Landscape complexity for the empirical risk of generalized linear models

Antoine Maillard, Gérard Ben Arous & Giulio Biroli

Mathematical and Scientific Machine Learning 2020

ENS | PSL
FONDATION CFM POUR LA RECHERCHE
NYU
ETH zürich

Universität Basel – October 20th 2021

Generalized Linear Models

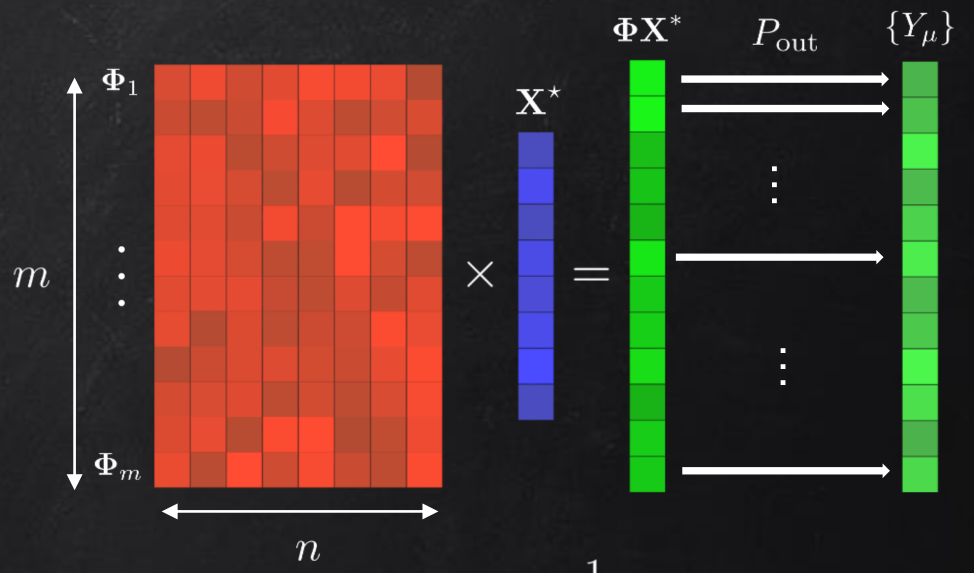
Goal : Recover $\mathbf{X}^\star \in \mathbb{R}^n$ from $\{\mathbf{\Phi}_\mu, Y_\mu\}_{\mu=1}^m$:

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

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

Channel: non-linearity
+ possible noise

Sensing matrix



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

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

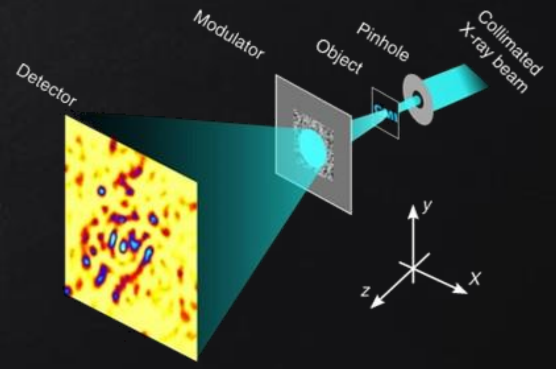


Image credits: [Zhang&al 16]

Goal : Fundamental limits of inference models with random input data in the typical case and in high dimension.

⚠ Different from "worst-case" injectivity studies. (e.g. [Bandeira&al '14])

"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).

Estimate X* ?

Bayesian estimators: Leverage the posterior distribution $\mathbb{P}(\mathbf{x}|\mathbf{Y},\mathbf{\Phi}) = \frac{\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{Y}|\mathbf{x},\mathbf{\Phi})}{\mathbb{P}(\mathbf{Y}|\mathbf{\Phi})}$

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

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

M-estimation:

$$\boxed{\hat{\mathbf{X}} \equiv \arg\min_{\mathbf{x}} \sum_{\mu=1}^{m} \rho(\mathbf{x}, \mathbf{\Phi}_\mu, y_\mu)}$$

$$y_\mu = \varphi\left[\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^{n} \Phi_{\mu i} X_i^\star\right]$$

(to simplify)

Example: *Empirical Risk Minimizer, e.g. square loss* $\boxed{\hat{\mathbf{X}} \equiv \arg\min_{\mathbf{x}} \left\{ \frac{1}{m} \sum_{\mu=1}^{m} \left[y_\mu - \varphi\left(\frac{1}{\sqrt{n}} \mathbf{\Phi} \cdot \mathbf{x} \right) \right]^2 \right\}}$ This presentation

Topology of high-dimensional landscapes

Empirical risk 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 input 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 ?

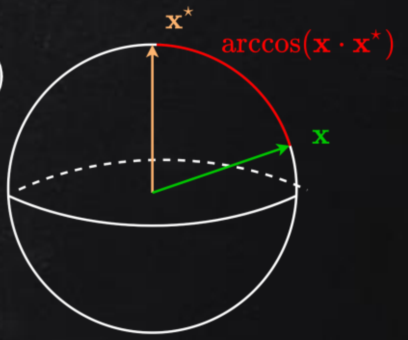
[Li&al '17]

"Complex" landscapes

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

Characterize complexity of a landscape

$\mathcal{S}^{n-1}(1)$



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

Counting critical points

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

Let $f : \mathcal{S}^{n-1}(1) \to \mathbb{R}$ a smooth function.

Count $N_f(\mathbf{u}) \equiv \#\{\mathbf{x} \in \mathcal{S}^{n-1}(1) \ : \ \mathrm{grad}\, f(\mathbf{x}) = \mathbf{u}\}$?

?

$$N_f(\mathbf{u}) = \int_{\mathrm{grad}\, f(\mathcal{S}^{n-1}(1))} \mathrm{d}\mathbf{v}\, \delta(\mathbf{v} - \mathbf{u}) \quad = \int_{\mathcal{S}^{n-1}(1)} \mathrm{d}\mathbf{x}\, \delta(\mathrm{grad}\, f(\mathbf{x}) - u)\, |\det \mathrm{Hess} f(\mathbf{x})|$$

Weak sense

Test function

Area formula $\boxed{\int \mathrm{d}\mathbf{u}\, g(u)\, N_f(\mathbf{u}) = \int_{\mathcal{S}^{n-1}(1)} \mathrm{d}\mathbf{x}\, g(\mathrm{grad}\, f(\mathbf{x}))\, |\det \mathrm{Hess}\, f(\mathbf{x})|}$ [Federer '59]

⚠ $L_2(\mathbf{x})$ is also a <u>random</u> function !

f is a.s. <u>Morse</u> (all critical points are non-degenerate)

+ technical regularity properties (hard for non-Gaussian functions)...

Strong sense at $\mathbf{u} = 0$

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

Density of the (random) gradient taken in 0

The Kac-Rice formula

e.g. [Adler&Taylor '09]

Theorem (Kac-Rice)

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

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

- Random differential geometry ⟹ Random matrix theory.
- Many possible refinements: fix the value of $f(\mathrm{x})$, 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, Belius&al '21, ...]

Annealed complexity

$\alpha = m/n = \Theta(1)$

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

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

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

Theorem

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

<u>Involved function</u> : 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!
- Generic result, applies to mixture of Gaussians, binary classification, …

Sketch of proof

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

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

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

$\mathbb{E}_{\boldsymbol{\Phi}}[\cdots] = \mathbb{E}_{\mathbf{y}}\mathbb{E}[\cdots|\mathbf{y}]$ 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 $\Longrightarrow$ $\boxed{\Sigma_\star^{(\mathrm{an})} = \sup_{\nu \in \mathcal{M}_1^+(\mathbb{R})} [\kappa_\alpha(\nu) + G(\nu) - \alpha H(\nu|\mathcal{N}(0,1))]}$ ∎

(Gradient density term in Kac-Rice)

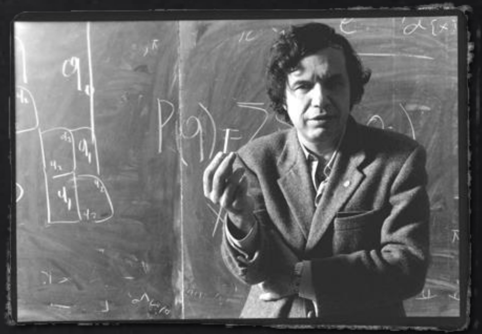
Extension 1: the quenched complexity

Replica trick

$$\mathbb{E}[\ln \mathrm{Crit}(f)] = \lim_{r \downarrow 0} \frac{\mathbb{E}\mathrm{Crit}(f)^r - 1}{r}$$

- Compute $\mathbb{E}[\mathrm{Crit}(f)^r]$ for $r \in \mathbb{N}$.
- Perform an analytical continuation to consider $r \downarrow 0$.

Physics 2021



Giorgio Parisi

Notably for describing the possible breaking of the symmetry between replicas in disordered systems

Replica theory is a prolific field of statistical physics

Refined Kac-Rice for $\mathbb{E}[\mathrm{Crit}(f)^r]$ + Replica trick

$$\Sigma_\star^{(\mathrm{qu.})}(B, Q) = \sup_{\nu \in \mathcal{M}_1(\mathbb{R})} [\cdots]$$

Closed formula for the quenched complexity

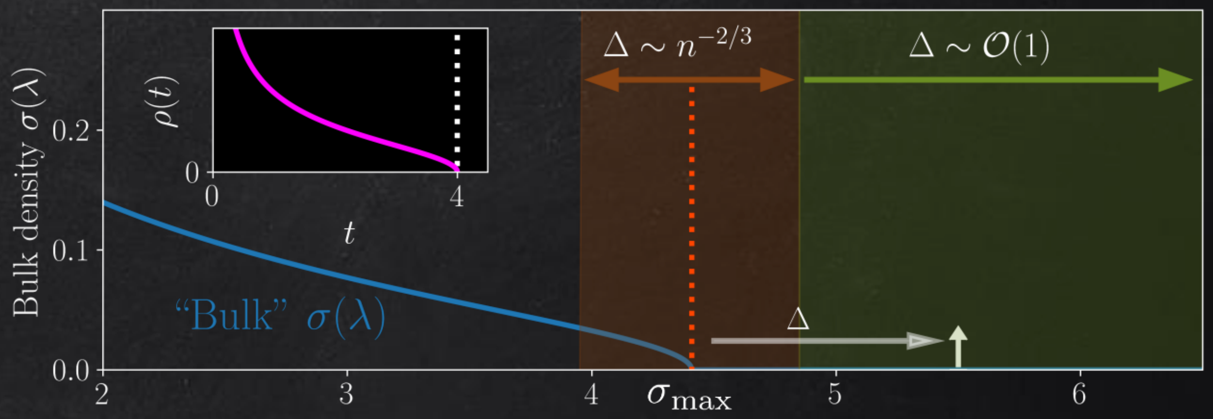
- Heuristic result, under a replica-symmetric ansatz.
- Drawback: Hard to solve, and even to interpret all terms so far…

Extension 2: Counting minima?

Goal: count only local minima, not all critical points !

Kac–Rice + restriction to local minima

Large deviations: $\frac{1}{n} \ln \mathbb{P}\left[\lambda_{\min}(\mathrm{Hess}\, f) \simeq x\right]$??



LDP for smallest/largest eigenvalue of "generalized covariance matrix" $\mathbf{M} = \frac{1}{m} \sum_{\mu=1}^{m} \rho_\mu \mathbf{z}_\mu \mathbf{z}_\mu^\dagger$?

Solved in [A.M. 21]

Real/complex

$$x \geq \sigma_{\max} : \quad \lim_{n \to \infty} \left\{ \frac{1}{n} \ln \mathbb{P}(\lambda_{\max}(\mathbf{M}) \simeq x) \right\} = -\frac{\beta}{2} \int_{\sigma_{\max}}^{x} [\overline{G}_\sigma(u) - G_\sigma(u)] \mathrm{d}u$$

$(G_\sigma(x), \overline{G}_\sigma(x))$ solutions to

$$x = \frac{1}{G} + \alpha \int \mathrm{d}t\, \rho(t) \frac{t}{\alpha - tG}$$

Dyson/Marchenko–Pastur equation

Using a technique based on a tilting of the measure.

[Biroli&Guionnet'20, Belinschi&al '20, Guionnet&al '20, Husson '20, Augeri&al '21]...

Summary & outlook

- First exact high-dimensional result obtained with Kac-Rice for non-Gaussian functions !
- Both annealed and quenched computations for the total complexity of critical points.
- Generalizes to other models: mixture of two Gaussians, binary linear classification... → Neural networks ?

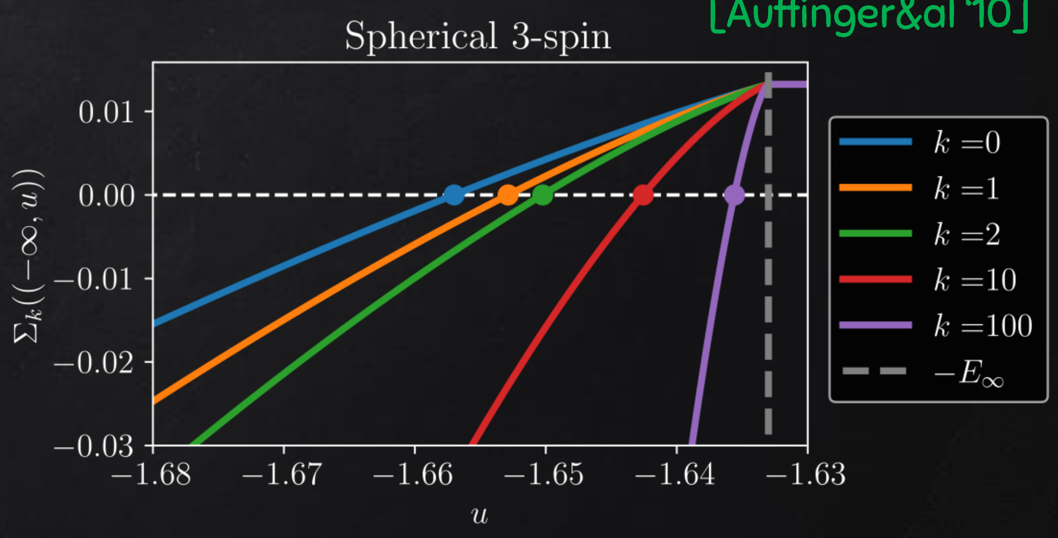
Physical discussion is still hard to reach. Many problems ahead:

- Numerically solve the variational problem? Hints in [A.M.&al '20]...

  Sign of the complexity given α, φ ? ...
- Count local minima ? We need a LDP for $\lambda_{\min}(\text{Hess } L(\mathbf{x}))$ ...

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

[Auffinger&al '10]



- Discrete systems ?
  - TAP approach ?
  - Recent algorithmic progress on F-RSB spin glasses [Subag '21, Montanari '21, El Alaoui & al '20].