距離空間上の深層汎化誤差評価[*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

本稿では，深層ネットワークを距離空間上の状態遷移モデルとして定式化し，その汎化誤差の深さ依存性を，語球 $B(k,F)$ の幾何から評価する枠組み [1] を解説する．中心となるのは，抽象仮説空間 $H_k = H \circ B(k,F)$ に対するバイアス–バリアンス分解と，Rademacher 複雑度の出力層–中間層分解である．前者により，実装誤差とモデル誤差はバイアス側に，深さ依存性は抽象クラス H_k のバリアンス（複雑度）側に分離される．後者により，深さ依存性は語球の被覆数の増大度に還元される．さらに，講演時に質問のあった“上界はどこまで鋭いか”という点に答えるため，Bernoulli 過程に対する Sudakov 型の下限評価を用いて，H が十分に高い表現力を持つときには，語球の指数増大から $\sqrt{k/n}$ 型の下限が従うことを示す．これにより，主定理の上界が少なくとも代表的な指数増大状況では本質的に鋭いことが分かる．

1　はじめに

深層学習の理論において，最も基本的な問いの一つは「なぜ深くすると良いのか」である．素朴には，合成により表現力が組合せ的に増大するため深さは有利に働くと期待される．しかし他方で，古典的な複雑度評価は，深さの増大が推定誤差を悪化させることも示唆する．したがって本質的な問題は，

> 近似誤差の減衰と推定誤差の増大の釣り合いが，深さ k に対してどのように決まるのか

を理解することである．

講演では，この問題を「距離空間上の状態遷移半群」の幾何へ還元する枠組み [1] を紹介した．そこでは入力空間を任意の距離空間 (X,d) とし，中間層を連続自己写像 $f\colon X \to X$，

[*1] RIMS 共同研究（公開型）関数空間論とその周辺 2025 年 12 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

出力層を実数値関数 $h\colon X \to \mathbb{R}$ とみなすことで，深さ k の抽象ネットワークを

$$H_k := H \circ B(k, F)$$

と定義する．ここで $F \subset C(X, X)$ は一層分の遷移のクラス，$B(k, F)$ は F の元を高々 k 回合成して得られる語球である．この枠組みでは，ReLU ネットワークやトランスフォーマーに限らず，反復推論や思考連鎖（chain-of-thought）のような逐次的情報処理も同じ数理構造で扱える．この点が，今回の報告の最大の利点である．

主結果は二段階である．第一に，経験誤差最小化に対する汎化誤差を，実装誤差・モデル誤差・Rademacher 複雑度に分解する（定理3.1）．第二に，その Rademacher 複雑度を出力層 H と語球 $B(k, F)$ のエントロピー積分に分解する（定理3.2）．これにより，深さ依存性は語球の被覆数 $\mathcal{N}(B(k, F), d_S, \varepsilon)$ の増大度へ還元される．これらはプレプリント [1] の Theorems 1,2 に対応する．

以下では，まず定式化を簡潔にまとめ，その後に主定理と証明スケッチを述べる．その後，講演時に出た「下からの評価はあるか」という質問に呼応して，Sudakov 型の下限評価を与え，適当な条件のもとで主結果が鋭いことを示す．

2　定式化

2.1　状態遷移モデルとしての深層ネットワーク

(X, d) を距離空間とし，$C(X)$ を連続実数値関数全体，$C(X, X)$ を連続自己写像全体とする．出力層のクラスを $H \subset C(X)$，一層の遷移のクラスを $F \subset C(X, X)$ と書く．

定義 2.1 (語球と深さ k の抽象ネットワーク)**.** $F^0 := \{\mathrm{Id}_X\}$,

$$F^\ell := \{f_\ell \circ \cdots \circ f_1 \mid f_j \in F\}, \qquad B(k, F) := \bigcup_{\ell=0}^{k} F^\ell$$

と定める．このとき

$$H_k := H \circ B(k, F) = \{h \circ f \mid h \in H,\ f \in B(k, F)\}$$

を深さ k の抽象ネットワークと呼ぶ．

この定義では，ネットワーク実装はまだ指定していない．実際の ReLU ネットワークや

CNN 等は，別の実装クラス H_{imp} と，抽象クラス H_k から H_{imp} への埋め込み

$$\mathrm{embed}\colon H_k \to H_{\mathrm{imp}}$$

によって扱う．ただし埋め込みの方法は問わない．一般にニューラルネットは高い表現力（稠密性，普遍近似性）をもつため，実装に伴う誤差は必要なだけ小さくとれるためです．この二段階構造により，「深さ」そのものに由来する効果と，「どの実装で近似したか」に由来する効果を分離できる．

2.2　学習アルゴリズムと汎化誤差

確率空間 $(X \times Y, P)$ 上で i.i.d. 標本

$$S_n = ((X_1, Y_1), \dots, (X_n, Y_n)) \sim P^n$$

を観測する．以下では $Y = \mathbb{R}$ とし，損失関数 $\ell\colon Y \times Y \to [0, b]$ は第 1 変数に関して 1-Lipschitz と仮定する．

集団損失と経験損失を

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

と書く．概念クラス C に対し

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

を Bayes risk の代わりに用いる．抽象クラス H_k 上で経験誤差最小化を行った解を

$$\hat{f} \in \arg\min_{f \in H_k} \widehat{L}_n[f]$$

とし，最終的な出力を $\hat{h} := \mathrm{embed}(\hat{f})$ とする．

汎化解析において本当に求めたい量は，$L[\hat{h}]$ である．ただし絶対値は情報が足りないため計算できないので，適当な基準から測った相対値を理論評価する．典型的には，余剰リスク $L[\hat{h}] - L_C$ と，汎化ギャップ $L[\hat{h}] - \widehat{L}_n[\hat{h}]$ である．これらはいずれも汎化誤差と呼ばれる．

2.3　Rademacher 複雑度と経験距離

独立な Rademacher 変数 $\sigma_1, \dots, \sigma_n \in \{\pm 1\}$ を導入し，実数値関数族 $\mathcal{G} \subset \mathbb{R}^X$ の経験 Rademacher 複雑度を

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) := \mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i)\right]$$

と定める．また，$h \in C(X)$ と $f, g \in C(X, X)$ に対し

$$\|h\|_S := \Big(\frac{1}{n}\sum_{i=1}^{n} |h(X_i)|^2\Big)^{1/2}, \qquad d_S(f,g) := \Big(\frac{1}{n}\sum_{i=1}^{n} d\big(f(X_i), g(X_i)\big)^2\Big)^{1/2}$$

を経験半ノルム・経験擬距離とする．

3　主結果

3.1　バイアス–バリアンス分解

まず，抽象クラス H_k と実装クラス H_{imp} を切り分けると，汎化誤差はバイアスとバリアンスに分解される．

定理 3.1 (バイアス–バリアンス分解)．ℓ は $[0,b]$ 値 *1-Lipschitz* 損失とする．

$$\varepsilon_{\mathrm{imp}} := \sup_{f \in H_k} \|f - \mathrm{embed}(f)\|_\infty, \qquad \varepsilon_{\mathrm{model}} := \inf_{h \in H_k} L[h] - L_C$$

とおく．このとき，任意の $\delta \in (0,1)$ に対して，確率少なくとも $1-\delta$ で

$$L[\hat{h}] - L_C \le \varepsilon_{\mathrm{imp}} + \varepsilon_{\mathrm{model}} + 4\widehat{\mathfrak{R}}_S(H_k) + 2b\sqrt{\frac{2\log(1/\delta)}{n}}, \tag{1}$$

$$L[\hat{h}] - \widehat{L}_n[\hat{h}] \le 2\varepsilon_{\mathrm{imp}} + 2\widehat{\mathfrak{R}}_S(H_k) + b\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{2}$$

ここで重要なのは，実装クラス H_{imp} の影響が $\varepsilon_{\mathrm{imp}}$ に押し込められ，バリアンス項が純粋に抽象クラス H_k の複雑度だけで決まることである．したがってこれ以降，深さ依存性の解析は $H_k = H \circ B(k,F)$ に集中して行えばよい．

3.2　出力層–中間層分解

次に，$\widehat{\mathfrak{R}}_S(H_k)$ を出力層 H と語球 $B(k,F)$ に分解する．

定理 3.2 (出力層–中間層分解)．$H \subset C(X)$，$F \subset C(X,X)$，$B_k := B(k,F)$ とする．各標本 $S = (X_i)_{i=1}^n$ に対し

$$Z_f(\sigma) := \sup_{h \in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h\big(f(X_i)\big), \qquad f \in B_k$$

と定める．また，$\mathrm{Id}_X \in F$ とし，中心化過程 $\widetilde{Z}_f := Z_f - Z_{\mathrm{Id}_X}$ が，ある定数 $L > 0$ に対して *subgaussian* 増分条件

$$\mathbb{E}_\sigma \exp\Big(\lambda(\widetilde{Z}_f - \widetilde{Z}_g)\Big) \le \exp\Big(\frac{\lambda^2 L^2 d_S(f,g)^2}{2n}\Big) \qquad (\lambda \in \mathbb{R}) \tag{3}$$

を満たすと仮定する．このとき普遍定数 $C > 0$ が存在して

$$\widehat{\mathfrak{R}}_S(H_k) \le \widehat{\mathfrak{R}}_S(H) + C\frac{L}{\sqrt{n}} \int_0^{\mathrm{diam}(B_k)} \sqrt{\log \mathcal{N}(B_k, d_S, \varepsilon)}\, d\varepsilon \tag{4}$$

が成り立つ．ここで $\mathrm{diam}(B_k) := \sup_{f \in B_k} d_S(f, \mathrm{Id}_X)/2$ である．

注意 3.3. 例えば H が RKHS の単位球で，対応する特徴量写像が d に関して Lipschitz であれば，Hilbert 値 Hoeffding 型不等式により (3) を検証しやすい．

式 (4) により，中間層の寄与は語球のエントロピー積分だけで測られる．例えば，

$$\mathcal{N}(B(k,F), d_S, \varepsilon) \lesssim (1 + k/\varepsilon)^D$$

なら $\widehat{\mathfrak{R}}_S(H_k) \lesssim \widehat{\mathfrak{R}}_S(H) + \sqrt{\log k/n}$ となり，一方

$$\mathcal{N}(B(k,F), d_S, \varepsilon_0) \gtrsim e^{ck}$$

なら上界だけからでも $\sqrt{k/n}$ 型の増大が予想される．以下ではこれを詳しく説明する．

4　定理3.1の証明スケッチ

まず，Rademacher 複雑度の標準的な使い方を復習する．

補題 4.1 (一様偏差評価)**.** $\mathcal{G} \subset \mathbb{R}^X$ を可測関数族とし，各 $g \in \mathcal{G}$ は $|g| \le b$ を満たすとする．このとき任意の $\delta \in (0,1)$ に対して，確率少なくとも $1-\delta$ で

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}g - \frac{1}{n}\sum_{i=1}^n g(X_i) \right| \le 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + b\sqrt{\frac{2\log(1/\delta)}{n}}$$

が成り立つ．

これは symmetrization と McDiarmid の不等式の組合せである．本稿では $\mathcal{G} = \ell \circ H_k$ に適用する．損失関数が第 1 変数に関して 1-Lipschitz なので，contraction lemma から

$$\widehat{\mathfrak{R}}_S(\ell \circ H_k) \le \widehat{\mathfrak{R}}_S(H_k)$$

が従う．

定理 3.1 の証明スケッチ． $f^\star \in \arg\min_{f\in H_k} L[f]$，$\hat{f} \in \arg\min_{f\in H_k} \widehat{L}_n[f]$ と書く．最終出力は $\hat{h} = \mathrm{embed}(\hat{f})$ である．まず余剰リスクについて

$$L[\hat{h}] - L_C = \underbrace{L[\hat{h}] - L[\hat{f}]}_{\le \varepsilon_{\mathrm{imp}}} + \underbrace{L[\hat{f}] - L[f^\star]}_{\text{推定誤差}} + \underbrace{L[f^\star] - L_C}_{=\varepsilon_{\mathrm{model}}}.$$

ここで中央項は

$$\begin{aligned} L[\hat{f}] - L[f^\star] &= \big(L[\hat{f}] - \widehat{L}_n[\hat{f}]\big) + \big(\widehat{L}_n[\hat{f}] - \widehat{L}_n[f^\star]\big) + \big(\widehat{L}_n[f^\star] - L[f^\star]\big) \\ &\le 2 \sup_{f\in H_k} \left| L[f] - \widehat{L}_n[f] \right| \end{aligned}$$

と評価される．補題 4.1 を $\ell \circ H_k$ に適用すれば

$$L[\hat{f}] - L[f^\star] \le 4\widehat{\mathfrak{R}}_S(H_k) + 2b\sqrt{\frac{2\log(1/\delta)}{n}}$$

となり，(1) が従う．

同様に汎化ギャップについては

$$\begin{aligned} L[\hat{h}] - \widehat{L}_n[\hat{h}] &= \big(L[\hat{h}] - L[\hat{f}]\big) + \big(L[\hat{f}] - \widehat{L}_n[\hat{f}]\big) + \big(\widehat{L}_n[\hat{f}] - \widehat{L}_n[\hat{h}]\big) \\ &\le 2\varepsilon_{\mathrm{imp}} + \sup_{f\in H_k} \left| L[f] - \widehat{L}_n[f] \right|, \end{aligned}$$

したがって補題 4.1 から (2) が従う． □

この議論の要点は，最適化アルゴリズムの詳細を使わず，ERM の最小性だけで証明が閉じることである．そのため，抽象クラス H_k と具体的実装 H_{imp} を分けて解析できる．

5 定理3.2の証明スケッチ

X 値写像族 $B_k \subset C(X, X)$ に対する Dudley 積分を導出する．

ステップ 1：アンカーで中心化する． Z_f の定義から

$$\widehat{\mathfrak{R}}_S(H_k) = \mathbb{E}_\sigma \sup_{f\in B_k} Z_f.$$

$\mathrm{Id}_X \in B_k$ なので，$\widetilde{Z}_f := Z_f - Z_{\mathrm{Id}_X}$ とおけば

$$\widehat{\mathfrak{R}}_S(H_k) \le \mathbb{E}_\sigma Z_{\mathrm{Id}_X} + \mathbb{E}_\sigma \sup_{f\in B_k} \widetilde{Z}_f = \widehat{\mathfrak{R}}_S(H) + \mathbb{E}_\sigma \sup_{f\in B_k} \widetilde{Z}_f.$$

したがって問題は中心化過程 $\{\widetilde{Z}_f\}_{f\in B_k}$ の上界へ帰着する．

ステップ 2: subgaussian increment を仮定する．定理の仮定 (3) は，擬距離

$$\rho_S(f,g) := \frac{L}{\sqrt{n}} d_S(f,g)$$

に関して $\widetilde{Z}_f$ 中心化 subgaussian 過程であることを意味する．したがって Dudley 型評価（see, e.g. [3]）を適用すると，普遍定数 $C>0$ が存在して

$$\mathbb{E}_\sigma \sup_{f\in B_k} \widetilde{Z}_f \le C \int_0^{\mathrm{diam}(B_k;\rho_S)} \sqrt{\log \mathcal{N}(B_k,\rho_S,\varepsilon)}\, d\varepsilon.$$

変数変換によりこれがちょうど

$$C\frac{L}{\sqrt{n}} \int_0^{\mathrm{diam}(B_k)} \sqrt{\log \mathcal{N}(B_k,d_S,\varepsilon)}\, d\varepsilon$$

となり，(4) が従う． □

注意 5.1 (なぜこの形が便利か)．通常の深層汎化解析では，B_k としてベクトル空間上に埋め込まれたデータを変換する具体的なネットワークを想定する．しかし，データをベクトル空間に埋め込むと，データの幾何構造が損なわれ，得られた評価と幾何構造の関係が見えにくくなる．本定理は「出力層の複雑度」だけを Rademacher 複雑度として残し，中間層側は語球の被覆数だけで評価する．このため，深さ依存性の解析は半群 $\langle F\rangle$ の幾何へ直接還元される．

6 Sudakov 型下限評価

H が十分に表現力を持ち，語球 B_k の離散構造を実数値関数として読み出せるならば，Rademacher 複雑度にも対応する下限が現れることを説明する．

6.1 Bernoulli 過程に対する Sudakov 型下限

まず，実数値関数族そのものに対する下限を述べる．

命題 6.1 (Bernoulli–Sudakov 型下限)**.** $\mathcal{G}\subset C(X)$ を $\|g\|_\infty \le B$ を満たす関数族とする．

このとき普遍定数 $c>0$ が存在して，任意の標本 $S=(X_i)_{i=1}^n$ と任意の $\varepsilon>0$ に対し

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) \ge c\min\left\{\varepsilon\sqrt{\frac{\log \mathcal{M}(\mathcal{G},\|\cdot\|_S,2\varepsilon)}{n}},\ \frac{\varepsilon^2}{B}\right\}. \tag{5}$$

特に *packing-covering duality* により

$$\mathcal{M}(\mathcal{G},\|\cdot\|_S,2\varepsilon) \ge \mathcal{N}(\mathcal{G},\|\cdot\|_S,2\varepsilon)$$

なので，右辺の対数項を被覆数で書いてもよい．

Proof. $M:=\mathcal{M}(\mathcal{G},\|\cdot\|_S,2\varepsilon)$ とし，2ε-packing $g_1,\dots,g_M\in\mathcal{G}$ を取る．各 j に対し

$$t_j:=\frac{1}{n}\big(g_j(X_1),\dots,g_j(X_n)\big)\in\mathbb{R}^n$$

と置く．すると，$j\neq\ell$ なら

$$\|t_j-t_\ell\|_{\ell_2}=\frac{1}{n}\Big(\sum_{i=1}^n |g_j(X_i)-g_\ell(X_i)|^2\Big)^{1/2}=\frac{1}{\sqrt{n}}\,\|g_j-g_\ell\|_S\ge\frac{2\varepsilon}{\sqrt{n}}.$$

また $\|t_j\|_{\ell_\infty}\le B/n$ である．したがって Bernoulli 過程に対する Sudakov minoration（[2], Theorem 6.4.1）を適用すると

$$\mathbb{E}_\sigma\sup_{1\le j\le M}\sum_{i=1}^n\sigma_i t_{j,i}\ge c\min\Big\{\frac{2\varepsilon}{\sqrt{n}}\sqrt{\log M},\ \frac{(2\varepsilon/\sqrt{n})^2}{B/n}\Big\}.$$

左辺は

$$\mathbb{E}_\sigma\sup_{1\le j\le M}\frac{1}{n}\sum_{i=1}^n\sigma_i g_j(X_i)\le\widehat{\mathfrak{R}}_S(\mathcal{G})$$

であるから，定数を吸収すれば (5) を得る． □

注意 6.2. これは Gauss 過程に対する古典的 Sudakov minoration の Bernoulli 版であり，Dudley 積分による上界の “逆向き” に当たる．式 (5) の第 1 項が通常の $\varepsilon\sqrt{\log M/n}$ 型で，第 2 項は Bernoulli 過程特有の ℓ_∞ 制御に由来する．深さ依存性を見る限り，第 1 項が本質的である．

6.2　語球から合成クラスへの下限の転換

次に，B_k の packing を $H_k = H \circ B_k$ へ移す．

定理 6.3 (表現力の高い H に対する下限の転換)**.** 標本 $S = (X_i)_{i=1}^n$ を固定する．各 k に対して写像

$$\Psi_k \colon B_k \to H_k$$

が存在し，ある定数 $\kappa, B > 0$ について

$$\|\Psi_k(f) - \Psi_k(g)\|_S \geq \kappa\, d_S(f,g), \tag{6}$$

$$\|\Psi_k(f)\|_\infty \leq B \qquad (f \in B_k) \tag{7}$$

を満たすと仮定する．このとき普遍定数 $c > 0$ が存在して，任意の $\varepsilon > 0$ に対し

$$\widehat{\mathfrak{R}}_S(H_k) \geq c \min\left\{ \kappa\varepsilon \sqrt{\frac{\log \mathcal{M}(B_k, d_S, 2\varepsilon)}{n}},\, \frac{\kappa^2\varepsilon^2}{B} \right\}. \tag{8}$$

したがって，$\mathcal{M}(B_k, d_S, 2\varepsilon_0) \geq e^{\alpha k}$ が成り立つなら

$$\widehat{\mathfrak{R}}_S(H_k) \geq c \min\left\{ \kappa\varepsilon_0 \sqrt{\alpha k/n},\, \kappa^2\varepsilon_0^2/B \right\}. \tag{9}$$

Proof. B_k の 2ε-packing $f_1, \dots, f_M$ を取る．仮定 (6) により

$$\|\Psi_k(f_j) - \Psi_k(f_\ell)\|_S \geq \kappa d_S(f_j, f_\ell) \geq 2\kappa\varepsilon \qquad (j \neq \ell)$$

だから，$\Psi_k(f_1), \dots, \Psi_k(f_M)$ は H_k の $2\kappa\varepsilon$-packing である．さらに (7) により一様有界性があるので，命題 6.1 を $\mathcal{G} = \{\Psi_k(f_j)\}_{j=1}^M$ に適用すると

$$\widehat{\mathfrak{R}}_S(H_k) \geq c \min\left\{ \kappa\varepsilon \sqrt{\frac{\log M}{n}},\, \frac{\kappa^2\varepsilon^2}{B} \right\}$$

を得る．$M = \mathcal{M}(B_k, d_S, 2\varepsilon)$ とすれば (8)，さらに $M \geq e^{\alpha k}$ を代入すれば (9) が従う． □

注意 6.4 (“H が十分に表現力を持つ” とは何か)**.** 定理 6.3 の仮定は，B_k を経験距離 d_S の意味で，合成クラス H_k の中へ下から Lipschitz に埋め込めることを要求している．$\Psi_k(f) = h_f \circ f$ と書けば，H は各 f ごとに適切な readout h_f を選ぶことで，語球の離散構造を実数値関数として読み出せる，という意味になる．これが本稿における “H が十分に高い表現力を持つ” という条件である．

6.3　具体的な十分条件：有限集合上の符号化

上の仮定はやや抽象的であるが，有限集合上の Lipschitz データを延長できるクラスでは比較的具体的に検証できる．

命題 6.5 (有限符号化による十分条件)．標本 $S = (X_i)_{i=1}^{n}$ を固定する．H が次の有限延長性を持つと仮定する：任意の有限集合 $A \subset X$ と任意の 1-*Lipschitz* 関数 $u\colon A \to [-B, B]$ に対し，その拡張 $h \in H$ が存在する．

さらに，$f_1, \dots, f_M \in B_k$ が存在して，各

$$A_j := \{f_j(X_1), \dots, f_j(X_n)\} \subset X$$

が互いに δ-分離しているとする，すなわち

$$\inf\{d(a, a') \mid a \in A_j,\ a' \in A_\ell\} \geq \delta \qquad (j \neq \ell).$$

このとき，ある $h \in H$ が存在して，$h \circ f_1, \dots, h \circ f_M$ は $\|\cdot\|_S$ に関して *pairwise* に $c_0\delta$-分離する．ただし $c_0 > 0$ は普遍定数である．

証明のアイデア．Gilbert–Varshamov 型の符号化により，長さ n の二値ベクトル

$$\omega^{(1)}, \dots, \omega^{(M)} \in \{0, \delta/2\}^n$$

を互いに Hamming 距離 $\geq n/8$ で取れる（$M \leq e^{cn}$ の範囲で十分）．和集合 $A := \bigcup_{j=1}^{M} A_j$ 上で

$$u\big(f_j(X_i)\big) := \omega_i^{(j)}$$

と定める．異なる群の間では距離が少なくとも δ あり，値の差は高々 $\delta/2$ なので，u は 1-Lipschitz である．仮定によりこれを拡張する $h \in H$ が存在する．すると Hamming 距離の条件から

$$\|h \circ f_j - h \circ f_\ell\|_S \geq c_0\delta \qquad (j \neq \ell)$$

が従う． □

この命題は，H が有限集合上のデータをほぼ自由に読み出せるなら，語球の packing をそのまま H_k の packing に持ち上げられることを示している．定理 6.3 の仮定 (6) を検証する一つの方法とみなせる．

表1　近似誤差と推定誤差の釣り合いから得られる典型的 regime

Regime	近似誤差 $\mathrm{bias}(k)$	推定誤差 $\mathrm{var}(k,n)$	典型的最適深さ k^*
EL	$e^{-\alpha k}$	$\sqrt{\log k/n}$	$\frac{1}{2\alpha}\log n$
EP	$e^{-\alpha k}$	$\sqrt{k^\gamma/n}$	$\frac{1}{2\alpha}\log n$
PL	$k^{-\beta}$	$\sqrt{\log k/n}$	$\big(n/\log n\big)^{1/(2\beta)}$
PP	$k^{-\beta}$	$\sqrt{k^\gamma/n}$	$n^{1/(2\beta+\gamma)}$

6.4　何が sharp になるのか

（下限を吟味するため，あえて大きい方を考える．）例えば条件 E1/E2 では，語球の被覆数は指数増大する：

$$\mathcal{N}(B_k, d_\infty, \varepsilon_0) \gtrsim r^k.$$

packing-covering duality と $d_S \le d_\infty$ を併用すれば，適切な標本に対して

$$\mathcal{M}(B_k, d_S, c\varepsilon_0) \gtrsim r^k$$

が期待される．このとき定理 6.3 は

$$\widehat{\mathfrak{R}}_S(H_k) \gtrsim \min\Big\{\sqrt{k/n}, 1\Big\}$$

型の下限を与える．したがって，定理 3.2 から得られる上界

$$\widehat{\mathfrak{R}}_S(H_k) \lesssim \sqrt{k/n}$$

は，少なくとも指数増大型の状況では本質的に sharp である．

7　深さ依存性と最適深さ

以上をまとめると，深さ依存性は語球のエントロピー増大度で決まり，それに応じて推定誤差の成長も決まる．代表的な 4 つの regime を表 1 に再掲する．

ここで EL は「指数近似 + 対数分散」であり，深いモデルを用いることの恩恵が最も大きい．例えば，収縮的 teacher–student 設定や反復的・階層的概念クラスがこれに対応する．一方，Hölder/Besov 空間のような古典的な概念空間 C では，Jackson 型下限のため近似誤差は通常多項式オーダーにとどまり，PL または PP に落ちる．

今回追加した下限評価の観点から言えば，EP/PP のように語球のエントロピーが指数増大する場合，$\sqrt{k/n}$ より良い一般上界は期待しにくい．逆に，P2 のような nilpotent growth では被覆数が多項式であり，Dudley 積分から $\sqrt{\log k/n}$ しか現れない．この差こそが「深さを深くしてもよい状況」と「深さが統計的に悪影響を及ぼす状況」を分けている．

8　おわりに

本稿では，距離空間上の状態遷移モデルという実装非依存の枠組みで，深層ネットワークの汎化誤差評価を解説した．

この視点では，深さ依存性はアーキテクチャ固有の細部ではなく，語球の増大度という半群の幾何へ還元される．関数空間論・調和解析・幾何群論的な展開として，次の方向性が考えられる:

- RKHS・ウェーブレット辞書・Besov 空間のような具体的関数空間で，仮定 (3) や定理 6.3 の lower-Lipschitz 埋め込みを具体化する．
- 半群の語球増大度と概念クラスの近似理論を組み合わせ，EL/EP/PL/PP の 4 regime を実際のモデルとして実現する．

参考文献

[1] 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.

[2] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes*, 2nd ed., Springer, 2021.

[3] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.