Exploring the random landscapes of inference

Part 1: Topological complexity

Gérard BEN AROUS

Courant Institute of Mathematical Sciences

ICTS, January 2020

January 8, 2020

Loss/Risk/Likelihood landscapes in statistical inference

- A common path to solve a statistical task, say parameter estimation (a.k.a learning), is through minimizing of a risk/loss (or maximizing a likelihood)

Loss/Risk/Likelihood landscapes in statistical inference

- A common path to solve a statistical task, say parameter estimation (a.k.a learning), is through minimizing of a risk/loss (or maximizing a likelihood)
- One typically would want to minimize the true risk/loss (or population risk/loss), which is given by

$$R(\theta) = E[L(X, \theta)] \tag{1}$$

Loss/Risk/Likelihood landscapes in statistical inference

- A common path to solve a statistical task, say parameter estimation (a.k.a learning), is through minimizing of a risk/loss (or maximizing a likelihood)
- One typically would want to minimize the true risk/loss (or population risk/loss), which is given by

$$R(\theta) = E[L(X, \theta)] \tag{1}$$

- Here θ is a parameter in R^N, to be estimated, by minimizing $R(\theta)$
- L is a loss function (for instance the negative of a log-likelihood), and X is the data, in R^D.

Loss/Risk/Likelihood landscapes in statistical inference

- Obviously the population risk in unknown.

Loss/Risk/Likelihood landscapes in statistical inference

- Obviously the population risk in unknown.
- But say an i.i.d sample of the distribution $(x_i)_{1\leq i\leq M}$, is given.

Loss/Risk/Likelihood landscapes in statistical inference

- Obviously the population risk in unknown.
- But say an i.i.d sample of the distribution $(x_i)_{1 \leq i \leq M}$, is given.
- So one tries to minimize the empirical risk, rather than the population risk

$$\hat{R}_M(\theta) = \frac{1}{M} \sum_{i=1}^{M} L(x_i, \theta) \tag{2}$$

Loss/Risk/Likelihood landscapes in statistical inference

- Obviously the population risk in unknown.
- But say an i.i.d sample of the distribution $(x_i)_{1 \leq i \leq M}$, is given.
- So one tries to minimize the empirical risk, rather than the population risk

$$\hat{R}_M(\theta) = \frac{1}{M} \sum_{i=1}^{M} L(x_i, \theta) \tag{2}$$

- Note: Here N is the dimension of the parameter, D the dimension of the data, and M the size of the sample.

Loss/Risk/Likelihood landscapes in statistical inference

Thus the natural questions are

- The Information Theory question (a la Shannon) : Is there any statistical procedure that can work? or is there an information theoretical threshold?

Loss/Risk/Likelihood landscapes in statistical inference

Thus the natural questions are

- The Information Theory question (a la Shannon) : Is there any statistical procedure that can work? or is there an information theoretical threshold?
- The statistical (or architecture) question: Is the minimization of our $R(\theta)$ a good strategy to estimate (learn) θ?

Loss/Risk/Likelihood landscapes in statistical inference

Thus the natural questions are

- The Information Theory question (a la Shannon) : Is there any statistical procedure that can work? or is there an information theoretical threshold?
- The statistical (or architecture) question: Is the minimization of our $R(\theta)$ a good strategy to estimate (learn) θ?
- i.e. Is our statistical procedure well chosen, what is its performance?, if we are ambitious: how to chose a near-optimal statistical procedure?

Loss/Risk/Likelihood landscapes in statistical inference

Thus the natural questions are

- The Information Theory question (a la Shannon) : Is there any statistical procedure that can work? or is there an information theoretical threshold?
- The statistical (or architecture) question: Is the minimization of our $R(\theta)$ a good strategy to estimate (learn) θ?
- i.e. Is our statistical procedure well chosen, what is its performance?, if we are ambitious: how to chose a near-optimal statistical procedure?
- In ML terms, this is the "architecture" question.

Loss/Risk/Likelihood landscapes in statistical inference

Loss/Risk/Likelihood landscapes in statistical inference

- Once the choice of the statistical procedure is done: Can the minimization of $\hat{R}_M(\theta)$ be a good substitute for the minimization of $R(\theta)$, in what regime for N,D and M?

Loss/Risk/Likelihood landscapes in statistical inference

- Once the choice of the statistical procedure is done: Can the minimization of $\hat{R}_M(\theta)$ be a good substitute for the minimization of $R(\theta)$, in what regime for N,D and M?
- in ML terms, this is the "generalization" question
- The algorithmic (or optimization) question: How can one find the minimum of $\hat{R}_M(\theta)$? Choice of the optimization algorithm? Guarantee of its performance?

Loss/Risk/Likelihood landscapes in statistical inference

- Once the choice of the statistical procedure is done: Can the minimization of $\hat{R}_M(\theta)$ be a good substitute for the minimization of $R(\theta)$, in what regime for N,D and M?
- in ML terms, this is the "generalization" question
- The algorithmic (or optimization) question: How can one find the minimum of $\hat{R}_M(\theta)$? Choice of the optimization algorithm? Guarantee of its performance?
- In ML terms this is the "training" question

Loss/Risk/Likelihood landscapes in statistical inference

- The function $\hat{R}_M$ is a smooth random function of many variables.

Loss/Risk/Likelihood landscapes in statistical inference

- The function $\hat{R}_M$ is a smooth random function of many variables.
- It is most often non-convex so its optimization might be difficult.

Loss/Risk/Likelihood landscapes in statistical inference

- The function $\hat{R}_M$ is a smooth random function of many variables.
- It is most often non-convex so its optimization might be difficult.
- A first question is then to understand the complexity of the topology/geometry of the random landscape defined by $\hat{R}_M(u)$

Loss/Risk/Likelihood landscapes in statistical inference

- The function $\hat{R}_M$ is a smooth random function of many variables.
- It is most often non-convex so its optimization might be difficult.
- A first question is then to understand the complexity of the topology/geometry of the random landscape defined by $\hat{R}_M(u)$
- Is there a glass phase? What is the topology of its level sets? How many critical points? How many minima? Are the deep minima close the true value of the parameter we need to estimate?

Loss/Risk/Likelihood landscapes in statistics

- The next question is to understand the performance of natural algorithms to explore this landscape and minimize $\hat{R}_M(u)$

Loss/Risk/Likelihood landscapes in statistics

- The next question is to understand the performance of natural algorithms to explore this landscape and minimize $\hat{R}_M(u)$
- For instance: Stochastic Gradient Descent, Gradient Descent, Langevin dynamics ?

Loss/Risk/Likelihood landscapes in statistics

- The next question is to understand the performance of natural algorithms to explore this landscape and minimize $\hat{R}_M(u)$
- For instance: Stochastic Gradient Descent, Gradient Descent, Langevin dynamics ?
- Role of algorithm, role of initialization, role of SNR (size of data)?

Topology of smooth functions: Morse Theory

- Consider a smooth real-valued function f on the N dimensional compact manifold M

Topology of smooth functions: Morse Theory

- Consider a smooth real-valued function f on the N dimensional compact manifold M
- How can one understand the topology of the "landscape" it defines?

Topology of smooth functions: Morse Theory

- Consider a smooth real-valued function f on the N dimensional compact manifold M
- How can one understand the topology of the "landscape" it defines?
- This question goes back to the 1920s, and Morse theory

Topology of smooth functions: Morse Theory

- Can one then understand the topology of its sub-level sets:

$$A(u) = \{x \in M, f(x) \leq u\} \tag{3}$$

Topology of smooth functions: Morse Theory

- Can one then understand the topology of its sub-level sets:

$$A(u) = \{x \in M, f(x) \leq u\} \tag{3}$$

- For instance, can one compute the Euler characteristic of the sub-level sets: $\chi(A(u))$?

Topology of smooth functions: Morse Theory

- Can one then understand the topology of its sub-level sets:

$$A(u) = \{x \in M, f(x) \leq u\} \tag{3}$$

- For instance, can one compute the Euler characteristic of the sub-level sets: $\chi(A(u))$?
- Can one compute the number of critical points of f, say $Crit^f_{N,k}(B)$ the number of critical points of f on the manifold M, of index k and with value in a subset B of the real line?

Topology of smooth functions: Morse Theory

- Can one then understand the topology of its sub-level sets:

$$A(u) = \{x \in M, f(x) \leq u\} \quad (3)$$

- For instance, can one compute the Euler characteristic of the sub-level sets: $\chi(A(u))$?
- Can one compute the number of critical points of f, say $Crit^f_{N,k}(B)$ the number of critical points of f on the manifold M, of index k and with value in a subset B of the real line?
- Morse inequalities give constraints on these numbers in terms of basic topological invariants (the Betti numbers), under the assumption that f is "generic", i.e. that the Hessian of f is non-denegerate at every critical point. (f is a Morse function).

Simple functions are not that simple

Simple functions are not that simple

- Consider the following (seemingly trivial) situation: the manifold M has a trivial topology, say M is the unit sphere S^{N-1} in N dimensions.

Simple functions are not that simple

- Consider the following (seemingly trivial) situation: the manifold M has a trivial topology, say M is the unit sphere S^{N-1} in N dimensions.
- Also assume that the function f is the simplest possible function, say a homogeneous polynomial of degree p.

Simple functions are not that simple

- Consider the following (seemingly trivial) situation: the manifold M has a trivial topology, say M is the unit sphere S^{N-1} in N dimensions.
- Also assume that the function f is the simplest possible function, say a homogeneous polynomial of degree p.
- How topologically complex can f be, if N diverges (and p is fixed)?

Simple functions are not that simple

- Consider the following (seemingly trivial) situation: the manifold M has a trivial topology, say M is the unit sphere S^{N-1} in N dimensions.
- Also assume that the function f is the simplest possible function, say a homogeneous polynomial of degree p.
- How topologically complex can f be, if N diverges (and p is fixed)?
- Obviously not very complex if $p = 1$ or $p = 2$!

Simple functions are not that simple

- But if $p \geq 3$, all hell breaks loose! Cubics can be terrible.

Simple functions are not that simple

- But if $p \geq 3$, all hell breaks loose! Cubics can be terrible.
- The maximal (finite) number of critical points for a homogeneous polynomial of degree p is $2[(p-1)^{N-1} + ... + (p-1) + 1]$

Simple functions are not that simple

- But if $p \geq 3$, all hell breaks loose! Cubics can be terrible.
- The maximal (finite) number of critical points for a homogeneous polynomial of degree p is $2[(p-1)^{N-1} + ... + (p-1) + 1]$
- There exists such a "worst-case" polynomial ! (Khozasov, 2018). So the worst homogeneous polynomials of degree $p \geq 3$ are exponentially complex! (the worst complexity is thus roughly $\log(p-1)$)

Sure, but this is only a worst-case!

- This is only a worst case scenario: let's be reasonable and pick the polynomial f to be more generic, say random! Assume the coefficients of f to be i.i.d Gaussian $N(0,1)$

Sure, but this is only a worst-case!

- This is only a worst case scenario: let's be reasonable and pick the polynomial f to be more generic, say random! Assume the coefficients of f to be i.i.d Gaussian $N(0, 1)$
- Then f is also exponentially complex when N diverges!

Sure, but this is only a worst-case!

- This is only a worst case scenario: let's be reasonable and pick the polynomial f to be more generic, say random! Assume the coefficients of f to be i.i.d Gaussian $N(0, 1)$
- Then f is also exponentially complex when N diverges!
- The (annealed) complexity is roughly half of the worst case.

Sure, but this is only a worst-case!

- This is only a worst case scenario: let's be reasonable and pick the polynomial f to be more generic, say random! Assume the coefficients of f to be i.i.d Gaussian $N(0, 1)$
- Then f is also exponentially complex when N diverges!
- The (annealed) complexity is roughly half of the worst case.
- How do we know? Two ingredients: the Kac-Rice formula plus Random Matrix Theory.

Sure, but this is only a worst-case!

- This is only a worst case scenario: let's be reasonable and pick the polynomial f to be more generic, say random! Assume the coefficients of f to be i.i.d Gaussian $N(0,1)$
- Then f is also exponentially complex when N diverges!
- The (annealed) complexity is roughly half of the worst case.
- How do we know? Two ingredients: the Kac-Rice formula plus Random Matrix Theory.
- For Kac-Rice formula, see the books by Azais and Wschebor, and by Adler and Taylor.

The Kac-Rice formula

- Consider a smooth random Gaussian function f on the N dimensional compact manifold M

The Kac-Rice formula

- Consider a smooth random Gaussian function f on the N dimensional compact manifold M
- The version of the Kac-Rice formula we will need reads

$$E[Crit_{N,k}^{f}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \tag{4}$$

The Kac-Rice formula

- Consider a smooth random Gaussian function f on the N dimensional compact manifold M
- The version of the Kac-Rice formula we will need reads

$$E[Crit^{f}_{N,k}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \tag{4}$$

- where

$$a_k(x,u) = E\Big[|\det \nabla^2(f)(x)|1_{i(x)=k}, \Big| f(x) = u, \nabla f(x) = 0\Big] \tag{5}$$

The Kac-Rice formula

- Consider a smooth random Gaussian function f on the N dimensional compact manifold M
- The version of the Kac-Rice formula we will need reads

$$E[Crit^{f}_{N,k}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \tag{4}$$

- where

$$a_k(x,u) = E\left[|\det \nabla^2(f)(x)| 1_{i(x)=k}, \middle| f(x) = u, \nabla f(x) = 0\right] \tag{5}$$

- where $i(x)$ is the index of the Hessian of f at x, and $\phi_x(u,v)$ is the density of the law of the gaussian vector $(f(x), \nabla f(x))$

The Kac-Rice formula

- Consider a smooth random Gaussian function f on the N dimensional compact manifold M
- The version of the Kac-Rice formula we will need reads

$$E[Crit_{N,k}^{f}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \qquad (4)$$

- where

$$a_k(x,u) = E\left[|\det \nabla^2(f)(x)|1_{i(x)=k}, \Big| f(x) = u, \nabla f(x) = 0\right] \qquad (5)$$

- where $i(x)$ is the index of the Hessian of f at x, and $\phi_x(u,v)$ is the density of the law of the gaussian vector $(f(x), \nabla f(x))$
- In fact Kac-Rice formulae also give higher moments of the number of critical points, and the Euler characteristic of level sets.

The link with Random Matrix Theory

- The important message: Kac-Rice (in high dimension) is a powerful link between questions of random geometry and Random Matrix Theory (RMT).

The link with Random Matrix Theory

- The important message: Kac-Rice (in high dimension) is a powerful link between questions of random geometry and Random Matrix Theory (RMT).
- The link with RMT is that this formula reduces the study of the moments of the number of critical points to the understanding of the distribution of the absolute value of the determinant of the Hessian of f at x conditionally on x being a critical point, and on $f(x) = u$

The link with Random Matrix Theory

- The important message: Kac-Rice (in high dimension) is a powerful link between questions of random geometry and Random Matrix Theory (RMT).
- The link with RMT is that this formula reduces the study of the moments of the number of critical points to the understanding of the distribution of the absolute value of the determinant of the Hessian of f at x conditionally on x being a critical point, and on $f(x) = u$
- This is the law of a NxN Gaussian random real symmetric matrix.

The link with Random Matrix Theory

- The important message: Kac-Rice (in high dimension) is a powerful link between questions of random geometry and Random Matrix Theory (RMT).
- The link with RMT is that this formula reduces the study of the moments of the number of critical points to the understanding of the distribution of the absolute value of the determinant of the Hessian of f at x conditionally on x being a critical point, and on $f(x) = u$
- This is the law of a NxN Gaussian random real symmetric matrix.
- Its covariance structure defines a 4-tensor, which is computable by differentiating the Covariance function C (plus some linear algebra to take the conditioning into account).

The link with Random Matrix Theory

The link with Random Matrix Theory

- So the covariance function C defines a Random Matrix model of Gaussian matrices, with dependent entries in general

The link with Random Matrix Theory

- So the covariance function C defines a Random Matrix model of Gaussian matrices, with dependent entries in general
- This class of random matrix models is hard in general. See the recent work by Laszlo Erdos and his collaborators.

The link with Random Matrix Theory

- So the covariance function C defines a Random Matrix model of Gaussian matrices, with dependent entries in general
- This class of random matrix models is hard in general. See the recent work by Laszlo Erdos and his collaborators.
- But for important classes of examples, the RMT model is tractable

The link with Random Matrix Theory

- So the covariance function C defines a Random Matrix model of Gaussian matrices, with dependent entries in general
- This class of random matrix models is hard in general. See the recent work by Laszlo Erdos and his collaborators.
- But for important classes of examples, the RMT model is tractable
- We will cover three classes of examples: Spherical Spin-Glasses, Tensor PCA, and Generalized Linear Models (one node networks).

The link with Random Matrix Theory

- So the covariance function C defines a Random Matrix model of Gaussian matrices, with dependent entries in general
- This class of random matrix models is hard in general. See the recent work by Laszlo Erdos and his collaborators.
- But for important classes of examples, the RMT model is tractable
- We will cover three classes of examples: Spherical Spin-Glasses, Tensor PCA, and Generalized Linear Models (one node networks).
- Hoping to understand more complex Machine Learning networks in the near future

Example 1: Spherical Spin Glasses energy landscapes

- The Hamiltonian of the pure p-spherical spin glass is given, for $x \in S^{N-1}(\sqrt{N})$, by

$$H(x) = \frac{1}{N^{(p-1)/2}} \sum_{i_1, i_2, \ldots, i_p} J_{i_1 \ldots i_p} x_{i_1} x_{i_2} \ldots x_{i_p} \tag{6}$$

  where the coupling constants J are i.i.d N(0,1).

Example 1: Spherical Spin Glasses energy landscapes

- The Hamiltonian of the pure p-spherical spin glass is given, for $x \in S^{N-1}(\sqrt{N})$, by

$$H(x) = \frac{1}{N^{(p-1)/2}} \sum_{i_1, i_2, \dots, i_p} J_{i_1 \dots i_p} x_{i_1} x_{i_2} \dots x_{i_p} \tag{6}$$

  where the coupling constants J are i.i.d N(0,1).
- This Hamiltonian is (up to trivial normalizations) the random homogeneous polynomial of degree p mentioned above!

Example 1: Spherical Spin Glasses energy landscapes

- The Hamiltonian of the pure p-spherical spin glass is given, for $x \in S^{N-1}(\sqrt{N})$, by

$$H(x) = \frac{1}{N^{(p-1)/2}} \sum_{i_1, i_2, \ldots, i_p} J_{i_1 \ldots i_p} x_{i_1} x_{i_2} \ldots x_{i_p} \tag{6}$$

  where the coupling constants J are i.i.d N(0,1).
- This Hamiltonian is (up to trivial normalizations) the random homogeneous polynomial of degree p mentioned above!
- We understand the (annealed) complexity of critical points of fixed index below a given level, the topology of the level sets, the quenched complexity at very low energy levels, the absolute minimum (the ground state), the lowest local minima... See Auffinger-BA-Cerny (2013), Subag(2015), Subag-Zeitouni (2017)

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- Understanding precisely the bottom of this energy landscape gives very sharp information on the Gibbs measure at low temperature. (See Subag 2017).

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- Understanding precisely the bottom of this energy landscape gives very sharp information on the Gibbs measure at low temperature. (See Subag 2017).
- It gives a precise detailed geometric picture of the 1 RSB phase (1 step Replica Symmetry Breaking)
- This description is much more precise than the one given by the Parisi description of the order parameter, i.e. the overlap distribution. For instance, it implies that there is no chaos in temperature.

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- This question can also be studied for certain general (i.e. mixed) spherical spin glass models. This corresponds to non homogeneous polynomials. See Auffinger-BA(2013), Subag (2018), BA-Subag-Zeitouni (2019)

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- This question can also be studied for certain general (i.e. mixed) spherical spin glass models. This corresponds to non homogeneous polynomials. See Auffinger-BA(2013), Subag (2018), BA-Subag-Zeitouni (2019)
- But in the mixed case, a lot more needs to be done.

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- This question can also be studied for certain general (i.e. mixed) spherical spin glass models. This corresponds to non homogeneous polynomials. See Auffinger-BA(2013), Subag (2018), BA-Subag-Zeitouni (2019)
- But in the mixed case, a lot more needs to be done.
- For instance, to understand the quenched complexity in general cases, and to cover Full RSB cases.

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- This question can also be studied for certain general (i.e. mixed) spherical spin glass models. This corresponds to non homogeneous polynomials. See Auffinger-BA(2013), Subag (2018), BA-Subag-Zeitouni (2019)
- But in the mixed case, a lot more needs to be done.
- For instance, to understand the quenched complexity in general cases, and to cover Full RSB cases.
- The problem here is that the Kac-Rice formula is "annealed", it deals only with moments.

Example 1: Spherical Spin Glasses Gibbs measures at low temperature

- This question can also be studied for certain general (i.e. mixed) spherical spin glass models. This corresponds to non homogeneous polynomials. See Auffinger-BA(2013), Subag (2018), BA-Subag-Zeitouni (2019)
- But in the mixed case, a lot more needs to be done.
- For instance, to understand the quenched complexity in general cases, and to cover Full RSB cases.
- The problem here is that the Kac-Rice formula is "annealed", it deals only with moments.
- If you want to jump in: try a mixture of degree 3 and degree 16.

Example 2: One important "hard" statistical example: Tensor PCA

- One observes an M-sample of a "noisy" p-tensor in N variables $T_i = \lambda v^{\otimes p} + Z_i$
- Here v is a fixed unknown vector on the unit sphere S^{N-1}, and the Z_i's are random i.i.d centered p-tensors.
- λ is a signal-to-noise ratio.
- The objective is to detect and recover (i.e. estimate) v.
- The same question could be asked if the signal were a low rank tensor (rather than a rank one tensor as here).

Example 2: One important "hard" example: Tensor PCA

- We assume that the noise Z is Gaussian, and that its entries are i.i.d N(0,1).
- We assume no prior information on the spike $v \in S^{N-1}$, i.e. our prior is the uniform measure on S^{N-1}

Example 2: One important "hard" example: Tensor PCA

- We assume that the noise Z is Gaussian, and that its entries are i.i.d N(0,1).
- We assume no prior information on the spike $v \in S^{N-1}$, i.e. our prior is the uniform measure on S^{N-1}
- The problem is well studied, for instance by Montanari-Reichman-Zeitouni 2015, Ge-Ma 2016, Montanari-Richard 2016, T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, L. Zdeborova, 2017, and Perry-Wein-Bandeira 2017

Example 2: The three thresholds for Tensor PCA

- Question 1: Detection, and the IT threshold.
- Is it possible to detect the signal? i.e the TV distance between the distribution with a SNR $\lambda > 0$ and the distribution with no signal $\lambda = 0$, going to 0 or not?

Example 2: The three thresholds for Tensor PCA

- Question 1: Detection, and the IT threshold.
- Is it possible to detect the signal? i.e the TV distance between the distribution with a SNR $\lambda > 0$ and the distribution with no signal $\lambda = 0$, going to 0 or not?
- Question 2: Recovery, and the statistical threshold.

Example 2: The three thresholds for Tensor PCA

- Question 1: Detection, and the IT threshold.
- Is it possible to detect the signal? i.e the TV distance between the distribution with a SNR $\lambda > 0$ and the distribution with no signal $\lambda = 0$, going to 0 or not?
- Question 2: Recovery, and the statistical threshold.
- Pick a statistical method (say MLE). Above what threshold for the SNR λ is the estimator of the signal v better than random? When is it converging to the true v (strong or full recovery)

Example 2: The three thresholds for Tensor PCA

- Question 1: Detection, and the IT threshold.
- Is it possible to detect the signal? i.e the TV distance between the distribution with a SNR $\lambda > 0$ and the distribution with no signal $\lambda = 0$, going to 0 or not?
- Question 2: Recovery, and the statistical threshold.
- Pick a statistical method (say MLE). Above what threshold for the SNR λ is the estimator of the signal v better than random? When is it converging to the true v (strong or full recovery)
- Question 3: Computation. The algorithmic or computational threshold.

Example 2: The three thresholds for Tensor PCA

- Question 1: Detection, and the IT threshold.
- Is it possible to detect the signal? i.e the TV distance between the distribution with a SNR $\lambda > 0$ and the distribution with no signal $\lambda = 0$, going to 0 or not?
- Question 2: Recovery, and the statistical threshold.
- Pick a statistical method (say MLE). Above what threshold for the SNR λ is the estimator of the signal v better than random? When is it converging to the true v (strong or full recovery)
- Question 3: Computation. The algorithmic or computational threshold.
- Pick a computational algorithm (say GD, or SGD), for what SNR does it recovers the signal, in short time scales?

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.
- Question 2: the statistical threshold for (partial) recovery by MLE.

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.
- Question 2: the statistical threshold for (partial) recovery by MLE.
- $\lambda_{MLE} = \lambda_{IT} = 1$

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.
- Question 2: the statistical threshold for (partial) recovery by MLE.
- $\lambda_{MLE} = \lambda_{IT} = 1$
- When it is possible to detect the signal, it is possible to recover it (partially) by the plain vanilla method, MLE!

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.
- Question 2: the statistical threshold for (partial) recovery by MLE.
- $\lambda_{MLE} = \lambda_{IT} = 1$
- When it is possible to detect the signal, it is possible to recover it (partially) by the plain vanilla method, MLE!
- This Question 2 is where the Kac Rice formula is useful.

Example 2: The three thresholds for Tensor PCA

- Question 1: the detection threshold.
- With the proper normalization $\lambda_{IT} = 1$ Below this threshold, the two distributions (with a signal or without) are indistinguishable. Above it, their distance remains (asymptotically) positive.
- Question 2: the statistical threshold for (partial) recovery by MLE.
- $\lambda_{MLE} = \lambda_{IT} = 1$
- When it is possible to detect the signal, it is possible to recover it (partially) by the plain vanilla method, MLE!
- This Question 2 is where the Kac Rice formula is useful.
- See BA-Montanari-Mei-Nica (2019), Ros-BA-Biroli-Cammarota (2019)

Example 2: The GAP

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.
- The simple optimization algorithms to find the MLE (GD, Langevin, SGD) work in short time scales only when the SNR is larger than $N^{(p-2)/2}$ (see BA-Ghessairi-Jagannath 2019).

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.
- The simple optimization algorithms to find the MLE (GD, Langevin, SGD) work in short time scales only when the SNR is larger than $N^{(p-2)/2}$ (see BA-Ghessairi-Jagannath 2019).
- This is the so-called computational/statistical GAP.

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.
- The simple optimization algorithms to find the MLE (GD, Langevin, SGD) work in short time scales only when the SNR is larger than $N^{(p-2)/2}$ (see BA-Ghessairi-Jagannath 2019).
- This is the so-called computational/statistical GAP.
- This threshold is also the threshold for many other algorithms (like Approximate Message Passing, see Lesieur et al 2017))

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.
- The simple optimization algorithms to find the MLE (GD, Langevin, SGD) work in short time scales only when the SNR is larger than $N^{(p-2)/2}$ (see BA-Ghessairi-Jagannath 2019).
- This is the so-called computational/statistical GAP.
- This threshold is also the threshold for many other algorithms (like Approximate Message Passing, see Lesieur et al 2017))
- But there are other less naive algorithms that work better, i.e. above a lower threshold $N^{(p-2)/4}$

Example 2: The GAP

- Question 3: the algorithmic or computational threshold and the computation/statistical GAP.
- The simple optimization algorithms to find the MLE (GD, Langevin, SGD) work in short time scales only when the SNR is larger than $N^{(p-2)/2}$ (see BA-Ghessairi-Jagannath 2019).
- This is the so-called computational/statistical GAP.
- This threshold is also the threshold for many other algorithms (like Approximate Message Passing, see Lesieur et al 2017))
- But there are other less naive algorithms that work better, i.e. above a lower threshold $N^{(p-2)/4}$
- These include: semidefinite relaxations (the SOS hierarchy) see Bandeira et al 2017, the Kikuchi hierarchy spectral algorithms (see Wein, Al Alaoui, Moore 2019), the replicated gradient descent (Ros, Biroli, Cammarota 2019)

Example 2: MLE for Tensor PCA

Example 2: MLE for Tensor PCA

- This estimator v_{MLE} is obtained by solving the following optimization problem
- Find the minimum of $\hat{R}_M(u) = -\frac{1}{M}\sum_{i=1}^{M} < T_i, u^{\otimes p} >$ for $u \in S^{N-1}$

Example 2: MLE for Tensor PCA

- This estimator v_{MLE} is obtained by solving the following optimization problem
- Find the minimum of $\hat{R}_M(u) = -\frac{1}{M}\sum_{i=1}^{M} < T_i, u^{\otimes p} >$ for $u \in S^{N-1}$
- $\hat{R}_M$ is thus the following homogeneous random polynomial of degree p

$$\hat{R}_M(u) = \lambda < v, u >^p + \sum_{i_1, i_2, \ldots, i_p} Z_{i_1, i_2, \ldots, i_p} u_{i_1} u_{i_2} \ldots u_{i_p} \tag{7}$$

Example 2: MLE for Tensor PCA

- This estimator v_{MLE} is obtained by solving the following optimization problem
- Find the minimum of $\hat{R}_M(u) = -\frac{1}{M}\sum_{i=1}^{M} < T_i, u^{\otimes p} >$ for $u \in S^{N-1}$
- $\hat{R}_M$ is thus the following homogeneous random polynomial of degree p

$$\hat{R}_M(u) = \lambda < v, u >^p + \sum_{i_1, i_2, \ldots, i_p} Z_{i_1, i_2, \ldots, i_p} u_{i_1} u_{i_2} \ldots u_{i_p} \tag{7}$$

- Obviously here one can assume that $M = 1$ by changing the signal-to-noise ratio λ to $\lambda\sqrt{M}$

Example 2: MLE for Tensor PCA

- This estimator v_{MLE} is obtained by solving the following optimization problem
- Find the minimum of $\hat{R}_M(u) = -\frac{1}{M}\sum_{i=1}^{M} < T_i, u^{\otimes p} >$ for $u \in S^{N-1}$
- $\hat{R}_M$ is thus the following homogeneous random polynomial of degree p

$$\hat{R}_M(u) = \lambda < v, u >^p + \sum_{i_1, i_2, \ldots, i_p} Z_{i_1, i_2, \ldots, i_p} u_{i_1} u_{i_2} \ldots u_{i_p} \qquad (7)$$

- Obviously here one can assume that $M = 1$ by changing the signal-to-noise ratio λ to $\lambda\sqrt{M}$
- We can also assume that the unknown signal is $v = e_1$ (by invariance by rotation of the distribution of the Gaussian noise, and of the uniform prior). so that

$$\hat{R}(u) = \lambda u_1^p + \sum_{i_1, i_2, \ldots, i_p} Z_{i_1, i_2, \ldots, i_p} u_{i_1} u_{i_2} \ldots u_{i_p} \qquad (8)$$

Example 2: MLE for Tensor PCA

Example 2: MLE for Tensor PCA

- Without a signal, i.e. when $\lambda = 0$, the function $\hat{R}_M(u)$ is well known. It is the p-spin spherical Hamiltonian! We know it is complex!

Example 2: MLE for Tensor PCA

- Without a signal, i.e. when $\lambda = 0$, the function $\hat{R}_M(u)$ is well known. It is the p-spin spherical Hamiltonian! We know it is complex!
- Does this likelihood function stay complex when the SNR is present? when it is large?

Example 2: MLE for Tensor PCA

- Without a signal, i.e. when $\lambda = 0$, the function $\hat{R}_M(u)$ is well known. It is the p-spin spherical Hamiltonian! We know it is complex!
- Does this likelihood function stay complex when the SNR is present? when it is large?
- We need to understand if the global minimum of $\hat{R}$ is close to the signal v to check partial recovery.

Example 2: MLE for Tensor PCA

- Without a signal, i.e. when $\lambda = 0$, the function $\hat{R}_M(u)$ is well known. It is the p-spin spherical Hamiltonian! We know it is complex!
- Does this likelihood function stay complex when the SNR is present? when it is large?
- We need to understand if the global minimum of $\hat{R}$ is close to the signal v to check partial recovery.
- But we will see that this function is very complex and has exponentially many local minima!

Example 2: MLE for Tensor PCA

- Without a signal, i.e. when $\lambda = 0$, the function $\hat{R}_M(u)$ is well known. It is the p-spin spherical Hamiltonian! We know it is complex!
- Does this likelihood function stay complex when the SNR is present? when it is large?
- We need to understand if the global minimum of $\hat{R}$ is close to the signal v to check partial recovery.
- But we will see that this function is very complex and has exponentially many local minima!
- Where is the global minimum? Close to the signal? or lost in the entropy of the equator?

Example 2: Landscape complexity for MLE in Tensor PCA

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!
- The RMT problem brought up by the Kac-Rice formula is here the following: understand the spectrum of a rank one perturbation of the GOE.

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!
- The RMT problem brought up by the Kac-Rice formula is here the following: understand the spectrum of a rank one perturbation of the GOE.
- Using Kac-Rice, the question becomes: when does the top eigenvalue of a spiked GOE get out of the bulk of the spectrum?

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!
- The RMT problem brought up by the Kac-Rice formula is here the following: understand the spectrum of a rank one perturbation of the GOE.
- Using Kac-Rice, the question becomes: when does the top eigenvalue of a spiked GOE get out of the bulk of the spectrum?
- This is well understood, as the BBP transition.

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!
- The RMT problem brought up by the Kac-Rice formula is here the following: understand the spectrum of a rank one perturbation of the GOE.
- Using Kac-Rice, the question becomes: when does the top eigenvalue of a spiked GOE get out of the bulk of the spectrum?
- This is well understood, as the BBP transition.
- There is a long series of works on this type of questions, first asked by I. Johnstone, and started in 2005 in BA-Baik-Peche, continued by S. Peche, M. Capitaine, Benaych-Georges, D. Feral, C. Donati-Martin.

Example 2: Landscape complexity for MLE in Tensor PCA

- Use Kac-Rice again!
- The RMT problem brought up by the Kac-Rice formula is here the following: understand the spectrum of a rank one perturbation of the GOE.
- Using Kac-Rice, the question becomes: when does the top eigenvalue of a spiked GOE get out of the bulk of the spectrum?
- This is well understood, as the BBP transition.
- There is a long series of works on this type of questions, first asked by I. Johnstone, and started in 2005 in BA-Baik-Peche, continued by S. Peche, M. Capitaine, Benaych-Georges, D. Feral, C. Donati-Martin.
- The important result here is an LDP for the top eigenvalue proven by M. Maida in 2007.

Example 2: Landscape complexity for MLE in Tensor PCA

- The total complexity is always exponential in N.

Example 2: Landscape complexity for MLE in Tensor PCA

- The total complexity is always exponential in N.
- Indeed, close enough to the equator, the function is close to a pure spherical spin glass. So it is complex there, whatever the value of the SNR..

Example 2: Landscape complexity for MLE in Tensor PCA

- The total complexity is always exponential in N.
- Indeed, close enough to the equator, the function is close to a pure spherical spin glass. So it is complex there, whatever the value of the SNR..
- When $\lambda > 1$ a ring of critical points appears at a positive latitude (closer to the signal than a random point) which contains the absolute minimum (the MLE estimator). The number of critical point on this ring is sub-exponential.

Example 2: Landscape complexity for MLE in Tensor PCA

- The total complexity is always exponential in N.
- Indeed, close enough to the equator, the function is close to a pure spherical spin glass. So it is complex there, whatever the value of the SNR..
- When $\lambda > 1$ a ring of critical points appears at a positive latitude (closer to the signal than a random point) which contains the absolute minimum (the MLE estimator). The number of critical point on this ring is sub-exponential.
- When λ grows, the latitude of this ring of critical points increases, and the ring converges to the signal (the north pole). The MLE converges to the signal (asymptotic strong recovery)

The complexity of the landscape of Tensor PCA

- For M a subset of $[-1,1]$ and E a subset of the real line, let $Crit(M,E)$ be the number of critical points $x \in S^{N-1}$ such that $x_1 \in M$ and $\hat{R}_M(x) \in E$
- Theorem (GBA, Mei-Montanari, Nica, 2018):

$$\limsup \frac{1}{N} \log E[Crit(M,E)] \leq -\inf(S(m,e), m \in \bar{M}, e \in \bar{E}) \tag{9}$$

$$\liminf \frac{1}{N} \log E[Crit(M,E)] \geq -\inf(S(m,e), m \in Int(M), e \in Int(E)) \tag{10}$$

The complexity of the landscape of Tensor PCA

- The function S(m,e) is given by

$$S(m,e) = U(m) + \Phi(e) - p\lambda^2(m^{2p-2}(1-m^2) - (e-\lambda m^p)^2 \tag{11}$$

- Where, for $|e| \leq 2$

$$\Phi(e) = \frac{e^2}{4} - \frac{1}{2} \tag{12}$$

- and, for $|e| \geq 2$

$$\Phi(e) = \frac{e^2}{4} - \frac{1}{2} - \frac{|e|}{4}\sqrt{e^2-4} + \log(\sqrt{\frac{e^2}{4} - 1} + \frac{|e|}{2}) \tag{13}$$

and

$$U(m) = \frac{1}{2}(\log(p-1) + 1) + \log(1-m^2)) \tag{14}$$

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.
- When λ grows this band grows

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.
- When λ grows this band grows
- When $\lambda_1 < \lambda < \lambda_2$ a sub-exponential number of critical points appear at a latitude $x_1 > 0$ away from this band, but the aboslute minimum is still at the equator.

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.
- When λ grows this band grows
- When $\lambda_1 < \lambda < \lambda_2$ a sub-exponential number of critical points appear at a latitude $x_1 > 0$ away from this band, but the aboslute minimum is still at the equator.
- When $\lambda_2 < \lambda$ the minimum is achieved at this higher latitude. Weak recovery is possible.

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.
- When λ grows this band grows
- When $\lambda_1 < \lambda < \lambda_2$ a sub-exponential number of critical points appear at a latitude $x_1 > 0$ away from this band, but the aboslute minimum is still at the equator.
- When $\lambda_2 < \lambda$ the minimum is achieved at this higher latitude. Weak recovery is possible.
- In fact λ_2 can be checked to be the IT threshold i.e. $\lambda_2 = 1$. So when detection is at all possible, the ML estimator can do it!

The complexity of the landscape of Tensor PCA

- An explicit result is also valid for the number of local minima.
- These results show an interesting transition
- When $\lambda < \lambda_1$, there is an exponential number of critical points (and local minima) in a band near the equator $x_1 = 0$, and nowhere else.
- When λ grows this band grows
- When $\lambda_1 < \lambda < \lambda_2$ a sub-exponential number of critical points appear at a latitude $x_1 > 0$ away from this band, but the aboslute minimum is still at the equator.
- When $\lambda_2 < \lambda$ the minimum is achieved at this higher latitude. Weak recovery is possible.
- In fact λ_2 can be checked to be the IT threshold i.e. $\lambda_2 = 1$. So when detection is at all possible, the ML estimator can do it!
- When λ tends to ∞ the latitude tends to 1: asymptotic strong recovery

Complexity is not the only hard problem for Tensor PCA

- The optimal threshold $N^{(p-2)/2}$ results for the Gradient Descent and Langevin dynamics for Tensor PCA model are obtained in "Thresholds for signal recovery via Langevin dynamics" 2019 with Reza Gheissairi and Aukosh Jagannath, and an upcoming work which give the same threshold for a SGD algorithm.

Complexity is not the only hard problem for Tensor PCA

- The optimal threshold $N^{(p-2)/2}$ results for the Gradient Descent and Langevin dynamics for Tensor PCA model are obtained in "Thresholds for signal recovery via Langevin dynamics" 2019 with Reza Gheissairi and Aukosh Jagannath, and an upcoming work which give the same threshold for a SGD algorithm.
- The performance of these simple optimization algorithms is hampered by another important problem, different from the landscape complexity, re initialization: " Escaping mediocrity"

Complexity is not the only hard problem for Tensor PCA

- The optimal threshold $N^{(p-2)/2}$ results for the Gradient Descent and Langevin dynamics for Tensor PCA model are obtained in "Thresholds for signal recovery via Langevin dynamics" 2019 with Reza Gheissairi and Aukosh Jagannath, and an upcoming work which give the same threshold for a SGD algorithm.
- The performance of these simple optimization algorithms is hampered by another important problem, different from the landscape complexity, re initialization: " Escaping mediocrity"
- Even in the simple phase of the topological transition, the weakness of the signal in the region of maximal entropy for the prior makes recovery impossible in polynomial time.

Complexity is not the only hard problem for Tensor PCA

- The optimal threshold $N^{(p-2)/2}$ results for the Gradient Descent and Langevin dynamics for Tensor PCA model are obtained in "Thresholds for signal recovery via Langevin dynamics" 2019 with Reza Gheissairi and Aukosh Jagannath, and an upcoming work which give the same threshold for a SGD algorithm.
- The performance of these simple optimization algorithms is hampered by another important problem, different from the landscape complexity, re initialization: " Escaping mediocrity"
- Even in the simple phase of the topological transition, the weakness of the signal in the region of maximal entropy for the prior makes recovery impossible in polynomial time.
- Indeed if the drift induced by the signal is too weak, the algorithm will linger too long near the equator, i.e. in an exponentially complex landscape, close to the spherical spin glass and will end up trapped by complexity for very long times.

Example 3: Landscape Complexity for the perceptron and Generalized Linear Models

- Consider now the following random loss function

$$L_1(x) = \frac{1}{M} \sum_{\mu=1}^{M} \phi(\xi_\mu . x) \tag{15}$$

- where $x \in S^{N-1}$, the data ξ_μ are i.i.d vectors in R^N and ϕ is a smooth activation function.

Example 3: Landscape Complexity for the perceptron and Generalized Linear Models

- Consider now the following random loss function

$$L_1(x) = \frac{1}{M} \sum_{\mu=1}^{M} \phi(\xi_\mu . x) \tag{15}$$

- where $x \in S^{N-1}$, the data ξ_μ are i.i.d vectors in R^N and ϕ is a smooth activation function.
- Define also the 'planted" version

$$L_2(x) = \frac{1}{M} \sum_{\mu=1}^{M} [\phi(\xi_\mu . x) - \phi(\xi_\mu . x^*)]^2 \tag{16}$$

- where $x^* \in S^{N-1}$ is a planted signal

Example 3: The perceptron and Generalized Linear Models

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...
- Huge literature, see for instance Barbier-Krzakala-Macris-Miolane-Zdeborova (2018)

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...
- Huge literature, see for instance Barbier-Krzakala-Macris-Miolane-Zdeborova (2018)
- Question 1: What is the complexity of these functions?

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...
- Huge literature, see for instance Barbier-Krzakala-Macris-Miolane-Zdeborova (2018)
- Question 1: What is the complexity of these functions?
- Joint work with G. Biroli and A.Maillard, 2020. Submitted tomorrow.

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...
- Huge literature, see for instance Barbier-Krzakala-Macris-Miolane-Zdeborova (2018)
- Question 1: What is the complexity of these functions?
- Joint work with G. Biroli and A.Maillard, 2020. Submitted tomorrow.
- Question 2: How hard is it to find the minimum in short time scales, with GD or SGD algorithms randomly initialized?

Example 3: The perceptron and Generalized Linear Models

- These two random loss functions cover many well studied estimation problems, depending an the activation function
- Linear regression (if ϕ is linear), phase retrieval (when ϕ is quadratic), GLM, teacher-student network or one node network, ...
- Huge literature, see for instance Barbier-Krzakala-Macris-Miolane-Zdeborova (2018)
- Question 1: What is the complexity of these functions?
- Joint work with G. Biroli and A.Maillard, 2020. Submitted tomorrow.
- Question 2: How hard is it to find the minimum in short time scales, with GD or SGD algorithms randomly initialized?
- Current work with R. Ghessairi and A. Jagannath

Example 3: Landscape complexity for GLMs

Example 3: Landscape complexity for GLMs

- Under natural smoothness assumptions on the activation function, we can compute the complexity of the functions L_1 and L_2, in the high dimensional setting, where both N and M tend to infinity with their ratio converging to α.

Example 3: Landscape complexity for GLMs

- Under natural smoothness assumptions on the activation function, we can compute the complexity of the functions L_1 and L_2, in the high dimensional setting, where both N and M tend to infinity with their ratio converging to α.
- We have to assume that the data is Gaussian. The ξ_μ are i.i.d standard Gaussian in R^N.

Example 3: Landscape complexity for GLMs

- Under natural smoothness assumptions on the activation function, we can compute the complexity of the functions L_1 and L_2, in the high dimensional setting, where both N and M tend to infinity with their ratio converging to α.
- We have to assume that the data is Gaussian. The ξ_μ are i.i.d standard Gaussian in R^N.
- Method: a simple extension of Kac-Rice to random functions which are not Gaussian but a smooth function of a Gaussian random function. (as in Azais-Wschebor)

Example 3: What is the RMT question for GLMs?

Example 3: What is the RMT question for GLMs?

- If we can indeed apply a version of Kac-Rice, what is the natural RMT question?

Example 3: What is the RMT question for GLMs?

- If we can indeed apply a version of Kac-Rice, what is the natural RMT question?
- Answer: Spiked Generalized Wishart matrices, i.e. rank-one perturbations of general Wishart (or Pastur-Marchenko) matrices

$$H = XDX^t \tag{17}$$

Example 3: What is the RMT question for GLMs?

- If we can indeed apply a version of Kac-Rice, what is the natural RMT question?
- Answer: Spiked Generalized Wishart matrices, i.e. rank-one perturbations of general Wishart (or Pastur-Marchenko) matrices

$$H = XDX^t \tag{17}$$

- where D is diagonal, and X is an NxM Random Gaussian Matrix, with i.i.d entries

Example 3: What is the RMT question for GLMs?

Example 3: What is the RMT question for GLMs?

- The spectrum of these matrices is well understood when the empirical measure of the entries of the diagonal D converge, say to a measure ν.
- It is linked to free probability: it converges to the free *multiplicative* convolution of the asymptotic spectral measure ν and the Marchenko-Pastur distribution (with ratio α).
- Spiking this matrix H by a rank one perturbation is also understood, in fact the BBP transition started in 2005 with the simplest version of this type of example, not with the spiked GOE.
- What is still missing, but not for long, is an LDP for the spiked eigenvalue. This will allow the understanding of the complexity of local minima.

Example 3: Results about annealed complexity

Example 3: Results about annealed complexity

- We compute, for both functions L_1 and L_2, the limiting annealed complexity in terms of a complicated variational principle in the space of probability measures on the real line.

Example 3: Results about annealed complexity

- We compute, for both functions L_1 and L_2, the limiting annealed complexity in terms of a complicated variational principle in the space of probability measures on the real line.
- Let B be a subset of the real line, and, as above, $Crit_{N,L_1}(B)$ the number of critical points of the function L_i with value in B

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_{N,L_1}(B)] = \sup_{\nu, \int \phi(t) d\nu(t) \in B} T_1(\nu) - \alpha H(\nu) \tag{18}$$

Example 3: Results about annealed complexity

- We compute, for both functions L_1 and L_2, the limiting annealed complexity in terms of a complicated variational principle in the space of probability measures on the real line.
- Let B be a subset of the real line, and, as above, $Crit_{N,L_1}(B)$ the number of critical points of the function L_i with value in B

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_{N,L_1}(B)] = \sup_{\nu, \int \phi(t) d\nu(t) \in B} T_1(\nu) - \alpha H(\nu) \tag{18}$$

- Here $H(\nu)$ is the relative entropy of ν w.r.t. the standard Gaussian measure $N(0,1)$

Example 3: Results about annealed complexity

Example 3: Results about annealed complexity

- T_1 is a rather involved functional on the space of probability measures on the real line

$$T_1(\nu) = \frac{1 + \log \alpha}{2} - \frac{1}{2} \log[\int \phi'(t)^2 d\nu(t)] + K(\nu) \tag{19}$$

Example 3: Results about annealed complexity

- T_1 is a rather involved functional on the space of probability measures on the real line

$$T_1(\nu) = \frac{1 + \log \alpha}{2} - \frac{1}{2} \log[\int \phi'(t)^2 d\nu(t)] + K(\nu) \qquad (19)$$

- $K(\nu)$ is defined as a logarithmic potential

Example 3: Results about annealed complexity

- T_1 is a rather involved functional on the space of probability measures on the real line

$$T_1(\nu) = \frac{1 + \log \alpha}{2} - \frac{1}{2} \log[\int \phi'(t)^2 d\nu(t)] + K(\nu) \qquad (19)$$

- $K(\nu)$ is defined as a logarithmic potential

$$K(\nu) = \int \log |x - \int t\phi'(t) d\nu(t)| d\mu_\nu(x) \qquad (20)$$

Example 3: Results about annealed complexity

- T_1 is a rather involved functional on the space of probability measures on the real line

$$T_1(\nu) = \frac{1 + \log \alpha}{2} - \frac{1}{2} \log[\int \phi'(t)^2 d\nu(t)] + K(\nu) \tag{19}$$

- $K(\nu)$ is defined as a logarithmic potential

$$K(\nu) = \int \log |x - \int t\phi'(t) d\nu(t)| d\mu_\nu(x) \tag{20}$$

- Here the measure μ_ν is defined as the free multiplicative convolution of the α Marchenko-Pastur distribution with the image of ν by the map ϕ''.

Example 3: Results about annealed complexity

Example 3: Results about annealed complexity

- One can check that this complexity is indeed zero for linear and quadratic activations! So these models are indeed not complex. It is known that optimization of L_1 is not hard in these cases.

Example 3: Results about annealed complexity

- One can check that this complexity is indeed zero for linear and quadratic activations! So these models are indeed not complex. It is known that optimization of L_1 is not hard in these cases.
- For what activations are these model complex? This question is now reduced to this deep variational problem.

Example 3: Results about annealed complexity

- One can check that this complexity is indeed zero for linear and quadratic activations! So these models are indeed not complex. It is known that optimization of L_1 is not hard in these cases.
- For what activations are these model complex? This question is now reduced to this deep variational problem.
- The paper also computes the annealed complexity of L_2 as an even richer variational problem, including the latitude (i.e. the correlation with the signal, as in the tensor PCA problem).

Example 3: Results about annealed complexity

- One can check that this complexity is indeed zero for linear and quadratic activations! So these models are indeed not complex. It is known that optimization of L_1 is not hard in these cases.
- For what activations are these model complex? This question is now reduced to this deep variational problem.
- The paper also computes the annealed complexity of L_2 as an even richer variational problem, including the latitude (i.e. the correlation with the signal, as in the tensor PCA problem).
- We also compute the quenched complexity L_1 and L_2 using the (non-rigorous) replicated Kac-Rice formula introduced in the work with Biroli, Ros and Cammarota on Tensor PCA. We will soon be able to compute the complexity of critical points with fixed index, and thus of local minima.

Example 3: Results about annealed complexity

- One can check that this complexity is indeed zero for linear and quadratic activations! So these models are indeed not complex. It is known that optimization of L_1 is not hard in these cases.
- For what activations are these model complex? This question is now reduced to this deep variational problem.
- The paper also computes the annealed complexity of L_2 as an even richer variational problem, including the latitude (i.e. the correlation with the signal, as in the tensor PCA problem).
- We also compute the quenched complexity L_1 and L_2 using the (non-rigorous) replicated Kac-Rice formula introduced in the work with Biroli, Ros and Cammarota on Tensor PCA. We will soon be able to compute the complexity of critical points with fixed index, and thus of local minima.
- What about a Computational-Statistical gap for cases for the landscape is complex? (ongoing work with Ghessairi-Jagannath)

THANKS!