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

60210412

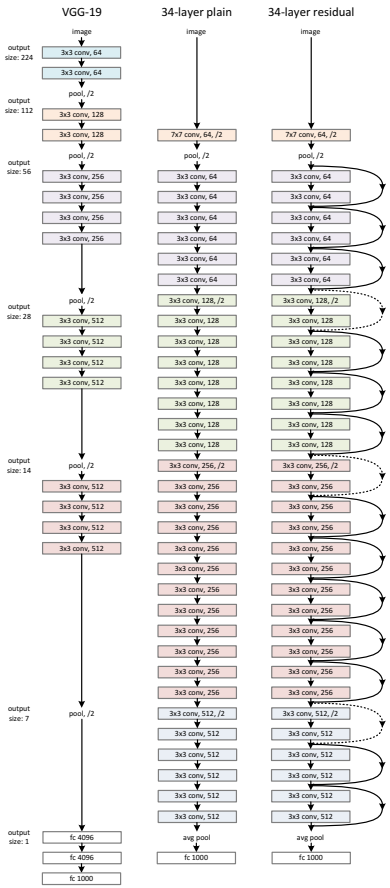
宮崎大学 工学部 情報通信工学プログラム

指導教員：伊達 章 准教授

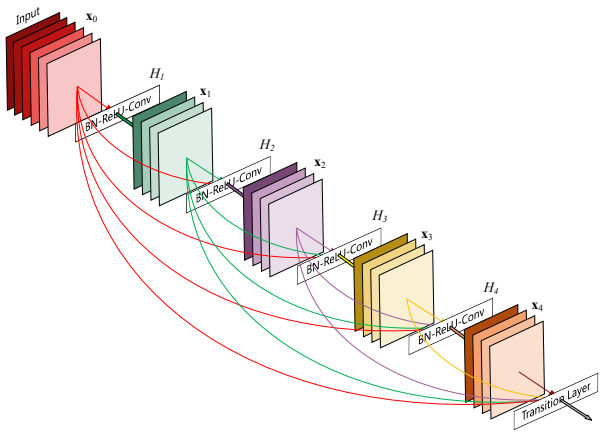
Outline

Outline

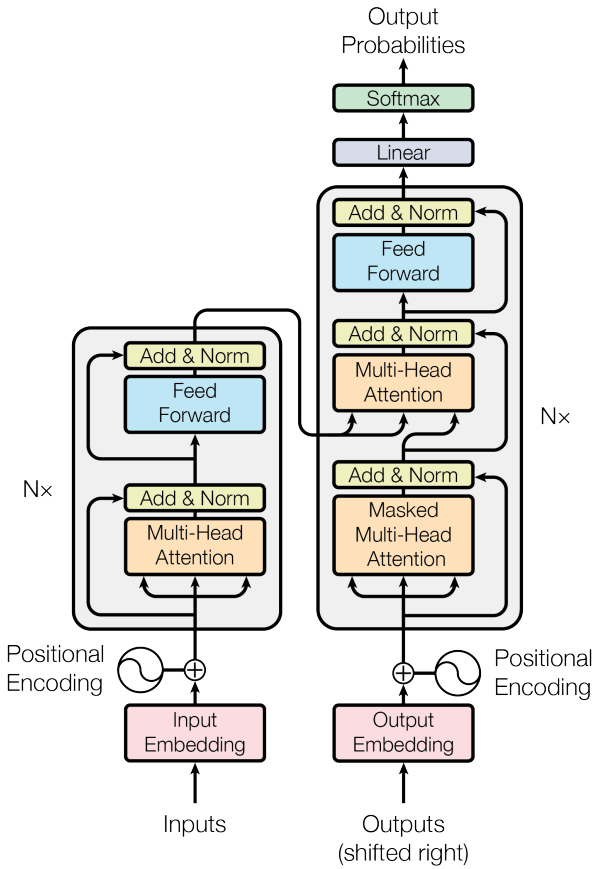
深層学習の成功



ResNet [He *et al.*, 2015]



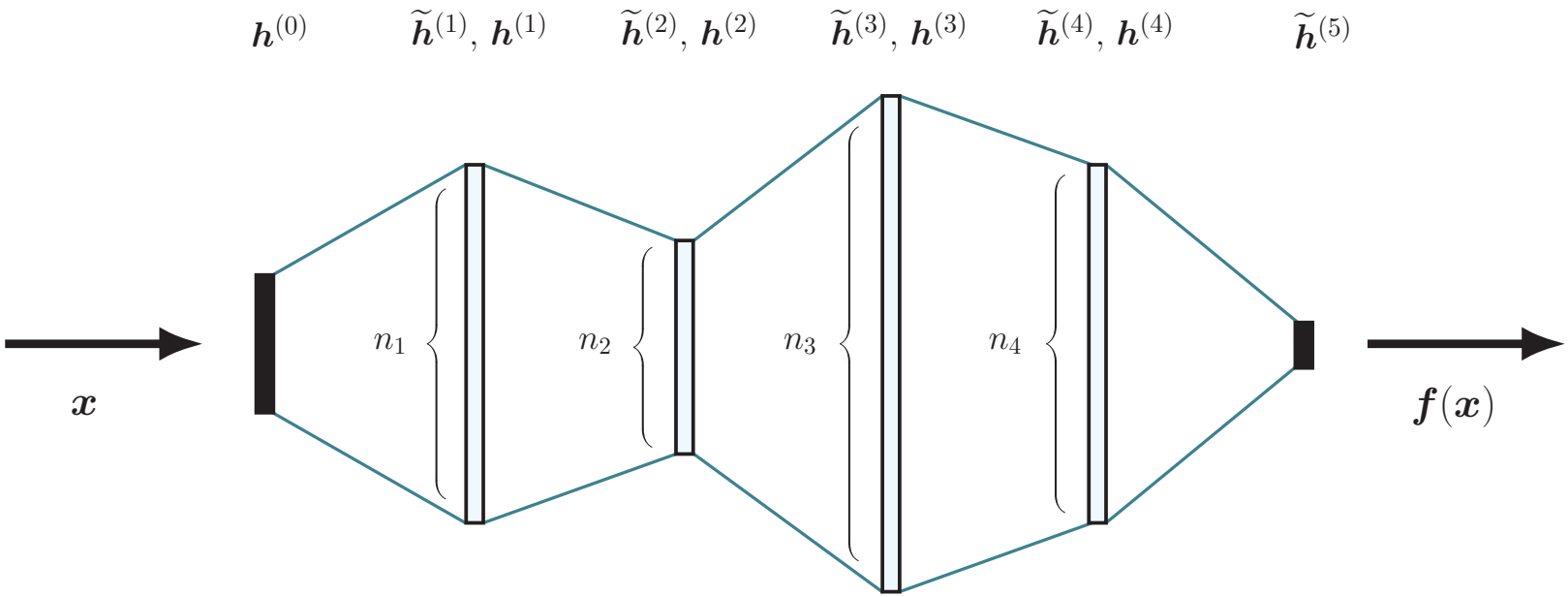
DenseNet [Huang *et al.*, 2017]



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〜）

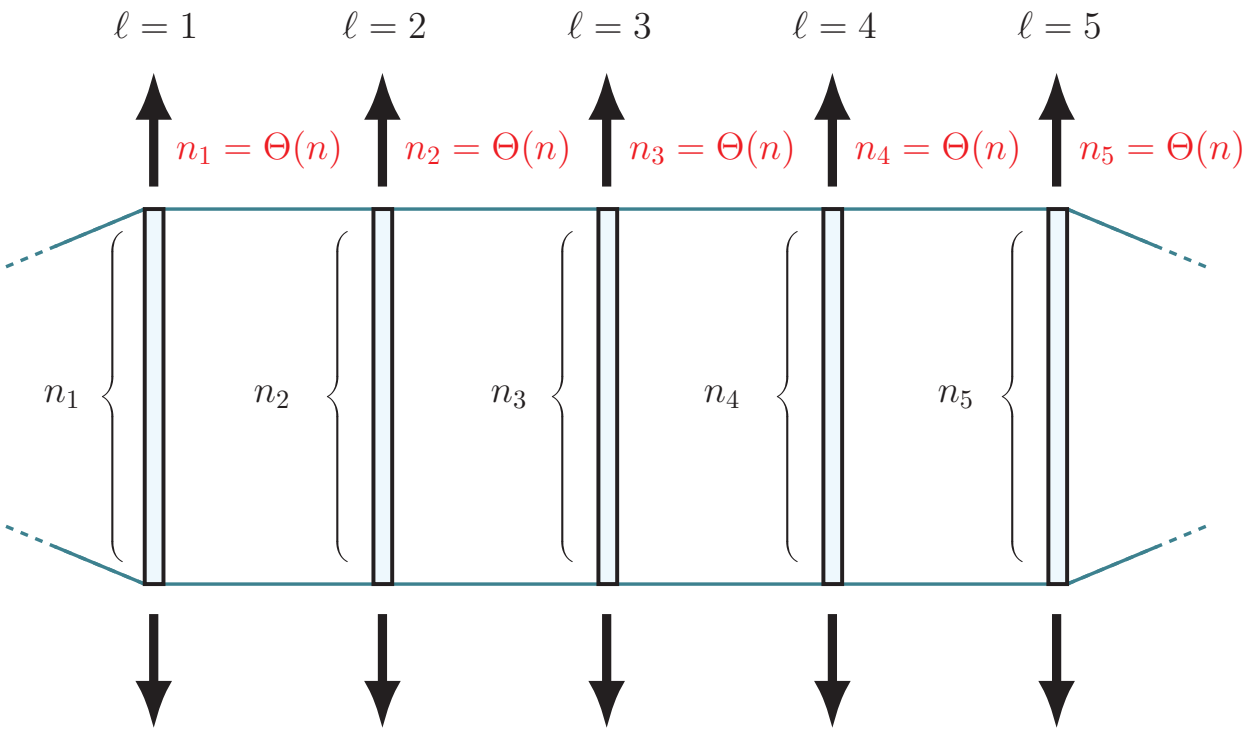
深層 NN の定義と学習の枠組み



$$
\begin{aligned}
\boldsymbol{f}(\boldsymbol{x}) &:= \widetilde{\boldsymbol{h}}^{(L)}, \\
\widetilde{\boldsymbol{h}}^{(\ell)} &= g_\ell \boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)}, && 1 \leq \ell \leq L, \\
\boldsymbol{h}^{(\ell)} &= \psi(\widetilde{\boldsymbol{h}}^{(\ell)}), && 1 \leq \ell \leq L-1, \\
\boldsymbol{h}^{(0)} &= \boldsymbol{x}.
\end{aligned}
$$

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

ニューロン数無限大の学習ダイナミクス



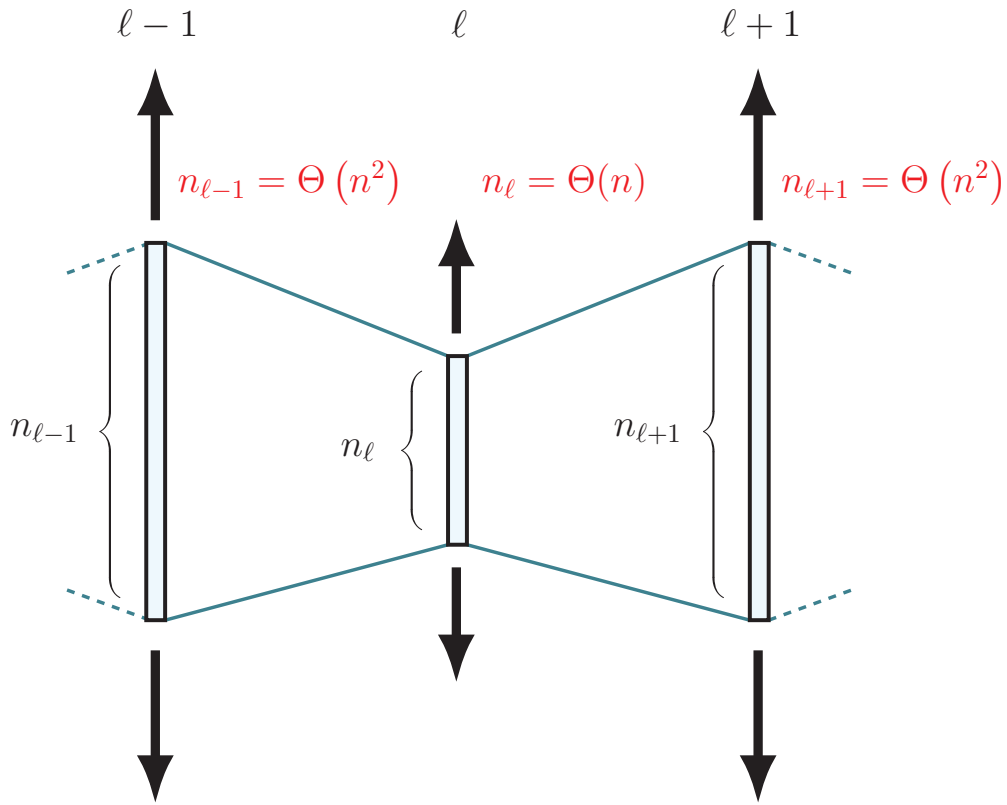
- **Maximal Update Parametrization** (μ**P**) の理論 [Yang *et al.*, 2021]
- $|\Delta \widetilde{h}_i^{(\ell)}| = \Theta(1)$ を実現する g_ℓ, σ_ℓ のとり方を提案
- $n := n_1 = n_2 = \cdots = n_L$ & すべて $\Theta(n)$ を前提
- $n_{\ell-1} \neq n_\ell$ の場合は？ $\Rightarrow$ **Spectral Parametrization** [Yang *et al.*, 2023]

研究課題

自然な問い

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：
$$n_{\ell-1} = \Theta\left(n^2\right)$$
$$n_{\ell} = \Theta\left(n\right)$$
$$n_{\ell+1} = \Theta\left(n^2\right)$$



- 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

順伝播と逆伝播のオーダー評価

$\widetilde{h}^{(\ell)}$ の各要素が消失や発散しない

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

損失 $\mathcal{L}$ が効果的に減少

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

... 計算を進めると，中間表現の更新量

$$\|\Delta \widetilde{\boldsymbol{h}}^{(1)}\|_2, \ \|\Delta \widetilde{\boldsymbol{h}}^{(2)}\|_2, \ \ldots, \ \|\Delta \widetilde{\boldsymbol{h}}^{(L-1)}\|_2$$

は同じオーダーでなければならないことが導かれる

σ_ℓ と g_ℓ が満たすべき条件

- $\left| \left\langle \dfrac{\partial \mathcal{L}}{\partial \widetilde{\boldsymbol{h}}^{(\ell)}},\ \Delta \widetilde{\boldsymbol{h}}^{(\ell)} \right\rangle \right| = \cdots = \Theta\left(g_\ell^2 \|\boldsymbol{h}^{(\ell-1)}\|_2^2 \dfrac{1}{\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2^2} \right)$
- $g_\ell \sqrt{n_{\ell-1}}\, \sigma_\ell \overset{\text{req.}}{=} \Theta(1)$
- $\|\Delta \widetilde{\boldsymbol{h}}^{(L)}\|_2 \overset{\text{req.}}{=} \Theta(1)$

中間層 $1 \le \ell \le L-1$：

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

出力層（第 L 層）：

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

勾配とモデル出力の発散を防ぐための基準

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

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

$$\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2 \overset{\text{req.}}{=} O(\sqrt{n_{\min}})$$

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

学習が安定に進むための基準

$$\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2 = \Theta\left(n_{\min}^{r}\right), \qquad r \in [0,\, 1/2]$$

提案手法

Dynamic Parametrization (DP):

$$
\begin{aligned}
\boldsymbol{f}(\boldsymbol{x}) &= \frac{1}{\sqrt{n_{L-1}}} \boldsymbol{W}^{(L)} \boldsymbol{h}^{(L-1)}, \\
\widetilde{\boldsymbol{h}}^{(\ell)} &= \frac{n_{\text{min}}^{r}}{\sqrt{n_{\ell-1}}} \boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)}, && 1 \leq \ell \leq L-1, \\
\boldsymbol{h}^{(\ell)} &= \psi(\widetilde{\boldsymbol{h}}^{(\ell)}), && 1 \leq \ell \leq L-1, \\
\widetilde{\boldsymbol{h}}^{(0)} &= \boldsymbol{x}.
\end{aligned}
$$

- ただし，$r \in [0,\, 1/2]$．重み行列の各要素は，

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

  で独立に初期化する

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

既存手法との比較

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

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

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

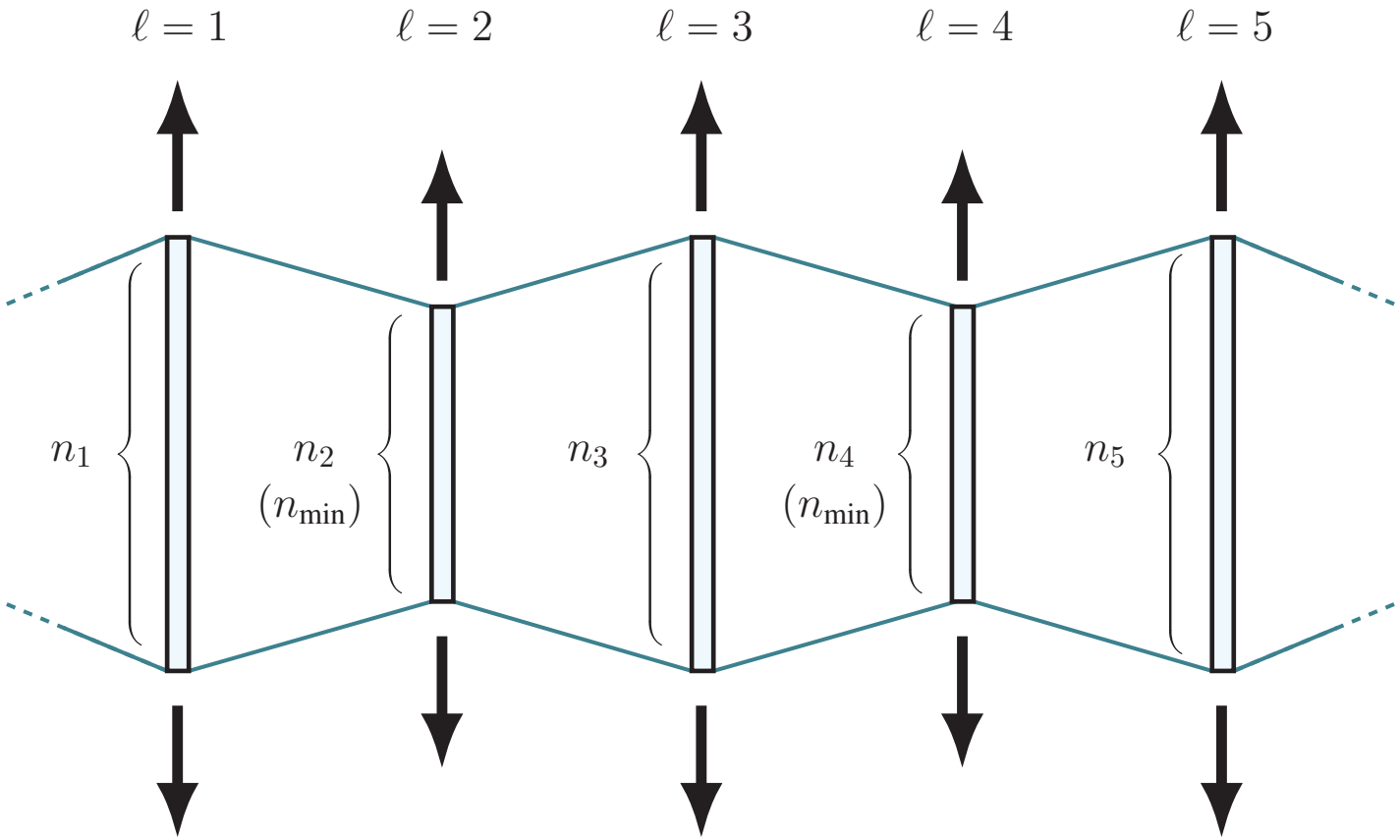
$\Delta\widetilde{\boldsymbol{h}}^{(\ell)}$ がニューロン数に依らず損失 $\mathcal{L}$ の減少に寄与することを意味する

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

が満たされないと考えられる

Outline

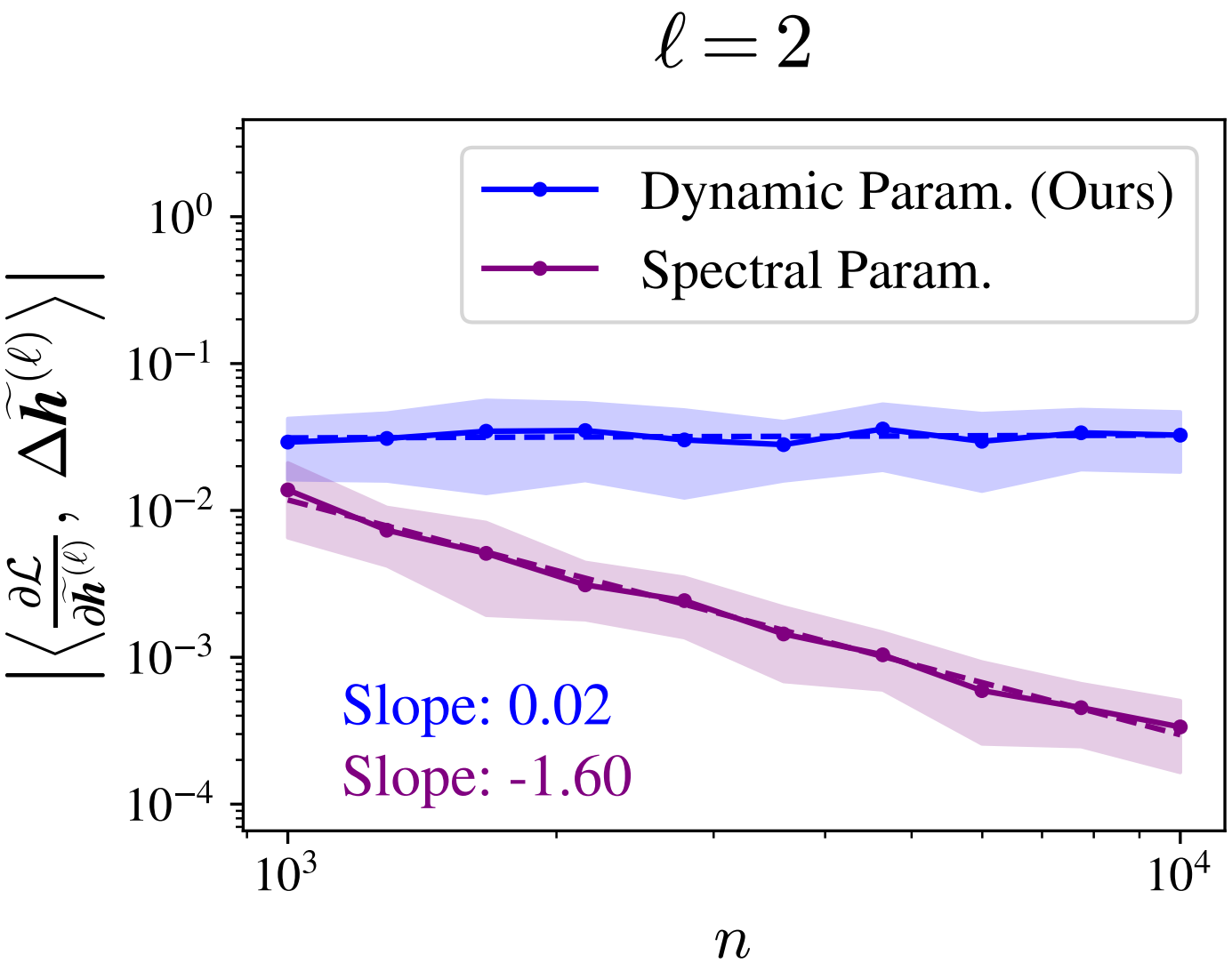
既存手法との比較



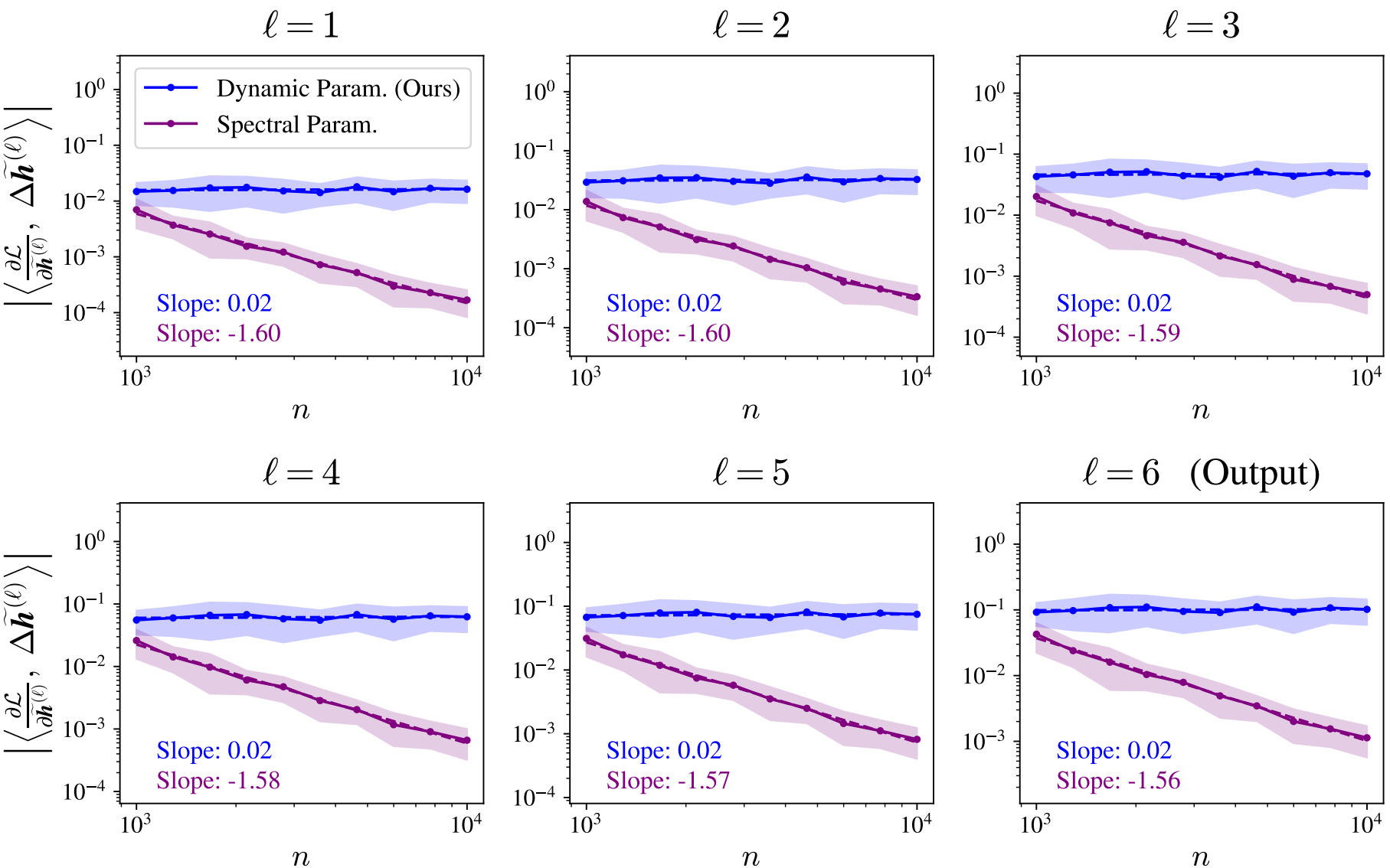
- $n \coloneqq n_1 = n_3 = n_5$
- $n_{\min} \coloneqq n_2 = n_4$

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

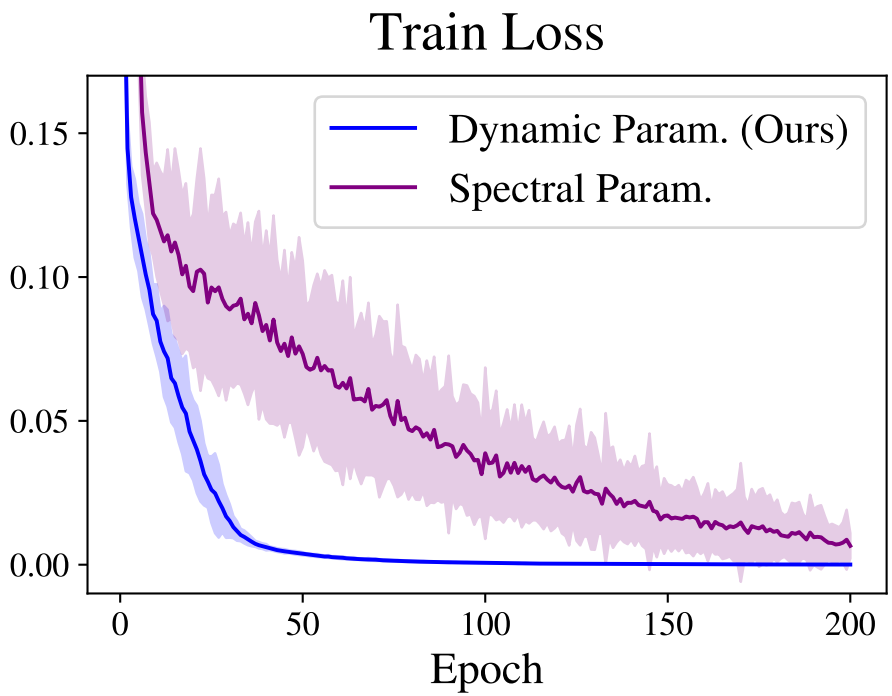
既存手法との比較

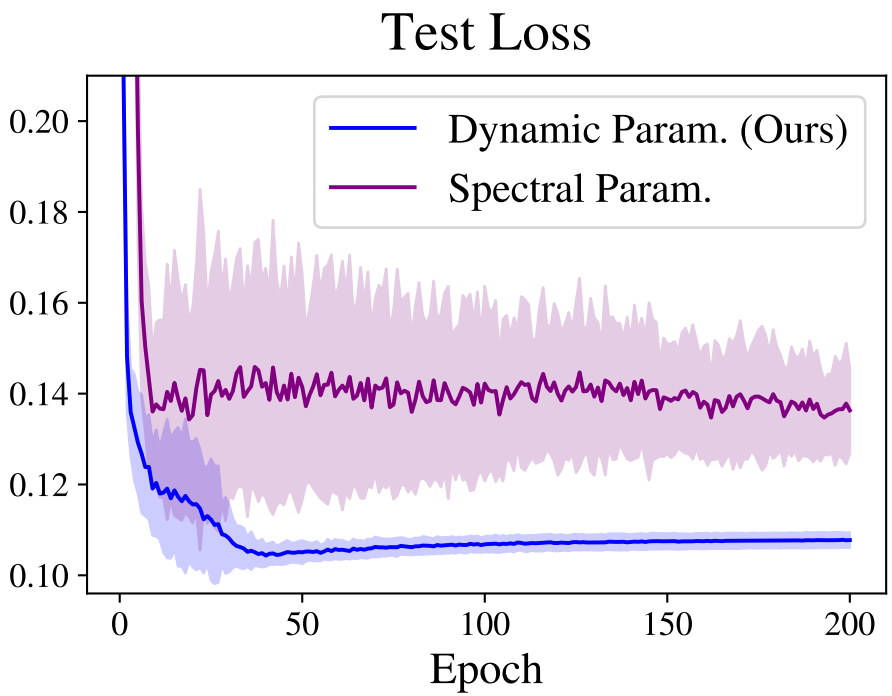

$\ell = 2$
Dynamic Param. (Ours)
Spectral Param.
Slope: 0.02
Slope: -1.60
$\left|\left\langle \frac{\partial \mathcal{L}}{\partial \widetilde{\boldsymbol{h}}^{(\ell)}}, \Delta \widetilde{\boldsymbol{h}}^{(\ell)} \right\rangle\right|$
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{3}
10^{4}
n

既存手法との比較


$\ell = 1$
$\ell = 2$
$\ell = 3$
$\ell = 4$
$\ell = 5$
$\ell = 6$ (Output)
$\left|\left\langle \frac{\partial \mathcal{L}}{\partial \widetilde{\boldsymbol{h}}^{(\ell)}}, \Delta \widetilde{\boldsymbol{h}}^{(\ell)} \right\rangle\right|$
n
Dynamic Param. (Ours)
Spectral Param.
Slope: 0.02
Slope: -1.60
Slope: 0.02
Slope: -1.60
Slope: 0.02
Slope: -1.59
Slope: 0.02
Slope: -1.58
Slope: 0.02
Slope: -1.57
Slope: 0.02
Slope: -1.56

性能への影響


Train Loss
Dynamic Param. (Ours)
Spectral Param.
0.15
0.10
0.05
0.00
0
50
100
150
200
Epoch


Test Loss
Dynamic Param. (Ours)
Spectral Param.
0.20
0.18
0.16
0.14
0.12
0.10
0
50
100
150
200
Epoch

Outline

まとめ

- 学習が安定に進むための基準として
  $\|\Delta\widetilde{\boldsymbol{h}}^{(\ell)}\|_2 = \Theta(n_{\min}^r),\ \ r \in [0, 1/2]$ 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):

$$
\begin{aligned}
\boldsymbol{f}(\boldsymbol{x}) &= \frac{1}{\sqrt{n_{L-1}}} \boldsymbol{W}^{(L)} \boldsymbol{h}^{(L-1)}, \\
\widetilde{\boldsymbol{h}}^{(\ell)} &= \frac{n_{\min}^{r}}{\sqrt{n_{\ell-1}}} \boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)}, && 1 \leq \ell \leq L-1, \\
\boldsymbol{h}^{(\ell)} &= \psi(\widetilde{\boldsymbol{h}}^{(\ell)}), && 1 \leq \ell \leq L-1, \\
\widetilde{\boldsymbol{h}}^{(0)} &= \boldsymbol{x}.
\end{aligned}
$$

- 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，$\gamma_1 = \cdots = \gamma_{L-1} = \sqrt{2}$ & $\gamma_L = 1$