Topology and Geometry of Half-Rectified Network Optimization

Daniel Freeman[1] and Joan Bruna[2]

[1]UC Berkeley
[2]Courant Institute and Center for Data Science, NYU

Math and Data

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Stats 385
Stanford
Nov 15th

Motivation

- We consider the standard ML setup:

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

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

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

$\ell(z)$ convex

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

Motivation

- We consider the standard ML setup:

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

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

$\ell(z)$ convex

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

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

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

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

Motivation

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

Motivation

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

- We first address how overparametrization affects the energy landscapes $E(\Theta), \hat{E}(\Theta)$.
- **Goal 1**: Study simple *topological* properties of these landscapes for half-rectified neural networks.
- **Goal 2**: Estimate simple *geometric* properties with efficient, scalable algorithms. Diagnostic tool.

Outline of the Lecture

- Topology of Deep Network Energy Landscapes
- Geometry of Deep Network Energy Landscapes
- Energy Landscapes, Statistical Inference and Phase Transitions.

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]

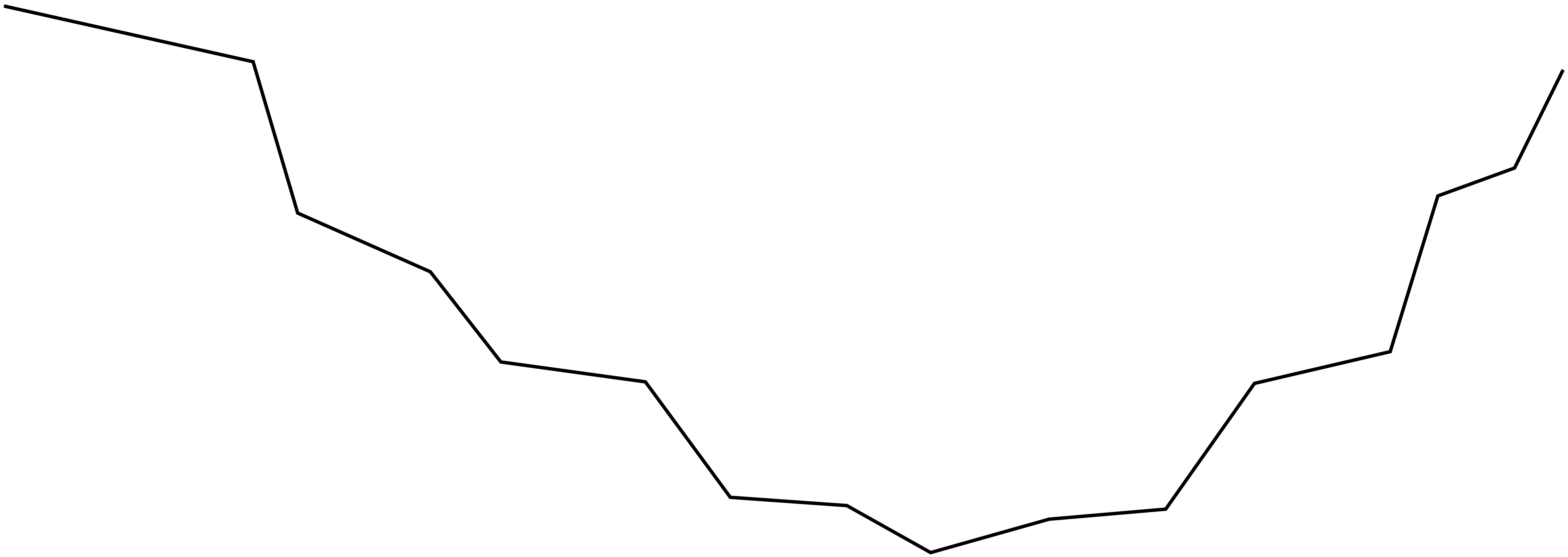
Prior Related Work

- 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

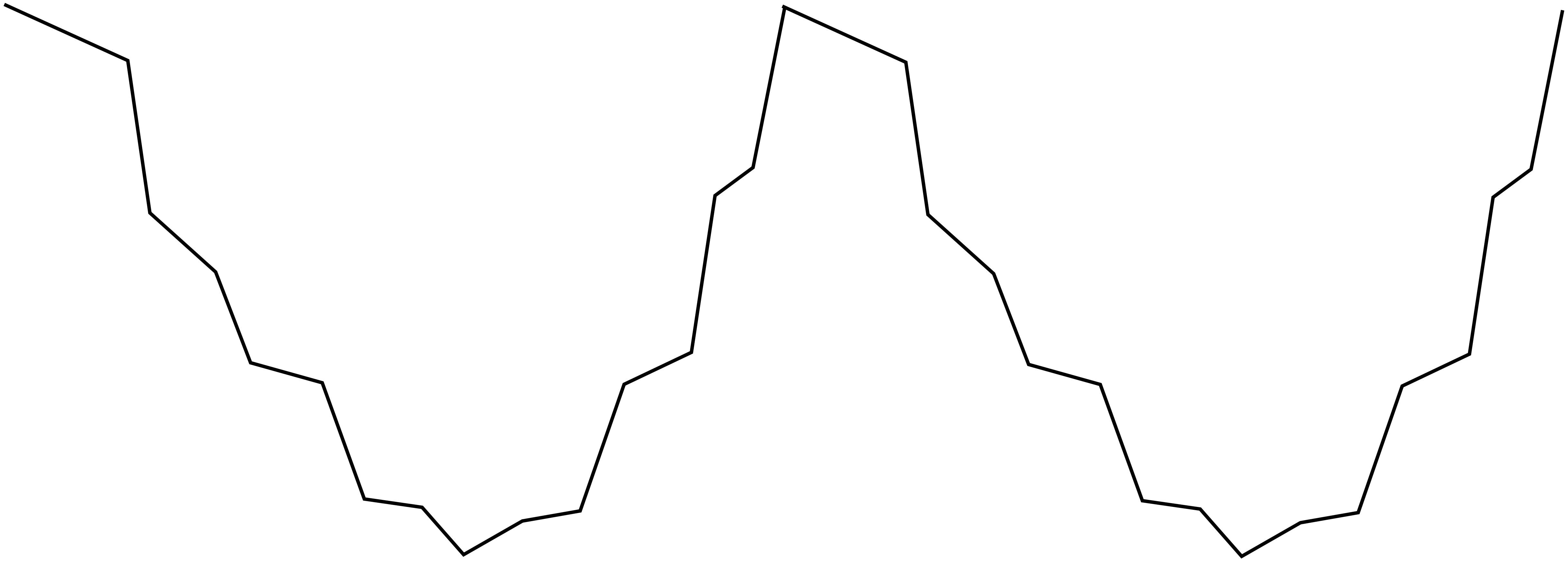
- Models from Statistical physics have been considered as possible approximations [Dauphin et al.'14, Choromanska et al.'15, Segun et al.'15]
- Tensor factorization models capture some of the non convexity essence [Anandukar et al'15, Cohen et al. '15, Haeffele et al.'15]
- [Shafran and Shamir,'15] studies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

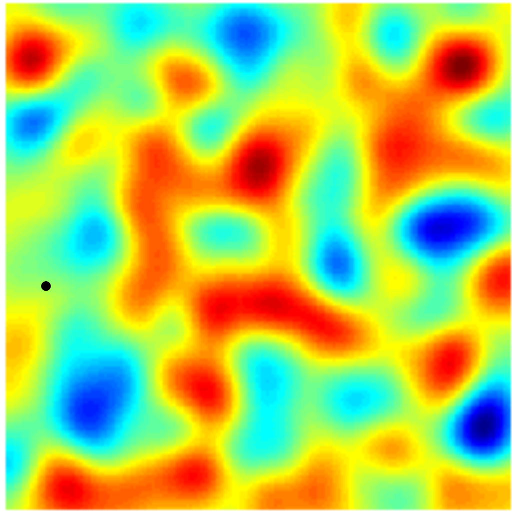


$$F(\theta) = F(g.\theta)\ ,\ g \in G \text{ compact.}$$

- 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.

Analysis of Non-convex Loss Surfaces

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

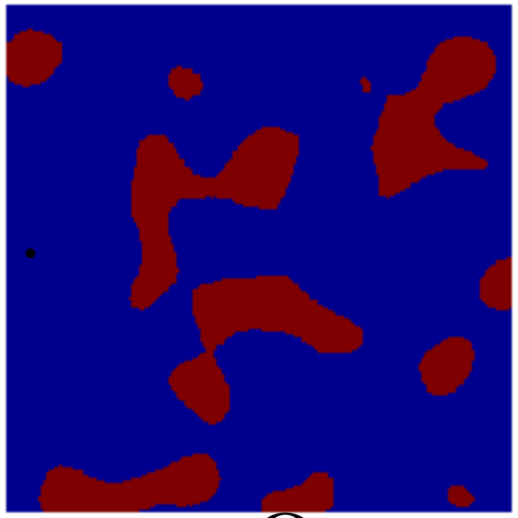


$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\} \ .$$

Analysis of Non-convex Loss Surfaces

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

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



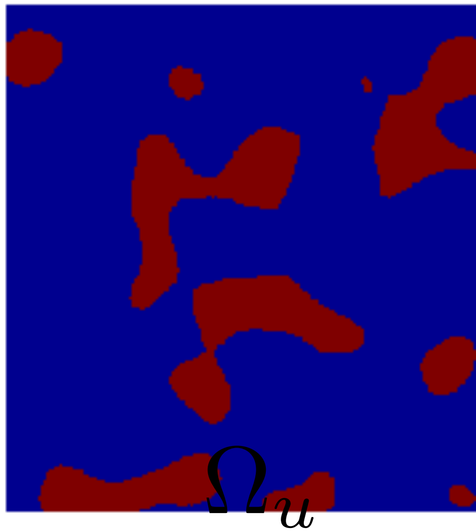
Ω_u

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

- 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

- A first notion we address is about the topology of the level sets .
  - In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?
- This is directly related to the question of global minima:

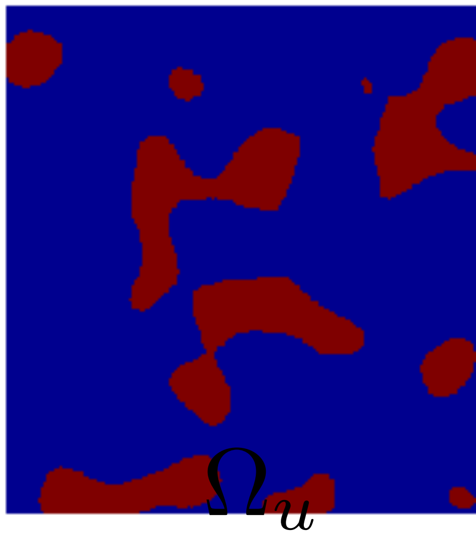


Proposition: If $N_u = 1$ for all u then E has no poor local minima.

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

Analysis of Non-convex Loss Surfaces

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

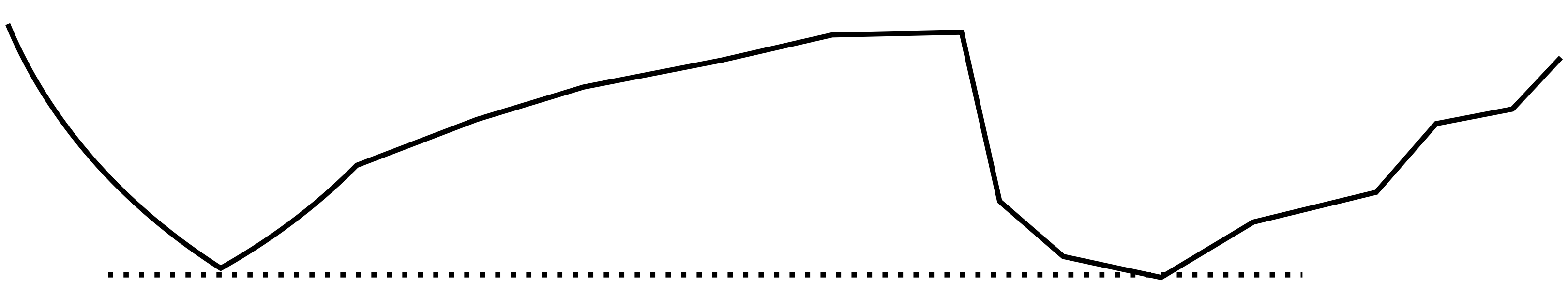


- This is directly related to the question of global minima:

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

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

  - We say E is *simple* in that case.
- The converse is clearly not true.



Linear vs Non-linear deep models

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

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

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

Linear vs Non-linear deep models

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

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

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

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

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

Linear vs Non-linear deep models

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

Proposition: [BF'16]

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

- We pay extra redundancy price to get simple topology.

Linear vs Non-linear deep models

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

Proposition: [BF'16]

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

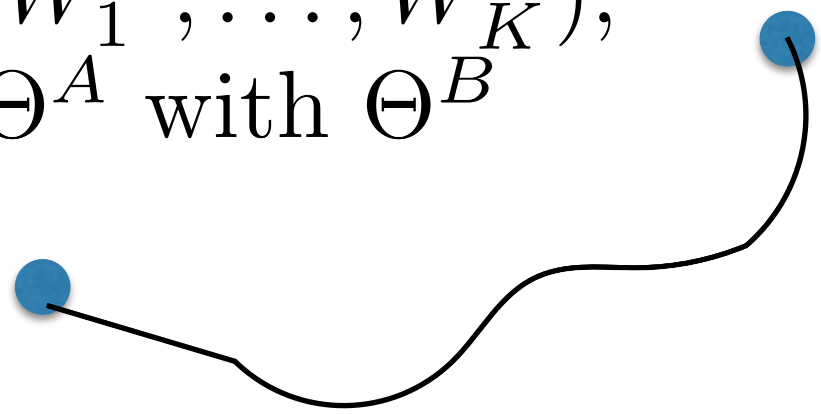
- We pay extra redundancy price to get simple topology.
- This simple topology is an "artifact" of the linearity of the network:

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

Proof Sketch

- Goal:

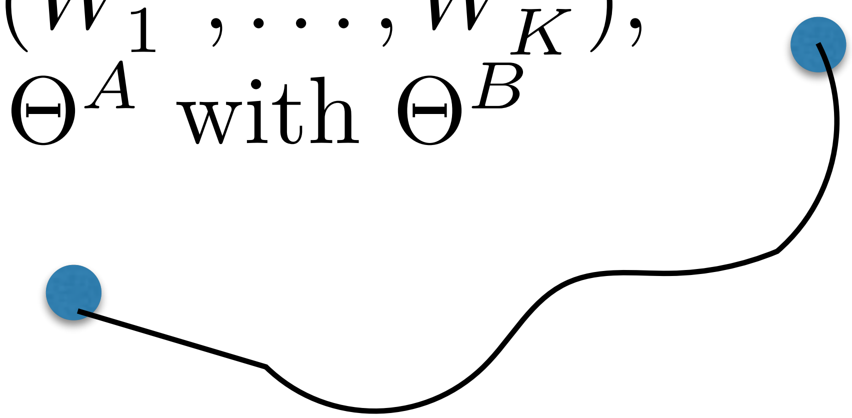
Given $\Theta^A = (W_1^A, \ldots, W_K^A)$ and $\Theta^B = (W_1^B, \ldots, W_K^B)$, we construct a path $\gamma(t)$ that connects Θ^A with Θ^B st $E(\gamma(t)) \leq \max(E(\Theta^A), E(\Theta^B))$.



Proof Sketch

- Goal:

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



- Main idea:

1. Induction on K.

2. *Lift* the parameter space to $\widetilde{W} = W_1 W_2$: the problem is convex $\Rightarrow$ there exists a (linear) path $\widetilde{\gamma}(t)$ that connects Θ^A and Θ^B.

3. Write the path in terms of original coordinates by *factorizing* $\widetilde{\gamma}(t)$.

- Simple fact:

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

Group Symmetries

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

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

Group Symmetries

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

- Q: 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 $n_k > \min(n, m)$
    - ❖ We do the same analysis in the quotient space defined by the equivalence relationship $W \sim \tilde{W} \Leftrightarrow W = \tilde{W}U \ , \ U \in GL(\mathbb{R}^n) \ .$

Group Symmetries

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

- Q: 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 $n_k > \min(n, m)$

  ❖ We do the same analysis in the quotient space defined by the equivalence relationship $W \sim \tilde{W} \Leftrightarrow W = \tilde{W}U \ , \ U \in GL(\mathbb{R}^n)$.

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

❖ Construct paths on the Grassmanian manifold of subspaces.

❖ Generalizes best known results for multilinear case (no assumptions on data 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":

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

- We have the following extension:

Proposition: If $M \geq 3N^2$, then the landscape of two-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

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":

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

- We have the following extension:

Proposition: If $M \geq 3N^2$, then the landscape of two-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

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

- *Open question:* Improve rate by exploiting Group symmetries? Currently we only win on the constants.

Asymptotic Connectedness of ReLU

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

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

$$\|w_{1,i}\|_2 \leq 1 \ , \ \ell_1 \text{ Regularization on } W_2 \ .$$

Asymptotic Connectedness of ReLU

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

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

$$\|w_{1,i}\|_2 \leq 1\ ,\ \ell_1 \text{ Regularization on } W_2\ .$$

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:

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

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

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

Asymptotic Connectedness of ReLU

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

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

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

- Overparametrisation "wipes-out" local minima (and group symmetries).
- The bound is cursed by dimensionality, ie exponential in n .
- *Open question:* polynomial rate using Taylor decomp of $\rho(z)$?

Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters $\mathbf{\Theta}$

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

to *canonical* parameters $\beta = \mathcal{A}(\mathbf{\Theta})$:

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

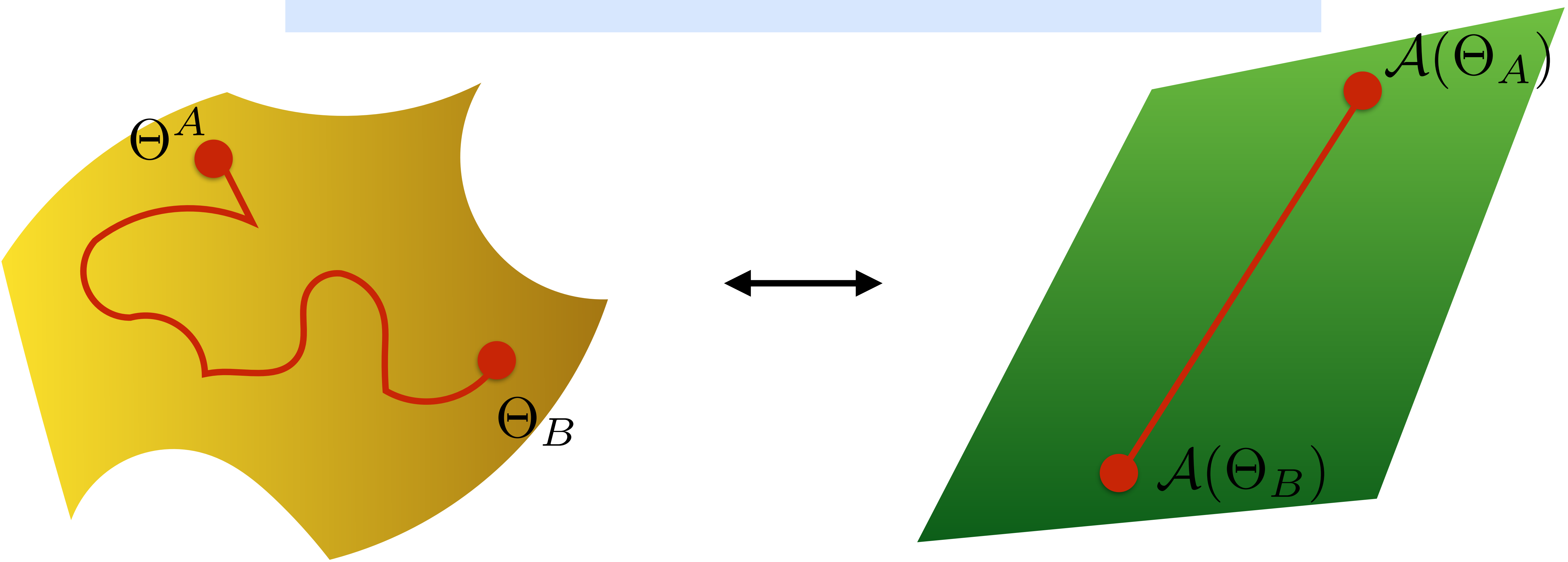
Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ

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

to *canonical* parameters $\beta = \mathcal{A}(\Theta)$

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



Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ

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

to *canonical* parameters $\beta = \mathcal{A}(\Theta)$:

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

- Second layer setup: $\rho(\langle w, X \rangle) = \langle \mathcal{A}(w), \Psi(X) \rangle$.

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

Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ

$$\Phi(x;\Theta) = W_k\rho(W_{k-1}\dots\rho(W_1X)))\ ,\ \Theta = (W_1,\dots W_k)\ ,$$

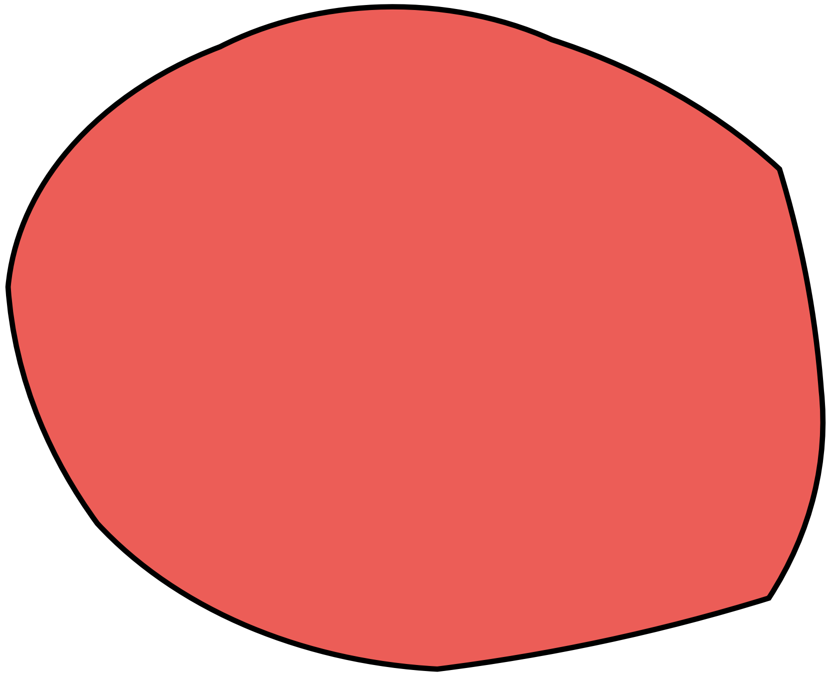
to *canonical* parameters $\beta = \mathcal{A}(\Theta)$

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

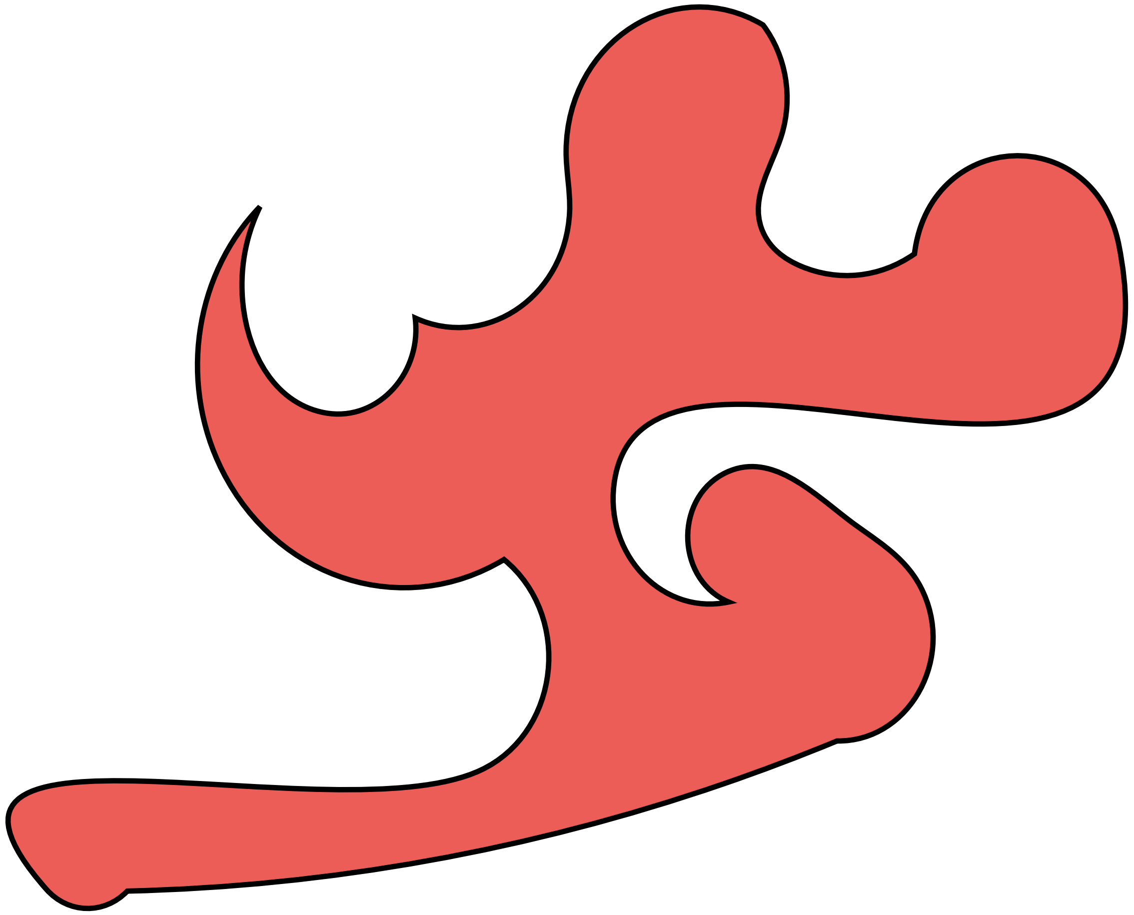
- This is precisely the formulation of ERM in terms of Reproducing Kernel Hilbert Spaces [Scholkopf, Smola, Gretton, Rosasco, …]
- Recent works developed RKHS for Deep Convolutional Networks
  - [Mairal et al.'17, Zhang, Wainwright & Liang '17]
  - See also F. Bach's talk tomorrow [Bach'15].
  - Open question: behavior of SGD in Θ in terms of canonical params? Progress on matrix factorization, e.g [Srebo'17]

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?



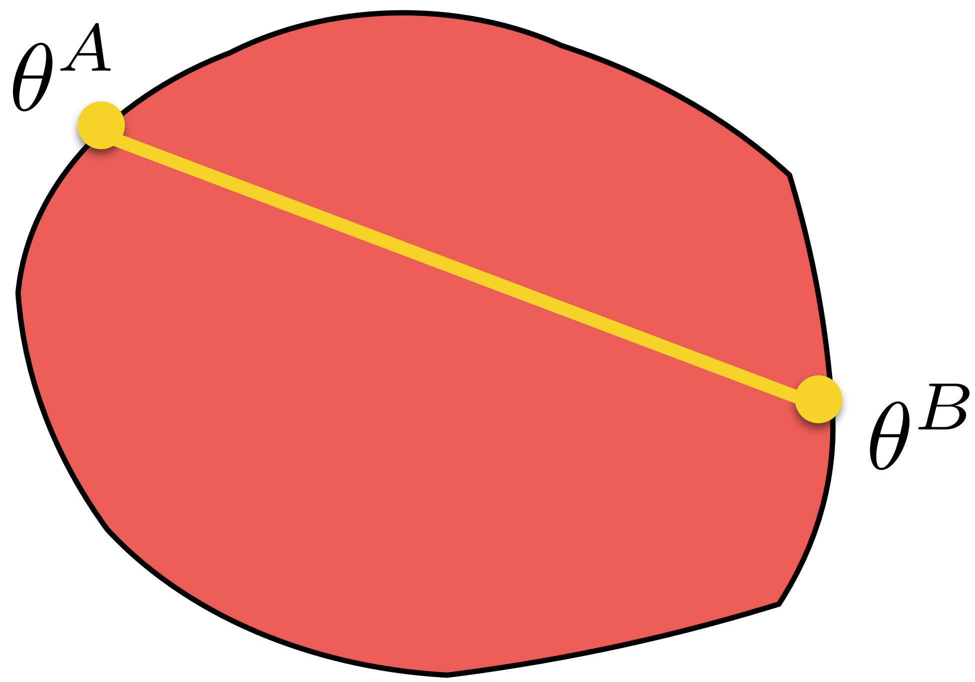
easy to move from one energy level to lower one



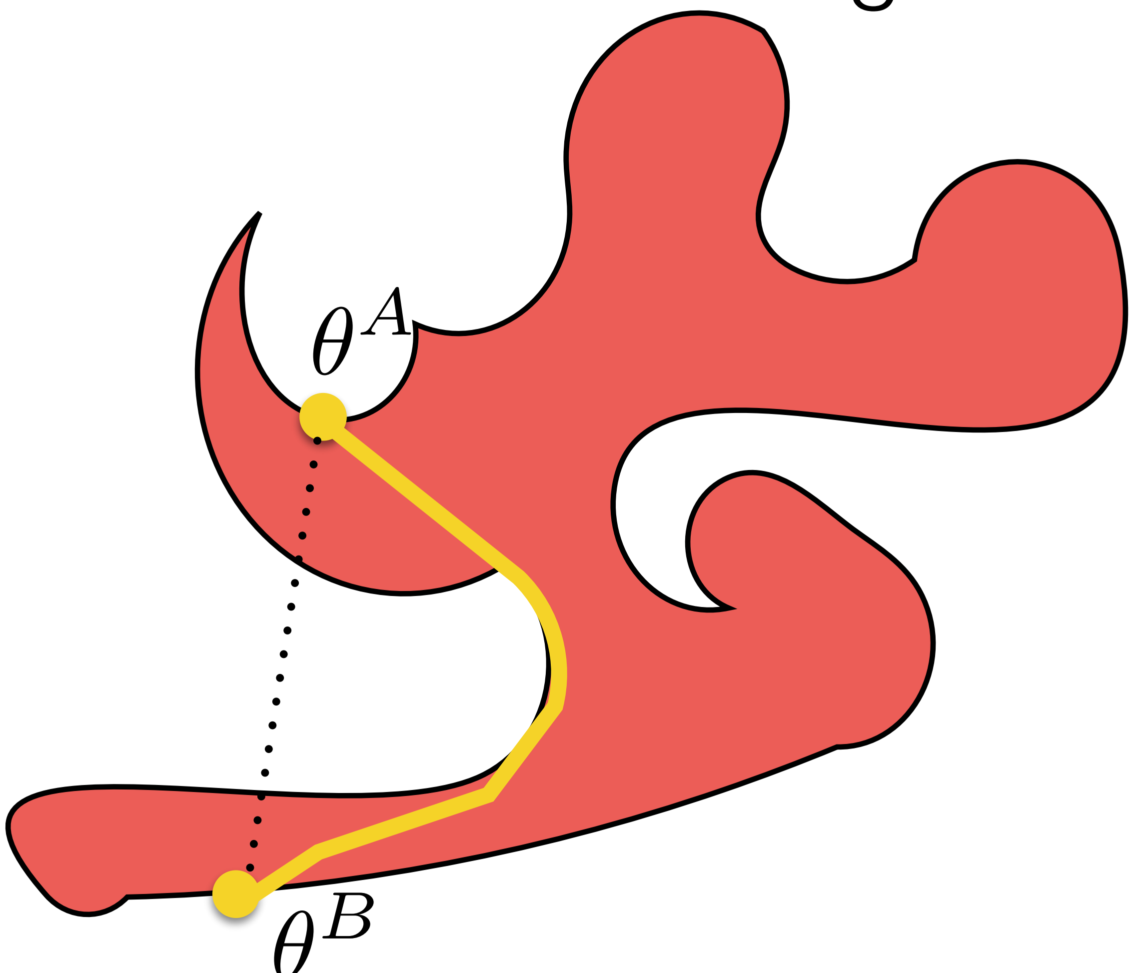
hard to move from one energy level to lower one

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.



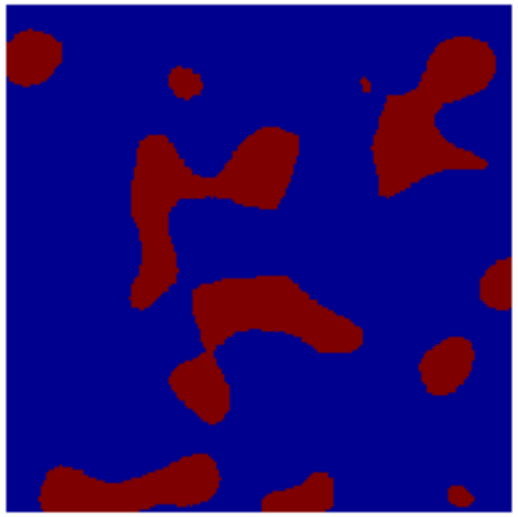
easy to move from one energy level to lower one



hard to move from one energy level to lower one

Finding Connected Components

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



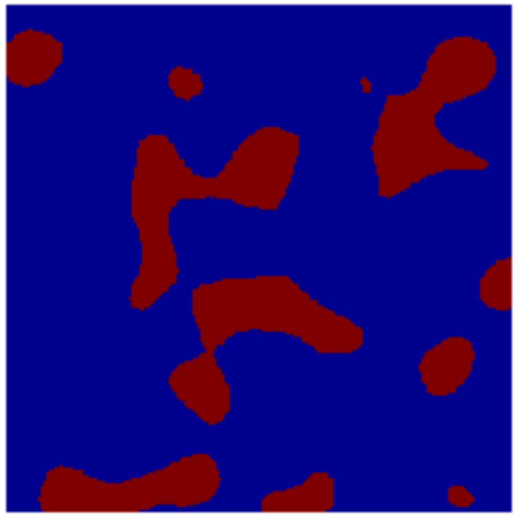
Ω_u

  - Moreover, we penalize the length of the path:

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

Finding Connected Components

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



Ω_u

  - Moreover, we penalize the length of the path:

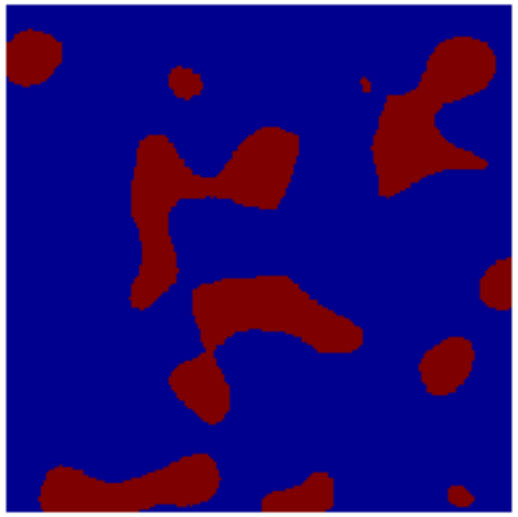
$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \ \text{ and } \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

Finding Connected Components

- Suppose θ_1, θ_2 are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that

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



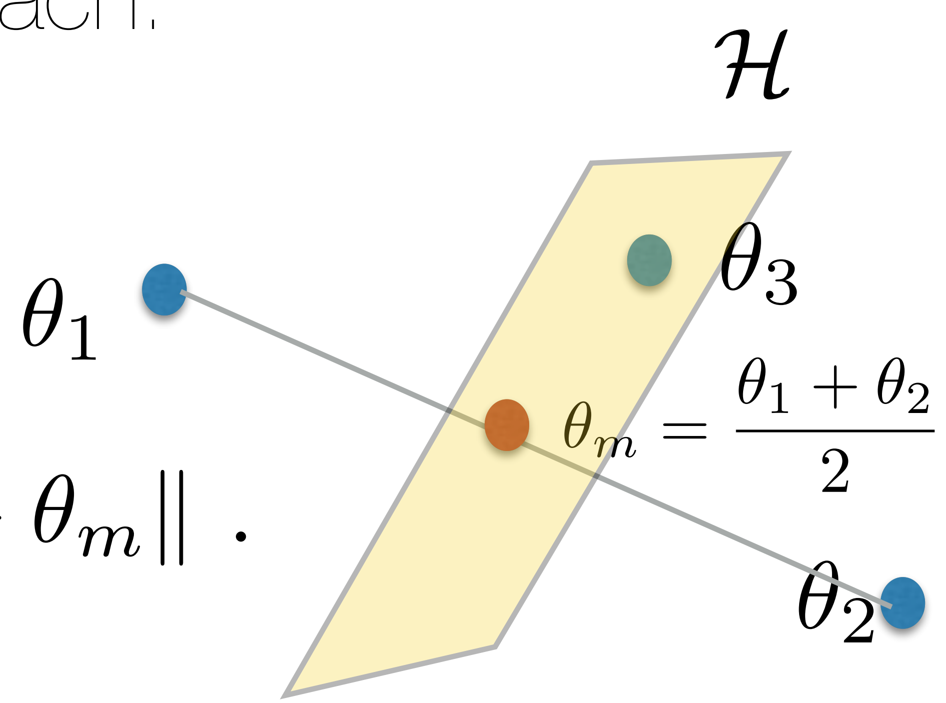
Ω_u

  - Moreover, we penalize the length of the path:

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

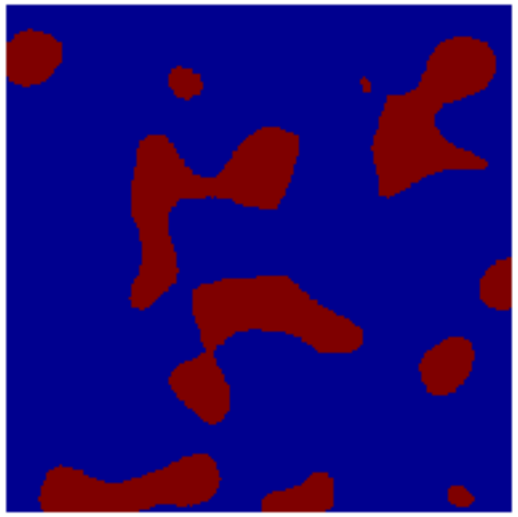
- Dynamic programming approach:

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



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



Ω_u

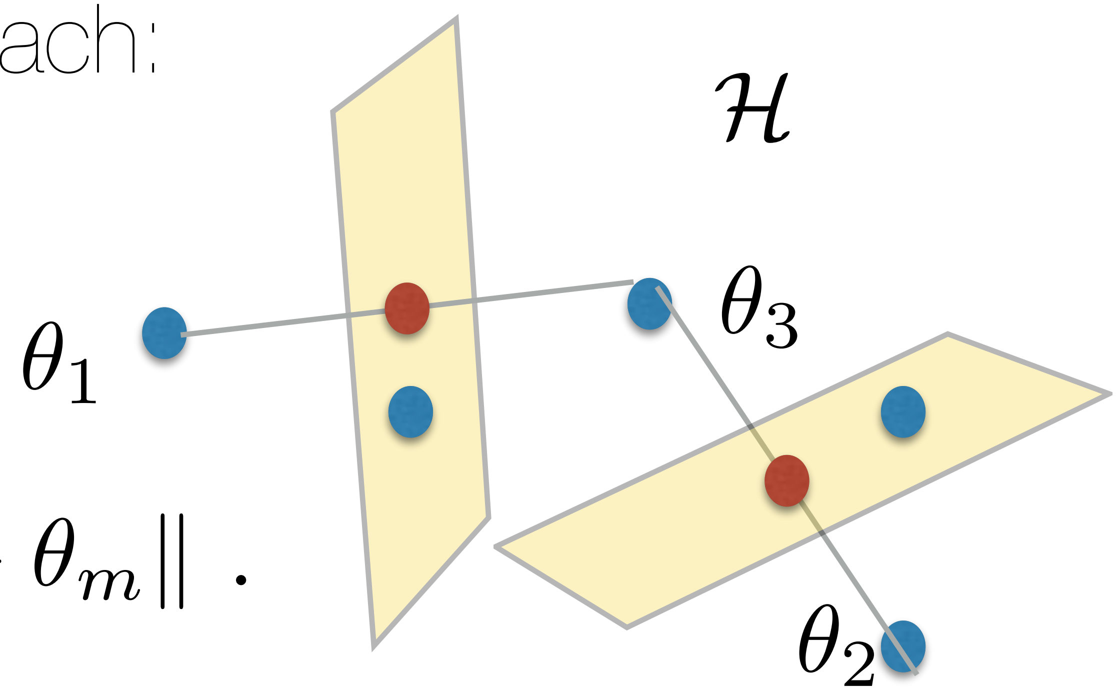
  - Moreover, we penalize the length of the path:

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

- Dynamic programming approach:

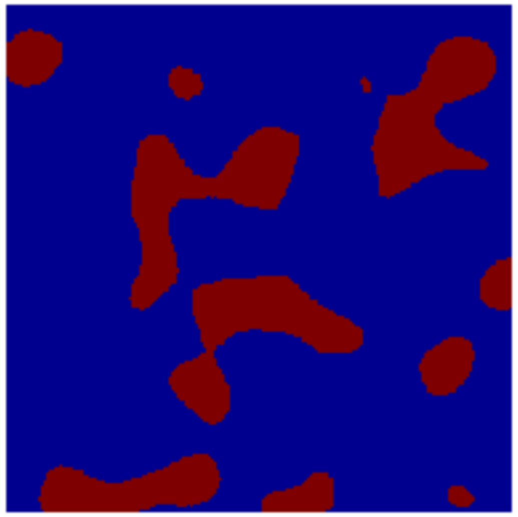
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

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



Finding Connected Components

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



Ω_u

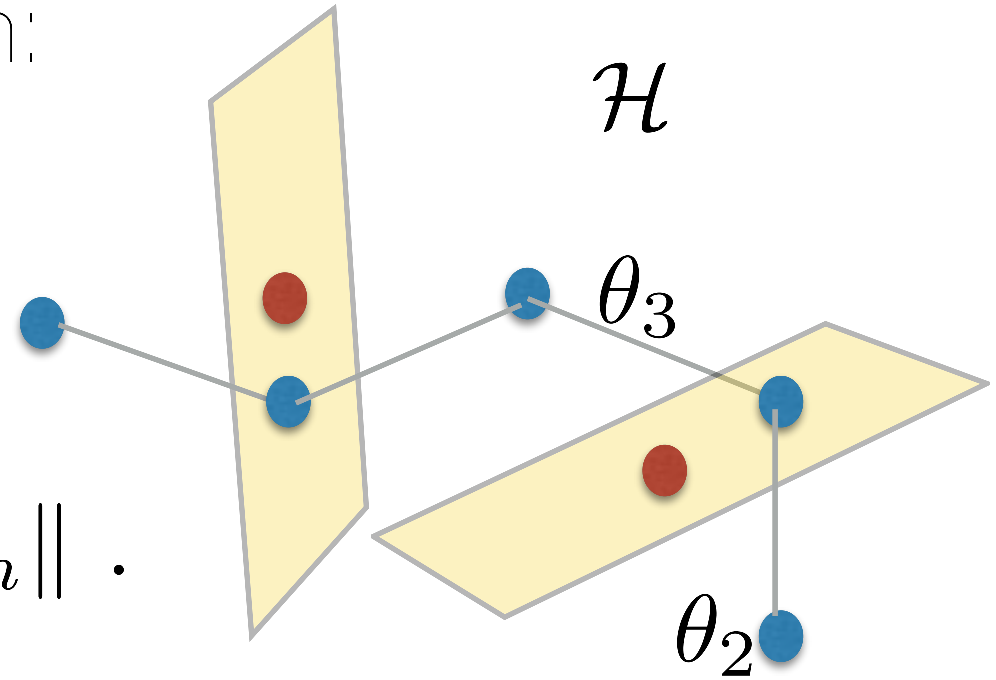
  - Moreover, we penalize the length of the path:

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

- Dynamic programming approach:

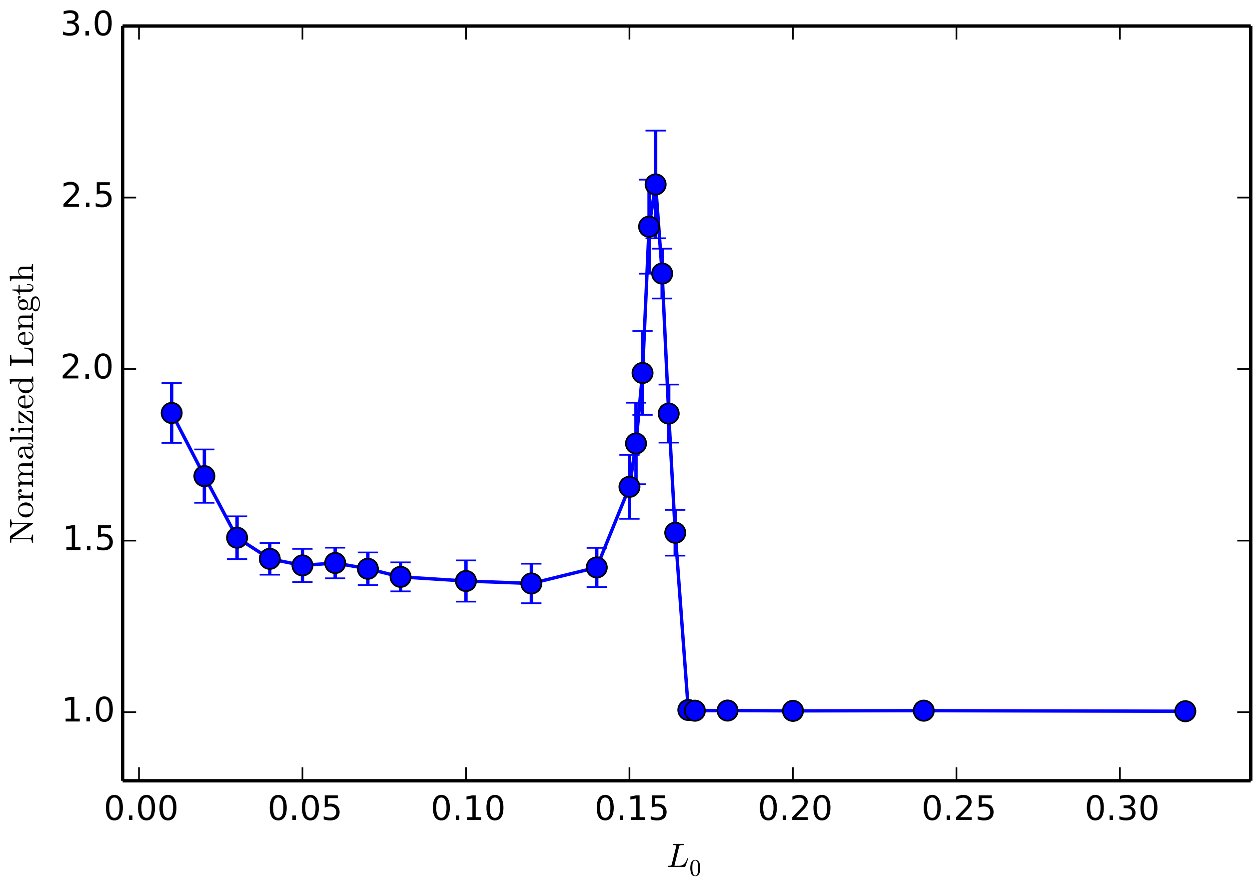
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

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

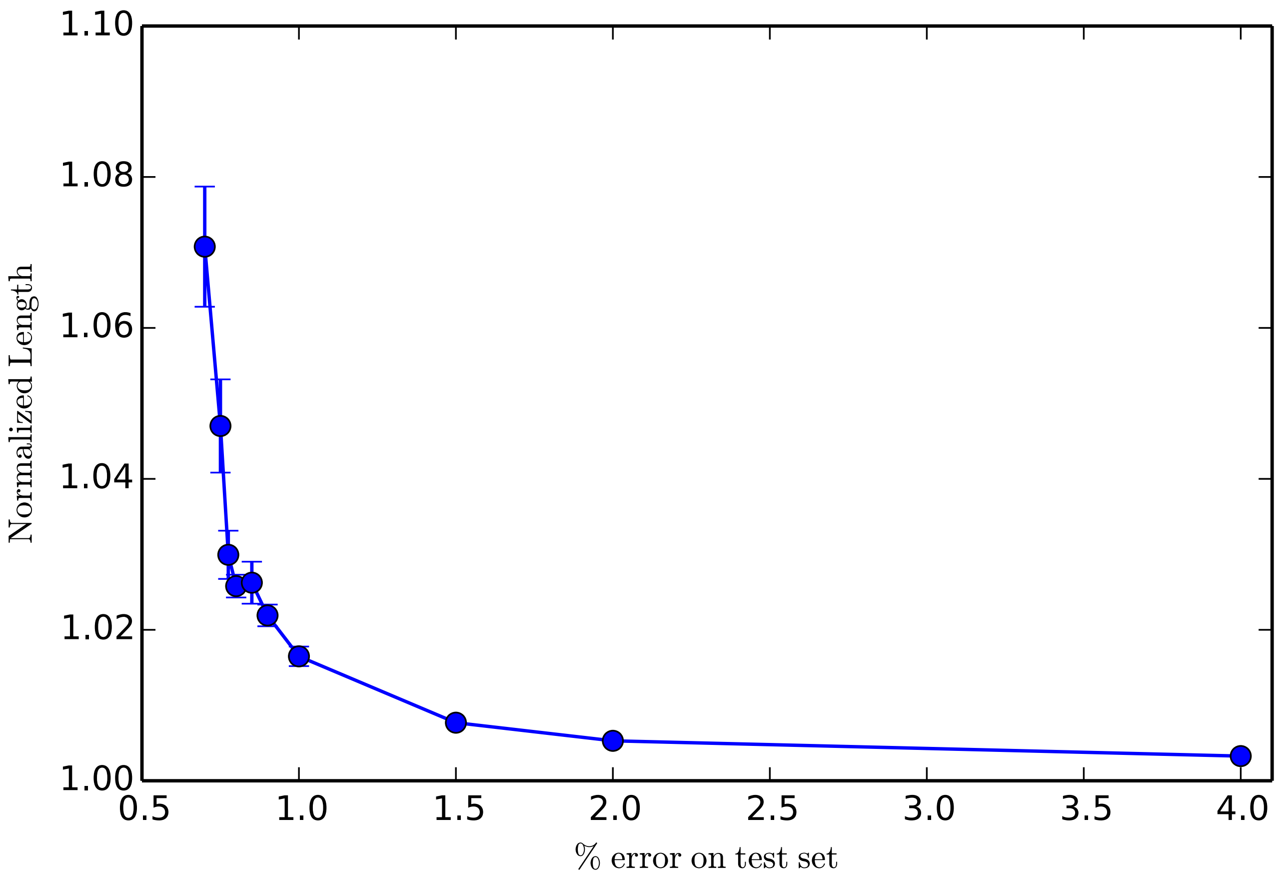


Numerical Experiments

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



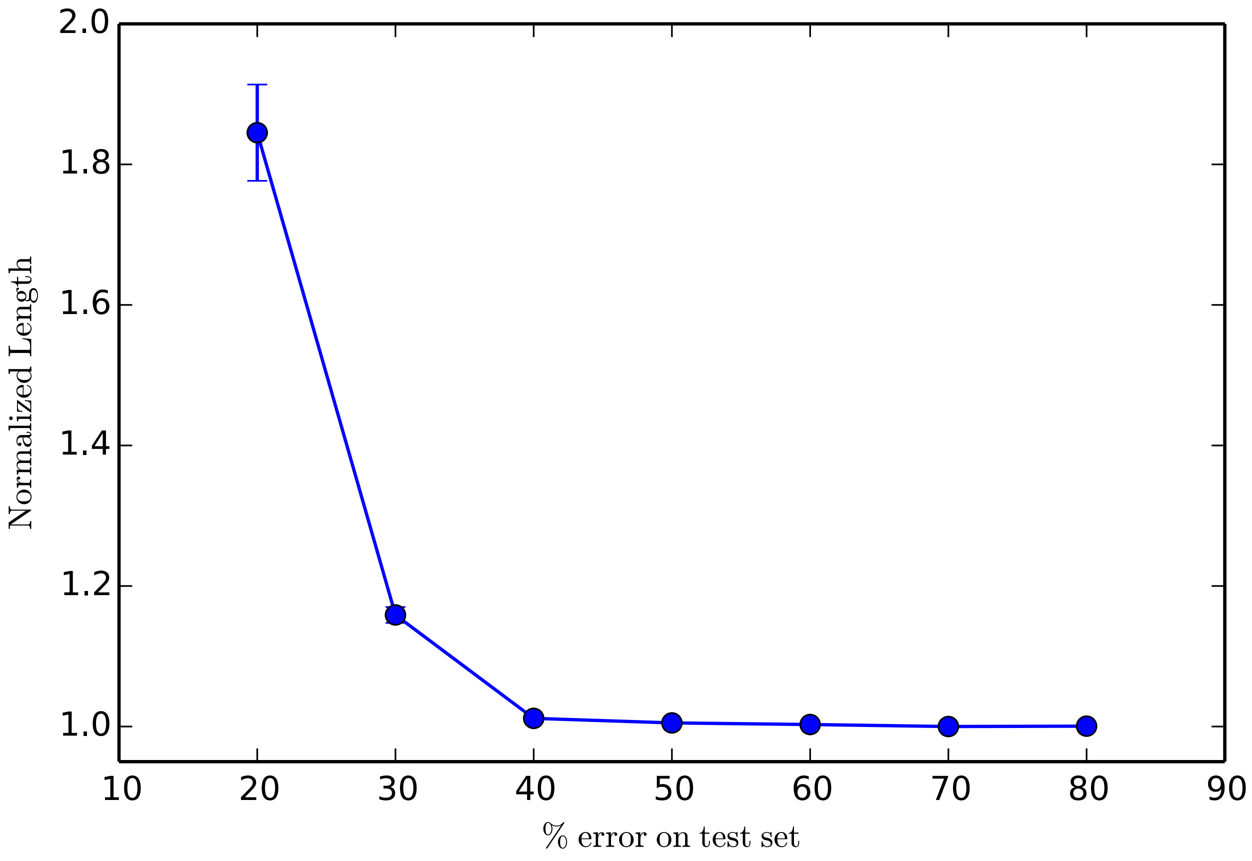
cubic polynomial



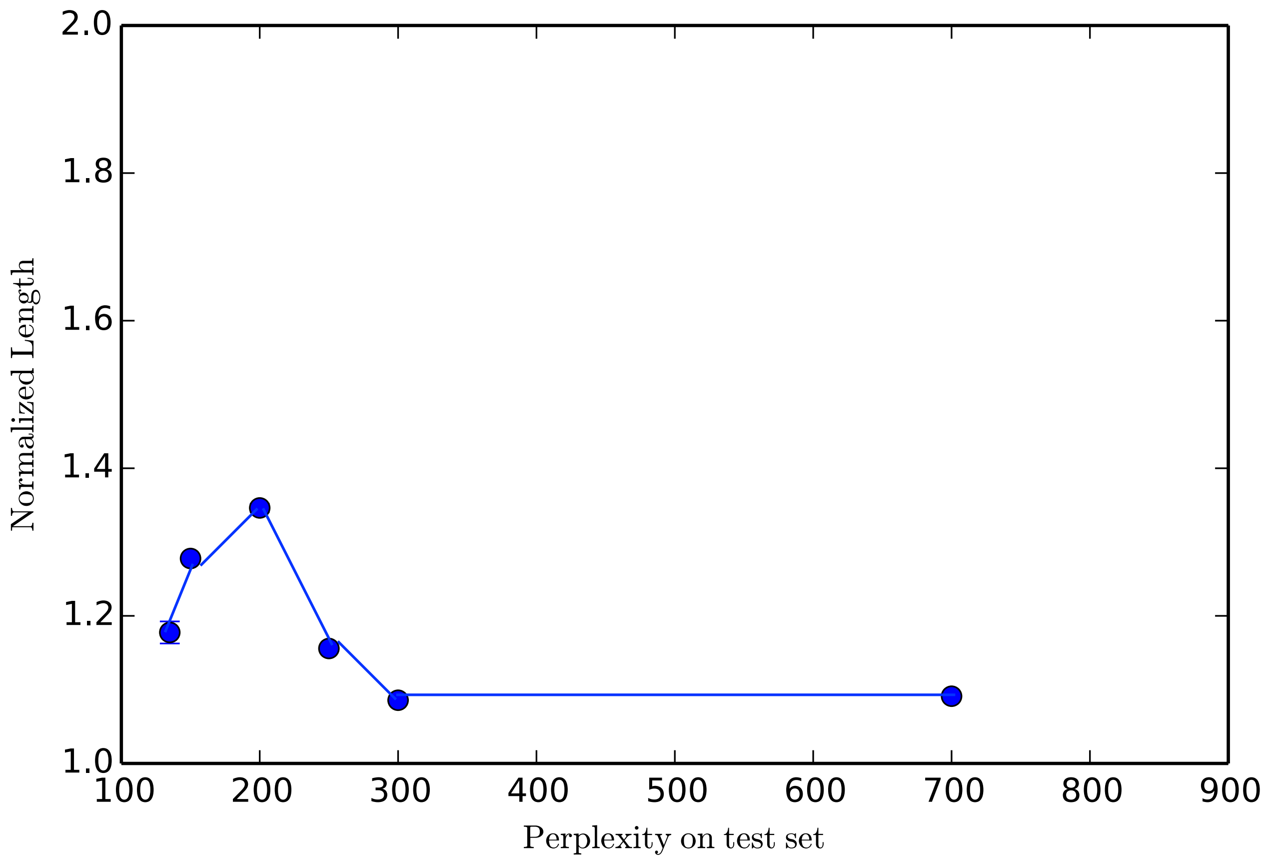
CNN/MNIST

Numerical Experiments

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



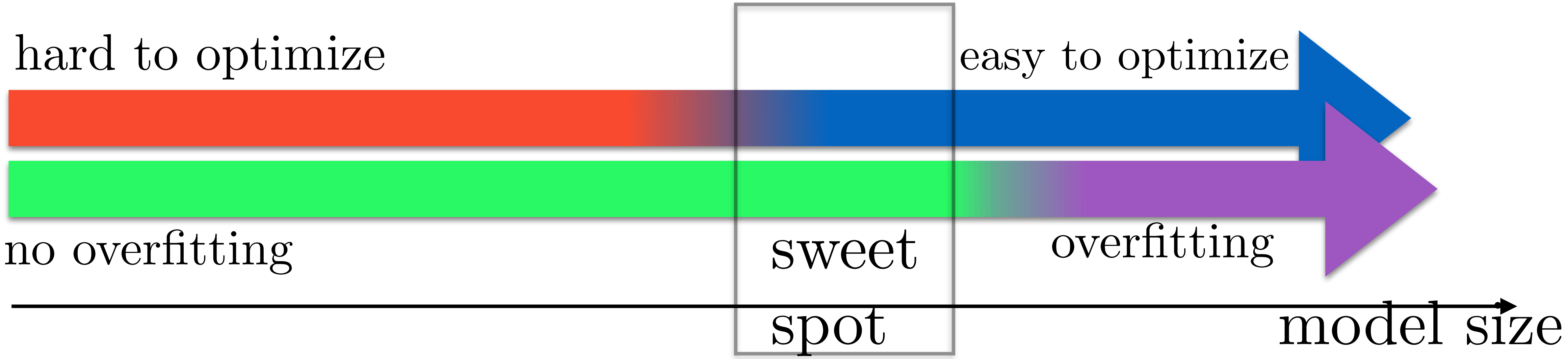
CNN/CIFAR-10



LSTM/Penn

Analysis and perspectives

- #of components does not increase: no detected poor local minima so far when using typical datasets and typical architectures (at energy levels explored by SGD).
- Level sets become more irregular as energy decreases.
- Presence of "energy barrier"?
- Kernels are back? CNN RKHS
- Open: "sweet spot" between overparametrisation and overfitting?
- Open: Role of Stochastic Optimization in this story?



Energy Landscapes, Statistical Inference, and Phase Transitions

Some Open/Current Directions

- The previous setup considered arbitrary classification/regression tasks, e.g object classification.
- We introduced a notion of *learnable hardness*, in terms of the topology and geometry of the Empirical/Population Risk Minimization.

Some Open/Current Directions

- The previous setup considered arbitrary classification/regression tasks, e.g object classification.
- We introduced a notion of *learnable hardness*, in terms of the topology and geometry of the Empirical/Population Risk Minimization.

- Q: How does this notion of hardness connect with other forms of hardness? e.g.
  - Statistical Hardness.
  - Computational Hardness.
- This suggests using Neural Networks on "classic" Statistical Inference.
  - Other motivations: faster inference? data adaptive?

Sparse Coding

- Consider the following inference problem.

Given $D \in \mathbb{R}^{n \times m}$ and $x \in \mathbb{R}^n$,

$$\min_z E(z) = \frac{1}{2}\|x - Dz\|^2 + \lambda \|z\|_1 \ .$$

- Long history in Statistics and Signal Processing:
  - Lasso estimator for variable selection [Tibshirani, '95].
  - Building block in many signal processing and machine learning pipelines [Mairal et al. '10]
- Problem is convex, unique solution for generic D, not strongly convex in general.

Sparse Coding and Iterative Thresholding

- A popular approach to solving SC is via iterative splitting algorithms [Bruck, Passty,70s]:

$$z^{(n)} = \rho_{\gamma\lambda}((\mathbf{1} - \gamma D^T D) z^{(n-1)} + \gamma D^T x) \ , \ \text{with}$$

$$\rho_t(x) = \text{sign}(x) \cdot \max(0, |x| - t)$$

- When $\gamma \leq \frac{1}{\|D\|^2}$, $z^{(n)}$ converges to a solution, in the sense that

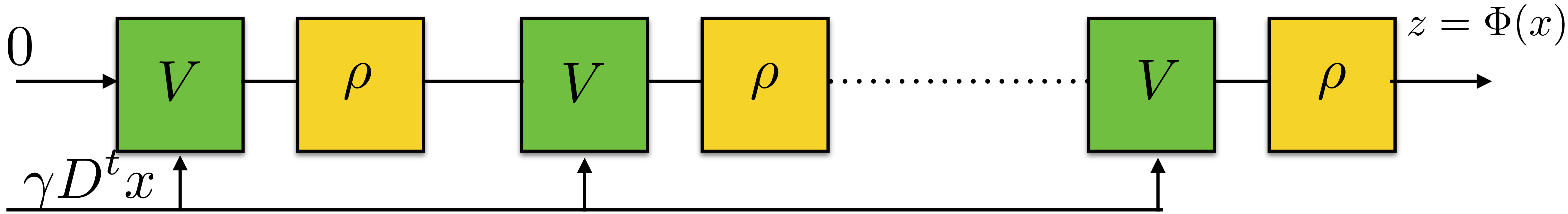
$$E(z^{(n)}) - E(z^*) \leq \frac{\gamma^{-1} \|z^{(0)} - z^*\|^2}{2n} \ .$$

[Beck, Teboulle,'09]

  – sublinear convergence due to lack of strong convexity.
  – however, linear convergence can be obtained under weaker conditions (e.g. RSC/RSM, [Argawal & Wainwright]).

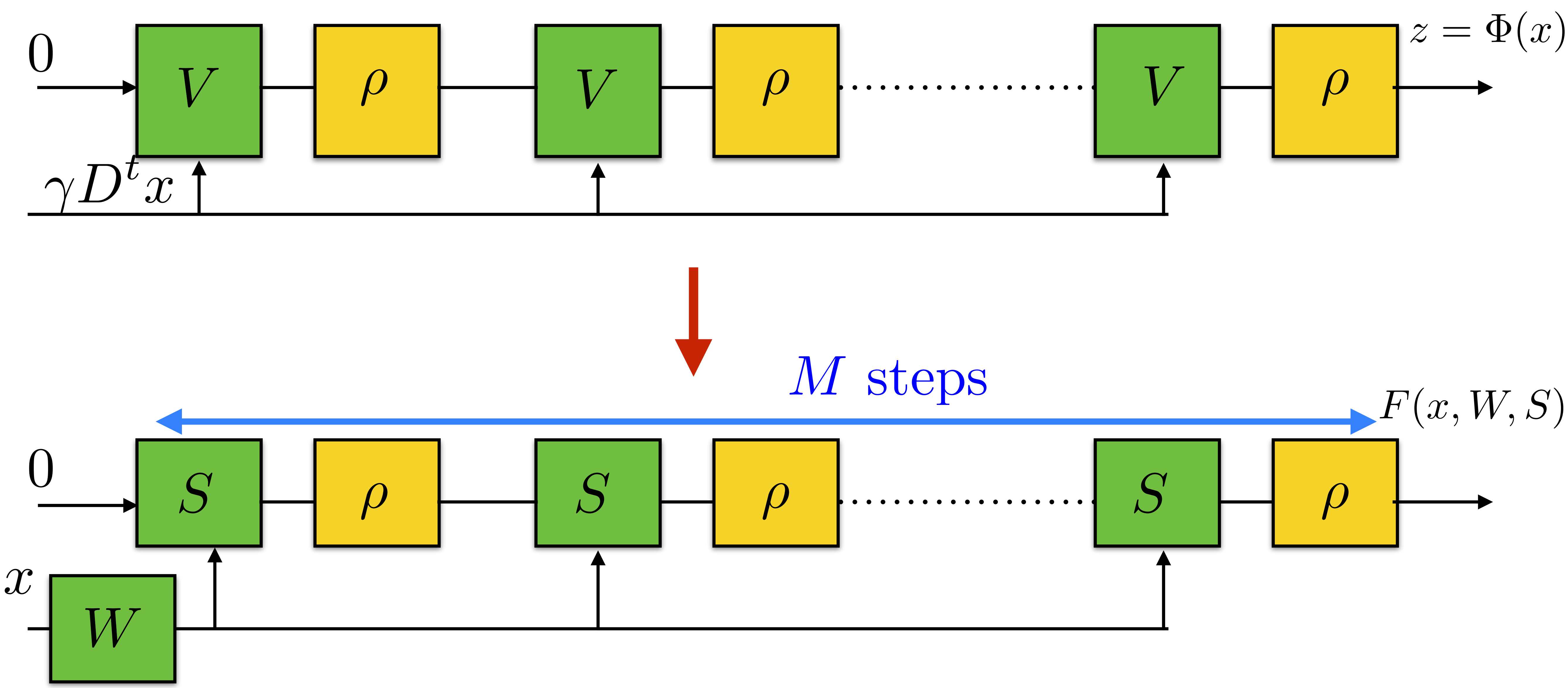
LISTA [Gregor & LeCun'10]

- The Lasso (sparse coding operator) can be implemented as a specific deep network with infinite, recursive layers.
- Can we accelerate the sparse inference with a shallower network, with trained parameters?



LISTA [Gregor & LeCun'10]

- The Lasso (sparse coding operator) can be implemented as a specific deep network with infinite, recursive layers.
- Can we accelerate the sparse inference with a shallower network, with trained parameters? In practice, yes.



Sparsity Stable Matrix Factorizations

[joint work with Th. Moreau (ENS)]

- Principle of proximal splitting: the regularization term $\|z\|_1$ is *separable* in the canonical basis:

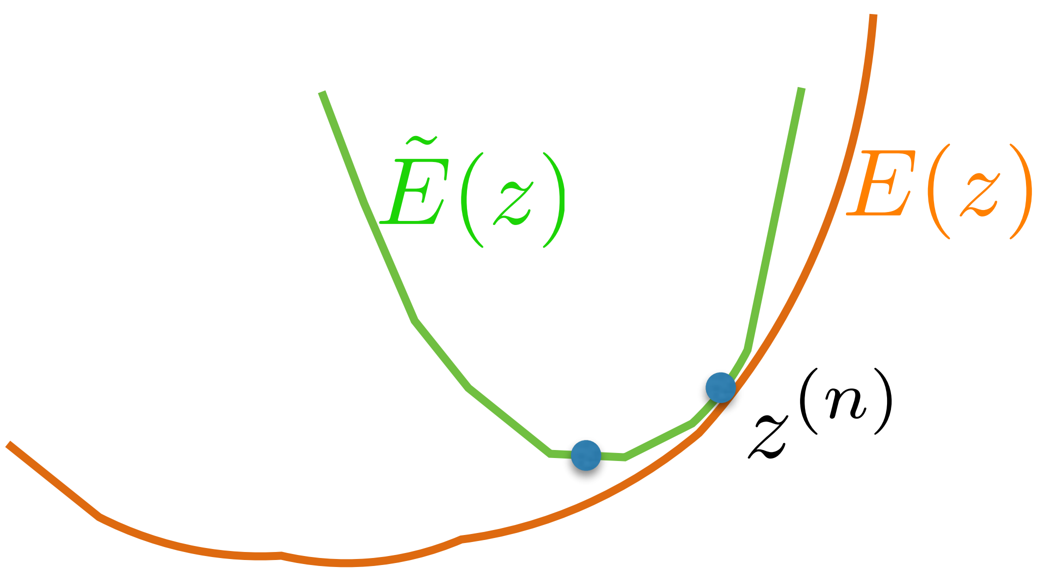
$$\|z\|_1 = \sum_i |z_i| \ .$$

- Using convexity we find an upper bound of the energy that is also separable:

$$E(z) \leq \tilde{E}(z; z^{(n)}) = E(z^{(n)}) + \langle B(z^{(n)} - y), z - z^{(n)} \rangle + Q(z, z^{(n)}) \ , \text{ with}$$

$$Q(z,u) = \frac{1}{2}(z-u)^T S(z-u) + \lambda \|z\|_1 \qquad B = D^T D \ , \ y = D^\dagger x$$

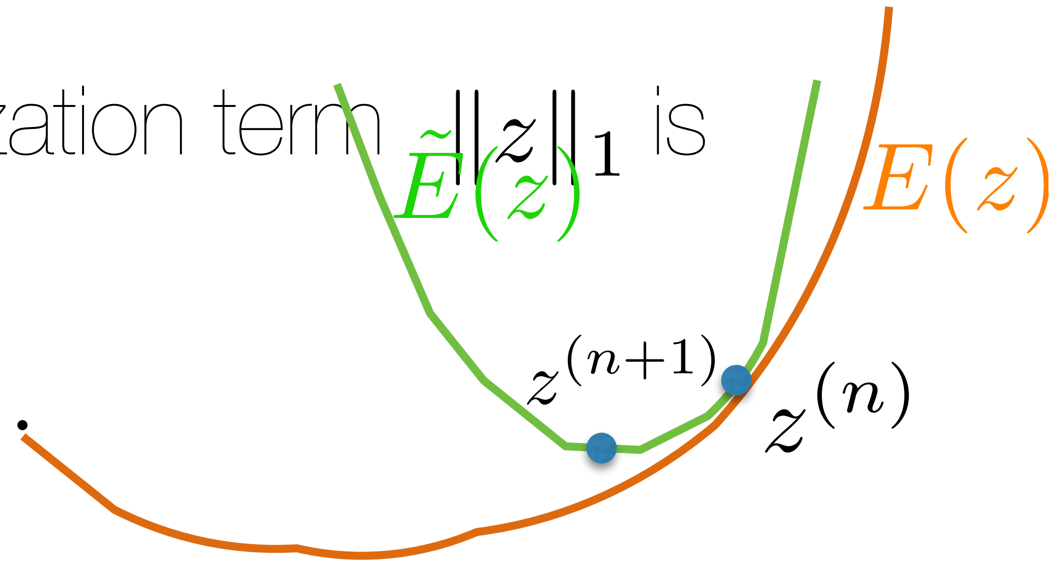
S diagonal such that $S - B \succ 0$.



Sparsity Stable Matrix Factorizations

- Principle of proximal splitting: the regularization term $\|z\|_1$ is *separable* in the canonical basis:

$$\|z\|_1 = \sum_i |z_i| \ .$$



- Using convexity we find an upper bound of the energy that is also separable:

$$E(z) \leq \tilde{E}(z; z^{(n)}) = E(z^{(n)}) + \langle B(z^{(n)} - y), z - z^{(n)} \rangle + Q(z, z^{(n)}) \ , \ \text{with}$$

$$Q(z, u) = \frac{1}{2}(z - u)^T S(z - u) + \lambda \|z\|_1 \qquad B = D^T D \ , \ y = D^\dagger x$$

S diagonal such that $S - B \succ 0$.

- Explicit minimization via the proximal operator:

$$\boxed{z^{(n+1)} = \arg\min_z \langle B(z^{(n)} - y), z - z^{(n)} \rangle + Q(z, z^{(n)}) \ .}$$

Sparsity Stable Matrix Factorizations

[joint work with Th. Moreau (ENS)]

- Consider now unitary matrix A and

$$E(z) \leq \tilde{E}_A(z; z^{(n)}) = E(z^{(n)}) + \langle B(z^{(n)} - y), z - z^{(n)} \rangle + Q(Az, Az^{(n)}) \ .$$

Sparsity Stable Matrix Factorizations

[joint work with Th. Moreau (ENS)]

- Consider now unitary matrix A and

$$E(z) \leq \tilde{E}_A(z; z^{(n)}) = E(z^{(n)}) + \langle B(z^{(n)} - y), z - z^{(n)} \rangle + Q(Az, Az^{(n)}) ~.$$

- <u>Observation</u>: $\tilde{E}_A(z; z^{(n)})$ still admits an explicit solution via a proximal operator:

$$\arg\min_z \tilde{E}_A(z; z^{(n)}) =$$

$$A^T \arg\min_z \left(\langle v, z \rangle + \frac{1}{2}(z - Az^{(n)})^T S (z - Az^{(n)}) + \lambda \|z\|_1 \right) ~.$$

- Q: How to choose the rotation A?

Sparsity Stable Matrix Factorizations

[joint work with Th. Moreau (ENS)]

- We denote

$$\delta_A(z) = \lambda(\|Az\|_1 - \|z\|_1) \ , \ R = A^T S A - B$$

- $\delta_A(z)$ measures the invariance of the ℓ_1 ball by the action of A.

Sparsity Stable Matrix Factorizations

[joint work with Th. Moreau (ENS)]

- We denote

$$\delta_A(z) = \lambda(\|Az\|_1 - \|z\|_1) \ , \ R = A^T S A - B$$

- $\delta_A(z)$ measures the invariance of the ℓ_1 ball by the action of A .

Proposition: If $R \succ 0$ and $z^{(n+1)} = \arg\min_z \tilde{E}_A(z; z^{(n)})$ then

$$E(z^{(n+1)}) - E(z^*) \leq \frac{1}{2}(z^* - z^{(n)})^T R(z^* - z^{(n)}) + \delta_A(z^*) - \delta_A(z^{(n+1)}) \ .$$

- We are thus interested in factorizations (A, S) such that
  - $\|R\|$ is small,
  - $|\delta_A(z) - \delta_A(z')|$ is small.
- Q: When are these factorizations possible? Consequences?

Certificate of Acceleration for Random Designs

- Let $D \in \mathbb{R}^{n \times m}$ be a generic dictionary with iid entries.
- Let $z_k \in \mathbb{R}^m$ be a current estimate of

$$z^* = \arg\min_z \frac{1}{2}\|x - Dz\|^2 + \lambda\|z\|_1 \ .$$

- **Theorem:** [Moreau, B'17] Then if

$$\lambda\|z_k\|_1 \leq \sqrt{\frac{m(m-1)}{n}}\|z_k - z^*\|_2^2$$

the upper bound is optimized away from $A = \mathbf{1}$.

Certificate of Acceleration for Random Designs

- Let $D \in \mathbb{R}^{n \times m}$ be a generic dictionary with iid entries.
- Let $z_k \in \mathbb{R}^m$ be a current estimate of

$$z^* = \arg\min_z \frac{1}{2}\|x - Dz\|^2 + \lambda\|z\|_1 \ .$$

- **Theorem:** [Moreau, B'17] Then if

$$\lambda\|z_k\|_1 \leq \sqrt{\frac{m(m-1)}{n}}\|z_k - z^*\|_2^2$$

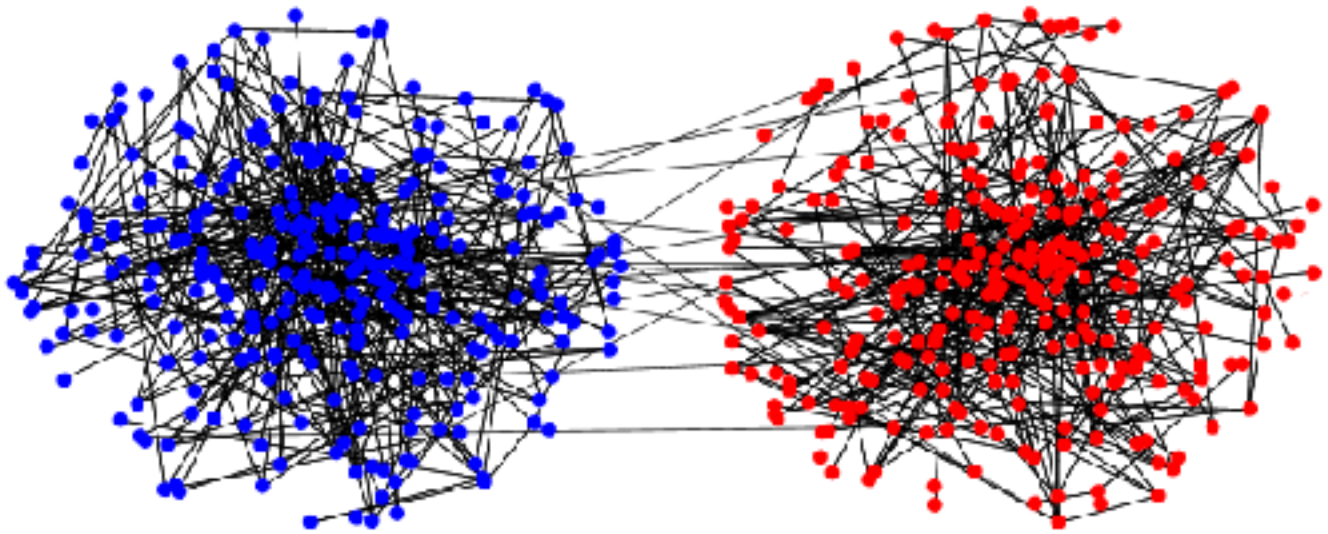
the upper bound is optimized away from $A = \mathbf{1}$.

- Remarks:
  - Transient Acceleration: only effective when far away from the solution.
  - Existence of acceleration improves as dimensionality increases.
  - Related to Sparse PCA [d'Aspremont, Rigollet, el Ganoui, et al.]

Statistical Inference on Graphs

[joint work with Lisha Li (UC Berkeley)]

- A related setup is spectral clustering / community detection:



- Detecting community structure as optimizing a constrained quadratic form (Min Cut / Max-Flow): $$\min_{y_i=\pm 1; \bar{y}=0} y^T \mathcal{A}(G) y \ .$$

- Detecting community by posterior inference on MRF:

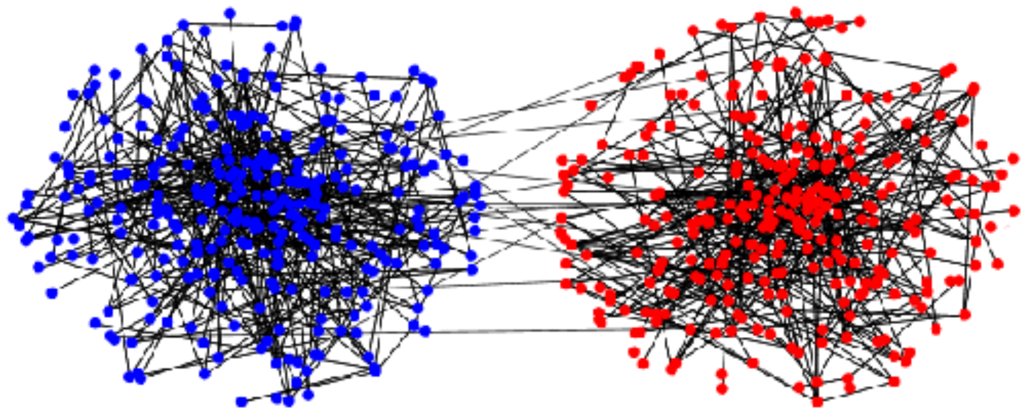
$$p(G \mid y) \propto \prod_{(i,j)\in E} \varphi(y_i, y_j) \prod_{i \in V} \psi_i(y_i) \ .$$

- Q: Can these algorithms be made data-driven? Why/ How ?

Data-Driven Community Detection

[joint work with Lisha Li (UC Berkeley)]

- A first setup is to consider the symmetric, binary Stochastic Block Model $W \sim \mathrm{SBM}(p, q)$



- Two recovery regimes:
  - *Exact* recovery: $\Pr(\hat{y} = y) \to 1 \ (n \to \infty)$ when

$$p = \frac{a \log n}{n}, \ q = \frac{b \log n}{n}, \sqrt{a} - \sqrt{b} \geq \sqrt{2} \ .$$

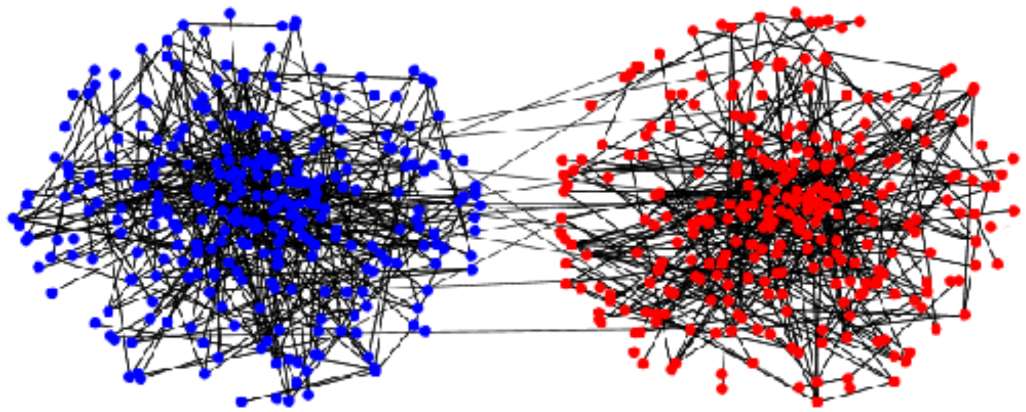
  - *Detection:* $\exists \epsilon > 0 \, ; \, \Pr(\hat{y} = y) > \frac{1}{2} + \epsilon \ (n \to \infty)$ when

$$p = \frac{a}{n}, \ q = \frac{b}{n}, \ (a - b)^2 > 2(a + b) \ .$$

Data-Driven Community Detection

[joint work with Lisha Li (UC Berkeley)]

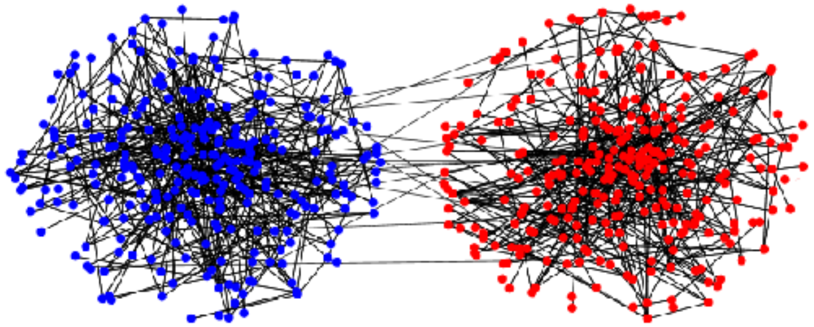
- A first setup is to consider the symmetric, binary Stochastic Block Model $$W \sim \mathrm{SBM}(p, q)$$



- Two recovery regimes:
  - *Exact* recovery: $\Pr(\hat{y} = y) \to 1 \ (n \to \infty)$ when
  $$p = \frac{a \log n}{n}, \ q = \frac{b \log n}{n}, \sqrt{a} - \sqrt{b} \geq \sqrt{2} \ .$$
  - *Detection:* $\exists \epsilon > 0 \, ; \, \Pr(\hat{y} = y) > \frac{1}{2} + \epsilon \ (n \to \infty)$ when
  $$p = \frac{a}{n}, \ q = \frac{b}{n}, \ (a-b)^2 > 2(a+b) \ .$$
- Algorithms to achieve *information-theoretic threshold*:
  - "Perturbed Spectral Methods" achieve the threshold on both regimes.
  - Loopy Belief propagation: thanks to the local-tree structure.

Data-driven Community Detection

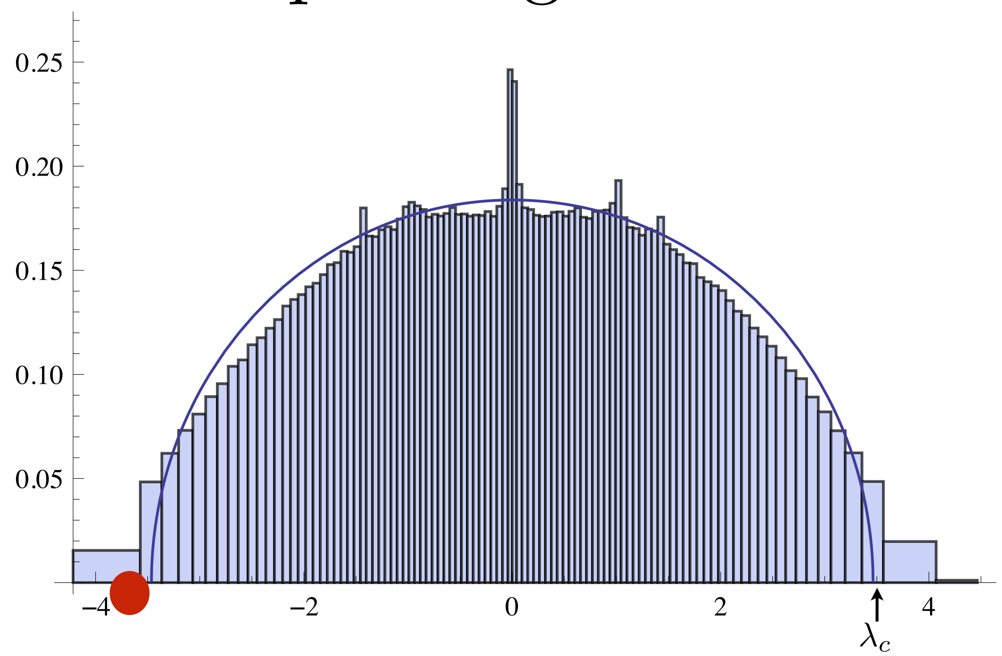
- $\mathcal{A}(G)$: linear operator defined on G, eg Laplacian $\Delta = D - A$.



- Spectral Clustering estimators:

$$\hat{y} = \mathrm{sign}\left(\mathrm{Fiedler}(\mathcal{A}(G))\right) \ ,$$

Fiedler(M): eigenvector corresponding to 2nd smallest eigenvalue

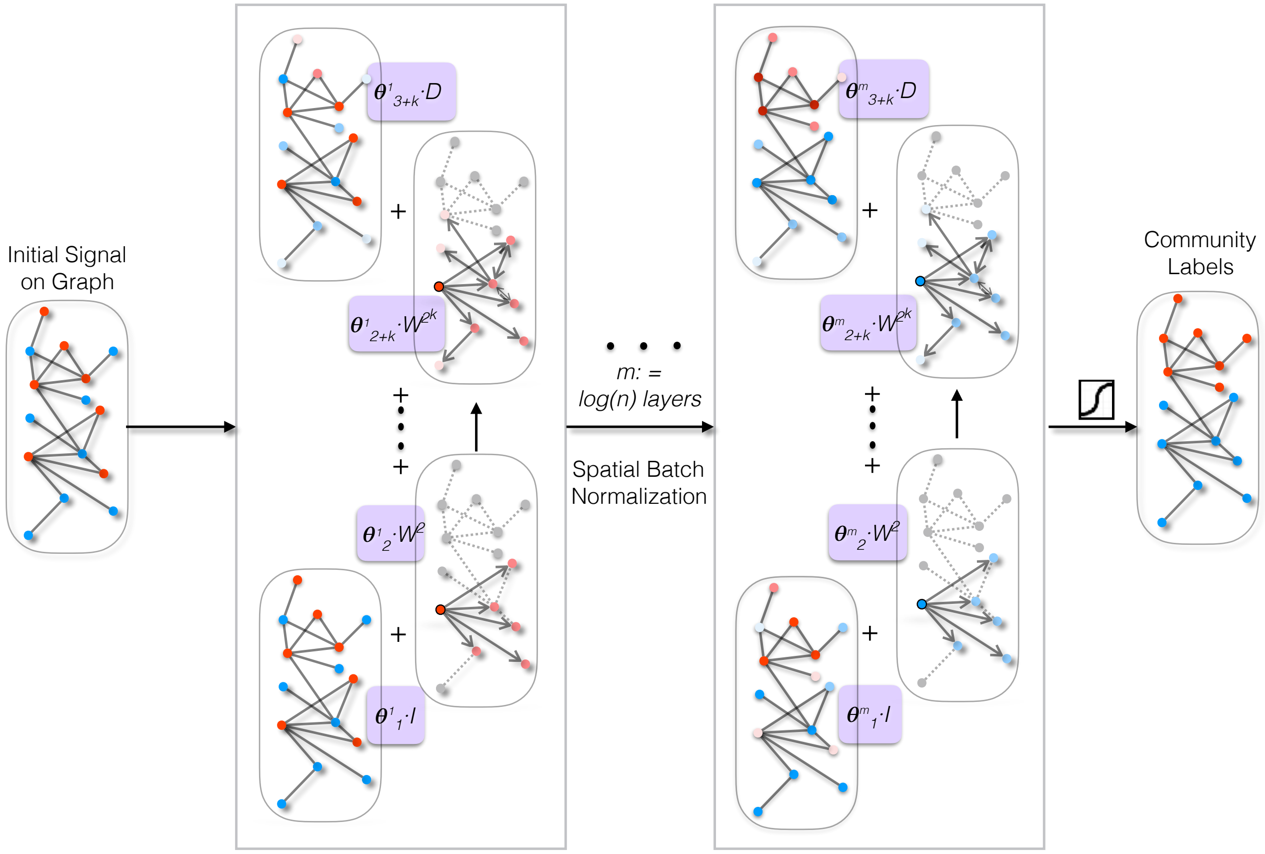


- Iterative algorithm: projected power iterations on shifted $\mathcal{A}(G)$:

$$M = \|\mathcal{A}(G)\|\mathbf{1} - \mathcal{A}(G)$$

Data-Driven Community Detection

- The resulting neural network architecture is a Graph Neural network [Scarselli et al.'09 , Bruna et al. '14] generated by operators $\{\mathbf{1}, A, D\}$: $\tilde{x} = \rho\left(\theta_1 x + \theta_2 D x + \theta_3 A x\right)$.

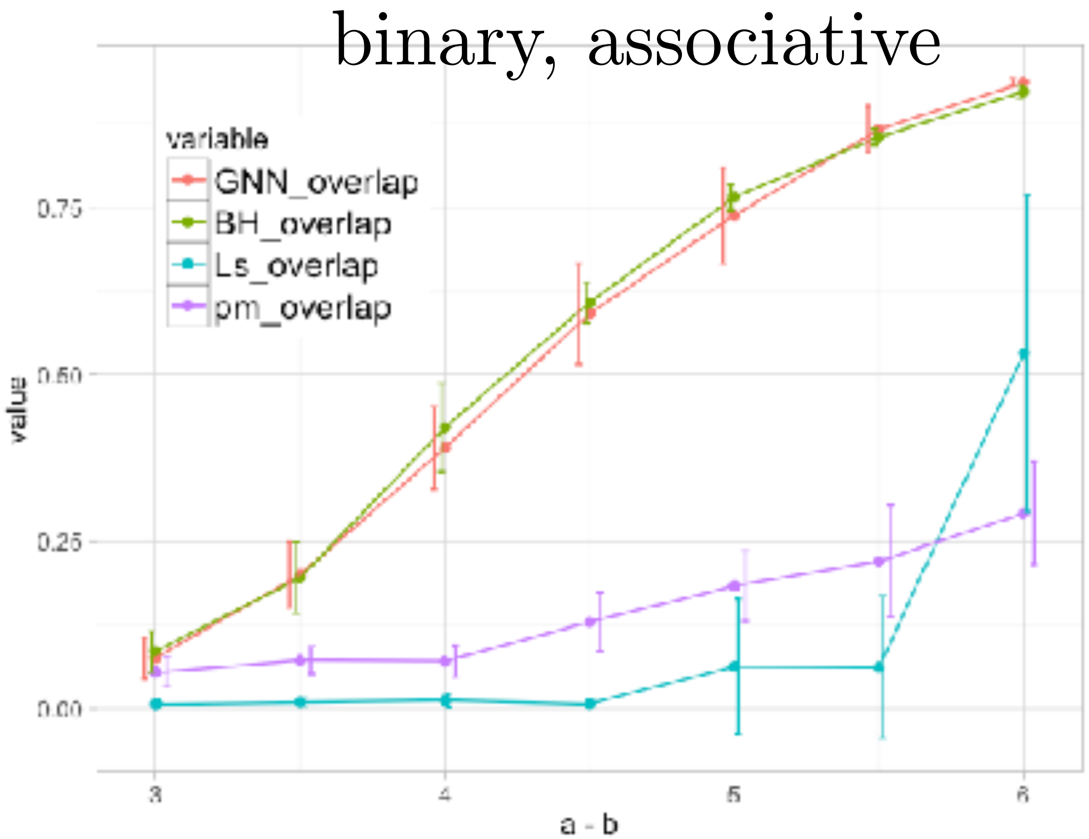


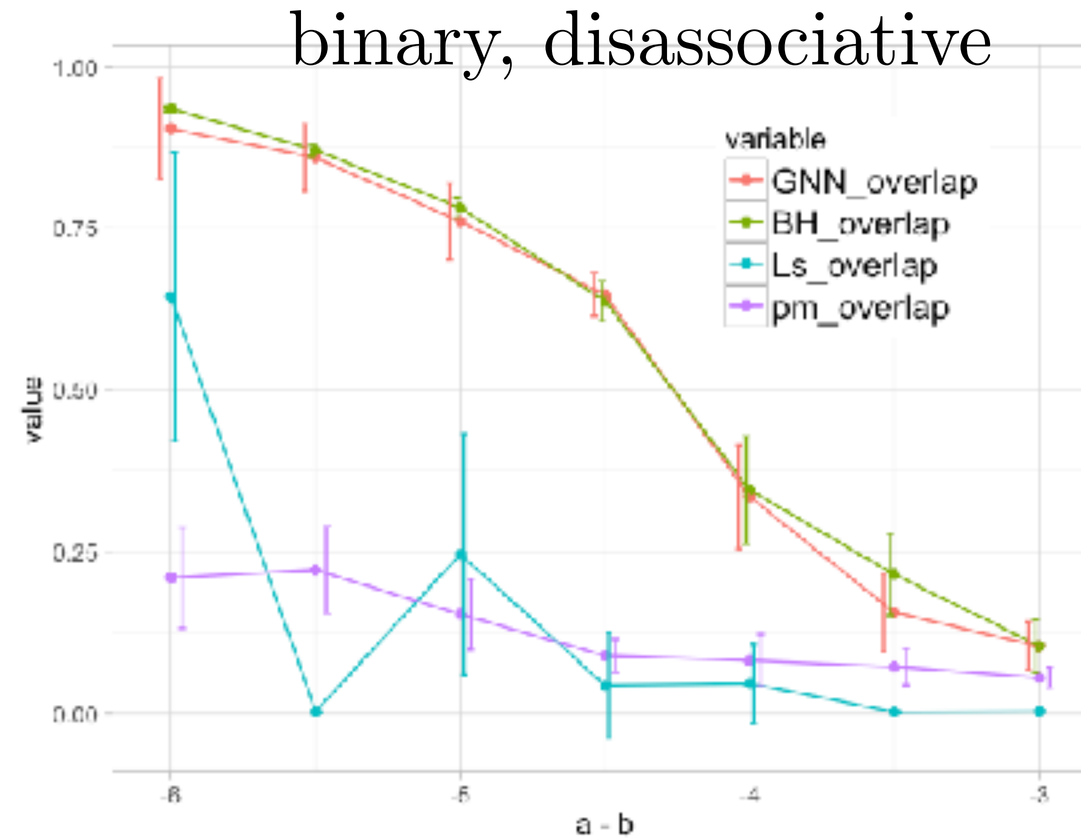
- We train it by back propagation using a loss that is globally invariant to label permutations:

$$E(\Theta) = \mathbb{E}_{W,y \sim \mathrm{SBM}} \ell(\Phi(W;\Theta), y) \ , \ \hat{E}(\Theta) = \frac{1}{L} \sum_{(W_l, y_l) \sim \mathrm{SBM}} \ell(\Phi(W_l;\Theta), y_l)$$

Reaching Detection Threshold on SBM

- Stochastic Block Model Results:





  – we reach the detection threshold, matching the specifically designed spectral method.

- Real-world community detection results:

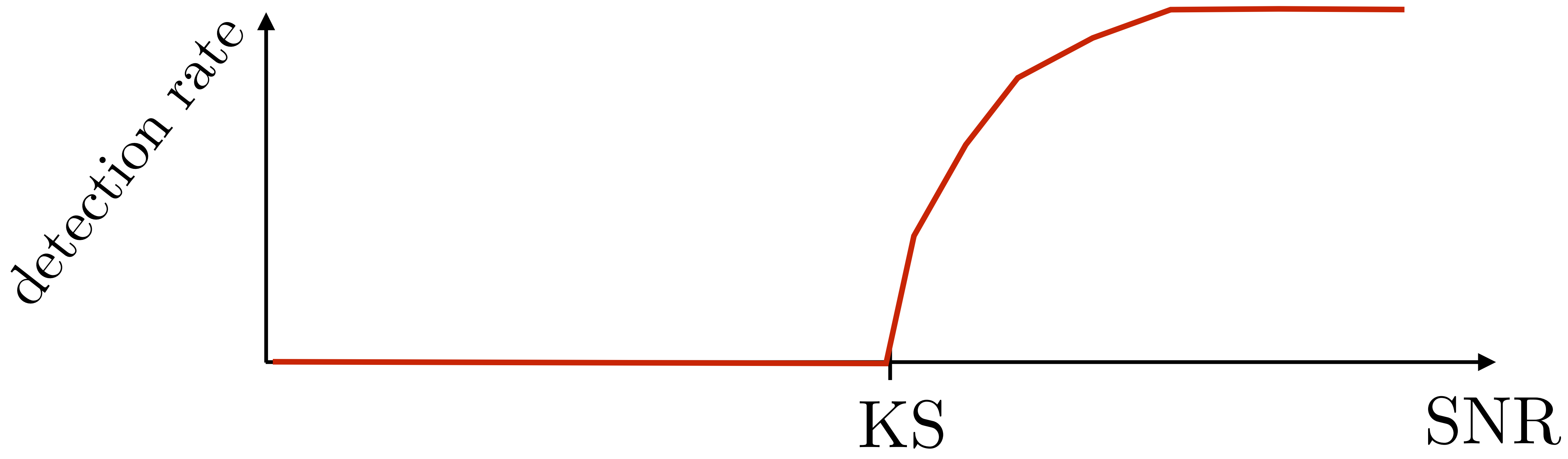
Table 1: Snap Dataset Performance Comparison between GNN and AGM

	Subgraph Instances			Overlap Comparison	
Dataset	(train/test)	Avg Vertices	Avg Edges	GNN	AGMFit
Amazon	315 / 35	60	346	$\mathbf{0.74 \pm 0.13}$	$\mathbf{0.76 \pm 0.08}$
DBLP	2831 / 510	26	164	$\mathbf{0.78 \pm 0.03}$	0.64 ± 0.01
Youtube	48402 / 7794	61	274	$\mathbf{0.9 \pm 0.02}$	0.57 ± 0.01

Phase Transitions in Learning

[with A. Bandeira, S. Villar, Z. Chen (NYU)]

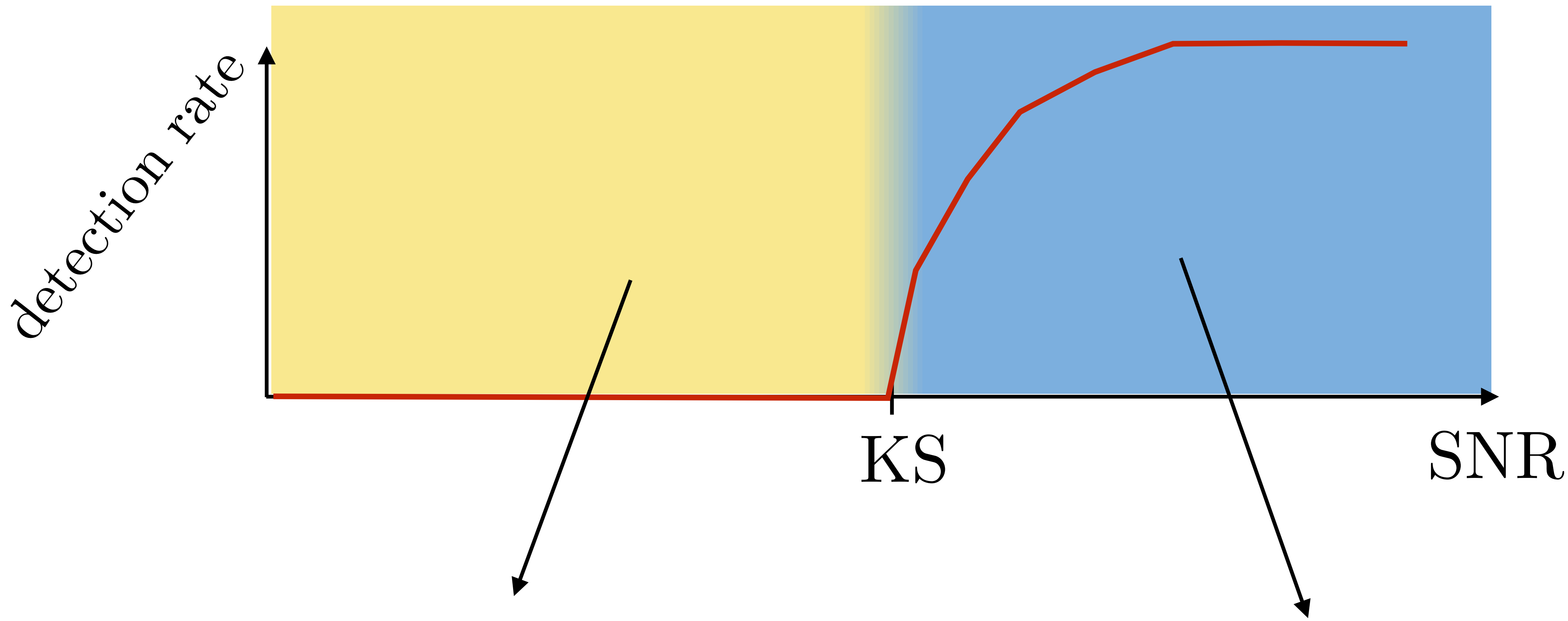
- In this binary setting, the computational threshold matches the IT threshold:



Phase Transitions in Learning

[with A. Bandeira, S. Villar, Z. Chen (NYU)]

- In this binary setting, the computational threshold matches the IT threshold:



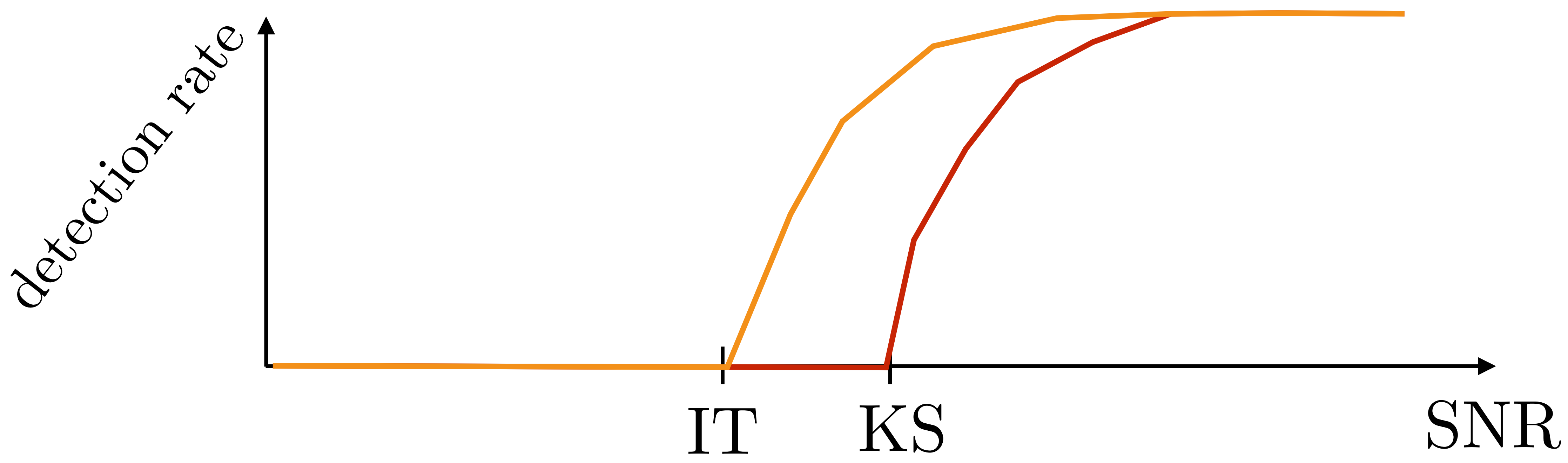
Landscape of $E(\Theta)$ simple/complex?
$\hat{E}(\Theta)$

- A priori, no reason why below IT threshold landscape should be more complex?

Phase Transitions in Learning

[with A. Bandeira, S. Villar, Z. Chen (NYU)]

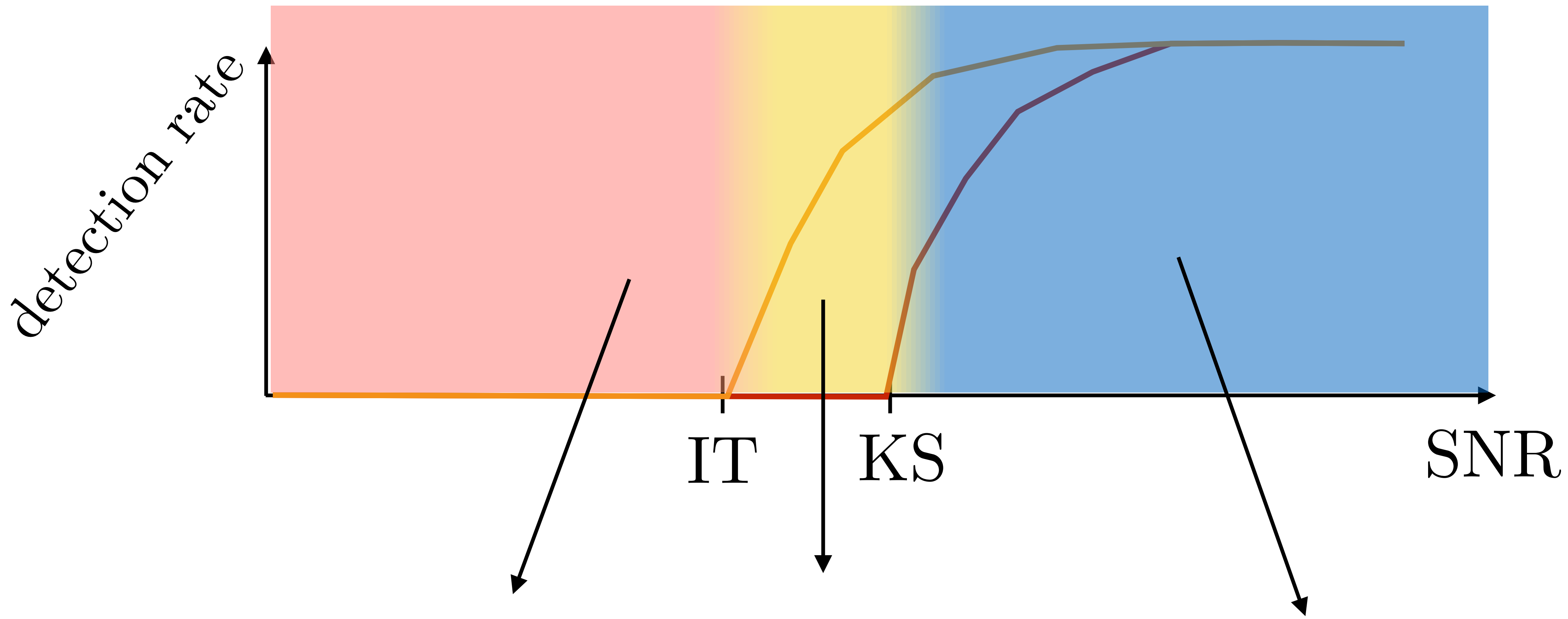
- For more general setups (k>3 communities), the computational threshold might not match IT threshold:



Phase Transitions in Learning

[with A. Bandeira, S. Villar, Z. Chen (NYU)]

- For more general setups (k>3 communities), the computational threshold might not match IT threshold:



Landscape of $E(\Theta)$ simple/complex?
$\hat{E}(\Theta)$

- Studying complexity of learning may inform about this gap?

Thank you!