

深層ニューラルネットワークにおける 学習ダイナミクスの初等的解析

宇都宮 幸大

所属：XXXXXX

指導教員：XXXXXX

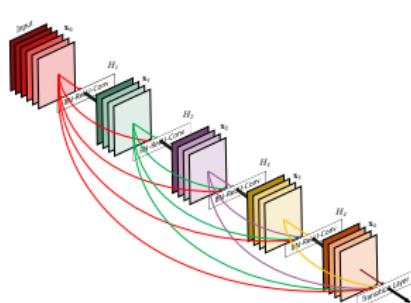
Outline

1. 研究背景 & 問題提起
2. 解析の概要
3. コンピュータシミュレーション
4. まとめ

Outline

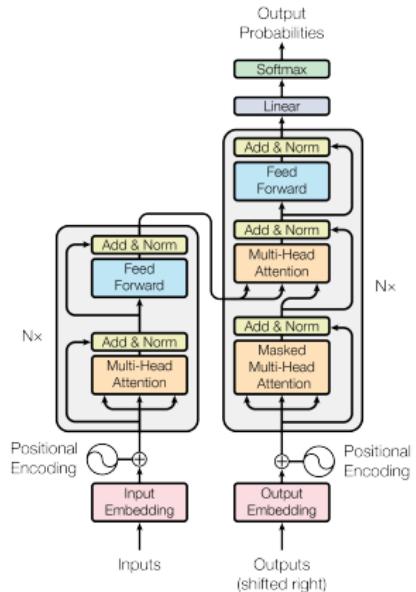
1. 研究背景 & 問題提起
2. 解析の概要
3. コンピュータシミュレーション
4. まとめ

深層学習の成功



DenseNet [Huang *et al.*, 2017]

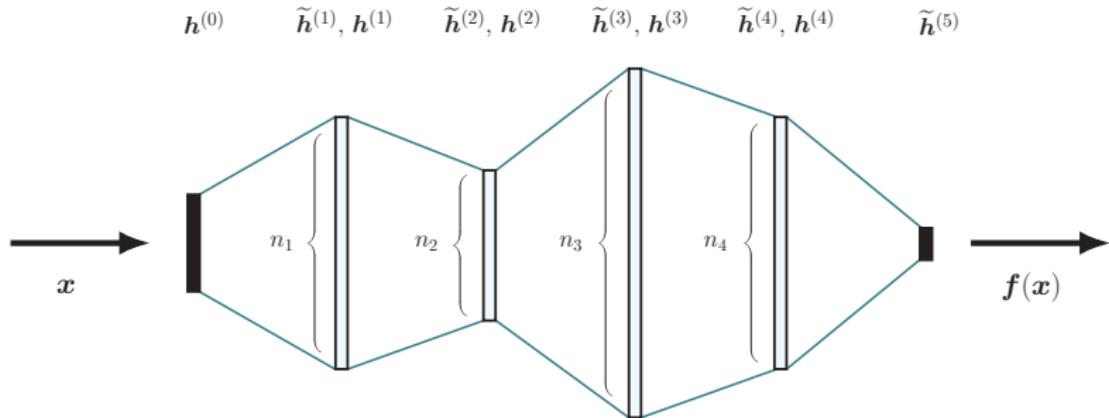
ResNet [He *et al.*, 2015]



Transformer [Vaswani *et al.*, 2017]

- 巨大な数の可変なパラメータを用いる。やってみると上手くいく
- 理論の後追い（2018年頃～）

深層NNの定義と学習の枠組み



$$f(x) := \tilde{h}^{(L)},$$

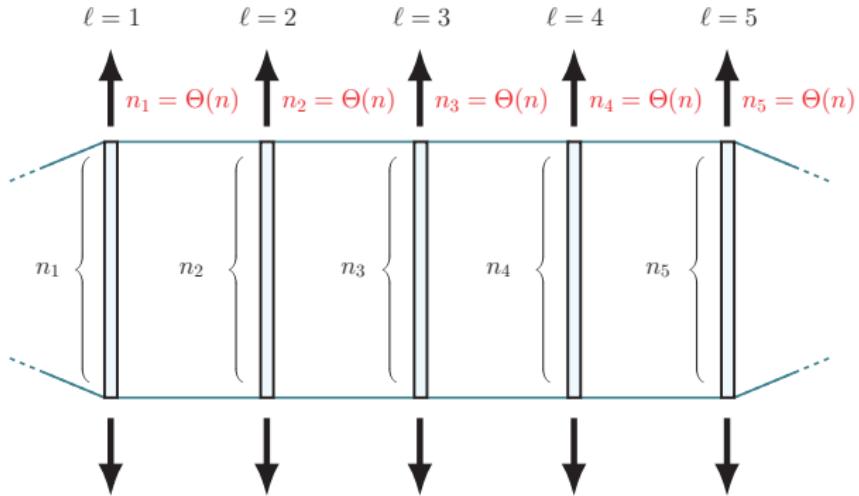
$$\tilde{h}^{(\ell)} = \textcolor{red}{g_\ell} W^{(\ell)} h^{(\ell-1)}, \quad 1 \leq \ell \leq L,$$

$$h^{(\ell)} = \psi(\tilde{h}^{(\ell)}), \quad 1 \leq \ell \leq L-1,$$

$$h^{(0)} = x.$$

- $W_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma_\ell^2)$ で独立に初期化. 損失 $\mathcal{L}(f(x), y)$ を最小化

ニューロン数無限大の学習ダイナミクス



- **Maximal Update Parametrization ($\mu\mathbf{P}$) の理論** [Yang *et al.*, 2021]
- $|\Delta \tilde{h}_i^{(\ell)}| = \Theta(1)$ を実現する g_ℓ , σ_ℓ のとり方を提案
- $n := n_1 = n_2 = \dots = n_L$ & すべて $\Theta(n)$ を前提
- $n_{\ell-1} \neq n_\ell$ の場合は? \Rightarrow **Spectral Parametrization** [Yang *et al.*, 2023]

研究課題

自然な問い合わせ

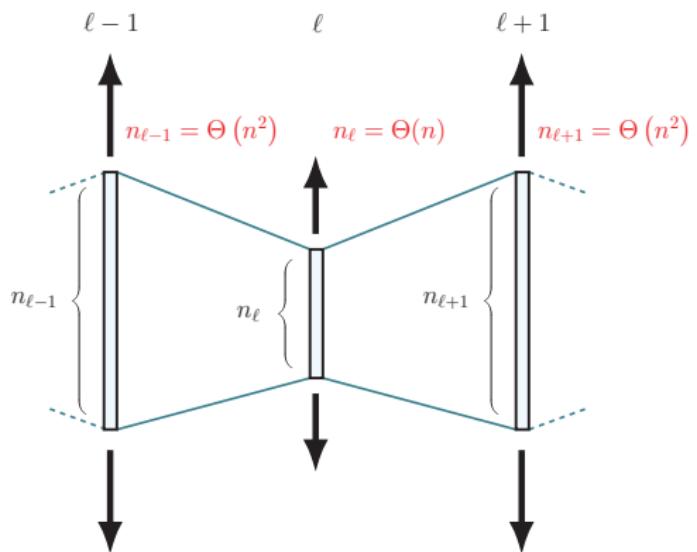
中間層のニューロン数が層によって異なるオーダーのときは？

例： $n_{\ell-1} = \Theta(n^2)$

$n_\ell = \Theta(n)$

$n_{\ell+1} = \Theta(n^2)$

- このときにも学習が安定に進むパラメータ設定を導出



Outline

1. 研究背景 & 問題提起
2. 解析の概要
3. コンピュータシミュレーション
4. まとめ

順伝播と逆伝播のオーダー評価

$\tilde{h}^{(\ell)}$ の各要素が消失や発散しない

$$|\tilde{h}_i^{(\ell)}| = \Theta(1)$$

損失 \mathcal{L} が効果的に減少

$$\left| \left\langle \frac{\partial \mathcal{L}}{\partial \tilde{h}^{(\ell)}}, \Delta \tilde{h}^{(\ell)} \right\rangle \right| \stackrel{\text{req.}}{=} \Theta(1)$$

... 計算を進めると、中間表現の更新量

$$\|\Delta \tilde{h}^{(1)}\|_2, \|\Delta \tilde{h}^{(2)}\|_2, \dots, \|\Delta \tilde{h}^{(L-1)}\|_2$$

は同じオーダーでなければならぬことが導かれる

σ_ℓ と g_ℓ が満たすべき条件

- $\left| \left\langle \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{h}}^{(\ell)}}, \Delta \tilde{\mathbf{h}}^{(\ell)} \right\rangle \right| = \dots = \Theta \left(g_\ell^2 \|\mathbf{h}^{(\ell-1)}\|_2^2 \frac{1}{\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2^2} \right)$
- $g_\ell \sqrt{n_{\ell-1}} \sigma_\ell \stackrel{\text{req.}}{=} \Theta(1)$
- $\|\Delta \tilde{\mathbf{h}}^{(L)}\|_2 \stackrel{\text{req.}}{=} \Theta(1)$

中間層 $1 \leq \ell \leq L-1$:

$$g_\ell = \Theta \left(\frac{\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2}{\sqrt{n_{\ell-1}}} \right), \quad \sigma_\ell = \Theta \left(\frac{1}{\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2} \right)$$

出力層 (第 L 層) :

$$g_L = \Theta \left(\frac{1}{\sqrt{n_{L-1}}} \right), \quad \sigma_L = \Theta \left(\frac{1}{\|\Delta \tilde{\mathbf{h}}^{(L-1)}\|_2} \right)$$

勾配とモデル出力の発散を防ぐための基準

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{h}}^{(\ell)}} = g_{\ell+1} \left(\mathbf{W}^{(\ell+1)} + \Delta \mathbf{W}^{(\ell+1)} \right)^\top \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{h}}^{(\ell+1)}}$$

- 勾配の発散を防ぐための基準 : $\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2 \stackrel{\text{req.}}{=} O(\sqrt{n_{\ell-1}})$
- $\|\Delta \tilde{\mathbf{h}}^{(1)}\|_2, \|\Delta \tilde{\mathbf{h}}^{(2)}\|_2, \dots, \|\Delta \tilde{\mathbf{h}}^{(L-1)}\|_2$ の条件を踏まえると

$$\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2 \stackrel{\text{req.}}{=} O(\sqrt{n_{\min}})$$

- モデル出力の発散を防ぐための基準 : $\|\Delta \tilde{\mathbf{h}}^{(L-1)}\|_2 \stackrel{\text{req.}}{=} \Omega(1)$

学習が安定に進むための基準

$$\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2 = \Theta(n_{\min}^r), \quad r \in [0, 1/2]$$

提案手法

Dynamic Parametrization (DP):

$$\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{n_{L-1}}} \mathbf{W}^{(L)} \mathbf{h}^{(L-1)},$$

$$\tilde{\mathbf{h}}^{(\ell)} = \frac{n_{\min}^r}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \mathbf{h}^{(\ell-1)}, \quad 1 \leq \ell \leq L-1,$$

$$\mathbf{h}^{(\ell)} = \psi(\tilde{\mathbf{h}}^{(\ell)}), \quad 1 \leq \ell \leq L-1,$$

$$\tilde{\mathbf{h}}^{(0)} = \mathbf{x}.$$

- ただし, $r \in [0, 1/2]$. 重み行列の各要素は,

$$W_{ij}^{(\ell)} \sim \mathcal{N} \left(0, \frac{\gamma_\ell^2}{n_{\min}^{2r}} \right)$$

で独立に初期化する

- $\psi(z) = \max(0, z)$ の場合, $\gamma_1 = \dots = \gamma_{L-1} = \sqrt{2}$ & $\gamma_L = 1$

既存手法との比較

Spectral Parametrization [Yang *et al.*, 2023]

$$g_\ell = \Theta\left(\sqrt{\frac{n_\ell}{n_{\ell-1}}}\right), \quad \sigma_\ell = \begin{cases} \Theta\left(\frac{1}{\sqrt{n_\ell}}\right), & \text{if } n_{\ell-1} \leq n_\ell, \\ \Theta\left(\frac{1}{\sqrt{n_{\ell-1}}}\right), & \text{if } n_{\ell-1} > n_\ell \end{cases}$$

- $|\Delta \tilde{h}_i^{(\ell)}| = \Theta(1)$ すなわち $\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2 = \Theta(\sqrt{n_\ell})$ を実現
- $\|\Delta \tilde{\mathbf{h}}^{(1)}\|_2, \|\Delta \tilde{\mathbf{h}}^{(2)}\|_2, \dots, \|\Delta \tilde{\mathbf{h}}^{(L-1)}\|_2$ もそれぞれ異なるオーダーに.

$\Delta \tilde{\mathbf{h}}^{(\ell)}$ がニューロン数に依らず損失 \mathcal{L} の減少に寄与することを意味する

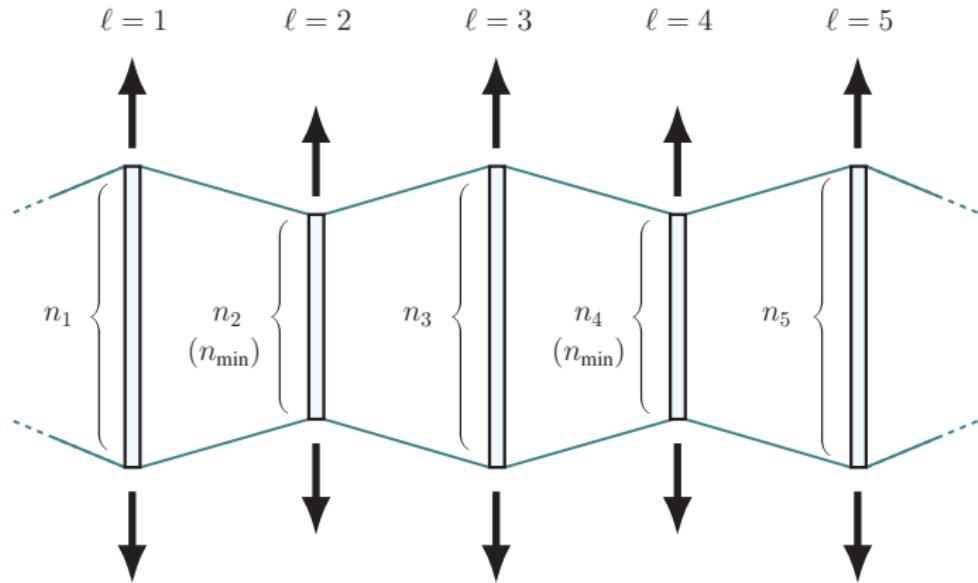
$$\left| \left\langle \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{h}}^{(\ell)}}, \Delta \tilde{\mathbf{h}}^{(\ell)} \right\rangle \right| = \Theta(1)$$

が満たされないと考えられる

Outline

1. 研究背景 & 問題提起
2. 解析の概要
3. コンピュータシミュレーション
4. まとめ

既存手法との比較

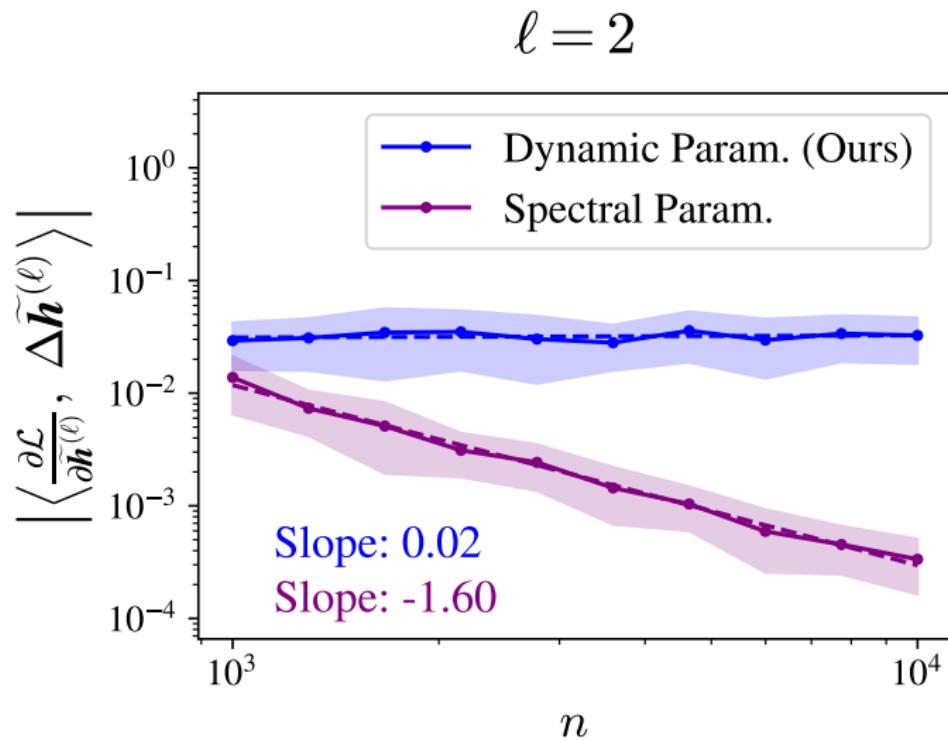


- $n := n_1 = n_3 = n_5$

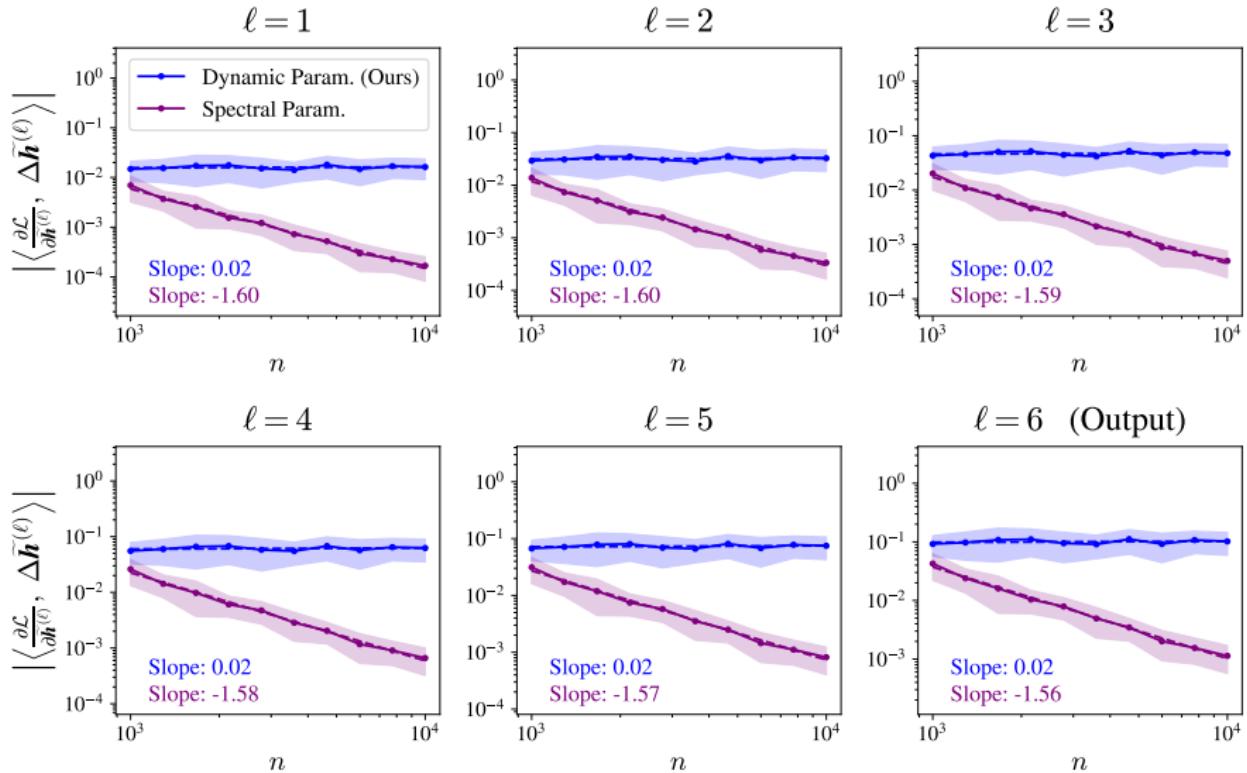
- $n_{\min} := n_2 = n_4$

$$n_{\min} = 150 n^{1/5}. \text{ すなわち } n_{\min} \text{ が } \Theta(n) \text{ ではなく } \Theta(n^{1/5})$$

既存手法との比較

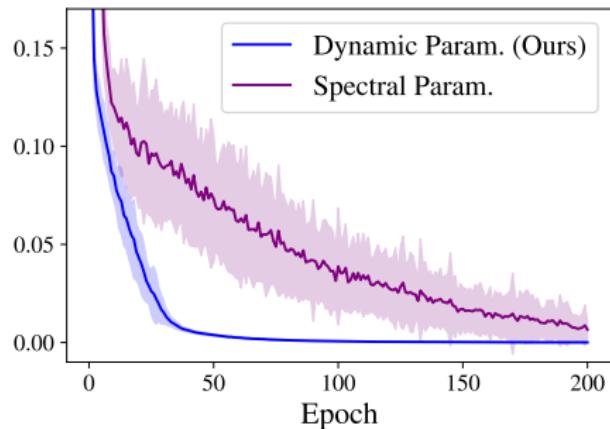


既存手法との比較

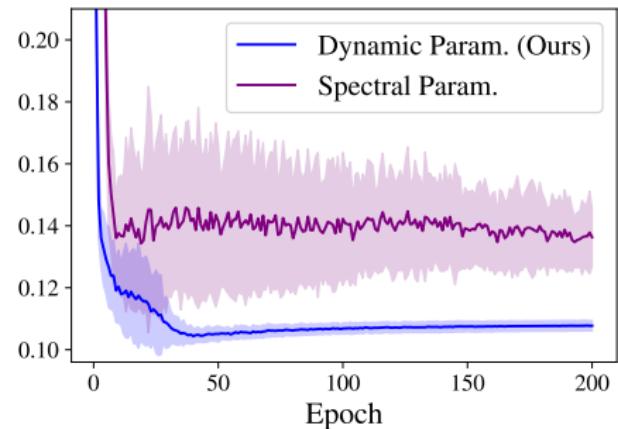


性能への影響

Train Loss



Test Loss



Outline

1. 研究背景 & 問題提起
2. 解析の概要
3. コンピュータシミュレーション
4. まとめ

まとめ

- 学習が安定に進むための基準として

$\|\Delta \tilde{\mathbf{h}}^{(\ell)}\|_2 = \Theta(n_{\min}^r)$, $r \in [0, 1/2]$ を導くとともに,

この基準を満たす具体的なパラメータ設定を提案

- 訓練性と汎化性能の向上が期待される

本研究の解析の限界

- 漸近的な議論における確率的な揺らぎ
- 学習によって獲得される具体的な表現

例 :

- 訓練データの特徴をどのような階層構造として抽出するか？
- ボトルネック部分でどのような情報圧縮のメカニズムが働くか？

【再掲】提案手法

Dynamic Parametrization (DP):

$$\mathbf{f}(\mathbf{x}) = \frac{1}{\sqrt{n_{L-1}}} \mathbf{W}^{(L)} \mathbf{h}^{(L-1)},$$

$$\tilde{\mathbf{h}}^{(\ell)} = \frac{n_{\min}^r}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \mathbf{h}^{(\ell-1)}, \quad 1 \leq \ell \leq L-1,$$

$$\mathbf{h}^{(\ell)} = \psi(\tilde{\mathbf{h}}^{(\ell)}), \quad 1 \leq \ell \leq L-1,$$

$$\tilde{\mathbf{h}}^{(0)} = \mathbf{x}.$$

- ただし, $r \in [0, 1/2]$. 重み行列の各要素は,

$$W_{ij}^{(\ell)} \sim \mathcal{N} \left(0, \frac{\gamma_\ell^2}{n_{\min}^{2r}} \right)$$

で独立に初期化する

- $\psi(z) = \max(0, z)$ の場合, $\gamma_1 = \dots = \gamma_{L-1} = \sqrt{2}$ & $\gamma_L = 1$