Dynamical aspects of learning linear neural networks

Second Symposium on Machine Learning and Dynamical Systems

Zhenyu Liao

Joint work with Y. Chitour and R. Couillet

Department of Statistics, University of California, Berkeley.

Outline

Objective: given (X, Y), find function $f_{\mathbf{W}}(\cdot)$ parameterized by $\mathbf{W}$ to minimize $\|Y - f_{\mathbf{W}}(X)\|^2$.

- First solution: if $f_{\mathbf{W}}(X) = WX$, linear regression $W_{LR} = YX^{\mathsf{T}}(XX^{\mathsf{T}})^{-1}$ if XX^{T} invertible. However, may not work well for difficult problems (e.g., cat & dogs classif, face recognition, etc): describe solely a linear transformation between X and Y.
- (Brain-inspired) linear neural network models (back to [Rosenblatt, 1958])

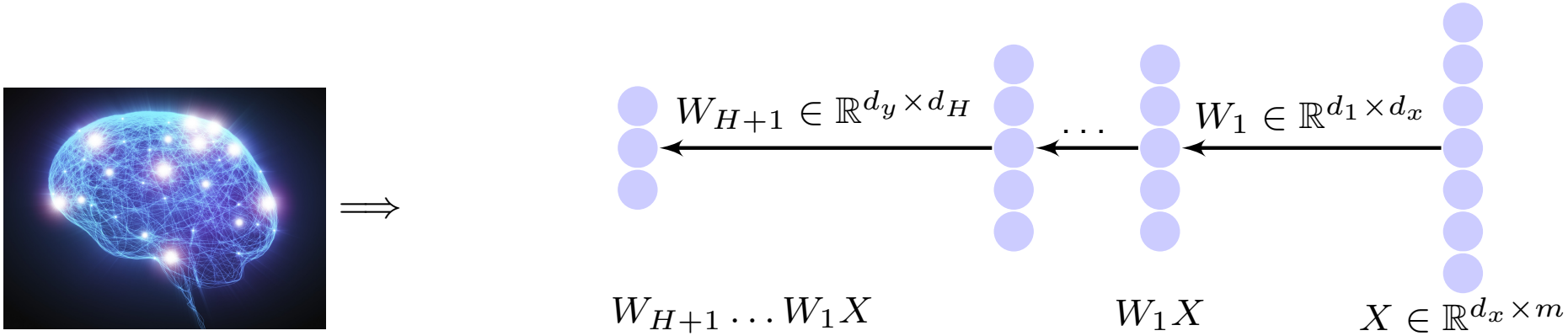


Figure: Illustration of H-hidden-layer linear neural network

Linear neural networks : $\boxed{f_{\mathbf{W}}(X) = W_{H+1} W_H W_1 X.}$

with $\mathbf{W} = (W_{H+1}, W_H, \cdots, W_1)$, equivalent to linear regression if $W_{H+1} W_H W_1 = W_{LR}$.

- Nonlinear neural networks:

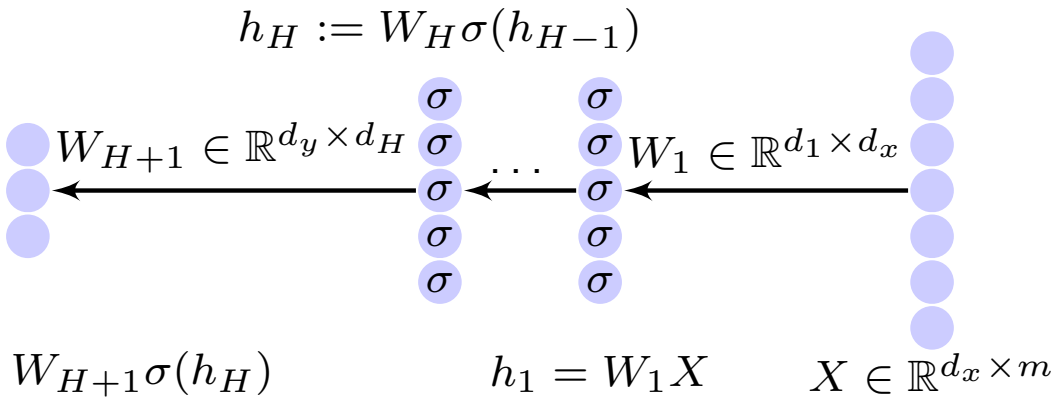


Figure: Illustration of H-hidden-layer nonlinear neural network

with nonlinear *activation function* $\sigma(z)$: ReLU $\max(z, 0)$, Leaky ReLU $\max(z, az)$ or sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$, $\arctan(z)$, $\tanh(z)$, etc.

$$\boxed{f_{\mathbf{W}}(X) = W_{H+1}\sigma(W_H\sigma(W_{H-1}\sigma(\cdots W_1 X))).}$$

Why "Deep" Neural Networks?

Practitioners find "deeper" structures bring better performance, e.g., for (simple) handwritten digits classification:

Figure: Samples from the MNIST dataset [LeCun et al. 1998].

Network	Classification error rate
$H = 0$ (linear regression)	12.0%
$H = 2$ [LeCun et al. 1998]	2.5%
$H = 4$ [LeCun et al. 1998]	0.8%

Table: Evolution of state of the art on MNIST dataset.

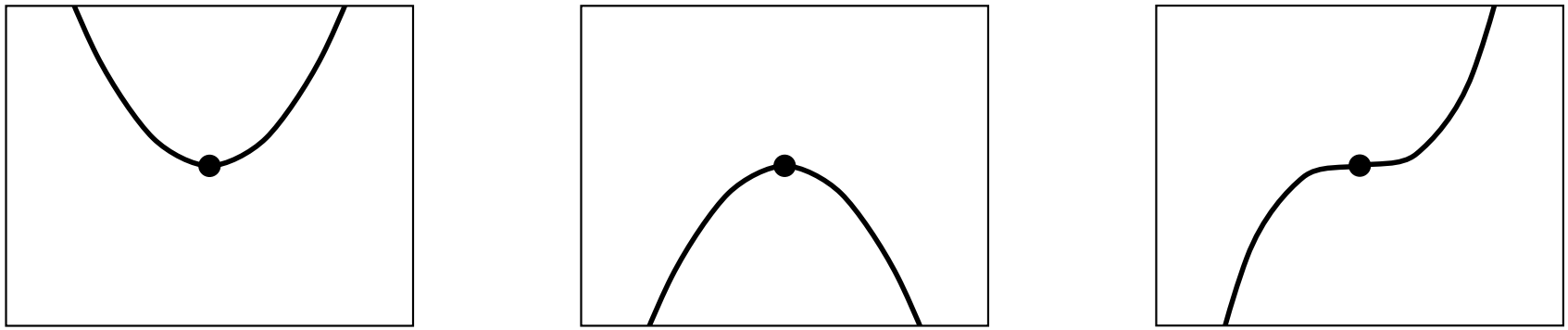
However, deep networks are computationally more challenging!

Challenges in Training Deep Neural Networks

1. huge demand of computational resources: LeNet in [LeCun et al. 1998] 5-layer with $60K$ parameters to ResNet in [He et al. 2015] 152-layer with $60M$ parameters.
2. Commonly trained with first-order optimization methods due to complexity constraints, e.g., (stochastic) gradient decent.
3. unfortunately non-convex optimization problem: e.g., in a single-hidden-layer linear network, use gradient descent to find (W_1, W_2) that minimizes

$$L(W_1, W_2) = \|Y - W_2 W_1 X\|_F^2.$$

   Since $L(W_1^*, W_2^*) = L(\alpha W_1^*, \frac{1}{\alpha} W_2^*) \Rightarrow (\alpha W_1^*, \frac{1}{\alpha} W_2^*)$ is as "good" as (W_1^*, W_2^*)!
4. can be local minima/maxima and saddle points! All depends on (X, Y) and network design.



(a) Local minimum (b) Local maximum (c) Saddle point

Figure: Illustration of three types of critical points in one dimension.

Non-convexity in Deep Neural Networks

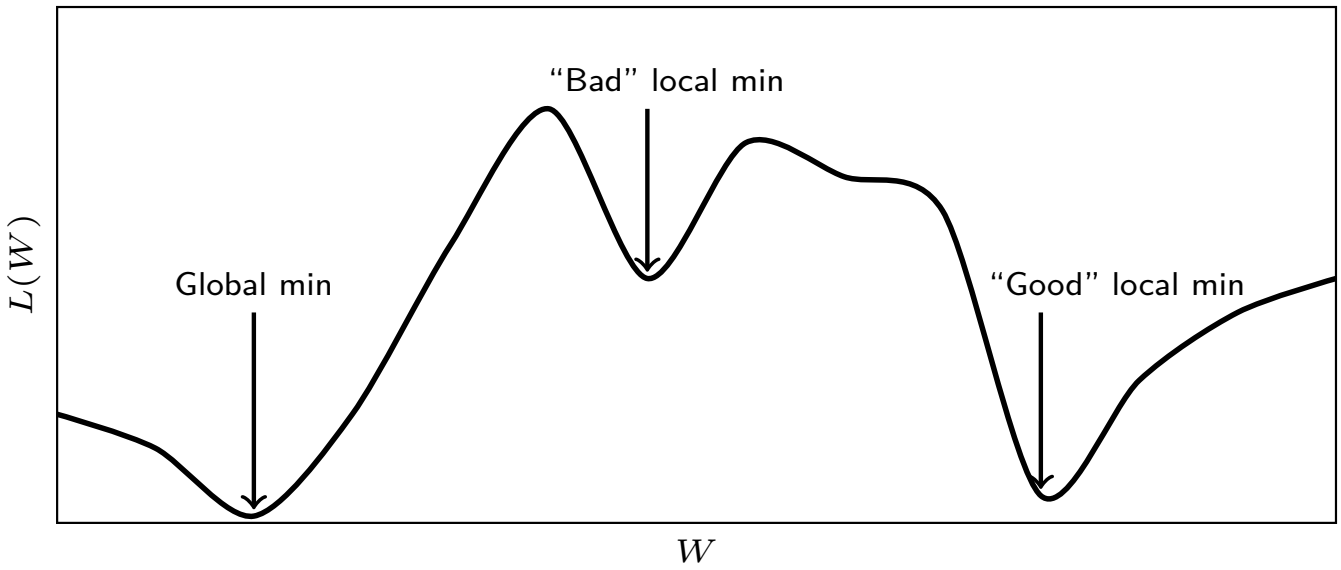


Figure: Examples of different local minima.

In non-convex case, performance of gradient descent can be sensitive to initialization!

Can we still obtain some general results in this difficult case?

On Linear Deep Neural Networks

Set $d_{H+1} := d_y$, $d_0 := d_x$ and consider

$$X \in \mathbb{R}^{d_0 \times m} \quad Y \in \mathbb{R}^{d_{H+1} \times m}.$$

Goal: find $\mathbf{W} = (W_{H+1}, \cdots, W_1)$ that minimizes (depending on (X, Y))

$$L(\mathbf{W}) := \|Y - W_{H+1} W_H \cdots W_1 X\|^2,$$

where

$$W_j \in \mathbb{R}^{d_j \times d_{j-1}}, \quad 1 \leq j \leq H+1.$$

Define state space $\mathcal{W}$ (recall $d_y = d_{H+1}$ and $d_0 = d_x$)

$$\mathcal{W} = \mathbb{R}^{d_{H+1} \times d_H} \times \cdots \mathbb{R}^{d_1 \times d_0}.$$

and **Gradient Descent Dynamic (GDD) associated with** L

$$(GDD)_{(X,Y)} \qquad \frac{d\mathbf{W}}{dt} = -\nabla L(\mathbf{W}), \quad \mathbf{W} \in \mathcal{W}.$$

Conjecture (Global Convergence to Global Minimum)

*For almost every (X, Y) and almost every $\mathbf{W}_0 \in \mathcal{W}$, trajectory of $(GDD)_{(X,Y)}$ starting at $\mathbf{W}_0$ converges to a **Global** minimum of L.*

Gradient Descent for Linear Neural Networks - First reductions

(Usual) working assumptions

$$X, Y \text{ full rank }, m \geq \max(d_i) \geq \min(d_i) = d_y.$$

Up to SVD and computations, can assume

$$X = I_{d_x} \text{ (i.e. } m = d_x), \quad Y = \begin{pmatrix} D_Y & 0 \end{pmatrix}, \quad D_Y \in \mathbb{R}^{d_y \times d_y} \text{ diagonal } > 0.$$

Notation

$$(\Pi W)_i^j = W_j \cdots W_i, \quad 1 \leq i \leq j \leq H+1, \quad M = Y - (\Pi W)_1^{H+1}.$$

Gradient Descent Dynamic, $1 \leq j \leq H+1$

$$(GDD) \qquad \frac{dW_j}{dt} = (\Pi W)_{j+1}^{H+1} M (\Pi W)_1^{j-1}.$$

Definition Critical points $\nabla L(\mathbf{W}) = 0$

$$\text{Crit}(L) = \{\mathbf{W} = (W_{H+1}, \cdots, W_1) \in \mathcal{W}, \ \ (\Pi W)_{j+1}^{H+1} M (\Pi W)_1^{j-1} = 0\}.$$

Candidates for limit points of GDD trajectories.

Gradient Descent for Linear Neural Networks - Convergence

Theorem (Chitour, Liao, Couillet '18)

Every trajectory of (GDD) converges to an element of $\mathrm{Crit}(L)$.

PROOF: (Obvious but) Key remark: (GDD) analytic $\Longrightarrow$ Lojasiewicz's theorem can be used.

Proposition (Lojasiewicz 50s')

Every bounded trajectory of analytic gradient system converges to a critical point.

Proof reduces to show that trajectories are bounded.

Proposition (Invariants)

For $1 \leq j \leq H$, following quantities are conserved along trajectory of (GDD)

$$W_{j+1}(t)^{\mathsf{T}} W_{j+1}(t) - W_j(t) W_j(t)^{\mathsf{T}} = (W_{j+1}^{\mathsf{T}} W_{j+1} - W_j W_j^{\mathsf{T}})\Big|_{t=0}.$$

$\Longrightarrow \|W_j(t)\|_F^2 = \|W_{H+1}(t)\|_F^2 + C_j \quad t \geq 0,\ 1 \leq j \leq H.$

Set $g(t) = \|W_{H+1}\|_F^2$. Given a trajectory of (GDD), one proves that there exists $C_0, C_1 > 0$

$$\frac{dg}{dt} \leq -C_0 g^{H+1}(t) + C_1 \Big(1 + g^H(t)\Big), \quad \forall t \geq 0.$$

Consequence of Key Invariant

Proposition (Exponential Convergence to Global Minimum)

Assume that $d_1 \geq d_2 \geq \ldots \geq d_H$ *and*

$$C_j := \left[W_{j+1}^{\mathsf{T}} W_{j+1} - W_j W_j^{\mathsf{T}}\right]_{t=0} \in \mathbb{R}^{d_j \times d_j}$$

has at least d_{j+1} *positive eigenvalues, then every trajectory of (GDD) converges to a global minimum at least at the rate of* $\exp(-2\alpha t)$ *with* $\alpha > 0$ *the* d_{j+1}*-smallest eigenvalue of* C_j*.*

Establish that

$$\frac{dL}{dt} \leq -2 \sum_{j=1}^{H+1} \prod_{k=1}^{j-1} \lambda_{\min}(W_k^{\mathsf{T}} W_k) \prod_{l=j+1}^{H+1} \lambda_{\min}(W_l W_l^{\mathsf{T}}) \cdot L$$

- find any $1 \leq j \leq H+1$ such that $\prod_{k=1}^{j-1} \lambda_{\min}(W_k^{\mathsf{T}} W_k) \prod_{l=j+1}^{H+1} \lambda_{\min}(W_l W_l^{\mathsf{T}}) \geq \alpha > 0$,
- for $j = 1$, becomes $\prod_{l=2}^{H+1} \lambda_{\min}(W_l W_l^{\mathsf{T}})$, closely connected to $\lambda(W_l^{\mathsf{T}} W_l)$ when $d_1 \geq d_2 \geq \ldots \geq d_H$ and can be controlled if C_{l-1} has $\uparrow$ positive eigenvalues.

Gradient Descent for Linear Neural Networks - Study of $\mathrm{Crit}(L)$

Definition

For $\mathbf{W} \in \mathrm{Crit}(L)$ define

$$R(\mathbf{W}) = (\Pi W)_2^{H+1}, \quad r_R(\mathbf{W}) = \operatorname{rank} R(\mathbf{W}) \in [0, d_y],$$

$$Z(\mathbf{W}) = (\Pi W)_2^{H} \quad r_Z(\mathbf{W}) = \operatorname{rank} Z(\mathbf{W}) \geq R(\mathbf{W}).$$

Then

$$\mathrm{Crit}(L) = \cup_{r=0}^{d_y} \mathrm{Crit}_r(L), \quad \mathrm{Crit}_r(L) = \{\mathbf{W} \in \mathrm{Crit}(L),\ r_R(\mathbf{W}) = r\}.$$

$CrV(L) =$ Set of critical values of $L = \{L(\mathbf{W}), \mathbf{W} \in \mathrm{Crit}(L)\}$.

Proposition (Landscape of Deep Linear Networks)

Assume Y has **distinct singular values** $S_Y = \{s_1, \cdots, s_{d_Y}\}$.

i) $CrV(L) = \{\frac{1}{2}\sum_{s\in I} s^2 \mid I \subset S_Y\}$, *finite.*

ii) $\mathrm{Crit}_{d_y}(L) =$ *set of local (and global) minima with* $L = 0$.

iii) For $0 \leq r \leq d_y - 1$, $\mathrm{Crit}_r(L)$ *algebraic variety of dimension* > 0 *made of saddle points. If* $r_Z > r \geq 0$, $Hessian(L)(\mathbf{W})$ *has at least one negative eigenvalue.*

Item $i)$ and $ii)$ as in previous efforts, e.g., [Kawaguchi '16],

Gradient Descent for Linear Neural Networks - Case $H = 1$

Reformulation of Conjecture (Global Convergence to Global Minimum):

Conjecture (New formulation of Conjecture)

For almost every (X, Y), the union of the basins of attraction of saddles points is of measure zero.

Proposition (Chitour, Liao, Couillet '18)

Conjecture is true for $H = 1$, i.e., in the case of single-hidden-layer linear network.

Argument relies on concept of **Normal Hyperbolicity** (due to Fenichel 1972).

Illustration of Hyperbolic Equilibrium Point in 3D

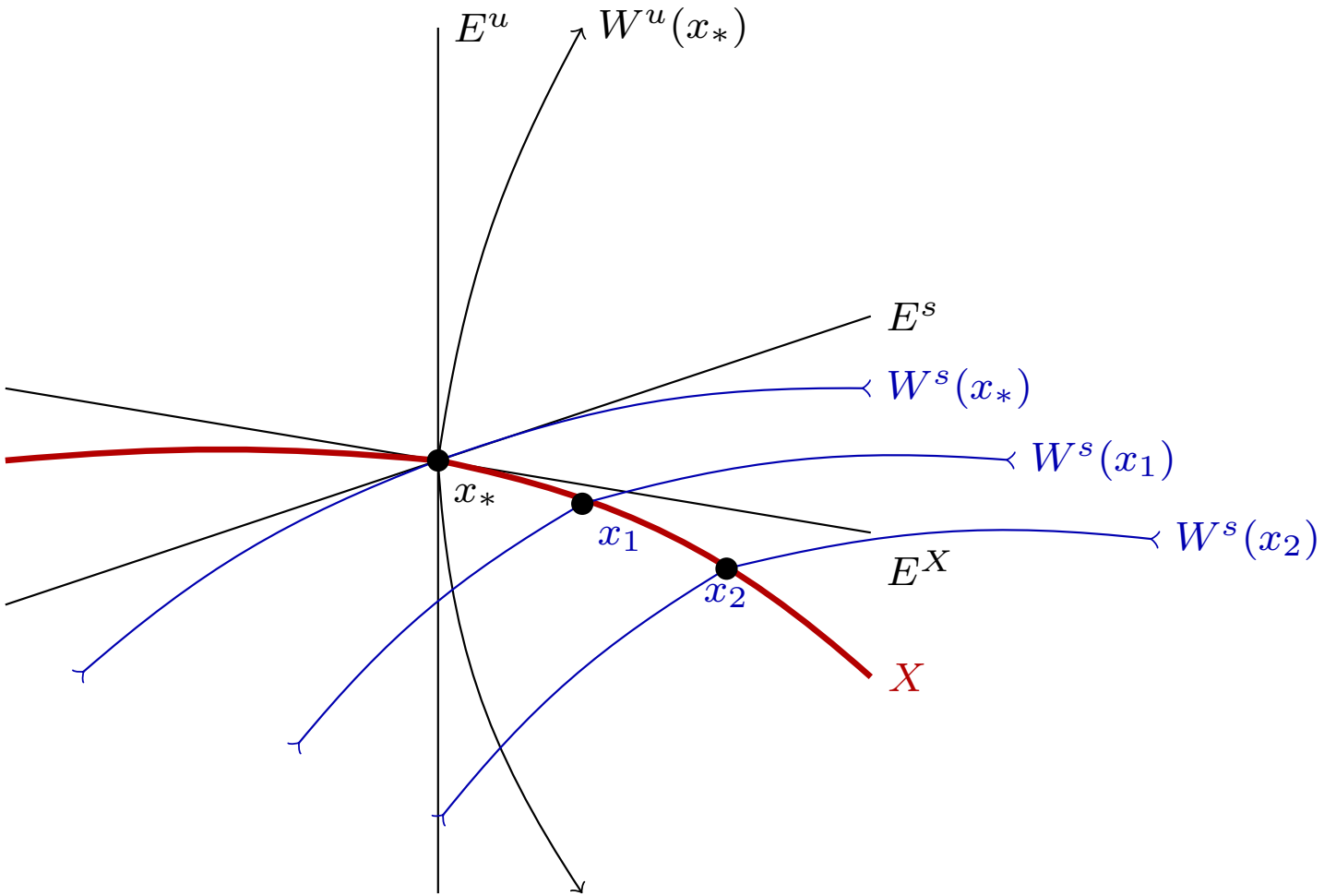


Figure: Illustration of Hyperbolic Equilibrium Point in 3D

Conclusion

Take-away message:

- invariant structure in Gradient Descent Dynamics of linear and nonlinear neural networks:
  - establish convergence, with up to *exponential rate* with properly chosen initialization
  - characterize specific *geometric property* of Gradient Descent Dynamics in neural networks
- linear networks: reduce to global analysis of union of basins of attraction for all saddle points
- continuous-time analysis without specific assumption on the gradient or the Hessian

Reference:

Y. Chitour, Z. Liao, and R. Couillet, A geometric approach of gradient descent algorithms in neural networks. arXiv preprint arXiv:1811.03568.

Thank you!