MATHEMATICS OF DEEP LEARNING

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

Lecture 11: Optimization Landscapes: spin glasses, tensors and deep networks

OBJECTIVES LECTURE 11

- Landscape of Spherical Spin Glasses

- Landscape of Tensor Decomposition and Spiked Tensor Model.

- Landscape of Neural Network Optimization

SPHERICAL SPIN GLASSES

- Consider the Hamiltonian of the Spherical Spin Glass of order p with N particles:

$$H_{N,p}(\sigma) = \frac{1}{N^{(p-1)/2}} \sum_{i_1,\dots,i_p=1}^{N} J_{i_1,\dots i_p}\sigma_{i_1}\dots\sigma_{i_p}\ ,$$

$$\sigma \in S^{N-1}(\sqrt{N})\ ,\ J_{i_1,\dots,i_p} \sim \mathcal{N}(0,1)\ .$$

- $H_{N,p}$ is the (only) Gaussian Process on the sphere whose covariance is given by

$$\mathbb{E}[H_{N,p}(\sigma)H_{N,p}(\sigma')] = N\left(\underbrace{N^{-1}\langle\sigma,\sigma'\rangle}_{\text{overlap}}\right)^p\ .$$

COMPLEXITY OF RANDOM HIGH-DIMENSIONAL FUNCTION

- In this case, the landscape is random. We want to understand notions of average-case complexity of this landscape.
- Consider the Random variable

$$\mathrm{Crt}_{N,k}(B) = \sum_{\sigma;\nabla H(\sigma)=0} \mathbf{1}\{H_{N,p}(\sigma) \in NB\}\mathbf{1}\{\mathrm{ind}[\nabla^2 H_{N,p}(\sigma)] = k\},$$

$$B \subset \mathbb{R} \qquad \mathrm{ind}[\nabla^2 H] = \text{ number of negative eigenvalues of } \nabla^2 H$$

  - it counts the number of configurations where $H_{N,p}$ has a critical point of a given index, at a given energy value.

COMPLEXITY OF SPHERICAL SPIN GLASS

- The first-order measure of complexity is to understand $\mathbb{E}\mathrm{Crt}_{N,k}(B)$ as $N \to \infty$, for fixed k, B and p.

- [Auffinger, Ben Arous & Cerny'10] exploited the Gaussian structure to prove

Theorem: [ABC'10]

$$\mathbb{E}[\mathrm{Crt}_{N,k}(B)] = 2\sqrt{\frac{2}{p}}(p-1)^{N/2}\mathbb{E}^N_{\mathrm{GOE}}\left\{e^{-N\frac{p-2}{2p}(\lambda^N_k)^2}\mathbf{1}\left[\lambda^N_k \in \sqrt{\frac{p}{2p-2}}B\right]\right\}$$

  - *GOE* is the Gaussian Orthogonal Ensemble, a random matrix where entries are independent Gaussians.
  - The joint distribution of eigenvalues is well-known and studied.

COMPLEXITY OF SPHERICAL SPIN GLASS

- Using this result, we obtain the following asymptotic behavior:

Theorem [ABC'10]: For all $p \geq 2$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\mathrm{Crt}_{N,k}(u) = \Theta_{k,p}(u)\,,$$

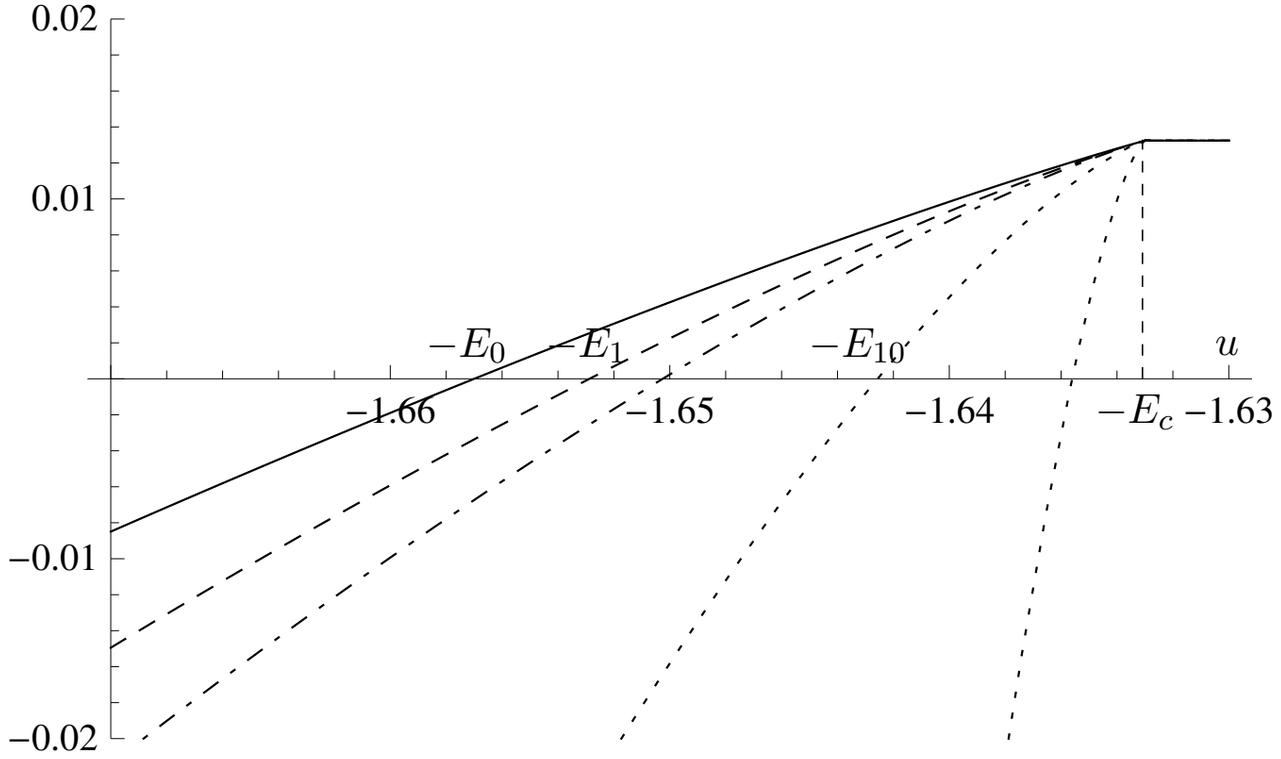
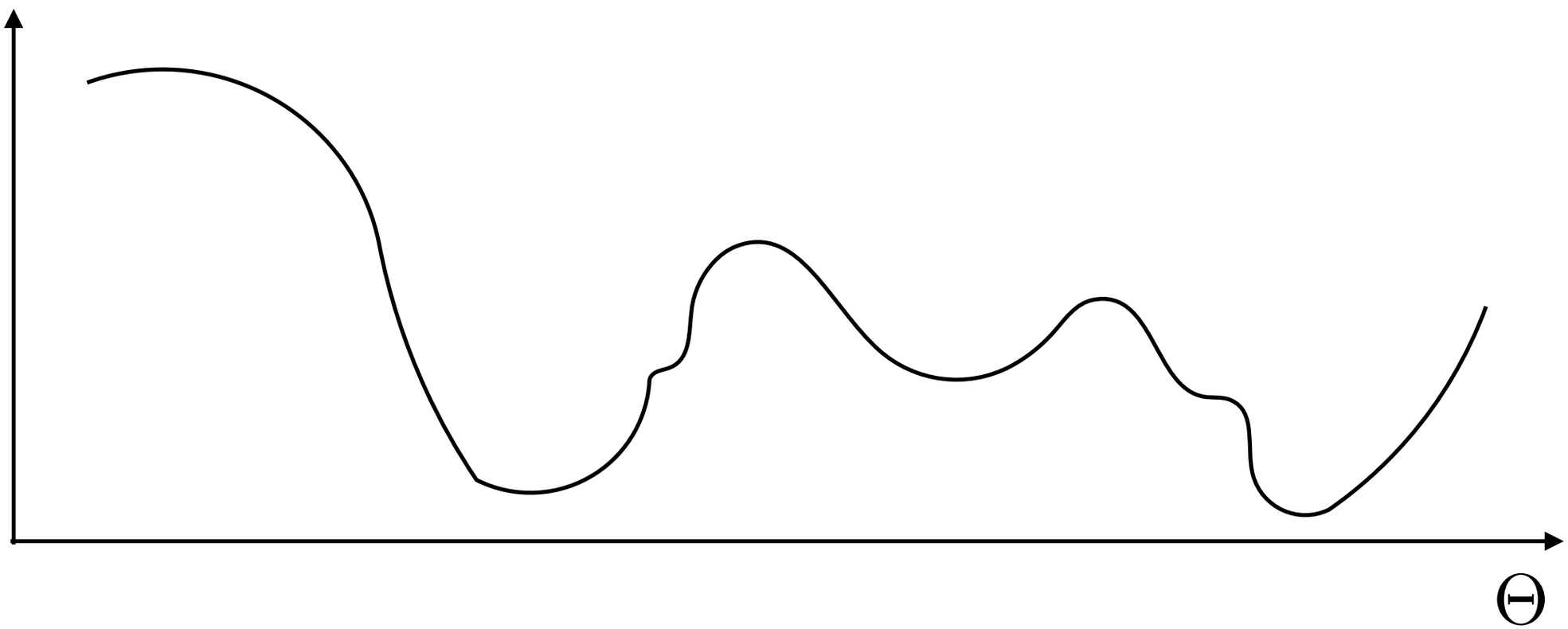


FIGURE 1. The functions $\Theta_{k,p}$ for $p = 3$ and $k = 0$ (solid), $k = 1$ (dashed), $k = 2$ (dash-dotted), $k = 10$, $k = 100$ (both dotted). All these functions agree for $u \geq -E_\infty$.

SPHERICAL SPIN GLASS

- [Auffinger et al '10] [Auffinger, Ben Arous'13], obtained a complete description of the behavior of critical points of spherical spin glasses.

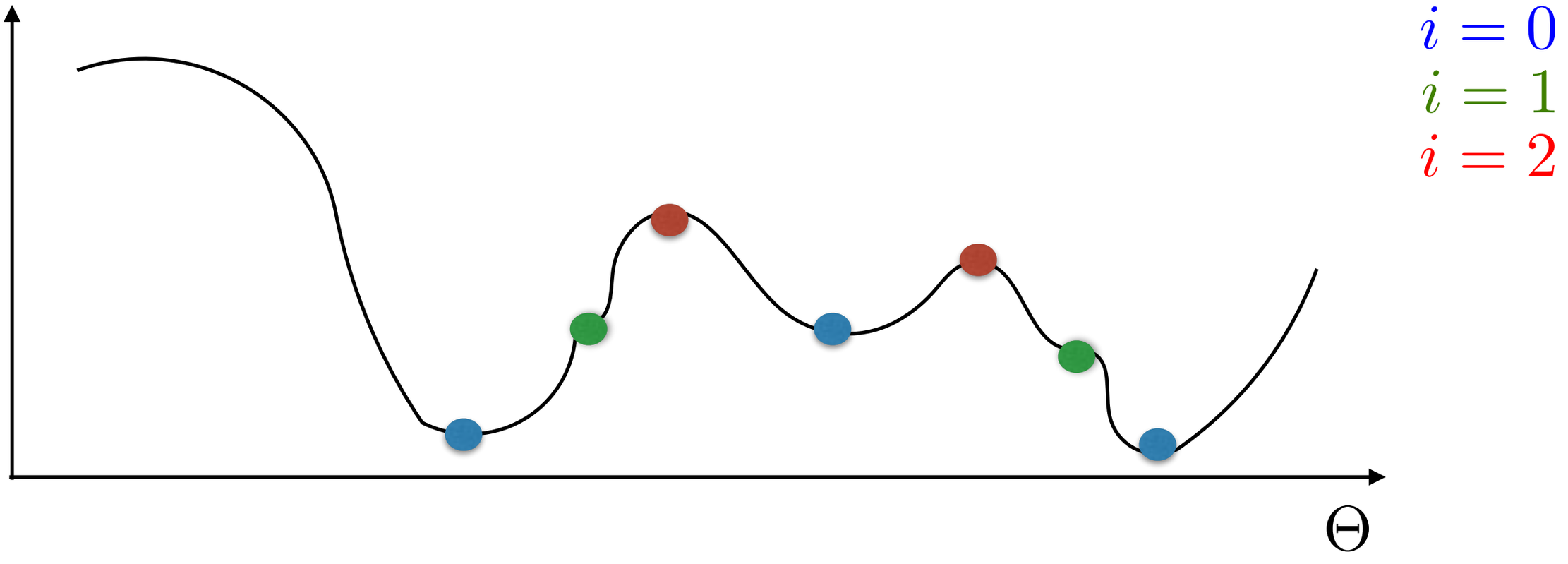
- In particular, critical points (ratio of negative to positive eigenvalues of the Hessian) occur at different energy bands:



SPHERICAL SPIN GLASS

- This result uncovers an important aspect of the landscape: layered structure of critical points according to their *index*.

- In particular, index of critical points (ratio of negative to positive eigenvalues of the Hessian) occur at different energy bands:



SPHERICAL SPIN GLASSES

- Let $E_k(p)$ be such that $\Theta_{k,p}(-E_k(p)) = 0$.
- All critical points of $H_{N,p}$ of finite index concentrate in the band $(-NE_0(p), -NE_\infty(p))$

Theorem 2.14. *Let for an integer $k \geq 0$ and $\varepsilon > 0$, $B_{N,k}(\varepsilon)$ be the event "there is a critical value of index k of the Hamiltonian $H_{N,p}$ above the level $-N(E_\infty(p) - \varepsilon)$", that is $B_{N,k}(\varepsilon) = \{\mathrm{Crt}_{N,k}((-E_\infty(p) + \varepsilon, \infty)) > 0\}$. Then for all $k \geq 0$ and $\varepsilon > 0$,*

$$\limsup_{N\to\infty} \frac{1}{N^2} \log \mathbb{P}(B_{N,k}(\varepsilon)) < 0. \tag{2.25}$$

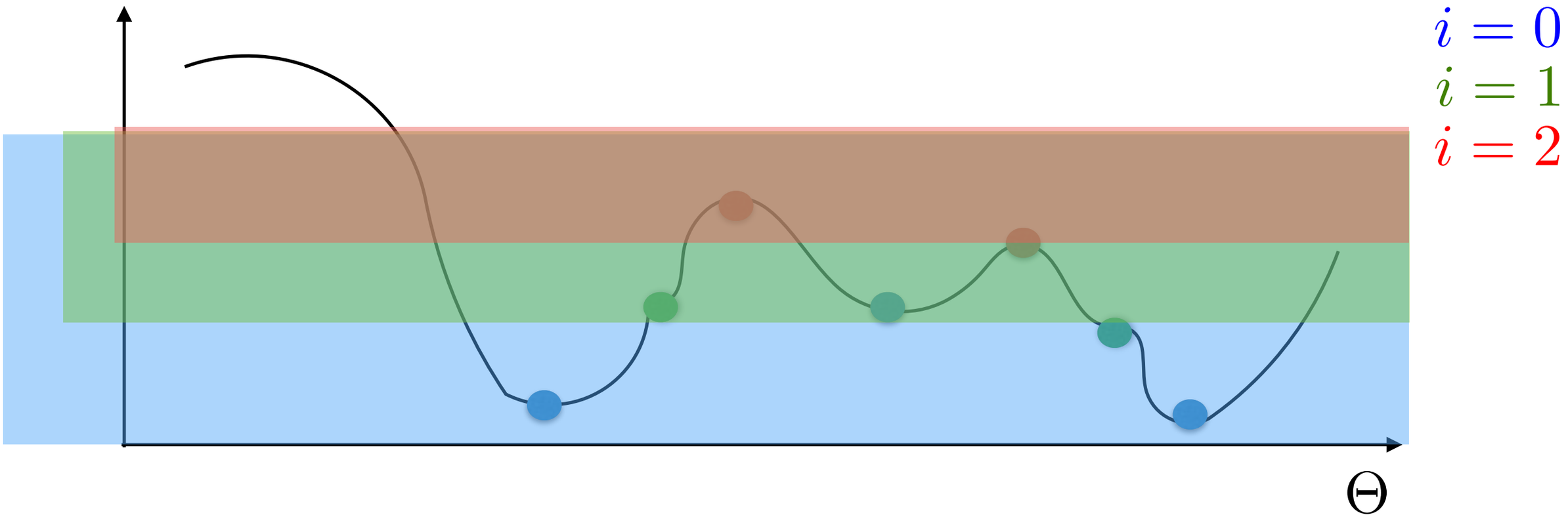
- They concentrate according to their index:

Theorem 2.15. *For $k \geq 0$ and $\varepsilon > 0$, let $A_{N,k}(\varepsilon)$ to be the event "there is a critical value of the Hamiltonian $H_{N,p}$ below the level $-N(E_k(p) + \varepsilon)$ and with index larger or equal to k", that is $A_{N,k}(\varepsilon) = \{\sum_{i=k}^{\infty} \mathrm{Crt}_{N,i}(-E_k(p) - \varepsilon) > 0\}$. Then for all $k \geq 0$ and $\varepsilon > 0$,*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(A_{N,k}(\varepsilon)) < 0. \tag{2.26}$$

SPHERICAL SPIN GLASSES

- As $N \to \infty$, the number of local minima dominate the rest of the indices.
- Moreover, in the energy range $(-NE_k(p), -NE_{k+1}(p))$, we only find critical points with index $\leq k$
- In particular, what is the behavior of the energy gap to move from one local minima to the next?



SPHERICAL SPIN GLASSES

- These results describe the *average* number of critical points at in terms of energy value and index.
- But do they capture the *typical* behavior of a given realization of the random function $H_{N,p}$?
- This requires control not only of $\log \mathbb{E}\mathrm{Crt}_{N,k}$, but rather of $\mathbb{E} \log \mathrm{Crt}_{N,k}$

- In statistical physics, this is referred as the *quenched entropy,* as opposed to the *annealed entropy*.

THE KAC-RICE FORMULA

- The key mathematical tool behind these results is the so-called *Kac-Rice formula*.
- It is a general tool to compute the expected number of points on a manifold that satisfy certain random functional constraints. Given $P : \mathbb{R}^d \to \mathbb{R}^d$, $Q : \mathbb{R}^d \to \mathbb{R}^k$, B open set in $\mathbb{R}^k$ the Kac-Rice formula counts the expected number of points $x \in \mathbb{R}^d$ that satisfy $P(x) = 0$, $Q(x) \in B$.
- When $P(x) = \nabla f(x), Q(x) = \{\nabla^2 f(x), x\}$, $B = \{A; 0 \succeq A\} \otimes E, (E \subset S^{d-1})$

Kac-Rice lemma: Let f be a random function defined on the unit sphere S^{d-1} and $E \subset S^{d-1}$. Denote by M_f the set of all local maxima of f. Then under appropriate regularity

$$\mathbb{E}\{M_f \cap E\} = \int_x \mathbb{E}\left[|\det(\nabla^2 f)|\mathbf{1}(0 \succeq \nabla^2 f) \cdot \mathbf{1}(x \in E)|\nabla f = 0\right] \rho_{\nabla f(x)}(0) dx\,.$$

NOISE + SIGNAL LANDSCAPE MODEL

- What is the effect of adding a deterministic smooth function on top of the previous random model?

- How to move from average complexity measures to typical complexity?

NOISE + SIGNAL LANDSCAPE MODEL

- [Ros et al.'18] consider the modified loss

$$H_{p,q,r}(\sigma) = -\sum_{i_1,\ldots,i_p=1}^{N} J_{i_1,\ldots,i_p}\sigma_{i_1}\ldots\sigma_{i_p} - rNf_q(N^{-1}\langle\sigma, v_0\rangle)$$

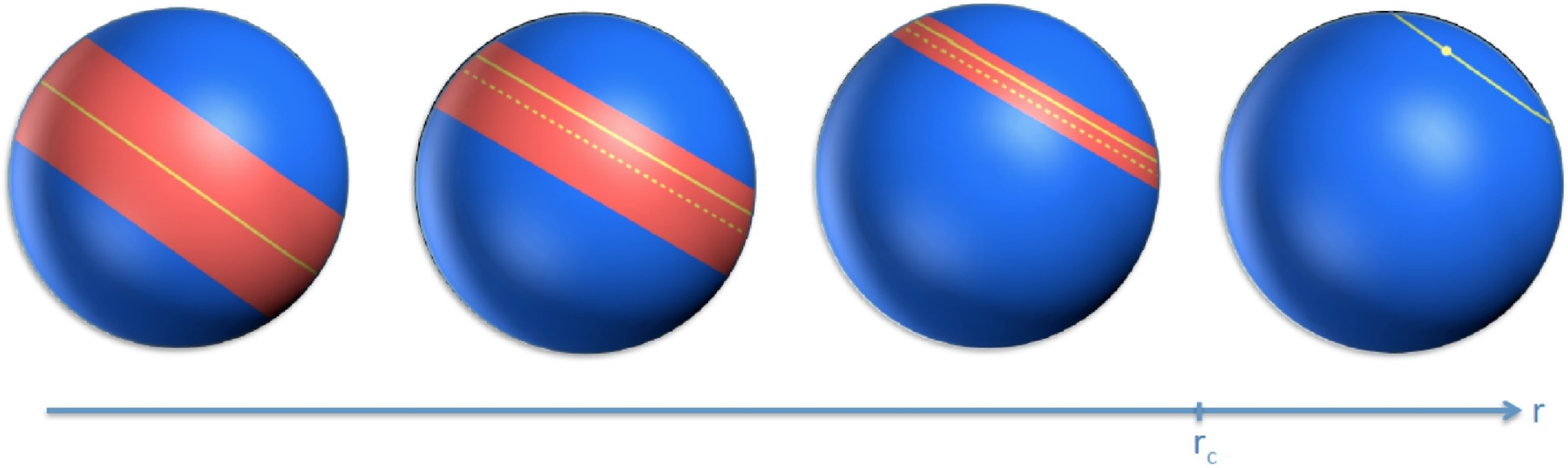
- First term is random as before
- Deterministic term: $f_q : [0,1] \to \mathbb{R}$ monotonic increasing, $s.t.\, \partial^r f(0) = 0 \,\forall\, r < q\,, \partial^q f(0) \neq 0\,.$
- It favors a fixed configuration v_0. r controls the SNR.
- If $f_q = \frac{x^q}{q}, q = p$, this is the *spiked tensor model*.

NOISE + SIGNAL LANDSCAPE MODEL

- There is a tension between the random and deterministic components.
- The random component contains exponential number of critical points (from previous analysis), whereas the deterministic component has only one global minima.
- The energy landscape is fundamentally controlled by both the SNR parameter r and by the order q.
- How does the qualitative behavior depend upon r and q?

NOISE + SIGNAL LANDSCAPE MODEL

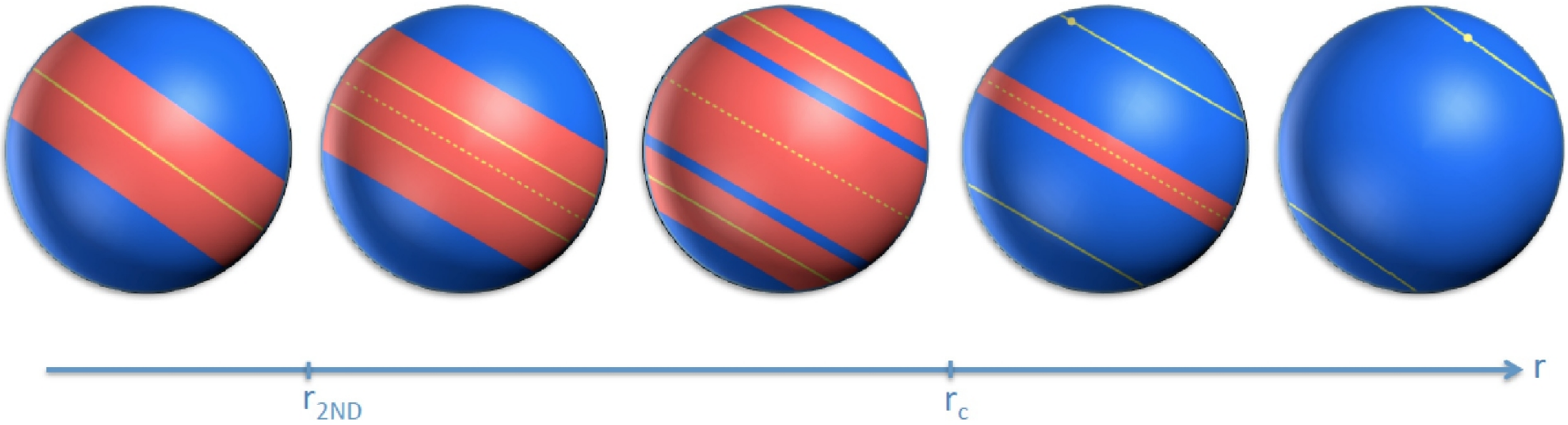
- First case: $q=1$



- bulk of local minima in red strip, deepest minima in yellow strip.
- At $r = r_c$ we encounter a phase-transition: only one minima survives, although not exactly at $\sigma = v_0$.

NOISE + SIGNAL LANDSCAPE MODEL

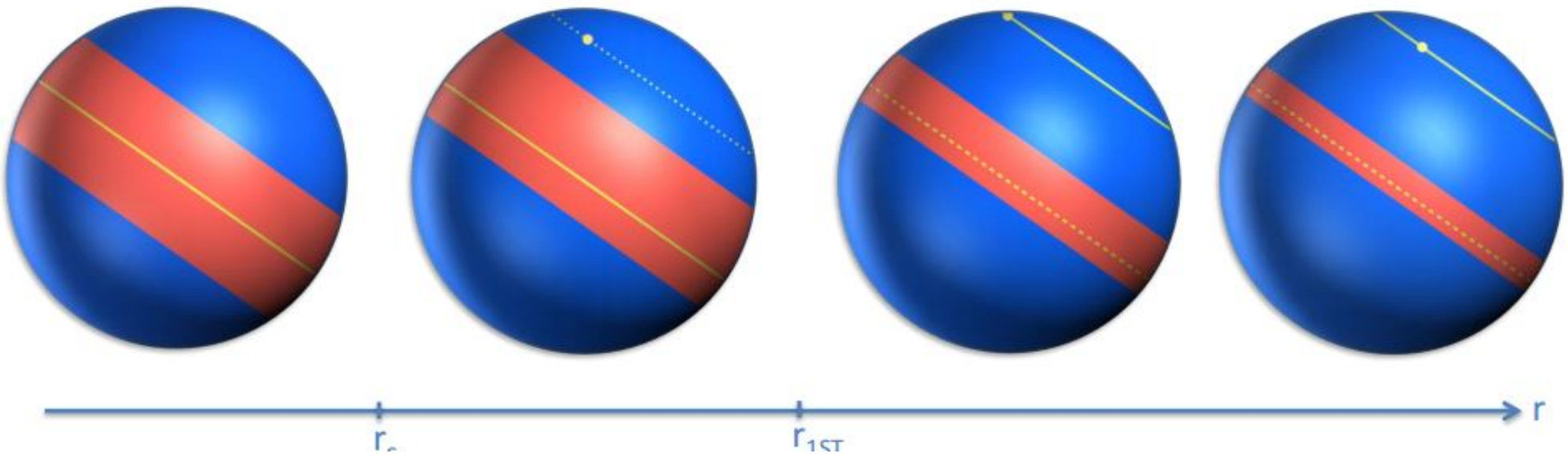
- Second case, $q=2$:

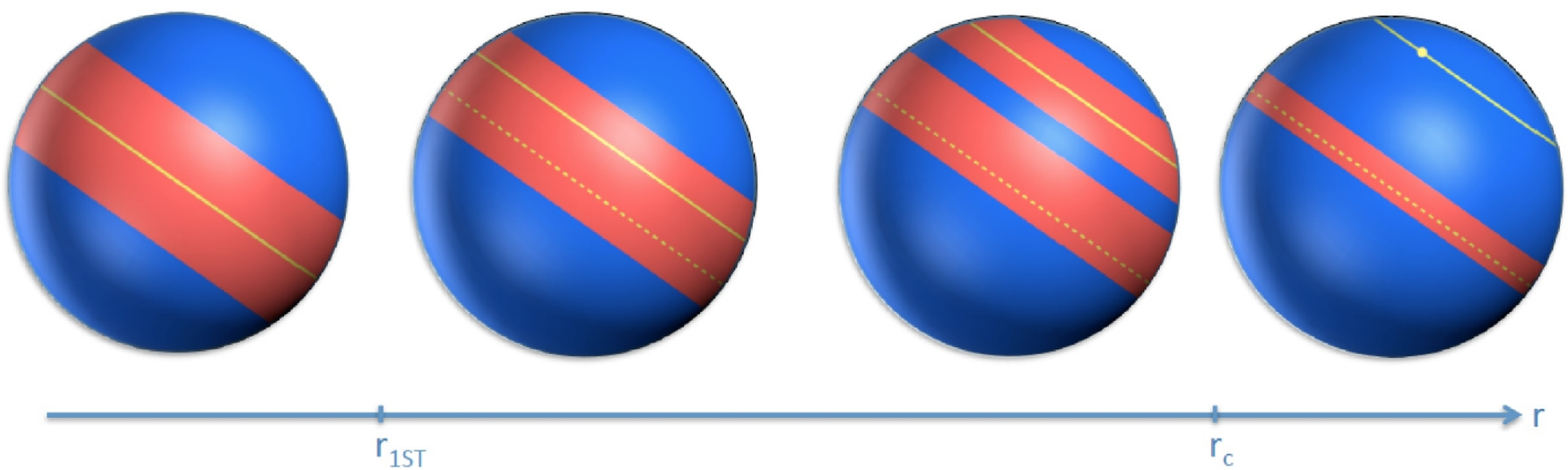


- Now there are two transitions: first the deepest local minima splits and moves towards poles, breaking the bulk of local minima in three bands.
- Next, the bands closer to poles turn into two isolated global minima, and equator band eventually disappears.

NOISE + SIGNAL LANDSCAPE MODEL

- General case, $q>2$: Now the deterministic term has no influence on Gradient and Hessian at equator.





THE REPLICA METHOD

- These results are obtained qualitatively using tools from Statistical Physics.
- One remarkable tool is the *replica method*.
- It starts by considering the system free energy at temperature β

$$f = -\frac{1}{\beta N}\mathbb{E}\log Z\,, \qquad Z = \sum_{\sigma}\exp(-\beta H(\sigma))\,.$$

- As the temperature $\beta \to 0$, the partition function is dominated by states close to global minima.
- But, how to estimate the expectation $\mathbb{E}\log Z$?

THE REPLICA METHOD

- The replica method computes instead the moments $\mathbb{E}Z^r$, $r \in \mathbb{N}$ and considers the analytic continuation

$$\mathbb{E}\log Z = \lim_{r \to 0} \frac{Z^r - 1}{r} .$$

- The moment $\mathbb{E}Z^r$ can be expressed in terms of r *replicas* $\sigma^{(1)}, \ldots, \sigma^{(r)}$:

$$\mathbb{E}[Z^r] = \sum_{\sigma^{(j)}} \mathbb{E}\exp\left(-\beta \sum_{j=1}^{r} H(\sigma^{(j)})\right) .$$

- The *Replica Symmetry Breaking* (RSB) formalism studies the geometry of critical points by studying how replicas overlap.

THE LANDSCAPE OF TENSOR DECOMPOSITION

- Consider a 4-th order tensor T of the form

$$T = \sum_{i=1}^{n} a_i \otimes a_i \otimes a_i \otimes a_i \ , \quad a_i \in \mathbb{R}^d \, .$$

- Focus on the case where rank $n \gg d$.
- Goal: recover components $a_1 \dots a_n$ from T.
- This is in general NP-hard.
- However, for a given problem distribution, we can consider the average-case recovery.

THE LANDSCAPE OF TENSOR DECOMPOSITION

- Ge and Ma focus on the Gaussian case $a_i \sim \mathcal{N}(0, I)$, using the loss

$$\max F(x) = \sum_{i,j,k,l=1}^{d} T_{i,j,k,l} x_i x_j x_k x_l = \langle T, x^{\otimes 4} \rangle_F = \sum_{i=1}^{d} \langle a_i, x \rangle^4 \,,\; x \in S^{d-1}.$$

- This objective finds the direction that mostly correlates with T
- Another random polynomial on the sphere, but now coefficients are not iid Gaussian.
- For $n \ll d^2$, global maxima of F are located around $\pm \frac{1}{\sqrt{d}} a_i$.

THE LANDSCAPE OF TENSOR DECOMPOSITION

- Ge and Ma use Kac-Rice formula to attack the over complete case:

Theorem [Ge & Ma]: Let $\epsilon, \zeta \in (0, 1/3)$, d sufficiently large, and assume $d^{1+\epsilon} < n < d^{2-\epsilon}$. Then whp the superlevel set $\{x \in S^{d-1}; F(x) \geq 3(1+\zeta)n\}$ contains exactly $2n$ local maxima, each of which is close to one of $\pm \frac{1}{\sqrt{d}} a_i$.

- No spurious local maxima in this super-level set.
- A random initialization on the sphere does not suffice:

$$\mathbb{E}[\langle a, x \rangle^4] = 3\,, a \sim \mathcal{N}(0, I), x \sim \mathrm{Unif}(S^{d-1})\,.$$

- The proof builds an estimator for the Kac-Rice formula.
- Open: spurious local maxima outside this level set?

SPIKED TENSOR MODEL

- A related problem is the recovery of a rank-one tensor

$$U = u^{\otimes k}, u \in \mathbb{R}^d$$

from noisy measurements: $Y = \lambda u^{\otimes k} + \frac{1}{\sqrt{2d}} W$

$W \in (\mathbb{R}^d)^{\otimes k}, W = \sum_{\pi \in S_d} G^\pi/(k!), G_{i_1,\ldots,i_k} \sim \mathcal{N}(0,1).$

- Similarly as before, we consider the (ML) objective

$$\max F(x) = \langle Y, x^{\otimes k} \rangle_F, s.t.\, x \in S^{d-1}.$$

- λ is the Signal-to-Noise ratio of the problem.
- Solving this objective is generally NP-hard for k>2.

SPIKED TENSOR MODEL

- In [Ben Arous, Mei, Montanari and Mica'18], the authors study the landscape of optimization for a typical realization of Y.

- Using again the Kac-Rice formula, they compute the expected number of critical points of the likelihood objective, as a function of SNR λ , order k and dimension d.

- The expected # of local maxima with $F(x) \approx s$, $\langle x, u \rangle \approx m$ is $\exp\{dS_0(m, s) + o(d)\}$.

SPIKED TENSOR MODEL

- In particular, the expected bulk of (exponential) local maxima are in the annulus $|\langle u, x\rangle| \leq \Theta(\lambda^{-1/(k-2)})$.
- This suggests that local ascent methods need to be initialized outside this annulus: $x_0\ s.t.\ |\langle u, x_0\rangle| > C\lambda^{-1/(k-2)}$
- But random initialization results in $|\langle u, x_0\rangle| = \Theta(d^{-1/2})$, thus one can escape local maxima provided

$$\lambda \geq Cn^{\frac{k-2}{2}}\ .$$

- This rate matches the best known polynomial time algorithm (power iteration algorithm by [Montanari et al.'14]).

FROM TENSORS TO NEURAL NETWORKS

- in [Mondelli & Montanari,'18], the authors reduce the learning of a 2-layer neural network with Gaussian input to the problem of tensor decomposition.

- in [Cohen, Shashua et al], network architecture (width and depth) is related to certain forms of tensor decomposition (parafac vs hierarchical). Depth provides more efficient tensor approximation.

DEEP NETWORKS AND SPIN GLASSES

- See also:
  - *"The effect of Gradient Noise on the Energy Landscape of Deep Networks"*, Chaudhari & Soatto. They study exterior magnitude field and its associated smoothing annealing schemes to reduce number of critical points.
  - *"Explorations on high dimensional landscapes"*, Sagun, Guney, Ben Arous, LeCun. Study the existence of a narrow band containing the bulk of the critical points of deep energy landscapes in the high-dimensional setting.

OPTIMIZATION LANDSCAPE OF NEURAL NETWORKS

- Relatively recent field, but many papers already.
  - Rate of papers increasing.
- Most papers focus on the deterministic aspect of the landscape: Given an input data distribution $(X, Y) \sim P$ and model architecture $\Phi(x; \theta)$, $\theta \in \Theta$, we consider the landscape

$$F(\theta) = \mathbb{E}_P[\ell(\Phi(X; \theta), Y)]$$

- This framework contains both population and empirical risk, by considering either P or $\hat{P} = \frac{1}{L}\sum_{l=1}^{L} \delta_{(X_l, Y_l)}$.
- Here, the *mean* landscape is expected to be of similar complexity than the *empirical* landscape for large L.

RELATED WORK

- Models from Statistical physics have been considered as possible approximations [Dauphin et al.'14, Choromanska et al.'15, Segun et al.'15]
- Tensor factorization models capture some of the non convexity essence [Anandukar et al'15, Cohen et al. '15, Haeffele et al.'15]
- [Shafran and Shamir,'15] studies bassins of attraction in neural networks in the overparametrized regime.
- [Soudry'16, Song et al'16] study Empirical Risk Minimization in two-layer ReLU networks, also in the over-parametrized regime.

RELATED WORK

- Models from Statistical physics have been considered as possible approximations [Dauphin et al.'14, Choromanska et al.'15, Segun et al.'15]
- Tensor factorization models capture some of the non convexity essence [Anandukar et al'15, Cohen et al. '15, Haeffele et al.'15]
- [Shafran and Shamir,'15] studies bassins of attraction in neural networks in the overparametrized regime.
- [Soudry'16, Song et al'16] study Empirical Risk Minimization in two-layer ReLU networks, also in the over-parametrized regime.
- [Tian'17] studies learning dynamics in a gaussian generative setting.
- [Chaudhari et al'17]: Studies local smoothing of energy landscape using the local entropy method from statistical physics.
- [Pennington & Bahri'17]: Hessian Analysis using Random Matrix Th.
- [Soltanolkotabi, Javanmard & Lee'17]: layer-wise quadratic NNs.

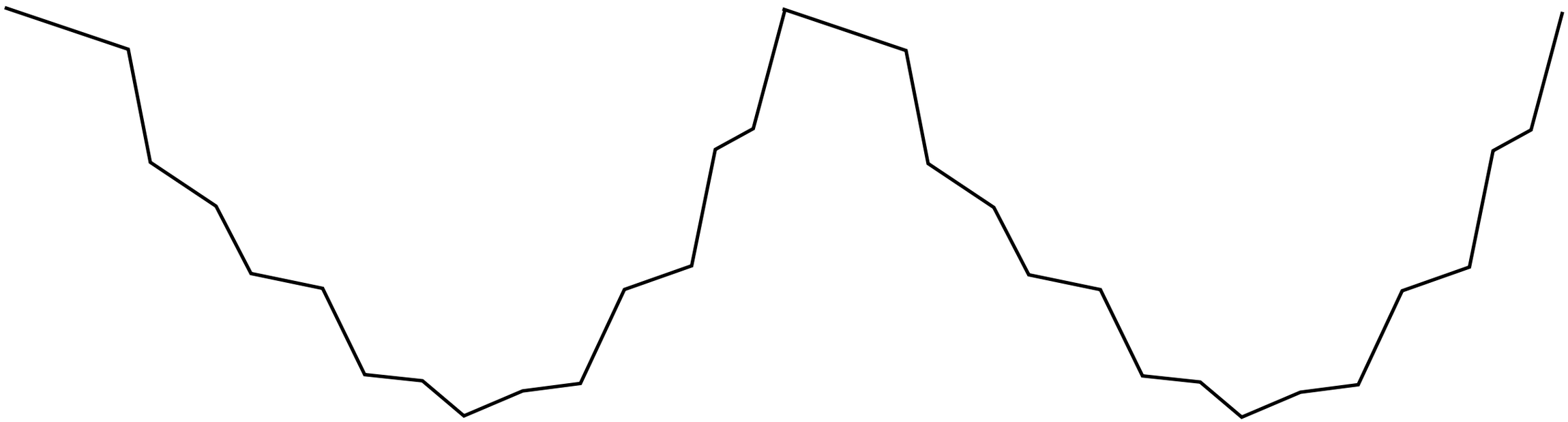
NON-CONVEXITY ≠ NOT OPTIMIZABLE

- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.

NON-CONVEXITY ≠ NOT OPTIMIZABLE

- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.
- In particular, deep models have internal symmetries.

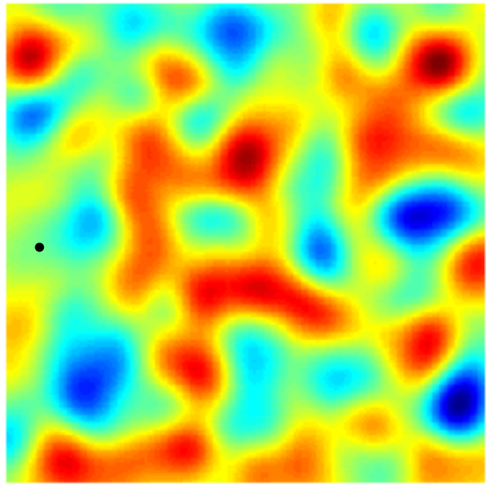
$$F(\theta) = F(g.\theta) \ , \ g \in G \text{ compact.}$$



ANALYSIS OF NON-CONVEX LOSS SURFACES

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

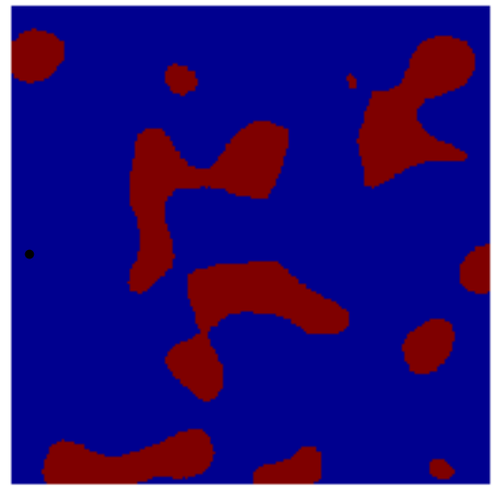
$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\} \ .$$



ANALYSIS OF NON-CONVEX LOSS SURFACES

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\}$$

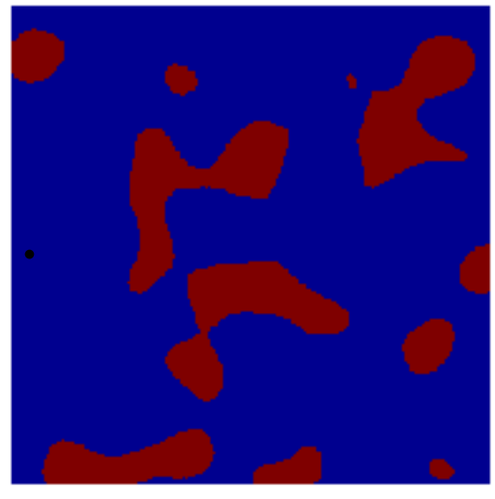


- A first notion we address is about the topology of the level sets Ω_u.
- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u ?

ANALYSIS OF NON-CONVEX LOSS SURFACES

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\}$$



- A first notion we address is about the topology of the level sets Ω_u.

- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u ?

- Related to presence of poor local minima:

Proposition: If $N_u = 1$ for all u then E has no poor local minima.

(i.e. no local minima y^* s.t. $E(y^*) > \min_y E(y)$)

SPURIOUS VALLEYS

- More generally, we are interested in certifying existence of descent paths.

SPURIOUS VALLEYS

- More generally, we are interested in certifying existence of descent paths.
- **Definition:** A *valley* is a connected component of the sublevel set Ω_u.

  Definition: A *spurious valley* is a connected component of the sublevel set Ω_u that does not contain a global minima.

SPURIOUS VALLEYS

- More generally, we are interested in certifying existence of descent paths.
- **Definition:** A *valley* is a connected component of the sublevel set Ω_u.

  Definition: A *spurious valley* is a connected component of the sublevel set Ω_u that does not contain a global minima.
- If a loss has no spurious valley, then one can continuously move from any point in parameter space to a global minima without increasing the loss:

  Given any initial parameter $\theta_0 \in \Theta$,
  $\exists$ continuous path $\theta : t \in [0,1] \mapsto \theta(t) \in \Theta$ st:
  $\theta(0) = \theta_0$, $\theta(1) \in \arg\min_{\theta} E(\theta)$, and
  $t \mapsto E(\theta(t))$ is non-increasing.

LINEAR VS NON-LINEAR DEEP MODELS

- Some authors have considered linear "deep" models as a first step towards understanding nonlinear deep models:

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2 \ .$$

$$X \in \mathbb{R}^n \ , \ Y \in \mathbb{R}^m \ , \ W_k \in \mathbb{R}^{n_k \times n_{k-1}} \ .$$

LINEAR VS NON-LINEAR DEEP MODELS

- Some authors have considered linear "deep" models as a first step towards understanding nonlinear deep models:

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2 \ .$$

$$X \in \mathbb{R}^n \ , \ Y \in \mathbb{R}^m \ , \ W_k \in \mathbb{R}^{n_k \times n_{k-1}} \ .$$

Theorem: [**Kawaguchi'16**] If $\Sigma = \mathbb{E}(XX^T)$ and $\mathbb{E}(XY^T)$ are full-rank and Σ has distinct eigenvalues, then $E(\Theta)$ has no poor local minima.

- studying critical points.
- later generalized in [Hardt & Ma'16, Lu & Kawaguchi'17]

LINEAR VS NON-LINEAR DEEP MODELS

$$E(W_1, \dots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \dots W_1 X - Y\|^2 \ .$$

Proposition: [BF'16]

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.

2. (2-layer case, ridge regression)
   $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$
   satisfies $N_u = 1 \ \forall \ u$ if $n_1 > \min(n, m)$.

- We pay extra redundancy price to get simple topology.

LINEAR VS NON-LINEAR DEEP MODELS

$$E(W_1, \dots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \dots W_1 X - Y\|^2 \ .$$

Proposition: [BF'16]

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.

2. (2-layer case, ridge regression)
   $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$
   satisfies $N_u = 1 \ \forall \ u$ if $n_1 > \min(n, m)$.

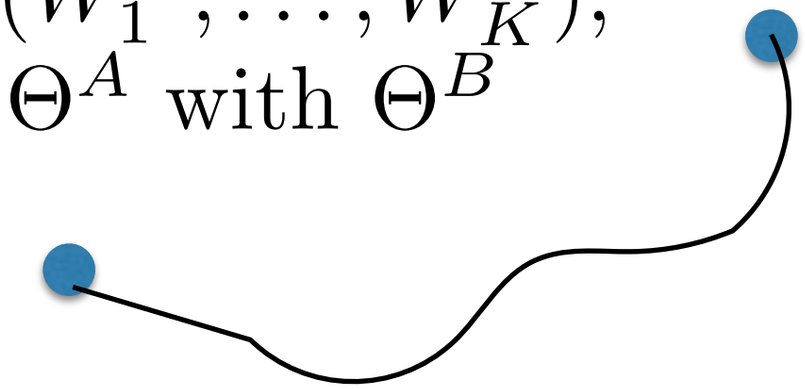
- We pay extra redundancy price to get simple topology.
- This simple topology is an "artifact" of the linearity of the network:

Proposition: [BF'16] For any architecture (choice of internal dimensions), there exists a distribution $P_{(X,Y)}$ such that $N_u > 1$ in the ReLU $\rho(z) = \max(0, z)$ case.

PROOF SKETCH

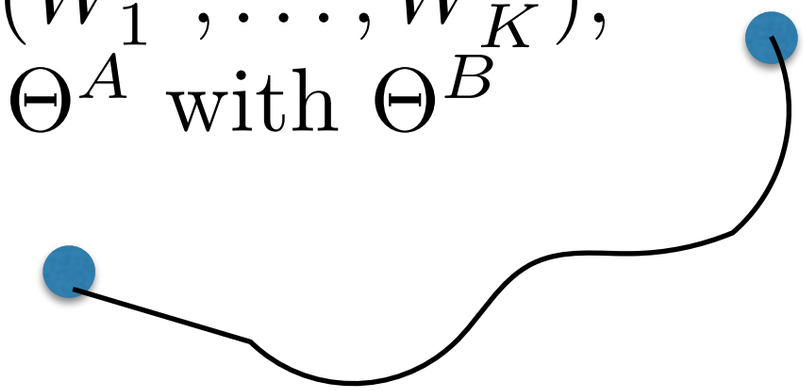
- Goal:

Given $\Theta^A = (W_1^A, \ldots, W_K^A)$ and $\Theta^B = (W_1^B, \ldots, W_K^B)$, we construct a path $\gamma(t)$ that connects Θ^A with Θ^B st $E(\gamma(t)) \leq \max(E(\Theta^A), E(\Theta^B))$.



PROOF SKETCH

➤ Goal:
Given $\Theta^A = (W_1^A, \ldots, W_K^A)$ and $\Theta^B = (W_1^B, \ldots, W_K^B)$,
we construct a path $\gamma(t)$ that connects Θ^A with Θ^B
st $E(\gamma(t)) \leq \max(E(\Theta^A), E(\Theta^B))$.



➤ Main idea:

1. Induction on K.
2. *Lift* the parameter space to $\widetilde{W} = W_1 W_2$: the problem is convex $\Rightarrow$ there exists a (linear) path $\widetilde{\gamma}(t)$ that connects Θ^A and Θ^B.
3. Write the path in terms of original coordinates by *factorizing* $\widetilde{\gamma}(t)$.

➤ Simple fact:
If $M_0, M_1 \in \mathbb{R}^{n \times n'}$ with $n' > n$,
then there exists a path $t : [0,1] \to \gamma(t)$
with $\gamma(0) = M_0$, $\gamma(1) = M_1$ and
$M_0, M_1 \in \text{span}(\gamma(t))$ for all $t \in (0,1)$.

MODEL SYMMETRIES

[with L. Venturi, A. Bandeira, '17]

- How much extra redundancy are we paying to achieve $N_u = 1$ instead of simply no poor-local minima?

MODEL SYMMETRIES

[with L. Venturi, A. Bandeira, '17]

- How much extra redundancy are we paying to achieve $N_u = 1$ instead of simply no poor-local minima?
  - In the multilinear case, we don't need $n_k > \min(n, m)$.

$$(W_1, W_2, \dots W_K) \sim (\widetilde{W}_1, \dots, \widetilde{W}_K) \Leftrightarrow \widetilde{W}_k = U_k W_k U_{k-1}^{-1} \ , \ U_k \in GL(\mathbb{R}^{n_k \times n_k}) \ .$$

MODEL SYMMETRIES [with L. Venturi, A. Bandeira, '17]

- How much extra redundancy are we paying to achieve $N_u = 1$ instead of simply no poor-local minima?
  - In the multilinear case, we don't need $n_k > \min(n, m)$

$$(W_1, W_2, \dots W_K) \sim (\widetilde{W}_1, \dots, \widetilde{W}_K) \Leftrightarrow \widetilde{W}_k = U_k W_k U_{k-1}^{-1} \ , \ U_k \in GL(\mathbb{R}^{n_k \times n_k}) \ .$$

- We do the same analysis in the quotient space defined by the equivalence relationship .

Theorem: [BVV'18] The Multilinear regression $\mathbb{E}_{(X,Y)\sim P} \|W_K \dots W_1 X - Y\|^2$ has no spurious valleys.

MODEL SYMMETRIES

[with L. Venturi, A. Bandeira, '17]

- How much extra redundancy are we paying to achieve $N_u = 1$ instead of simply no poor-local minima?
  - In the multilinear case, we don't need $n_k > \min(n, m)$

$$(W_1, W_2, \dots W_K) \sim (\widetilde{W}_1, \dots, \widetilde{W}_K) \Leftrightarrow \widetilde{W}_k = U_k W_k U_{k-1}^{-1} \ , \ U_k \in GL(\mathbb{R}^{n_k \times n_k}) \ .$$

- We do the same analysis in the quotient space defined by the equivalence relationship .

Theorem: [BVV'18] The Multilinear regression $\mathbb{E}_{(X,Y)\sim P} \|W_K \dots W_1 X - Y\|^2$ has no spurious valleys.

- Construct paths on the Grassmanian manifold of linear subspaces
- Generalizes best known results for multilinear case (no assumptions on covariance).

BETWEEN LINEAR AND RELU: POLYNOMIAL NETS

- Quadratic nonlinearities $\rho(z) = z^2$ are a simple extension of the linear case, by lifting or "kernelizing":

$$\rho(Wx) = \mathcal{A}_W X \ , \ X = xx^T \ , \ \mathcal{A}_W = (W_k W_k^T)_{k \leq M} \ .$$

BETWEEN LINEAR AND RELU: POLYNOMIAL NETS

- Quadratic nonlinearities $\rho(z) = z^2$ are a simple extension of the linear case, by lifting or "kernelizing":

$$\rho(Wx) = \mathcal{A}_W X \ , \ X = xx^T \ , \ \mathcal{A}_W = (W_k W_k^T)_{k \leq M} \ .$$

- Level sets are connected with sufficient overparametrisation:

Proposition: If $M_k \geq 3N^{2^k} \ \forall \ k \leq K$, then the landscape of K-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

BETWEEN LINEAR AND RELU: POLYNOMIAL NETS

- Quadratic nonlinearities $\rho(z) = z^2$ are a simple extension of the linear case, by lifting or "kernelizing":

$$\rho(Wx) = \mathcal{A}_W X \ , \ X = xx^T \ , \ \mathcal{A}_W = (W_k W_k^T)_{k \leq M} \ .$$

- Level sets are connected with sufficient overparametrisation:

Proposition: If $M_k \geq 3N^{2^k} \ \forall \ k \leq K$, then the landscape of K-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

- No poor local minima with much better bounds in the scalar output two-layer case:

Theorem: [BBV'18] The two-layer quadratic network regression $\mathbb{E}_{(X,Y)\sim P}|U(WX)^2 - Y|^2$ has no spurious valleys if $M > 2N$.

ASYMPTOTIC CONNECTEDNESS OF RELU

- Good behavior is recovered with nonlinear ReLU networks, provided they are sufficiently overparametrized:

- Setup: two-layer ReLU network:

$$\Phi(X;\Theta) = W_2\rho(W_1X)\ ,\ \rho(z) = \max(0,z).W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^m$$

ASYMPTOTIC CONNECTEDNESS OF RELU

- Good behavior is recovered with nonlinear ReLU networks, provided they are sufficiently overparametrized:

- Setup: two-layer ReLU network:

$$\Phi(X;\Theta) = W_2\rho(W_1X)\ ,\ \rho(z) = \max(0,z). W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^m$$

Theorem [BF'16]: For any $\Theta^A, \Theta^B \in \mathbb{R}^{m\times n}, \mathbb{R}^m$, with $E(\Theta^{\{A,B\}}) \leq \lambda$, there exists path $\gamma(t)$ from Θ^A and Θ^B such that $\forall\ t\ , E(\gamma(t)) \leq \max(\lambda, \epsilon)$ and $\epsilon \sim m^{-\frac{1}{n}}$.

ASYMPTOTIC CONNECTEDNESS OF RELU

- Good behavior is recovered with nonlinear ReLU networks, provided they are sufficiently overparametrized:

- Setup: two-layer ReLU network:

$$\Phi(X;\Theta) = W_2\rho(W_1X)\ ,\ \rho(z) = \max(0,z). W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^m$$

Theorem [BF'16]: For any $\Theta^A, \Theta^B \in \mathbb{R}^{m\times n}, \mathbb{R}^m$, with $E(\Theta^{\{A,B\}}) \leq \lambda$, there exists path $\gamma(t)$ from Θ^A and Θ^B such that $\forall\ t\ , E(\gamma(t)) \leq \max(\lambda, \epsilon)$ and $\epsilon \sim m^{-\frac{1}{n}}$.

- Overparametrisation "wipes-out" local minima (and group symmetries).

- The bound is cursed by dimensionality, ie exponential in n .

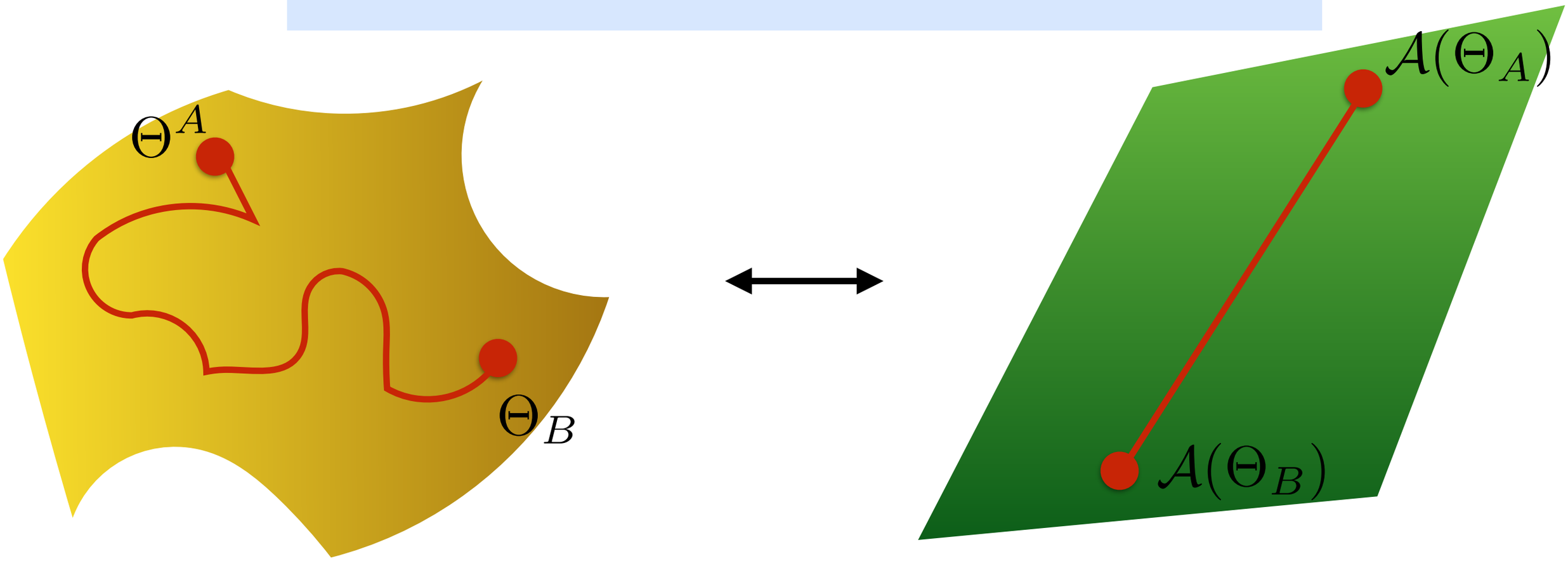
- Result is based on local linearization of the ReLU kernel (hence exponential price).

KERNELS ARE BACK?

$$\Phi(x;\Theta) = W_k \rho(W_{k-1} \dots \rho(W_1 X))) \ , \ \Theta = (W_1, \dots W_k) \ ,$$

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ to *canonical* parameters $\beta = \mathcal{A}(\Theta)$:

$$\Phi(X;\Theta) = \langle \Psi(X), \mathcal{A}(\Theta) \rangle \ .$$



KERNELS ARE BACK?

$$\Phi(x;\Theta) = W_k \rho(W_{k-1} \dots \rho(W_1 X))) \ , \ \Theta = (W_1, \dots W_k) \ ,$$

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ to *canonical* parameters $\beta = \mathcal{A}(\Theta)$

$$\Phi(X;\Theta) = \langle \Psi(X), \mathcal{A}(\Theta) \rangle \ .$$

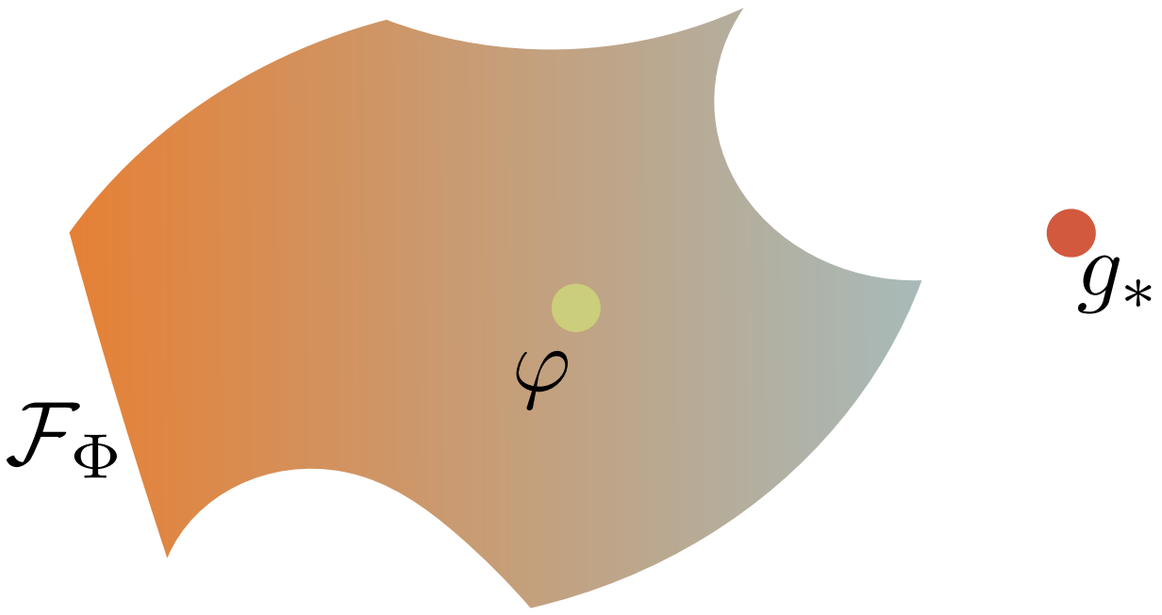
Theorem: [BVV'18] If $\dim\{\mathcal{A}(w),\, w \in \mathbb{R}^n\} = q < \infty$, then $L(U,W) = \mathbb{E}\|U\rho(WX) - Y\|^2$ has no spurious valley if $M \geq q$.

- This includes Empirical Risk Minimization (since RKHS is only queried on finite # of datapoints), and polynomial activations.
- See [Bietti&Mairal'17,Zhang et al'17, Bach'17] for related work.

PARAMETRIC VS MANIFOLD OPTIMIZATION

- This suggests thinking about the problem in the functional space generated by the model:

$$\mathcal{F}_\Phi = \{\varphi : \mathbb{R}^n \to \mathbb{R}^m \,; \varphi(x) = \Phi(x; \Theta) \text{ for some } \Theta\} \,.$$



$$\min_{\varphi \in \mathcal{F}_\Phi} \|\varphi - g_*\|_p$$

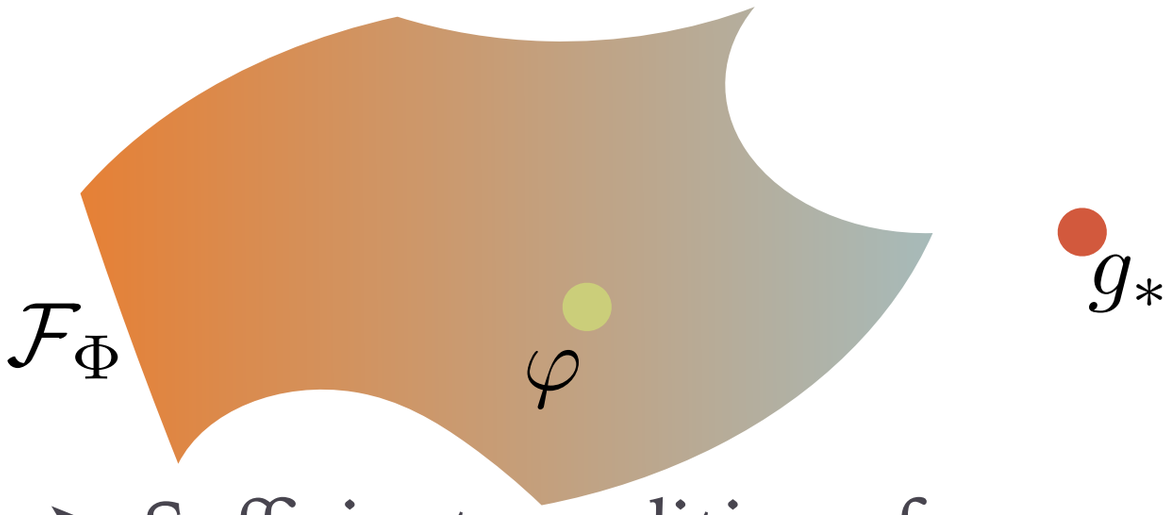
$$g_* : x \mapsto \mathbb{E}(Y|x)$$

$$\langle f, g \rangle_p := \mathbb{E}\{f(X)g(X)\} \,.$$

PARAMETRIC VS MANIFOLD OPTIMIZATION

- This suggests thinking about the problem in the functional space generated by the model:

$$\mathcal{F}_\Phi = \{\varphi : \mathbb{R}^n \to \mathbb{R}^m \,; \varphi(x) = \Phi(x; \Theta) \text{ for some } \Theta\} \,.$$



$$\min_{\varphi \in \mathcal{F}_\Phi} \|\varphi - g_*\|_p$$

$$g_* : x \mapsto \mathbb{E}(Y|x)$$

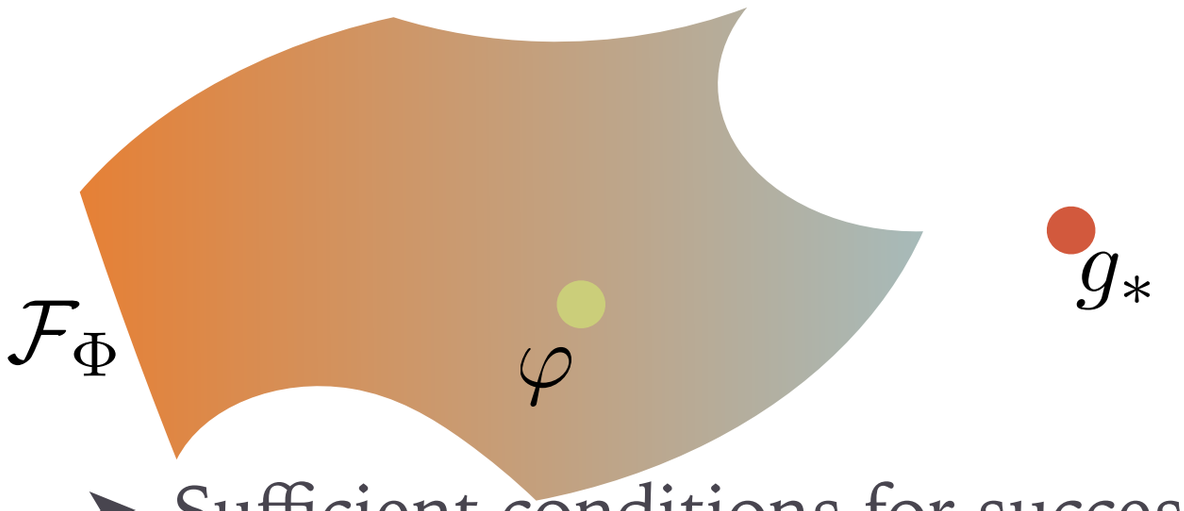
$$\langle f, g \rangle_p := \mathbb{E}\{f(X)g(X)\} \,.$$

- Sufficient conditions for success so far:
  - $\mathcal{F}_\Phi$ convex and Θ sufficiently large so that we can move freely within.
- Necessary condition: $\mathcal{F}_\Phi$ is *ball-connected:*
  $\mathcal{F}_\Phi \cap B_p(R, \epsilon)$ are connected for all p, R, ϵ.

PARAMETRIC VS MANIFOLD OPTIMIZATION

- This suggests thinking about the problem in the functional space generated by the model:

$$\mathcal{F}_\Phi = \{\varphi : \mathbb{R}^n \to \mathbb{R}^m \,; \varphi(x) = \Phi(x;\Theta) \text{ for some } \Theta\} \,.$$



$$\min_{\varphi \in \mathcal{F}_\Phi} \|\varphi - g_*\|_p$$

$$g_* : x \mapsto \mathbb{E}(Y|x)$$

$$\langle f, g \rangle_p := \mathbb{E}\{f(X)g(X)\} \,.$$

- Sufficient conditions for success so far:
  - $\mathcal{F}_\Phi$ convex and Θ sufficiently large so that we can move freely within.
- Necessary condition: $\mathcal{F}_\Phi$ is *ball-connected:*

$$\mathcal{F}_\Phi \cap B_p(R, \epsilon) \text{ are connected for all } p, R, \epsilon.$$

- What happens when the model is not sufficiently overparametrised?

FROM SIMPLE LANDSCAPES TO ENERGY BARRIER?

- Does a similar macroscopic picture arise in our setting?
- Given $\rho(z)$ homogeneous, assume
  - $\tilde{\rho}(\langle w, X\rangle) = \langle A_w, \psi(X)\rangle$, with $\dim(\psi(X)) = f(N)$.
- Define

$$\beta(M,N) = \inf_{S;\dim(S)=f^{-1}(M)} \inf_{\substack{U\in\mathbb{R}^{m\times M}\\ W\in\mathbb{R}^{M\times f^{-1}(M)}}} \sup_{\substack{\mathbb{E}\|Z\|\le N-f^{-1}(M),\\ P_S Z=0}} \mathbb{E}\|U\rho(WP_S X+Z)-Y\|^2$$

  - Best loss obtained by first projecting the data onto the best possible subspace of dimension $f^{-1}(M)$and adding bounded noise in the complement.
  - $\beta(M,N)$ decreases with M and $\beta(f(N),N) = \min_{U,W} E(U,W)$.

FROM SIMPLE LANDSCAPES TO ENERGY BARRIER

- Does a similar macroscopic picture arise in our setting?
- Given $\rho(z)$ homogeneous, assume
  - $\tilde{\rho}(\langle w, X\rangle) = \langle A_w, \psi(X)\rangle$, with $\dim(\psi(X)) = f(N)$.
- Define

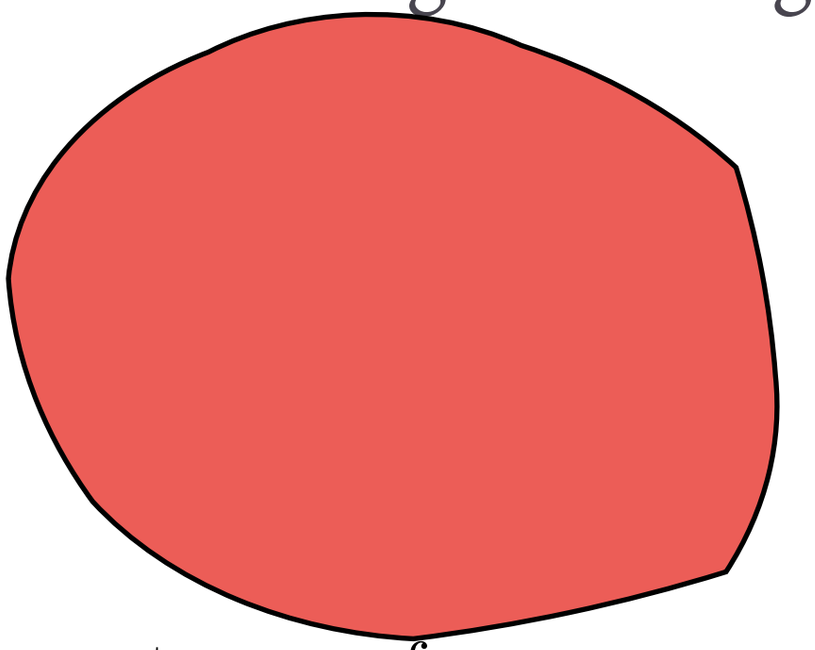
$$\beta(M,N) = \inf_{S;\dim(S)=f^{-1}(M)} \inf_{\substack{U\in\mathbb{R}^{m\times M}\\ W\in\mathbb{R}^{M\times f^{-1}(M)}}} \sup_{\substack{\mathbb{E}\|Z\|\le N-f^{-1}(M),\\ P_S Z=0}} \mathbb{E}\|U\rho(WP_S X+Z)-Y\|^2$$

  - Best loss obtained by first projecting the data onto the best possible subspace of dimension $f^{-1}(M)$and adding bounded noise in the complement.
  - $\beta(M,N)$ decreases with M and $\beta(f(N),N) = \min_{U,W} E(U,W)$.

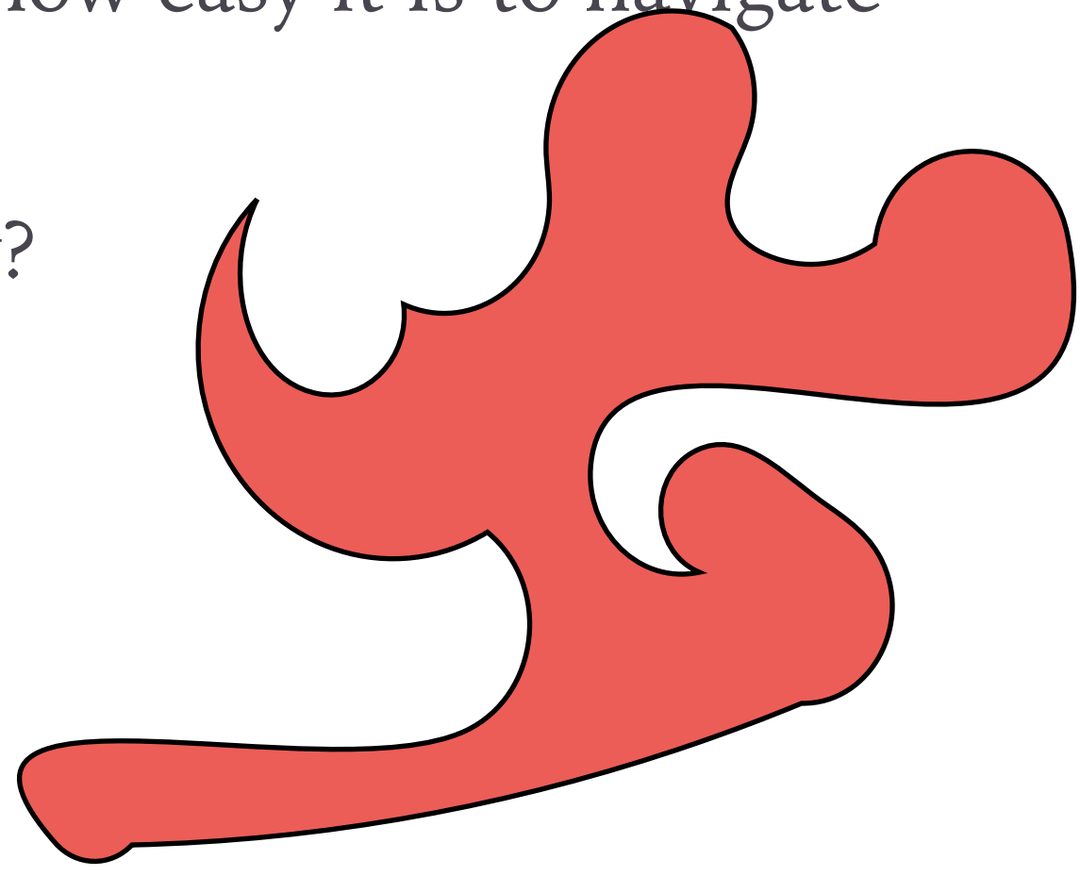
Conjecture [LBB'18]: The loss $L(U,W) = \mathbb{E}\|U\rho(WX)-Y\|^2$ has no poor local minima above the energy barrier $\beta(M,N)$.

FROM TOPOLOGY TO GEOMETRY

- The next question we are interested in is conditioning for descent.
- Even if level sets are connected, how easy it is to navigate through them?
- How "large" and regular are they?



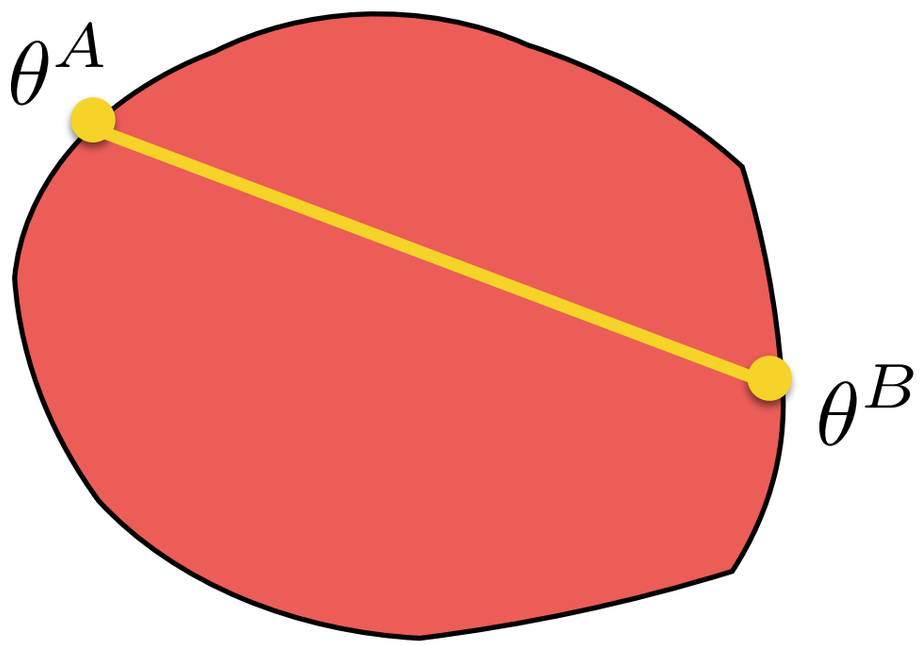
easy to move from one energy level to lower one



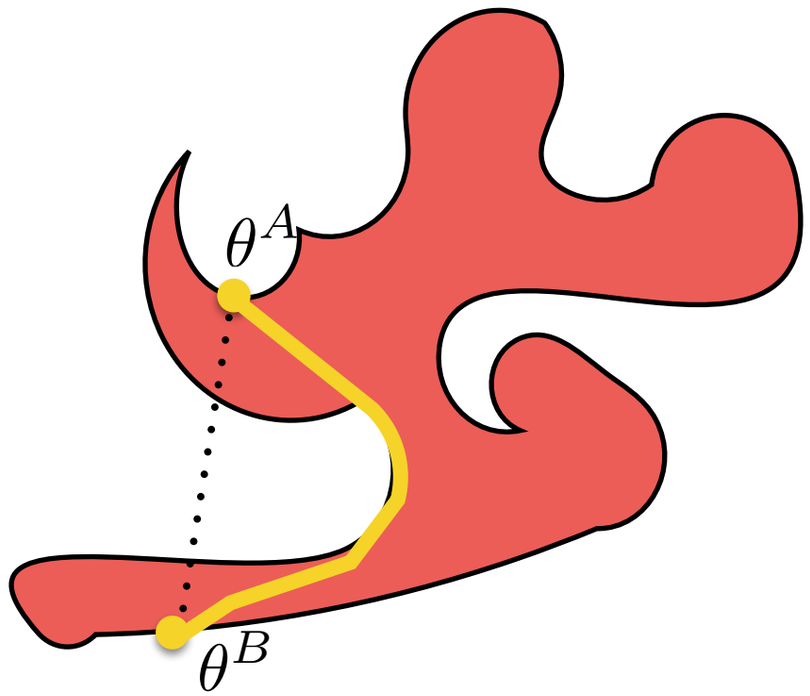
hard to move from one energy level to lower one

FROM TOPOLOGY TO GEOMETRY

- The next question we are interested in is conditioning for descent.
- Even if level sets are connected, how easy it is to navigate through them?
- We estimate level set geodesics and measure their length.

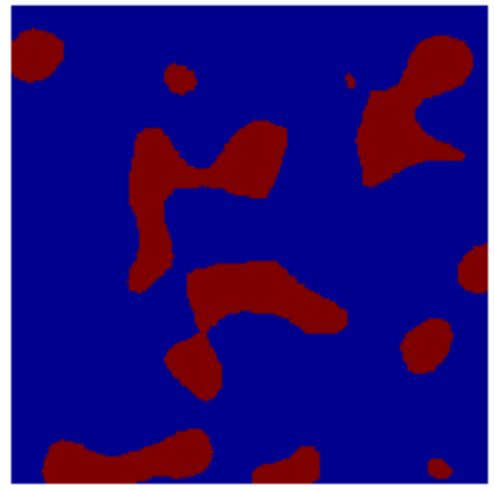


easy to move from one energy level to lower one



hard to move from one energy level to lower one

FINDING CONNECTED COMPONENTS



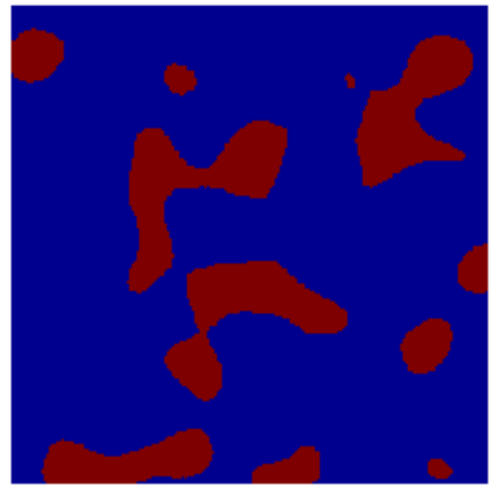
- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

    there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
    $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

- Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

FINDING CONNECTED COMPONENTS



- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

    there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
    $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

  - Moreover, we penalize the length of the path:

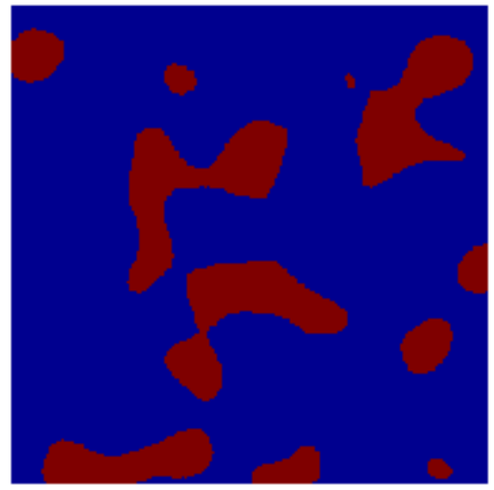
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

θ_1

θ_2

FINDING CONNECTED COMPONENTS



- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

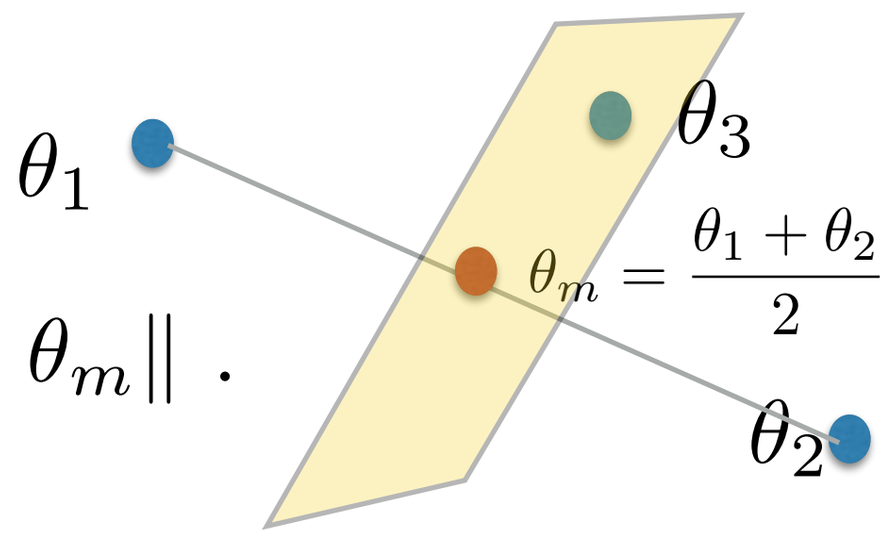
    there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

  - Moreover, we penalize the length of the path:

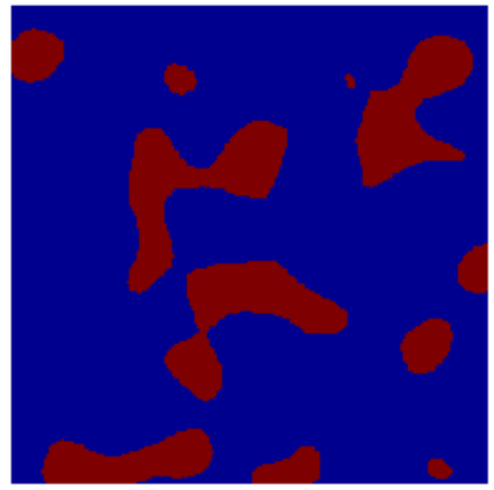
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int_{\mathcal{H}} \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\, E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$



FINDING CONNECTED COMPONENTS



- Suppose θ_1, θ_2 are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

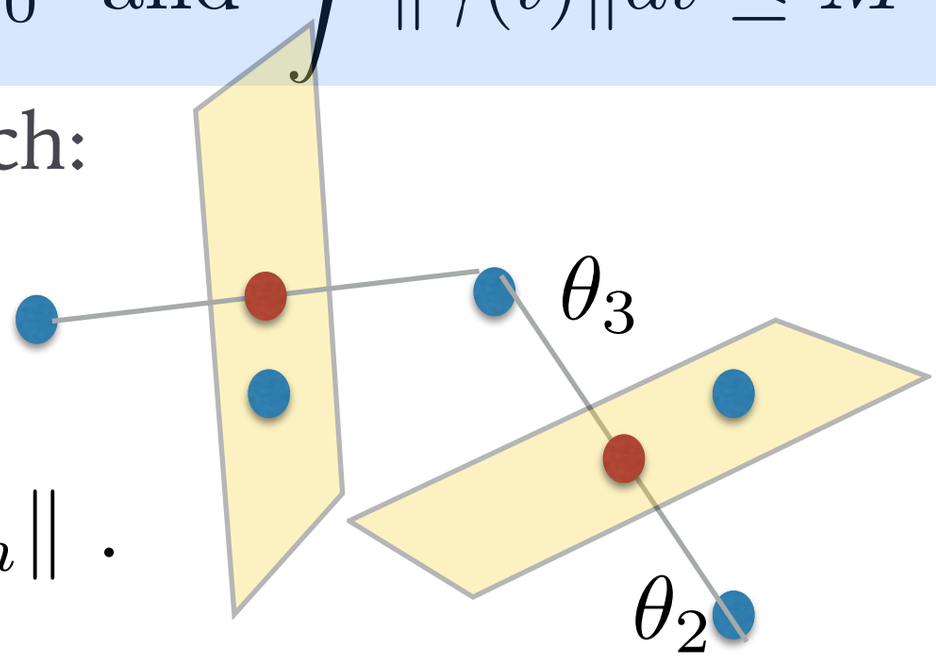
- Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

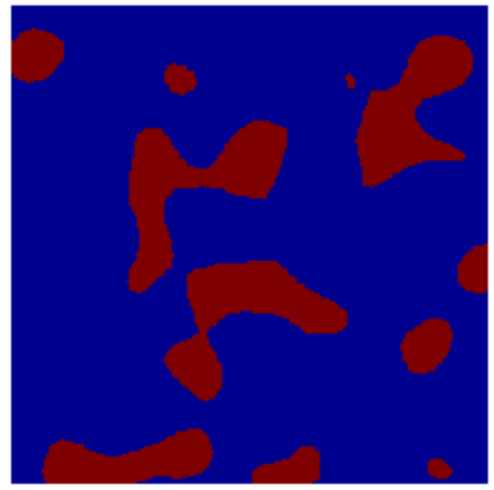
- Dynamic programming approach:

$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\, E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$



FINDING CONNECTED COMPONENTS



- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

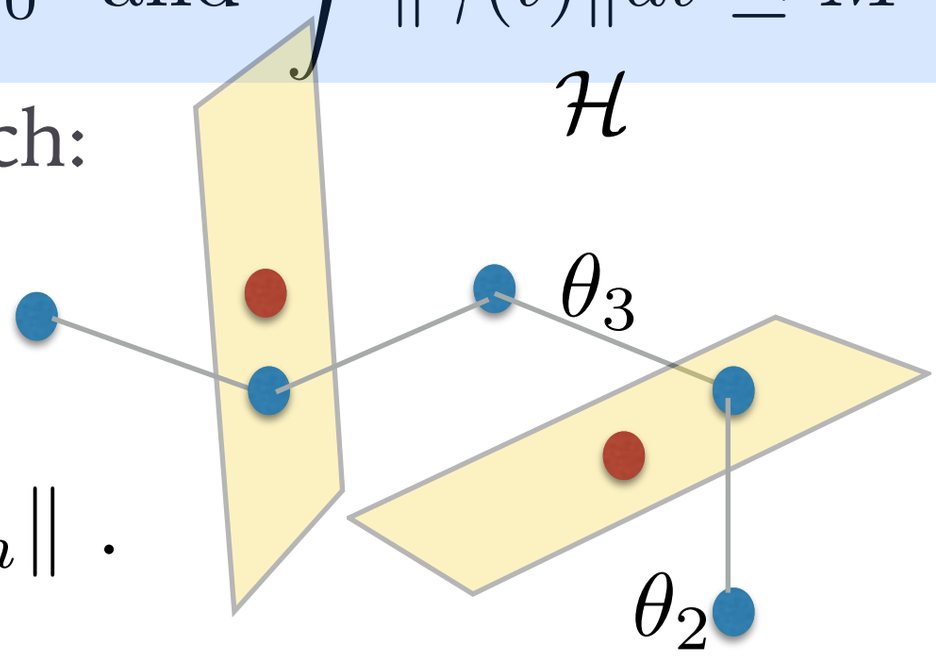
- Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \int_{\mathcal{H}} \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

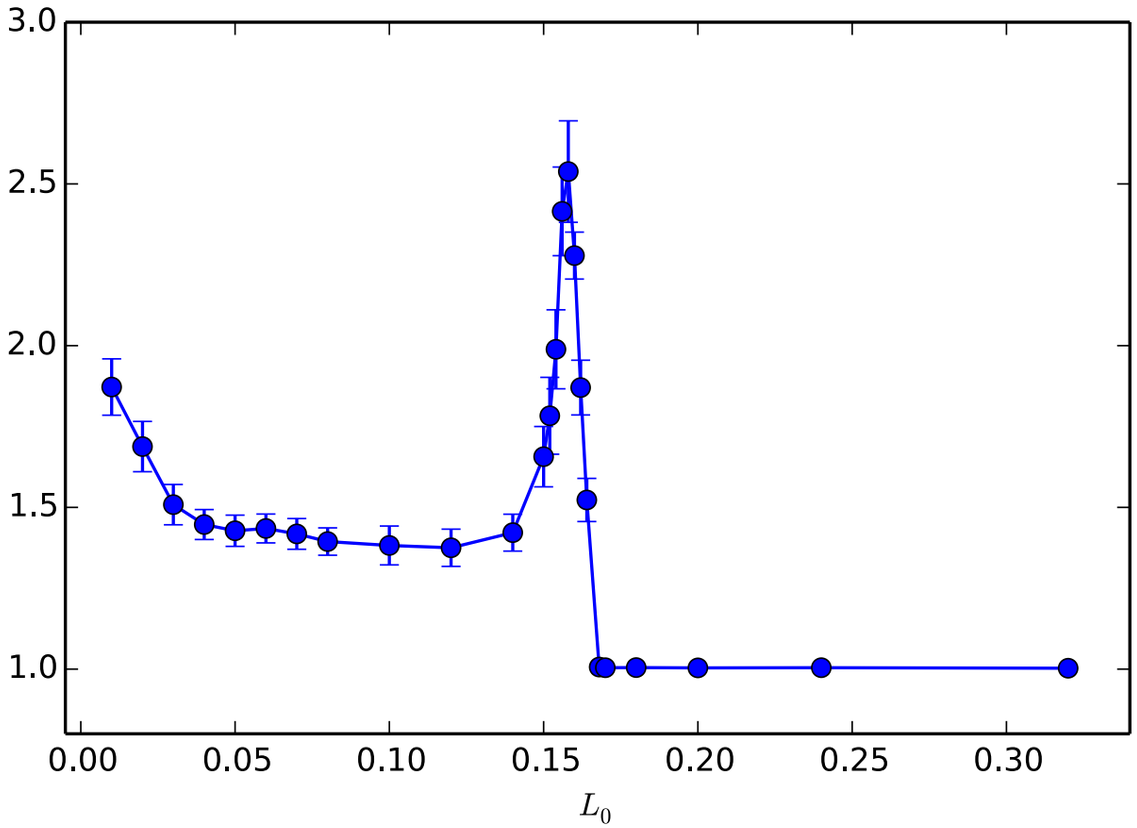
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\ E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$

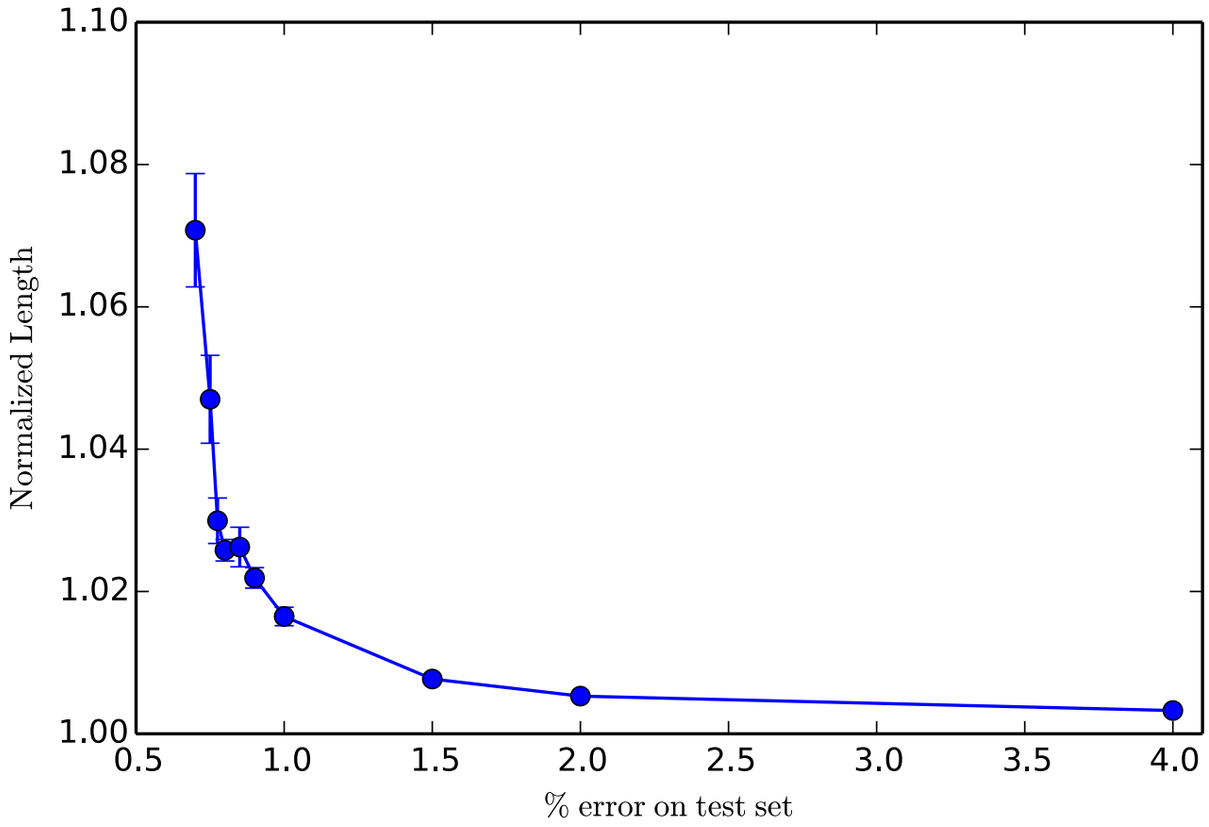


NUMERICAL EXPERIMENTS

- Compute length of geodesic in Ω_u obtained by the algorithm and normalize it by the Euclidean distance. Measure of curviness of level sets.



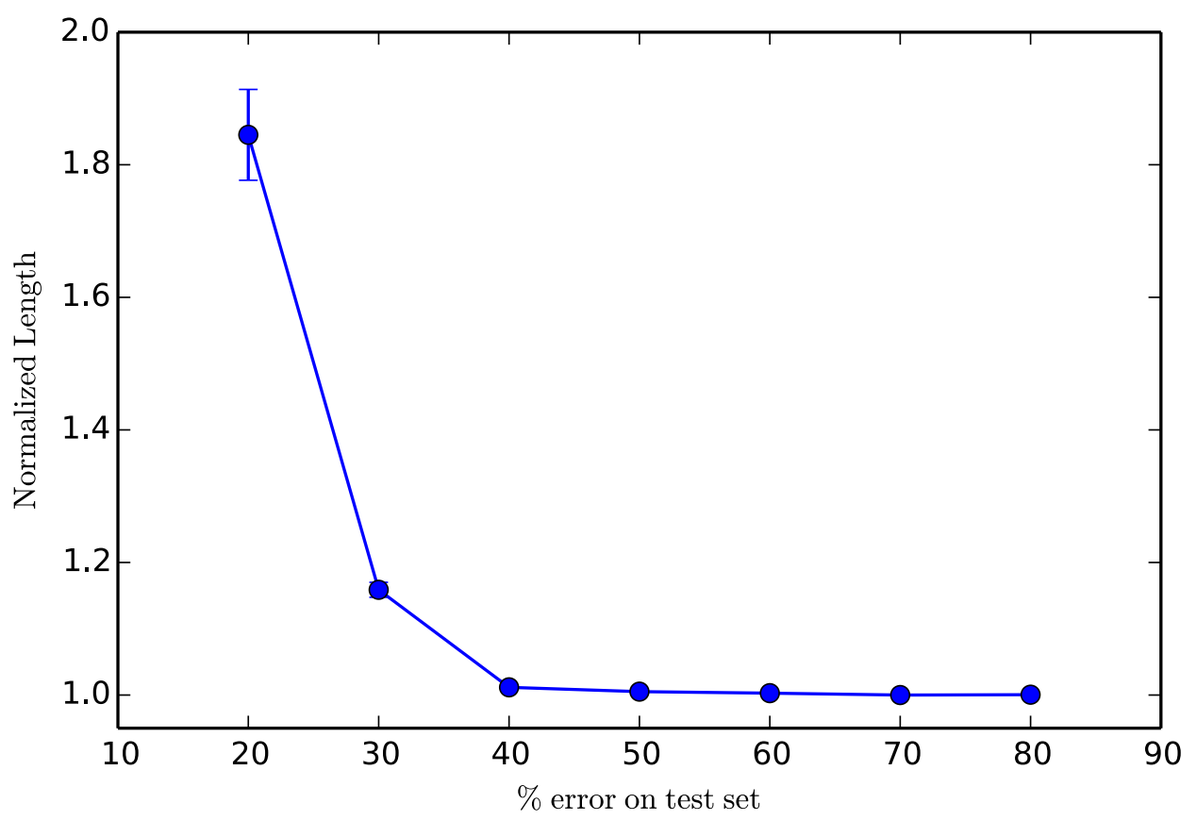
cubic polynomial



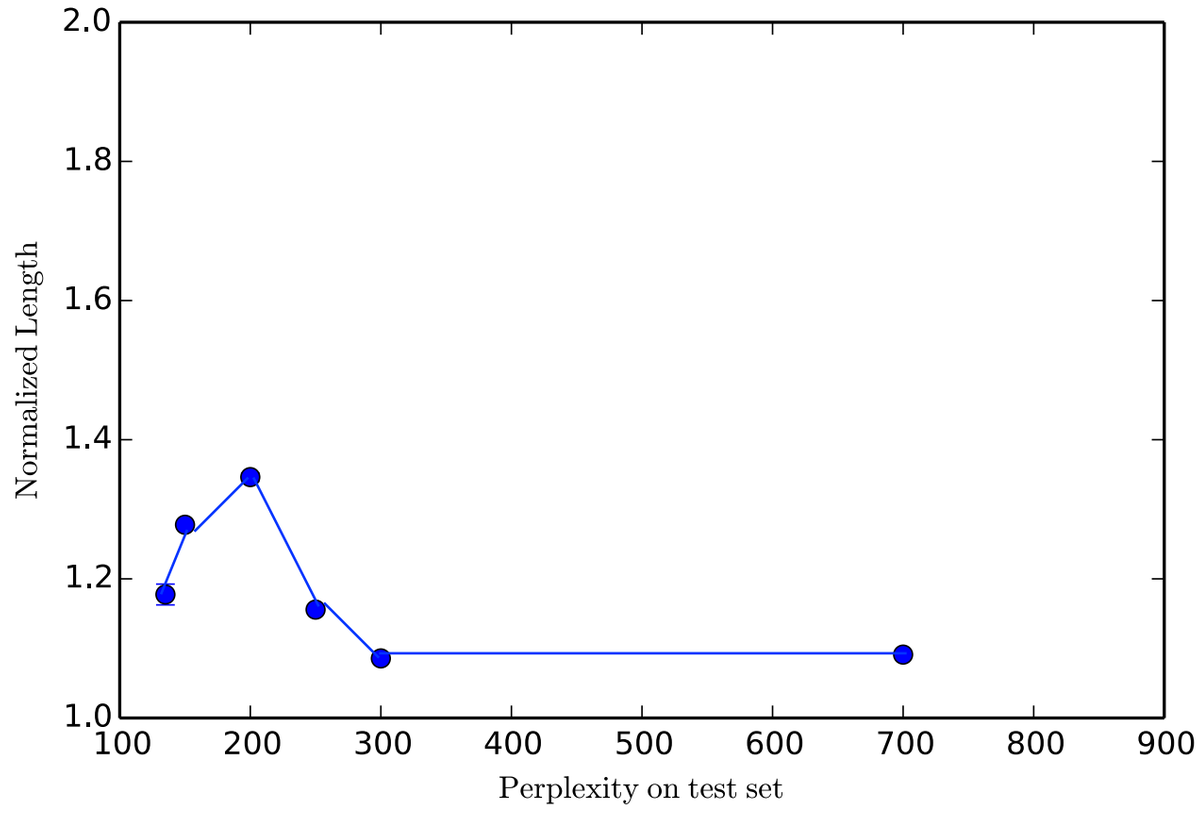
CNN/MNIST

NUMERICAL EXPERIMENTS

- Compute length of geodesic in Ω_u obtained by the algorithm and normalize it by the Euclidean distance. Measure of curviness of level sets.



CNN/CIFAR-10



LSTM/Penn

FOLLOW-UPS

- recent icml paper
-

ANALYSIS AND PERSPECTIVES

- #of components does not increase: no detected poor local minima so far when using typical datasets and typical architectures (at energy levels explored by SGD).
- Level sets become more irregular as energy decreases.
- Presence of “energy barrier”? extend to truncated Taylor?
- Kernels are back? CNN RKHS
- Open: “sweet spot” between overparametrisation and overfitting?
- Open: Robustness to noise in specification of activation function. Connection with Stochastic Gradient descent?