Approximation and Estimation Ability of Transformers for Sequence-to-sequence Function with Infinite Dimensional Input
（無限次元入力トランスフォーマーの近似及び推定能力について）

数理情報学専攻 48226214　高倉 将吉
指導教員　鈴木 大慈 准教授

1　はじめに

トランスフォーマーは自己注意機構を用いた深層学習アーキテクチャであり，自然言語処理や画像処理などの特に入力の次元が高いタスクにおいて高い性能を示している．本研究では，無限次元入出力のノンパラメトリック回帰問題において，トランスフォーマーの近似能力および推定能力を解析した．特に，真の関数の滑らかさが入力の方向ごとに異なる場合について入出力の次元に依存しない収束レートを示し，トランスフォーマーが次元の呪いを回避できることを示した．さらに，滑らかな方向が入力ごとに異なるような関数クラスを提案し，同様の収束レートを示した．

2　既存研究

有限次元入出力の場合については学習可能な位置符号化を用いたトランスフォーマーが任意の連続関数を近似できることを示している [4]．しかしながら，必要なパラメータ数は入力次元に対して指数関数的に増大してしまい，これは回避できない [4]．よって，意味のある収束レートを得るためには，関数クラスを適切に制限する必要がある．この観点から，入力の一部にのみ依存するブール値入力の関数クラス [1] や階層的な構造を持った関数クラス [2] についてトランスフォーマーの近似能力および推定能力が解析されている．しかし，これらの研究では収束レートが入力次元に依存しており，無限次元入力を含む非常に入力の次元が高い設定では意味のある結果を得ることができない．一方，畳み込みニューラルネットワークについては，無限次元入力の設定において非等方的な滑らかさを持つ関数クラスに対する収束レートが示されている [3]．

3　問題設定

3.1　ノンパラメトリック回帰問題

入力トークンの次元 d に対して，$[0,1]^{d\times\infty}$ 上の確率測度 P_X にしたがう確率変数 $X^{(i)}=\left\{x_j^{(i)}\right\}_{j=-\infty}^{\infty}$ を考える．ある無限次元入出力の真の関数 $F^\circ:[0,1]^\infty\to\mathbb{R}^\infty$ が存在して，出力が $Y^{(i)}:=F^\circ(X^{(i)})+\xi^{(i)}$ と生成されるとする．ここで，$\xi_j^{(i)}(j\in\mathbb{N})$ は正規分布 $N(0,\sigma^2)$ にしたがう独立な確率変数である．

推定量 $\hat{F}$ の評価には，以下のような平均二乗誤差を用いる．

$$R_{l,r}(\hat{F},F^\circ)=\frac{1}{r-l+1}\sum_{i=l}^{r}\mathbb{E}\left[\left\|\hat{F}_i-F_i^\circ\right\|_{2,P_X}^2\right],$$

本研究では特に，観測された n 個の入出力データ $\left\{(X^{(i)},Y^{(i)})\right\}_{i=1}^n$ に対してトランスフォーマーのクラスの中で経験損失 $\sum_{i=1}^n\sum_{j=l}^r\left(F(X^{(i)})_j-Y_j^{(i)}\right)^2$ を最小化する推定量 $\hat{F}$ の性能を解析する．

3.2　トランスフォーマーモデル

トランスフォーマーは (1) 全結合層，(2) 自己注意機構，(3) 位置符号化の 3 つの要素から構成される．まず，横幅 W の L 層の全結合層 $f(x):=(A_L\eta(\cdot)+b_L)\circ\cdots\circ(A_1x+b_1)$ であって，$\|A_l\|_\infty,\|b_l\|_\infty\le B,\sum_{i=1}^L\|A_i\|_0+\|b_i\|_0\le S$ を満たすものからなるクラスを $\Psi(L,W,S,B)$ とする．次に埋め込み次元 D, 窓幅 U の自己注意機構 $g(X)_i:=x_i+\sum_{h=1}^H V_hX[i-U:i+U]\,\mathrm{Softmax}((K_hX[i-U:i+U])^\top(Q_hx_i))$ であって，$\|K_h\|_\infty,\|Q_h\|_\infty,\|V_h\|_\infty\le B$ を満たすものからなるクラスを $\mathcal{A}(U,D,H,B)$ とする．ただし，$X[i-U,i+U]=[x_{i-U},\dots,X_{i+U}]$ と定めた．最後に，位置符号化を $\mathrm{Enc}_P(X)=EX+P$ と定める．ここで，P は固定された位置符号化である．以上により，トランスフォーマーのクラスは以下のように定義される．

$$\begin{aligned}&\mathcal{T}(M,U,D,H,L,W,S,B)\\&:=\{f_M\circ g_M\circ\cdots\circ f_1\circ g_1\circ\mathrm{Enc}_P\mid\|E\|_\infty\le B,\\&\qquad f_i\in\Psi(L,W,S,B),g_i\in\mathcal{A}(U_i,D,H,B)\},\end{aligned}$$

以下では解析の都合上，推定量の有界性を保証した $\mathcal{T}_R:=\left\{\tilde{F}_i(X)=\min\{R,\max\{-R,F_i(X)\}\}\mid F\in\mathcal{T}\right\}$ というクラスを考えることにする．

3.3 γ 平滑空間

$$\psi_{r_{ij}}(x) := \begin{cases} \sqrt{2}\cos(2\pi|r_{ij}|x) & (r_{ij} < 0), \\ 1 & (r_{ij} = 0), \\ \sqrt{2}\sin(2\pi|r_{ij}|x) & (r_{ij} > 0), \end{cases}$$
$$\psi_r(X) := \prod_{i=1}\prod_{j=1} \psi_{r_{ij}}(X_{ij})$$
$$\delta_s(f) = \sum_{r\in\mathbb{Z}_0^{d\times\infty}, \lfloor 2^{s_{ij}-1}\rfloor \le r_{ij} < 2^{s_{ij}}} \langle f, \psi_r\rangle \psi_r.$$

とする．このとき，$p \ge 2, \theta \ge 1$ に対して γ 平滑空間は以下のように定義される．

$$\mathcal{F}^{\gamma}_{p,\theta} := \left\{ f \in L^2([0,1]^{d\times\infty}) \mid \|f\|_{\mathcal{F}^{\gamma}_{p,\theta}} < \infty \right\}.$$

ここで，$\|f\|_{\mathcal{F}^{\gamma}_{p,\theta}} := \left(\sum_{s\in\mathbb{N}_0^{d\times\infty}} 2^{\theta\gamma(s)}\|\delta_s(f)\|^{\theta}_{p,P_X}\right)^{1/\theta}$ は $\mathcal{F}^{\gamma}_{p,\theta}$ のノルムである．γ は各方向の滑らかさを表す関数であり，本研究では特にある数列 $a \in \mathbb{R}^{d\times\infty}_{>0}$ を用いて (1) 異方平滑 $\gamma(s) = \langle a, s\rangle$，(2) 混合平滑 $\gamma = \max\{a_{ij}s_{ij} \mid i \in [d], j \in \mathbb{Z}\}$ と表される場合を扱う．ここで，a は各方向の滑らかさを表すパラメータであり，a_{ij} が大きいほどその方向の滑らかであると言える．一方で，a_{ij} が小さいほど関数はその方向に大きく変化するため，重要な方向であると言える．滑らかさパラメータ a に関して $\bar{a}$ を a_{ij} を昇順に並び替えた数列として，$\|a\|_{wl^{\alpha}} := \sup_j j^{\alpha}\bar{a}_j^{-1}, \tilde{a} := \left(\sum_{i=1}^{\infty}\bar{a}_i^{-1}\right)^{-1}$ と定める．また，簡単のため混合平滑の場合には $a^{\dagger} = \tilde{a}$，異方平滑の場合には $a^{\dagger} = \bar{a}_1$ と定める．

3.4 区分 γ 平滑関数

γ 平滑空間では滑らかさは入力によらず一定であり，重要な方向は変化しない．しかし，実際の応用では入力ごとに重要な方向が異なると考えられる．そこで各領域ごとに滑らかな方向が異なる区分 γ 平滑関数を次のように定義する．まず，分布のサポート Ω の分割 $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ が与えられているとする．$V \in \mathbb{N}$ に対して，$\{1,\dots,2V+1\}$ から $\{-V,\dots,V\}$ への全単射 π_λ に対してトークンを並び替える写像 Π_λ と Π を次のように定める．

$$\Pi_\lambda([x_{-V},\dots,x_V]) := [x_{\pi_\lambda(1)},\dots,x_{\pi_\lambda(2V+1)}],$$
$$\Pi(X) := \Pi_\lambda([x_{-V},\dots,x_V]) \quad (X \in \Omega_\lambda).$$

以上を用いて区分 γ 平滑関数は

$$\mathcal{P}^{\gamma}_{p,\theta}(\Omega) := \left\{ g = f\circ\Pi \mid f \in \mathcal{F}^{\gamma}_{p,\theta}, \|g\|_{\mathcal{P}^{\gamma}_{p,\theta}} < \infty \right\},$$

と定義される．ここで，$\|g\|_{\mathcal{P}^{\gamma}_{p,\theta}} := \left(\sum_{s\in\mathbb{N}_0^{d\times[-V:V]}} 2^{\theta\gamma(s)}\|\delta_s(f)\circ\Pi\|^{\theta}_{p,P_X}\right)^{1/\theta}$ は $\mathcal{P}^{\gamma}_{p,\theta}$ のノルムである．また，領域の分割 $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ への正則条件として，ある定数 $c > 0$ に対して以下の条件を満たす重要度関数 $\mu : \Omega \to \mathbb{R}^{\infty}$ が存在することを仮定する．

$$\Omega_\lambda = \{X \in \Omega \mid \mu(X)_{\pi_\lambda(1)} > \cdots > \mu(X)_{\pi_\lambda(2V+1)}\},$$
$$\mu(X)_{\pi_\lambda(i)} \ge \mu(X)_{\pi_\lambda(i+1)} + ci^{-\beta}$$

4 主結果

本稿では推定誤差に関する結果のみを紹介する．

4.1 推定誤差解析

真の関数 F° がシフト同変であり 0 番目のトークンに対応する出力 F_0° が（区分）γ 平滑空間の単位球に含まれており，ある定数 R に対して $\|F_0\|_{\infty} \le R$ を満たすとする．また，ある $0 < \alpha < \infty$ について滑らかさパラメータ a が $\|a\|_{wl^{\alpha}} \le 1$ と $a_{ij} = \Omega(\log(|j|+1))$ (γ 平滑空間の場合), $a_{ij} = \Omega(j^{\alpha})$ (区分 γ 平滑空間の場合) を満たすとする．特に F° が区分 γ 平滑関数の場合，重要度関数 μ が γ 平滑関数に関する仮定を満たすとする．さらに，混合平滑の場合には $\bar{a}_1 < \bar{a}_2$ を仮定する．

以上の仮定のもとで，ハイパーパラメータを適切に設定した経験損失最小化推定量 $\hat{F}$ の推定誤差は以下のように評価できる．

- γ 平滑空間の場合
$$R_{l,r}(\hat{F}, F) \lesssim n^{-\frac{2a^{\dagger}}{2a^{\dagger}+1}}(\log n)^{2/\alpha+2+\max\{4/\alpha,4\}}.$$
- 区分 γ 平滑空間の場合
$$R_{l,r}(\hat{F}, F) \lesssim n^{-\frac{2a^{\dagger}}{2a^{\dagger}+1}}(\log n)^{5/\alpha+2+\max\{4/\alpha,4\}}.$$

どちらの場合も主要項は $n^{-\frac{2a^{\dagger}}{2a^{\dagger}+1}}$ であり，$a^{\dagger}$ は滑らかさのみに依存するため，推定誤差は入力や出力の次元に依存しないことがわかる．

参考文献

[1] Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive Biases and Variable Creation in Self-Attention Mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

[2] Iryna Gurevych, Michael Kohler, and Gözde Gül Şahin. On the rate of convergence of a classifier based on a Transformer encoder. *IEEE Transactions on Information Theory*, 2022.

[3] Sho Okumoto and Taiji Suzuki. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022.

[4] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.