クープマン作用素を用いた深層学習モデルの汎化誤差解析 *

NTT 株式会社 † 橋本　悠香
Yuka Hashimoto, NTT, Inc.

1　はじめに

深層ニューラルネットワークの汎化特性を理解することは，機械学習分野における最大の課題の一つである．汎化とは，モデルが未知のデータに対してどの程度適合できるかを表す．古典的には，汎化誤差は VC 次元 [7, 1] を用いて上から抑えられる．また，ノルムに基づく上界 [13, 3, 6, 12, 15, 11, 10, 4] や圧縮に基づく上界 [2, 14] も研究されてきた．ノルムに基づく上界は重み行列の行列 (p, q) ノルムに依存し，圧縮に基づく上界はネットワークをどの程度圧縮できるかを調べることで導出される．これらの上界は，低ランクの重み行列や特異値が小さい（すなわちほぼ低ランクな）重み行列が汎化に好影響を与えることを示唆している．

一方，高ランクで特異値が大きい重み行列を持つモデルがよく汎化するという現象が，経験的に観察されている [5]．ノルムに基づく上界や圧縮に基づく上界は低ランクおよびほぼ低ランクの場合のみに着目しているため，これらの現象を説明することができない．この現象を理論的に説明するために，クープマンに基づく上界が提案された [9]．クープマン作用素は非線形関数の合成を表す線形作用素であり，ニューラルネットワークの本質的な構造である合成の解析に重要な役割を果たす．この既存の上界は，各重み行列のノル

* 本論文の内容は，文献 [8] に基づく．
† 〒180-8585 東京都武蔵野市緑町

ムと行列式の比として以下のように表される：

$$O\Big(\prod_{l=1}^{L}\frac{G_l\|K_{\sigma_l}\|_{H_l}\|W_l\|^{s_{l-1}}}{\sqrt{S}\det(W_l^*W_l)^{1/4}}\Big). \tag{1}$$

ここで，S はサンプルサイズ，s_l は第 l 層の滑らかさを表し，G_l は第 $l\sim L$ 層によって決まる定数，K_{σ_l} は活性化関数 σ_l に関するクープマン作用素，$\|\cdot\|_{H_l}$ はソボレフ空間 H_l における作用素ノルムを表す．行列式の項が上界の分母に現れるため，重み行列が高ランクで特異値が大きくても，この上界は小さくなりうる．クープマン作用素に基づく上界は，高ランクの重み行列を持つニューラルネットワークがなぜ汎化するのかを理論的に説明している．

しかし，クープマン作用素に基づく上界の既存の解析は，モデルの滑らかさとデータ空間の非有界性に強く依存しており，有界なデータ空間を持つ現実的なモデルや，hyperbolic tangent，sigmoid，ReLU 型の非滑らかな活性化関数を使用するモデルは対象外となる．加えて，上界の活性化関数への依存性が明確でない．実際，上界 (1) における $\|K_{\sigma_l}\|_{H_l}$ および G_l の評価は多くの場合困難である．

本論文では，既存のクープマンに基づく上界の問題点を解決する新たな上界を提案する．提案する上界は以下のように表される：

$$O\Big(\prod_{l=1}^{L}\frac{G_l\|K_{\sigma_l}\|_{\mathcal{L}_l}}{\sqrt{S}\det(W_l^*W_l)^{1/4}}\Big).$$

ここで，$\|\cdot\|_{\mathcal{L}_l}$ は L^2 関数空間における作用素ノルムを表す．既存のクープマンに基づく上界と同様に，提案する上界は高ランクのニューラルネットワークがなぜ汎化するのかを説明する．一方，提案する上界は，関数空間 $\mathcal{L}_l$ が H_l と異なることが大きな利点である．$\mathcal{L}_l$ は H_l より大きな空間であり，$\mathcal{L}_l$ によって滑らかでない深層モデルや有界なデータ空間の解析が可能となる．さらに，$\|K_{\sigma_l}\|_{\mathcal{L}_l}$ および G_l の評価が容易になり（補題 2.3–2.5 参照），活性化関数が深層モデルに与える影響を理解できる．結果として，提案する上界は，より広範なモデルへの適用と深いモデル理解を可能にするという点で，既存の上界を大幅に改善するものである．

■記法　$d\in\mathbb{N}$ およびルベーグ測度空間 $\mathcal{X}\subseteq\mathbb{R}^d$ に対し，$L^2(\mathcal{X})$ を $\mathcal{X}$ 上の複素数値二乗ルベーグ可積分関数の空間とする．$\mathcal{X}$ 上のルベーグ測度を $\mu_{\mathcal{X}}$ と表す．ヒルベルト空間 $\mathcal{H}$ に対し，$\langle\cdot,\cdot\rangle_{\mathcal{H}}$ を $\mathcal{H}$ における内積とする．文脈から明らかな場合は添え字 $\mathcal{H}$ を省略する．$\mathcal{H}_1$ から $\mathcal{H}_2$ への有界線形作用素の空間を $B(\mathcal{H}_1,\mathcal{H}_2)$ と表す．特に，$B(\mathcal{H},\mathcal{H})=B(\mathcal{H})$ とする．

2 準備

2.1 クープマン作用素

クープマン作用素は，非線形関数の合成を表す線形作用素である．ニューラルネットワークは合成から構成されるため，クープマン作用素はニューラルネットワークの解析に本質的な役割を果たす．$\mathcal{X} \subseteq \mathbb{R}^d$ をルベーグ測度空間とする．クープマン作用素は以下のように定義される．

定義 2.1 (クープマン作用素)**.** $\tilde{\mathcal{X}} \subseteq \mathbb{R}^{d_1}$ および $\mathcal{X} \subseteq \mathbb{R}^{d_2}$ とする．写像 $\sigma : \tilde{\mathcal{X}} \to \mathcal{X}$ に関するクープマン作用素 K_σ は，$L^2(\mathcal{X})$ から $L^2(\tilde{\mathcal{X}})$ への線形作用素であり，$h \in L^2(\tilde{\mathcal{X}})$ に対して $K_\sigma h(x) = h(\sigma(x))$ と定義される．

本論文では，活性化関数に関するクープマン作用素を考える．全体を通して，これらのクープマン作用素が有界であることを仮定する．

仮定 2.2 (クープマン作用素の有界性)**.** 写像 σ に関するクープマン作用素 K_σ は有界である，すなわち $\|K_\sigma\| = \sup_{\|h\|=1} \|K_\sigma h\|$ として定義される作用素ノルムが有限である．

実際，クープマン作用素の有界性に関する十分条件について，次の補題が成り立つ．

補題 2.3. $\sigma : \tilde{\mathcal{X}} \to \mathcal{X}$ が全単射で，σ^{-1} が微分可能であり，σ^{-1} のヤコビアンが $\mathcal{X}$ 上で有界であると仮定する．このとき，$\|K_\sigma\| \leq \sup_{x \in \mathcal{X}} |J\sigma^{-1}(x)|^{1/2}$ が成り立つ．ここで $J\sigma^{-1}$ は σ^{-1} のヤコビアンである．特に，クープマン作用素 K_σ は有界である．

次の補題は，$\sigma([x_1, \ldots, x_d]) = [\tilde{\sigma}(x_1), \cdots, \tilde{\sigma}(x_d)]$（$\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$）として定義される，よく知られた要素ごとの活性化関数の有界性に関するものである．

補題 2.4. $\tilde{\mathcal{X}} = [a_1, b_1] \times \cdots \times [a_d, b_d] \subseteq \mathbb{R}^d$ を有界矩形領域，$\mathcal{X} = \sigma(\tilde{\mathcal{X}})$ とする．σ が $\tilde{\sigma}(x) = \tanh(x)$ として定義される要素ごとの hyperbolic tangent であるとき，$\mathcal{X} \subset [-1, 1]^d$ および $\|K_\sigma\| \leq (\prod_{i=1}^d \sup_{x \in \tilde{\sigma}([a_i, b_i])} 1/(1 - x^2))^{1/2}$ が成り立つ．σ が $\tilde{\sigma}(x) = 1/(1 + \mathrm{e}^{-x})$ として定義される要素ごとの sigmoid であるとき，$\mathcal{X} \subset [-1, 1]^d$ および $\|K_\sigma\| \leq (\prod_{i=1}^d \sup_{x \in \tilde{\sigma}([a_i, b_i])} 1/(x - x^2))^{1/2}$ が成り立つ．

σ が微分可能でなくても，クープマン作用素は有界であり，その上界を評価できる場合もある．

補題 2.5. $\tilde{\mathcal{X}} = \mathcal{X} = \mathbb{R}^d$ とする．σ を，$a > 0$ として $x \leq 0$ のとき $\tilde{\sigma}(x) = ax$，$x > 0$ のとき $\tilde{\sigma}(x) = x$ と定義される要素ごとの Leaky ReLU とする．このとき，$\|K_\sigma\| \leq \max\{1, 1/a^d\}^{1/2}$ が成り立つ．

3　クープマン作用素を用いた深層学習モデルの汎化誤差解析

3.1　クープマン作用素による深層学習モデルの代数的表現

ヒルベルト空間 $\mathcal{H}_0, \ldots, \mathcal{H}_{L-1}, \tilde{\mathcal{H}}_1, \ldots, \tilde{\mathcal{H}}_L$ を考える．Θ_l をパラメータの集合，$l = 1, \ldots, L$ に対して $\eta_l : \Theta_l \to B(\tilde{\mathcal{H}}_l, \mathcal{H}_{l-1})$ とする．また，$v \in \tilde{\mathcal{H}}_L$ および $A_l \in B(\mathcal{H}_l, \tilde{\mathcal{H}}_l)$ を固定する．$\theta_l \in \Theta_l$（$l = 1, \ldots, L$）に対して，次のモデルを考える：

$$f(\theta_1, \ldots, \theta_L) = \eta_1(\theta_1) A_1 \eta_2(\theta_2) \cdots A_{L-1} \eta_L(\theta_L) v. \tag{2}$$

モデル (2) を直接考える代わりに，データ空間 $\mathcal{X}_0$ 上でパラメータ $c > 0$ を持つ次の正則化モデルを考える：

$$F_c(\theta_1, \ldots, \theta_L, x) = \langle \eta_1(\theta_1) A_1 \eta_2(\theta_2) \cdots A_{L-1} \eta_L(\theta_L) v, p_{c,x} \rangle_{\mathcal{H}_0}. \tag{3}$$

ここで，$x \in \mathcal{X}_0$ および $c > 0$ に対して $p_{c,x} \in \mathcal{H}_0$ であり，$E(c) > 0$ に対して $\|p_{c,x}\| \leq E(c)$ とする．この正則化は，ラデマッハ複雑度の上界を導出するために必要である．しかし，次の例に示すように，正則化モデル (3) は元のモデル (2) を十分に近似する．

例 3.1. $\mathcal{X}_0 = \mathbb{R}^d$，$\mathcal{H}_0 = L^2(\mathbb{R}^d)$ とする．$c > 0$ および $x \in \mathcal{X}_0$ に対して $p_{c,x}(y) = (c/\pi)^{d/2} \mathrm{e}^{-c\|y-x\|^2}$ とする．このとき，$p_{c,x} \in L^2(\mathbb{R}^d)$ および $\|p_{c,x}\|^2 = (2c/\pi)^{d/2}$ である．$c \to \infty$ のとき $p_{c,x}$ は x を中心とするディラックのデルタ関数に近づくため，任意の $x \in \mathbb{R}^d$ に対して $\lim_{c\to\infty} F_c(\theta_1, \ldots, \theta_L, x) = f(\theta_1, \ldots, \theta_L)(x)$ が成り立つ．したがって，c が十分大きければ，$F_c(\theta_1, \ldots, \theta_L, x)$ は $f(\theta_1, \ldots, \theta_L)(x)$ をよく近似する．

3.2　単射な重み行列を持つニューラルネットワーク

$d_l \in \mathbb{N}$，$\Theta_l = \{W \in \mathbb{C}^{d_l \times d_{l-1}} \mid W \text{ は単射 }\}$ とする．$\mathcal{X}_0 \subset \mathbb{R}^{d_0}$ とし，$l = 1, \ldots, L$ に対して，$\mathcal{X}_l$，$\tilde{\mathcal{X}}_l$ は，$W_l \sigma_{l-1}(W_{l-1} \cdots W_2 \sigma_1(W_1 \mathcal{X}_0)) \subseteq \tilde{\mathcal{X}}_l \subseteq \mathbb{R}^{d_l}$，および $\mu_{\mathbb{R}^{d_l}}(\mathcal{X}_l) > 0$，$\sigma_l(W_l \cdots W_2 \sigma_1(W_1 \mathcal{X}_0)) \subseteq \mathcal{X}_l \subseteq \mathbb{R}^{d_l}$，および $\mu_{\mathbb{R}^{d_l}}(\tilde{\mathcal{X}}_l) > 0$ を満たすとする．

$\mathcal{X}_0$ から出発して，$l = 1, \ldots, L$ に対して $\tilde{\mathcal{X}}_l$ と $\mathcal{X}_l$ を逐次的に構成する．W_l は単射であるため，空間 $W_l \mathcal{X}_{l-1}$ は d_{l-1} 次元である．$d_l > d_{l-1}$ のとき，測度 $\mu_{\mathbb{R}^{d_l}}(W_l \mathcal{X}_{l-1})$ は 0 と

なり，$\tilde{\mathcal{X}}_l = W_l\mathcal{X}_{l-1}$ と設定すると解析が無意味になる．そのため，$W_l\mathcal{X}_{l-1}$ を含み，かつ $\mu_{\mathbb{R}^{d_l}}(\tilde{\mathcal{X}}_l) > 0$ となる空間 $\tilde{\mathcal{X}}_l$ を設定する．$\tilde{\mathcal{H}}_l = L^2(\tilde{\mathcal{X}}_l)$, $\mathcal{H}_l = L^2(\mathcal{X}_l)$, $\eta_l(W_l) = K_{W_l}$ を W_l に関する $\tilde{\mathcal{H}}_l$ から $\mathcal{H}_{l-1}$ へのクープマン作用素とする．また，$A_l = K_{\sigma_l}$ を，仮定 2.2 を満たす活性化関数 $\sigma_l : \tilde{\mathcal{X}}_l \to \mathcal{X}_l$ に関する $\mathcal{H}_l$ から $\tilde{\mathcal{H}}_l$ へのクープマン作用素とする．このとき，

$$f(W_1, \ldots, W_L)(x) = v(W_L\sigma_{L-1}(W_{L-1}\sigma_{L-2}(\cdots\sigma_1(W_1x)))).$$

が成り立つ．

$D > 0$, $\mathcal{F}_c = \{F_c(\theta_1, \ldots, \theta_L, \cdot) \mid |\det W_1^*W_1|^{-1/4}, \ldots, |\det W_L^*W_L|^{-1/4} \leq D\}$ とする．$h \in \tilde{\mathcal{H}}_l$ に対して $\alpha(h) = (\int_{W_l\mathcal{X}_{l-1}} |h(x)|^2 d\mu_{\mathcal{R}(W_l)}(x) / \int_{\tilde{\mathcal{X}}_l} |h(x)|^2 \mathrm{d}\mu_{\mathbb{R}^{d_l}}(x))^{1/2}$ とする．この値は $W_l\mathcal{X}_{l-1}$ と比較して $\tilde{\mathcal{X}}_l$ をどの程度大きく設定するかに依存しており，$\tilde{\mathcal{X}}_l$ を十分大きく設定することで，合理的な仮定のもとで 1 で上から抑えることができる（詳細は注意 3.3 参照）．次の上界が得られる．

定理 3.2. $f_l = v \circ W_L \circ \sigma_{L-1} \circ \cdots \circ W_{l+1} \circ \sigma_l$ とする．このとき，

$$\hat{R}(\mathcal{F}_c, x_1, \ldots, x_S) \leq \sup_{|\det W_l^*W_l|^{-1/4} \leq D} \frac{E(c)\|v\| \prod_{l=1}^{L-1} \|A_l\| \alpha(f_l)}{\sqrt{S} \prod_{l=1}^{L} |\det W_l^*W_l|^{1/4}} \tag{4}$$

が成り立つ．

$\|A_l\|$ については，$A_l = K_{\sigma_l}$ であるため，補題 2.3 によって $\|A_l\|$ の上界を評価できる．例えば，$\mathcal{X}_0$ が有界であれば，sigmoid と hyperbolic tangent に対して補題 2.4 を適用できる．

注意 3.3. $a, b > 0$ が存在して $a \leq |f_l(x)|^2 \leq b$ が成り立つと仮定する．$b \cdot \mu_{\mathcal{R}(W_l)}(W_l\mathcal{X}_{l-1}) \leq a \cdot \mu_{\mathbb{R}^{d_l}}(\tilde{\mathcal{X}}_l)$ となるように $\tilde{\mathcal{X}}_l$ を十分大きく設定する．このとき，

$$\alpha(f_l)^2 = \frac{\int_{W_l\mathcal{X}_{l-1}} |f_l(x)|^2 \mathrm{d}\mu_{\mathcal{R}(W_l)}(x)}{\int_{\tilde{\mathcal{X}}_l} |f_l(x)|^2 d\mu_{\mathbb{R}^{d_l}}(x)} \leq \frac{b \cdot \mu_{\mathcal{R}(W_l)}(W_l\mathcal{X}_{l-1})}{a \cdot \mu_{\mathbb{R}^{d_l}}(\tilde{\mathcal{X}}_l)} \leq 1$$

が成り立つ．

注意 3.4. 上界 (4) の分母と分子の大きさの間にはトレードオフが存在する．hyperbolic tangent や sigmoid のように，$\|x\| \to \infty$ のとき $\sigma_l(x)$ が定数に近づく場合，$\|x\|$ の大きさが大きくなるにつれて $\sigma_l^{-1}(x)$ の導関数が大きくなる傾向がある．この場合，補題 2.3 によれば，$\det W_l$ が大きければ $\mathcal{X}_l$ の体積が大きくなるため $\|A_l\|$ も大きくなる．この

場合，活性化関数が複雑さを増大させる上で重要な役割を担う．Leaky ReLU のように $\sigma_1, \ldots, \sigma_{L-1}$ が非有界な場合，$\det W_1, \ldots, \det W_L$ が大きければ $\tilde{\mathcal{X}}_L$ が大きくなり，$\|v\|$ が大きくなる．この場合，最終的な非線形変換 v が複雑さを増大させる上で重要な役割を担う．

3.3 一般のニューラルネットワーク

W が単射でない場合，クープマン作用素 K_W は非有界となる．そのため，標準的なクープマン作用素の代わりに重み付きクープマン作用素を考えることで同様な解析を行うことができる．詳しくは文献 [8] を参照されたい．

4 結論

本論文では，クープマン作用素に基づく新たなラデマッハ複雑度の上界を導出した．既存のクープマン作用素に基づく上界と同様に，提案する上界は高ランクの重み行列を持つニューラルネットワークが汎化できることを示している．既存のクープマン作用素に基づく上界は関数空間の滑らかさとデータ空間の非有界性に依存しており，滑らかで非有界な活性化関数を持つ限られたニューラルネットワークモデルに対してのみ有効であった．本論文では，ニューラルネットワークモデルの代数的表現を導入し，元のモデルを近似する正則化モデルを導入することでこの問題を解決した．提案する上界は，hyperbolic tangent，sigmoid，Leaky ReLU 活性化関数を持つモデルなど，幅広いモデルに対して有効である．提案する枠組みは，ラデマッハ複雑度の上界に関するクープマン作用素に基づく理論と実用的な状況との間のギャップを埋めるための出発点となるものである．

参考文献

[1] Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[2] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[3] Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st Conference on*

Neural Information Processing Systems (NIPS), 2017.

[4] Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55:369–406, 2022.

[5] Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein. Truth or backpropaganda? an empirical investigation of deep learning theory. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[6] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 2018 Conference On Learning Theory (COLT)*, 2018.

[7] Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, pages 1064–1068, 2017.

[8] Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Why high-rank neural networks generalize?: An algebraic framework with rkhss. In *The 14th International Conference on Learning Representations (ICLR)*, 2026.

[9] Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *The 12th International Conference on Learning Representations (ICLR)*, 2024.

[10] Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with Hessian-based generalization guarantees. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[11] Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(04):1352–1368, 2021.

[12] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[13] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 2015 Conference on Learning Theory (COLT)*, 2015.

[14] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[15] Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.