M A D

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

MATHEMATICS OF DEEP LEARNING

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

Lecture 10: Gradient Descent in non-convex Landscapes.

OBJECTIVES LECTURE 10

- Gradient Descent (stochastic and deterministic) Converges to local Minimisers.

- Role of Saddle points: exponential slowing-down. Efficiently escaping saddles with perturbations.

- Critical Point Analysis in Deep Learning: linear nets, spin glasses.

- Beyond Hessian analysis: sub-level set analysis

COMMENTS ON [RAGINSKY, RAKHLIN, TELGARSKY]

- **Theorem [Raginsky et al. '17, informal]:** The former decomposition terms are bounded as

$$\mathbb{E}g(\hat{\theta}) - \mathbb{E}g(\hat{\theta}_*) \simeq \epsilon \mathrm{Poly}\left(\beta^{-1}, d, \lambda_*^{-1}\right)$$

for $K \geq \mathrm{Poly}(\beta^{-1}, d, \lambda_*^{-1})\epsilon^{-4}$, and $\gamma \leq \dfrac{\epsilon^4}{\log(\epsilon^{-1})^4}$,

$$\mathbb{E}g(\hat{\theta}_*) - \mathbb{E}g_x(\hat{\theta}_*) \simeq \frac{(\beta^{-1} + d)^2}{\lambda_* n},$$

$$\mathbb{E}g_x(\hat{\theta}_*) - \inf_\theta g(\theta) \simeq \beta d \log(\beta^{-1} + 1).$$

- Only first term involves SGLD, other two only Gibbs sampler.
- λ_* is the spectral gap governing convergence of Markov chain.
  (it is independent of n)

COMMENTS ON [RAGINSKY, RAKHLIN, TELGARSKY]

- However, this spectral gap is exponential!
- We will see examples later where convergence to global optima is provably exponential.

NON-CONVEX OPTIMIZATION

- If optimization gradient-based algorithms converge to global minima in the convex case, what can we expect in non-convex optimization?

- Convergence to local minimisers?
- How many of those are global?
- How much can we be slowed down by saddle points?
- How much noise to add to gradient descent to escape those saddles?

NON-CONVEX OPTIMIZATION

- Vanilla Gradient descent scheme: $\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$
- Its equilibrium points satisfy $\nabla g(\theta_*) = 0$.

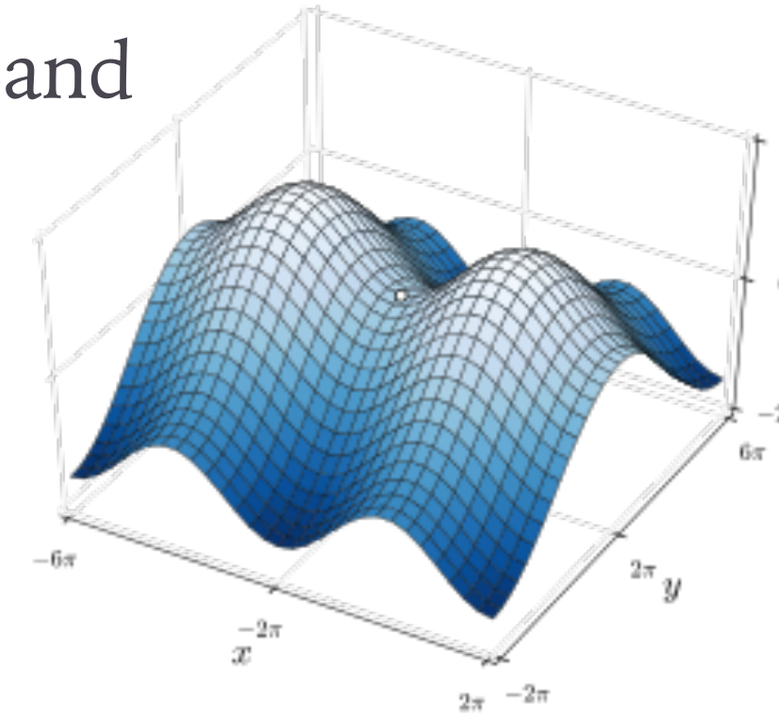
NON-CONVEX OPTIMIZATION

- Vanilla Gradient descent scheme: $\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$
- Its equilibrium points satisfy $\nabla g(\theta_*) = 0$.
- When g is convex, 1st order critical points are global minima.
- However, for general g, stationary points of gradient descent are not necessarily global minima.
- How can they be classified?

INDEX OF CRITICAL POINTS

- Suppose g is in $\mathcal{C}^2$ defined on $\mathbb{R}^d$.
- A *critical point* $\theta^* \in \mathbb{R}^d$ of g is such that $\nabla g(\theta^*) = 0$.
- θ^* is a *strict saddle point* if it is a critical point and its Hessian satisfies

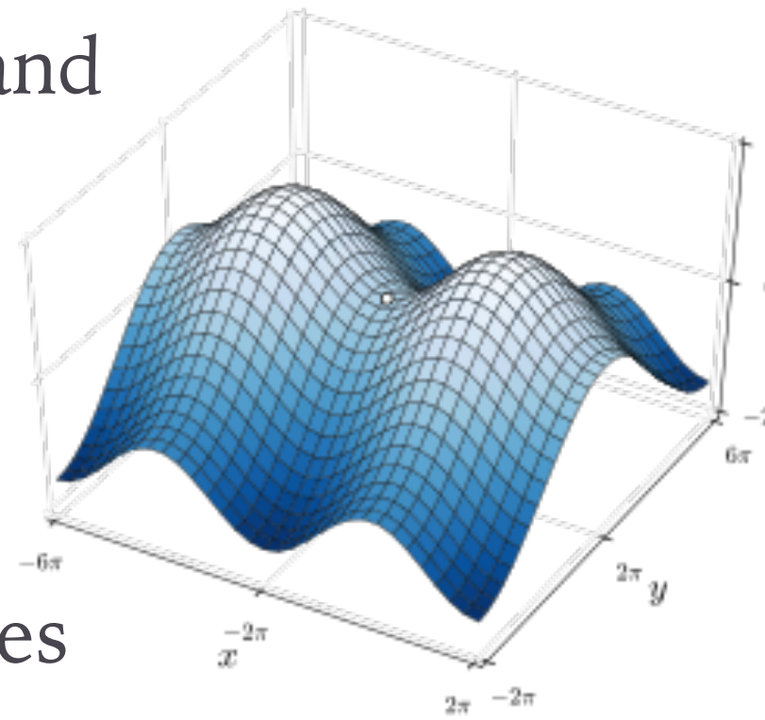
$$\lambda_{\min}(\nabla^2 g(\theta_*)) < 0\,.$$



INDEX OF CRITICAL POINTS

- Suppose g is in $\mathcal{C}^2$ defined on $\mathbb{R}^d$.
- A *critical point* $\theta^* \in \mathbb{R}^d$ of g is such that $\nabla g(\theta^*) = 0$.
- θ^* is a *strict saddle point* if it is a critical point and its Hessian satisfies

$$\lambda_{\min}(\nabla^2 g(\theta_*)) < 0\,.$$



- A local minima (not necessarily strict) satisfies $\lambda_{\min}(\nabla^2 g(\theta_*)) \geq 0\,.$
- While every critical point is an equilibrium point of gradient descent, which are *stable* equilibria?

CHARACTERIZING STABLE EQUILIBRIA OF GRADIENT DESCENT

- Our intuition is that strict saddles are unstable equilibria, and thus it is unlikely that gradient descent reaches those points.
- We can find worst-case initializations of gradient descent that provably converge to saddle-points [Nesterov,'04].

CHARACTERIZING STABLE EQUILIBRIA OF GRADIENT DESCENT

- Our intuition is that strict saddles are unstable equilibria, and thus it is unlikely that gradient descent reaches those points.
- We can find worst-case initializations of gradient descent that provably converge to saddle-points [Nesterov,'04].
- On the other hand, we have just seen that adding noise to gradients can be used to escape saddle points.
- But, how likely is it that randomly initialized gradient descent gets stuck in a saddle point?
- We can study the region of attraction associated to each critical point.

STABLE MANIFOLD

- A discrete-time optimization algorithm is a mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^d$
  - ex for gradient descent, $\varphi(\theta) = \theta - \gamma \nabla g(\theta)$.
  - iterate k is thus obtained as $\theta_k = \varphi^k(\theta_0)$.
- Denote by $\mathcal{X}^*$ the set of strict saddle points of g.
- **Definition** (Global Stable Set): The Global Stable Set of the strict saddles is

$$S_\varphi = \{\theta_0; \lim_{k \to \infty} \varphi^k(\theta_0) \in \mathcal{X}^*\} .$$

INTUITION: QUADRATIC CASE

- Consider the non-convex quadratic function

$$g(\theta) = \theta^\top H\theta\,, H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)\,, \lambda_1, \ldots \lambda_s > 0,\ \lambda_{s+1}, \ldots, \lambda_d < 0\,.$$

- A single critical point, $\theta^* = 0$, which is a strict saddle.
- Gradient descent initialised from θ_0 produces

$$\theta_{k+1} = \varphi(\theta_k) = \sum_{i=1}^{d} (1 - \gamma\lambda_i)^{k+1} \theta_{0,i} e_i\,.$$

- Suppose $\gamma < \dfrac{1}{\max_i |\lambda_i|}$. Then $\|\theta_k\| \to 0$ iff $\theta_0 \in \mathrm{span}\{e_1, \ldots, e_s\}$.
- It follows that if θ_0 is chosen at random, then we will avoid the strict saddle with probability 1.

INTUITION: GENERAL CASE

- In the case of gradient descent, in a neighborhood of a critical point θ_*, the local attractive set is well approximated by the span of eigenvectors corresponding to positive eigenvectors of the Hessian $\nabla^2 g(\theta_*)$.
- This local attractive set has zero measure within a small neighborhood, thus with probability 1 an initialization sufficiently close to θ_* will leave this neighborhood.
- From local stability to global stability?

INTUITION: GENERAL CASE

- In the case of gradient descent, in a neighborhood of a critical point θ_*, the local attractive set is well approximated by the span of eigenvectors corresponding to positive eigenvectors of the Hessian $\nabla^2 g(\theta_*)$.
- This local attractive set has zero measure within a small neighborhood, thus with probability 1 an initialization sufficiently close to θ_* will leave this neighborhood.
- From local stability to global stability?
- Not trivial: locally escaping a given saddle point does not guarantee we won't fall into another one later.

STABLE MANIFOLD THEOREM

- Given a diffeomorphism φ, we define an unstable fixed point set as

$$\mathcal{A}^*_\varphi = \{\theta;\, \varphi(\theta) = \theta;\, \max_i |\lambda_i(D\varphi(\theta))| > 1\}.$$

- **Theorem [Stable Manifold]:** Let φ be a C^1 mapping from $\mathcal{X} \to \mathcal{X}$ and $\det(D\varphi(\theta)) \neq 0$ for all $\theta \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero:

$$\mu\left(\{\theta_0 : \lim_k \varphi^k(\theta_0) \in \mathcal{A}^*_\varphi\}\right) = 0\,.$$

STABLE MANIFOLD THEOREM

- Given a diffeomorphism φ, we define an unstable fixed point set as

$$\mathcal{A}^*_\varphi = \{\theta;\, \varphi(\theta) = \theta;\, \max_i |\lambda_i(D\varphi(\theta))| > 1\}.$$

- **Theorem [Stable Manifold]:** Let φ be a C^1 mapping from $\mathcal{X} \to \mathcal{X}$ and $\det(D\varphi(\theta)) \neq 0$ for all $\theta \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero:

$$\mu\left(\{\theta_0 : \lim_k \varphi^k(\theta_0) \in \mathcal{A}^*_\varphi\}\right) = 0\,.$$

- **Corollary:** If $\mathcal{X}^* \subseteq \mathcal{A}^*_\varphi$, then $\mu(S_\varphi) = 0$.

GRADIENT DESCENT AVOIDS STRICT SADDLES

- Assume ∇g is Lipschitz, with $\|\nabla^2 g(\theta)\| \leq L$.
- Fact 1: every strict saddle point of g is an unstable fixed point of gradient descent: $\mathcal{X}^* \subseteq \mathcal{A}_\varphi^*$

GRADIENT DESCENT AVOIDS STRICT SADDLES

- Assume ∇g is Lipschitz, with $\|\nabla^2 g(\theta)\| \leq L$.
- Fact 1: every strict saddle point of g is an unstable fixed point of gradient descent: $\mathcal{X}^* \subseteq \mathcal{A}_\varphi^*$
- Fact 2: Moreover, if step size satisfies $\gamma < L^{-1}$, then $\det(D\varphi(\theta)) \neq 0 \, \forall \theta$.

GRADIENT DESCENT AVOIDS STRICT SADDLES

- Assume ∇g is Lipschitz, with $\|\nabla^2 g(\theta)\| \leq L$.
- Fact 1: every strict saddle point of g is an unstable fixed point of gradient descent: $\mathcal{X}^* \subseteq \mathcal{A}_\varphi^*$
- Fact 2: Moreover, if step size satisfies $\gamma < L^{-1}$, then $\det(D\varphi(\theta)) \neq 0 \, \forall \theta$.
- Consequence: [Lee et al.'16] Under these assumptions, the stable set of strict saddle points has measure zero.
- This result also holds for proximal point, coordinate descent, mirror descent.

COMMENTS

- This result establishes that gradient descent escapes strict saddles. It does not directly imply that it converges to local minimizers!
- The required additional property is that $\lim_k \theta_k$ exists.
- Two sufficient conditions are discussed in [Lee et al.'16]:
  - Isolated critical points and compact sublevel sets
  - Satisfies the local Lojasiewicz inequality

$$\|\nabla g(\theta)\| \geq m|g(\theta) - g(\theta^*)|^a \, , a < 1$$

- What happens in the stochastic case?

FIRST-ORDER STATIONARY POINTS IN GRADIENT DESCENT

- An *ϵ-first-order stationary point* of g is such that $\|\nabla g(\theta^*)\| \leq \epsilon$.
- We know that plain gradient descent converges to first-order stationary points in the class of L-smooth functions:

Theorem [Nesterov'98]: For any $\epsilon > 0$, gradient descent with step size $\gamma = L^{-1}$ requires $\frac{L(g(\theta_0)-g^*)}{\epsilon^2}$ iterations to reach an ϵ-first order stationary point of g.

- This result is not cursed by dimensionality.
- Whereas first-order stationary points are sufficient in convex optimization, not anymore when g is non-convex.

ESCAPING FROM SADDLE POINTS

- We are interested in the problem of finding ϵ-*second order stationary points:* Assuming $\nabla^2 g$ is Lipschitz, such points satisfy

$$\|\nabla g(\theta^*)\| \leq \epsilon \text{ and } \lambda_{\min}(\nabla^2 g(\theta^*)) \geq -\sqrt{\epsilon\rho}\,.$$

$$\rho = \mathrm{Lip}(\nabla^2 g)$$

- If all saddle points are strict, then second-order stationary points are all local minima.

ESCAPING FROM SADDLE POINTS

- Owing to the instability of strict saddle points, stochastic gradient descent is also robust:

Theorem [Ge et al.'15]: If $g(\theta)$ satisfies the strict saddle, then Noisy Stochastic Gradient Descent converges to a local minimum in polynomial time.

- The theorem requires robust strict saddle (away from 0).
- They consider the noisy SGD setting:

$$\theta_{k+1} = \theta_k - \gamma(\nabla g(\theta_k) + \epsilon_k + Z)\,, Z \sim \mathcal{N}(0, I)\,.$$

  - Gaussian noise guarantees exploration in all directions.
- However, Time is dimension-dependent $(\Omega(d^4))$.

ESCAPING FROM SADDLE POINTS

- We have established that both gradient descent and stochastic gradient descent escape (strict) saddles.

- How fast do they escape?

- Does the noise improve or decrease the efficiency?

- Dependency on input dimensionality of the problem?

ESCAPING FROM SADDLE POINTS WITH SGD

- [Jin et al.'17] study a similar version of SGD, where noise is added conditionally:
  - When the norm of the current gradient is smaller than a threshold, they add a small random perturbation, uniformly sampled from a d-dimensional ball.
- **Theorem [Jin et al.'17]:** Assume ∇g is L-Lipschitz and $\nabla^2 g$ is ρ-Lipschitz. Then with appropriate parameters, noisy Gradient Descent will output an ϵ-second order stationary point whp with

$$\tilde{O}\left(\frac{L(g(\theta_0)-g^*)}{\epsilon^2}\right) \text{ (up to polylog factors).}$$

ESCAPING FROM SADDLE POINTS WITH SGD

- The dimension appears in the rate as a $\log^4 d$ factor. It could be improved to a single log, but most probably not removed [open problem].
- When a current estimate θ_k is not an ϵ-second order stationary point, two things can happen:
  - $\|\nabla g(\theta_k)\|$ is large: not yet a 1st order stationary point.
  - $\nabla^2 g(\theta_k)$ has a large negative eigenvalue: noise added to updates enables descent in that case.
- Adding noise to the gradient is thus sufficient to escape saddles. Is noise also necessary to efficiently escape?

SMOOTH COUNTER-EXAMPLE

- Adversarially initializing Gradient Descent can stop any progress (for example, initialize exactly at a saddle point).
- As it turns out, even with uniform initialization, one can find smooth functions for which GD requires exponential time to escape:

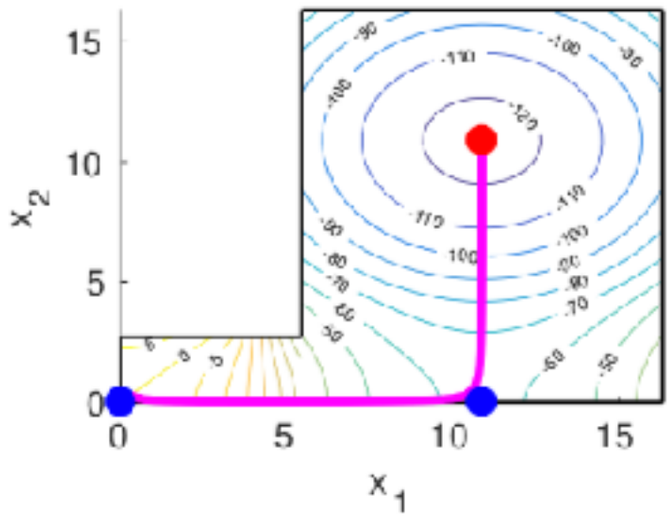
Theorem [Du et al.'17]: Suppose we initialize using uniform distribution on $[-1, 1]^d$. There exists a smooth function $g \in \mathbb{R}^d$ such that:

- with probability 1, GD will be $\Omega(1)$ distance away from any local minima for any $K \leq e^{\Omega(d)}$.
- for any $\epsilon > 0$, whp noisy GD finds a point θ such that $\|\theta - \theta^*\| \leq \epsilon$ for some local minimum θ^* in $\text{poly}(d, \epsilon^{-1})$ iterations.

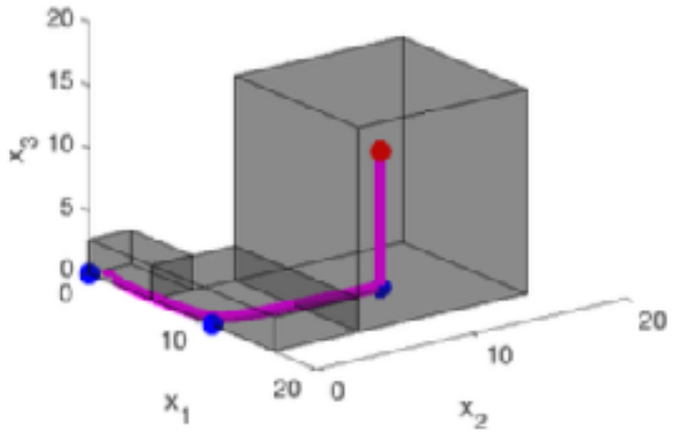
INTUITION

- The result can be generalized for initializations under any distribution that has most of its mass within a compact support (e.g. Gaussian).
- Key intuition: consider the function

$$g(\theta_1, \theta_2) = \begin{cases} -\gamma\theta_1^2 + L\theta_2^2 & \text{if } \theta_1 \in [0,1], \theta_2 \in [0,1] \\ L(\theta_1 - 2)^2 - \gamma\theta_2^2 & \text{if } \theta_1 \in [1,3], \theta_2 \in [0,1] \\ L(\theta_1 - 2)^2 + L(\theta_2 - 2)^2 - c & \text{if } \theta_1 \in [1,3], \theta_2 \in [1,3] \end{cases}$$

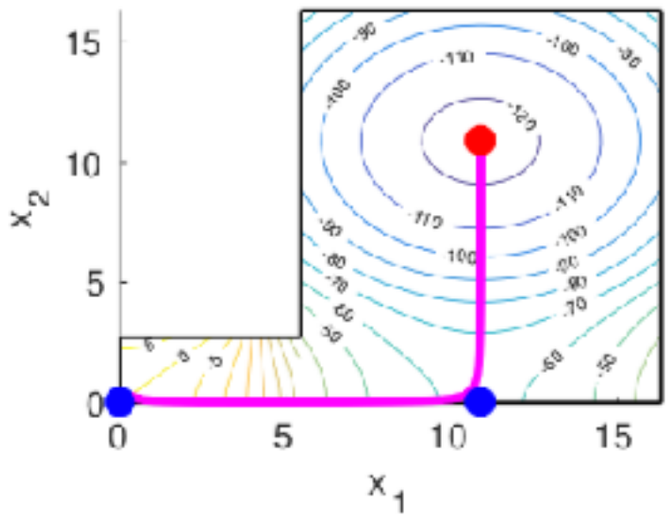


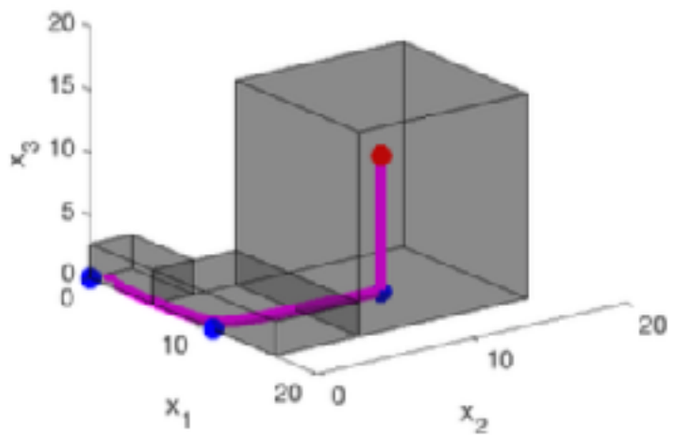
(a) Contour plot of the objective function and tube defined in 2D.



(b) Trajectory of gradient descent in the tube for $d = 3$.

INTUITION





$$R_1 = [0,1] \times [0,1].$$
$$R_2 = [1,3] \times [0,1].$$
$$R_3 = [1,3] \times [1,3].$$

- This function has minimum at (2,2) and saddle points at (0,0) and (2,0).
- Suppose initialization at R1; t_1:time to move from R1 to R2.
- Key observation: the time t_2 to transition R2->R3 is $t_2 \geq \kappa t_1$, with $\kappa > 1$.
- Therefore, escaping d saddle points requires $\Omega(\kappa^d)$ iterations.

BEYOND FIRST-ORDER METHODS

- There are other algorithms that efficiently escape saddle points, if allowed to use higher-order gradient information

- [Nesterov Polyak'06]: cubic regularization using Hessian.
- [Carmon et al,'16][Agarwal et al'16][Carmon & Duchi'16] showed convergence to second-order stationary points using only Hessian-vector product oracle.

LANDSCAPE OF CRITICAL POINTS

- Thus, studying the critical points of non-convex landscapes, and proving the strict-saddle property is a guarantee that gradient descent converges to global minima (eventually).
- Stochastic Gradient is theoretically motivated by escape time analysis (although not always with the correct noise).
- Successful analysis of strict saddles:
  - Tensor Decomposition [Lee et al.]
  - Matrix Completion [Ma et al.]
  - Deep Learning [Kawaguchi ‘16, Soltanolkotabi et al.’17]

TENSOR METHODS IN DEEP LEARNING

- Optimizing the training error with a generic deep network is a non-convex problem.

$$\min_{\Theta} \frac{1}{n} \sum_{i<n} \ell(y_i, \Phi(x_i; \Theta)) + \mathcal{R}(\Theta) \ .$$

- Consider a network of depth d with ReLU nonlinearities. Seen as a function of its parameters $\Theta, \Phi(x; \Theta)$ resembles a homogeneous piece-wise polynomial:

$$\Theta = \{\Theta^1, \ldots, \Theta^d\} \ .$$

$$\Phi(x; \Theta) = \sum_{p} \pi(x; \Theta) x_{p(1)} \prod_{j=1}^{d} \Theta^j_{p(j)} \ , \ \pi(x; \Theta) = \{0, 1\} \ .$$

TENSOR METHODS IN DEEP LEARNING

- Optimizing the training error with a generic deep network is a non-convex problem.

$$\min_{\Theta} \frac{1}{n} \sum_{i \leq n} \ell(y_i, \Phi(x_i; \Theta)) + \mathcal{R}(\Theta) \ .$$

- Consider a network of depth d with ReLU nonlinearities. Seen as a function of its parameters Θ, $\Phi(x; \Theta)$ resembles a homogeneous "piece-wise" polynomial:

$$\Phi(x; \Theta) = \sum_{p} \pi(x; \Theta) x_{p(1)} \prod_{j=1}^{d} \Theta^{j}_{p(j)} \ , \ \ \pi(x; \Theta) = \{0, 1\} \ .$$

$$\Theta = \{\Theta^1, \ldots, \Theta^d\} \ .$$

- The dependencies on Θ are partly captured by the d-order tensor

$$\Theta^1 \otimes \Theta^2 \ldots \otimes \Theta^d \ .$$

TENSOR METHODS

$$\min_{\Theta^1,\ldots,\Theta^d} F(Y, \Psi_X(\Theta^1,\ldots,\Theta^d)) + \mathcal{R}(\Theta^1,\ldots,\Theta^d) \ .$$

- Tensor factorizations are a broad class of non-convex optimization problems.

TENSOR METHODS

$$\min_{\Theta^1,\ldots,\Theta^d} F(Y, \Psi_X(\Theta^1,\ldots,\Theta^d)) + \mathcal{R}(\Theta^1,\ldots,\Theta^d) \ .$$

- Tensor factorizations are a broad class of non-convex optimization problems.

- A particularly famous instance is the matrix factorization problem:

$$\min_{U,V} \ \ell(Y, UV^T) + \mathcal{R}(U,V) \ , \ Y \in \mathbb{R}^{n\times m}, U \in \mathbb{R}^{n\times d}, V \in \mathbb{R}^{m\times d} \ .$$

  - Low-rank factorizations (e.g. PCA)
  - Sparse factorizations (Dictionary Learning, NMF)

MOTIVATION: MATRIX FACTORIZATION

- Example: low-rank factorization.

$$\min_{U,V} \ell(Y, UV^T) \ , \ \text{s.t. } \operatorname{rank}(UV^T) \leq r \ .$$

  - When $\ell(Y, X) = \|Y - X\|_{op}$, $\ell(Y, X) = \|Y - X\|_F$OK

- We can *lift* the problem and relax the constraint:

$$\min_X \ell(Y, X) + \lambda \|X\|_* \ , \quad \|X\|_* = \text{Nuclear norm of } X.$$

- Factorized and relaxed formulations are connected via a variational principle:

$$\|X\|_* = \min_{UV^T = X} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) \ .$$

MOTIVATION: MATRIX FACTORIZATION

- Example: low-rank factorization.

$$\min_{U,V} \ell(Y, UV^T) \ , \ \text{s.t.} \ \operatorname{rank}(UV^T) \leq r \ .$$

  - When $\ell(Y, X) = \|Y - X\|_{op}$, $\ell(Y, X) = \|Y - X\|_F$ OK
  - We can *lift* the problem and relax the constraint:

$$\min_X \ell(Y, X) + \lambda \|X\|_* \ , \qquad \|X\|_* = \text{Nuclear norm of } X.$$

  - Factorized and relaxed formulations are connected via a variational principle:

$$\|X\|_* = \min_{UV^T = X} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) \ .$$

- Q: General case?

TENSOR NORMS [BACH, HAEFFELE&VIDAL]

- A first generalization is the tensor norm

$$\|X\|_{u,v} = \inf_r \min_{UV^T=X} \frac{1}{2}\left(\sum_i \|U_i\|_u^2 + \|V_i\|_v^2\right) .$$

Theorem [H-V]: A local minimizer of the factorized problem $\min_{U,V} \ell(Y, UV^T) + \lambda \sum_{i<r} \|U_i\|_u \|V_i\|_v$ such that for some i $U_i = V_i = 0$ is a global minimizer of the convex problem $\min_X \ell(Y, X) + \lambda \|X\|_{u,v}$ as well as the factorized problem.

TENSOR NORMS [BACH, HAEFFELE&VIDAL]

- A first generalization is the tensor norm

$$\|X\|_{u,v} = \inf_r \min_{UV^T=X} \frac{1}{2}\left(\sum_i \|U_i\|_u^2 + \|V_i\|_v^2\right) .$$

Theorem [H-V]: A local minimizer of the factorized problem $\min_{U,V} \ell(Y, UV^T) + \lambda \sum_{i<r} \|U_i\|_u \|V_i\|_v$ such that for some i $U_i = V_i = 0$ is a global minimizer of the convex problem $\min_X \ell(Y, X) + \lambda \|X\|_{u,v}$ as well as the factorized problem.

- This produces an *optimality certificate:* we use a surrogate convex problem to obtain a guarantee that a non-convex problem is solved optimally.

FROM TENSOR FACTORIZATIONS TO DEEP NETS

- We start by generalizing a multilinear mapping (tensor) to homogeneous maps $\phi(\Theta^1, \ldots, \Theta^d)$:

s: degree of homogeneity.

$$\forall\ \Theta, \forall\ \alpha \geq 0\ ,\ \phi(\alpha\Theta^1, \ldots, \alpha\Theta^d) = \alpha^s \phi(\Theta^1, \ldots, \Theta^d)\ .$$

Ex: ReLU $\rho(x) = \max(0, x)$ is homogeneous of degree 1.

FROM TENSOR FACTORIZATIONS TO DEEP NETS

- We start by generalizing a multilinear mapping (tensor) to homogeneous maps $\phi(\Theta^1, \ldots, \Theta^d)$:

$$\forall\ \Theta\,, \forall\ \alpha \geq 0\ ,\ \phi(\alpha\Theta^1, \ldots, \alpha\Theta^d) = \alpha^s \phi(\Theta^1, \ldots, \Theta^d)\ .$$

s: degree of homogeneity.

Ex: ReLU $\rho(x) = \max(0, x)$ is homogeneous of degree 1.

- We construct models by adding r copies of homogenous maps:

$$\Phi_r(\Theta^1, \ldots, \Theta^d) = \sum_{i \leq r} \phi(\Theta_i^1, \ldots, \Theta_i^d)\ .$$

FROM TENSOR FACTORIZATIONS TO DEEP NETS

- We start by generalizing a multilinear mapping (tensor) to homogeneous maps $\phi(\Theta^1, \ldots, \Theta^d)$:

$$\forall\ \Theta, \forall\ \alpha \geq 0\ ,\ \ \phi(\alpha\Theta^1, \ldots, \alpha\Theta^d) = \alpha^s \phi(\Theta^1, \ldots, \Theta^d)\ .$$

Ex: ReLU $\rho(x) = \max(0, x)$ is homogeneous of degree 1.

- We construct models by adding r copies of homogenous maps:

$$\Phi_r(\Theta^1, \ldots, \Theta^d) = \sum_{i \leq r} \phi(\Theta_i^1, \ldots, \Theta_i^d)\ .$$

- We consider

$$\min_{\Theta^1, \ldots, \Theta^d} \ell(Y, \Phi_r(\Theta^1, \ldots, \Theta^d)) + \lambda \mathcal{R}(\Theta^1, \ldots, \Theta^d)\ ,$$

Key assumption: $\mathcal{R}$ is positively homogeneous of the same degree as Φ.

FROM TENSOR FACTORIZATIONS TO DEEP NETS

$$\Phi_r(\Theta^1, \ldots, \Theta^d) = \sum_{i=1}^{r} \phi(\Theta^1, \ldots, \Theta^d) \ .$$

Examples

Matrices: $\Phi(U, V) = UV^T = \sum_{i=1}^{r} U_i V_i^T \ \ (\phi(U_i, V_i) = U_i V_i^T) \ .$

Higher-order Tensors:

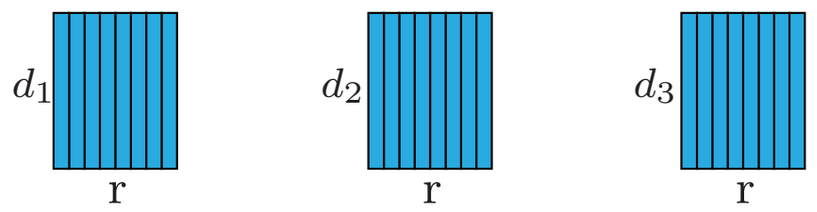
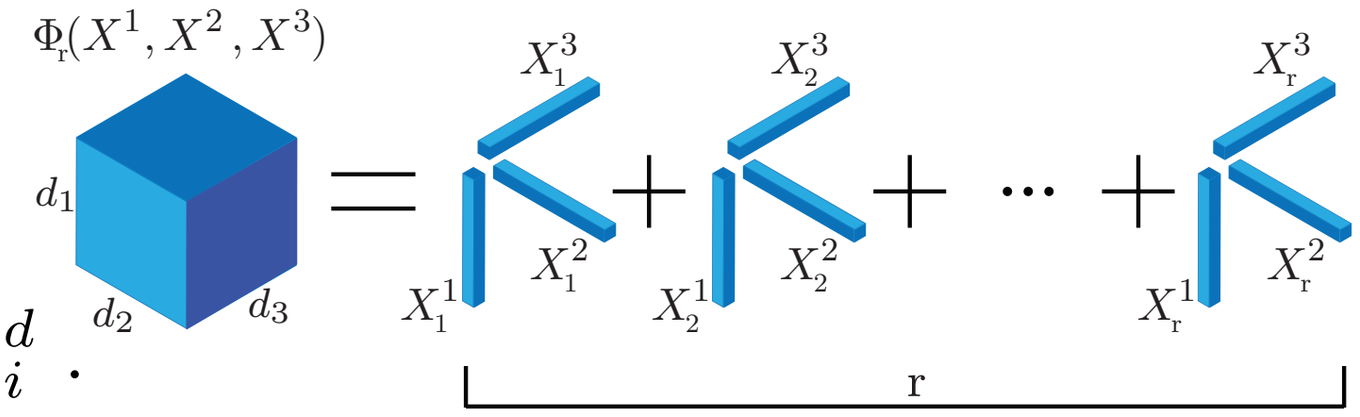


figure credit: R. Vidal



$$\phi(\Theta_i^1, \ldots, \Theta_i^d) = \Theta_i^1 \otimes \cdots \otimes \Theta_i^d \ .$$

Candecomp/Parafac (CP) Tensor decomposition.

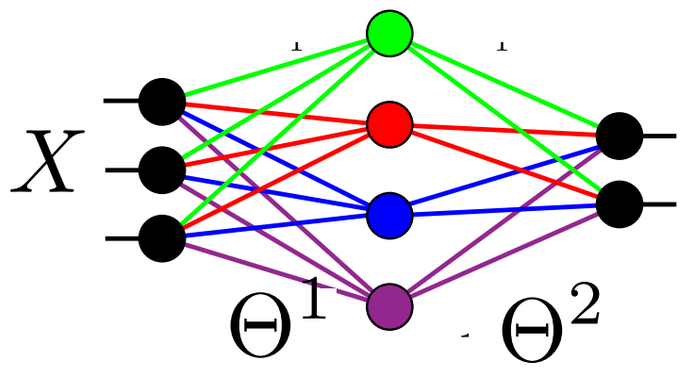
ADAPTATION TO DEEP MODELS

$$\Phi_r(\Theta^1, \ldots, \Theta^d) = \sum_{i=1}^{r} \phi(\Theta^1, \ldots, \Theta^d) \ .$$

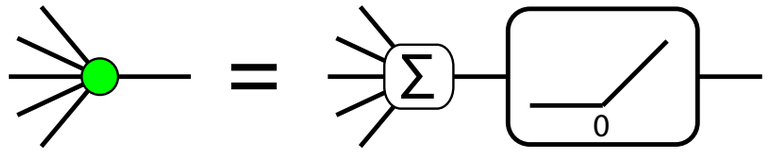
➤

ReLU Network:

ReLU Network with One Hidden Layer



Rectified Linear Unit (ReLU)



$\phi(\Theta^1, \Theta^2)$

Multilayer ReLU
Parallel Network

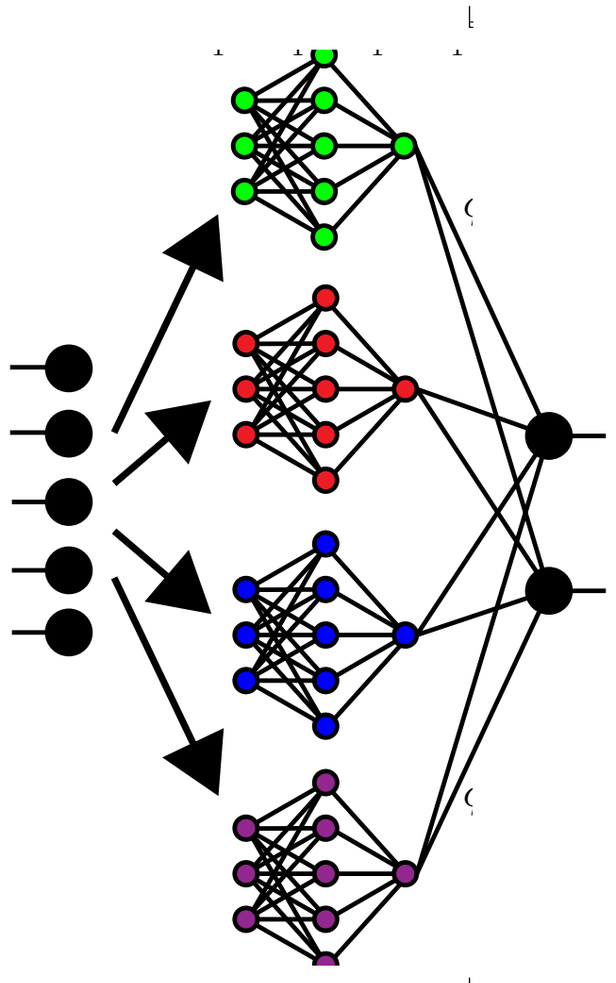


figure credit: R. Vidal

$$\Phi(\Theta^1, \ldots, \Theta^d) = \sum_i \phi(\Theta_i^1, \ldots, \Theta_i^d) \ .$$

ADAPTATION TO DEEP MODELS

- In the matrix case, the variational principle was

$$\|X\|_{u,v} = \min_{UV^T=X} \sum_{i\leq r} \|U_i\|_u \|V_i\|_v \ .$$

ADAPTATION TO DEEP MODELS

- In the matrix case, the variational principle was

$$\|X\|_{u,v} = \min_{UV^T=X} \sum_{i\leq r} \|U_i\|_u \|V_i\|_v \ .$$

- This is generalized to

$$\mathcal{R}(\Theta) = \min_{\Theta^1,\ldots,\Theta^d} \sum_{i\leq r} g(\Theta_i^1,\ldots,\Theta_i^d) \ , \ s.t. \ \Phi_r(\Theta^1,\ldots,\Theta^d) = \Theta \ .$$

Proposition [H-V]: $\mathcal{R}$ is convex.
Also, if g is positively homogeneous of degree s, so is $\mathcal{R}$.

ADAPTATION TO DEEP MODELS

Theorem [H-V]: A local minimizer of the factorized problem

$$\min_{\Theta^k} \ell(Y, \sum_{i \leq r} \phi_r(\Theta_i^k)) + \lambda \sum_{i \leq r} g(\Theta_i^k)$$

such that for some i and all k $\Theta_i^k = 0$ is a global minimizer for both factorized problem and the convex formulation

$$\min_{\Theta} \ell(Y, \Theta) + \lambda \mathcal{R}(\Theta).$$

ADAPTATION TO DEEP MODELS

Theorem [H-V]: A local minimizer of the factorized problem

$$\min_{\Theta^k} \ell(Y, \sum_{i \leq r} \phi_r(\Theta_i^k)) + \lambda \sum_{i \leq r} g(\Theta_i^k)$$

such that for some i and all k $\Theta_i^k = 0$ is a global minimizer for both factorized problem and the convex formulation

$$\min_{\Theta} \ell(Y, \Theta) + \lambda \mathcal{R}(\Theta).$$

- Global optimality certificate for a broad class of non-convex optimization problems, including some form of deep learning architectures.
- Q: How to use this certificate in practice?

ADAPTATION TO DEEP MODELS

- Pros
  - Global optimality certificate, easy to check
  - Inclues nonlinear models as long as they are homogeneous.
  - Provides a possible meta-algorithm: increase the lifting value r progressively is local optimum does not very condition.

- Cons
  - How much do we need to increase r in practice?
  - How stringent is the homogenous regularization condition?

TENSOR DECOMPOSITIONS AND NEURAL NETS

- Suppose a label generating model of the form

$$\mathbb{E}(y|x) = f_0(x) = \langle a_2, \sigma(A_1 x + b_1) \rangle + b_2 \ ,$$

$\sigma(\cdot)$: point-wise nonlinearity
$A_1 \in \mathbb{R}^{d \times k}$.

TENSOR DECOMPOSITIONS AND NEURAL NETS

- Suppose a label generating model of the form

$$\mathbb{E}(y|x) = f_0(x) = \langle a_2, \sigma(A_1 x + b_1) \rangle + b_2 \ ,$$

$\sigma(\cdot)$: point-wise nonlinearity
$A_1 \in \mathbb{R}^{d \times k}$.

- Q: Given training samples $\{(x_i, y_i)\,;\, y_i = f_0(x_i)\}_{i \leq n}$, can we estimate the parameters a_2, A_1, b_1, b_2 with provable risk?
- Q: Using a computationally efficient algorithm?

BREAKING THE PERILS OF (...) [Janzamin, Sedghi, Anandkumar]

- If one assumes knowledge of the input distribution $p(x)$, then one can exploit the relationship between score functions and conditional expectations:

Def: The m-th order score function $S_m(x)$ is the m-th order tensor

$$S_m(x) = (-1)^m \frac{\nabla^m p(x)}{p(x)} \ .$$

Proposition: If $f(x) = \mathbb{E}(y|x)$, then

$$\mathbb{E}(y \cdot S_3(x)) = \mathbb{E}(\nabla^3 f(x)) \ .$$

BREAKING THE PERILS OF (. . .) [Janzamin, Sedghi, Anandkumar]

- If one assumes knowledge of the input distribution $p(x)$, then one can exploit the relationship between score functions and conditional expectations.

- It follows that when $\mathbb{E}(y|x) = f_0(x)$, we have

$$\mathbb{E}(y \cdot S_3(x)) = \sum_{j \leq k} \lambda_j (A_1)_j \otimes (A_1)_j \otimes (A_1)_j \ \in \mathbb{R}^{d \times d \times d}, \ \lambda_j \in \mathbb{R} \ .$$

BREAKING THE PERILS OF (...) [Janzamin, Sedghi, Anandkumar]

➤ Learning generalization bound in the "realizable" setting:

Theorem: The tensor algorithm *NN-Lift* learns the target function $\mathbb{E}(y|x) = f_0(x)$ up to error ϵ when the number of samples is of the order of

$$n \geq O\left(\frac{kd^3}{\epsilon^2}\frac{\lambda_{max}(A_1)^2}{\lambda_{min}(A_1)^6}\right) .$$

(k: size of hidden layer)
(d: input dimension)

➤ Comments:

- ➤ Polynomial sample complexity.
- ➤ Algorithm has polynomial complexity as well.
- ➤ Extension to non-realizable setting (see paper for details).

BREAKING THE PERILS OF (. . .) [Janzamin, Sedghi, Anandkumar]

- Pros
  - Statistical Guarantees that also incorporate computational feasibility.
  - Learning is essentially reduced to finding low-rank tensor factorizations.

- Cons
  - very strong hypothesis: knowledge of p(x).
  - only a particular Neural network architecture (one hidden layer so far).
  - restrictive class of nonlinearities? : the proof requires

$$\mathbb{E}(\sigma'''(z)) \ , \ \mathbb{E}(\sigma''(z))$$

DEEP NETWORKS AND SPIN GLASSES

- Suppose we have a linear deep network:

$$\Phi(x; \Theta_1, \ldots, \Theta_K) = \Theta_K \Theta_{K-1} \ldots \Theta_1 x \ .$$

- And suppose we train using least squares regression:

$$E(\Theta) = \frac{1}{n} \sum_{i \leq n} \|y_i - \Phi(x_i; \Theta)\|^2 \ .$$

DEEP NETWORKS AND SPIN GLASSES

- Suppose we have a linear deep network:

$$\Phi(x; \Theta_1, \ldots, \Theta_K) = \Theta_K \Theta_{K-1} \ldots \Theta_1 x \ .$$

- And suppose we train using least squares regression:

$$E(\Theta) = \frac{1}{n} \sum_{i<n} \left\| y_i - \Phi(x_i; \Theta) \right\|^2 \ .$$

$$(\Theta_1 x)^j = \sum_l \Theta_1^{j,l} x^l \ ,$$

- In coordinates,

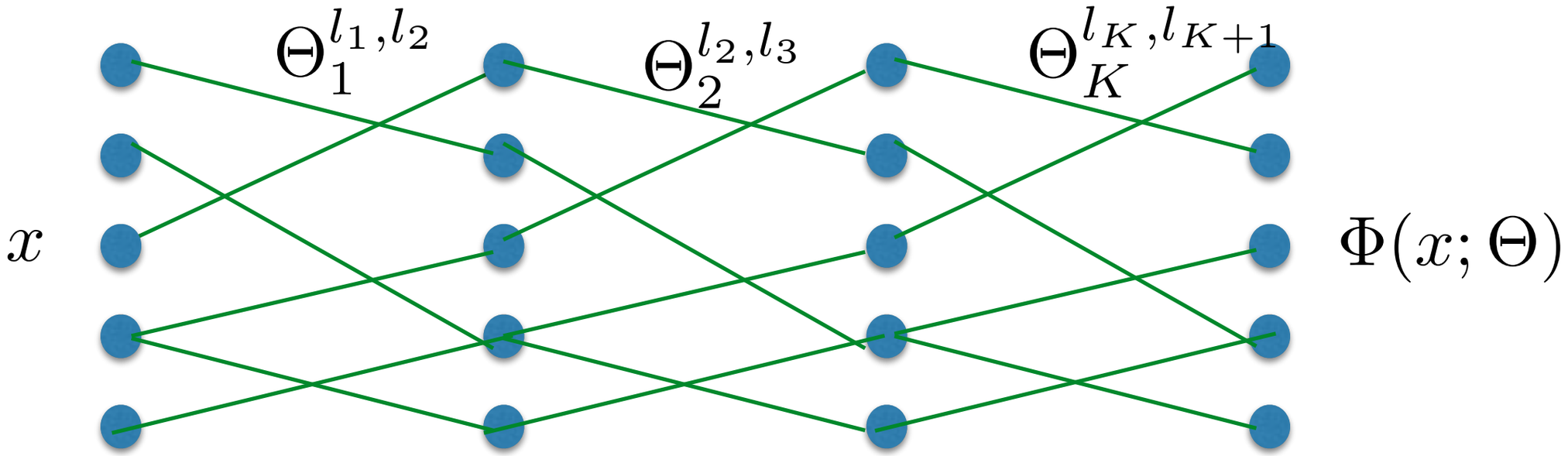
$$(\Theta_2 \Theta_1 x)^j = \sum_{l_1, l_2} \Theta_2^{j,l_2} \Theta_1^{l_2,l_1} x^{l_1} \ ,$$

$$(\Theta_K \ldots \Theta_2 \Theta_1 x)^j = \sum_{l_1, \ldots l_K} x^{l_1} \Theta_K^{j,l_K} \prod_{k=2}^{K-1} \Theta_k^{l_k, l_{k-1}} \ .$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Equivalently, we can define *paths*



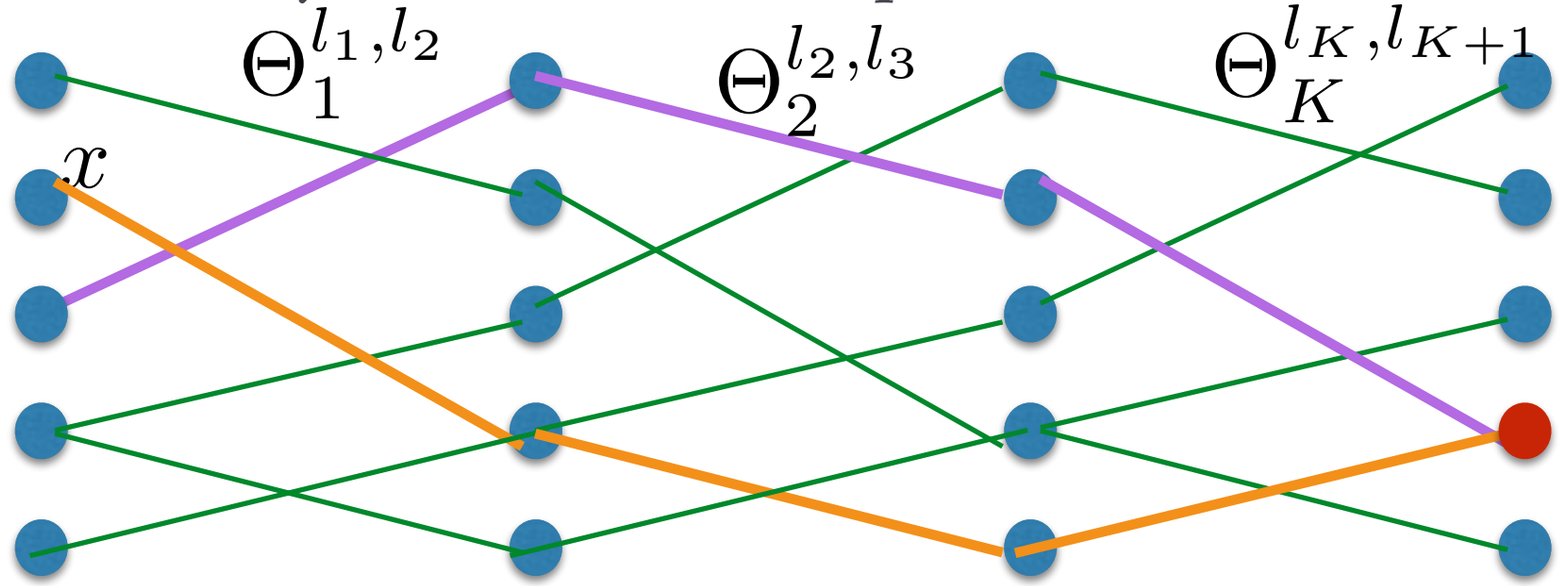
$$p = (l_0, l_1, \ldots, l_{K+1})$$

$$\mathcal{P} = \{p = (l_0, \ldots, l_{K+1}); 1 \leq l_k \leq M_k\}$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Equivalently, we can define *paths*



$\Phi(x;\Theta)$

$$p = (l_0, l_1, \ldots, l_{K+1})$$

$$\mathcal{P} = \{p = (l_0, \ldots, l_{K+1}); 1 \leq l_k \leq M_k\}$$

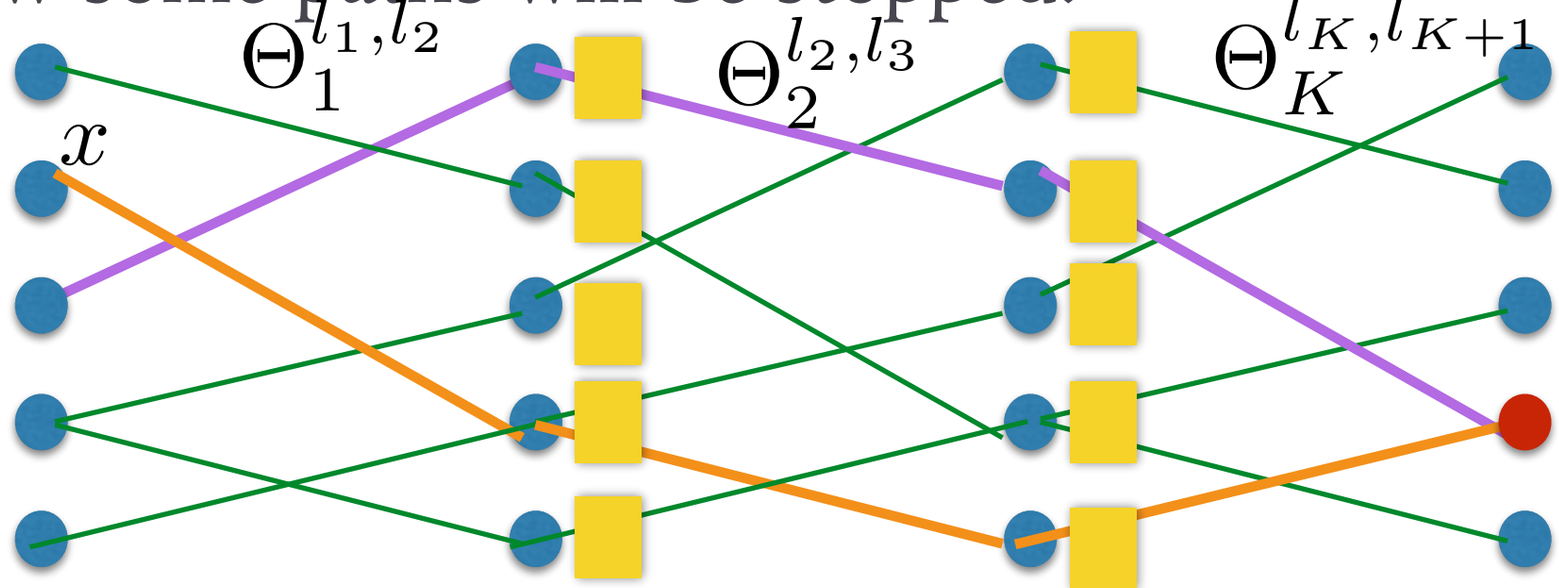
- Homogeneous polynomial on Θ.
- Q: What about a ReLU network instead?

$$\Phi(x;\Theta)^j = \sum_{p \in \mathcal{P}; p(K+1)=j} x^{p(1)} \prod_{k \leq K} \Theta_k^{p(k),p(k+1)} .$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Now some paths will be stopped:



-

$$\Phi(x;\Theta)^j = \sum_{p\in\mathcal{P};p(K+1)=j} \pi(p,x,\Theta)\cdot x^{p(1)} \prod_{k\leq K} \Theta_k^{p(k),p(k+1)} \ , \ \pi(p,x,\Theta) = \{0,1\}$$

$$p = (l_0,\ldots,l_K) \ , \ \tilde{p} = (l_0,\ldots,l_{K-1})$$

- Biases produce low-order terms (we ignore them for now)

$$\pi(p,x,\Theta) = \pi(\tilde{p},x,\Theta)\cdot\left(\sum_{p'\in\tilde{\mathcal{P}};p'(K)=p(K)} \pi(p',x,\Theta)\prod_{k<K}\Theta_k^{p'(k),p'(k+1)} > 0\right) .$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Loss becomes

$$E(\Theta) = \frac{1}{n}\sum_{i\leq n} \|y_i - \Phi(x_i;\Theta)\|^2$$

$$= \frac{1}{n}\sum_{i\leq n}\sum_{j=1}^{M_K}\left(y_i^j - \sum_{p\in\mathcal{P};p(K+1)=j} \pi(p,x_i,\Theta)\cdot x_i^{p(1)}\prod_{k\leq K}\Theta_k^{p(k),p(k+1)}\right)^2$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Loss becomes

$$E(\Theta) = \frac{1}{n}\sum_{i\leq n} \|y_i - \Phi(x_i;\Theta)\|^2$$

$$= \frac{1}{n}\sum_{i\leq n}\sum_{j=1}^{M_K}\left(y_i^j - \sum_{p\in\mathcal{P};p(K+1)=j} \pi(p,x_i,\Theta)\cdot x_i^{p(1)}\prod_{k\leq K}\Theta_k^{p(k),p(k+1)}\right)^2$$

$$\stackrel{n\to\infty}{\to} C + \sum_{p\in\mathcal{P}} q(X,Y,\Theta,p)\prod_{k\leq K}\Theta_k^{p(k),p(k+1)}$$

$$+ \sum_{p,p'\in\mathcal{P}} Q(X,\Theta,p,p')\prod_{k\leq K}\Theta_k^{p(k),p(k+1)}\Theta_k^{p'(k),p'(k+1)} \ , \text{ with}$$

$$q(X,Y,\Theta,p) = \mathbb{E}_{X,Y}\left(\pi(p,X,\Theta)Y^{p(K)}X^{p(1)}\right) \ ,$$

$$Q(X,\Theta,p,p') = \mathbb{E}_X\left(\pi(p,X,\Theta)\pi(p',X,\Theta)X^{p(1)}X^{p'(1)}\right) \ .$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- The loss "looks" like a polynomial in Θ provided we **break the dependency** of $\pi(p, x, \Theta)$ with respect to Θ .
  - It means that thresholding is independent of Θ.
- For large enough n (assuming iid samples), it results that

$$q(X, Y, p) \sim \mathcal{N}(\mu_p, \sigma_p^2) \ ,$$

$$Q(X, p, p') \sim \mathcal{N}(\mu_{p,p'}, \sigma_{p,p'}^2) \ ,$$

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- Furthermore, if one also assumes *redundancy* (weights shared across layers), *uniformity* (same weights are not used too often along surviving paths) and *normalized weights,* authors arrive at

$$E(\Theta) \simeq \mathcal{L}_{\Lambda,K}(\Theta) = \frac{1}{\Lambda^{(K-1)/2}} \sum_{l_1,\ldots,l_K=1}^{\Lambda} Z_{l_1,\ldots,l_K}\Theta_{l_1}\ldots\Theta_{l_K} \ ,$$

with $\|\Theta\|^2 = \Lambda$.

$\mathcal{L}_H(\Theta)$: Hamiltonian of the H-spin spherical spin glass model.

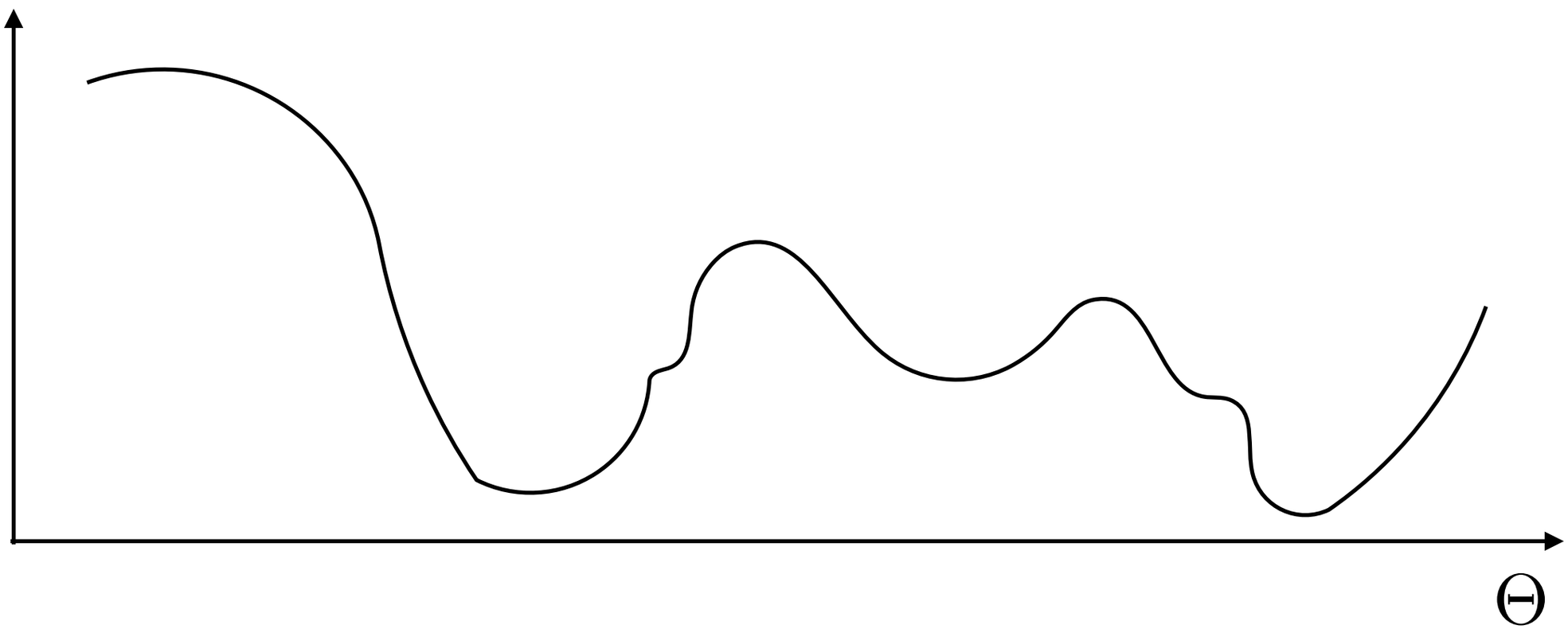
$Z_p \sim \mathcal{N}(0,\sigma^2)$.

[Choromaska, Henaff, Mathieu, LeCun, Ben Arous,'14]

DEEP NETWORKS AND SPIN GLASSES

- [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:



LANDSCAPE OF SPHERICAL SPIN GLASSES

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

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

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

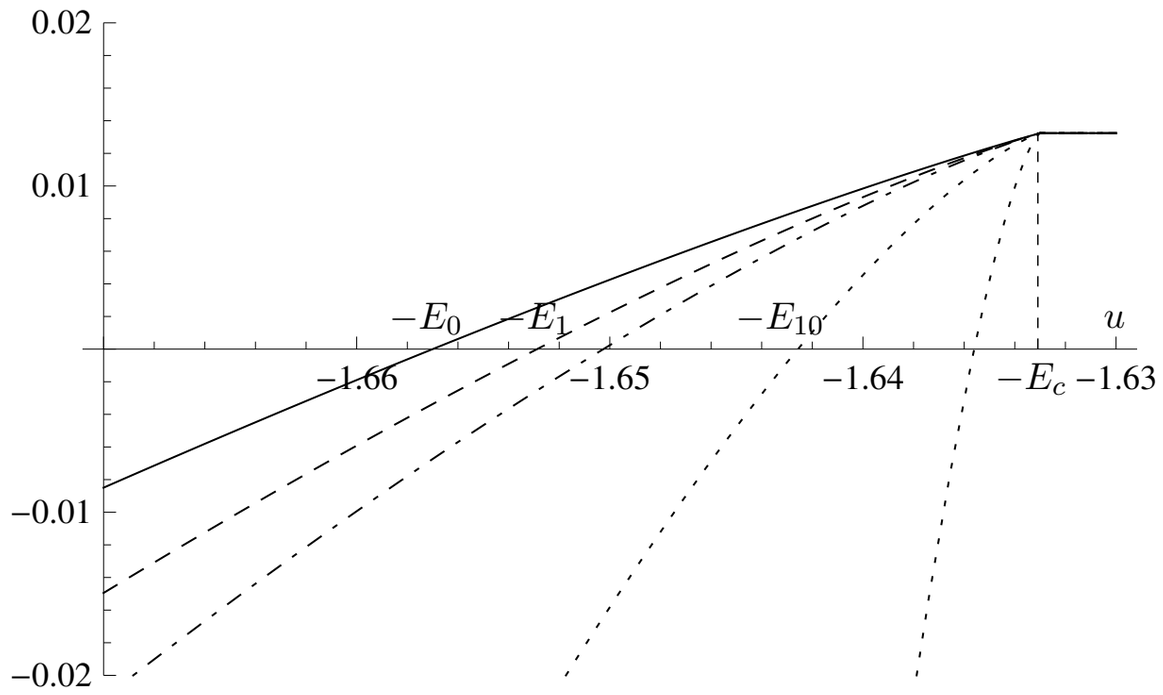


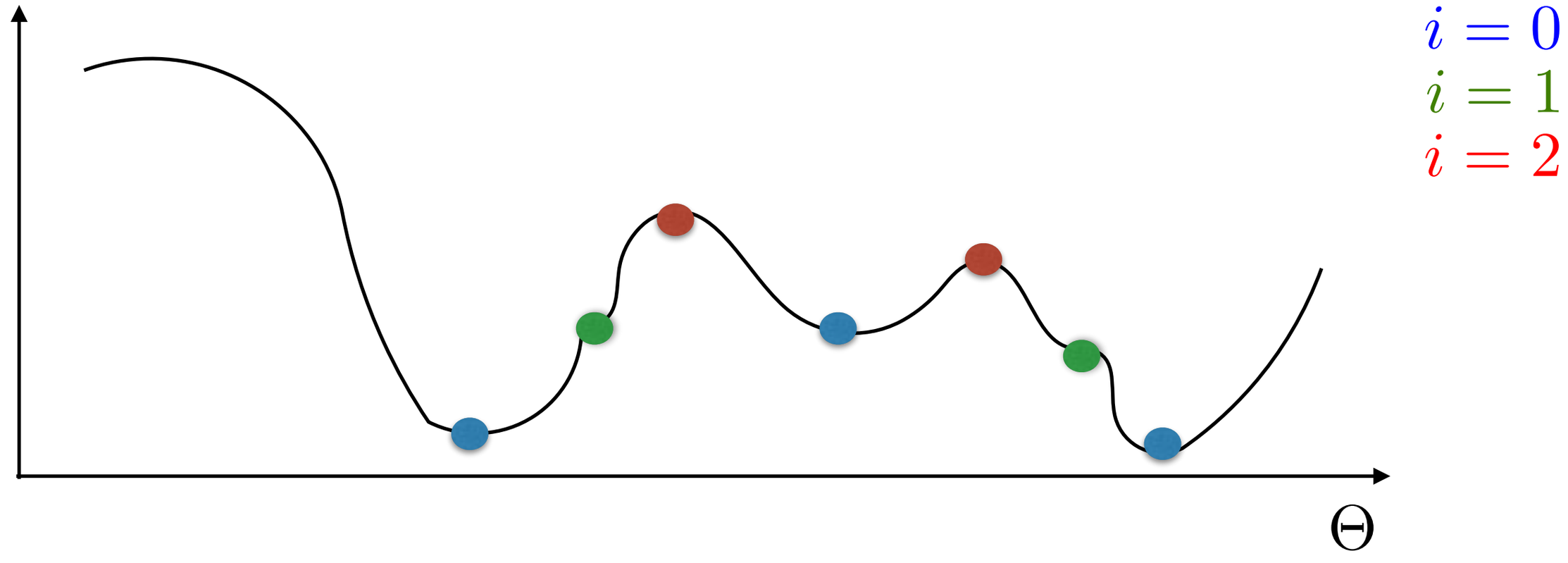
FIGURE 1. The functions $\Theta_{k,p}$ for $p = 3$ and $k = 0$ (solid), $k = 1$ (dashed), $k = 2$ (dash-dotted), $k = 10$, $k = 100$ (both dotted). All these functions agree for $u \geq -E_\infty$.

[Auffinger, Ben Arous Cerny,'11]

DEEP NETWORKS AND SPIN GLASSES

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

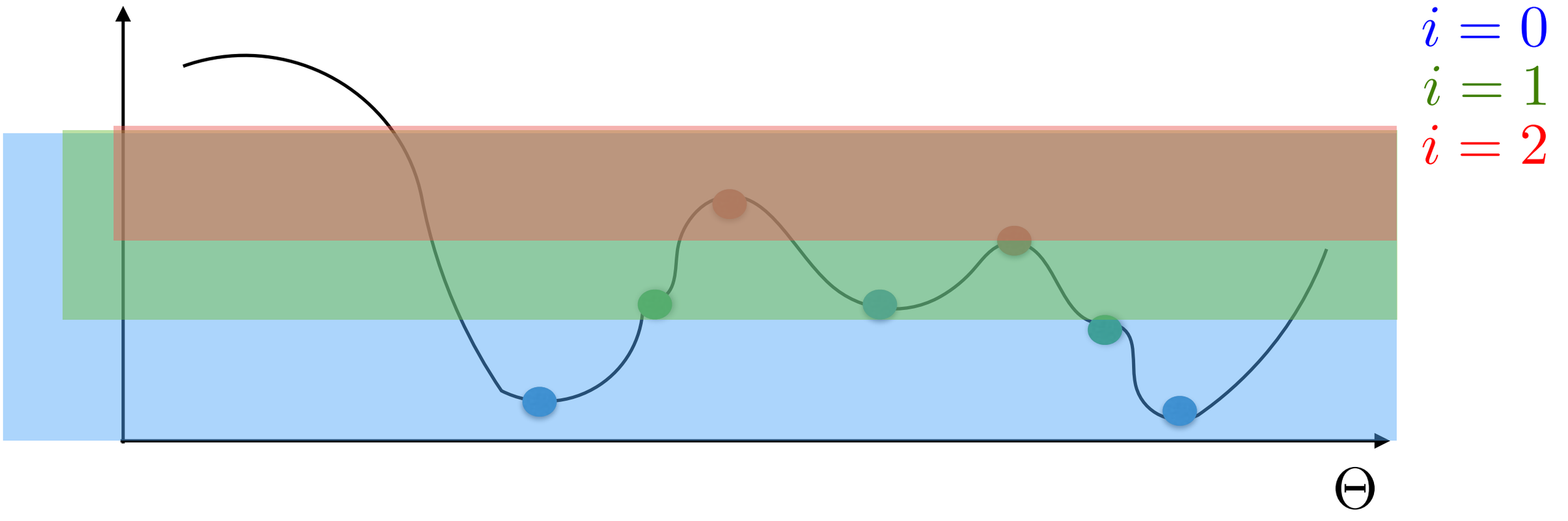
- In particular, index of critical points (ratio of negative to positive eigenvalues of the Hessian) occur at different energy bands:



DEEP NETWORKS AND SPIN GLASSES

- As $\Lambda \to \infty$, the distributions concentrate along different bands: each index concentrates in different bands.

- As $\Lambda \to \infty$, the number of local minima dominate the rest of the indices.



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.

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.

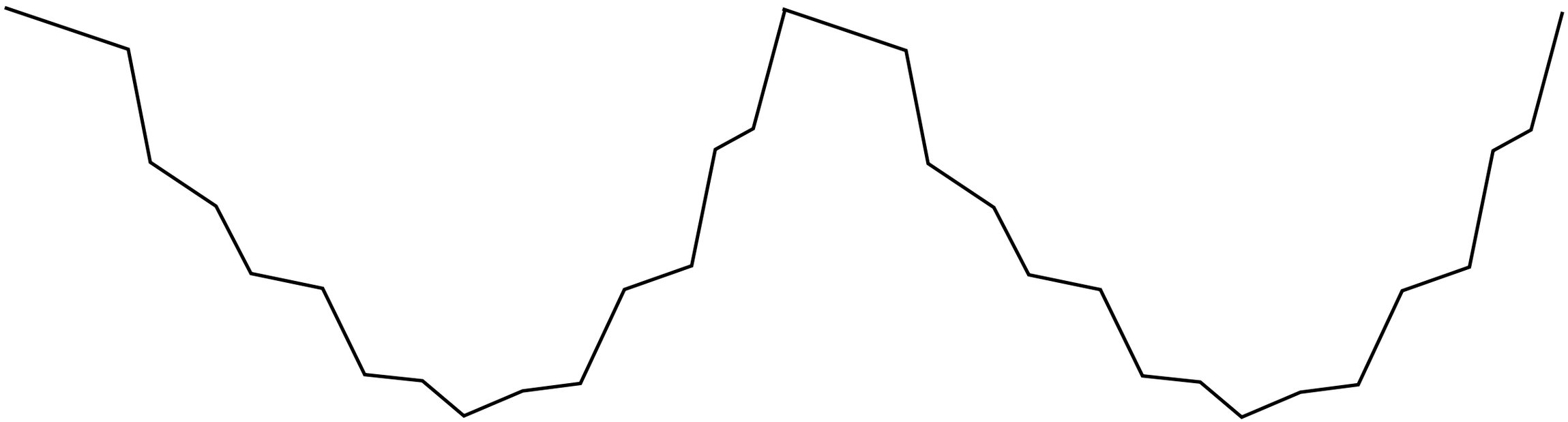
NON-CONVEXITY ≠ NOT OPTIMIZABLE

- 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.

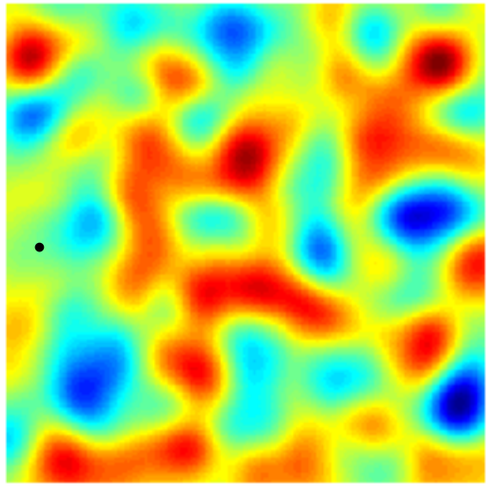
$$F(\theta) = F(g.\theta) \ , \ g \in G \text{ compact.}$$



ANALYSIS OF NON-CONVEX LOSS SURFACES

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

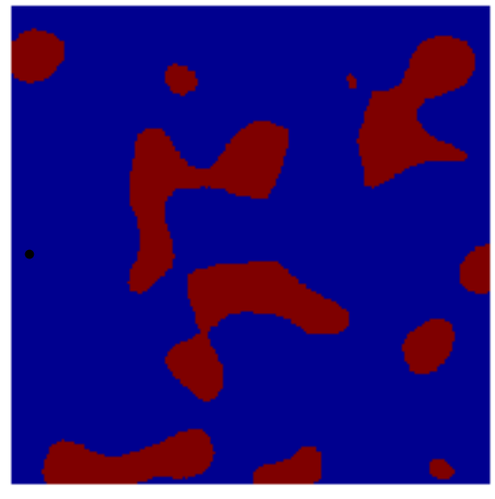
$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\} \ .$$



ANALYSIS OF NON-CONVEX LOSS SURFACES

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

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

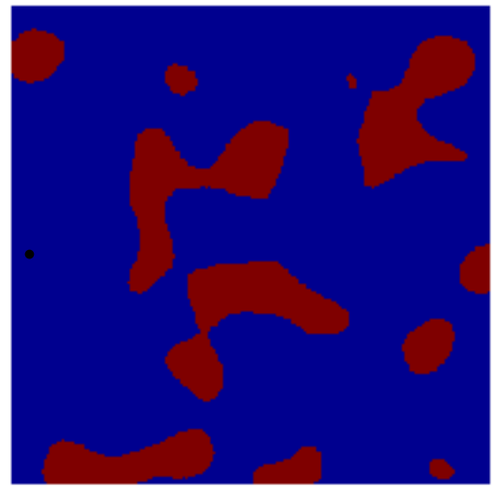


- A first notion we address is about the topology of the level sets Ω_u.
- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u ?

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

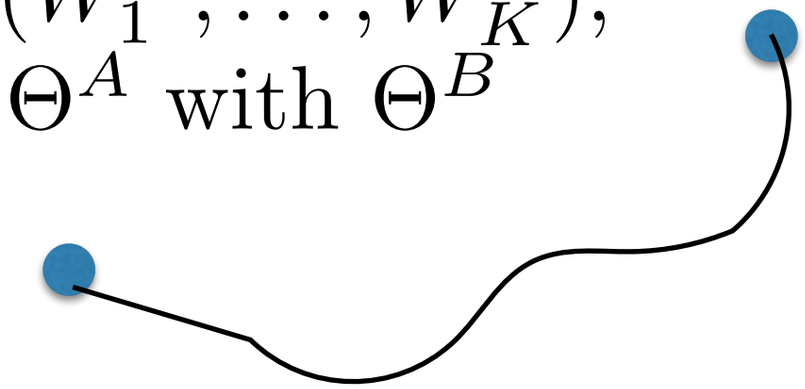
- We pay extra redundancy price to get simple topology.
- This simple topology is an "artifact" of the linearity of the network:

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

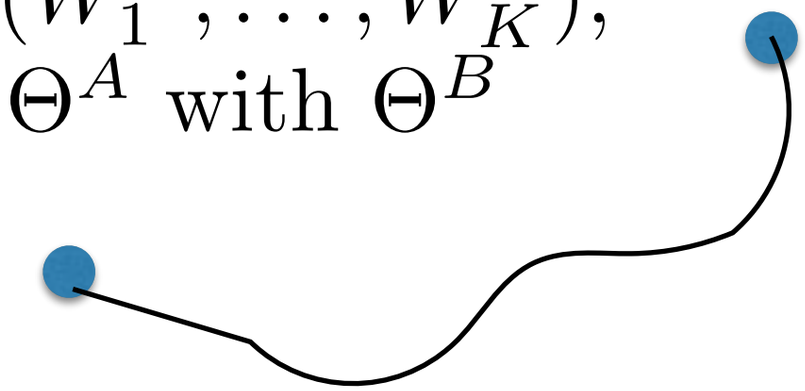
- Goal:

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 .

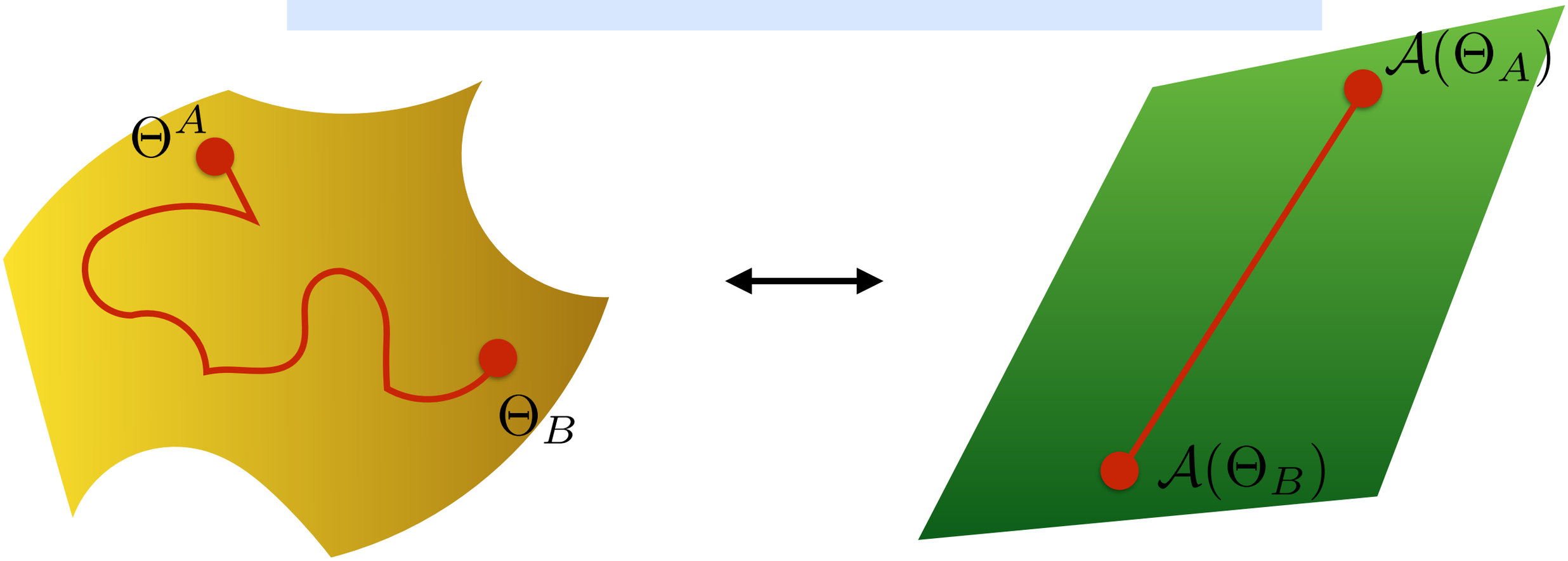
- Result is based on local linearization of the ReLU kernel (hence exponential price).

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

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

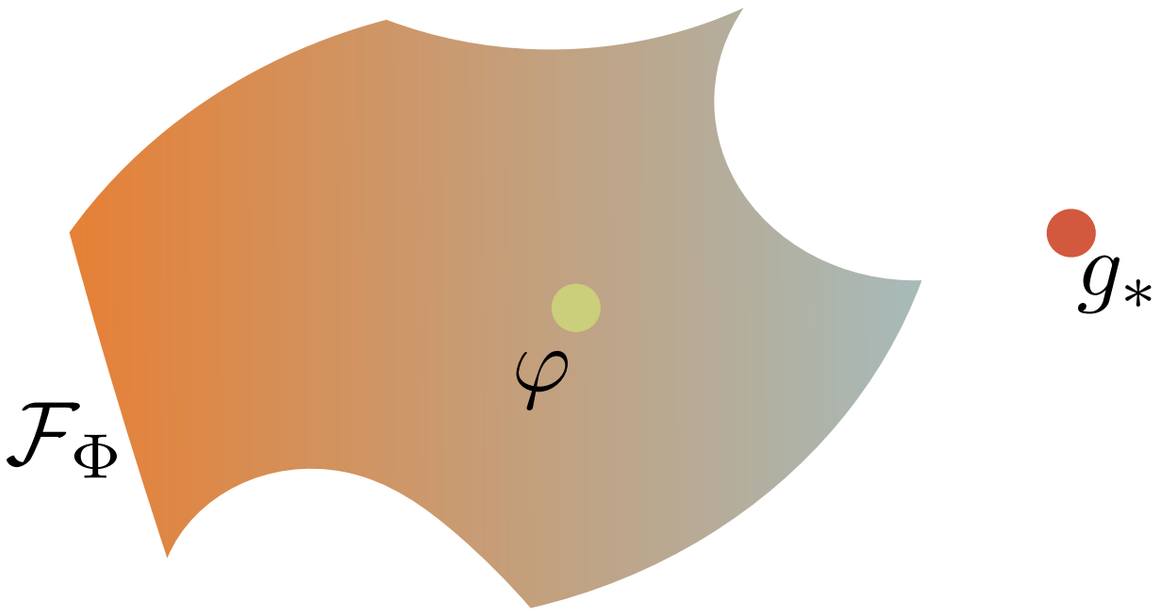
Theorem: [BVV'18] If $\dim\{\mathcal{A}(w),\, w \in \mathbb{R}^n\} = q < \infty$, then $L(U,W) = \mathbb{E}\|U\rho(WX) - Y\|^2$ has no spurious valley if $M \geq q$.

- 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:

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



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

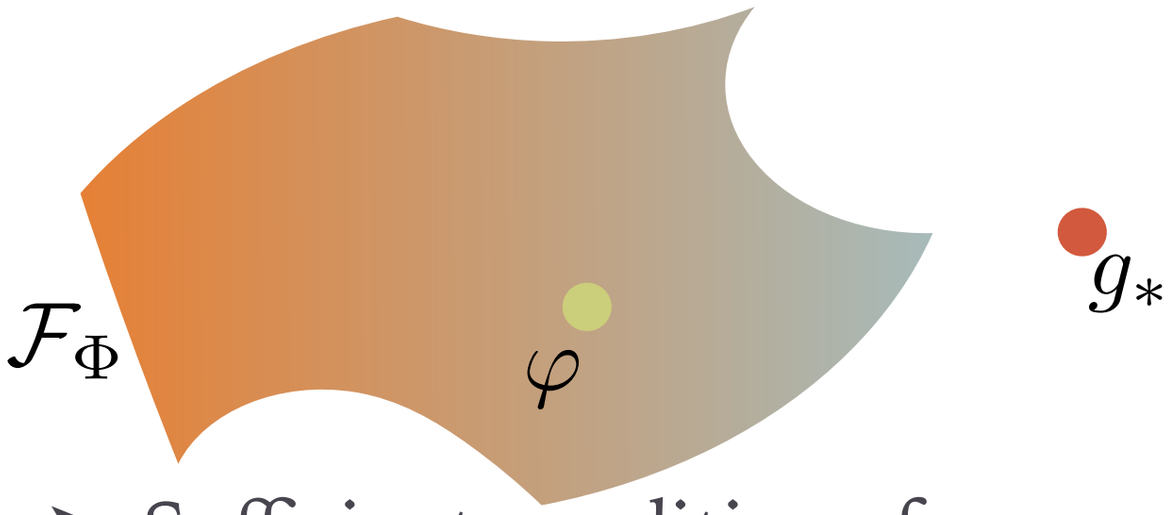
$$g_* : x \mapsto \mathbb{E}(Y|x)$$

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

PARAMETRIC VS MANIFOLD OPTIMIZATION

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

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



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

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

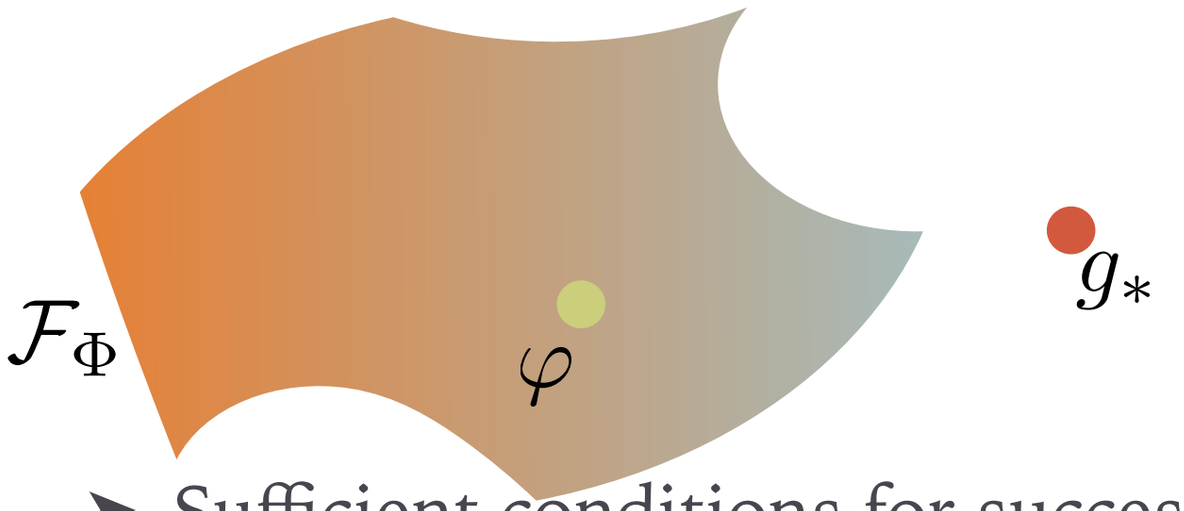
$$\langle f, g \rangle_p := \mathbb{E}\{f(X)g(X)\} \,.$$

- 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

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

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

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

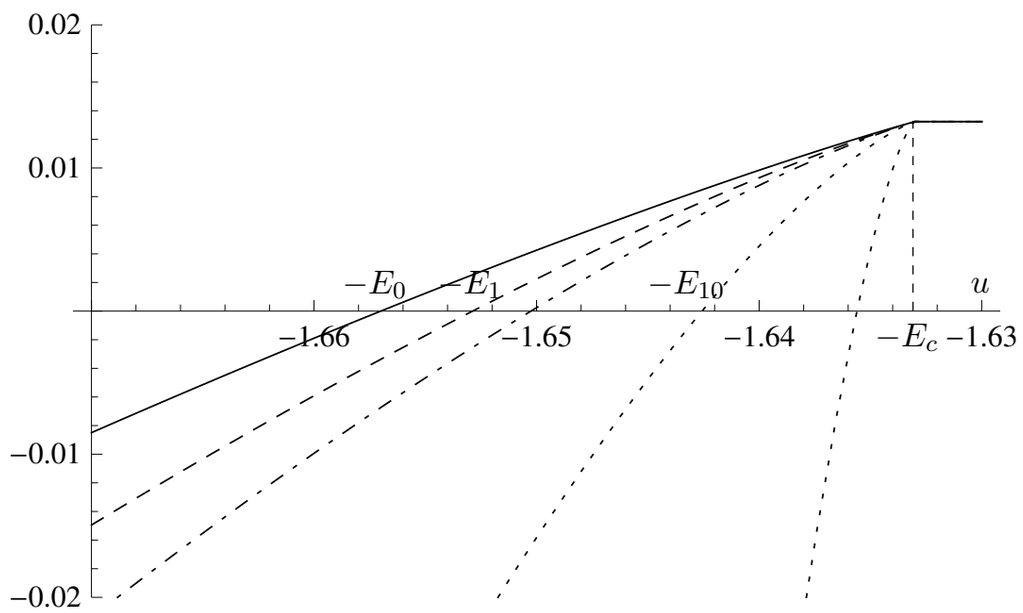


FIGURE 1. The functions $\Theta_{k,p}$ for $p = 3$ and $k = 0$ (solid), $k = 1$ (dashed), $k = 2$ (dash-dotted), $k = 10$, $k = 100$ (both dotted). All these functions agree for $u \geq -E_\infty$.

[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

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

  - 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

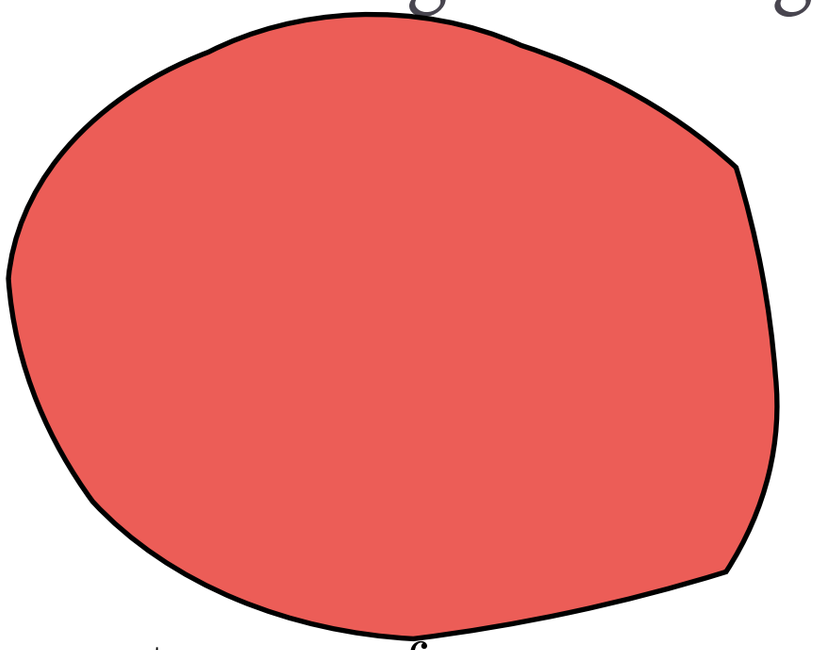
$$\beta(M,N) = \inf_{S;\dim(S)=f^{-1}(M)} \inf_{\substack{U\in\mathbb{R}^{m\times M}\\ W\in\mathbb{R}^{M\times f^{-1}(M)}}} \sup_{\substack{\mathbb{E}\|Z\|\le N-f^{-1}(M),\\ P_S Z=0}} \mathbb{E}\|U\rho(WP_S X+Z)-Y\|^2$$

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

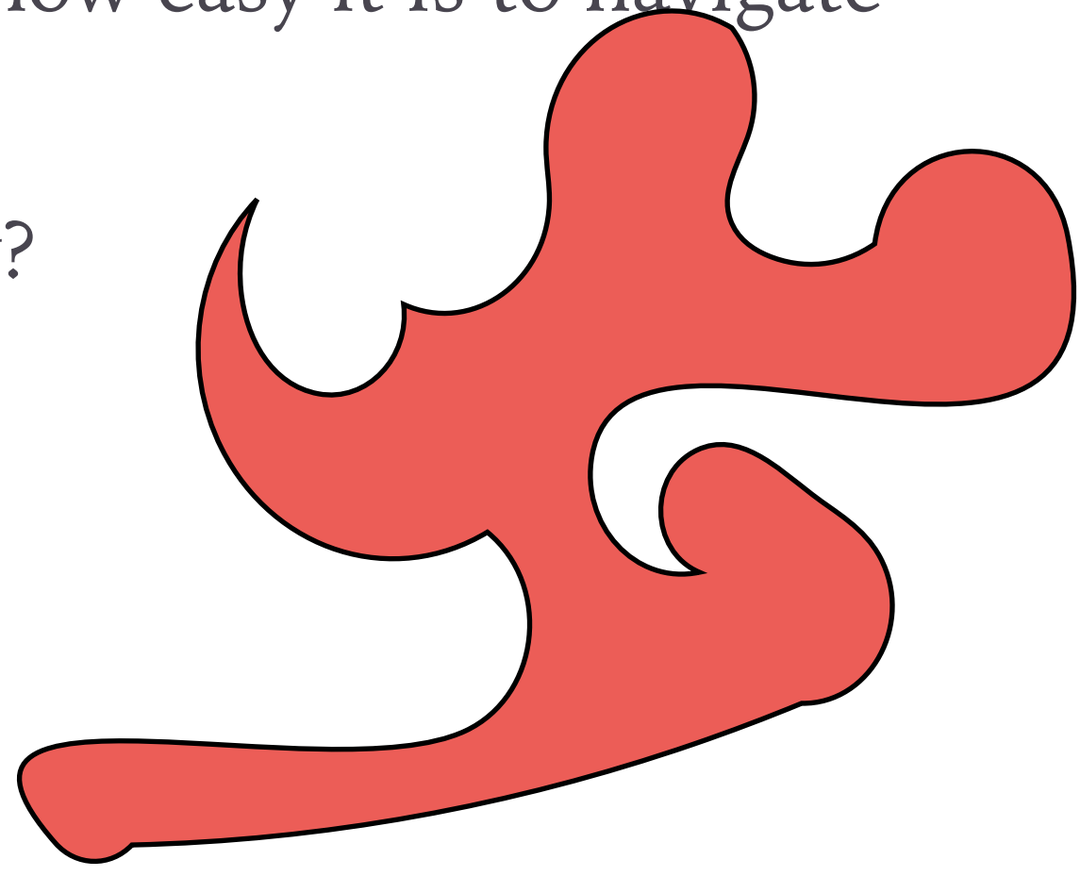
Conjecture [LBB'18]: The loss $L(U,W) = \mathbb{E}\|U\rho(WX)-Y\|^2$ has no poor local minima above the energy barrier $\beta(M,N)$.

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?



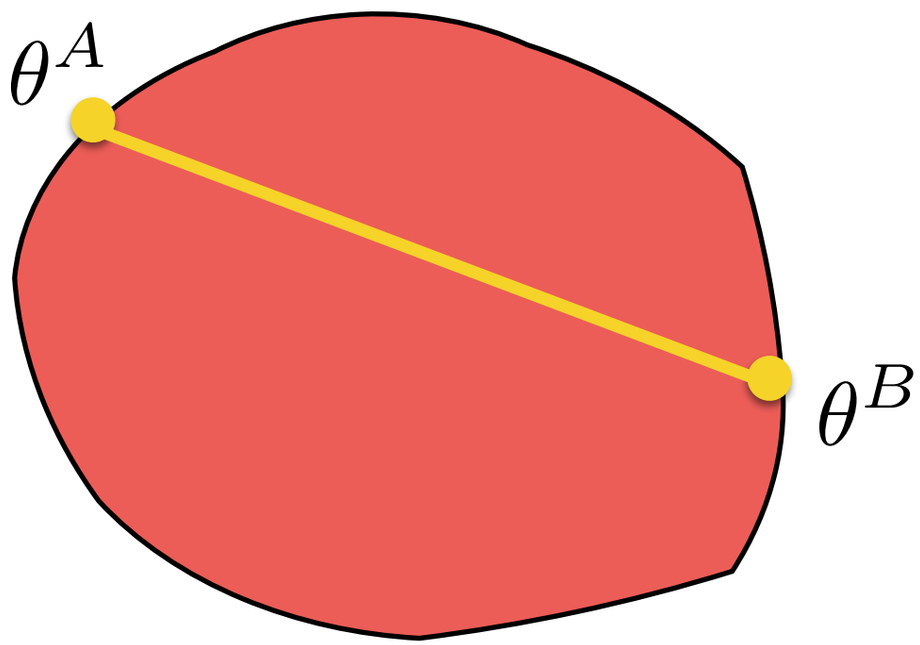
easy to move from one energy level to lower one



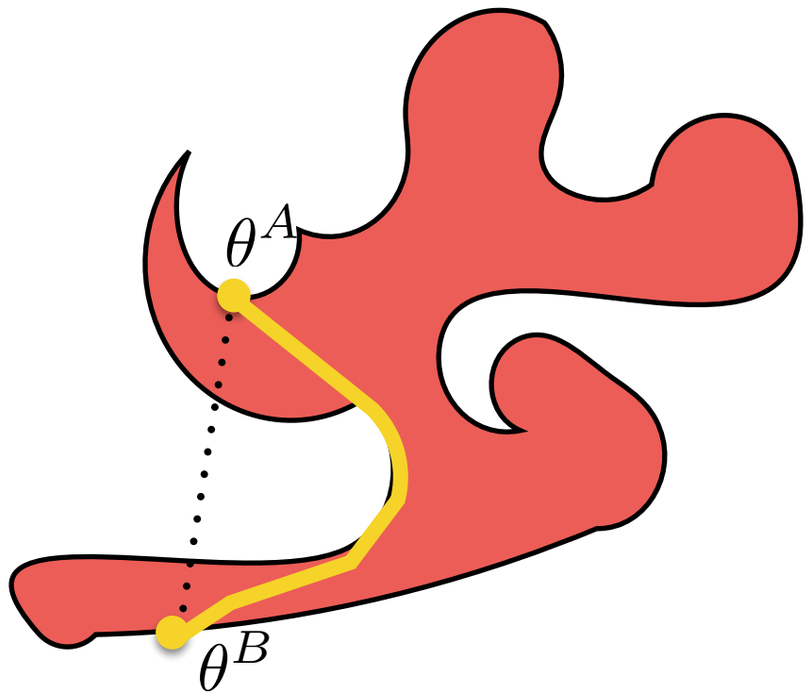
hard to move from one energy level to lower one

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.

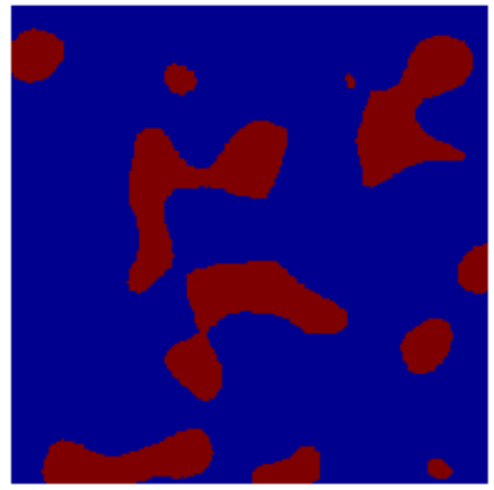


easy to move from one energy level to lower one



hard to move from one energy level to lower one

FINDING CONNECTED COMPONENTS



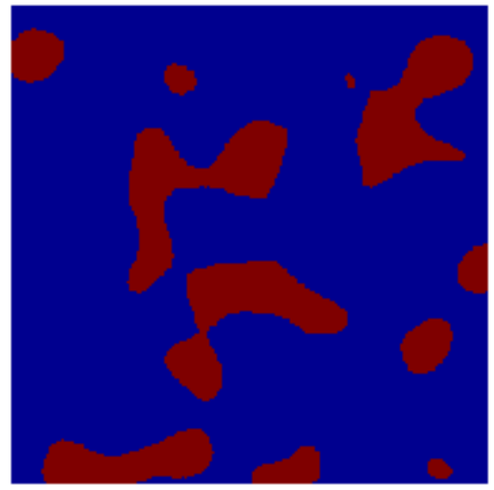
- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff

    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:

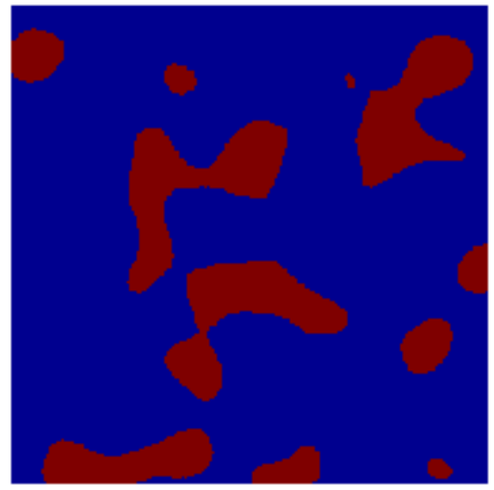
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- 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

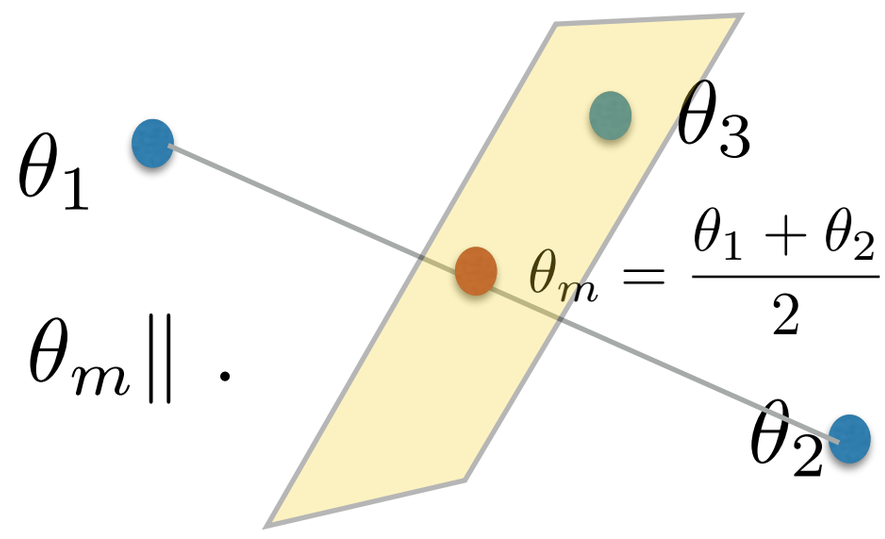
    there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$

  - Moreover, we penalize the length of the path:

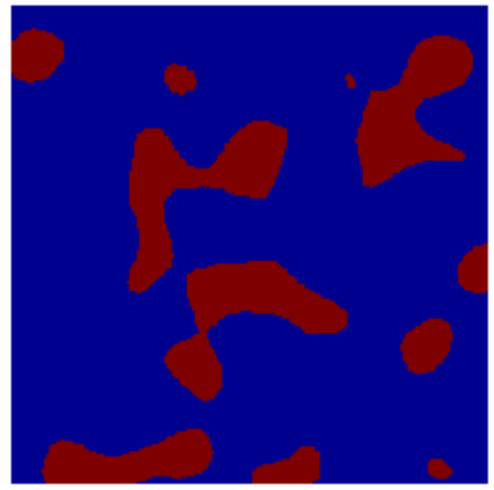
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int_{\mathcal{H}} \|\dot{\gamma}(t)\| dt \leq M\ .$$

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

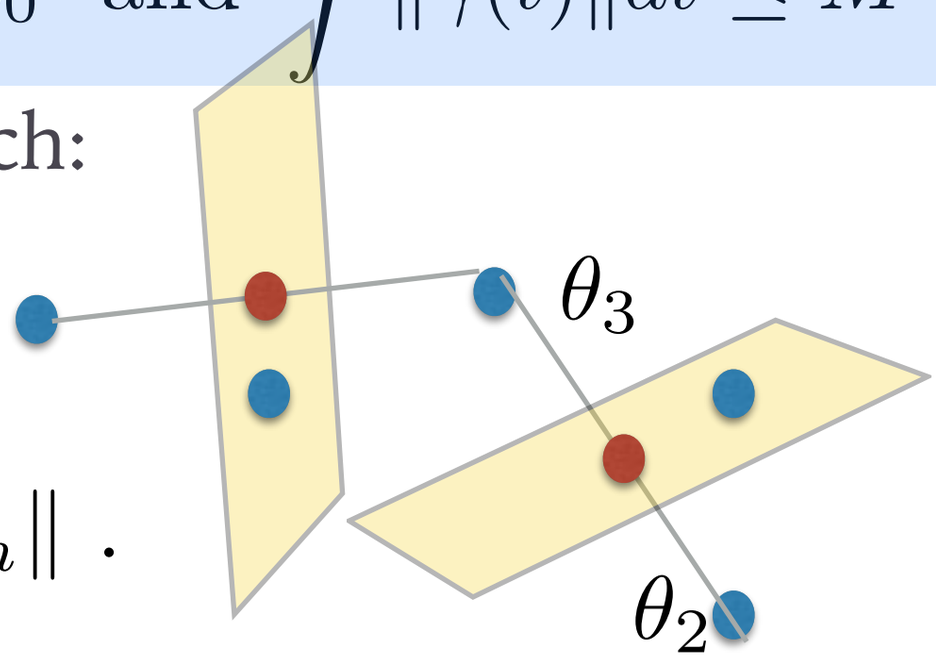
- Moreover, we penalize the length of the path:

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

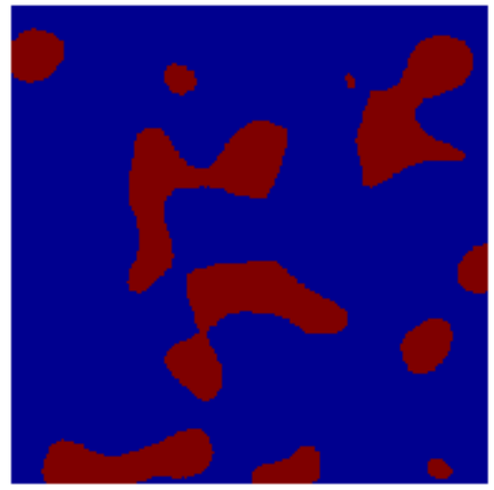
- Dynamic programming approach:

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

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



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

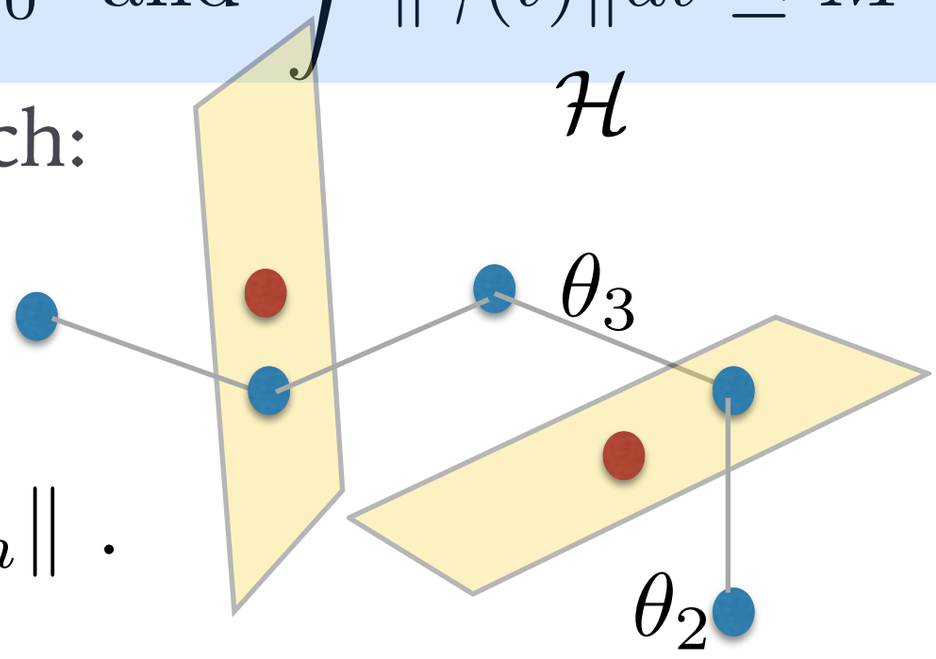
- Moreover, we penalize the length of the path:

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

- Dynamic programming approach:

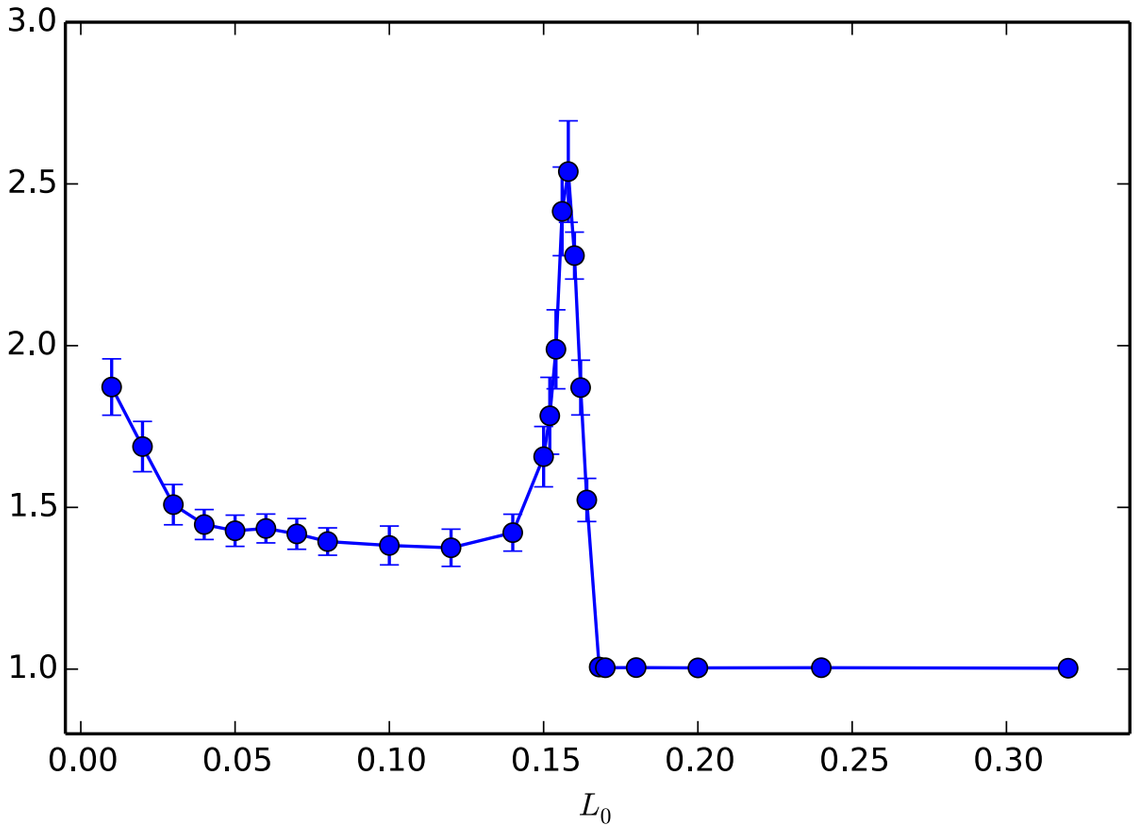
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

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

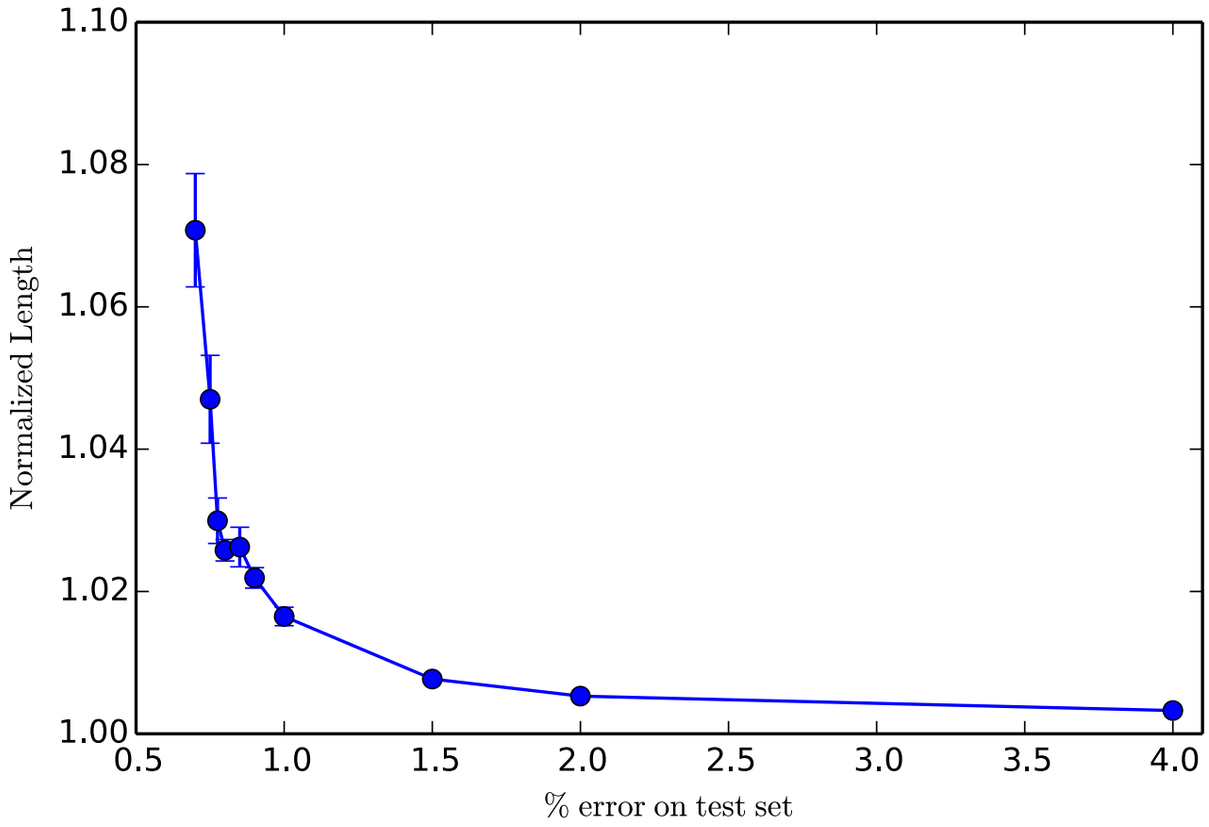


NUMERICAL EXPERIMENTS

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



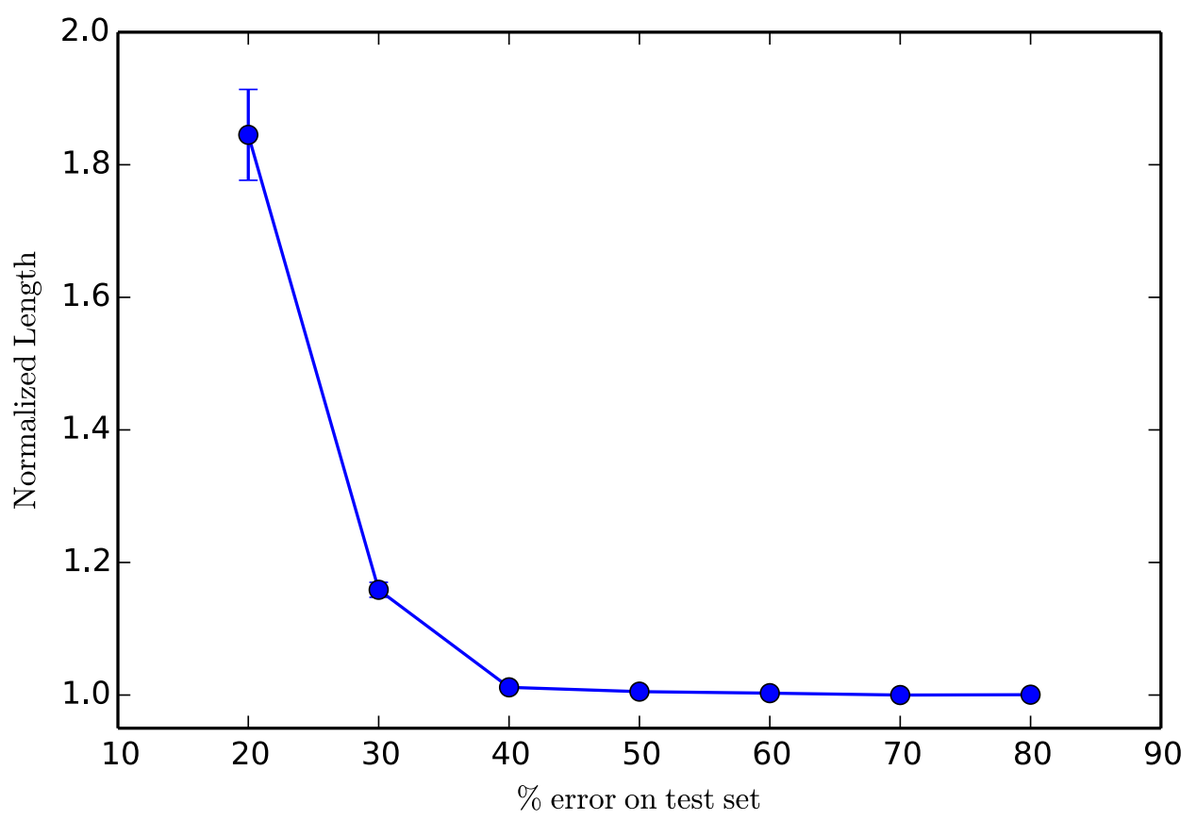
cubic polynomial



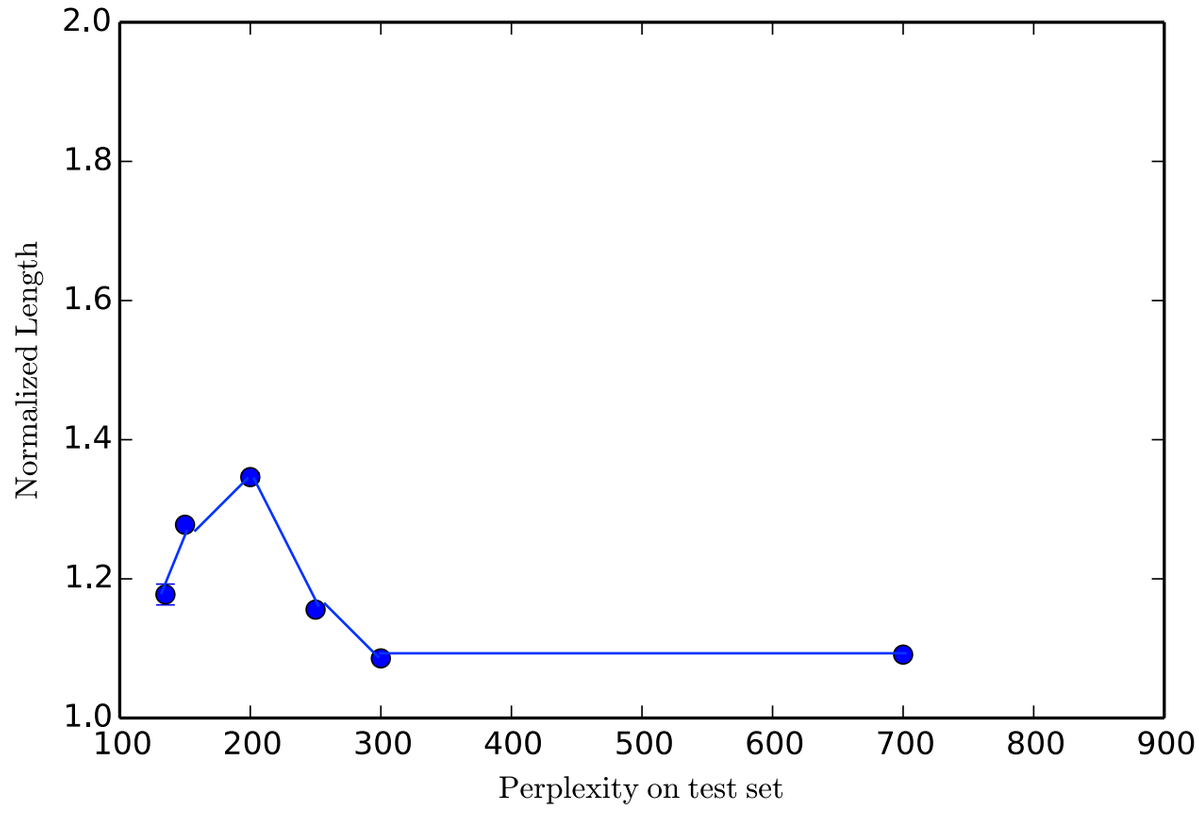
CNN/MNIST

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

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?