Landscape complexity for the empirical risk of generalized linear models

A.M, Gérard Ben Arous, Giulio Biroli

arXiv:1912.02143

1 Landscape complexity

Main goal : Understand the landscape of the empirical risk in statistical estimation

To analyze local optimization algorithms
(e.g. gradient descent and stochastic variants)

In many nonconvex problems, one can provably find a regime in which the optimization landscape is 'easy' (matrix decomposition, tensor factorization, neural nets...) [Soudry&al '16, Ge&al '16, Ge&Ma '17, and many others]

In practice, algorithms work far beyond these regimes ! Why ?

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

→ Optimization algorithms can work in a 'hard' regime (i.e. many spurious local minima) ?

Can we analyze the topological transition in the landscape, and characterize the 'hard' regime ?

2 Generalized linear models

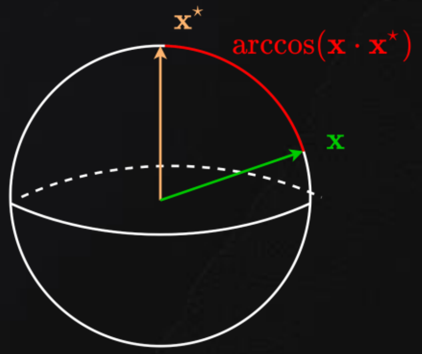
“Perceptron” energy

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

Squared loss of a generalized linear model

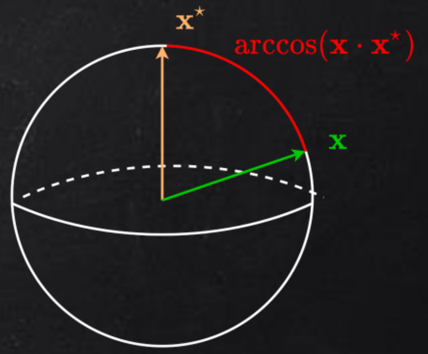
$$L_2(\mathbf{x}) = \frac{1}{2m}\sum_{\mu=1}^{m} \left[\phi(\boldsymbol{\xi}_\mu \cdot \mathbf{x}) - \phi(\boldsymbol{\xi}_\mu \cdot \mathbf{x}^\star)\right]^2$$

“Teacher–student” setup



- High dimensional limit $m, n \to \infty$ with $\alpha = m/n = \Theta(1)$.
- Can we count the number of critical points of these functions with a definite “energy” $L(\mathbf{x})$ and overlap $q = \mathbf{x} \cdot \mathbf{x}^\star$?

1. Landscape complexity



Number of critical points with "energy" $L(\mathbf{x})$ and overlap $q = \mathbf{x}\cdot\mathbf{x}^\star$?

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

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 $\Sigma_\star^{(\mathrm{qu.})}(B,Q) \equiv \lim_{n\to\infty} \frac{1}{n} \mathbb{E} \ln \mathrm{Crit}_\star(B,Q).$

⚠ In general, quenched and annealed are very different ! (Few exceptions [Subag '17])

3 The Kac-Rice formula

For a 1D function f(x), one would like to write : $\#\{x \text{ s.t.} f(x) = 0\} \stackrel{?}{=} \int \delta(f(x))|f'(x)|\mathrm{d}x$

The Kac-Rice formula makes this intuition precise $\Longrightarrow$ mean number of critical points of random $f : \mathbb{R}^n \to \mathbb{R}$:

$$\boxed{\mathbb{E}\mathrm{Crit}(f) = \int_{||\mathbf{x}||^2=1} \mathrm{d}\sigma(\mathbf{x}) \varphi_{\mathrm{grad} f(\mathbf{x})}(0) \mathbb{E}[|\det \mathrm{Hess}\ f(\mathbf{x})||\mathrm{grad}\ f(\mathbf{x}) = 0]}$$

Density of the gradient taken in 0

- Turns a random differential geometry problem into a random matrix theory problem.
- One can also fix the value of f(x), get higher-order moments...

We need the distribution of the Hessian, conditioned by the gradient being zero : intractable for "generic" functions !

→ Applications limited so far to Gaussian random functions

Pure p-spin and variants

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

- Large literature on the applications to Gaussian models, in physics and mathematics [Bray&Moore '80, Crisanti&al '95, Fyodorov&al '07, Auffinger&al '13, Ros&al '19, ...]. See [Adler&Taylor '07, Azais&Wschebor '09] for a mathematical introduction to Kac-Rice.

4 Main result

A first high-dimensional exact result with the Kac-Rice formula for a non-Gaussian random function !

Theorem (for L_1) $$\Sigma_\star^{(\mathrm{an.})}(B) = \frac{1+\ln\alpha}{2} + \sup_{\substack{\nu \in \mathcal{M}_1^+(\mathbb{R}) \\ \int \nu(\mathrm{d}t)\phi(t) \in B}} \Big[-\frac{1}{2}\ln\Big\{\int \nu(\mathrm{d}t)\phi'(t)^2\Big\} - \underbrace{\alpha H(\nu|\mathcal{N}(0,1))}_{\text{Relative entropy}} + \kappa_\alpha(\nu)\Big].$$

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

There is a similar theorem for L_2 : $\Sigma_\star^{(\mathrm{an})}(B,Q) = \sup_{q\in Q} \sup_{\nu \in \mathcal{M}_1^+(\mathbb{R}^2)} [\cdots]$

- Because of $\kappa_\alpha(\nu)$: very hard to solve in general ! (apart from trivial activation functions)
- We derive a heuristic closed formula of $\kappa_\alpha(\nu)$, based on [Marchenko&Pastur '67, Silverstein&Bai '95], which leads to simpler scalar fixed point equations (for L_1 and L_2) ⟶ Numerical solutions ? (Ongoing work)

5 Overview of the proof

We focus on L_1 (same method for L_2)

Kac–Rice formula $\Longrightarrow$ We need to study the Hessian, conditioned by the gradient being zero.

Main idea : Condition (gradient, Hessian) by the i.i.d. Gaussian random variables $y_\mu \equiv \boldsymbol{\xi}_\mu \cdot \mathbf{x}$

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

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

$\mathbf{z}_\mu$: i.i.d. standard Gaussian vectors
"Generalized" version of a sample covariance matrix
($\phi''(y_\mu)$ can be negative).

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

(Gradient density term In Kac–Rice)

6 Conclusion & Perspectives

Additional results

- We derive a similar closed formula for the quenched complexity $\Sigma_\star^{(\mathrm{qu.})}(B,Q)$. The derivation is based on the Kac-Rice formula for higher order moments and the non-rigorous replica method [Parisi&al '87, Ros&al '19].
- We generalize our results to other models: mixture of two Gaussians, binary linear classification, a simple unsupervised learning problem → Can we generalize to neural networks? (Open question)

Some future directions

- We present a first theoretical step: can we solve the variational problem and obtain numerical curves? (Ongoing work, cf the "Main Result" slide)
- We only derived formulas for the total number of critical points. Can we count only the local minima? To do so, we need the large deviations of the lowest eigenvalue of the Hessian.
  → We believe we can! (Ongoing work [A.M., to appear in 2020])

THANK YOU!