ON THE OPTIMIZATION LANDSCAPE OF NEURAL NETWORKS

JOAN BRUNA , CIMS + CDS, NYU

in collaboration with D.Freeman (UC Berkeley), Luca Venturi & Afonso Bandeira (NYU)

MOTIVATION

- We consider the standard Empirical Risk Minimization setup:

$$\hat{E}(\Theta) = \mathbb{E}_{(X,Y)\sim\hat{P}}\ell(\Phi(X;\Theta),Y) + \mathcal{R}(\Theta)$$

$$E(\Theta) = \mathbb{E}_{(X,Y)\sim P}\ \ell(\Phi(X;\Theta),Y)\ .$$

$\ell(z)$ convex
$\mathcal{R}(\Theta)$: regularization

$$\hat{P} = \frac{1}{n}\sum_{i\leq}\delta_{(xi,y_i)}\ .$$

MOTIVATION

- We consider the standard Empirical Risk Minimization setup:

$$\hat{E}(\Theta) = \mathbb{E}_{(X,Y)\sim\hat{P}}\ell(\Phi(X;\Theta),Y) + \mathcal{R}(\Theta)$$

$$E(\Theta) = \mathbb{E}_{(X,Y)\sim P}\ \ell(\Phi(X;\Theta),Y)\ .$$

$\ell(z)$ convex

$\mathcal{R}(\Theta)$: regularization

$$\hat{P} = \frac{1}{L}\sum_{l\leq L}\delta_{(x_l,y_l)}$$

- Population loss decomposition (*aka* "fundamental theorem of ML"):

$$E(\Theta^*) = \underbrace{\hat{E}(\Theta^*)}_{\text{training error}} + \underbrace{E(\Theta^*) - \hat{E}(\Theta^*)}_{\text{generalization gap}}\ .$$

- Long history of techniques to provably control generalization error via appropriate regularization.
- Generalization error and optimization are entangled [Bottou & Bousquet]

MOTIVATION

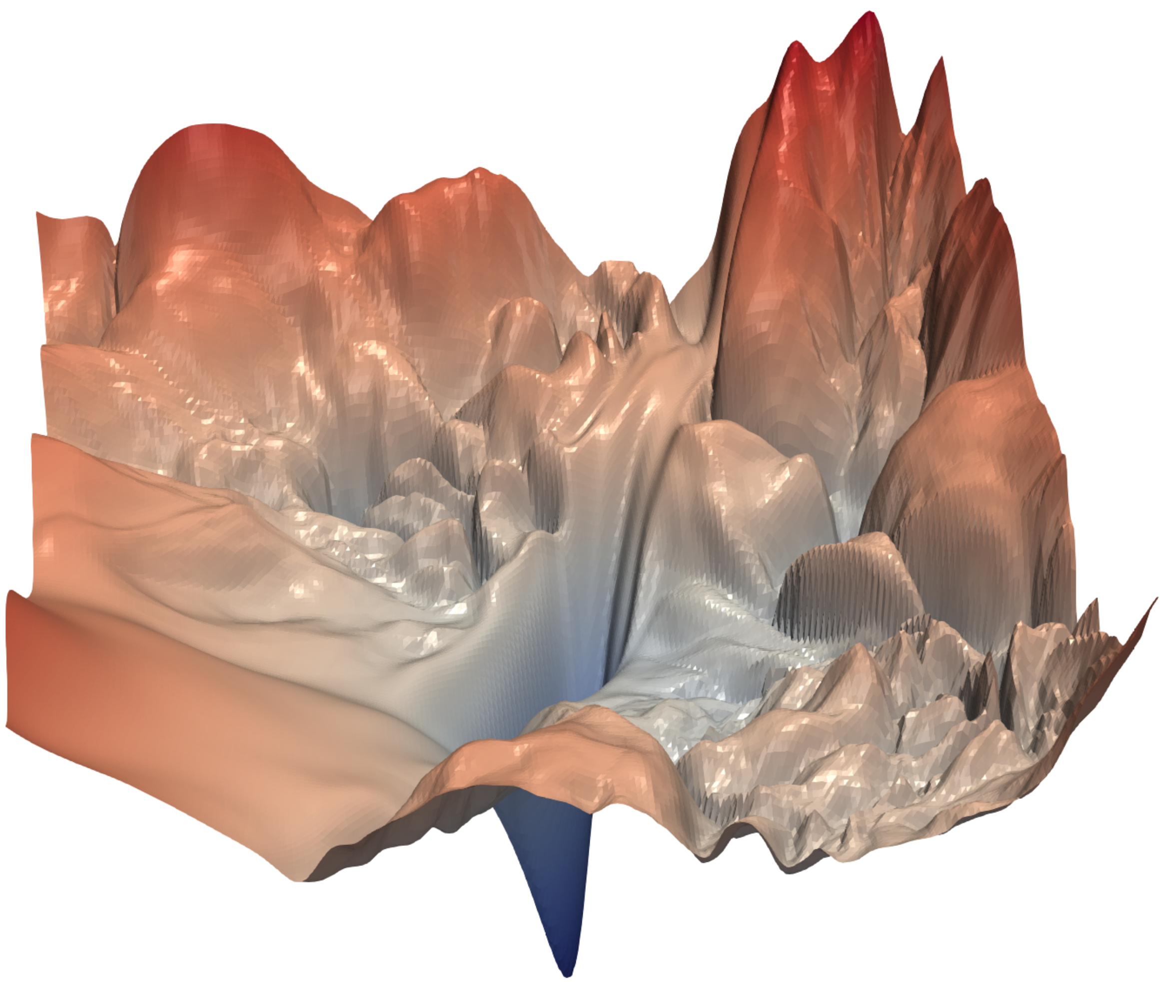
- However, when $\Phi(X;\Theta)$ is a large, deep network, current best mechanism to control generalization gap has two key ingredients:
  - Stochastic Optimization
    - *"During training, it adds the sampling noise that corresponds to empirical-population mismatch"* [Léon Bottou].
  - Make the model *convolutional* and *very large*.
    - see e.g. "Understanding Deep Learning Requires Rethinking Generalization", [Ch. Zhang *et al*, ICLR'17].

MOTIVATION

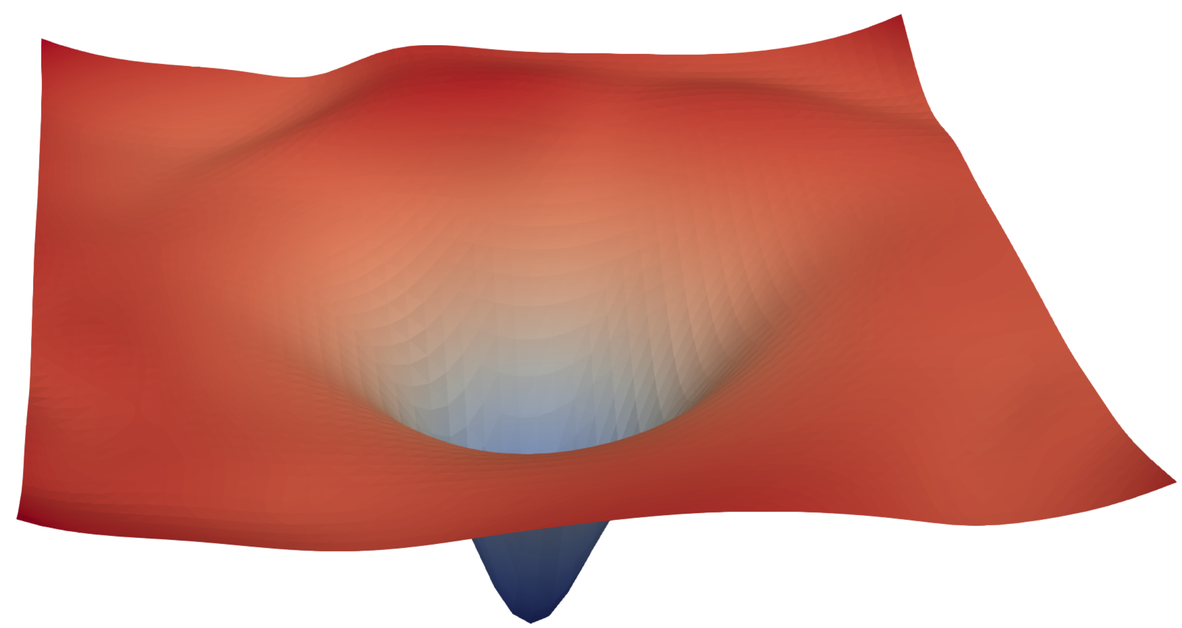
- However, when $\Phi(X;\Theta)$ is a large, deep network, current best mechanism to control generalization gap has two key ingredients:
  - Stochastic Optimization
  - Make the model *convolutional* and *as large as possible*.
- We first address how *overparametrization* affects the energy landscapes.
- **Goal 1**: Study simple *topological* properties of these landscapes $E(\Theta), \hat{E}(\Theta)$for half-rectified neural networks.
- **Goal 2**: Estimate simple *geometric* properties with efficient, scalable algorithms. Diagnostic tool.

OUTLINE

- Topology of Neural Network Energy Landscapes
- Geometry of Neural Network Energy Landscapes



(a) without skip connections



(b) with skip connections

[Li et al.'17]

PRIOR 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]

PRIOR RELATED WORK

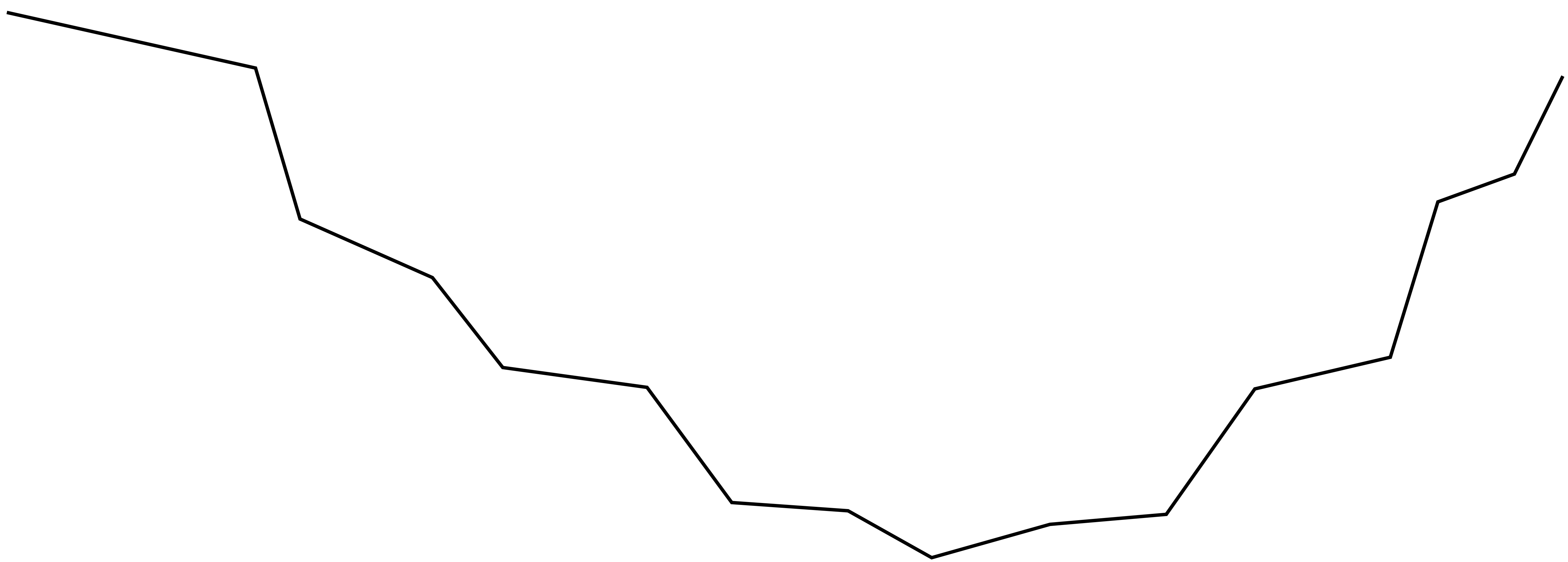
- Models from Statistical physics have been considered as possible approximations [Dauphin et al.'14, Choromanska et al.'15, Segun et al.'15]
- Tensor factorization models capture some of the non convexity essence [Anandukar et al'15, Cohen et al. '15, Haeffele et al.'15]
- [Shafran and Shamir,'15] studies 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.

PRIOR 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

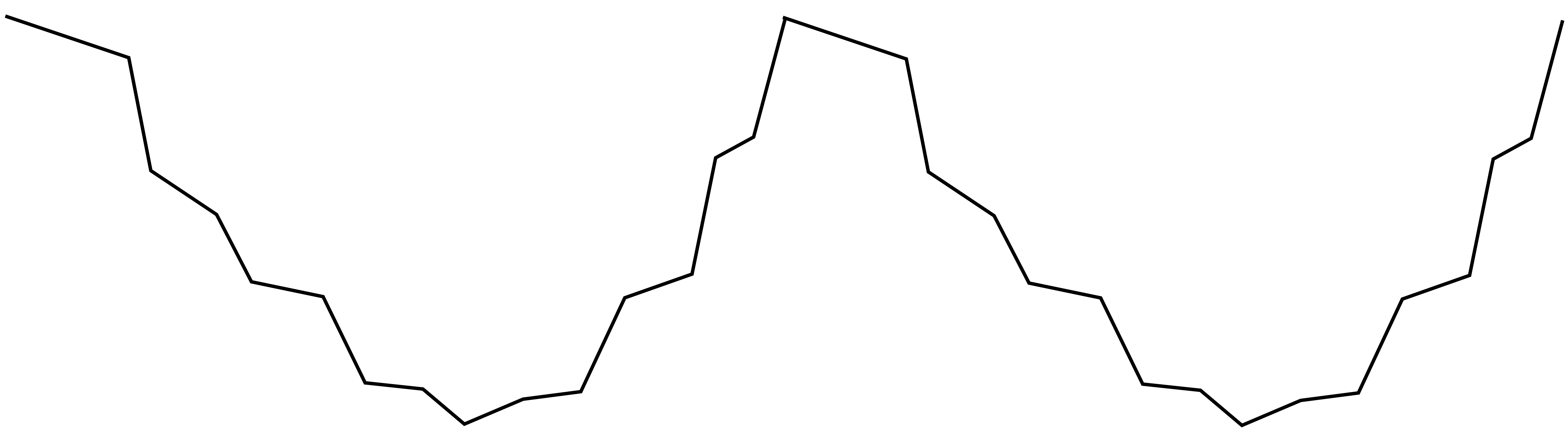
- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.



NON-CONVEXITY ≠ NOT OPTIMIZABLE

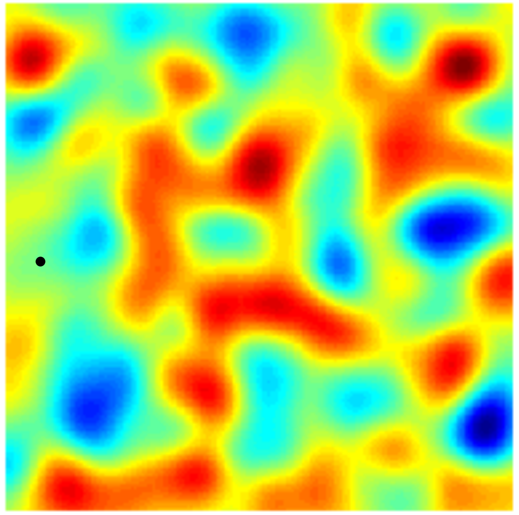
- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.
- In particular, deep models have internal symmetries.

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



ANALYSIS OF NON-CONVEX LOSS SURFACES

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

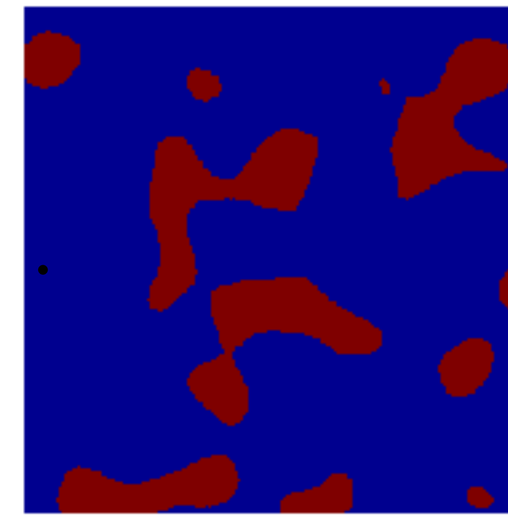


$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \le u\} \ .$$

ANALYSIS OF NON-CONVEX LOSS SURFACES

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

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



Ω_u

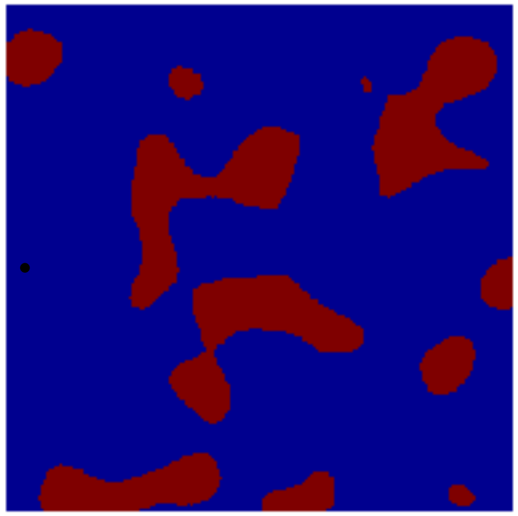
- A first notion we address is about the topology of the level sets Ω_u.

- 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:

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



Ω_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. $E(y^*) > \min_y E(y)$)

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:

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

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

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

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

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

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

Proposition: [BF'16]

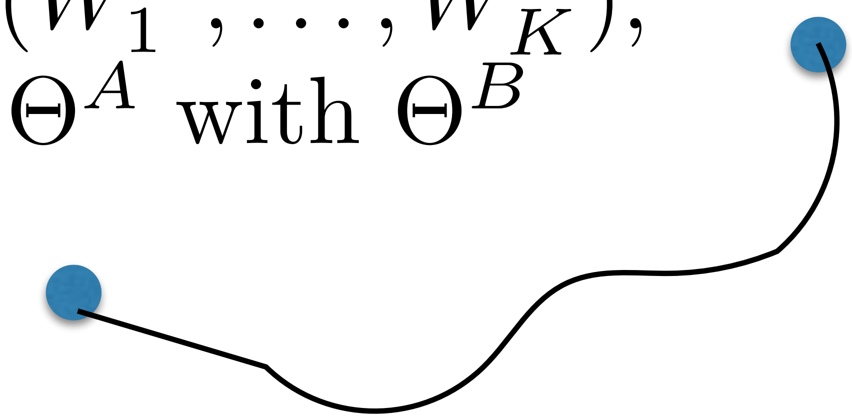
1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.
2. (2-layer case, ridge regression) $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$ satisfies $N_u = 1 \ \forall \ u$ if $n_1 > \min(n, m)$.

- 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

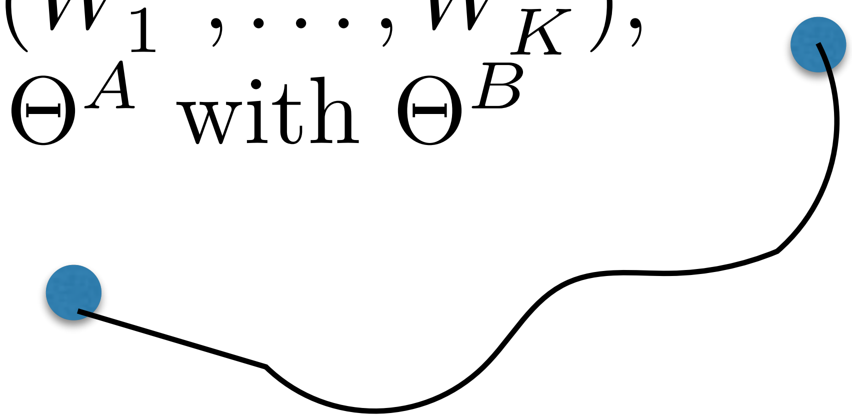
- Goal:
Given $\Theta^A = (W_1^A, \ldots, W_K^A)$ and $\Theta^B = (W_1^B, \ldots, W_K^B)$, we construct a path $\gamma(t)$ that connects Θ^A with Θ^B st $E(\gamma(t)) \leq \max(E(\Theta^A), E(\Theta^B))$.



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 $n_k > \min(n, m)$

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

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

Theorem [LBB'17]: The Multilinear regression $\mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2$ has no poor local minima.

  - 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 [LBB'17]: The two-layer quadratic network optimization $L(U, W) = \mathbb{E}_{(X,Y)\sim P} \|U(WX)^2 - Y\|^2$ has no poor local minima if $M \geq 2N$.

ASYMPTOTIC CONNECTEDNESS OF RELU

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

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

ASYMPTOTIC CONNECTEDNESS OF RELU

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

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

> **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 $\epsilon \sim m^{-\frac{1}{n}}$.

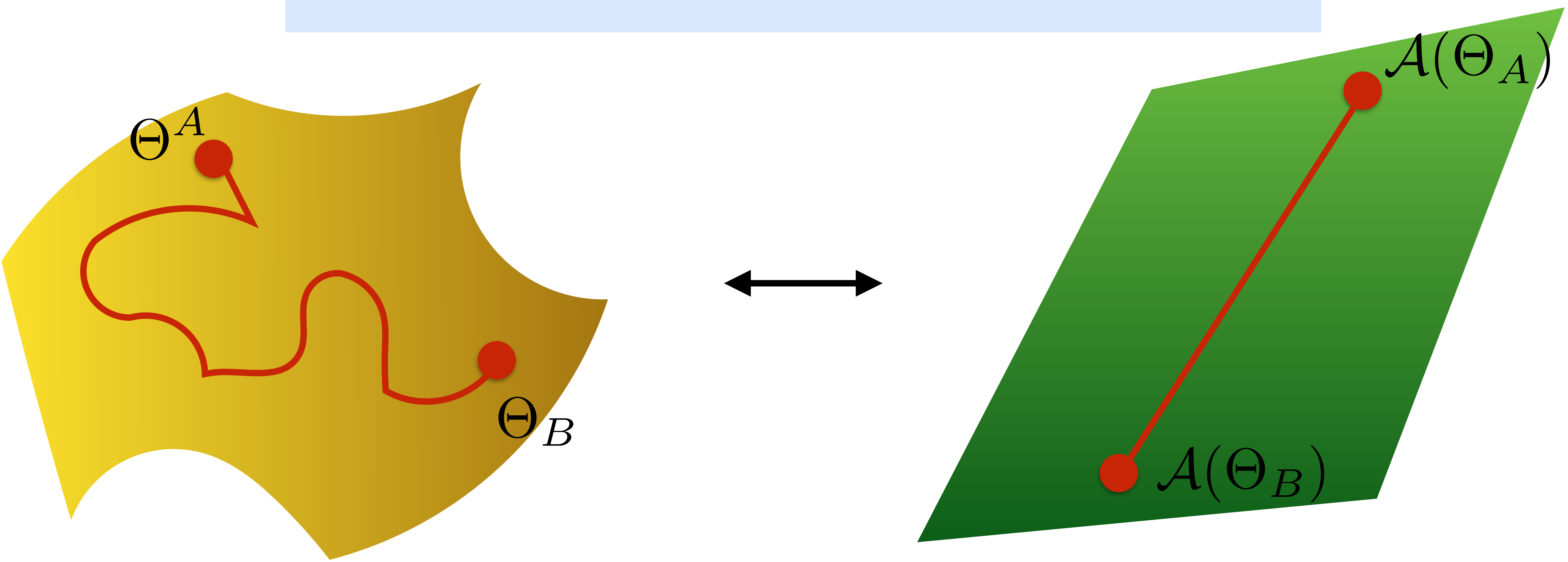
- Overparametrisation "wipes-out" local minima (and group symmetries).
- The bound is cursed by dimensionality, ie exponential in n .
- Result is based on local linearization of the ReLU kernel (hence exponential price).

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 $\beta = \mathcal{A}(\Theta)$

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

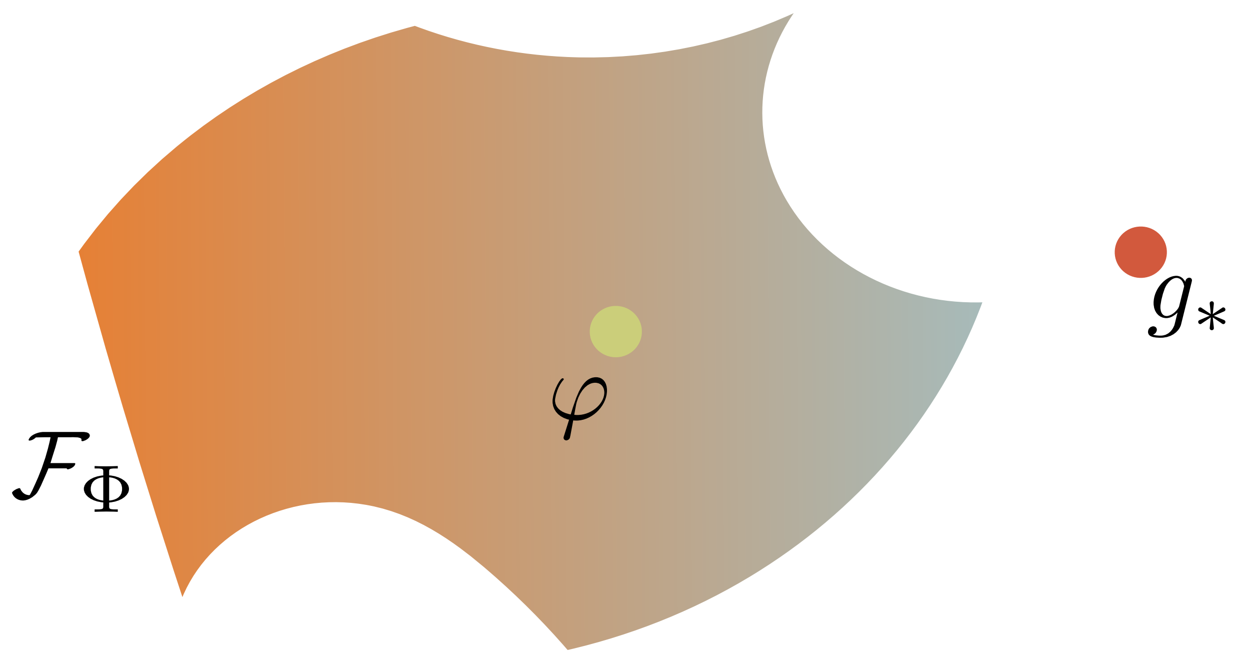
Corollary: [BBV'17] If $\dim\{\mathcal{A}(w), w \in \mathbb{R}^n\} = q < \infty$ and $M \geq 2q$, then $E(W,U) = \mathbb{E}|U\rho(WX) - Y|^2$, $W \in \mathbb{R}^{M \times N}$ has no poor local minima if $M \geq 2q$.

- This includes Empirical Risk Minimization (since RKHS is only queried on finite # of datapoints).
- 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:

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



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

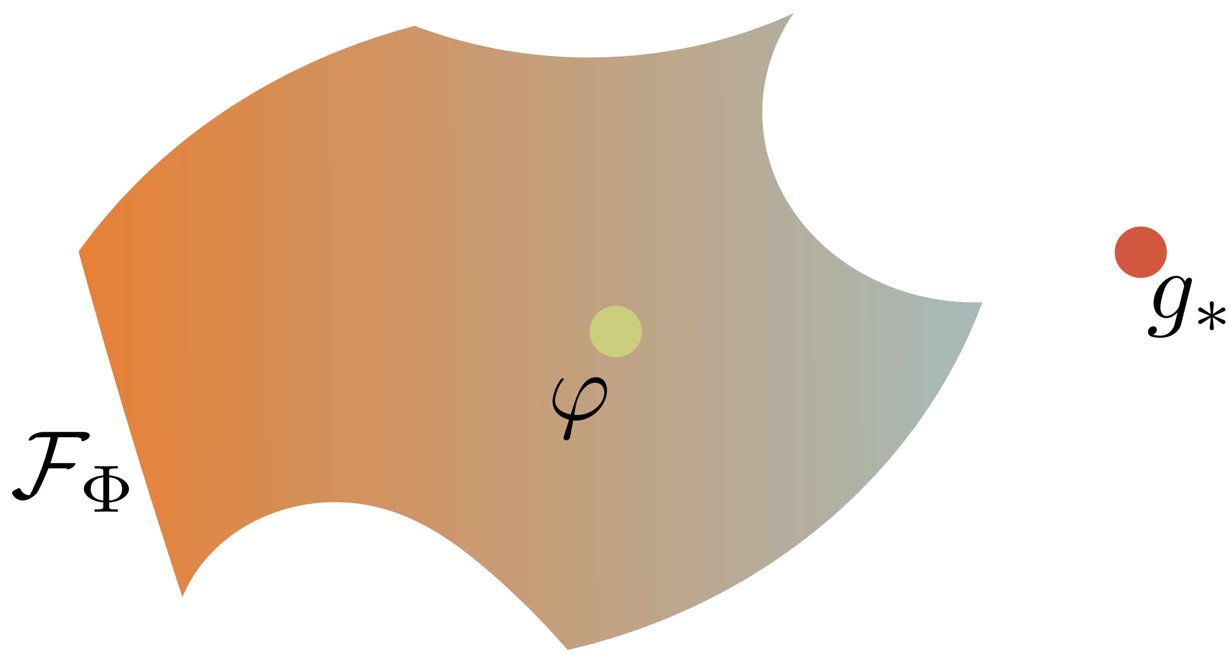
$$g_* : x \mapsto \mathbb{E}(Y|x)$$

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

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.
- What happens when the model is not overparametrised?

FROM SIMPLE LANDSCAPES TO ENERGY BARRIER

- The energy landscape of several prototypical models in statistical physics exhibit a so-called *energy barrier,* e.g. spherical spin glasses:

$$H_{N,p}(\sigma) = N^{-(p-1)/2} \sum_{i_1,\ldots i_p=1}^{N} J_{i_1,\ldots,i_p}\sigma_{i_1}\cdots\sigma_{i_p} \ , \ \sigma \in S^{N-1}(\sqrt{N}) \ , J_i \sim \mathcal{N}(0,1).$$

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

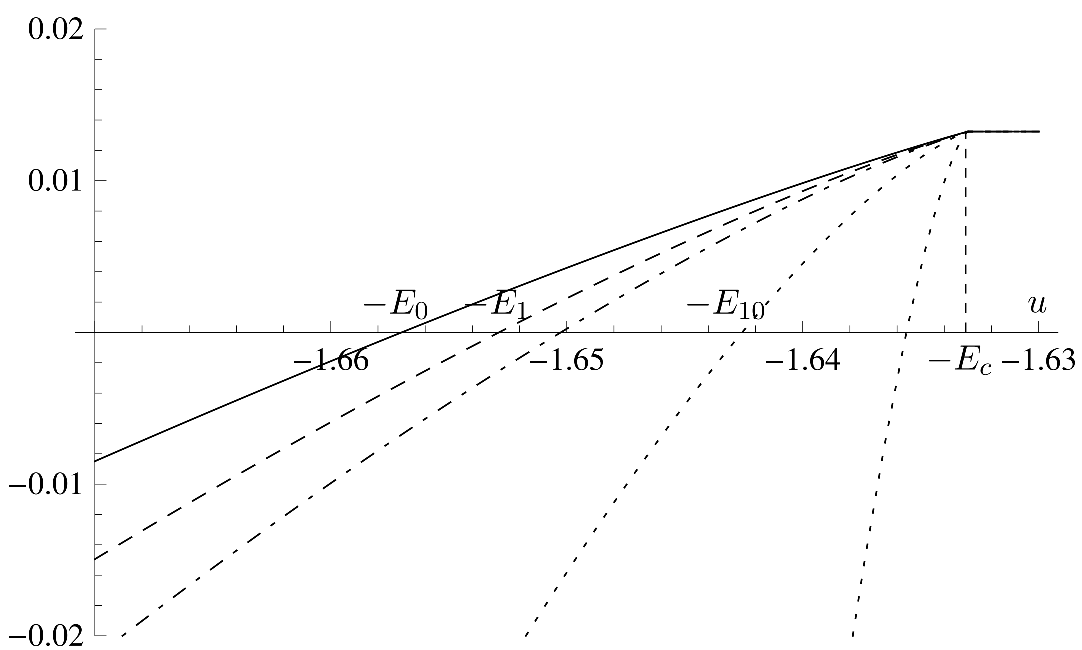


FIGURE 1. The functions $\Theta_{k,p}$ for $p = 3$ and $k = 0$ (solid), $k = 1$ (dashed), $k = 2$ (dash-dotted), $k = 10$, $k = 100$ (both dotted). All these functions agree for $u \geq -E_\infty$.

[Auffinger, Ben Arous Cerny,'11]

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

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

  - 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

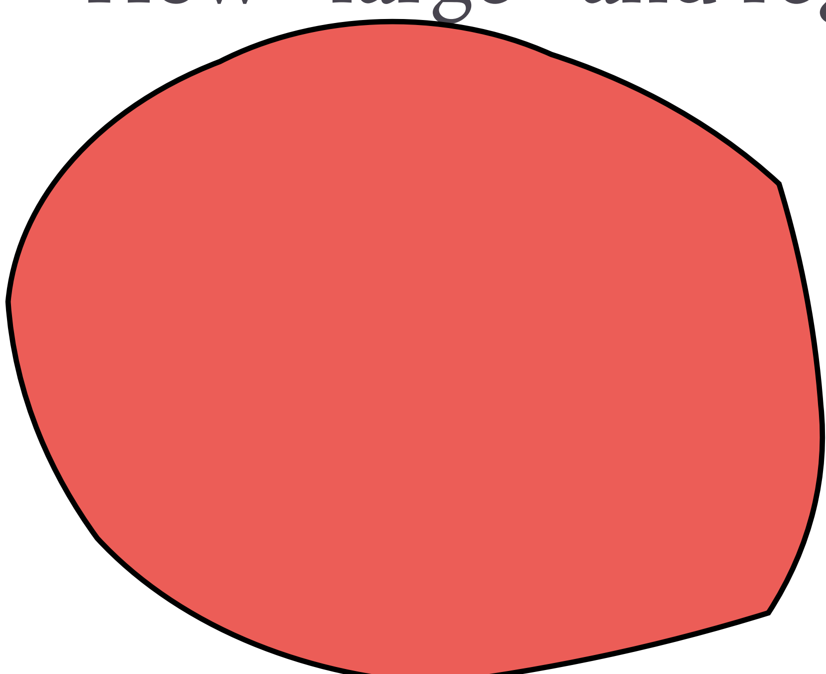
$$\beta(M,N) = \inf_{S; \dim(S)=f^{-1}(M)} \inf_{\substack{U \in \mathbb{R}^{m \times M} \\ W \in \mathbb{R}^{M \times f^{-1}(M)}}} \sup_{\substack{\mathbb{E}\|Z\| \leq N - f^{-1}(M), \\ P_S Z = 0}} \mathbb{E}\|U\rho(WP_S X + Z) - Y\|^2$$

  - 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?

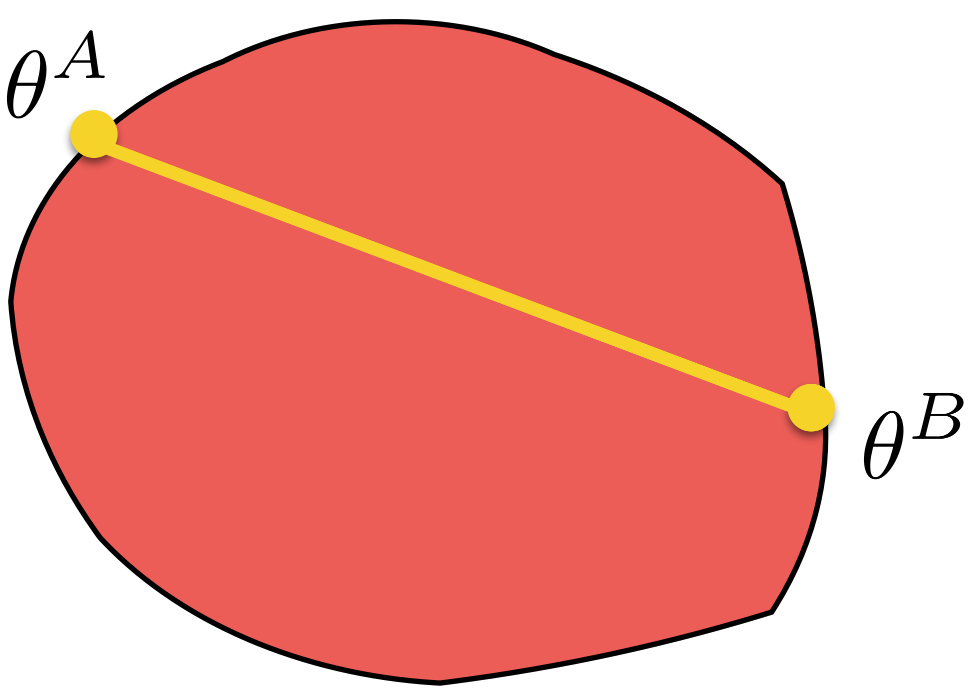


easy to move from one energy level to lower one

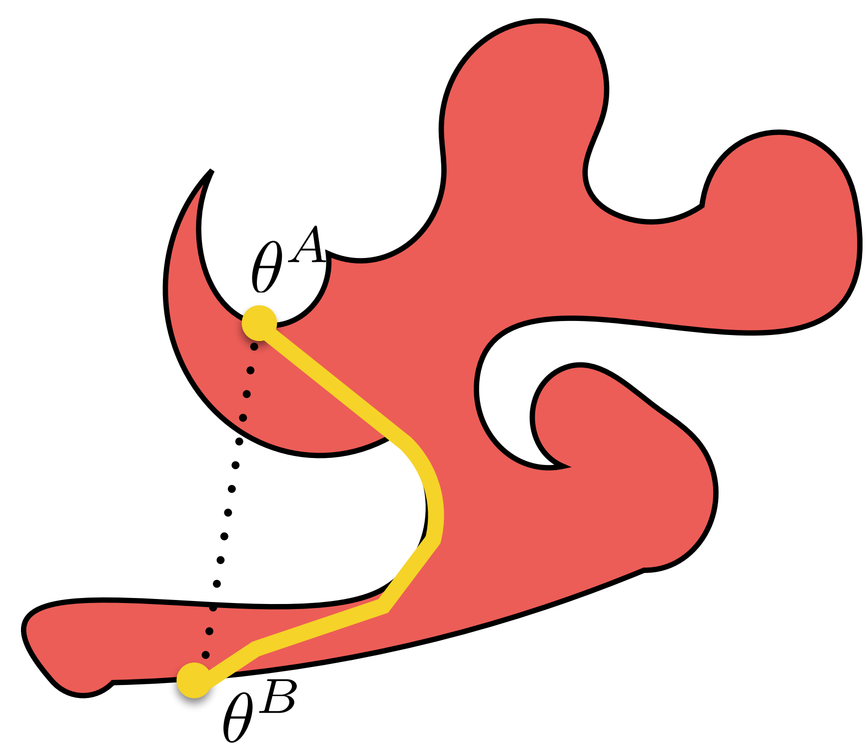
hard to move from one energy level to lower one

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.

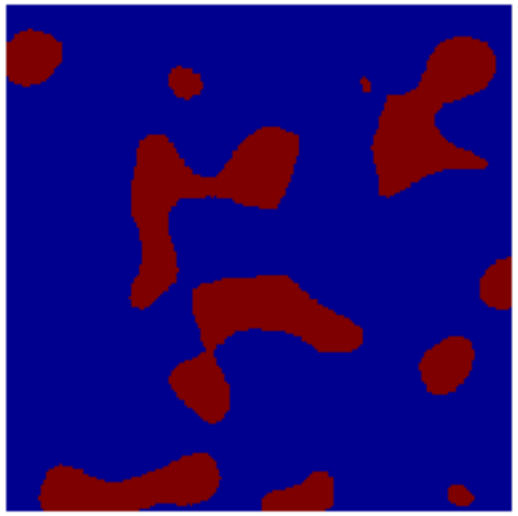


easy to move from one energy level to lower one



hard to move from one energy level to lower one

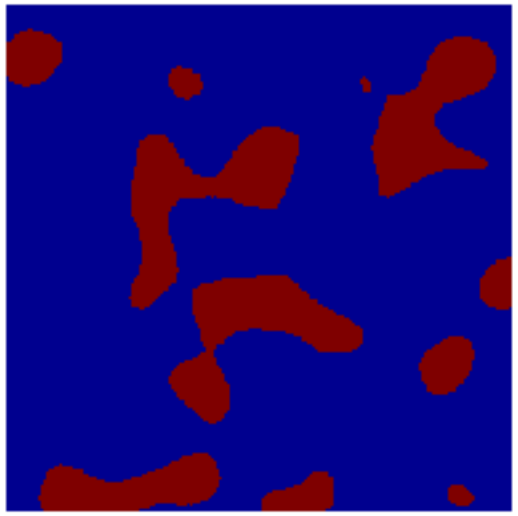
FINDING CONNECTED COMPONENTS



- 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:

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

FINDING CONNECTED COMPONENTS



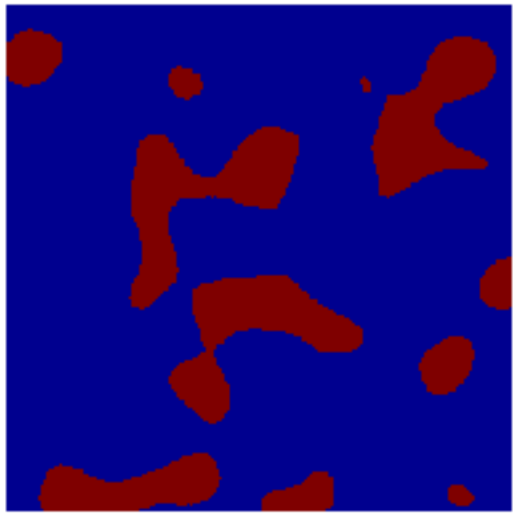
- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

    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:

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

- Dynamic programming approach:

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

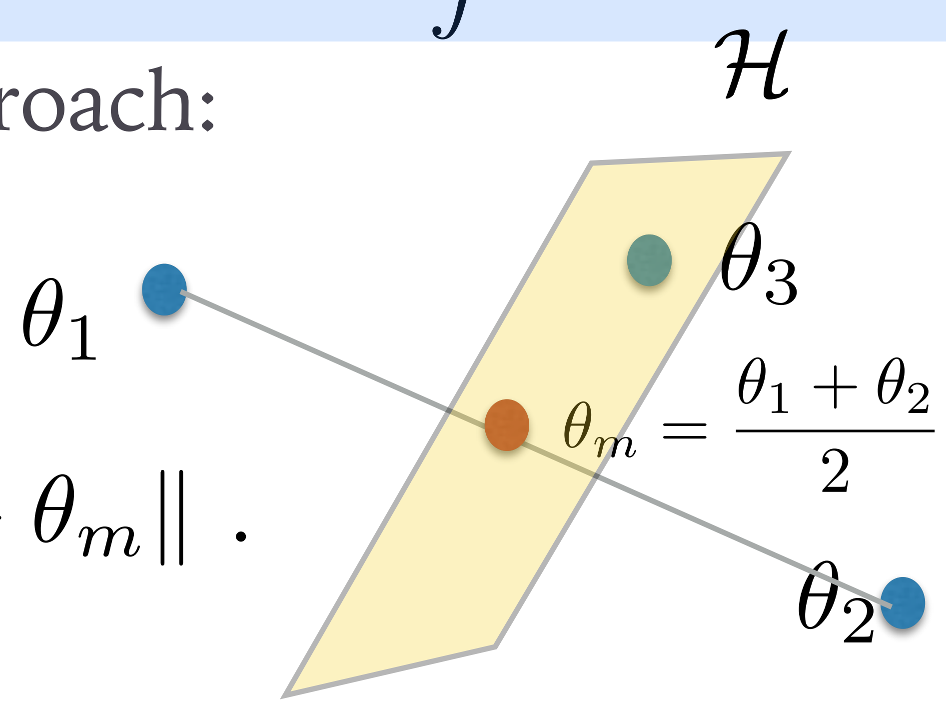
    there is a path $\gamma(t), \; \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
    $\forall \; t \in (0,1) \; , \; E(\gamma(t)) \leq u_0 \; .$

  - Moreover, we penalize the length of the path:

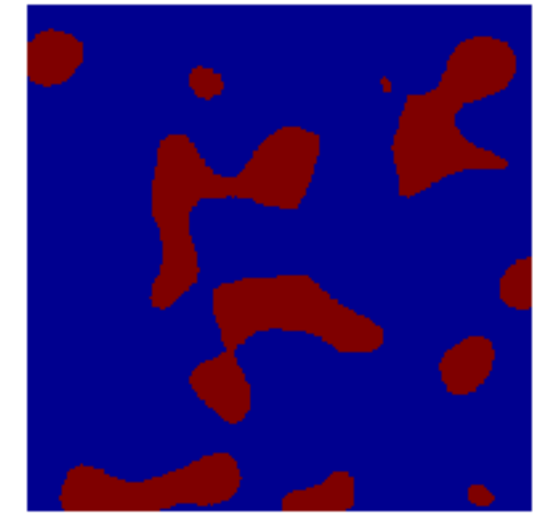
$$\forall \; t \in (0,1) \; , \; E(\gamma(t)) \leq u_0 \; \text{ and } \int \|\dot{\gamma}(t)\| dt \leq M \; .$$

- Dynamic programming approach:

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



FINDING CONNECTED COMPONENTS



- Suppose $\theta_1,\ \theta_2$ 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
    $$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$$

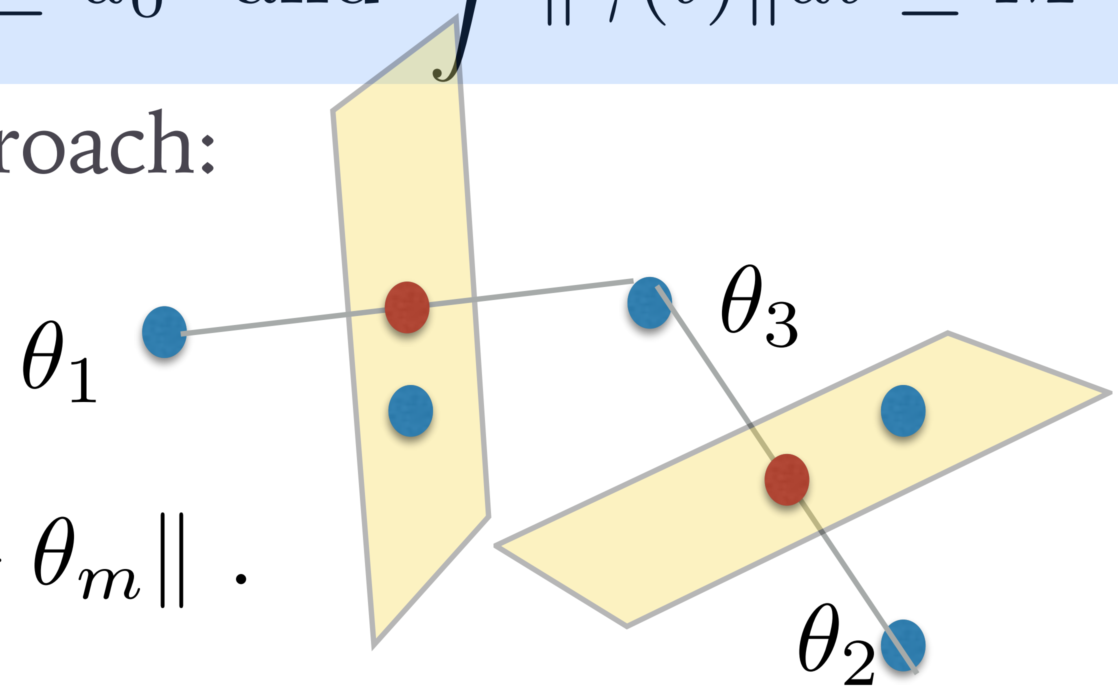
  - Moreover, we penalize the length of the path:

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

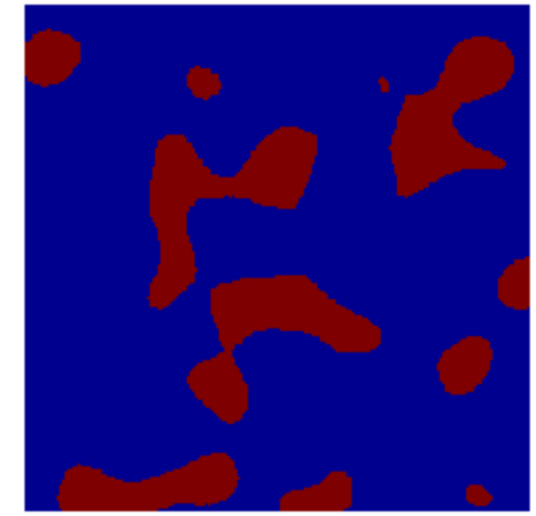
- Dynamic programming approach:

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

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



FINDING CONNECTED COMPONENTS



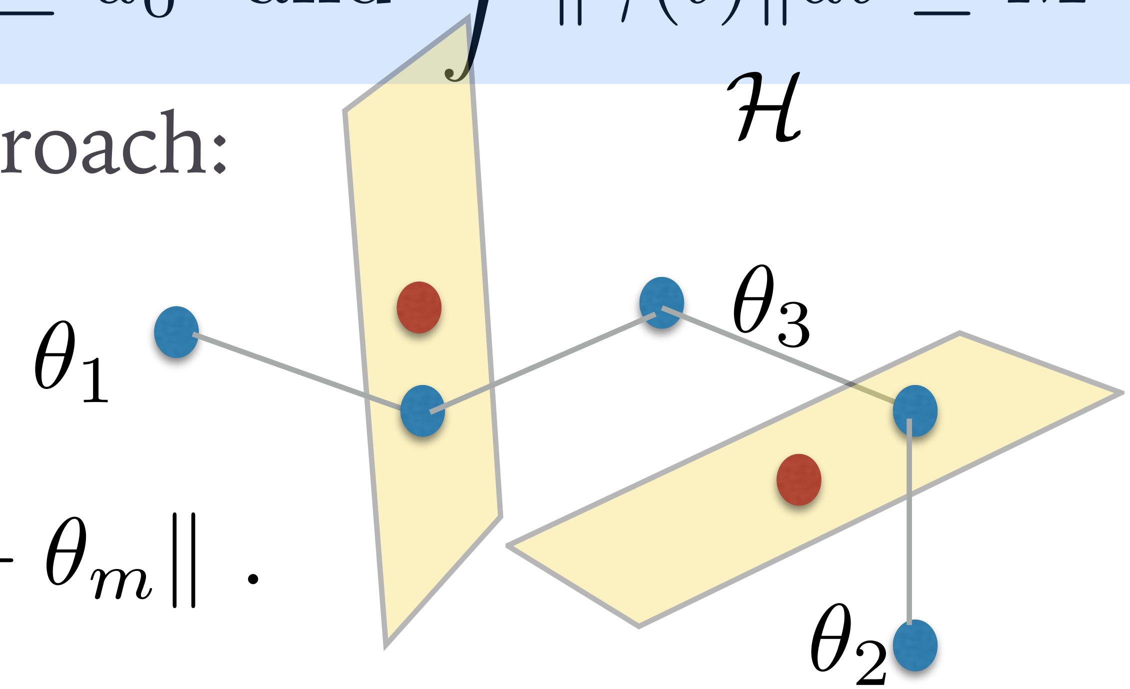
- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$
  - Moreover, we penalize the length of the path:

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

- Dynamic programming approach:

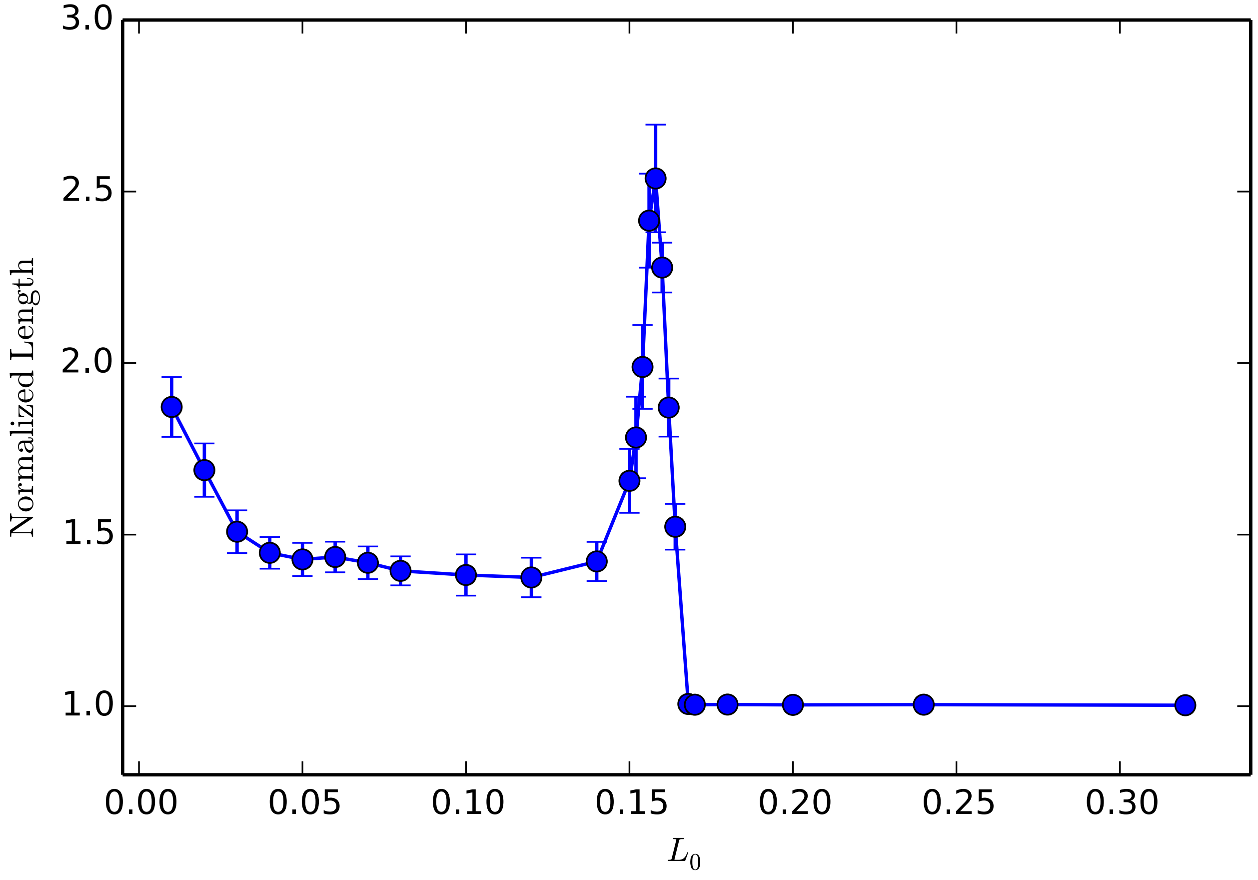
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

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

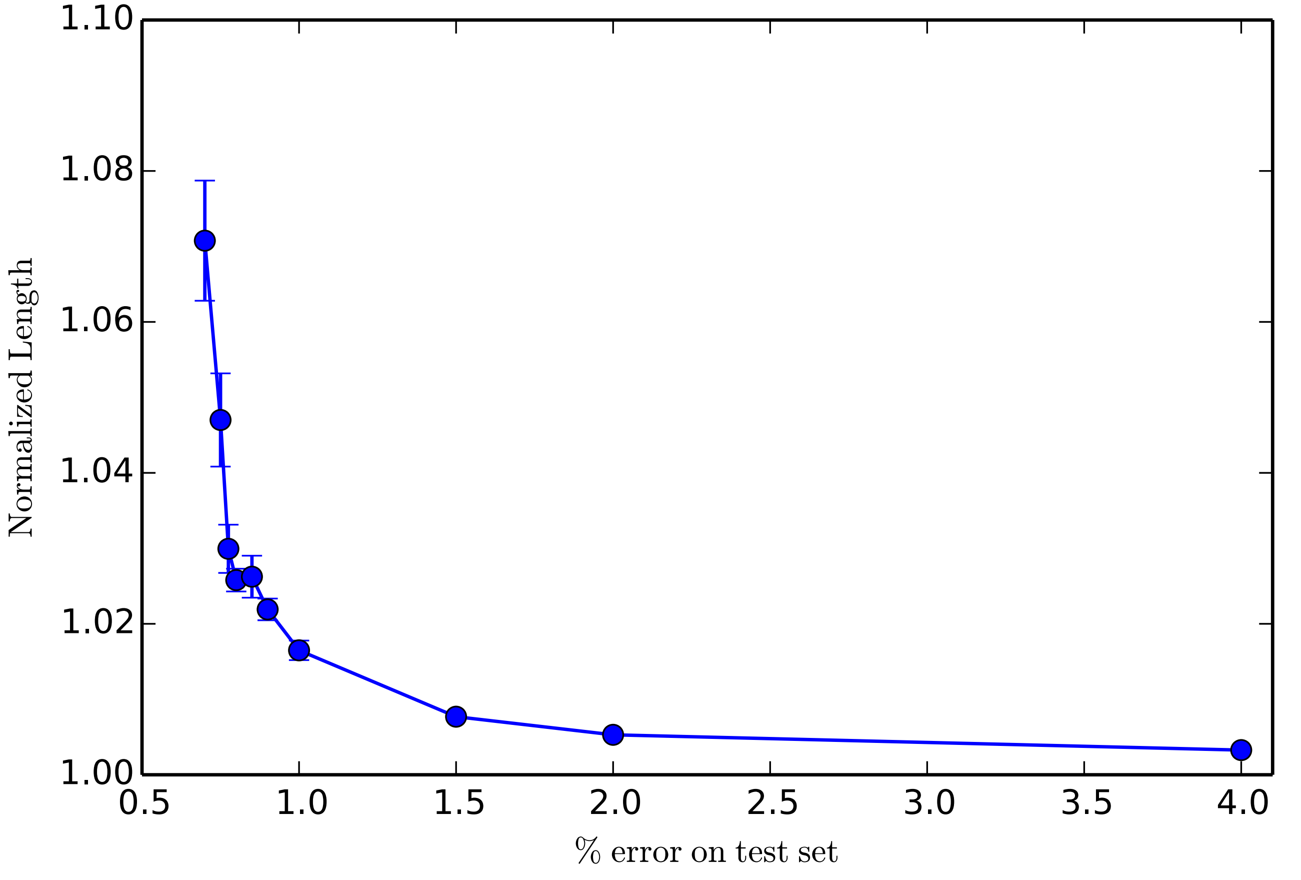


NUMERICAL EXPERIMENTS

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



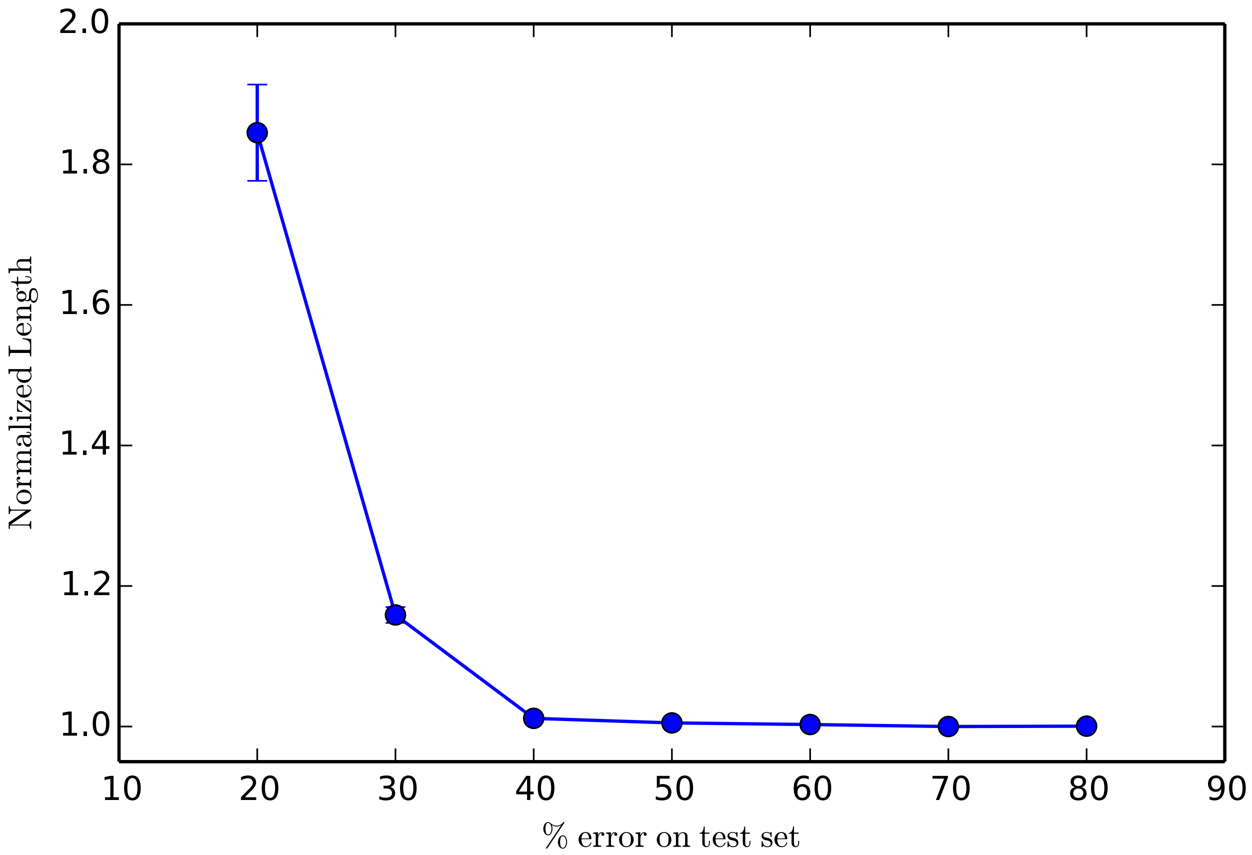
cubic polynomial



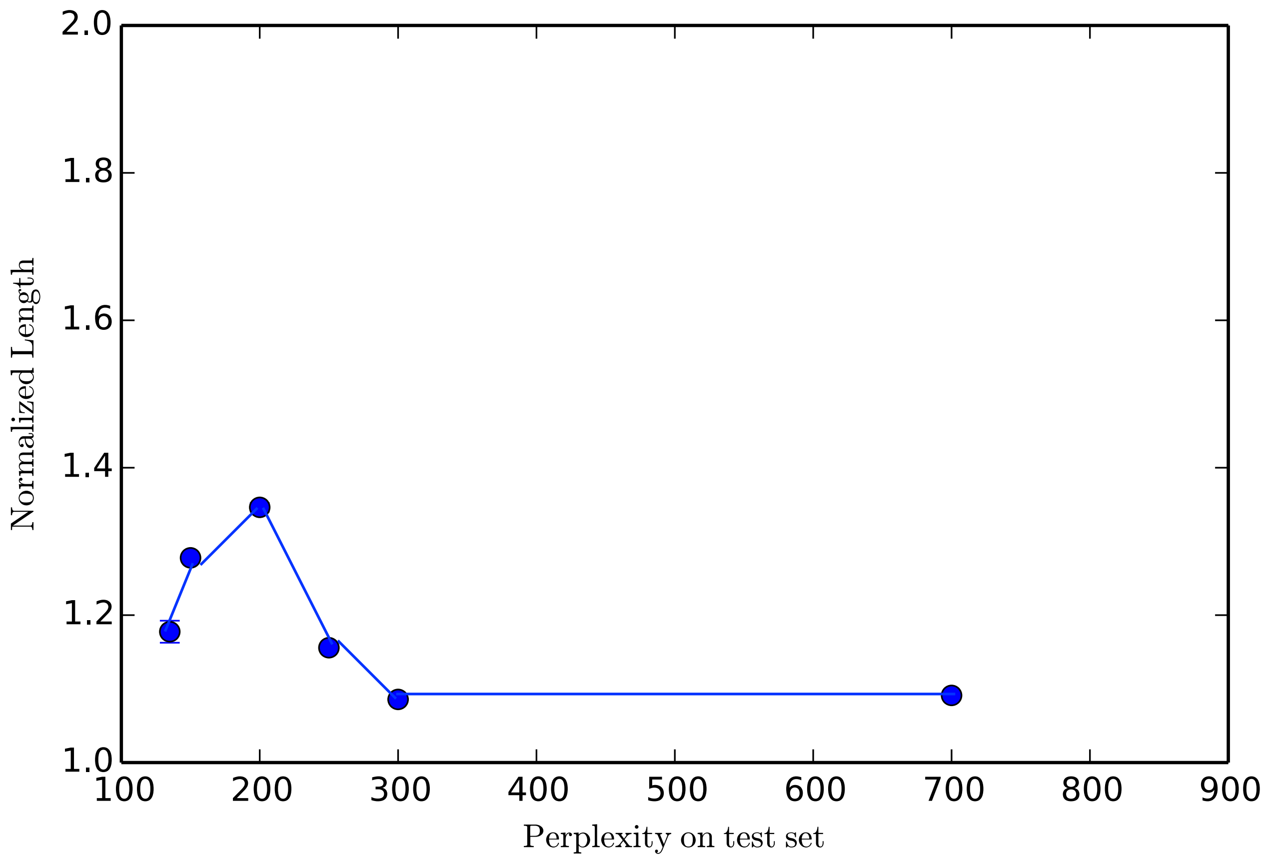
CNN/MNIST

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

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: Role of Stochastic Optimization in this story?

THANKS!