Everything You Wanted to Know About the Loss Surface but Were Afraid to Ask – The Talk

Boris Hanin

Texas A&M

August 21, 2018

Plan

Plan

1. Linear models:
    - Baldi-Hornik (1989)
    - Kawaguchi-Lu (2016)

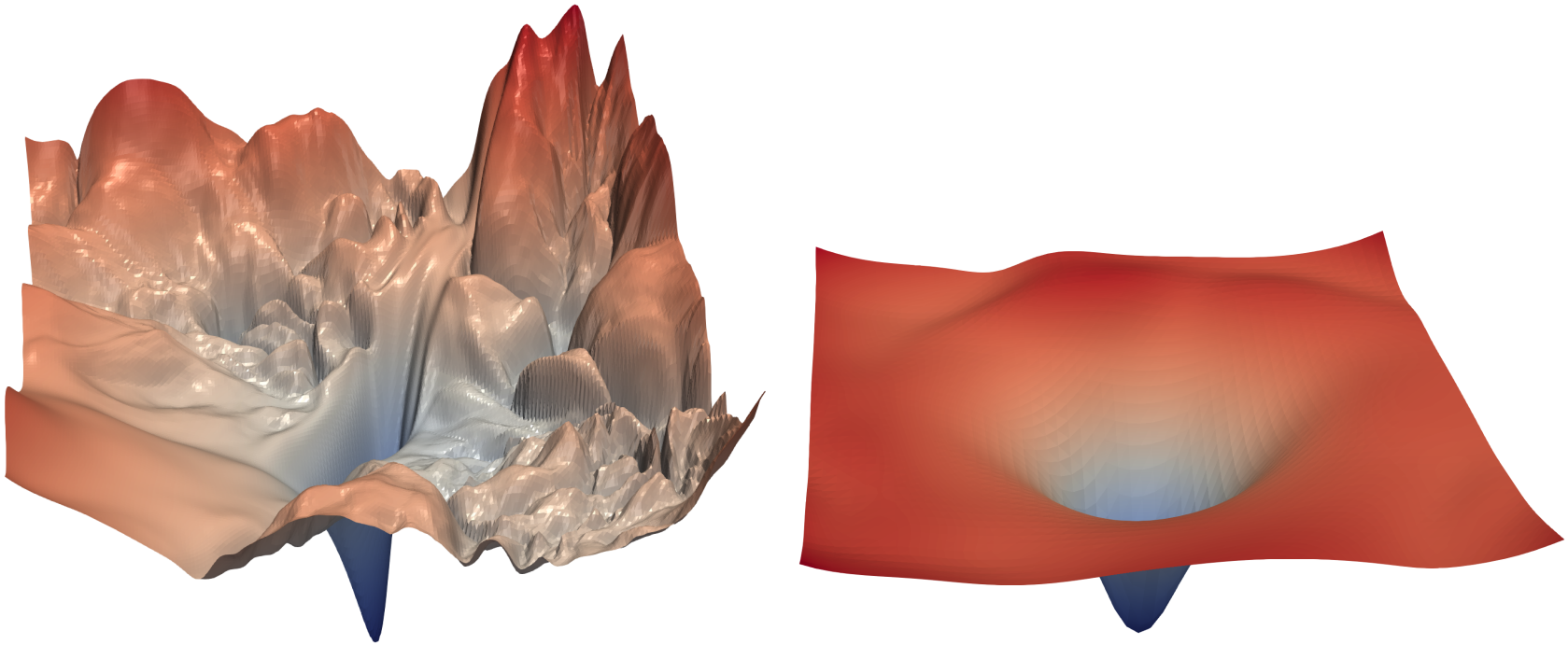
Plan

1. Linear models:
   - Baldi-Hornik (1989)
   - Kawaguchi-Lu (2016)
2. One hidden layer:
   - Ge-Lee-Ma (2016)
   - Mei-Montanari-Nguyen (2018)
   - Venturi-Bandeira-Bruna (2018)

Plan

1. Linear models:
   - Baldi-Hornik (1989)
   - Kawaguchi-Lu (2016)
2. One hidden layer:
   - Ge-Lee-Ma (2016)
   - Mei-Montanari-Nguyen (2018)
   - Venturi-Bandeira-Bruna (2018)
3. More than one hidden layer:
   - Gori-Tesi (1992) + Nguyen-Hein (2017)

But First ... Turns out Skip Connections are Good



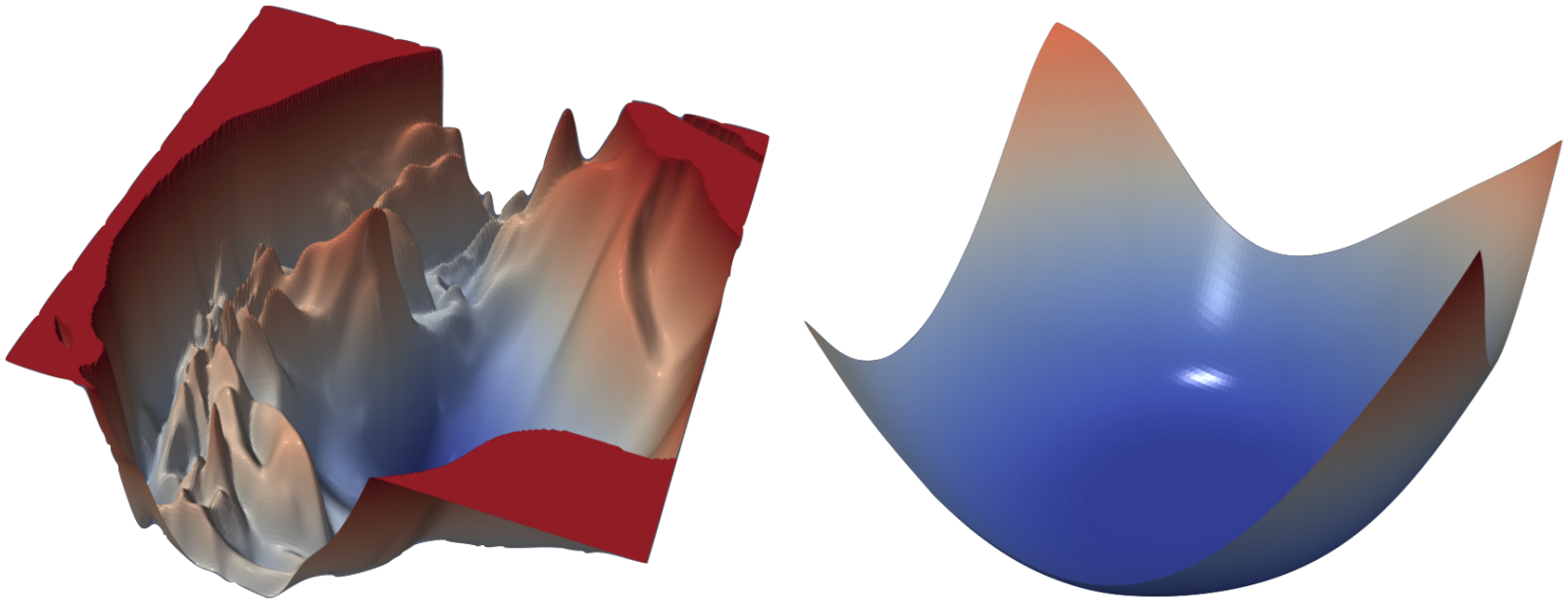
(a) without skip connections

(b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The vertical axis is logarithmic to show dynamic range. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Figure: From Visualizing the Loss Landscape of Neural Nets by Li, Xu, Taylor, Studer, Goldstein

But First ... Turns out Skip Connections are Good



(a) 110 layers, no skip connections

(b) DenseNet, 121 layers

Figure 6: (left) The loss surfaces of ResNet-110-noshort, without skip connections. (right) DenseNet, the current state-of-the-art network for CIFAR-10.

Figure: From Visualizing the Loss Landscape of Neural Nets by Li, Xu, Taylor, Studer, Goldstein

Setup to Baldi-Hornik (1989)

Setup to Baldi-Hornik (1989)

- Suppose

$$X = (x_j, 1 \leq j \leq N), \quad Y = (y_j, 1 \leq j \leq N), \ x_j \in \mathbb{R}^n, \ y_j \in \mathbb{R}^m.$$

Setup to Baldi-Hornik (1989)

- Suppose

$$X = (x_j,\, 1 \le j \le N)\,, \quad Y = (y_j,\, 1 \le j \le N)\,,\ x_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}^m.$$

- Define empirical (co)-variances:

$$\Sigma_{YX} = \sum_j y_j x_j^T, \qquad \Sigma_{XX} = \sum_j x_j x_j^T, \qquad \Sigma_{YY} = \sum_j y_j y_j^T.$$

Setup to Baldi-Hornik (1989)

- Suppose

$$X = (x_j,\, 1 \le j \le N)\,, \quad Y = (y_j,\, 1 \le j \le N)\,,\ x_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}^m.$$

- Define empirical (co)-variances:

$$\Sigma_{YX} = \sum_j y_j x_j^T, \qquad \Sigma_{XX} = \sum_j x_j x_j^T, \qquad \Sigma_{YY} = \sum_j y_j y_j^T.$$

- Classical least squares regression

$$A^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2 = \sum_j \|Ax_j - y_j\|^2$$

is a convex problem and can be solved by GD or analytically by

$$A^* = \Sigma_{YX}\Sigma_{XX}^{-1}.$$

Setup to Baldi-Hornik (1989)

Setup to Baldi-Hornik (1989)

- **Q.** What about rank-constrained least squares:

$$A_k^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2, \qquad \operatorname{rank}(A) \leq k < n?$$

Setup to Baldi-Hornik (1989)

- **Q.** What about rank-constrained least squares:

$$A_k^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2, \qquad \operatorname{rank}(A) \leq k < n?$$

- **A.** A_k^* is top k principal components of A^*.

Setup to Baldi-Hornik (1989)

- **Q.** What about rank-constrained least squares:

$$A_k^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2, \qquad \operatorname{rank}(A) \leq k < n?$$

- **A.** A_k^* is top k principal components of A^*.

- **Q.** Note that if A is $n \times m$ then

$$\operatorname{rank}(A) \leq k \quad \Rightarrow \quad A = BC, \;\; B - n \times k, \;\; C - k \times m.$$

Setup to Baldi-Hornik (1989)

- **Q.** What about rank-constrained least squares:

$$A_k^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2, \qquad \operatorname{rank}(A) \leq k < n?$$

- **A.** A_k^* is top k principal components of A^*.

- **Q.** Note that if A is $n \times m$ then

$$\operatorname{rank}(A) \leq k \quad \Rightarrow \quad A = BC, \quad B - n \times k, \quad C - k \times m.$$

- RC least squares $\leftrightarrow$ linear net with one layer and L^2 loss:

$$(A^*, B^*) = \operatorname{argmin} \mathcal{L}(A, B) = \|ABX - Y\|_F^2$$

up to $A \mapsto AC$, $B \mapsto C^{-1}B$, $\quad C \in GL_k$.

Setup to Baldi-Hornik (1989)

- **Q.** What about rank-constrained least squares:

$$A_k^* = \operatorname{argmin} \mathcal{L}(A) = \|AX - Y\|_F^2, \qquad \operatorname{rank}(A) \leq k < n?$$

- **A.** A_k^* is top k principal components of A^*.

- **Q.** Note that if A is $n \times m$ then

$$\operatorname{rank}(A) \leq k \quad \Rightarrow \quad A = BC, \;\; B - n \times k, \;\; C - k \times m.$$

- RC least squares $\leftrightarrow$ linear net with one layer and L^2 loss:

$$(A^*, B^*) = \operatorname{argmin} \mathcal{L}(A, B) = \|ABX - Y\|_F^2$$

up to $A \mapsto AC$, $B \mapsto C^{-1}B$, $\quad C \in GL_k$.

- **Key.** $\mathcal{L}(A, B)$ is not convex, so unclear if can solve by GD.

Results of Baldi-Hornik (1989)

Results of Baldi-Hornik (1989)

- Suppose Σ_{XX} invertible and $\Sigma = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ has full rank.
- **Thm.** All local minima of $\mathcal{L}(A, B)$ are global minima. Thus, GD should be fine.

Results of Baldi-Hornik (1989)

- Suppose Σ_{XX} invertible and $\Sigma = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ has full rank.
- **Thm.** All local minima of $\mathcal{L}(A, B)$ are global minima. Thus, GD should be fine.
- **Thm.** The saddle points of $\mathcal{L}(A, B)$ correspond to

$$AB = \mathrm{proj}_k(A^*),$$

where proj_k is projection onto some k principal component directions of A^*.

Results of Baldi-Hornik (1989)

- Suppose Σ_{XX} invertible and $\Sigma = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ has full rank.
- **Thm.** All local minima of $\mathcal{L}(A, B)$ are global minima. Thus, GD should be fine.
- **Thm.** The saddle points of $\mathcal{L}(A, B)$ correspond to

$$AB = \mathrm{proj}_k(A^*),$$

  where proj_k is projection onto some k principal component directions of A^*.
- The proofs are by explicit computation.

Proof Idea Baldi-Hornik (1989)

- Write $\mathcal{L}(A, B) = \operatorname{vec}(ABX - Y)^T \operatorname{vec}(ABX - Y)$.

Proof Idea Baldi-Hornik (1989)

- Write $\mathcal{L}(A, B) = \operatorname{vec}(ABX - Y)^T \operatorname{vec}(ABX - Y)$.
- Use identity

$$\operatorname{vec}(ABC) = \left(C^T \otimes A\right) \operatorname{vec}(B)$$

to differentiate with respect to $\operatorname{vec}(A), \operatorname{vec}(B)$.

Proof Idea Baldi-Hornik (1989)

- Write $\mathcal{L}(A, B) = \operatorname{vec}(ABX - Y)^T \operatorname{vec}(ABX - Y)$.
- Use identity

$$\operatorname{vec}(ABC) = \left(C^T \otimes A\right) \operatorname{vec}(B)$$

to differentiate with respect to $\operatorname{vec}(A), \operatorname{vec}(B)$.

- Check that at critical point $W = (A, B)$, we must have

$$P_A \Sigma P_A = P_A \Sigma = P_A \Sigma,$$

where $P_A = \operatorname{proj}_{\operatorname{col}(A)}$.

Kawaguchi-Lu (2016)

Kawaguchi-Lu (2016)

- Want to understand local minima of

$$\mathcal{L}(W_1, \ldots, W_d) = \|W_d \cdots W_1 X - Y\|_F^2,$$

when $W_1, \ldots, W_d$ have fixed shape.

Kawaguchi-Lu (2016)

- Want to understand local minima of

$$\mathcal{L}(W_1, \ldots, W_d) = \|W_d \cdots W_1 X - Y\|_F^2,$$

  when $W_1, \ldots, W_d$ have fixed shape.

- **Thm.** All local minima of $\mathcal{L}$ are global minima.

Kawaguchi-Lu (2016)

- Want to understand local minima of

$$\mathcal{L}(W_1, \ldots, W_d) = \|W_d \cdots W_1 X - Y\|_F^2,$$

  when $W_1, \ldots, W_d$ have fixed shape.

- **Thm.** All local minima of $\mathcal{L}$ are global minima.

- Idea of Proof:

  1. Every local minimum $\{W_j\}$ can be perturbed to local minimum $\left\{\widehat{W}_j\right\}$ with same value of loss and all $\widehat{W}_j$ of full rank.

Kawaguchi-Lu (2016)

- Want to understand local minima of

$$\mathcal{L}(W_1, \ldots, W_d) = \|W_d \cdots W_1 X - Y\|_F^2,$$

  when $W_1, \ldots, W_d$ have fixed shape.

- **Thm.** All local minima of $\mathcal{L}$ are global minima.

- Idea of Proof:

  1. Every local minimum $\{W_j\}$ can be perturbed to local minimum $\left\{\widehat{W}_j\right\}$ with same value of loss and all $\widehat{W}_j$ of full rank.
  2. Perturbations of $\widehat{W}_j's$ give all rank $\leq \min\{\text{layer widths}\}$ pertrubations of $\prod_j \widehat{W}_j$.

Kawaguchi-Lu (2016)

- Want to understand local minima of

$$\mathcal{L}(W_1, \ldots, W_d) = \|W_d \cdots W_1 X - Y\|_F^2,$$

  when $W_1, \ldots, W_d$ have fixed shape.

- **Thm.** All local minima of $\mathcal{L}$ are global minima.

- Idea of Proof:

  1. Every local minimum $\{W_j\}$ can be perturbed to local minimum $\left\{\widehat{W}_j\right\}$ with same value of loss and all $\widehat{W}_j$ of full rank.
  2. Perturbations of $\widehat{W}_j's$ give all rank $\leq \min\{\text{layer widths}\}$ pertrubations of $\prod_j \widehat{W}_j$.
  3. Thus, $\left\{\prod_{j \neq \min} \widehat{W}_j, \widehat{W}_{\min}\right\}$ is local minimum of Baldi-Hornik problem. Hence, global minimum.

Loss for One Layer, Gaussian Inputs: Ge-Lee-Ma (2016)

Loss for One Layer, Gaussian Inputs: Ge-Lee-Ma (2016)

- **Input/Output Distribution.** Inputs $x_j \sim N(0, \mathrm{Id}_n)$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = {a^*}^T \sigma\left(B^* x\right) = \sum_{k=1}^{n} a_k^* \sigma(\langle b_k^*, x_j \rangle), \quad \|b_k^*\| = \|a^*\| = 1.$$

Loss for One Layer, Gaussian Inputs: Ge-Lee-Ma (2016)

- **Input/Output Distribution.** Inputs $x_j \sim N(0, \mathrm{Id}_n)$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = {a^*}^T \sigma\left(B^* x\right) = \sum_{k=1}^{n} a_k^* \sigma(\langle b_k^*, x_j \rangle), \quad \|b_k^*\| = \|a^*\| = 1.$$

- **Thm.** If $\|a\| = \|b_k\| = 1$, then L^2 population risk

$$\mathbb{E}\left[\|y(x, a, B) - y(x, a^*, B^*)\|^2\right]$$

is a sum of rank m tensor-norms:

$$\sum_{m \geq 0} \widehat{\sigma}_m^2 \left\| \sum_{k=1}^{n} a_k b_k^{\otimes m} - \sum_{k=1}^{n} a_k^* (b_k^*)^{\otimes m} \right\|_F^2,$$

where $\widehat{\sigma}_m$ are the Hermite coefficients of σ :

$$\sigma(t) = \sum_{m \geq 0} \frac{\widehat{\sigma}_m}{m!} H_m(t).$$

Proof Ingredients for Ge-Lee-Ma (2016)

Proof Ingredients for Ge-Lee-Ma (2016)

1. Orthogonality:

$$\int_{\mathbb{R}} H_n(x) H_m(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \delta_{n=m} n!.$$

Proof Ingredients for Ge-Lee-Ma (2016)

1. Orthogonality:

$$\int_{\mathbb{R}} H_n(x) H_m(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \delta_{n=m} n!.$$

2. Sum-product formula for Hermite polynomials:

$$H_n(x \cdot y) = \sum_{|p|=n} \binom{n}{p} \prod_{j=1}^{d} H_{p_j}(x_j) y^{p_j}, \qquad x, y \in \mathbb{R}^d, \ \|y\| = 1,$$

Proof Ingredients for Ge-Lee-Ma (2016)

1. Orthogonality:

$$\int_{\mathbb{R}} H_n(x) H_m(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \delta_{n=m} n!.$$

2. Sum-product formula for Hermite polynomials:

$$H_n(x \cdot y) = \sum_{|p|=n} \binom{n}{p} \prod_{j=1}^{d} H_{p_j}(x_j) y^{p_j}, \qquad x, y \in \mathbb{R}^d, \ \|y\| = 1,$$

where the sum is over all multi-indices p.

3. Wick formula

$$\mathbb{E}\left[\frac{H_k(v^T x)}{k!} \frac{H_j(w^T x)}{j!}\right] = \delta_{j=k} \quad \frac{1}{k!} \langle v, w \rangle^k .$$

SGD for One Layer: Mei-Montanari-Nguyen (2018)

SGD for One Layer: Mei-Montanari-Nguyen (2018)

- **Input/Output Distribution.** Inputs $x_j \sim \mathbb{P}$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle x, w_i\rangle + b_i\right).$$

SGD for One Layer: Mei-Montanari-Nguyen (2018)

- **Input/Output Distribution.** Inputs $x_j \sim \mathbb{P}$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle x, w_i\rangle + b_i\right).$$

- L^2 loss $R_N(\theta) = \frac{1}{2}\mathbb{E}[\|y_N - \widehat{y}\|^2]$ is pair interaction in potential:

$$R(\rho) := \int V(\theta) d\,\rho^{(N)}(\theta) \;+\; \frac{1}{2}\int U(\theta_1,\theta_2)\, d\rho^{(N)}(\theta_1) d\rho^{(N)}(\theta_2),$$

SGD for One Layer: Mei-Montanari-Nguyen (2018)

- **Input/Output Distribution.** Inputs $x_j \sim \mathbb{P}$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle x, w_i\rangle + b_i\right).$$

- L^2 loss $R_N(\theta) = \frac{1}{2}\mathbb{E}[\|y_N - \widehat{y}\|^2]$ is pair interaction in potential:

$$R(\rho) := \int V(\theta) d\,\rho^{(N)}(\theta) + \frac{1}{2}\int U(\theta_1,\theta_2)\, d\rho^{(N)}(\theta_1) d\rho^{(N)}(\theta_2),$$

plus a constant where $d\rho^{(N)}(\theta) = \sum_{i=1}^{N} \delta_{\theta_i}$ and

$$V(\theta) = -\mathbb{E}\left[y\sigma_*(x,\theta)\right], \quad U(\theta_1,\theta_2) = \mathbb{E}\left[\sigma_*(x,\theta_1)\sigma_*(x,\theta_2)\right].$$

SGD for One Layer: Mei-Montanari-Nguyen (2018)

- **Input/Output Distribution.** Inputs $x_j \sim \mathbb{P}$ i.i.d. and outputs y_j produced by net with one hidden layer:

$$y_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle x, w_i\rangle + b_i\right).$$

- L^2 loss $R_N(\theta) = \frac{1}{2}\mathbb{E}[\|y_N - \widehat{y}\|^2]$ is pair interaction in potential:

$$R(\rho) := \int V(\theta) d\,\rho^{(N)}(\theta) + \frac{1}{2}\int U(\theta_1,\theta_2)\, d\rho^{(N)}(\theta_1) d\rho^{(N)}(\theta_2),$$

plus a constant where $d\rho^{(N)}(\theta) = \sum_{i=1}^{N} \delta_{\theta_i}$ and

$$V(\theta) = -\mathbb{E}\left[y\sigma_*(x,\theta)\right], \quad U(\theta_1,\theta_2) = \mathbb{E}\left[\sigma_*(x,\theta_1)\sigma_*(x,\theta_2)\right].$$

- **Q.** What does SGD looks like in $\rho-$space?

SGD for One Layer: Mei-Montanari-Nguyen (2018)

SGD for One Layer: Mei-Montanari-Nguyen (2018)

- **Thm.** Consider step sizes $s_k = \epsilon\xi(k\epsilon)$ and write $\rho_k^{(N)}$ for empirical measure after k GD steps. Then

$$\rho_{t/\epsilon}^{(N)} \quad \Longrightarrow \quad \rho_t$$

as $N \to \infty$ and $\epsilon \to 0$, where ρ_t evolves under gradient flow for Wasserstein metric

$$\partial_t \rho_t = \xi(t) \nabla_\theta \left(\rho_t \nabla_\theta \Psi(\theta, \rho_t) \right)$$
$$\Psi(\theta, \rho) := V(\theta) + \int U(\theta, \theta') d\rho(\theta').$$

In particular, have a good approximation when $k = t/\epsilon$ with $\epsilon, N^{-1} \ll 1/\text{input-dim}$.

L^2–loss in Deep and Wide Networks: Nguyen-Hein (2017)

L^2-loss in Deep and Wide Networks: Nguyen-Hein (2017)

- σ – smooth, analytic, bounded
- N distinct training samples with L^2-loss

L^2-loss in Deep and Wide Networks: Nguyen-Hein (2017)

- σ – smooth, analytic, bounded
- N distinct training samples with L^2-loss
- **Thm.** Fix a critical point. Suppose
  1. $\exists$ layer k with $n_k \geq N-1$ neurons with Hessian of loss restricted to parameters in layers $\geq k+1$ is non-degenerate
  2. weight matrices in layers $\geq k+1$ all have full column rank.

  Then this critical point is a global minimum.
- **Rmk.** Second condition requires "pyramidal" structure.

Proof Idea of Nguyen-Hein (2017)

Proof Idea of Nguyen-Hein (2017)

- **Thm.** (Gori–Tesi 1992) If inputs are linearly independent, layer widths monotonically decreasing, then any crit at which the weight matrices have full rank is a global min.

Proof Idea of Nguyen-Hein (2017)

- **Thm.** (Gori−Tesi 1992) If inputs are linearly independent, layer widths monotonically decreasing, then any crit at which the weight matrices have full rank is a global min.

- **Idea.** Backprop relates derivative of loss with respect to activations at successive layers:

$$\frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j)}} = \frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j+1)}} \underbrace{D^{(j)} W^{(j)}}_{Z^{(j)}},$$

where

$$D^{(j)} = \operatorname{Diag}\left(\phi'\left(\operatorname{preact}_{\beta}^{(j)}\right),\ \beta = 1, \ldots, n_j\right).$$

Proof Idea of Nguyen-Hein (2017)

- **Thm.** (Gori−Tesi 1992) If inputs are linearly independent, layer widths monotonically decreasing, then any crit at which the weight matrices have full rank is a global min.

- **Idea.** Backprop relates derivative of loss with respect to activations at successive layers:

$$\frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j)}} = \frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j+1)}} \underbrace{D^{(j)} W^{(j)}}_{Z^{(j)}},$$

  where

$$D^{(j)} = \operatorname{Diag}\left(\phi'\left(\operatorname{preact}_{\beta}^{(j)}\right),\ \beta = 1, \ldots, n_j\right).$$

- If $\phi' \neq 0$ and $W^{(j)}$ has full rank, then can do "forward prop"

$$\frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j+1)}} = \frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j)}} Z^{(j)\,T} \left[Z^{(j)} Z^{(j)\,T}\right]^{-1}.$$

Proof Idea of Nguyen-Hein (2017)

- **Thm.** (Gori−Tesi 1992) If inputs are linearly independent, layer widths monotonically decreasing, then any crit at which the weight matrices have full rank is a global min.

- **Idea.** Backprop relates derivative of loss with respect to activations at successive layers:

$$\frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j)}} = \frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j+1)}} \underbrace{D^{(j)} W^{(j)}}_{Z^{(j)}},$$

  where

$$D^{(j)} = \operatorname{Diag}\left(\phi'\left(\operatorname{preact}_{\beta}^{(j)}\right),\ \beta = 1, \ldots, n_j\right).$$

- If $\phi' \neq 0$ and $W^{(j)}$ has full rank, then can do "forward prop"

$$\frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j+1)}} = \frac{\partial \mathcal{L}}{\partial \operatorname{Act}^{(j)}} Z^{(j)\,T} \left[Z^{(j)} Z^{(j)\,T}\right]^{-1}.$$

- Now just transplant this condition to some intermediate layer.

Finite Intrinsic Dimension: Venturi-Bandeira-Bruna (2018)

Finite Intrinsic Dimension: Venturi-Bandeira-Bruna (2018)

- Consider one layer network with no bias, L^2 loss, and hidden layer of width p

$$\Phi(x,\theta) = U\rho(Wx), \qquad L(\theta) := \mathbb{E}\left[\|\Phi(X,\theta) - Y\|^2\right].$$

Finite Intrinsic Dimension: Venturi-Bandeira-Bruna (2018)

- Consider one layer network with no bias, L^2 loss, and hidden layer of width p

$$\Phi(x,\theta) = U\rho(Wx), \qquad L(\theta) := \mathbb{E}\left[\|\Phi(X,\theta) - Y\|^2\right].$$

- Consider span of possible features

$$V := \mathrm{Span}\left\{\psi_w, \quad w \in \mathbb{R}^n\right\}, \qquad \psi_w := \rho(\langle w, \cdot \rangle)$$

Finite Intrinsic Dimension: Venturi-Bandeira-Bruna (2018)

- Consider one layer network with no bias, L^2 loss, and hidden layer of width p

$$\Phi(x,\theta) = U\rho(Wx), \qquad L(\theta) := \mathbb{E}\left[\|\Phi(X,\theta) - Y\|^2\right].$$

- Consider span of possible features

$$V := \mathrm{Span}\{\psi_w, \quad w \in \mathbb{R}^n\}, \qquad \psi_w := \rho(\langle w, \cdot \rangle)$$

- **Thm.** If $p \geq \dim(V)$, then there are no bad local minima. If $p \geq 2\dim(V)$, then all local=global minima are connected.

Finite Intrinsic Dimension: Venturi-Bandeira-Bruna (2018)

- Consider one layer network with no bias, L^2 loss, and hidden layer of width p

$$\Phi(x,\theta) = U\rho(Wx), \qquad L(\theta) := \mathbb{E}\left[\|\Phi(X,\theta) - Y\|^2\right].$$

- Consider span of possible features

$$V := \mathrm{Span}\{\psi_w, \quad w \in \mathbb{R}^n\}, \qquad \psi_w := \rho(\langle w, \cdot\rangle)$$

- **Thm.** If $p \geq \dim(V)$, then there are no bad local minima. If $p \geq 2\dim(V)$, then all local=global minima are connected.
- **Ex.** $\rho =$ polynomial, linear

Setup for Proof in Venturi-Bandeira-Bruna (2018)

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):
  - Choose basis $\{\psi_{w_j}\}$ and define $\langle \cdot, \cdot \rangle$ with $\{\psi_{w_j}\}_j = \text{ONB}$

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):
  - Choose basis $\{\psi_{w_j}\}$ and define $\langle \cdot, \cdot \rangle$ with $\{\psi_{w_j}\}_j = \text{ONB}$
  - Define reproducing kernel $K_V(x, y) := \sum_j \psi_{w_j}(x)\psi_{w_j}(y)$.

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):
  - Choose basis $\{\psi_{w_j}\}$ and define $\langle \cdot, \cdot \rangle$ with $\{\psi_{w_j}\}_j = \text{ONB}$
  - Define reproducing kernel $K_V(x, y) := \sum_j \psi_{w_j}(x)\psi_{w_j}(y)$.
  - $\phi(x) := K_V(x, \cdot) \in V$ is kernel of evaluation at x

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):
  - Choose basis $\{\psi_{w_j}\}$ and define $\langle \cdot, \cdot \rangle$ with $\{\psi_{w_j}\}_j = \text{ONB}$
  - Define reproducing kernel $K_V(x, y) := \sum_j \psi_{w_j}(x)\psi_{w_j}(y)$.
  - $\phi(x) := K_V(x, \cdot) \in V$ is kernel of evaluation at x
  - Then $\psi_w(x) = \langle \psi_w, \phi(x) \rangle$

Setup for Proof in Venturi-Bandeira-Bruna (2018)

- Output $\Phi(x, \theta) \in V$.
- V is reproducing kernel Hilbert space (RKHS):
  - Choose basis $\{\psi_{w_j}\}$ and define $\langle \cdot, \cdot \rangle$ with $\{\psi_{w_j}\}_j = \text{ONB}$
  - Define reproducing kernel $K_V(x, y) := \sum_j \psi_{w_j}(x)\psi_{w_j}(y)$.
  - $\phi(x) := K_V(x, \cdot) \in V$ is kernel of evaluation at x
  - Then $\psi_w(x) = \langle \psi_w, \phi(x) \rangle$
- Obtain $\Phi = \langle U\psi(W), \phi(x) \rangle$