深層学習理論の代数的側面[*1]

理化学研究所／株式会社サイバーエージェント　園田　翔[*2]

Sho Sonoda

RIKEN / CyberAgent Inc

概要

本稿は，RIMS 共同研究における講演内容とその基礎にあるプレプリント [7] に基づき，深層学習の “深さ” を数学的にどう理解するかを，汎化誤差の観点から解説するものである．とくに，深さ依存性を実装に依らず議論するために，深層ネットワークを距離空間上の状態遷移モデル $\mathcal{H}_k = H \circ B(k, F)$ として定式化し，中間層が生成する半群の語球の増大度と汎化誤差との関係を説明する．前半では，概念空間・仮説空間・損失関数・汎化ギャップなどの基本概念を整理し，後半ではプレプリントの Table 1 と Section 4 の内容を中心に，Arzelà–Ascoli 型の飽和，冪零的状況における多項式増大，自由半群や ping–pong 型構成による指数増大を概観する．最後に，近似誤差と推定誤差の釣合いから現れる四つのレジーム EL/EP/PL/PP と最適深さについて述べ，“なぜ，いつ深い方が良いのか” という問いに対する現時点での数理的答えをまとめる．

1　はじめに

深層学習の成功を目の当たりにすると，“なぜ層を深くすると良いのか” という素朴な疑問が生じる．近似理論の側からは，関数合成を繰り返すことで表現力が増し，少ないパラメータで複雑な写像を表せることが期待される．しかし統計的学習理論の古典的な見方では，モデルを複雑にすれば過学習しやすくなるので，深さはむしろ汎化を悪化させる方向に働くはずである．この緊張関係が，いわゆる深層学習の汎化パラドックスである．

2010 年代以降，ReLU ネットワークを中心として，VC 次元，ノルム，マージン，圧縮率，PAC-Bayes，非パラメトリック回帰など多様な観点から深さ依存性の解析が進んだ．その結果，古典的な悲観論が与える指数依存よりはかなり穏やかな，線形・多項式・対数・定数といった依存性が得られる場合があることが分かってきた．ただし，その多くは個別の実装に強く依存しており，仮定と結論の対応が見えにくい．

本稿で紹介する最近の枠組み [7] の特徴は，ネットワークの実装を抽象化し，深さを “状

[*1] RIMS 共同研究（公開型）機械学習の数理的研究 2025 年 11 月
本研究は JSPS 科研費 24K21316, 25H01453, JST BOOST JPMJBY24E2, および JST CREST JPMJCR2015 の助成を受けたものです．

[*2] sho.sonoda@riken.jp

態遷移の合成回数” として捉えるところにある．この見方では，中間層の集合 F が生成する半群の語球 $B(k, F)$ の幾何学と代数が，推定誤差の深さ依存を支配する．とくに，

$$\text{自由度が高い（自由半群に近い）} \Longrightarrow \text{語球の増大が速い}$$
$$\text{自由度が低い（可換・冪零に近い）} \Longrightarrow \text{語球の増大が遅い}$$

という対応が，深層学習理論における “代数的側面” を端的に表している．

以下では，学習理論に馴染みの薄い読者を念頭に置いて，まず汎化の基本概念を整理する．その上で，状態遷移モデル，Rademacher 複雑度と被覆数による評価，そして語球の増大度の分類を説明する．証明の細部には立ち入らず，何が本質的な数学的対象なのかをできるだけ明確にすることを目的とする．

2　汎化とは何か

2.1　概念空間・仮説空間・標本空間

機械学習の基本的な登場人物は三つの空間と一つの写像で整理できる [5]．まず，データを生成する真の機構の属する空間を概念空間 $\mathcal{C}$ とする．概念空間の元 $c \in \mathcal{C}$ は未知であり，われわれはそこから得られた有限データしか観測できない．一方，学習機械の候補全体を仮説空間 $\mathcal{H}$ とする．標本空間を $\mathcal{S}$ と書けば，学習アルゴリズムとは

$$A : \mathcal{S} \to \mathcal{H}$$

という写像にほかならない．

この枠組みにおいて,「表現力」と「複雑度」の区別にも触れておく．表現力とは，仮説空間 $\mathcal{H}$ が概念空間 $\mathcal{C}$ をどれだけうまく近似できるかという “相対的な大きさ” であり，複雑度とは，$\mathcal{H}$ 自体がどれだけ大きい集合であるかという “絶対的な大きさ” である．浅いネットワークの普遍近似定理は表現力についての主張であり，VC 次元やパラメータ数，Rademacher 複雑度は複雑度についての主張である．この区別は，後で述べる近似誤差と推定誤差の分解にもつながる．

2.2　損失関数・汎化誤差・汎化ギャップ

以後，入力空間を可測空間 $\mathcal{X}$，出力空間を $\mathcal{Y} = \mathbb{R}$ とし，$(X, Y) \sim P$ を $\mathcal{X} \times \mathcal{Y}$ 上のデータ分布とする．損失関数 $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, b]$ を有界かつ第 1 変数に関して Lipschitz と仮定

すると，各仮説 $h : \mathcal{X} \to \mathbb{R}$ に対し

$$L[h] := \mathbb{E}_{(X,Y)\sim P}[\ell(h(X), Y)]$$

を期待損失（population risk）と呼ぶ．訓練標本 $S_n = ((X_i, Y_i))_{i=1}^n \sim P^n$ に対しては

$$\widehat{L}_n[h] := \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

を経験損失（empirical risk）と呼ぶ．

本当に知りたいのは $L[h]$ であるが，P は未知なので直接計算できない．そこで $\widehat{L}_n[h]$ を用いて近似する．学習済み仮説 $\widehat{h} = A(S_n)$ に対し，

$$L[\widehat{h}] - \widehat{L}_n[\widehat{h}]$$

を汎化ギャップと呼ぶ．これは “訓練時には良く見えたが，未知データではどれくらい悪くなるか” を表す最も基本的な量である．また，概念空間上で達成可能な最良のリスク

$$L_{\mathcal{C}} := \inf_{c\in\mathcal{C}} L[c]$$

を用いれば，余剰リスク $L[\widehat{h}] - L_{\mathcal{C}}$ も考えられる．両者はいずれも汎化の尺度であり，前者は訓練データとの乖離，後者は真の概念に対する近さを測っている．

2.3　なぜ深さが問題になるのか

古典的な一様収束型の議論では，複雑度が大きいほど汎化ギャップの上界は悪くなる．例えば VC 次元 d に基づく評価は，概形として

$$\mathrm{gen}(\mathcal{H}, n) \lesssim \sqrt{\frac{d \log n}{n}}$$

の形をとる．深いネットワークは一般にパラメータ数も VC 次元も大きいから，この種の評価だけを見ると，深さは不利にしか見えない．それにもかかわらず実際には，深いモデルほど高性能を示す場面が多い．これが深さを巡る理論上の困難である．

ここで重要なのは，“深い方が表現力は高い” と “深い方が複雑度は大きい” が同時に成り立つことである．前者は近似誤差を減らし，後者は推定誤差を増やす．したがって，深さの本当の役割を知るには，この二つを分けて評価しなければならない．

3　状態遷移として見た深層ネットワーク

3.1　抽象的な深さの定式化

[7] の出発点は，深層ネットワークの本質を実装ではなく関数合成に見ることである．状態空間を距離空間 $(\mathcal{X}, d)$ とし，出力層のクラスを $H \subset C(\mathcal{X})$，中間層のクラスを $F \subset C(\mathcal{X}, \mathcal{X})$ とする．このとき，深さ k の抽象モデルを

$$\mathcal{H}_k := H \circ B(k, F)$$

で定義する．ここで $B(k, F)$ は，F の元を高々 k 回合成して得られる写像全体，すなわち語球（word ball）である．より正確には，各 $g \in C(\mathcal{X}, \mathcal{X})$ に対し

$$|g|_{\mathcal{F}} := \min\{m \in \mathbb{N} \mid g = f_m \circ \cdots \circ f_1,\ f_j \in F\}$$

と語長を定め，

$$B(k, F) := \{g \in C(\mathcal{X}, \mathcal{X}) \mid |g|_{\mathcal{F}} \le k\}$$

と書く．

この定式化の利点はきわめて大きい．ReLU ネットワーク，畳み込みネットワーク，変換器，拡散モデル，test-time computation，chain-of-thought のような反復的推論まで，すべて “状態遷移の合成” として同じ型に載る．深さとは，この抽象的な遷移の合成回数である．

3.2　実装と抽象構造の分離

実際の学習では，抽象クラス $\mathcal{H}_k$ の元をそのまま実装できるわけではない．そこで [7] では，実装をもつネットワークのクラス $\mathcal{H}_{\mathrm{imp}}$ を別に用意し，学習を二段階に分ける．まず抽象クラス $\mathcal{H}_k$ 上で経験誤差最小化を行い，つぎにその解を実装クラスで近似する．

$$\widehat{h} = \mathrm{embed}\Big(\arg\min_{f \in \mathcal{H}_k} \widehat{L}_n[f]\Big).$$

この分離は，“推定誤差は深さそのものに由来するのか，それとも特定の実装に由来するのか” を見極めるための装置である．以下の定理は，推定誤差の主要項が抽象クラス $\mathcal{H}_k$ のみに依存することを示す．

定理 3.1 (バイアス–バリアンス分解の骨格)**.** 損失関数が有界かつ *Lipschitz* であるとする．抽象クラス $\mathcal{H}_k$ での最良近似誤差を $\varepsilon_{\mathrm{model}}$，実装クラスへの埋め込み誤差を $\varepsilon_{\mathrm{imp}}$ と書くと，高確率で

$$L[\widehat{h}] - L_{\mathcal{C}} \lesssim \varepsilon_{\mathrm{imp}} + \varepsilon_{\mathrm{model}} + \widehat{\mathfrak{R}}_S(\mathcal{H}_k) + \sqrt{\frac{1}{n}}$$

および

$$L[\widehat{h}] - \widehat{L}_n[\widehat{h}] \lesssim \varepsilon_{\mathrm{imp}} + \widehat{\mathfrak{R}}_S(\mathcal{H}_k) + \sqrt{\frac{1}{n}}$$

が成り立つ．

ここで $\widehat{\mathfrak{R}}_S(\mathcal{H}_k)$ は経験 Rademacher 複雑度である．定理の読み方は単純で，

$$\text{近似誤差} = \text{バイアス}, \qquad \text{Rademacher 複雑度} = \text{推定誤差の主要項}$$

である．特に，推定誤差の本体は抽象的な $\mathcal{H}_k$ によって決まり，実装 $\mathcal{H}_{\mathrm{imp}}$ には直接依存しない．これは “実装に依らない深さの理論” の出発点である．

3.3 Rademacher 複雑度の出力–中間分解

実数値関数族 $\mathcal{H}$ に対しては，Rademacher 複雑度は “ランダム符号 $\sigma_i \in \{\pm 1\}$ に対する最悪相関” として定義される：

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) := \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)\Big].$$

これは，モデルが無意味なランダムノイズにどれだけ適応できるかを測る量だと理解するとよい．複雑度が高いほど，ランダム符号にもよく適応してしまう．

問題は，中間層 $f : \mathcal{X} \to \mathcal{X}$ は $\mathcal{X}$ 値写像であり，そのままでは Rademacher 複雑度が定義できないことである．ここで Dudley のエントロピー積分が効く．$\mathcal{X}$ 値写像に対しても被覆数は定義できるので，中間層の複雑度をエントロピーで測ればよい．その結果，概形として次の分解が得られる：

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \le \widehat{\mathfrak{R}}_S(H) + \frac{12L}{\sqrt{n}} \int_0^{\mathrm{diam}(B(k,F))} \sqrt{\log N(B(k,F), d_S, \varepsilon)}\, d\varepsilon. \tag{1}$$

ここで d_S は標本 S が誘導する経験距離である．この式が示す本質は，出力層 H の寄与と中間層 F の寄与が分離していることである．以後の深さ依存性は，ほとんどすべて

$$N(B(k,F), d, \varepsilon)$$

表1　Section 4 の簡略化された対応表

	$\mathrm{Lip}F<1$（縮小）	$\mathrm{Lip}F=1$（等長）	$\mathrm{Lip}F>1$（拡大）
$\mathcal{X}$ コンパクト	飽和（P1）	飽和（P1）	多項式（P2）または指数（E2）
$\mathcal{X}$ 非コンパクト	飽和（P1’）	多項式（P2）または指数（E1）	超指数・二重指数（E3）

の k 依存性の問題に帰着される．

4　語球の増大度と代数的分類

4.1　なぜ代数が現れるのか

深さ k の増加に伴って語球 $B(k,F)$ がどれくらい大きくなるかは，関数合成の代数的性質に強く依存する．極端な場合を考えよう．もし異なる語が常に異なる写像を与えるなら，F は自由半群のように振る舞い，語球の大きさは指数的に増大する．反対に，可換性や冪零性があれば，異なる合成順序が同じ写像に潰れるため，語球の増大は大幅に抑えられる．

群論では，有限生成群の語球の多項式増大と冪零性が Gromov (1981) の定理で結び付いている．本稿で現れるのは群ではなく自己写像の半群であり，しかも生成系は典型的にはコンパクト無限集合である．それでもなお，“自由に近いほど速く増大し，冪零に近いほど遅く増大する” という原則は保たれる．

4.2　Section 4 の見取り図

プレプリントの Section 4 では，語球の被覆数 $N(B(k,F),d_\infty,\varepsilon)$ の k 依存性を支配する条件が整理されている．簡略化した見取り図を表 1 に示す．ここで d_∞ は一様距離であり，$d_S\le d_\infty$ なので，これを評価すれば (1) の右辺も制御できる．

この表はあくまで “見取り図” であり，実際の判定には分離条件や埋め込み条件など追加仮定が必要である．しかし本質は明快である．

1. コンパクトかつ非拡大的なら，Arzelà–Ascoli 型の議論により増大は飽和する．
2. 冪零 Lie 群への埋め込みがあれば，多項式増大に抑えられる．
3. 自由半群や ping–pong 型の作用があれば，指数増大が現れる．

4. 一様拡大が層ごとにエントロピーを増幅すれば，超指数・二重指数も起こりうる．

以下で順に見ていく．

4.3　P1 と P1’：Arzelà–Ascoli による飽和

まず最も安定な場合は，$\mathcal{X}$ がコンパクトで，生成された半群 $\langle F\rangle$ が非拡大的あるいは一様 Lipschitz である場合である．このとき写像族は同程度連続（equicontinuous）となり，Arzelà–Ascoli の定理により相対コンパクトとなる．したがって各固定スケール $\varepsilon>0$ に対し

$$N(B(k,F),d_\infty,\varepsilon)\le N(\overline{\langle F\rangle},d_\infty,\varepsilon)<\infty$$

であり，右辺は k に依存しない．すなわち被覆数は深くしても増えない．

縮小写像系がコンパクトアトラクタに収束する場合も同様である．典型例は Cantor 集合を生む反復関数系であり，十分深い語はアトラクタ上のほぼ定数写像になるため，深さ依存は固定スケールで消えてしまう．この意味で，縮小は表現力の観点では有利であっても，推定誤差の観点では極めて穏やかである．

4.4　P2：冪零性と多項式増大

次は，半群が冪零 Lie 群に双 Lipschitz に埋め込まれる場合である．アーベル群や Heisenberg 群が代表例であり，このとき語球の被覆数は

$$N(B(k,F),d_\infty,\varepsilon)\lesssim\left(1+\frac{k}{\varepsilon}\right)^{D}$$

と多項式的に増大する．指数が D で，これは Guivarc’h–Bass の斉次次元に対応する．群論に馴染みのある読者には，有限生成群に対する多項式増大と virtually nilpotent の関係の連続版だと考えると分かりやすい．

被覆数が多項式増大であれば，その対数は $\log k$ 程度なので，Dudley 積分を通した推定誤差は

$$\mathrm{var}(k,n)\lesssim\sqrt{\frac{\log k}{n}}$$

と対数増大になる．これは深さ依存としてはかなり弱い．ReLU ネットワークに対するノルム評価で見られる “深さにほとんど依らない” 振る舞いは，この側に属する現象とみなせる．

4.5　E1, E2, E3：自由性・ping–pong・一様拡大

これに対し，語の違いが写像の違いとしてはっきり観測できるとき，被覆数は指数増大する．E1 はその最も単純な形で，$F = \{f_1, \dots, f_r\}$ が自由半群を生成し，ある基点 x_* において異なる長さ k の語 u, v が

$$d(f_u(x_*), f_v(x_*)) \geq \delta$$

で一様に分離しているなら，

$$N(B(k, F), d_\infty, \varepsilon) \geq r^k \qquad (\varepsilon < \delta/2)$$

が従う．少なくとも一点，識別性の高い点を見つけることで，語の本数そのものが被覆数に転写できるのである．

E2 は ping–pong 条件である．互いに離れた chamber U_i と anchor a_i があり，各 f_i が自分の chamber 上では拡大しつつ全 chamber に到達し，外ではほぼ定数 a_i にリセットする．このとき語の右端で最初に食い違う場所が，最終出力に一様な差を生む．自由群の古典的な ping–pong 補題が，ここでは写像半群の被覆数下界として現れているわけである．

さらに E3 では，一様拡大写像 A が層ごとにエントロピーを増幅するため，超指数あるいは二重指数の増大すら可能である．これは一般の深層学習にとって“典型的”というよりは，最悪の場合の鋭さを示す結果として理解するとよい．

4.6　推定誤差への帰結

以上を (1) に戻すと，大雑把には次の対応が得られる：

$$N(B(k, F), d_\infty, \varepsilon) \sim \begin{cases} \text{定数または多項式} & \Longrightarrow \mathrm{var}(k, n) \lesssim \sqrt{\log k / n}, \\ \exp(ck) & \Longrightarrow \mathrm{var}(k, n) \lesssim \sqrt{k/n}, \end{cases}$$

すなわち，被覆数の指数増大は推定誤差の多項式増大を生み，被覆数の多項式増大は推定誤差の対数増大しか生まない．古典的な“深さに対する指数爆発”は，各層の Lipschitz 定数を逐次掛け合わせる粗い評価に由来するものであり，語球の幾何そのものを見れば，そこまで悲観的である必要はないことが分かる．

表2　四つのレジームと最適深さの概形

Regime	bias(k)	var(k,n)	最適深さ k^*	最適誤差の概形
EL	$e^{-\alpha k}$	$\sqrt{\log k/n}$	$\dfrac{\log n}{2\alpha}+\widetilde{o}(1)$	$\asymp\sqrt{\log\log n/n}$
EP	$e^{-\alpha k}$	$\sqrt{k^{\gamma}/n}$	$\dfrac{\log n}{2\alpha}+\widetilde{O}(1)$	$\asymp\sqrt{(\log n)^{\gamma}/n}$
PL	$k^{-\beta}$	$\sqrt{\log k/n}$	$\asymp(n/\log n)^{1/(2\beta)}$	$\asymp\sqrt{\log n/n}$
PP	$k^{-\beta}$	$\sqrt{k^{\gamma}/n}$	$\asymp n^{1/(2\beta+\gamma)}$	$\asymp n^{-\beta/(2\beta+\gamma)}$

5　最適深さと四つのレジーム

5.1　近似誤差と推定誤差の釣合い

深さを増すと，表現力の上昇により近似誤差は減少するが，仮説空間の増大により推定誤差は増加する．したがって最適な深さは，二つの釣合いとして定まる．[7] では，近似誤差が

$$\mathrm{bias}(k)=e^{-\alpha k}\qquad\text{または}\qquad\mathrm{bias}(k)=k^{-\beta}$$

のように減衰し，推定誤差が

$$\mathrm{var}(k,n)=\sqrt{\frac{\log k}{n}}\qquad\text{または}\qquad\mathrm{var}(k,n)=\sqrt{\frac{k^{\gamma}}{n}}$$

のように増大する四つの組合せを調べている．

表 2 から読み取れる最も重要なメッセージは三つある．第一に，どのレジームでも一般に $k^*>1$ であり，浅いモデルが最適とは限らない．これが深さ優位（depth supremacy）の意味である．第二に，データ数 n が増えるほど最適深さも増える．ビッグデータと深いモデルの相性の良さは，この意味で理論的にも自然である．第三に，最も有利なのは EL である．すなわち近似誤差は指数的に減るが，推定誤差は対数程度にしか増えない，という状況である．

5.2　レジームの解釈

EL が最も有利であるという事実は，近似側の構造がどれほど重要かを示している．推定誤差が多少悪くても，近似誤差が指数的に減れば全体として深さの恩恵は大きい．逆に PP

では，近似誤差も推定誤差も多項式級数でしか改善・悪化しないため，深くする効果は限定的になる．

この観点から見ると，深層学習が特に成功しやすいのは，データ生成機構そのものが反復的・階層的である場合だと言える．微分方程式の時間発展，拡散過程の反復的 denoising，段階的な推論過程などは，まさにそのような構造をもつ．

6　例と関連話題

6.1　縮小型 teacher–student 設定（EL）

最も分かりやすい例は，教師と生徒が同じ合成構造を共有する teacher–student 設定である．生徒を $\mathcal{H}_k = H \circ B(k, F)$，教師を無限深さ極限 $\mathcal{C} = \mathcal{H}_\infty := H \circ B(\infty, F)$ とし，各 $f \in F$ が $\mathrm{Lip}(f) \le \lambda < 1$ を満たすとする．このとき教師の “尻尾” を深さ k で打ち切ることにより，近似誤差は λ^k で指数減衰する．

一方，縮小性により語球の被覆数は固定スケールで飽和するから，推定誤差は対数増大よりもさらに良い振る舞いを示す．したがってこの状況は EL に属し，深さの効果が最大化される．

6.2　Neural Operator と反復スキーム

数値解析に近い例として，非線形放物型 PDE の解作用素を学習する Neural Operator を挙げられる．Furuya–Taniguchi–Okuda [3] の結果では，Picard 反復を一層の中間写像に対応させることで，深さに関して準指数的な近似誤差が得られる．Picard 反復は縮小写像であり，しかも単生成で可換的なので，語球の増大は多項式的に抑えられる．したがって推定誤差は対数増大であり，EL に近い非常に有利なレジームが期待される．

この例は，“連続時間発展を深さ方向に離散化する” という現代的見方をよく表している．Neural ODE や拡散モデルも同様に，深さが物理的・確率的時間に対応するため，反復構造そのものが概念空間側に組み込まれている．

6.3　Hölder/Sobolev/Besov 空間上の ReLU ネットワーク（PP/PL）

一方，概念空間を Hölder や Sobolev，Besov 空間のような古典的 smoothness class にとると状況は変わる．Jackson 型評価により，一般には近似誤差はパラメータ数や深さに

対して多項式的にしか減衰しない．Yarotsky の論法 [10] により ReLU ネットワークは（Jackson 型よりも速い）超収束（super-convergence）近似レートを達成できることが知られているが，それでも多項式減衰の域を出ない．

他方，推定誤差については VC 次元評価や圧縮評価により多項式増大が知られている [1, 2, 4, 8, 9]．したがって典型的には PP，より構造をうまく使える場合でも PL に属する．ここでは深さの利益は確かにあるが，teacher–student 型や反復型の問題ほど劇的ではない．

6.4　階層的合成クラスと PL

Schmidt–Hieber [6] の複合関数クラスは，深層学習が本質的に有利になる古典的な例である．概念空間そのものが Hölder 写像の合成でできているため，ReLU ネットワークはその階層性に適応してほぼ minimax 最適の収束率を達成する．この場合，近似誤差には合成構造に由来する指数的改善が部分的に現れ，推定誤差は被覆数評価により対数増大に抑えられる．純粋な意味での PL ではないにせよ，“階層性が深さの恩恵を増幅する” ことを示す代表例である．

7　おわりに

本稿では，深層学習理論の代数的側面を，中間層が生成する状態遷移半群の語球という対象を通して説明した．要点をまとめると次の通りである．

- 深層ネットワークを $\mathcal{H}_k = H \circ B(k, F)$ と抽象化すると，深さ依存性は実装から切り離して議論できる．
- 汎化誤差の主要な推定誤差項は，語球の被覆数のエントロピーによって支配される．
- 語球の増大度は，自由性・冪零性・拡大性といった代数的性質と密接に関係する．
- その結果，近似誤差と推定誤差の釣合いとして最適深さが現れ，一般に $k^* > 1$ となる．

深さの理論を “実装の細部” から “半群の幾何と代数” へと引き上げる視点は，現代の多様な AI モデルを横断的に捉えるうえで有力である．深層学習理論と幾何学，群論，力学系との接点は，今後さらに豊かな展開を見せるだろう．

今後の課題は多い．第一に，本稿の理論は経験誤差最小化と有界 Lipschitz 損失を前提としており，実際の最適化ダイナミクスや暗黙の正則化までは織り込んでいない．第二に，語

球の増大度を具体的な実装の幾何とどう結び付けるかは，依然として大きな問題である．第三に，拡散モデルや chain-of-thought のような反復的システムが本当に EL 型の有利なレジームに入るのかを，個別のモデルで厳密に示すことも今後の重要課題である．

参考文献

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, Stronger generalization bounds for deep nets via a compression approach, in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 254–263.

[2] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks, *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[3] T. Furuya, K. Taniguchi, and S. Okuda, Quantitative approximation for neural operators in nonlinear parabolic equations, in *The Thirteenth International Conference on Learning Representations*, 2025.

[4] N. Golowich, A. Rakhlin, and O. Shamir, Size-independent sample complexity of neural networks, *Information and Inference*, 9(2):473–504, 2020.

[5] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2nd edition, 2018.

[6] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with ReLU activation function, *The Annals of Statistics*, 48(4):1875–1897, 2020.

[7] S. Sonoda, Y. Hashimoto, I. Ishikawa, and M. Ikeda, Why and when deep is better than shallow: an implementation-agnostic state-transition view of depth supremacy, arXiv:2505.15064v3, 2025.

[8] T. Suzuki, Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality, in *International Conference on Learning Representations*, 2019.

[9] T. Suzuki, H. Abe, and T. Nishimura, Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network, in *International Conference on Learning Representations*, 2020.

[10] D. Yarotsky, Error bounds for approximations with deep ReLU networks, *Neural Networks*, 94:103–114, 2017.