The complexity of random functions of many variables

From Statistical physics to Machine learning

Gérard BEN AROUS

Courant Institute of Mathematical Sciences, New York University

Conference 59 Christian-Patrick, Toulouse

June 9, 2017

A short summary

- A smooth random function of many variables can be exponentially complex.

A short summary

- A smooth random function of many variables can be exponentially complex.
- The basic mathematical tool to study the complexity of random functions is a link with Random Matrix Theory (RMT), through the classical tools of random geometry, i.e Kac-Rice formulae (cf the books by Azais-Wschebor or Adler-Taylor).

A short summary

- This is well understood for random smooth functions on the sphere in high dimensions (aka spherical Spin Glasses), since the RMT model is a simple modification of the Gaussian Orthogonal Ensemble (GOE), i.e. NxN real symmetric random matrices, where the entries are i.i.d Gaussian . I will review recent progress for this case.

A short summary

- This is well understood for random smooth functions on the sphere in high dimensions (aka spherical Spin Glasses), since the RMT model is a simple modification of the Gaussian Orthogonal Ensemble (GOE), i.e. NxN real symmetric random matrices, where the entries are i.i.d Gaussian . I will review recent progress for this case.
- I will indicate how this complexity question for more general models of random functions, boils down, once translated into the RMT framework, to specific questions about Gaussian real symmetric random matrices, with correlated entries, and how recent progress in this direction (Dec 2016, by Erdos and al) can help.

A short summary

A short summary

- The general dictionary opens a new set of questions for the performance of optimization algorithms (and thus statistical algorithms) for very high dimensional models

A short summary

- The general dictionary opens a new set of questions for the performance of optimization algorithms (and thus statistical algorithms) for very high dimensional models
- In particular, there is a wide ranging phase transition in topology/complexity, illustrated here for the problem of Spiked Tensor PCA

A short summary

- The general dictionary opens a new set of questions for the performance of optimization algorithms (and thus statistical algorithms) for very high dimensional models
- In particular, there is a wide ranging phase transition in topology/complexity, illustrated here for the problem of Spiked Tensor PCA
- This topological transition is the case for instance for questions of statistical estimation in very high dimensional statistics (aka Big Data), and in particular, for the loss landscapes of the multilayered networks of deep learning.

A SIMPLE QUESTION TO BEGIN: MINIMIZING CUBICS

A simple question to begin: minimizing cubics

- Consider a homogeneous polynomial f of degree 3 in N variables.
- Since it is homogeneous, restrict it to the unit sphere S^{N-1}.

A simple question to begin: minimizing cubics

- Consider a homogeneous polynomial f of degree 3 in N variables.
- Since it is homogeneous, restrict it to the unit sphere S^{N-1}.
- Question: How easy is it to find its minimum on S^{N-1}?

A simple question to begin: minimizing cubics

- Consider a homogeneous polynomial f of degree 3 in N variables.
- Since it is homogeneous, restrict it to the unit sphere S^{N-1}.
- Question: How easy is it to find its minimum on S^{N-1}?
- Of course the answer depends on the polynomial!
- For instance, if $f(x) = x_1^3$, the problem is trivial.

A simple question: minimizing cubics

- What if f is less special and more "generic"? say f is chosen randomly?

A simple question: minimizing cubics

- What if f is less special and more "generic"? say f is chosen randomly?
- Choose

$$f(x) = \sum_{i,j,k=1}^{N} J_{i,j,k} x_i x_j x_k$$

where the coefficients J are i.i.d $N(0,1)$

A simple question: minimizing cubics

- What if f is less special and more "generic"? say f is chosen randomly?
- Choose
$$f(x) = \sum_{i,j,k=1}^{N} J_{i,j,k} x_i x_j x_k$$
where the coefficients J are i.i.d $N(0,1)$
- What is the minimum value of f on S^{N-1}?

A simple question: minimizing cubics

- What if f is less special and more "generic"? say f is chosen randomly?
- Choose
$$f(x) = \sum_{i,j,k=1}^{N} J_{i,j,k} x_i x_j x_k$$
where the coefficients J are i.i.d $N(0,1)$
- What is the minimum value of f on S^{N-1}?
- Is it easy to minimize f through your preferred algorithm (say a gradient descent, a stochastic gradient descent, a Langevin dynamics)?

A simple question: minimizing cubics

- What if f is less special and more "generic"? say f is chosen randomly?
- Choose

$$f(x) = \sum_{i,j,k=1}^{N} J_{i,j,k} x_i x_j x_k$$

  where the coefficients J are i.i.d $N(0,1)$
- What is the minimum value of f on S^{N-1}?
- Is it easy to minimize f through your preferred algorithm (say a gradient descent, a stochastic gradient descent, a Langevin dynamics)?
- Will the algorithm get to (or near to)the minimum or stay stuck above it (in finite time)?
- If it gets stuck, where?

A simple question: minimizing cubics. Some answers

A simple question: minimizing cubics. Some answers

- The minimum m_N is of order $\sqrt{N}$.

A simple question: minimizing cubics. Some answers

- The minimum m_N is of order $\sqrt{N}$.
- More precisely
$$\lim_{N\to\infty} \frac{m_N}{\sqrt{N}} = -E_0$$
with $E_0 \sim 1.657$.

A simple question: minimizing cubics. Some answers

- The minimum m_N is of order $\sqrt{N}$.
- More precisely
$$\lim_{N\to\infty} \frac{m_N}{\sqrt{N}} = -E_0$$
  with $E_0 \sim 1.657$.
- BUT : a minimization algorithm will probably get stuck (in finite time) at the threshold $-E_\infty\sqrt{N}$, with $E_\infty \sim 1.633$, or rather slightly above it.

A simple question: minimizing cubics. Some answers

- The minimum m_N is of order $\sqrt{N}$.
- More precisely
$$\lim_{N\to\infty} \frac{m_N}{\sqrt{N}} = -E_0$$
with $E_0 \sim 1.657$.
- BUT : a minimization algorithm will probably get stuck (in finite time) at the threshold $-E_\infty\sqrt{N}$, with $E_\infty \sim 1.633$, or rather slightly above it.
- Don't trust me, try it! The Facebook research team around Yann Le Cun did with a stochastic gradient descent algorithm. (see in our AISTATS 2015 paper)

A simple question: minimizing cubics

- In order to understand the question above, we need geometric information.

A simple question: minimizing cubics

- In order to understand the question above, we need geometric information.
- How does the (random) function looks like near its low values?

A simple question: minimizing cubics

- In order to understand the question above, we need geometric information.
- How does the (random) function looks like near its low values?
- Many minima? Multiple wells? separated by high barriers?

A simple question: minimizing cubics

- In order to understand the question above, we need geometric information.
- How does the (random) function looks like near its low values?
- Many minima? Multiple wells? separated by high barriers?
- What can we say about the topology of the sub-level set $A_u := \{x \in M, f(x) \leq u\}$?

A simple question: minimizing cubics. What is going on?

A simple question: minimizing cubics. What is going on?

- A (random) cubic is in fact a very complex function!!

A simple question: minimizing cubics. What is going on?

- A (random) cubic is in fact a very complex function!!
- The number of local minima is exponentially large in N.

A simple question: minimizing cubics. What is going on?

- A (random) cubic is in fact a very complex function!!
- The number of local minima is exponentially large in N.
- All local minima, and in fact all critical points of finite index, have values between $-E_0\sqrt{N}$ and $-E_\infty\sqrt{N}$

A simple question: minimizing cubics. What is going on?

- A (random) cubic is in fact a very complex function!!
- The number of local minima is exponentially large in N.
- All local minima, and in fact all critical points of finite index, have values between $-E_0\sqrt{N}$ and $-E_\infty\sqrt{N}$
- Among the critical points with values in any interval $(a\sqrt{N}, b\sqrt{N})$ in this range, the local minima dominate exponentially!

A simple question: minimizing cubics. What is going on?

A simple question: minimizing cubics. What is going on?

- Above the threshold level $-E_{\infty}\sqrt{N}$ there are only critical points of diverging index $k = cN$, and thus no local minima (good news)

A simple question: minimizing cubics. What is going on?

- Above the threshold level $-E_\infty\sqrt{N}$ there are only critical points of diverging index $k = cN$, and thus no local minima (good news)
- But below that threshold level, there are "extensive barriers" to cross, of order $c\sqrt{N}$ (bad news).

A simple question: minimizing cubics. What is going on?

- Above the threshold level $-E_\infty\sqrt{N}$ there are only critical points of diverging index $k = cN$, and thus no local minima (good news)
- But below that threshold level, there are "extensive barriers" to cross, of order $c\sqrt{N}$ (bad news).
- Below that threshold, the Euler characteristic of the sub-level set A_u is exponentially large in N. This is compatible with the image of a union of an exponentially large number of small caps around the local minima.

A simple question: minimizing cubics. What is going on?

- Above the threshold level $-E_\infty\sqrt{N}$ there are only critical points of diverging index $k = cN$, and thus no local minima (good news)
- But below that threshold level, there are "extensive barriers" to cross, of order $c\sqrt{N}$ (bad news).
- Below that threshold, the Euler characteristic of the sub-level set A_u is exponentially large in N. This is compatible with the image of a union of an exponentially large number of small caps around the local minima.
- Above that threshold, the Euler characteristic of the sub-level set A_u oscillates wildly between $e^{c(u)N}$ and $-e^{c(u)N}$!!

Just in case you wonder: what about other degrees?

- Complexity only begins with cubics: in fact this theorem is valid for any degree $p \geq 3$.

Just in case you wonder: what about other degrees?

- Complexity only begins with cubics: in fact this theorem is valid for any degree $p \geq 3$.
- But obviously not in degree $p \leq 2$.

Just in case you wonder: what about other degrees?

- Complexity only begins with cubics: in fact this theorem is valid for any degree $p \geq 3$.
- But obviously not in degree $p \leq 2$.
- An analogous result holds for non homogeneous polynomials, but the results are more intricate.

THE LINK WITH RANDOM MATRIX THEORY

The Kac-Rice formula

The Kac-Rice formula

- The classical Kac-Rice formula (for the first moment) counts the mean number of critical points of a random Gaussian function f on a (compact) manifold M.
- Define $Crit^{f}_{N,k}(B)$ to be 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.

The Kac-Rice formula

- The classical Kac-Rice formula (for the first moment) counts the mean number of critical points of a random Gaussian function f on a (compact) manifold M.
- Define $Crit^{f}_{N,k}(B)$ to be 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.
- The version of the Kac-Rice formula we will need reads

$$E[Crit^{f}_{N,k}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \qquad (1)$$

The Kac-Rice formula

- The classical Kac-Rice formula (for the first moment) counts the mean number of critical points of a random Gaussian function f on a (compact) manifold M.
- Define $Crit^{f}_{N,k}(B)$ to be 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.
- The version of the Kac-Rice formula we will need reads

$$E[Crit^{f}_{N,k}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \tag{1}$$

- where

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

The Kac-Rice formula

- The classical Kac-Rice formula (for the first moment) counts the mean number of critical points of a random Gaussian function f on a (compact) manifold M.
- Define $Crit^f_{N,k}(B)$ to be 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.
- The version of the Kac-Rice formula we will need reads

$$E[Crit^f_{N,k}(B)] = \int_B \int_M a_k(x,u)\phi_x(u,0)dxdu \tag{1}$$

- where

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

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

The Kac-Rice formula

The Kac-Rice formula

- The classical Kac-Rice formula can also count higher (factorial) moments of the number of critical points of a random Gaussian function f on a (compact) manifold M

The Kac-Rice formula

- The classical Kac-Rice formula can also count higher (factorial) moments of the number of critical points of a random Gaussian function f on a (compact) manifold M
- It can also compute the moments of the Euler characteristic of sub-level sets.
- Denote the sub-level set below u by

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

The Kac-Rice formula

- The classical Kac-Rice formula can also count higher (factorial) moments of the number of critical points of a random Gaussian function f on a (compact) manifold M
- It can also compute the moments of the Euler characteristic of sub-level sets.
- Denote the sub-level set below u by

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

- Then the KR formula for the Euler characteristic reads

$$E[\chi(A(u)] = \int_{-\infty}^{u} \int_{M} b(x, v)\phi_x(v, 0)dxdv \tag{4}$$

The Kac-Rice formula

- The classical Kac-Rice formula can also count higher (factorial) moments of the number of critical points of a random Gaussian function f on a (compact) manifold M
- It can also compute the moments of the Euler characteristic of sub-level sets.
- Denote the sub-level set below u by

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

- Then the KR formula for the Euler characteristic reads

$$E[\chi(A(u)] = \int_{-\infty}^{u} \int_{M} b(x, v)\phi_x(v, 0)dxdv \tag{4}$$

- where

$$b(x, v) = E\left[\det \nabla^2 f(x) \middle| f(x) = v, \nabla f(x) = 0\right] \tag{5}$$

The link with Random Matrix Theory

The link with Random Matrix Theory

- The link with RMT is thus that this formula reduces the study of the moments of the number of critical points or of the Euler characteristics of the sub-levels sets, 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 link with RMT is thus that this formula reduces the study of the moments of the number of critical points or of the Euler characteristics of the sub-levels sets, 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 link with RMT is thus that this formula reduces the study of the moments of the number of critical points or of the Euler characteristics of the sub-levels sets, 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).

$$C(x, y) = E[f(x)f(y)] - E[f(x)]E[f(y)] \tag{6}$$

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
- In fact, in order to understand the higher moments of the number of critical points, and thus study its possible concentration and show that the first moment gives the correct behavior, one needs to understand a field of correlated Gaussian random matrices, given by the Hessians of f at the critical points.
- This class of random matrix models is hard in general. The first broad result of convergence dates back to Dec 6 2016(Erdos and al, Arxiv).

The link with Random Matrix Theory

The link with Random Matrix Theory

- But our spherical case corresponds to one of the most studied and simplest class of random matrices, i.e the GOE (Gaussian Orthogonal Ensemble)!

The link with Random Matrix Theory

- But our spherical case corresponds to one of the most studied and simplest class of random matrices, i.e the GOE (Gaussian Orthogonal Ensemble)!
- The GOE is the distribution of a symmetric real random Gaussian matrix M where the entries are independent and Gaussian centered, with variances

$$E[M_{i,j}^2] = \frac{1 + \delta_{i,j}}{2N} \tag{7}$$

The link with Random Matrix Theory

- But our spherical case corresponds to one of the most studied and simplest class of random matrices, i.e the GOE (Gaussian Orthogonal Ensemble)!
- The GOE is the distribution of a symmetric real random Gaussian matrix M where the entries are independent and Gaussian centered, with variances

$$E[M_{i,j}^2] = \frac{1+\delta_{i,j}}{2N} \tag{7}$$

The link with Random Matrix Theory

- A trivial computation shows that for a random homogeneous polynomial of degree p, the covariance is given by

$$C(x,y) = E[f(x)f(y)] = (x.y)^p \tag{8}$$

The link with Random Matrix Theory

- A trivial computation shows that for a random homogeneous polynomial of degree p, the covariance is given by

$$C(x,y) = E[f(x)f(y)] = (x.y)^p \tag{8}$$

- By differentiation of C one sees that the distribution of the Hessian at x conditioned by the the fact x is a critical point and by the value $f(x) = u$ is a shifted GOE; $cuId - M$, where the shift is a function of the critical value $c = c(p)$ and where M is a GOE matrix.

The link with Random Matrix Theory

- Using the Kac-Rice formula above, we must compute

The link with Random Matrix Theory

- Using the Kac-Rice formula above, we must compute

$$E[|det(M - XId)|1_{i(M-XId)=k}1_{X\in B}] \tag{9}$$

- where M is a GOE (N-1) matrix, and X is an independent Gaussian variable

The link with Random Matrix Theory

- Using the Kac-Rice formula above, we must compute

$$E[|det(M - XId)|1_{i(M-XId)=k}1_{X\in B}] \tag{9}$$

- where M is a GOE (N-1) matrix, and X is an independent Gaussian variable
- or, for the Euler characteristic,

$$E[det(M - XId)1_{X\leq u}] \tag{10}$$

The link with Random Matrix Theory

- Theorem (Auffinger-Ben Arous-Cerny)

  If B is a subset of the real line,

$$E[Crit_k(p, \sqrt{N}B)] = c_p E_{GOE(N)}[e^{-N\frac{p-2}{2p}\lambda_k^2} 1_{\lambda_k \in c_p' B}]$$

  where $c_p = 2\sqrt{(}2/p)(p-1)^{N/2}$ *and* $c_p' = \frac{p}{2(p-1)}^{1/2}$ *and*

$$\lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{N-1}$$

  are the eigenvalues of a GOE matrix of size N.

A natural tool : Large Deviations for Random Matrices

M is a GOE_N matrix, if it is a real symmetric NxN matrix, whose entries are i.i.d centered Gaussian (above the diagonal) with

$$E(M_{i,j}^2) = \frac{(1+\delta_{i,j})}{2N}$$

and $\lambda_0 \leq \lambda_1 \cdots \leq \lambda_{N-1}$ its eigenvalues.
It is well known that the empirical spectral measure $\mu_N = \frac{1}{N}\sum \delta_{\lambda_i}$ converges to the semi-circle distribution (with radius $\sqrt{2}$).
Moreover a LDP is proved with rate N^2 (BA-Guionnet)

$$P[\mu_N \in A] \sim e^{-N^2 \inf_{\mu \in A} I(\mu)} \tag{11}$$

From this we deduce a LDP for the k-th eigenvalue λ_k, with rate N^2 if it is in the bulk: $k \sim cN$

$$P[\lambda_k \in A] \sim e^{-N^2 \inf_{\mu \in A} I_c(\mu)} \tag{12}$$

Large Deviations for Random Matrices

It is also well known that the smallest eigenvalue λ_0 converges to the left edge $-\sqrt{2}$ of the semi-circle. Moreover a LDP is proved in the scale N (BA-Dembo-Guionnet) for deviations to the left

$$P[\lambda_0 \leq -\sqrt{2} - u] \sim e^{-NJ(u)} \tag{13}$$

and in the scale N^2 for deviations to the right

$$P[\lambda_0 \geq -\sqrt{2} + u] \sim e^{-N^2 J'(u)} \tag{14}$$

Similarly LDP are proven for eigenvalues λ_k, with k fixed as N tends ∞.

Complexity of random homogeneous polynomials

- Theorem (Auffinger-Ben Arous-Cerny)

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_k(-\infty, -Nu)] = \theta_{k,p}(u)$$

Here the (annealed) complexity function $\theta_k(u)$ is continuous, non-decreasing and explicit.

Complexity of random homogeneous polynomials

- Theorem (Auffinger-Ben Arous-Cerny)

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_k(-\infty, -Nu)] = \theta_{k,p}(u)$$

Here the (annealed) complexity function $\theta_k(u)$ is continuous, non-decreasing and explicit.

- *For $u \leq -E_\infty$*

$$\theta_{k,p}(u) = 1/2 \log(p-1) - \frac{p-2}{4(p-1)}u^2 - (k+1)I(\frac{u}{E_\infty})$$

with $E_\infty = 2\sqrt{\frac{p-1}{p}}$ and $I(v) = 2\int_v^1 \sqrt{v^2-1}dv$

Complexity of random homogeneous polynomials

- Theorem (Auffinger-Ben Arous-Cerny)

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_k(-\infty, -Nu)] = \theta_{k,p}(u)$$

Here the (annealed) complexity function $\theta_k(u)$ is continuous, non-decreasing and explicit.

- *For $u \leq -E_\infty$*

$$\theta_{k,p}(u) = 1/2 \log(p-1) - \frac{p-2}{4(p-1)} u^2 - (k+1) I(\frac{u}{E_\infty})$$

with $E_\infty = 2\sqrt{\frac{p-1}{p}}$ and $I(v) = 2\int_v^1 \sqrt{v^2-1} dv$

- *For $u \geq -E_\infty$*

$$\theta_{k,p}(u) = 1/2 \log(p-1) - \frac{p-2}{p}$$

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_\infty)$, for any k

$$\theta_0(u) > \theta_1(u) > \cdots > \theta_k(u)$$

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_\infty)$, for any k

$$\theta_0(u) > \theta_1(u) > \cdots > \theta_k(u)$$

- So that on this interval, the local minima dominate exponentially

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_\infty)$, for any k

$$\theta_0(u) > \theta_1(u) > \cdots > \theta_k(u)$$

- So that on this interval, the local minima dominate exponentially
- On the interval $(-E_\infty, \infty)$, the functions θ_k are all constant

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_\infty)$, for any k

$$\theta_0(u) > \theta_1(u) > \cdots > \theta_k(u)$$

- So that on this interval, the local minima dominate exponentially
- On the interval $(-E_\infty, \infty)$, the functions θ_k are all constant
- So (whp) there are no critical points of finite index with values above $-E_\infty N$.

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_\infty)$, for any k

$$\theta_0(u) > \theta_1(u) > \cdots > \theta_k(u)$$

- So that on this interval, the local minima dominate exponentially
- On the interval $(-E_\infty, \infty)$, the functions θ_k are all constant
- So (whp) there are no critical points of finite index with values above $-E_\infty N$.
- In fact, the probability to find one such critical point is of order e^{-cN^2}

Complexity of random homogeneous polynomials

- Define E_k by $\theta_k(E_k) = 0$.

Complexity of random homogeneous polynomials

- Define E_k by $\theta_k(E_k) = 0$.
- E_k is the threshold of positive complexity of critical points of index k, i.e the level above which the average number is exponentially large.
- Obviously the sequence E_k is strictly increasing. It converges to E_∞.

Complexity of random homogeneous polynomials

- Define E_k by $\theta_k(E_k) = 0$.
- E_k is the threshold of positive complexity of critical points of index k, i.e the level above which the average number is exponentially large.
- Obviously the sequence E_k is strictly increasing. It converges to E_∞.
- On the interval $(-E_0 N, -E_\infty N)$, the average number of local minima is exponentially large.

Complexity of random homogeneous polynomials

- Define E_k by $\theta_k(E_k) = 0$.
- E_k is the threshold of positive complexity of critical points of index k, i.e the level above which the average number is exponentially large.
- Obviously the sequence E_k is strictly increasing. It converges to E_∞.
- On the interval $(-E_0 N, -E_\infty N)$, the average number of local minima is exponentially large.
- On the interval $(-E_0 N, -E_1 N)$, the average number of local minima is exponentially large and the number of critical points of positive index is exponentially small. This hints at the existence of high (extensive) barriers.

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_0 N)$ the average number of local minima is exponentially small, so that the minimum m_N is larger than $(-E_0 - \epsilon)\sqrt{N}$ with high probability.

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_0 N)$ the average number of local minima is exponentially small, so that the minimum m_N is larger than $(-E_0 - \epsilon)\sqrt{N}$ with high probability.
- Getting an upper bound on m_N, and proving that $m_N \sim -\sqrt{N} E_0$ is much more delicate.

Complexity of random homogeneous polynomials

- On the interval $(-\infty, -E_0 N)$ the average number of local minima is exponentially small, so that the minimum m_N is larger than $(-E_0 - \epsilon)\sqrt{N}$ with high probability.
- Getting an upper bound on m_N, and proving that $m_N \sim -\sqrt{N} E_0$ is much more delicate.
- It can be done as in the original paper using the Parisi formula for the free energy of spherical spin glasses (and its one step replica symmetry breaking (1RSB) at low temperature for pure p-spins), or with probabilistic tools, as in the recent work of E. Subag, by **controlling the second moment sharply.**

Pure p-spins: SUBAG's results

Pure p-spins: SUBAG's results

- The results above are about the mean number of critical points, It is a first moment method.

Pure p-spins: SUBAG's results

- The results above are about the mean number of critical points, It is a first moment method.
- They have been improved through a second moment analysis, valid for low values of the energy (Subag 2015). Through a study of the second moment Kac-Rice formula, the complexity questions boils down to question about a pair of correlated GOE matrices.

Pure p-spins: SUBAG's results

- The results above are about the mean number of critical points, It is a first moment method.
- They have been improved through a second moment analysis, valid for low values of the energy (Subag 2015). Through a study of the second moment Kac-Rice formula, the complexity questions boils down to question about a pair of correlated GOE matrices.

- Theorem (Subag 2015)

  For $u < -E_\infty$

$$\lim_{N\to\infty} \frac{1}{N} \log E[Crit_0(-\infty, Nu)^2] = 2\theta_{0,p}(u) \tag{15}$$

$$\lim_{N\to\infty} \frac{Crit_0(-\infty, Nu)}{E[Crit_0(-\infty, Nu)]} = 1 \tag{16}$$

Pure p-spins: SUBAG's results

- This approach gives a full asymptotic description of the extremal process (Subag-Zeitouni 2016) as a Poisson-Gumbel process

Pure p-spins: SUBAG's results

- This approach gives a full asymptotic description of the extremal process (Subag-Zeitouni 2016) as a Poisson-Gumbel process
- And even a very detailed description of the Gibbs measure at very low temperature, which exhibits the TAP states (Subag 2016) These TAP states are centered on the "deepest" wells, their bottoms are near the ground state.

Pure p-spins: SUBAG's results

- This approach gives a full asymptotic description of the extremal process (Subag-Zeitouni 2016) as a Poisson-Gumbel process
- And even a very detailed description of the Gibbs measure at very low temperature, which exhibits the TAP states (Subag 2016) These TAP states are centered on the "deepest" wells, their bottoms are near the ground state.
- This analysis at very low temperature shows the **absence of "temperature chaos"**

Pure p-spins: Open questions

- The second moment analysis is valid only below the threshold $-E_\infty$ for the total number of critical points or for local minima, since they dominate there. What about other saddle points in this region?
- The TAP like description is valid only for very low temperatures (as long as a harmonic approximation is valid). It should be for higher temperatures (below the static transition). And there should be no temperature chaos there.
- What about the phase above this where the Gibbs measure is carried by an exponential number of wells centered on deep wells centered on local minima above the ground state. This phase is RS but complex.

Pure p-spins: Dynamics questions

- What do these result mean for Langevin dynamics?
- In finite time scales, the situation is well understood (Cugliandolo-Kurchan, Dembo-Guionnet-GBA) The dynamics can only reach the dynamic threshold $-E_\infty$.
- What about longer time scales? Mixing time? Metastability? Aging?

Pure p-spins: Dynamics questions

- Recent result: Jagannath-BA, 2017, valid for hard spins and spherical cases

Theorem

The mixing time is exponential is N as soon as the so-called Franz-Parisi potential has a "secondary minimum", and for instance as soon as the support of the Parisi measure is not connected. So in particular for this 1 RSB situation.

Pure p-spins: Dynamics questions

- This result is based on a Cheeger type estimate applied to a pair of replicated dynamics, using a new large deviation principle for the distribution of the overlap of two replica, in tunr based on the recent and veyr deep work of Panchenko on the so-called 2d Guerra-Talagrand bounds.

Pure p-spins: Dynamics questions

- This result is based on a Cheeger type estimate applied to a pair of replicated dynamics, using a new large deviation principle for the distribution of the overlap of two replica, in tunr based on the recent and veyr deep work of Panchenko on the so-called 2d Guerra-Talagrand bounds.
- Aging: the classical Fractional kinetics type aging (a la Bouchaud) should be valid for Langevin dynamics in exponential time scales. (proven for Hard p-spins for the simpler Random Hopping TIme (RHT) or Bouchaud dynamics (Bovier-Cerny-BA, 2009), and for Metropolis dynamics for REM (Cerny-Wassmer, Gayrard 2016).

An important phase transition; the topological transition

- What if the Gaussian process has a non zero mean $m_N(x)$, which is non complex, i.e has only a few critical points

An important phase transition; the topological transition

- What if the Gaussian process has a non zero mean $m_N(x)$, which is non complex, i.e has only a few critical points
- We study this case on the sphere with G. Biroli and C. Cammarota by adding a mean to the Gaussian process

$$m(x) = CN(\frac{\sum x_i n_i}{N})^k \tag{17}$$

where n is a fixed point on the sphere, say $n = e_1$, to a random homogeneous polynomial of degree p (a p-spin spherical spin glass)

An important phase transition; the topological transition

- What if the Gaussian process has a non zero mean $m_N(x)$, which is non complex, i.e has only a few critical points
- We study this case on the sphere with G. Biroli and C. Cammarota by adding a mean to the Gaussian process

$$m(x) = CN(\frac{\sum x_i n_i}{N})^k \tag{17}$$

  where n is a fixed point on the sphere, say $n = e_1$, to a random homogeneous polynomial of degree p (a p-spin spherical spin glass)
- Then there is an important phase transition depending on the strength of the signal-to-noise ratio C, with different behaviors if $k = 1$, $k = 2$ and $k > 2$

A denoising problem: Tensor PCA

- You observe a p-tensor in N variables $X = Cv^{\otimes p} + Z$
- Here v is a fixed unknown vector on the unit sphere S_N, and Z is a random centered p-tensor.

A denoising problem: Tensor PCA

- You observe a p-tensor in N variables $X = Cv^{\otimes p} + Z$
- Here v is a fixed unknown vector on the unit sphere S_N, and Z is a random centered p-tensor.
- We assume that the noise Z is Gaussian, and that its entries are i.i.d.

A denoising problem: Tensor PCA

- You observe a p-tensor in N variables $X = Cv^{\otimes p} + Z$
- Here v is a fixed unknown vector on the unit sphere S_N, and Z is a random centered p-tensor.
- We assume that the noise Z is Gaussian, and that its entries are i.i.d.
- The objective is to reconstruct (estimate) v. This is done by the maximum likelihood estimator, v_{ML} obtained by solving the following optimization problem
- Find the maximum of $< X, u^{\otimes p} >$ for $u \in S_N$

A denoising problem: Tensor PCA

- You observe a p-tensor in N variables $X = Cv^{\otimes p} + Z$
- Here v is a fixed unknown vector on the unit sphere S_N, and Z is a random centered p-tensor.
- We assume that the noise Z is Gaussian, and that its entries are i.i.d.
- The objective is to reconstruct (estimate) v. This is done by the maximum likelihood estimator, v_{ML} obtained by solving the following optimization problem
- Find the maximum of $< X, u^{\otimes p} >$ for $u \in S_N$
- When $p \leq 3$, this optimization problem is NP hard (Hillar and Lim 2013)

A denoising problem: Tensor PCA

- Of course when $p = 2$ this is the usual spiked Principal Component Aanalysis model.

A denoising problem: Tensor PCA

- Of course when $p = 2$ this is the usual spiked Principal Component Aanalysis model.
- It is well known that there is an important transition depending on the value of C,for various models of noise.

A denoising problem: Tensor PCA

- Of course when $p = 2$ this is the usual spiked Principal Component Aanalysis model.
- It is well known that there is an important transition depending on the value of C,for various models of noise.
- This BBP transition is the basis of many practical "shrinking" algorithms (Johnstone, Donoho ...)

A denoising problem: Tensor PCA

- Of course when $p = 2$ this is the usual spiked Principal Component Aanalysis model.
- It is well known that there is an important transition depending on the value of C,for various models of noise.
- This BBP transition is the basis of many practical ”shrinking” algorithms (Johnstone, Donoho ...)
- We want to explore if something similar happens for tensors (recent works by A. Montanari et al, Ge et al, A. Bandeira et al...)

A denoising problem: Tensor PCA

- Of course when $p = 2$ this is the usual spiked Principal Component Aanalysis model.
- It is well known that there is an important transition depending on the value of C,for various models of noise.
- This BBP transition is the basis of many practical "shrinking" algorithms (Johnstone, Donoho ...)
- We want to explore if something similar happens for tensors (recent works by A. Montanari et al, Ge et al, A. Bandeira et al...)
- But this is exactly our problem above, if one restricts to the case $k = p$!

The topological/complexity phase transition (when $k > 2$)

The topological/complexity phase transition (when $k > 2$)

- When $C < C_1$ is small, the model is exponentially complex and the critical points are on the equator

The topological/complexity phase transition (when $k > 2$)

- When $C < C_1$ is small, the model is exponentially complex and the critical points are on the equator
- When $C_1 < C < C_2$, then the model is still exponentially complex but there are exponentially many local minima in a band inside the north hemisphere

The topological/complexity phase transition (when $k > 2$)

- When $C < C_1$ is small, the model is exponentially complex and the critical points are on the equator
- When $C_1 < C < C_2$, then the model is still exponentially complex but there are exponentially many local minima in a band inside the north hemisphere
- When $C > C_2$, the model is no longer complex and the (now unique) local minimum gets closer the north pole
- Joint works with C. Cammarota and G. Biroli (using the physics replica method), and exact computation of complexity with M.Nica (based on the BBP transition and recent work by M. Maida).

The topological/complexity transition (when $k > 2$)

- What are the dynamical consequences? For a gradient descent algorithm? For Langevin dynamics? For a stochastic gradient descent?
- Joint work with C. Cammarota and G. Biroli, and numerical work in progress with L.Sagun and M. Baity-Jesy.

The topological/complexity transition (when $k > 2$)

- What are the dynamical consequences? For a gradient descent algorithm? For Langevin dynamics? For a stochastic gradient descent?
- Joint work with C. Cammarota and G. Biroli, and numerical work in progress with L.Sagun and M. Baity-Jesy.
- In the third (non complex) phase, gradient descent finds the minimum in finite time. The signal is recovered almost fully.

The topological/complexity transition (when $k > 2$)

- What are the dynamical consequences? For a gradient descent algorithm? For Langevin dynamics? For a stochastic gradient descent?
- Joint work with C. Cammarota and G. Biroli, and numerical work in progress with L.Sagun and M. Baity-Jesy.
- In the third (non complex) phase, gradient descent finds the minimum in finite time. The signal is recovered almost fully.
- In the first (complex) phase, gradient descent as well as Langevin dynamics cannot escape the equator, even in exponential time. Converges to a measure carried by the equator (i.e. the signal is fully lost).

The intermediate phase of topological/complexity transition

- In the intermediate phase, gradient descent cannot escape the equator, except if it is given a warm start. Langevin dynamics stops in one of the exponentially many local minima, in a band near the pole, but takes exponential time to get there.
- If the signal to noise ratio diverges as a power of N (should be $k - 2/4$), then one finds this band in finite time
- Instead if one has K samples, then the on-line stochastic gradient descent algorithm finds this band after the crossing of a small barrier (note the signal to noise ratio also diverges)

THE NATURAL QUESTIONS IN A MORE GENERAL CONTEXT

The natural questions

- Consider a random smooth function $f_N(x)$ on the manifold M_N of large dimension N.
- Assume f_N is Gaussian, and thus characterized by its mean function $m_N(x)$ and its covariance $C_N(x, y)$.

The natural questions

- Consider a random smooth function $f_N(x)$ on the manifold M_N of large dimension N.
- Assume f_N is Gaussian, and thus characterized by its mean function $m_N(x)$ and its covariance $C_N(x,y)$.
- Question 1: Compute the number of critical points? at (or below) a fixed value? with a fixed index?

$$Crit_k(B) = \#(x \in M_N, \nabla f_N(x) = 0, i(\nabla^2 f_N(x)) = k, f_N(x) \in B) \tag{18}$$

The natural questions

- Consider a random smooth function $f_N(x)$ on the manifold M_N of large dimension N.
- Assume f_N is Gaussian, and thus characterized by its mean function $m_N(x)$ and its covariance $C_N(x, y)$.
- Question 1: Compute the number of critical points? at (or below) a fixed value? with a fixed index?

$$Crit_k(B) = \#(x \in M_N, \nabla f_N(x) = 0, i(\nabla^2 f_N(x)) = k, f_N(x) \in B) \tag{18}$$

- Question 2: Understand the topology of this landscape, for instance the Euler characteristic of its sub-level sets?

$$\chi(u) = \chi(\{x \in M_N, f_N(x) \leq u\}) \tag{19}$$

The natural questions

- Question 3: How can this geometric information be used to understand the behavior of Gibbs measures at low temperature?

$$\mu_{\beta,N}(dx) = \frac{1}{Z_N(\beta)} e^{-\beta f_N(x)} dx \tag{20}$$

The natural questions

- Question 3: How can this geometric information be used to understand the behavior of Gibbs measures at low temperature?

$$\mu_{\beta,N}(dx) = \frac{1}{Z_N(\beta)} e^{-\beta f_N(x)} dx \tag{20}$$

- Question 4: How can it be used for dynamical properties? Either for natural Langevin dynamics, or for optimization algorithms used in statistics of big data, like stochastic gradient descent.

TWO IMPORTANT MODELS

The isotropic models of Gaussian Smooth Functions

- Assume that the manifold M_N is Riemannian, call d_N the Riemannian distance.
- Assume that random function f_N is centered $m_N(x) = 0$ and that the covariance is a function of the distance.

$$C(x, y) = cov(f(x), f(y)) = g(d_N(x, y)) \tag{21}$$

The isotropic models of Gaussian Smooth Functions

- Assume that the manifold M_N is Riemannian, call d_N the Riemannian distance.
- Assume that random function f_N is centered $m_N(x) = 0$ and that the covariance is a function of the distance.

$$C(x, y) = cov(f(x), f(y)) = g(d_N(x, y)) \tag{21}$$

- The variance is constant: $Var(f(x) = g(0)$ (and f_N is independent from its gradient)
- wlog we assume that $Var(f(x) = g(0) = 1$

The isotropic models of Gaussian Smooth Functions

- Assume that the manifold M_N is Riemannian, call d_N the Riemannian distance.
- Assume that random function f_N is centered $m_N(x) = 0$ and that the covariance is a function of the distance.

$$C(x, y) = cov(f(x), f(y)) = g(d_N(x, y)) \tag{21}$$

- The variance is constant: $Var(f(x) = g(0)$ (and f_N is independent from its gradient)
- wlog we assume that $Var(f(x) = g(0) = 1$
- The metric induced by the Gaussian process is topologically equivalent to the Riemannian metric.

$$E[(f(x) - f(y))^2] = 2(1 - g)(d_N(x, y)) \tag{22}$$

The isotropic models on the Sphere

- If the manifold M is the unit sphere S^{N-1}, the functions g such that $g(d_M(x,y))$ defines a covariance (i.e. is positive definite) have been characterized by Schoenberg in 1942.
- If we assume that g is independent of the dimension N, then there exists a sequence $a_p \geq 0$ such that

$$g(d) = \sum_{p=1}^{\infty} a_p (\cos d)^p \tag{23}$$

- Another way to write this is to introduce the function $\nu(r) = \sum_{p=1}^{\infty} a_p r^p$ and

$$C_N(x,y) = \nu(< x, y >) \tag{24}$$

The isotropic models on the Sphere

The isotropic models on the Sphere

- Define, for a sequence $a_p \geq 0$, the random function

$$f_{N,\nu}(x) = \sum_{p=2}^{\infty} \sqrt{a}_p f_{N,p}(x) = \sum_{p=2}^{\infty} \sqrt{a}_p \sum_{i_1,i_2,\ldots,i_p=1}^{N} J_{i_1,\ldots 1_p} x_{i_1} \ldots x_{i_p} \tag{25}$$

  where the J's are i.i.d Gaussian $N(0,1)$.

- It is easy to see that

$$cov(f_{N,\nu}(x), f_{N,\nu}(y)) = \sum a_p < x, y >^p = \nu(< x, y >) \tag{26}$$

- With this realization, it is easy to see that, if the a_p decay fast enough (say exponentially), then the random function $f_{N,\nu}(x)$ is a.s. smooth (and Morse)

The isotropic models on the Sphere

- The isotropic models on the sphere are exactly the mixed Spherical Spin Glass models.
- The general mixed Spherical Spin Glass hamiltonian is defined on $S^{N-1}(\sqrt{N})$, the sphere of radius $\sqrt{N}$ by

$$H_{N,\nu}(\sigma) = \sqrt{N} f_{N,\nu}(\frac{\sigma}{\sqrt{N}}) \tag{27}$$

Minimizing risk or loss in statistics

- One of the most important tasks in statistics is parameter estimation.
- In parametric models of statistics, one is given a family of probability measures P_θ on R^D, depending on the parameter $\theta \in R^N$
- The "true" parameter, say θ_0 is unknown, and one wants to estimate it using the data, i.e an M-sample of i.i.d random variables $X_1, ..., X_M$ in R^D

Minimizing risk or loss in statistics

- The most common path to parameter estimation is through minimizing of a risk/loss (or maximizing a likelihood)
- The true risk (or population risk) is unaccessible and is given by

$$R(\theta) = \int L(x, \theta)] dP(x) \tag{28}$$

- Here θ is a parameter in R^N, to be estimated, by minimizing $R(\theta)$
- x is the data, in R^d, whose distribution P is unknown.

Minimizing risk or loss in statistics

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

$$R_M(\theta) = \frac{1}{M} \sum_{i=1}^{M} L(x_i, \theta) \tag{29}$$

Minimizing risk or loss in statistics

- Thus the question becomes: can the minimization of $R_M(\theta)$ be a good substitute for the minimization of $R(\theta)$?
- The answer is yes, in classical parametric statistics where N is fixed and not large and M tends to ∞
- The answer is also yes, when both N and M are large and of the same order. This has been explored massively in the last ten years, see Candes-Tao for instance, and see recent work by Montanari and his coauthors (among many others).
- But what if the data is really very high dimensional, and N is much larger than M? Which is the case for the most important questions of machine learning usually solved through deep learning or deep convolutional networks.

Minimizing risk or loss in statistics

- $$R_M(\theta) \approx R(\theta) + \frac{1}{\sqrt{M}} G_M(\theta) \tag{30}$$

  Where G_M is a fluctuation term, which one may assume to be a Gaussian function of the parameter $\theta \in R^p$
- So the question is now to understand the critical points and the local/global minima of a Gaussian function of many variables (non centered).
- Typically one assumes that the mean (or the signal) $R(\theta)$ is not complex, and the question becomes: can the random term bring much larger complexity?
- The next step is then: can this complexity impede the minimization procedure?

The Loss landscape of multilayered networks

- Consider a multilayered network, with internal parameters $x = (J, \theta)$. Here the synaptic strength parameters of the network are denoted by J and the thresholds are denoted by θ. The dimension of the space of internal parameters will be noted N.
- Let *In* be the space of inputs, and *Out* the space of outputs.
- The network is fed with M random inputs i_k, sampled i.i.d from an unknown distribution ν on the input space *In*
- For any input $i \in In$, the network computes an output $O(i, x) = O(i, J, \theta)$, by linear maps using the parameters J (on the edges) and thresholding using the parameters θ (on the nodes).

The Loss landscape of multilayered networks

- Let T be the function being learned from a space *In* (in puts) to a space *Out* (outputs). i.e., if $i \in In$ is an input, then $T(i) \in Out$ is the "true value of the output".
- The Loss function is the total error made on the sample, for given parameters $x = (J, \theta)$

  $$Loss(x) = 1/M \sum_{k=1,..,M} Err(O(i_k, x), T(i_k)) \qquad (31)$$

  where *Err* is the error function used to rate the network.
- The goal is to minimize this random Loss function.

The Loss landscape of multilayered networks

- Its mean is $m_N(x) = \int Err(O(i,x), T(i)) d\nu(i)$
- Its covariance is

$$C_N(x,y) = K_N(x,y) - m_N(x)m_N(y) \tag{32}$$

where

$$K_N(x,y) = \int Err(O(i,x), T(i)) Err(O(i,y), T(i)) d\nu(i) \tag{33}$$

The Loss landscape of multilayered networks

- Its mean is $m_N(x) = \int Err(O(i,x), T(i))d\nu(i)$
- Its covariance is

$$C_N(x,y) = K_N(x,y) - m_N(x)m_N(y) \tag{32}$$

where

$$K_N(x,y) = \int Err(O(i,x), T(i))Err(O(i,y), T(i))d\nu(i) \tag{33}$$

- If the size M of the sample is large, it might be reasonable to assume that the Loss function is a Gaussian function of $x = (J, \theta)$

$$L_N(x) \asymp m_N(x) + \frac{1}{\sqrt{M}} G_N(x) \tag{34}$$

where $G_N(x)$ is a centered Gaussian function with covariance C_N on the space M_N of internal parameters.

The Loss landscape of multi-layered networks

- The question is thus: how complex is this Gaussian landscape, and how hard is it to minimize this function?
- Can we learn from the Spherical case?
- Can we prove the same type of results?
- What is the role of the "signal", here the unknown function m_N? and of the signal to noise ratio, i.e the relative size of the sample M (the length of the supervised learning phase) and the dimension N?

Loss surfaces

"The vast majority of practical applications of deep learning use supervised learning with very deep networks. The supervised loss function (usually a cross entropy) is minimized using some form of stochastic gradient descent (SGD). The general shape of the loss function is very poorly understood. Several researchers experimenting with larger networks and SGD had noticed that while multilayer nets do have many local minima, the results of multiple experiments consistently give very similar performance. This suggests that,while local minima are numerous, they are relatively easy to find, and they are more or less equivalent in terms of performance on the test set."

WHAT WOULD THAT PICTURE MEAN IN STATISTICS AND DATA SCIENCE

- If we could indeed understand well enough the random matrix model defined by a statistical model as above, and if we could understand the analog of the phase transition just described, then the following scenario would be possible
- When the learning phase is too short (the sample is too small), then the empirical risk optimization is exponentially complex, and the model learns only "noise"
- When the learning phase is longer (the sample is larger), then the optimization task is still exponentially complex, the result is NOT close to the absolute minimum of the population risk (i.e. the estimation is not consistent!), but still the performance is a positive fraction of the best possible.
- When the sample is longer the estimation becomes consistent.

References

1. Random matrices and complexity of spin glasses, Communications in Pure and Applied Mathematics 2013, with Antonio AUFFINGER (Northwestern), Jirí CERNÝ (Vienna),
2. Complexity of random smooth functions on the high-dimensional sphere, Annals of Probability 2013, with Antonio AUFFINGER (Northwestern)
3. The complexity of spherical p-spin models: a second moment approach, Arxiv 2015, Eliran SUBAG (Weizmann Institute),
4. The extremal process of critical points of the p-spin spherical spin glass model, Arxiv 2016, Eliran SUBAG and Ofer ZEITOUNI (Weizman)
5. The Geometry of the the Gibbs measure of pure Spherical spin glasses, Arxiv 2016, Eliran SUBAG

References

1. A simple model of deep learning, joint work with Giulio BIROLI (CEA, Saclay), Chiara CAMMAROTA (London)
2. joint ongoing work with Eliran SUBAG (Weizmann) and Ofer ZEITOUNI (Weizmann and Courant), in progress
3. Joint work with Paul Bourgade (CIMS)
4. How many equilibria will a large complex system have? joint work with Yan FYODOROV(King's College, London), Boris KHORUZHENKO (Queen Mary University, London), in progress

References (Machine Learning)

1. The loss surfaces of multilayer networks, AISTATS 2015, with Anna Choromanska, Mikael Henaff, Michael Mathieu, Yann Le Cun (Courant and Facebook)
2. The landscape of loss surfaces of multilayer networks, COLT 2015, with Anna Choromanska and Yann Le Cun (Courant and Facebook)
3. Explorations on high dimensional landscapes, ICLR 2015, with Levent Sagun (Courant), V.Ugur Guney (CUNY), Yann Le Cun (Courant and Facebook)