Embedding Principle of Loss Landscape of Deep Neural Networks

张耀宇

Institute of Natural Sciences&School of Mathematical Sciences
Shanghai Jiao Tong University

机器学习联合研讨计划, c2sml.cn

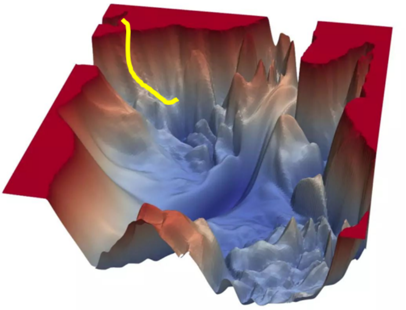
Outline

Outline

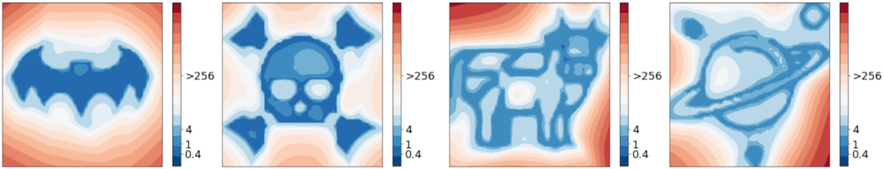
Loss landscape

$$R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{y}_i)$$

Model: $\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta})$
Data: $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$
Loss: $\ell(\cdot, \cdot)$



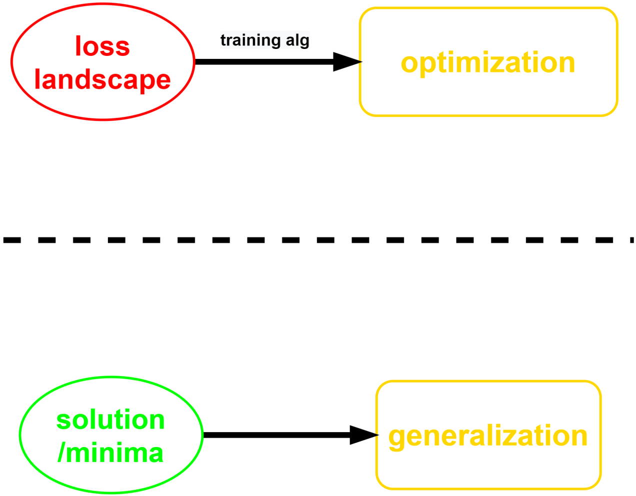
DNN loss landscape is complex



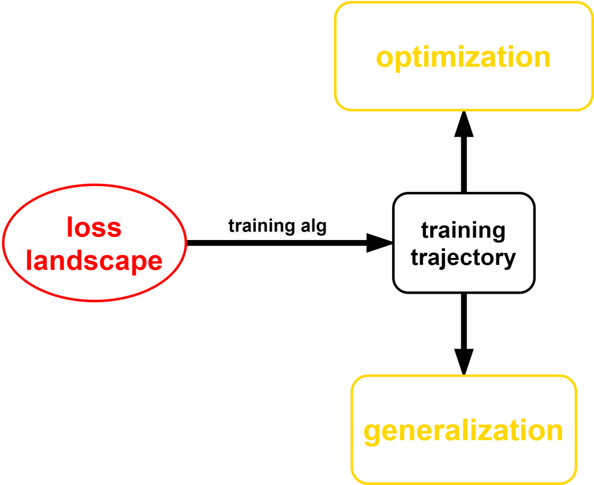
(a) Loss surface on FashionMNIST dataset (b) Loss surface on CIFAR10 dataset

I. Skorokhodov, M. Burtsev, 2019

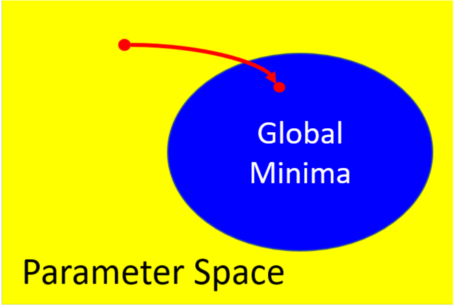
Role of loss landscape in conventional ML



Role of loss landscape in deep learning



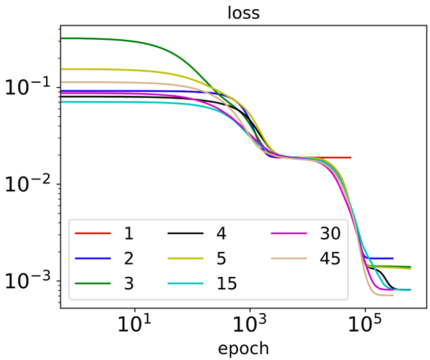
Global minima degenerate for overparameterized NN

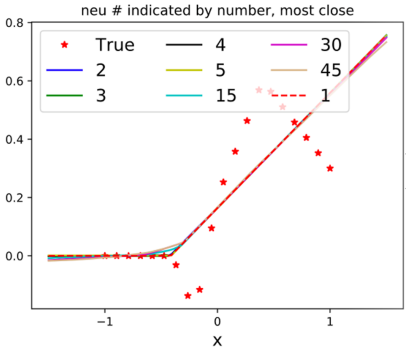


Yaim Cooper, 2018:
Global minima is usually a $M - n$ dimensional submanifold of $\mathbb{R}^M$, M is number of model parameters, n is number of data points.

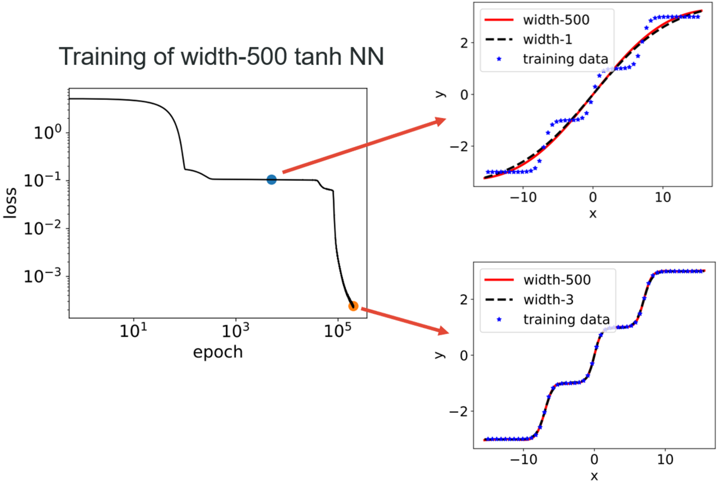
Question1: (degeneracy) properties of other critical points?

Intriguing experimental phenomenon in width

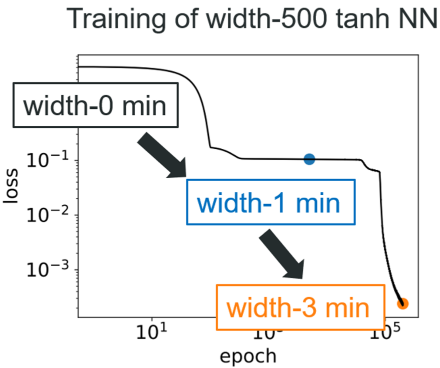




Training process of wide NN beyond NTK



Implicit regularization towards "simple" critical points



Question2: Are there "simple" critical points in wide NN?

"simple": output function can be realized by a narrow NN

Outline

Embedding Principle

Embedding Principle

The loss landscape of any network "contains" all critical points of all narrower networks.

$R_S(\theta_{\rm wide})$ "contain" $\theta^{\rm c}_{\rm narr}$: $\exists \theta^{\rm c}_{\rm wide}$, s.t. $\boldsymbol{f}_{\theta^{\rm c}_{\rm narr}} = \boldsymbol{f}_{\theta^{\rm c}_{\rm wide}}$.

Reference

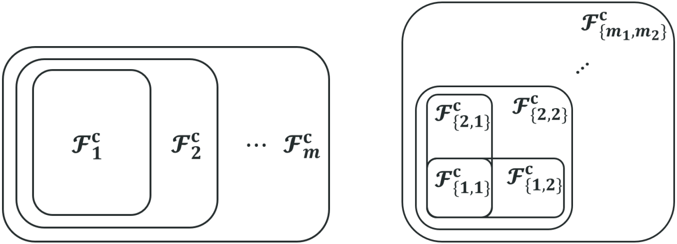
Yaoyu Zhang*, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu*, Embedding Principle of Loss Landscape of Deep Neural Networks. arXiv:2105.14573, 2021. (accepted by NeurIPS 2021 as spotlight)

Answer to Question2

Question2: Are there "simple" critical points in wide NN?
Answer: Yes!

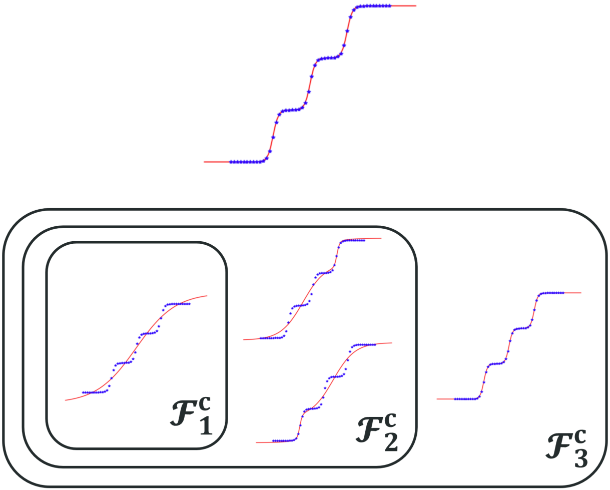
Embedding Principle[1]

The loss landscape of any network "contains" all critical points of all narrower networks.
$R_S(\theta_{\text{wide}})$ "contain" $\theta^{\text{c}}_{\text{narr}}$: $\exists \theta^{\text{c}}_{\text{wide}}$, s.t. $f_{\theta^{\text{c}}_{\text{narr}}} = f_{\theta^{\text{c}}_{\text{wide}}}$.



[1] **Zhang***, Zhang, Luo, Xu*, Embedding Principle of Loss Landscape of Deep Neural Networks. arXiv:2105.14573, 2021.

Example：critical points of width-3 tanh NN



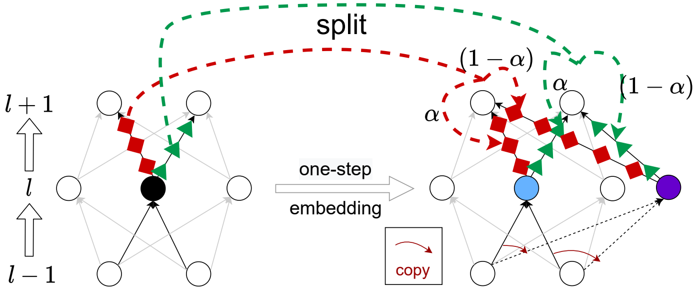
Key to our proof of embedding principle

Key: discovering critical embedding

Discover embedding $\mathcal{T}:\mathbb{R}^{M_{\text{narr}}}\to\mathbb{R}^{M_{\text{wide}}}$ such that for $\theta_{\text{wide}}=\mathcal{T}(\theta_{\text{narr}})$
(i) **output preserving:** $\boldsymbol{f}_{\theta_{\text{narr}}}=\boldsymbol{f}_{\theta_{\text{wide}}}$;
(ii) **criticality preserving:** if θ_{narr} is a critical point, then θ_{wide} is also a critical point.

critical embedding exists $\Rightarrow$ Embedding Principle

One-step embedding



One-step embedding $\mathcal{T}_{l,s}^{\alpha}$.

One-step embedding is critical embedding

Proposition (output and representation preserving)

For any point θ_{narr} of a DNN, a point θ_{wide} of a wider DNN obtained from θ_{narr} by **one-step embedding** satisfies

$$\boldsymbol{f}_{\theta_{\text{narr}}}(\boldsymbol{x}) = \boldsymbol{f}_{\theta_{\text{wide}}}(\boldsymbol{x}) \text{ for any } \boldsymbol{x}.$$

定理 (criticality preserving)

*For any critical point θ_{narr} of a DNN, a point θ_{wide} of a wider DNN obtained from θ_{narr} by **one-step embedding** is a critical point.*

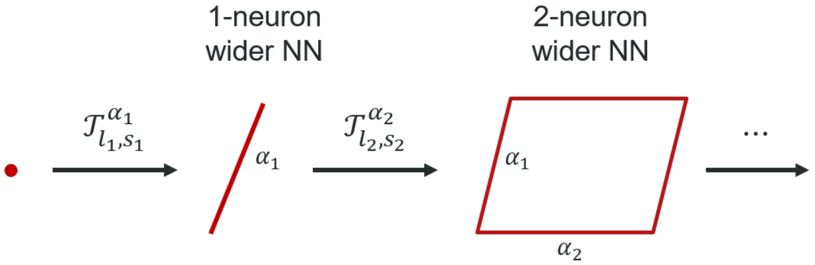
Remark: Obviously, **multi-step embedding**, i.e., composition of **one-step embedding**, is also critical embedding.

Answer to Question1

Question1: (degeneracy) properties of other critical points?

定理 (informal)

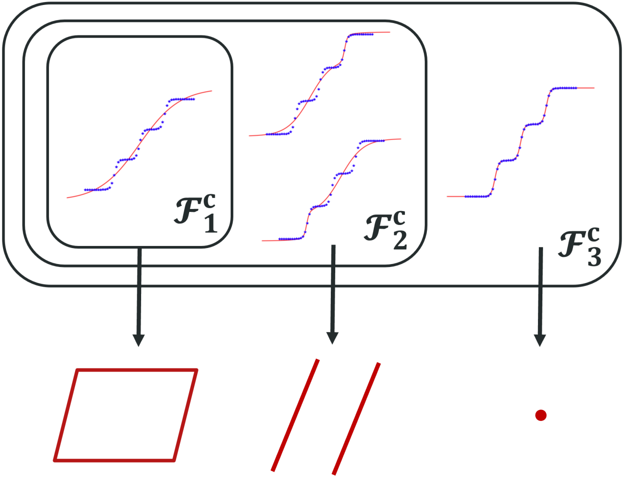
(Under mild assumption) Any critical point θ^c of a DNN can be embedded to K-dimensional critical affine subspaces of a K-neuron wider DNN.



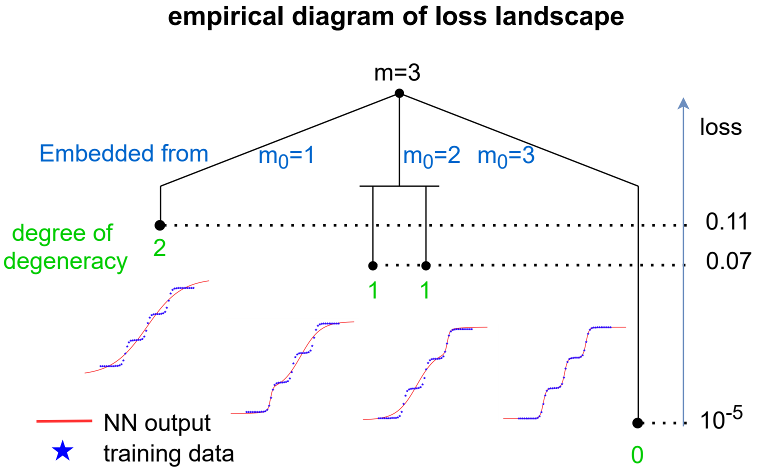
Critical points/submanifolds of width-3 tanh NN

functions of
critical points

critical points/
submanifolds



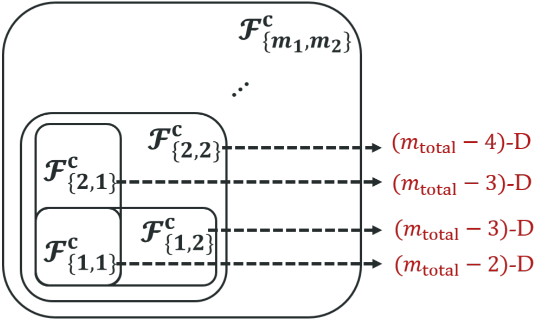
Numerical verification



Insight to degeneracy supplement to Yaim's work

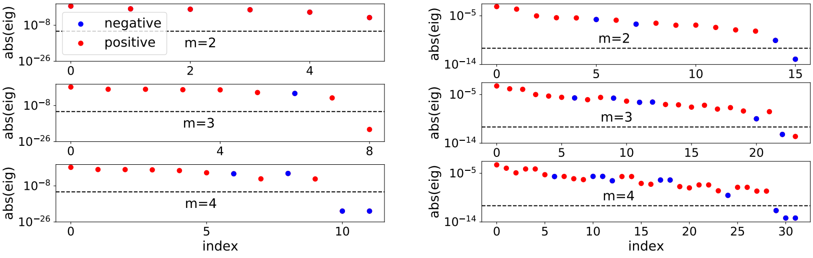
Overparameterization induced degeneracy [Yaim Cooper 2018]
Global minima is $(M - n)$-D, M: #parameters, n: #data.

Embedding (neuron redundancy) induced degeneracy [our work]
Global minima is at least $(m_{\mathrm{total}} - m_{\mathrm{min}})$-D,
m_{total}: #neurons, m_{min}: minimum #neurons for interpolation.



Outline

Potential relevance to optimization



(a) synthetic data

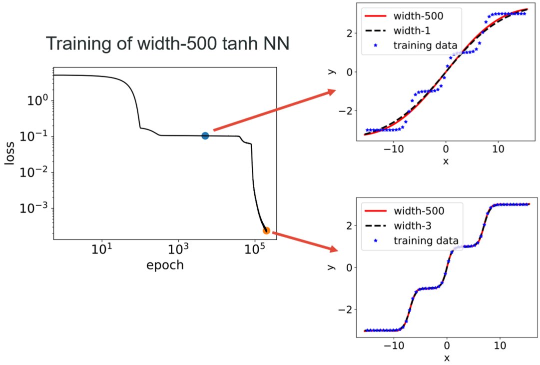
(b) Iris data

Eigenvalues of Hessian of critical points. BLue: Negative, Red: Positive.

Hint

A local min of narrow NN may become strict saddle points in wider NNs.

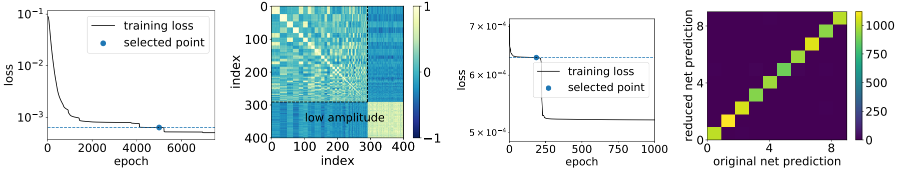
Potential relevance to generalization



Hint

A NN may be guided by "simple" critical points towards a "simple" global min.

Potential relavance to pruning



Width-400 ReLU NN pruned to width-58 NN near a critical point for MNIST.

Hint

NN training may experience "simple" critical points with great pruning potential.

Outline

Embedding Principle sheds light to DNN loss landscape

Embedding Principle

The loss landscape of any network “contains” all critical points of all narrower networks.

Understanding

- Prevalence of "simple" critical points/affine subspaces
- Degeneracy $\geq$ neuron redundancy

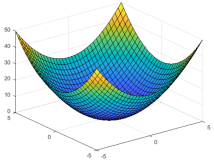
A new perspective

- Different width DNN loss landscapes as a unified object.

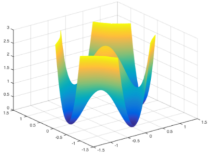
A new tool

- Critical embedding is an important tool for the analysis of width effect regarding both optimization and generalization.

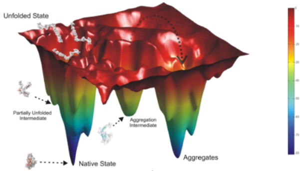
Discussion: mysterious easy optimization of DNN



Mathworks



Ma, 2021



Quintas, 2013

?

convex

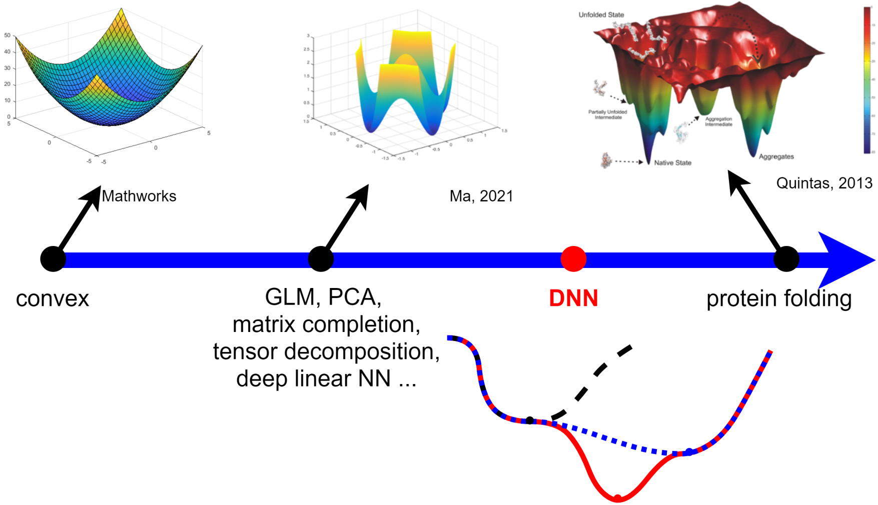
GLM, PCA,
matrix completion,
tensor decomposition,
deep linear NN ...

DNN

protein folding

optimization difficulty

Discussion: mysterious easy optimization of DNN



Conjecture for the easy optimization mystery

Conjecture

Bad local min/critical point may be common, but truly bad one is rare.

定义 (Truly bad critical point)

Given any critical point $\boldsymbol{\theta}$, if $\mathcal{T}\boldsymbol{\theta}$ is not a strict saddle for any critical embedding $\mathcal{T}$, then it is a truly bad critical point.

Our lines of research

- **Frequency Principle (training dynamics&implicit bias)**
  - **Experiment:** (1) Xu, Zhang, Xiao, ICONIP 2019, (2) Xu, Zhang, Luo, Xiao, Ma, CiCP, 2020, (3) Xu, Zhou, AAAI 2020.
  - **Theory:** (4) Zhang, Luo, Ma, Xu, CPL, 2021, (5) Luo, Ma, Xu, Zhang, CSIAM, 2021, (6) Luo, Ma, Xu, Zhang, arXiv, 2020, (7) Luo, Ma, Wang, Xu, Zhang, arXiv, 2020.
- **Phase diagram (training dynamics&regime)**
  - (8) Zhang, Xu, Luo, Ma, MSML 2020, (9) Luo, Xu, Ma, Zhang, JMLR, 2021, (10) Xu, Zhou, Luo, Zhang, arXiv, 2021.
- **Embedding Principle (loss landscape)**
  - (11) Zhang, Zhang, Tao Luo, Xu, arXiv:2105.14573, 2021 (accepted by NeurIPS 2021 as spotlight)

合作者（上海交通大学）

同事：许志钦(Zhi-Qin John Xu)，罗涛(Tao Luo)，马征(Zheng Ma)
学生：张众望(Zhongwang Zhang)

Outline

Settings and assumptions

- **Data**
  $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i = \boldsymbol{f}^*(\boldsymbol{x}_i))\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{y}_i \in \mathbb{R}^{d'}$.
- **Neural network**
  a L-layer ($L \geq 2$) fully-connected DNN $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$.
  $\boldsymbol{\theta} = \left(\boldsymbol{\theta}|_1, \cdots, \boldsymbol{\theta}|_L\right) = \left(\boldsymbol{W}^{[1]}, \boldsymbol{b}^{[1]}, \ldots, \boldsymbol{W}^{[L]}, \boldsymbol{b}^{[L]}\right) \in \mathrm{Tuple}_{\{m_l\}}$,
  $\boldsymbol{f}_{\boldsymbol{\theta}}^{[0]}(\boldsymbol{x}) = \boldsymbol{x}$,
  $\boldsymbol{f}_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x}) = \sigma(\boldsymbol{W}^{[l]}, \boldsymbol{f}_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l]})$, $l \in [L-1]$,
  $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{\theta}}^{[L]}(\boldsymbol{x}) = \boldsymbol{W}^{[L]} \boldsymbol{f}_{\boldsymbol{\theta}}^{[L-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[L]}$.
- **Activation function**
  $\sigma(\cdot)$ is a (weakly) differentiable function.
- **Loss function**
  $R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{f}^*(\boldsymbol{x}_i)) = \mathbb{E}_S \ell(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{f}^*(\boldsymbol{x}))$.
  $\ell(\cdot, \cdot)$ is a (weakly) differentiable function.

Definitions

定义 (critical point)

Parameter vector $\boldsymbol{\theta}$ is a critical point of the landscape of R_S if $\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) = \mathbf{0}$.

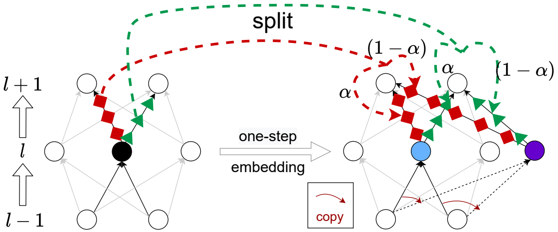
定义 (critical submanifold/affine subspace)

A critical submanifold or affine subspace $\mathcal{M}$ is a connected submanifold or affine subspace of the parameter space $\mathbb{R}^M$, such that each $\boldsymbol{\theta} \in \mathcal{M}$ is a critical point of loss with the same loss value.

定义 (degree of degeneracy)

The degree of degeneracy of point $\boldsymbol{\theta}$ in the landscape of R_S is the corank of Hessian matrix $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} R_S$, i.e., number of the zero eigenvalues.

One-step embedding



One-step embedding: $\mathcal{T}^{\alpha}_{l,s}(\boldsymbol{\theta}) = (\mathcal{T}_{l,s} + \alpha\mathcal{V}_{l,s})(\boldsymbol{\theta})$.

$$\mathcal{T}_{l,s}(\boldsymbol{\theta})|_k = \boldsymbol{\theta}|_k \text{ for } k \neq l, l+1, \quad \mathcal{T}_{l,s}(\boldsymbol{\theta})|_l = \left(\begin{bmatrix} \boldsymbol{W}^{[l]} \\ \boldsymbol{W}^{[l]}_{s,[1:m_{l-1}]} \end{bmatrix}, \begin{bmatrix} \boldsymbol{b}^{[l]} \\ \boldsymbol{b}^{[l]}_s \end{bmatrix}\right),$$

$$\mathcal{T}_{l,s}(\boldsymbol{\theta})|_{l+1} = \left(\left[\boldsymbol{W}^{[l+1]}, \boldsymbol{0}\right], \boldsymbol{b}^{[l+1]}\right);$$

$$\mathcal{V}_{l,s}(\boldsymbol{\theta})|_k = (\boldsymbol{0}, \boldsymbol{0}) \text{ for } k \neq l, l+1, \quad \mathcal{V}_{l,s}(\boldsymbol{\theta})|_l = (\boldsymbol{0}, \boldsymbol{0}),$$

$$\mathcal{V}_{l,s}(\boldsymbol{\theta})|_{l+1} = \left(\left[\boldsymbol{0}, -\boldsymbol{W}^{[l+1]}_{[1:m_{l+1}],s}, \boldsymbol{0}, \boldsymbol{W}^{[l+1]}_{[1:m_{l+1}],s}\right], \boldsymbol{0}\right).$$

Illustration

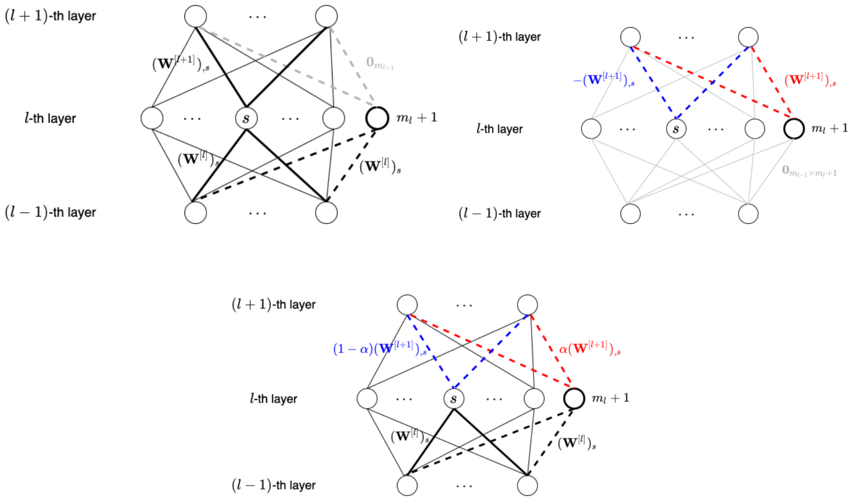


图: Illustration of $\mathcal{T}_{l,s}$, $\mathcal{V}_{l,s}$, and $\mathcal{T}_{l,s}^{\alpha}$.

Properties of one-step embedding

Remark on $\mathcal{T}^{\alpha}_{l,s}$

$\mathcal{T}^{\alpha}_{l,s}$: $\{\mathrm{Tuple}_{\{m_0,\cdots,m_L\}} | L > l, m_l \geq s\} \to \{\mathrm{Tuple}_{\{m_0,\cdots,m_l+1,\cdots,m_L\}} | L > l, m_l \geq s+1\}$ is a linear injective operator.

Proposition (**output and representation preserving**)

For any point $\boldsymbol{\theta}_{\mathrm{narr}}$ of a DNN, a point $\boldsymbol{\theta}_{\mathrm{wide}}$ of a wider DNN obtained from $\boldsymbol{\theta}_{\mathrm{narr}}$ by one-step embedding satisfies
(i) $\boldsymbol{f}_{\boldsymbol{\theta}_{\mathrm{narr}}}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{\theta}_{\mathrm{wide}}}(\boldsymbol{x})$ for any $\boldsymbol{x}$;
(ii) representation of the wide DNN at $\boldsymbol{\theta}_{\mathrm{wide}}$, i.e., the set of all different response functions of neurons, is the same as representation of the narrow DNN at $\boldsymbol{\theta}_{\mathrm{narr}}$.

定理 (**criticality preserving**)

For any critical point $\boldsymbol{\theta}_{\mathrm{narr}}$ of a DNN, a point $\boldsymbol{\theta}_{\mathrm{wide}}$ of a wider DNN obtained from $\boldsymbol{\theta}_{\mathrm{narr}}$ by one-step embedding is a critical point.

Sketch of proof

定义

$$\boldsymbol{f}_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x}) = \sigma(\boldsymbol{W}^{[l]}\boldsymbol{f}_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l]}), \quad \boldsymbol{g}_{\boldsymbol{\theta}}^{[l]} = \sigma^{(1)}\left(\boldsymbol{W}^{[l]}\boldsymbol{f}_{\boldsymbol{\theta}}^{[l-1]} + \boldsymbol{b}^{[l]}\right),$$

$$\boldsymbol{z}_{\boldsymbol{\theta}}^{[L]} = \nabla_{\boldsymbol{f}_{\boldsymbol{\theta}}^{[L]}}\ell(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}^*), \quad \boldsymbol{z}_{\boldsymbol{\theta}}^{[l]} = (\boldsymbol{W}^{[l+1]})^\intercal \boldsymbol{z}_{\boldsymbol{\theta}}^{[l+1]} \circ \boldsymbol{g}_{\boldsymbol{\theta}}^{[l+1]}.$$

Gradient of loss with respect to network parameters of each layer can be computed as follows

$$\nabla_{\boldsymbol{W}^{[l']}} R_S(\boldsymbol{\theta}) = \nabla_{\boldsymbol{W}^{[l']}} \mathbb{E}_S \ell(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{f}^*(\boldsymbol{x})) = \mathbb{E}_S\left(\boldsymbol{z}_{\boldsymbol{\theta}}^{[l']} \circ \boldsymbol{g}_{\boldsymbol{\theta}}^{[l']} (\boldsymbol{f}_{\boldsymbol{\theta}}^{[l'-1]})^\intercal\right),$$

$$\nabla_{\boldsymbol{b}^{[l']}} R_S(\boldsymbol{\theta}) = \nabla_{\boldsymbol{b}^{[l]}} \mathbb{E}_S \ell(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{f}^*(\boldsymbol{x})) = \mathbb{E}_S(\boldsymbol{z}_{\boldsymbol{\theta}}^{[l']} \circ \boldsymbol{g}_{\boldsymbol{\theta}}^{[l']}).$$

Effect of one-step embedding

引理

Given a L-layer ($L \geq 2$) fully-connected neural network with width $(m_0, \ldots, m_L)$, for any network parameters $\boldsymbol{\theta} = (\boldsymbol{W}^{[1]}, \boldsymbol{b}^{[1]}, \cdots, \boldsymbol{W}^{[L]}, \boldsymbol{b}^{[L]})$ and for any $l \in [L-1]$, $s \in [m_l]$, we have the expressions for $\boldsymbol{\theta}' := \mathcal{T}^{\alpha}_{l,s}(\boldsymbol{\theta})$

(i) feature vectors in $\boldsymbol{F}_{\theta'}$: $\boldsymbol{f}^{[l']}_{\theta'} = \boldsymbol{f}^{[l']}_{\theta'}$, $l' \neq l$ and $\boldsymbol{f}^{[l]}_{\theta'} = \left[(\boldsymbol{f}^{[l]}_{\theta})^{\intercal}, (\boldsymbol{f}^{[l]}_{\theta})_s\right]^{\intercal}$;

(ii) feature gradients in $\boldsymbol{G}_{\theta'}$: $\boldsymbol{g}^{[l']}_{\theta'} = \boldsymbol{g}^{[l']}_{\theta}$, $l' \neq l$ and $\boldsymbol{g}^{[l]}_{\theta'} = \left[(\boldsymbol{g}^{[l]}_{\theta})^{\intercal}, (\boldsymbol{g}^{[l]}_{\theta})_s\right]^{\intercal}$;

(iii) error vectors in $\boldsymbol{Z}_{\theta'}$: $\boldsymbol{z}^{[l']}_{\theta'} = \boldsymbol{z}^{[l']}_{\theta}$, $l' \neq l$ and $\boldsymbol{z}^{[l]}_{\theta'} = \left[(\boldsymbol{z}^{[l]}_{\theta})^{\intercal}_{[1:s-1]}, (1-\alpha)(\boldsymbol{z}^{[l]}_{\theta})_s, (\boldsymbol{z}^{[l]}_{\theta})^{\intercal}_{[s+1:m_l]}, \alpha(\boldsymbol{z}^{[l]}_{\theta})_s\right]^{\intercal}$.

定理 (criticality preserving)

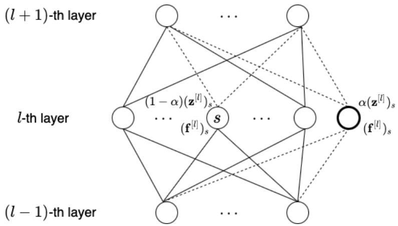
Given a L-layer ($L \geq 2$) fully-connected neural network with width $(m_0, \ldots, m_L)$, for any network parameters $\boldsymbol{\theta} = (\boldsymbol{W}^{[1]}, \boldsymbol{b}^{[1]}, \cdots, \boldsymbol{W}^{[L]}, \boldsymbol{b}^{[L]})$ and for any $l \in [L-1]$, $s \in [m_l]$, $\alpha \in \mathbb{R}$,

$$\text{if } \nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) = \mathbf{0}, \text{ then } \nabla_{\boldsymbol{\theta}} R_S(\mathcal{T}^{\alpha}_{l,s}(\boldsymbol{\theta})) = \mathbf{0}.$$

Illustration:

$$\nabla_{\boldsymbol{W}^{[l']}} R_S(\boldsymbol{\theta}) = \nabla_{\boldsymbol{W}^{[l']}} \mathbb{E}_S \ell(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{f}^*(\boldsymbol{x})) = \mathbb{E}_S \left(\boldsymbol{z}_{\boldsymbol{\theta}}^{[l']} \circ \boldsymbol{g}_{\boldsymbol{\theta}}^{[l']} (\boldsymbol{f}_{\boldsymbol{\theta}}^{[l'-1]})^{\intercal} \right),$$

$$\nabla_{\boldsymbol{b}^{[l']}} R_S(\boldsymbol{\theta}) = \nabla_{\boldsymbol{b}^{[l]}} \mathbb{E}_S \ell(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{f}^*(\boldsymbol{x})) = \mathbb{E}_S (\boldsymbol{z}_{\boldsymbol{\theta}}^{[l']} \circ \boldsymbol{g}_{\boldsymbol{\theta}}^{[l']}).$$



Degeneracy

定义 (degree of degeneracy)

The degree of degeneracy of point θ in the landscape of R_S is the corank of Hessian matrix $\nabla_\theta \nabla_\theta R_S$, i.e., number of the zero eigenvalues.

Remark

For loss and activation with only first-order differentiability, we adopt the following relation to study the degree of degeneracy:
For any critical point θ belonging to a K-dimensional critical submanifold $\mathcal{M}$, its degree of degeneracy is at least K.

引理 (increment of the degree of degeneracy)

Given a L-layer ($L \geq 2$) fully-connected neural network with width $(m_0, \ldots, m_L)$, if there exists $l \in [L-1]$, $s \in [m_l]$, and a d-dimensional manifold $\mathcal{M}$ consisting of critical points of R_S such that for any $\boldsymbol{\theta} \in \mathcal{M}$, $\boldsymbol{W}^{[l+1]}_{[1:m_{l+1}],s} \neq \mathbf{0}$, then

$$\mathcal{M}' := \{\mathcal{T}^{\alpha}_{l,s}(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \mathcal{M}, \alpha \in \mathbb{R}\} \text{ is a } (d+1)\text{-dimensional submanifold}$$

consisting of critical points for the corresponding L-layer fully-connected neural network with width $(m_0, \ldots, m_{l-1}, m_l + 1, m_{l+1}, \ldots, m_L)$.

定理 (informal)

If output weights of neurons in each layer of a DNN at a critical point θ_{narr} are not all zero, then, for any K-neuron wider DNN, θ_{narr} can be embedded to a K-dimensional critical affine subspace.

定理 (formal)

Consider two L-layer ($L \geq 2$) fully-connected neural networks $\mathrm{NN}_A(\{m_l\}_{l=0}^{L})$ and $\mathrm{NN}_B(\{m'_l\}_{l=0}^{L})$ which is K-neuron wider than NN_A. Suppose that the critical point $\boldsymbol{\theta}_A = (\boldsymbol{W}^{[1]}, \boldsymbol{b}^{[1]}, \cdots, \boldsymbol{W}^{[L]}, \boldsymbol{b}^{[L]})$ satisfy $\boldsymbol{W}^{[l]} \neq \boldsymbol{0}$ for each layer $l \in [L]$. Then the parameters $\boldsymbol{\theta}_A$ of NN_A can be critically embedded to a K-dimensional critical affine subspace of loss landscape of NN_B

$$\mathcal{M}_B = \{\boldsymbol{\theta}_B + \textstyle\sum_{i=1}^{K} \alpha_i \boldsymbol{v}_i | \alpha_i \in \mathbb{R}\},$$

where $\boldsymbol{\theta}_B = (\prod_{i=1}^{K} \mathcal{T}_{l_i,s_i})(\boldsymbol{\theta}_A)$ and $\boldsymbol{v}_i = \mathcal{T}_{l_K,s_K} \cdots \mathcal{V}_{l_i,s_i} \cdots \mathcal{T}_{l_1,s_1} \boldsymbol{\theta}_A$.