深層神経回路の数理

Fujitsu Research

ランダム行列・無限次元近似・自由確率論

Hayase 早瀬 Tomohiro 友裕

(~~Fujitsu Laboratories~~)

supported by JST ACT-X

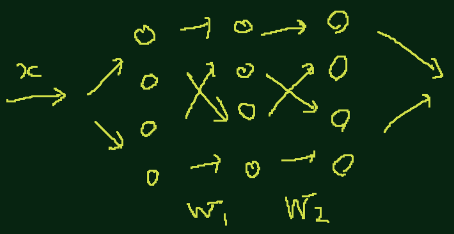
Table of Contents

§1 Overview

Deep Neural Networks

Rossenblatt '57 ~ 62 Multilayer perceptron

(x, y) : given data.



$z = f_{w_1, w_2}(x)$

$L(f_{w_1, w_2}) = | y - \underbrace{f_{w_1, w_2}}_{\text{parameter}}(x) |^2$

Stochastic gradient decsent Amari, Tsypkin '66 ~ 67

Error Back Propagation '76

§1 Overview

A standard setting

1. Multilayer Perceptron is a parametric family $(f_\theta)_{\theta \in \Theta}$

w/ $f_\theta : \mathbb{R}^M \to \mathbb{R}^M \quad (M \in \mathbb{N})$

$$= \varphi^L \circ g^L_{\theta^L} \circ \cdots \circ \varphi^1 \circ g^1_{\theta^1}$$

$$g^\ell_{\theta^\ell}(x) = \underbrace{W^\ell}_{M \times M} x + \underbrace{b^\ell}_{\in \mathbb{R}^M} \quad (\theta^\ell = (W^\ell, b^\ell))$$

activation $\varphi^\ell \in C(\mathbb{R})$: differentiable except for finite number of points

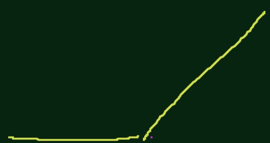
$$\varphi^\ell(x) = \begin{pmatrix} \varphi^\ell(x_1) \\ \vdots \\ \varphi^\ell(x_n) \end{pmatrix}$$

$$x = x^0 \xrightarrow{W^1} h^1 \xrightarrow{\varphi^1} x^1 \xrightarrow{W^2} h^2 \xrightarrow{\varphi^2} x^2 \cdots \to x^L = f_\theta(x)$$

Examples of Activation

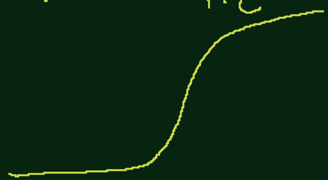
ReLU '15

$\varphi(x) = \max(0, x)$

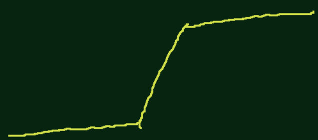


Sigmoid

$\varphi(x) = \dfrac{1}{1+e^{-x}}$

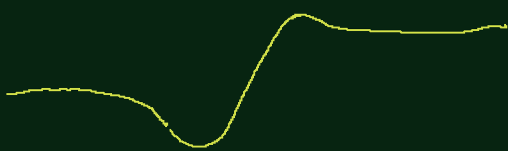


Hard tanh



SiLU (Sigmoid Linear Unit) '17
(or Swish)

$\varphi(x) = \dfrac{x}{1+e^{-x}}$



testing dataset : training dataset に含まれていないが あるていど近い dataset.

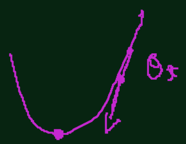
2. A training dataset is a set of pairs $(x_n, y_n)_{n=1}^{N}$

3. Gradient Descent

$$L(f_\theta) = \frac{1}{2N}\sum_{n=1}^{N} \| f_\theta(x_n) - y_n \|_2^2 \quad (\text{MSE})$$

$t = 0, 1, 2, \cdots$

$$\theta_{t+1} = \theta_t - \eta_t \left.\frac{d}{d\theta} L(f_\theta)\right|_{\theta=\theta_t} \quad (\eta_t > 0)$$

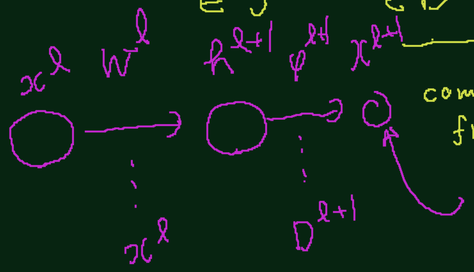


※ In practice (x_n, y_n) are picked randomly from the training dataset.

4. Error Back-Propagation

$$\frac{\partial L}{\partial W^l} = \underbrace{(f_\theta(x) - y)^T}_{\varepsilon^T} \underbrace{\frac{\partial f_\theta}{\partial x^{l+1}}}_{J^{l+1}} D^{l+1} x^l$$

$$\varepsilon^T J^l = \varepsilon^T D^L W^L \cdots D^l W^l$$

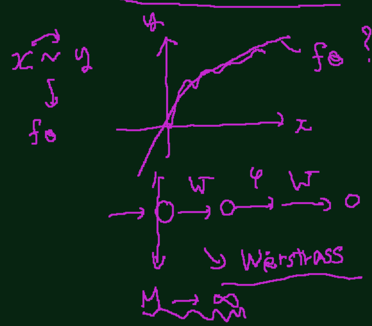


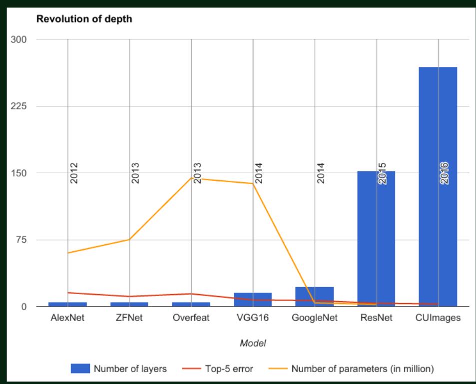
compute & store from L to 1 : Error Back propagation

Deep

- ILSVRC 2012
  (ImageNet Large Scale Visual Recognition Challenge)
  : AlexNet (Hinton e.t.c)
  ImageNet Classification with Deep Convolutional Neural Networks
  (NIPS 2012)

- Deepである必要
  - 表現能力の指数的向上
  - 性能は上昇する(?)





§2. Dynamical Isometry

Fix L (= the number of Layers)

$$J = D^L W^L \cdots D^1 W^1$$

D^l : activation の Jacobian.

$W^1, \cdots, W^L$: initial state

- ✓ Uniform
- ✓ Gaussian
- ✓ Haar orthogonal (uniformly picking up from $O(M)$)

Error back propagation $\| \varepsilon J \|_2$ can explode / vanish as $L \to \infty$

"Expoding/Vanishing Gradient Problem"

Pennington, Schoenholz, Ganguli [NIPS '17, AISTATS '18]

「Ensuring J's singular values ($\sqrt{JJ^T}$'s eigenvalues) $\sim O(1)$ as $L \to \infty$ (Dynamical Isometry) is essential for avoiding the exploding/vanishing of gradients」

「Gaussian Initialization ✕

Orthogonal Initialization + normalization of ρ $\Rightarrow$ D.I.

$$S_{\lim_{M\to\infty} \mu_{JJ^T}}(z) \xrightarrow{L\to\infty} \exp\left(-\frac{z\sigma_0^2}{1+z}\right) \quad \text{(Voiculescu)}$$

free prob.

← Free Infinite Multipricative Infinite Divisible

limit spectral distribution

Sketch

Assume that $(W_1, W_1^*), \cdots, (W_L, W_L^*)$, $(D_1, \cdots, D_L)$ are asymptotic free as $M \to \infty$.

Let $(u_1, u_1^*), \cdots (u_L, u_L^*)$, $(d_1, \cdots, d_L)$ be free family in a C^*-prob. sp. $(\mathcal{A}, \tau)$.

free probability

$$j := d_L u_L \cdots d_1 u_1$$

$$jj^* = d_L u_L \underbrace{(d_{L-1} u_{L-1} \cdots d_1^2 \cdots u_{L-1} d_{L-1})}_{j_{L-1} j_{L-1}^*} u_L^* d_L$$

$$S_{jj^*}(z) = S_{d_L^2}(z) S_{j_{L-1} j_{L-1}^*}(z)$$

$$= S_{d_1^2}(z) \cdots S_{d_L^2}(z)$$

← unitary だから消えた

§3. Fisher Information Matrix & Neural Tangent Kernel

$$I(\theta) = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial f_\theta(x_n)}{\partial \theta}^T \frac{\partial f_\theta(x_n)}{\partial \theta}$$

: $M^2L \times M^2L$ 行列

(Empirical)
Fisher Information Matrix

$b^2=0$とした場合

$D_{KL}(P_\theta \| P_{\theta+d\theta})$
$\approx (d\theta)^T I(\theta)\, d\theta$

(Information Geometry (Amari))

$P_\theta = \exp(-L(f_\theta))/Z.$

Neural Tangent Kernel

$$\Theta = \frac{\partial f_\theta(x)}{\partial \theta} \frac{\partial f_\theta(x)}{\partial \theta}^T$$

: M×M 行列
C×C (C<<M) : 出力の次元

NTK describes learning dynamics

Learning dynamics of parameters is given by:

$$\frac{d\theta_t}{dt} = \eta(\nabla_\theta f_{\theta_t})^T (y - f_{\theta_t})$$

Learning dynamics of DNN is given by:

$$\frac{df_{\theta_t}}{dt} = \eta\Theta_t(y - f_{\theta_t})$$

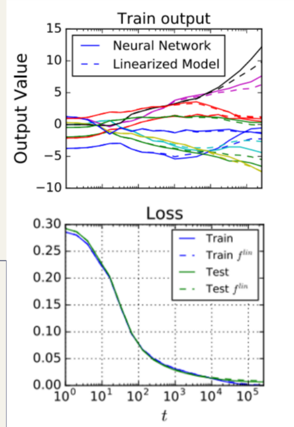
where
$$\Theta_t = \nabla_\theta f_{\theta_t}(\nabla_\theta f_{\theta_t})^T$$

Informal [Jacot+NeurIPS2018, Lee+NeurIPS2019]: Under the wide limit M \to \infty, the learning of the DNN is approximated by

$$\frac{df_{\theta_t}}{dt} = \eta\Theta(y - f_{\theta_t})$$

where
$$\Theta = \nabla_\theta f_\theta(\nabla_\theta f_\theta)^T$$

(Neural Tangent Kernel)



14

θ : 初期値

(?) orthogonal

✓ Gaussian

✓ $C = O(1)$ as $M \to \infty$

M?

✓ D.I (?)

Spectrum of $\overset{\text{per sample}}{\frac{\partial f_\theta}{\partial \theta}}$ $\left(= D^L \frac{\partial x^L}{\partial \theta}\right)$ arXiv: 2006.07814

T.H. & Ryo Karakida [AISTATS '21]

$$H_L := \frac{1}{M} \frac{\partial x^L}{\partial \theta} \frac{\partial x^{L}}{\partial \theta}^{*}$$ ..... A dual of Fisher Information Matrix per sample

... NTK

Then limit spectral distribution of H_L

concentrates on the point q^L as $M \to \infty$

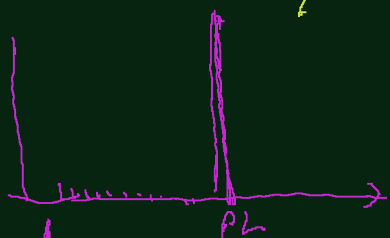
$$\text{w/ } q = \lim_{M\to\infty} \frac{1}{M} \|x^\ell\|_2^2$$

under some assumptions to achieve dynamical isometry.

Sketch

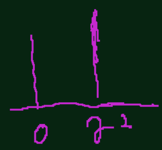
$$\frac{1}{M}\|x^\ell\|_2^2$$

$$H_{\ell+1} = \hat{q}_\ell + W_{\ell+1} D_\ell H_\ell D_\ell W_{\ell+1}^{*}$$



asymptotic freeness

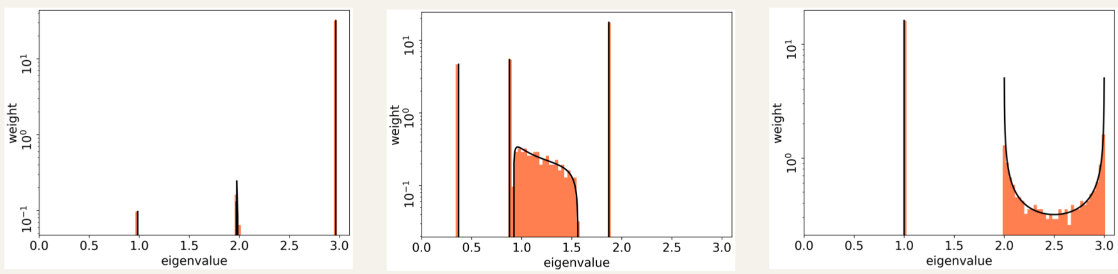
$$\Rightarrow \underbrace{\mu_{\ell+1}}_{H_{\ell+1}} = (q + \cdot)_* (\underbrace{\nu_\ell}_{D_\ell^2} \boxtimes \mu_\ell)$$



Point: If activation function is ReLU-like ftn,

- μ_ℓ has an atom as a maximal point spectrum.
- $\nu \boxtimes \mu(\{ab\}) = \nu(\{a\}) \cancel{\mu(\{b\})} + \mu(\{b\}) - 1$.

Limit spectral distributions : L = 3



D → P   P Q   P, Q : free

↓

0 or 1

Training Dynamics

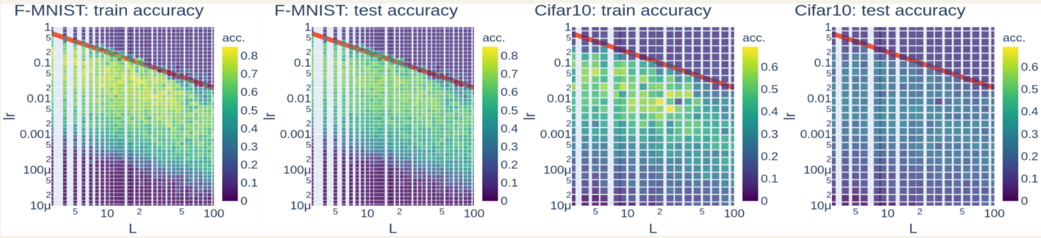
- $\eta > \frac{1}{\lambda_{max}(H_L)} \Rightarrow$ The training dynamics may not converge

Training under D. Isometry

Red line (the boarder line of the exploding gradients) :

$$\eta = 2/L$$

This line is expected by our theory !



Conclusion

- Random Matrices appear in the theory of deep neural networks (e.g. dynamical isometry, Fisher Information, NTK)
- Since Jacobian (input or parameter) are (noncommutative) polynomial of Random Matrices,

Free Probability Theory provides tools for handling them!

To use this, we have to prove asymptotic freeness $\begin{cases} \text{a.s.} \\ \text{expected.} \\ \text{strong (operator norm)} \end{cases}$

Gaussian $\begin{cases} \text{Hanin-Nica '19} \\ \text{Yang '19, '20} \\ \text{Pastur '20} \end{cases}$

Orthogonal → H.) (under preparation)

'67

gradient independence

[illegible]

$\begin{cases} W^1, \cdots, W^L \\ D^1, \cdots, D^L \end{cases}$ independent と思って

freeness → $\mathrm{tr}(Q(W^1, \cdots, W^L, D^1, \cdots, D^L))$: ok (?)