahead:

➢ Numerically solve the variational problem? Sign of the complexity given α, φ ? …

➢ Count local minima ? We need a LDP for $\lambda_{\min}(\mathrm{Hess}\, L(\mathbf{x}))$ …

→ Obtained in [A.M. , EPL 2021] ! To be continued…

"Tilting" of the measure [Biroli&Guionnet'20, Belinschi&al '20, Guionnet&al '20, Husson '20, Augeri&al '21]…



➢ Discrete systems ?

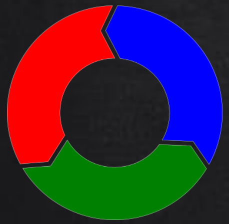
❖ TAP approach ?

❖ Recent algorithmic progress on F-RSB spin glasses [Subag '21, Montanari '21, El Alaoui & al '20].

Concluding remarks

Theory of inference/learning

Data structure



Architecture of the model

Algorithms

Diversified toolbox

Statistical physics
Replica, message-passing, Plefka expansions, DMFT…

Probabilistic methods
Guerra interpolation, concentration identities…

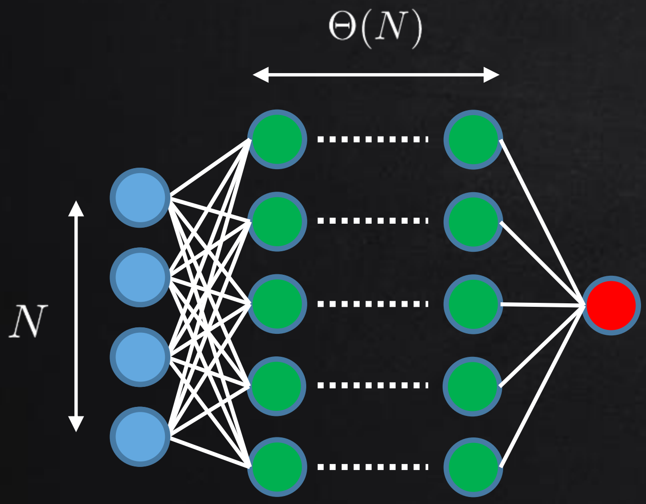
Topological approach
Kac-Rice, large deviations, random matrix theory…,

To name a few…

Many exciting questions, e.g.:

- Interplay in more involved learning models ?
- What if we do not know how the data was generated ?

Realistic <u>deep</u> networks ?



Or

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^\intercal + \mathbf{Z}$$

- ~~Classical statistical physics approach.~~
- Extensive-rank HCIZ ? [Matytsin '93];[Guionnet&Zeitouni '02]
- Plefka expansion ? [Parisi&Potters '95];[A.M.&al '19], …

The "extensive-rank problem"

One challenge among many….

Conclusion