MAD

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

MATHEMATICS OF DEEP LEARNING

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

Lecture 13: Optimization Landscapes of deep networks. Open Problems

OBJECTIVES LECTURE 13

- Landscape of Optimization of Neural Networks

- Some Open Challenges

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 $n_1 > \min(n, m)$.

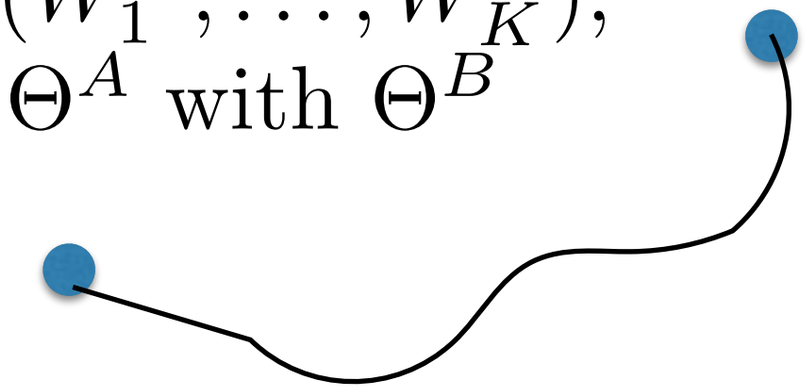
- We pay extra redundancy price to get simple topology.
- This simple topology is an "artifact" of the linearity of the network:

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

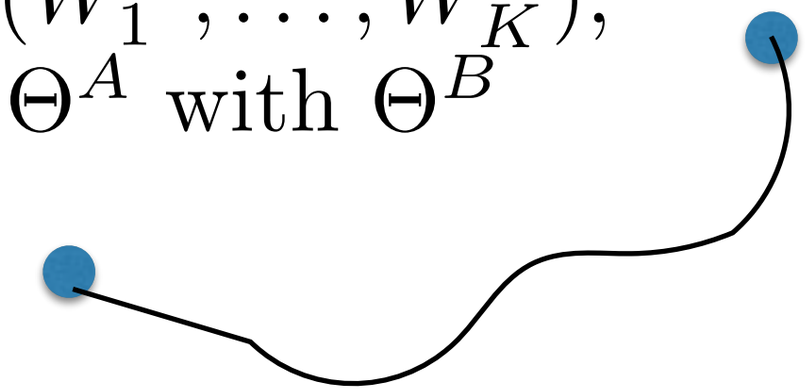
- Goal:

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 .

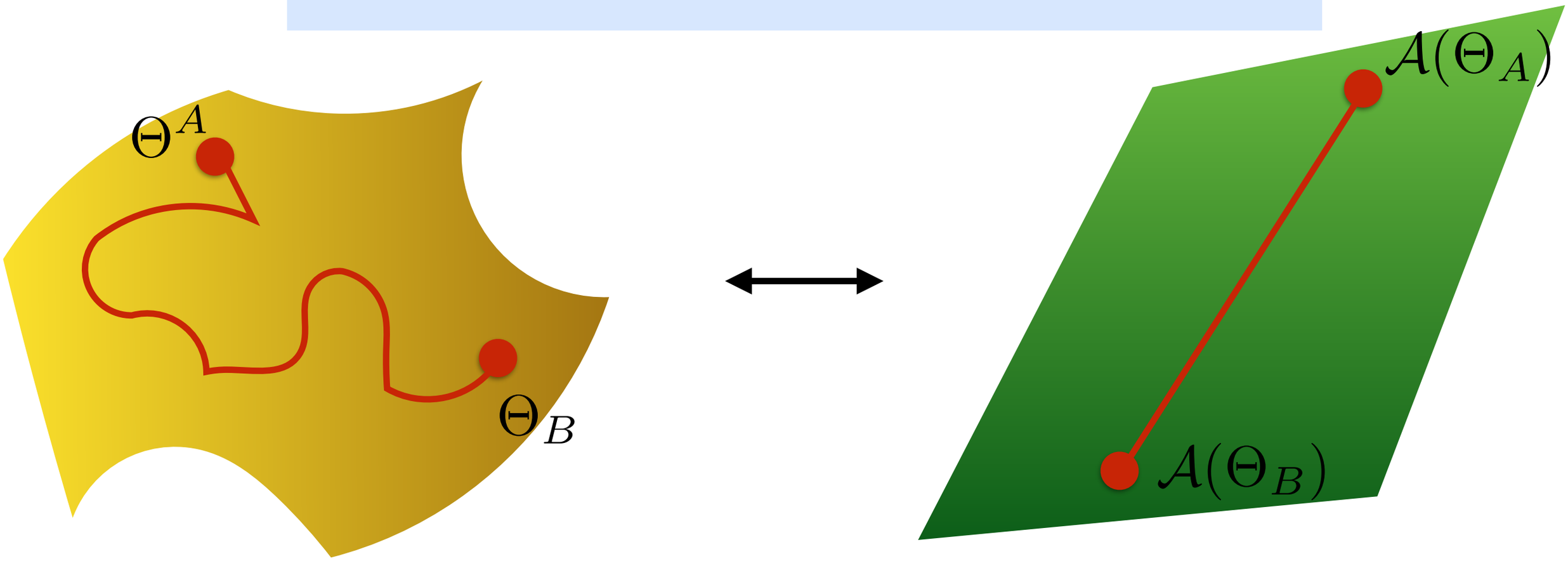
- Result is based on local linearization of the ReLU kernel (hence exponential price).

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

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

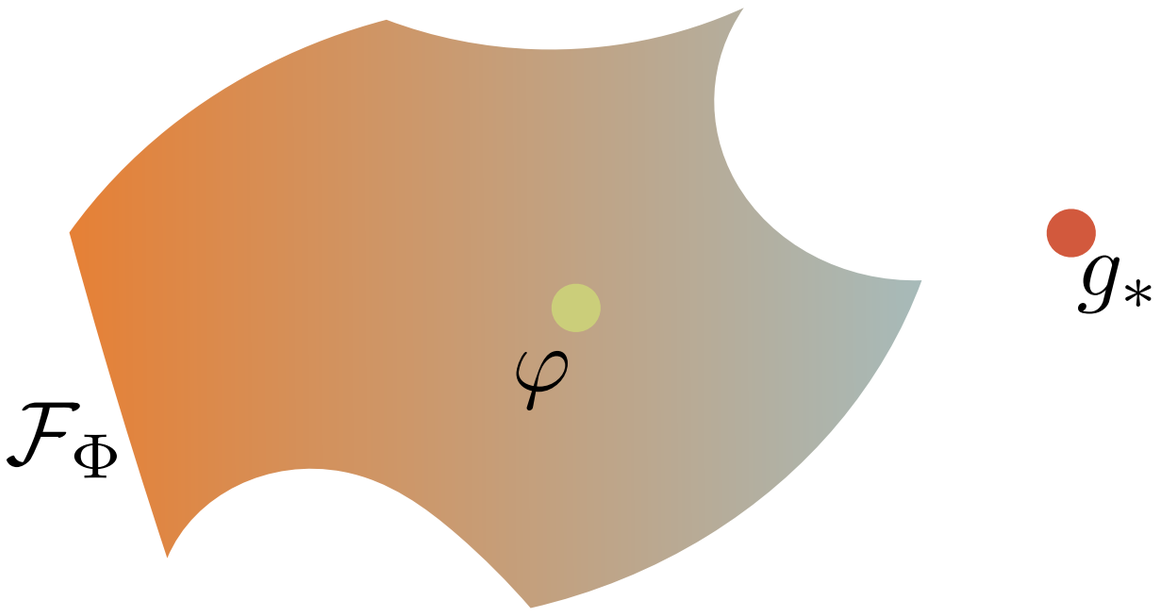
Theorem: [BVV'18] If $\dim\{\mathcal{A}(w),\, w \in \mathbb{R}^n\} = q < \infty$, then $L(U,W) = \mathbb{E}\|U\rho(WX) - Y\|^2$ has no spurious valley if $M \geq q$.

- 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:

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



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

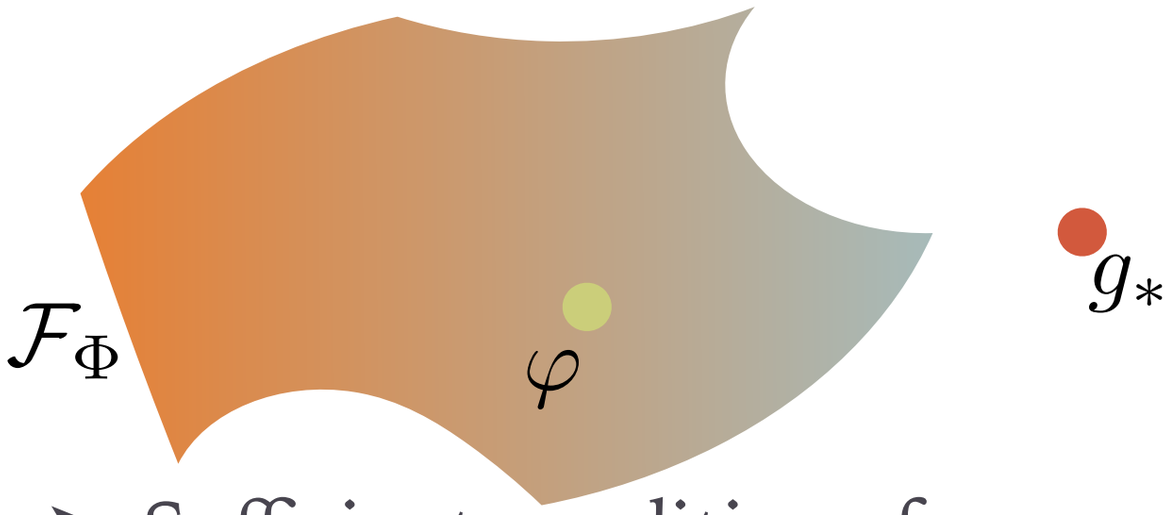
$$g_* : x \mapsto \mathbb{E}(Y|x)$$

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

PARAMETRIC VS MANIFOLD OPTIMIZATION

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

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



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

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

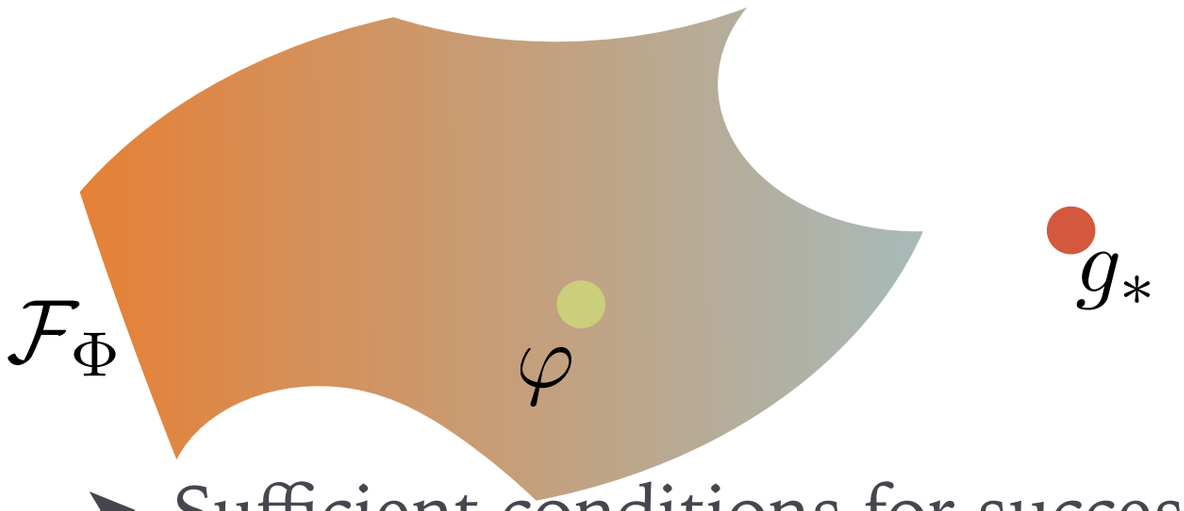
$$\langle f, g \rangle_p := \mathbb{E}\{f(X)g(X)\} \,.$$

- 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

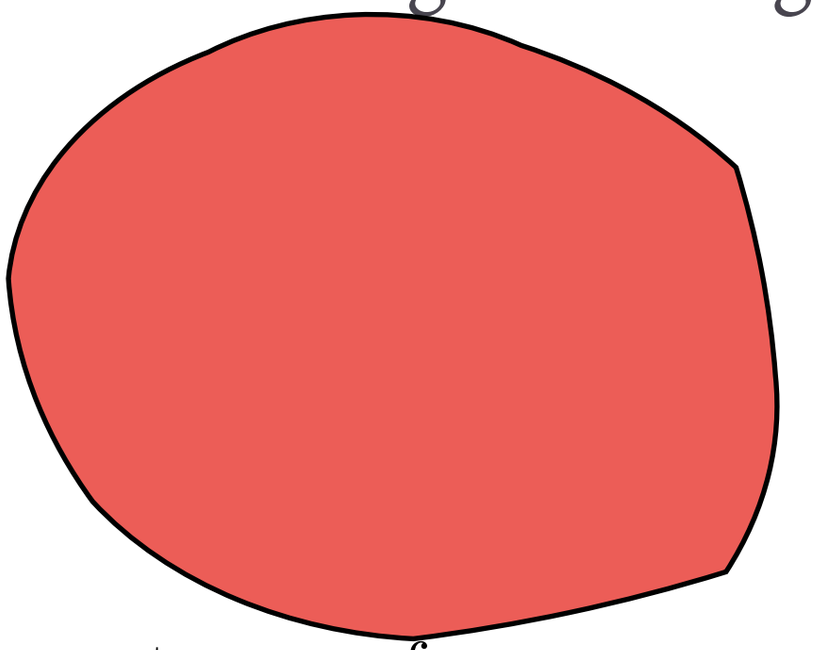
$$\beta(M,N) = \inf_{S;\dim(S)=f^{-1}(M)} \inf_{\substack{U\in\mathbb{R}^{m\times M}\\ W\in\mathbb{R}^{M\times f^{-1}(M)}}} \sup_{\substack{\mathbb{E}\|Z\|\le N-f^{-1}(M),\\ P_S Z=0}} \mathbb{E}\|U\rho(WP_S X+Z)-Y\|^2$$

  - 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 $\beta(f(N),N) = \min_{U,W} E(U,W)$.

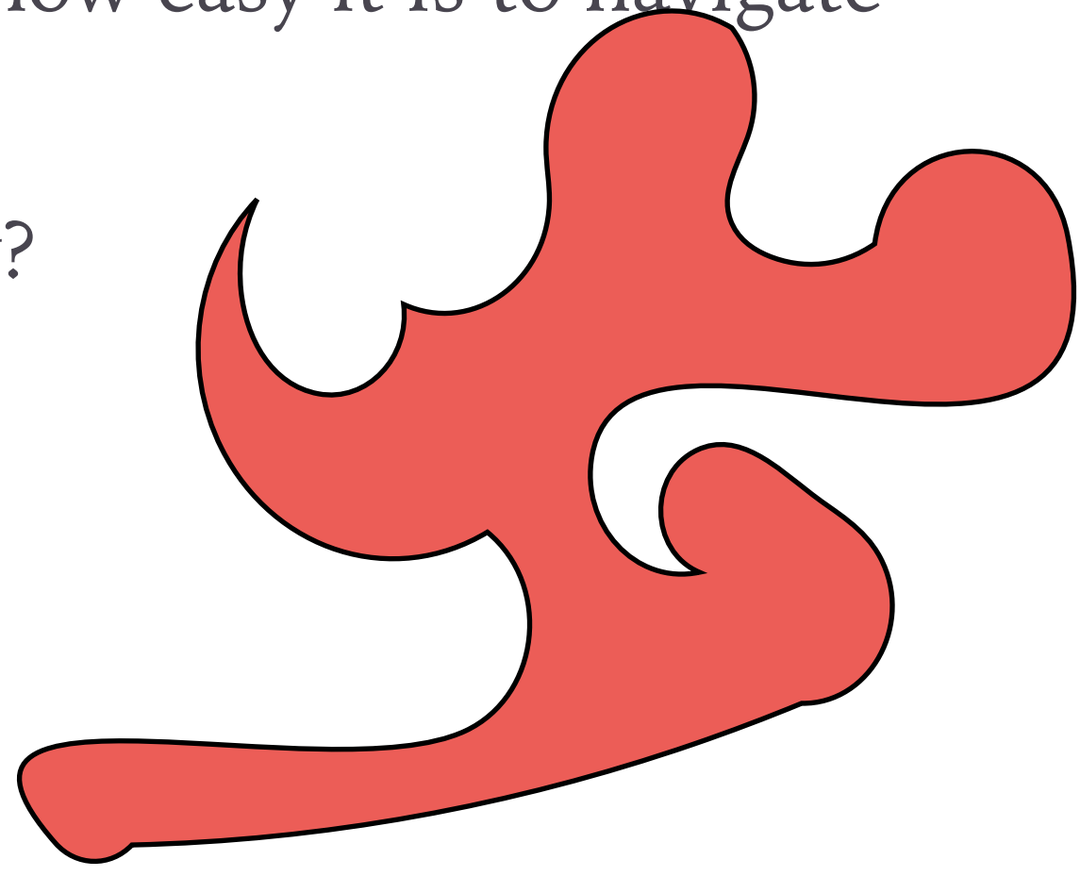
Conjecture [LBB'18]: The loss $L(U,W) = \mathbb{E}\|U\rho(WX)-Y\|^2$ has no poor local minima above the energy barrier $\beta(M,N)$.

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?



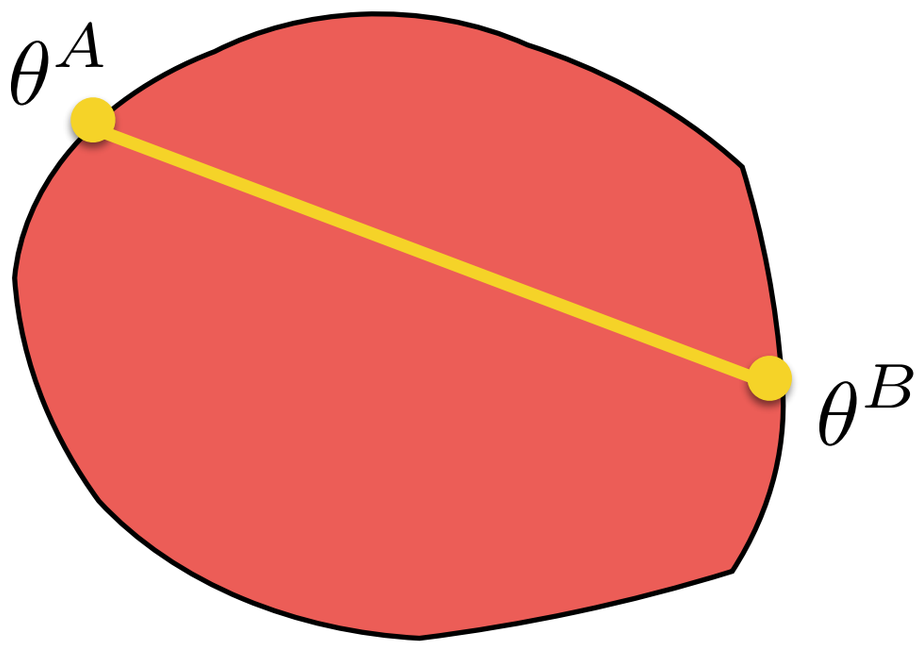
easy to move from one energy level to lower one



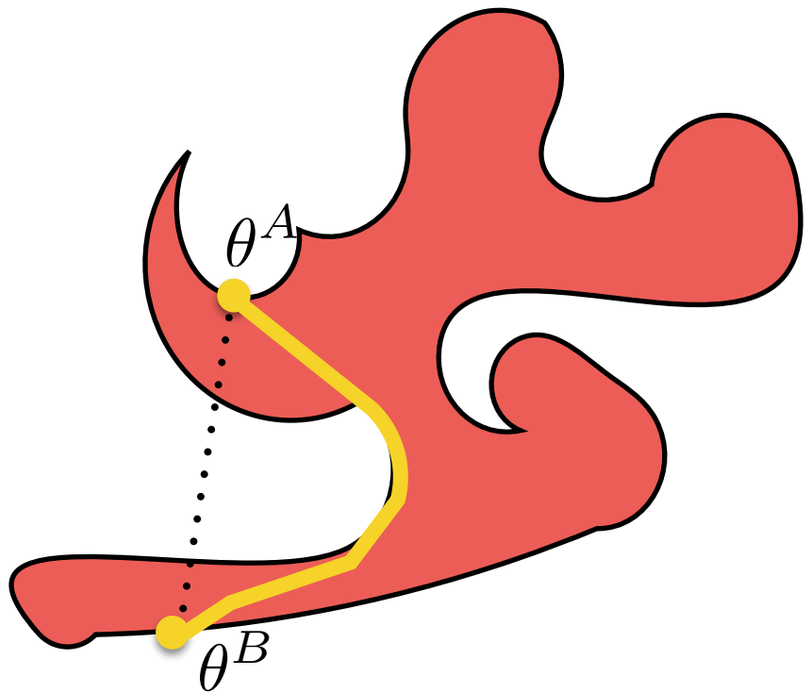
hard to move from one energy level to lower one

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.

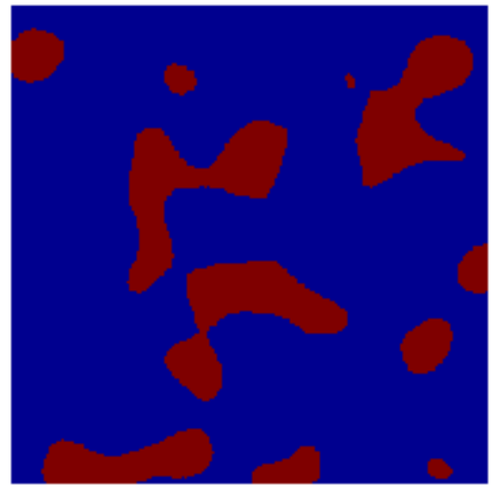


easy to move from one energy level to lower one



hard to move from one energy level to lower one

FINDING CONNECTED COMPONENTS



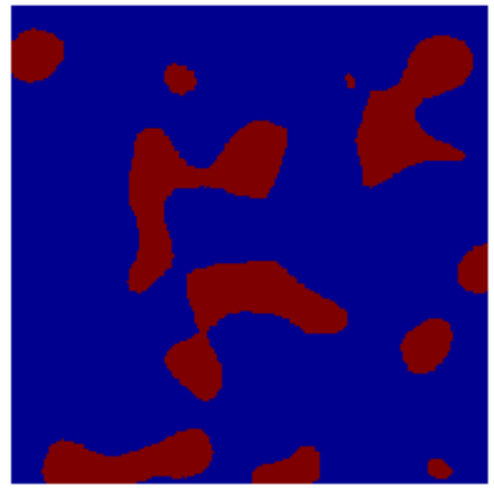
- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

    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:

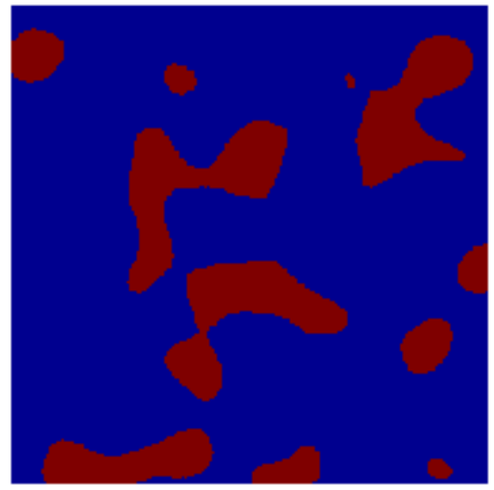
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- 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

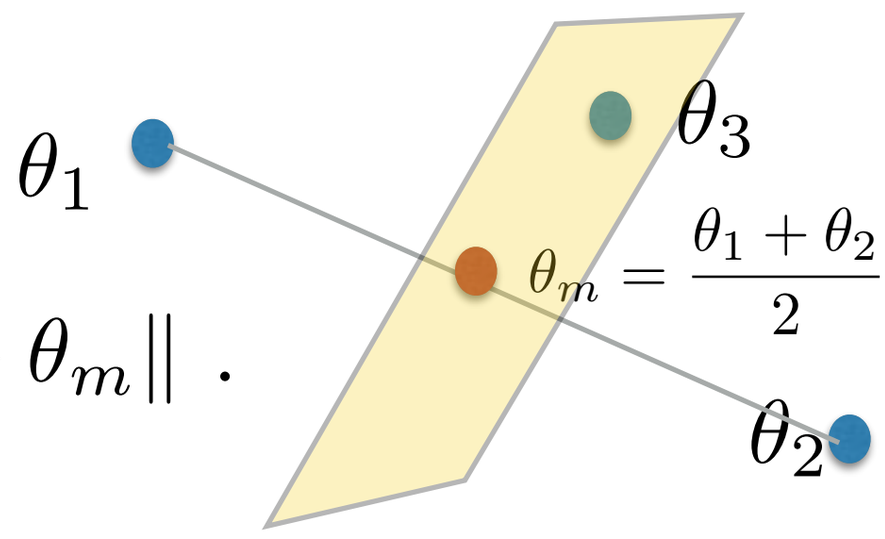
    there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

  - Moreover, we penalize the length of the path:

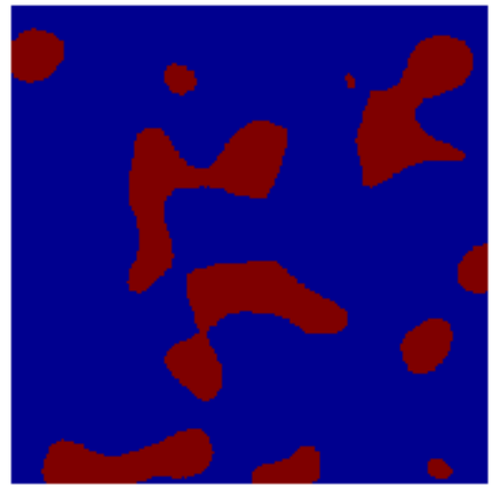
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int_{\mathcal{H}} \|\dot{\gamma}(t)\| dt \leq M\ .$$

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

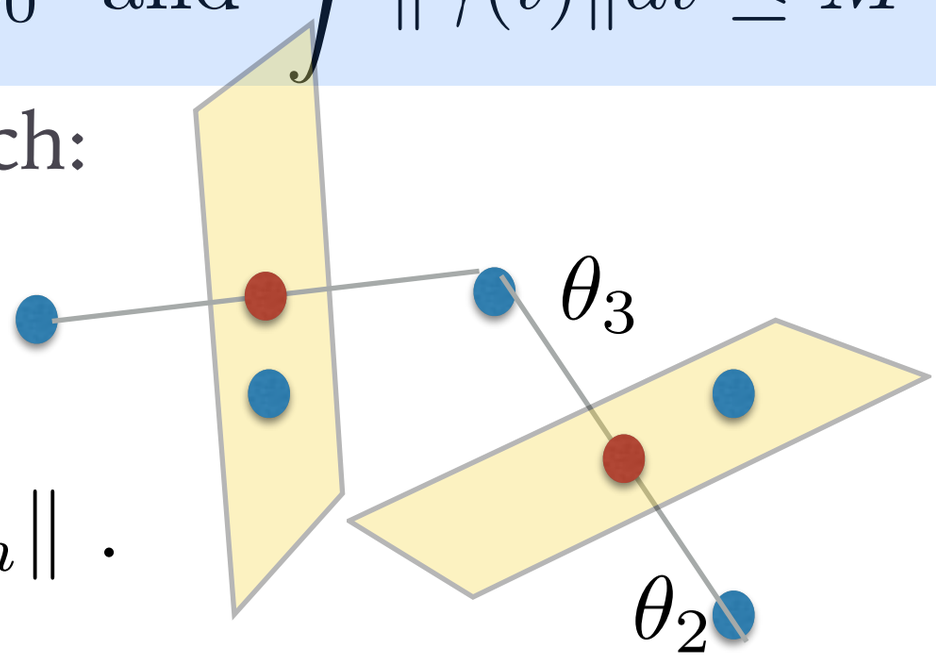
- Moreover, we penalize the length of the path:

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

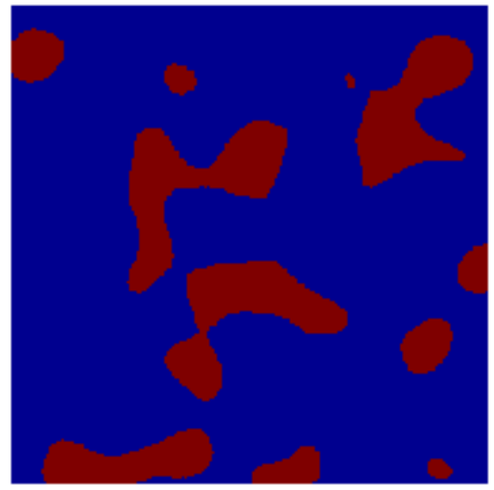
- Dynamic programming approach:

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

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



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
$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

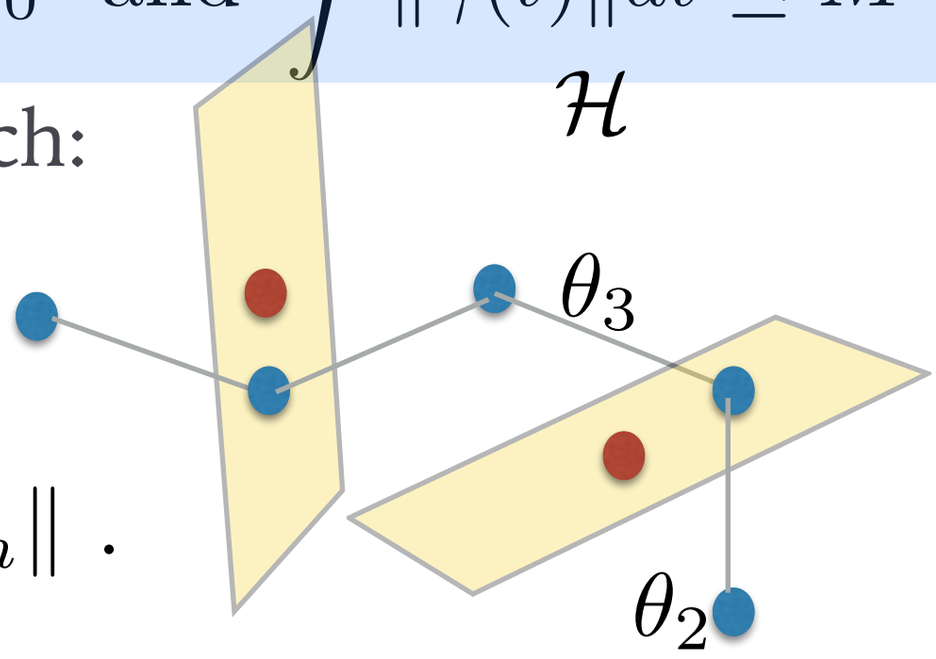
- Moreover, we penalize the length of the path:

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

- Dynamic programming approach:

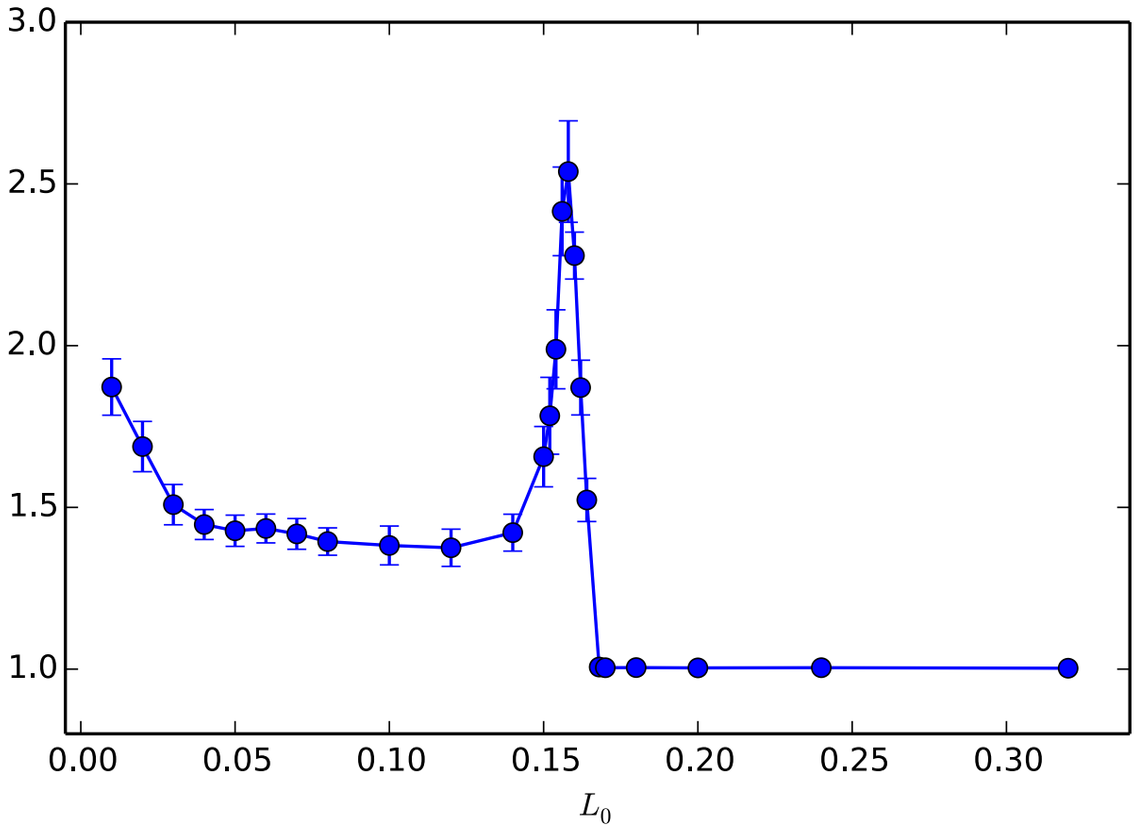
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

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

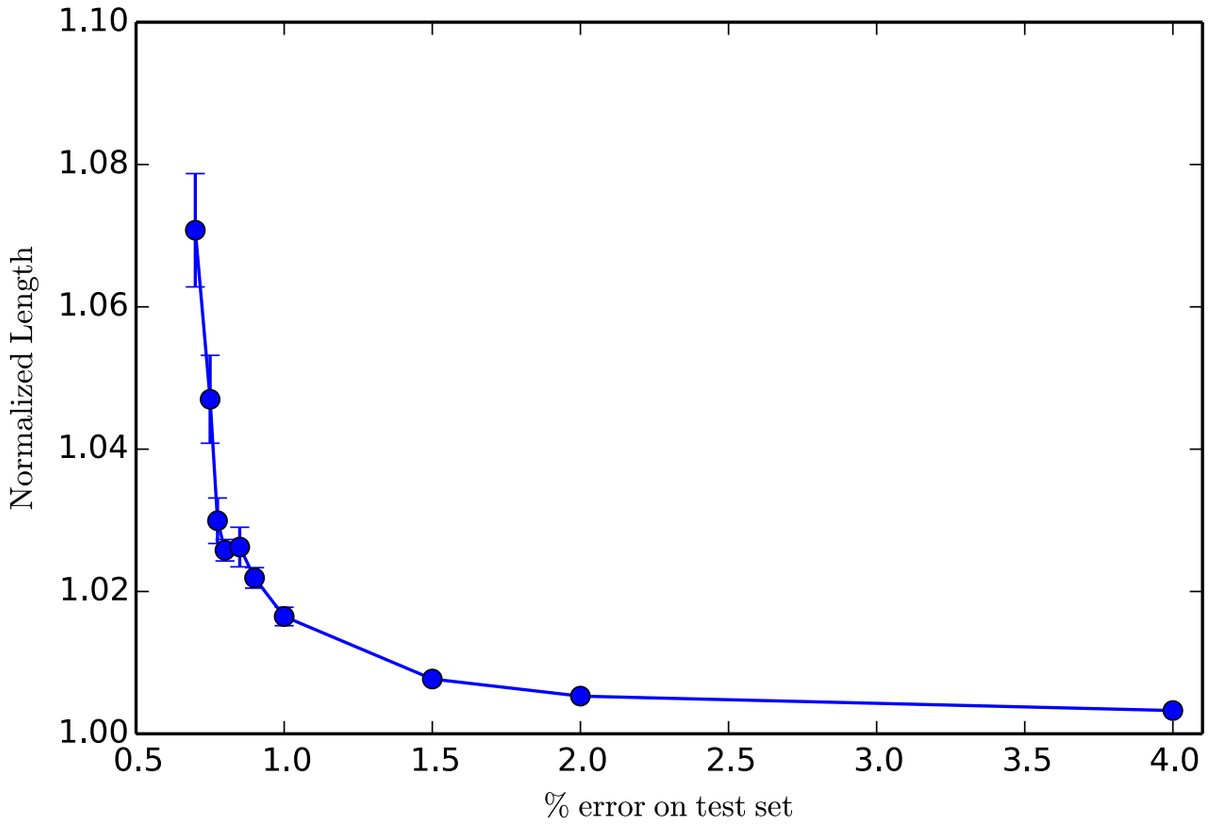


NUMERICAL EXPERIMENTS

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



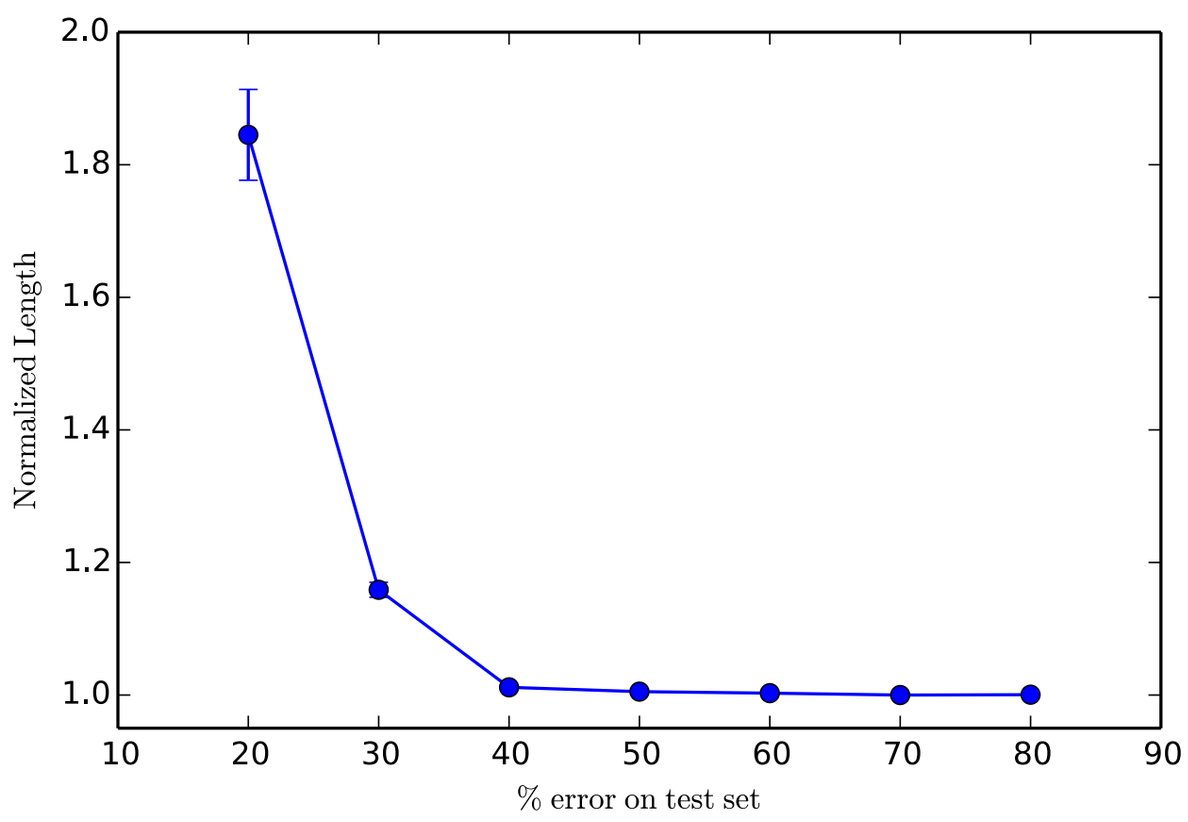
cubic polynomial



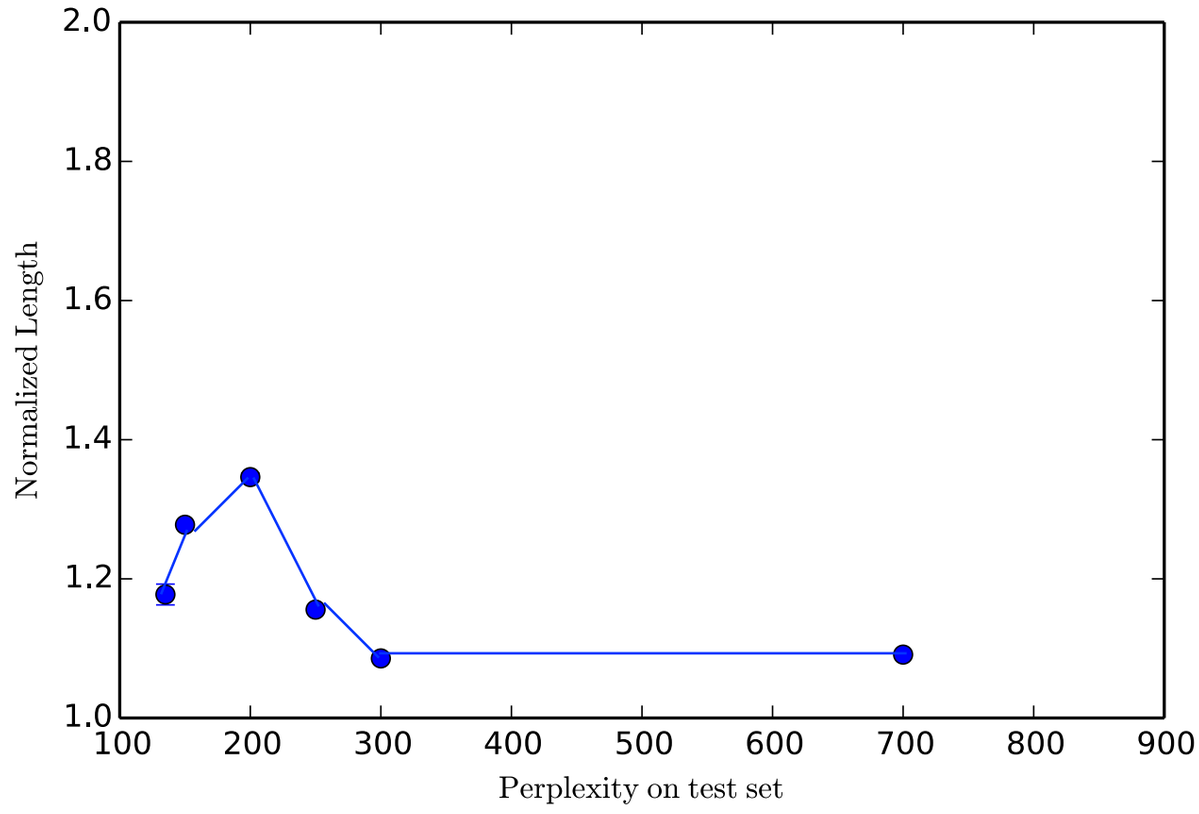
CNN/MNIST

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.



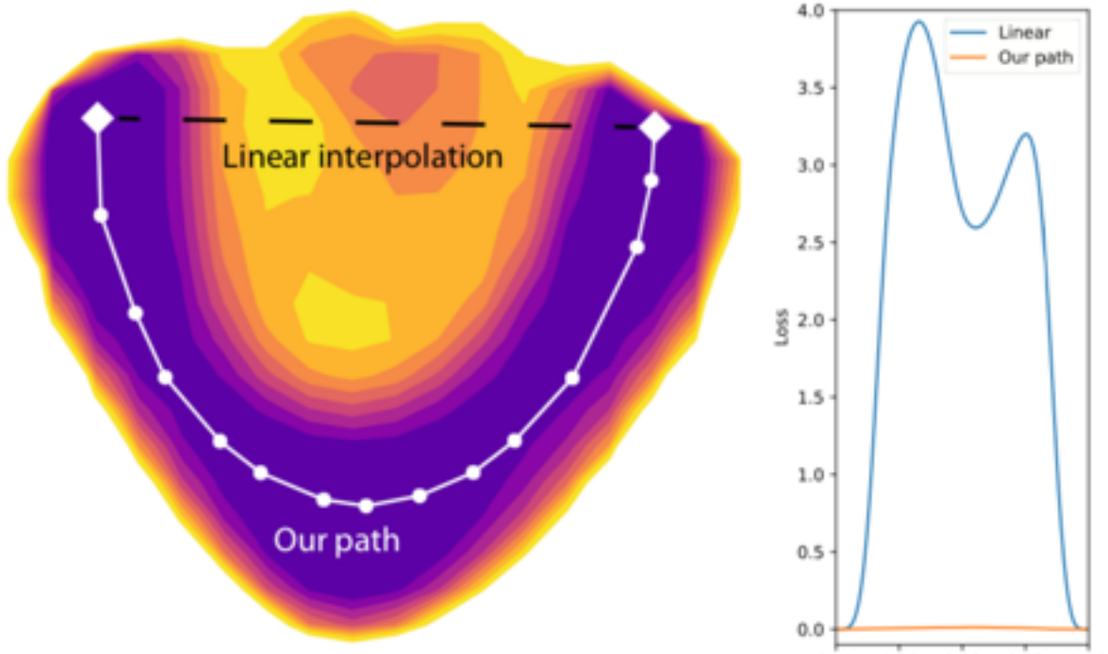
CNN/CIFAR-10



LSTM/Penn

FOLLOW-UPS

- Extensions to larger models and using a more sophisticated path interpolation method in [Draxler et al.'18]



- Using *Nudged Elastic Band* method.

BEYOND FINITE INTRINSIC DIMENSION

- For generic activation function and data distribution, should one expect presence of spurious valleys/minima?
  - Yes [more later]
- What characterization should one pursue in that case?

BEYOND FINITE INTRINSIC DIMENSION

- For generic activation function and data distribution, should one expect presence of spurious valleys/minima?
  - Yes [more later]
- What characterization should one pursue in that case?
- *Towards energy barrier characterizations*: certify that spurious minima lie below a certain energy level.

PRESENCE OF POOR LOCAL MINIMA [YUN ET AL.18]

- Consider a one-hidden layer nonlinear network trained with mean-squared error:

$$L(U, W) = \widehat{\mathbb{E}}\|U\rho(WX) - Y\|^2, \text{ with}$$

$$\rho(x) = \begin{cases} s_+ x & , x \geq 0 \\ s_- x & , x < 0 \end{cases}$$

- Assume $Y \neq RX + b$ for any $R \in \mathbb{R}^{d_y \times d_x}$, $b \in \mathbb{R}^{d_y}$.

PRESENCE OF POOR LOCAL MINIMA [YUN ET AL.18]

- Consider a one-hidden layer nonlinear network trained with mean-squared error:

$$L(U, W) = \widehat{\mathbb{E}}\|U\rho(WX) - Y\|^2, \text{ with}$$

$$\rho(x) = \begin{cases} s_+ x & , x \geq 0 \\ s_- x & , x < 0 \end{cases}$$

- Assume $Y \neq RX + b$ for any $R \in \mathbb{R}^{d_y \times d_x}$, $b \in \mathbb{R}^{d_y}$.
- **Theorem [YSJ'18]:** If samples are distinct and $m > 1$, then there are poor local minima of L whose empirical risk is the same as the linear fitting.

PRESENCE OF POOR LOCAL MINIMA

- Linear fitting is a very weak property: any interesting dataset won't be linearly solvable.
- The proof relies heavily in the piece-wise linear structure of the activation function.

PRESENCE OF POOR LOCAL MINIMA

- Linear fitting is a very weak property: any interesting dataset won't be linearly solvable.
- The proof relies heavily in the piece-wise linear structure of the activation function.
- What is the situation for generic activations?
- We advance towards results somewhat between positive and negative: the landscape is "nice" above a certain energy level.

RESNETS VS LINEAR MODELS

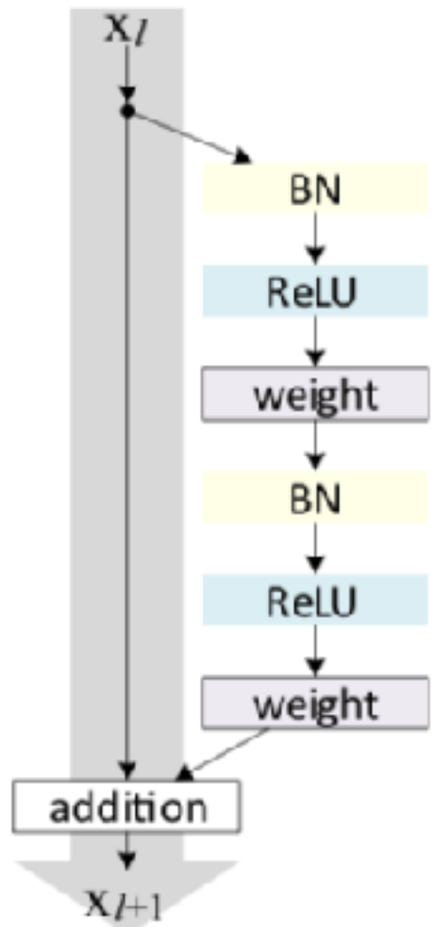
- [Shamir '18] studies the landscape of optimization of generic *Residual Networks* with linear preactivations:

$$f_{k+1}(x) = f_k(x) + h(f_k(x); \theta_k)$$

- The resulting network and loss becomes

$$\Phi_\Theta(x) = w^\top (x + V F_\theta(x)) \,, \; \Theta = \{w, V, \theta\} \,.$$

$$L(w, V, \theta) = \mathbb{E}[\ell(\Phi_\Theta(x), y)] \,.$$



[He et al.'16]

RESNETS VS LINEAR MODELS

- Linear models correspond to a specific slice of the parameter space:

$$L(w, 0, \theta) = \mathbb{E}[\ell(w^\top x, y)] \,.$$

RESNETS VS LINEAR MODELS

- Linear models correspond to a specific slice of the parameter space:

$$L(w, 0, \theta) = \mathbb{E}[\ell(w^\top x, y)] \,.$$

- **Theorem [Shamir'18]:** Suppose L is twice-differentiable and suppose M is a subset of the parameter domain such that $\max_M(\|w\|, \|V\|) \leq b$. Then any $(w, V, \theta) \in M$ which is an ϵ-second order stationary point satisfies

$$L(w, V, \theta) \leq \min_{\|w'\| \leq r} L(w', 0, \theta) + (\epsilon + \epsilon^{1/4})\text{poly}(b, r, \mu) \,.$$

(μ: Lipschitz constants of L and its derivatives)

- Consequence for spurious local minima?

RESNETS VS LINEAR MODELS

- Taking $\epsilon = 0$, we obtain

Corollary [Shamir'18]: Suppose L is twice-differentiable and continuous Hessians at any compact subset of its domain. Then every local minimum (w, V, θ) of L satisfies

$$L(w, V, \theta) \leq \inf_{w'} L(w', 0, 0) .$$

RESNETS VS LINEAR MODELS

- Taking $\epsilon = 0$, we obtain

Corollary [Shamir'18]: Suppose L is twice-differentiable and continuous Hessians at any compact subset of its domain. Then every local minimum (w, V, θ) of L satisfies

$$L(w, V, \theta) \leq \inf_{w'} L(w', 0, 0) .$$

- Does this directly imply that SGD will provably converge efficiently to a model outperforming linear regression?

RESNETS VS LINEAR MODELS

- Taking $\epsilon = 0$, we obtain

Corollary [Shamir'18]: Suppose L is twice-differentiable and continuous Hessians at any compact subset of its domain. Then every local minimum (w, V, θ) of L satisfies

$$L(w, V, \theta) \leq \inf_{w'} L(w', 0, 0) \, .$$

- Does this directly imply that SGD will provably converge efficiently to a model outperforming linear regression?
- No, because the bound depends on the norm of the stationary point (parameter b).
- Also, this energy barrier is typically very high.

CONCLUSIONS ON OPTIMIZATION LANDSCAPE

- Lots of recent results covering both positive, negative and *mixed* cases.
- Positive results:
  - Many results are distribution-independent, but not architecture independent. Universality should contain a bit of both.
- Negative results:
  - Most construct *planted* solutions: $Y = \Phi(X, \theta_0)$ for some θ_0. Towards more realistic negative examples?
- Mixed cases: energy barriers are a reasonable mathematical picture of realistic landscapes with the right universality properties.

SOME OPEN CHALLENGES

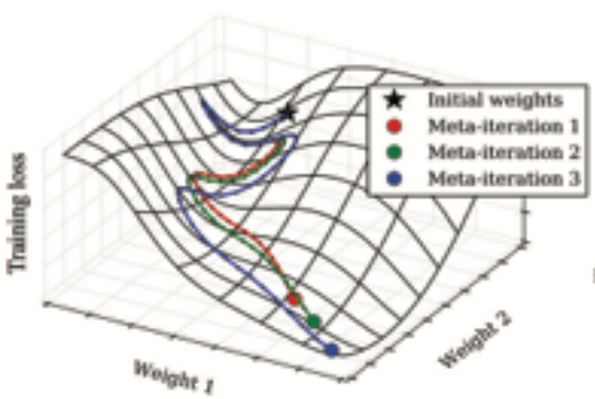
CONVOLUTIONS AND GENERALIZATION IN SUPERVISED LEARNING

- We saw last week that size of the model does not directly impact its generalization error…
- … provided it has the appropriate architecture.
- In the case of natural images, CNNs capture the correct inductive bias.

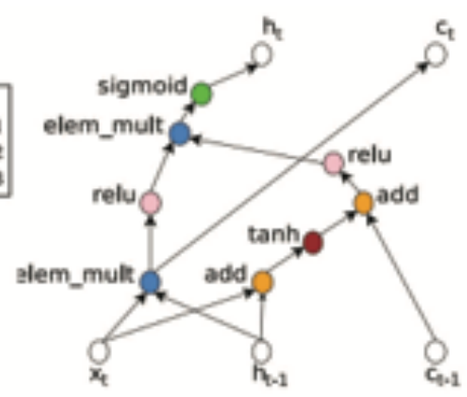
- Open: **construct statistical models for natural images that explain this mathematically.**
- Need to combine tools from signal processing (e.g. wavelets) with statistical learning.
- Geometric measures of generalization.

NON-STATIONARY AND FEW-SHOT LEARNING

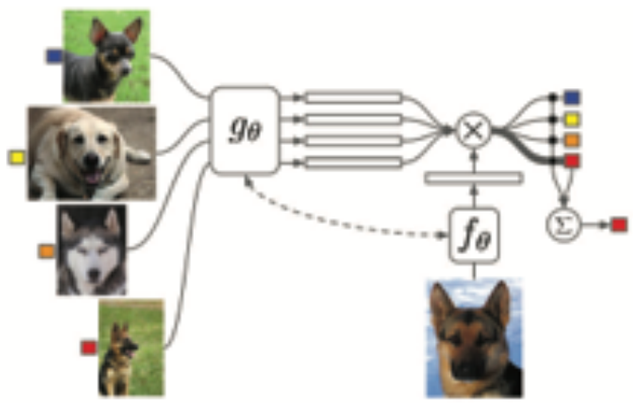
- Many practical scenarios fall outside the traditional iid setup of a large training set $X_i \sim P$.
  - Few-shot learning [Lake et al.'15]
  - Meta-learning [Schmidhuber,'90s, ...]
  - "Learning to Learn"



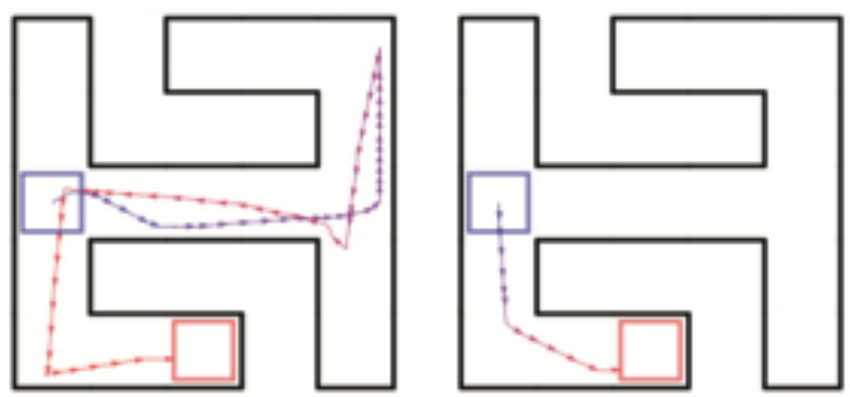
hyperparameter optimization
Maclaurin et al. '15



learned recurrent cell
Zoph & Le '17



few-shot image classifier
Vinyals et al. '16


learning to quickly navigate new mazes
Duan et al. '16

- *C. Finn, Bair blog*

NON-STATIONARY AND FEW-SHOT LEARNING

- Most of these instances can be reduced to a supervised learning task by lifting the input to be the set

$$S_i = (\{x_1, y_1, \ldots x_K, y_K, x_t\}, y_t)$$

  - and then assuming one has an iid collection of such sets.

- How to formalize this setup and quantify the resulting bias-variance tradeoffs?
- Which properties of high-dimensional landscapes and stability translate into this setting?

DEEP LEARNING APPLICATIONS

- In numerical methods
  - Develop Data-driven Integrators that replace fixed integrators (e.g. Projective Dynamics, RK).
  - Fast Multipole Methods.
  - **Need certificate guarantees.**
- In statistical/computational complexity
  - Tool for high-dimensional algorithmic search?
  - **Need ability to interpret.**

DEEP LEARNING IN PHYSICAL SCIENCES

- In Quantum mechanics, current surge of interest for exploiting deep learning models to approximate wave functions Ψ.
- In an N-body system, this wave function describes the quantum state of the system and is defined in a (complex) Hilbert space of dimension N.
- Observables correspond to the spectrum o adjoint operators of this Hilbert space.
- Challenge: construct tractable *ansatz* with approximation properties

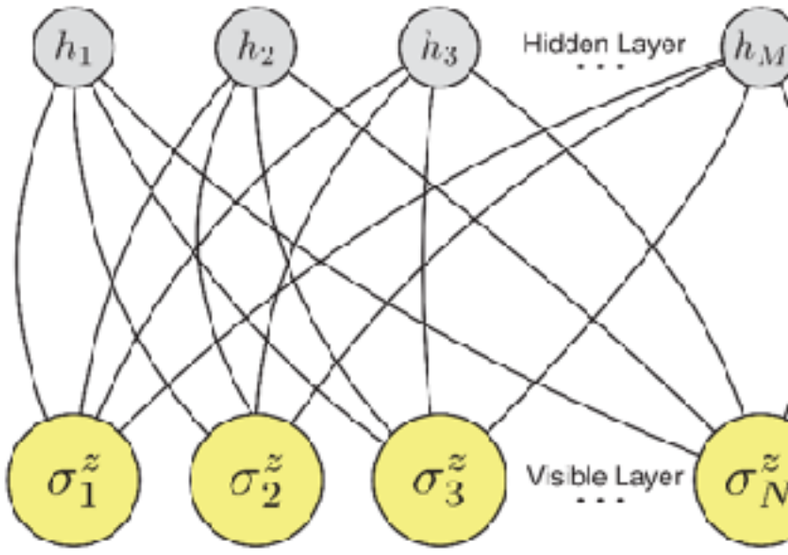
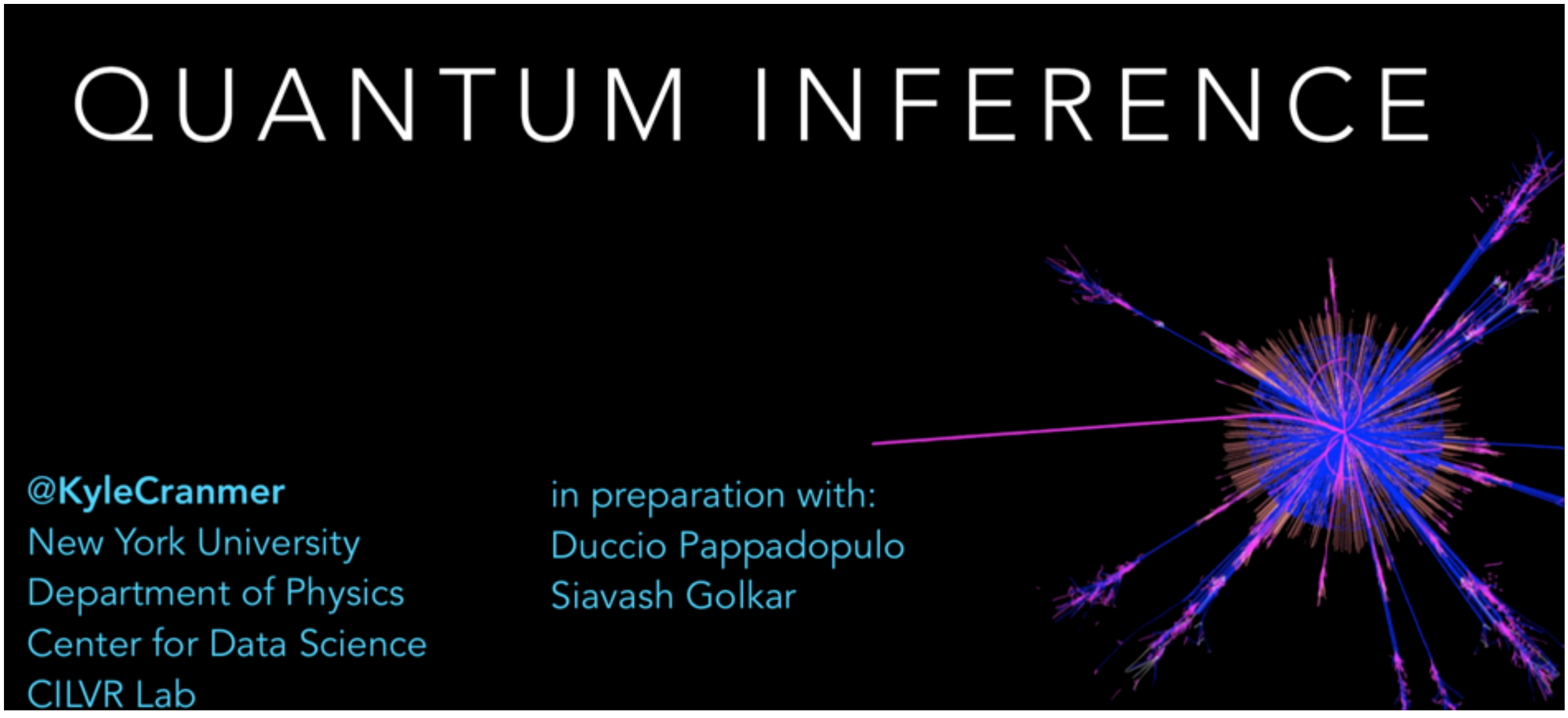


Fig. 1. Artificial neural network encoding a many-body quantum state of N spins. A Boltzmann machine architecture that features a set of N visible artificial neurons (yellow d set of M hidden neurons (gray dots) is shown. For each value of the many body spin co $\mathcal{S} = (\sigma_1^z, \sigma_2^z, \ldots, \sigma_N^z)$, the artificial neural network computes the value of the wave function

Carleo et al, Science'17

DEEP LEARNING IN PHYSICAL SCIENCES

- In particular, how to perform quantum variational inference?



MULTI-AGENT SYSTEMS

- Beyond potential systems: not even a landscape of optimization, but a more general dynamical system:

$$\dot{\theta}_k = \nabla_{\theta_k} f_k(\theta_1, \ldots, \theta_N) \ , \ (k \leq N) \, .$$

The Mechanics of n-Player Differentiable Games

David Balduzzi [1] Sébastien Racaniere [1] James Martens [1] Jakob Foerster [2] Karl Tuyls [1] Thore Graepel [1]

- Framework that goes beyond matrix games (e.g. prisoners dilemma)?
- High-dimensional dynamics, no Nash equilibria in general.

THANKS FOR LISTENING!