Approximation Theory for Lipschitz Continuous Transformers

Takashi Furuya

Doshisha University, RIKEN AIP, Japan

Joint work with
Davide Murari (University of Cambridge)
Carola-Bibiane Schönlieb (University of Cambridge)

Motivation

- Standard Transformers do not preserve Lipschitz continuity.
- How arbitrary Lipschitz maps can be approximated by Lipschitz Transformers ?
- How can we construct such Transformers while preserving Lipschitz continuity?

Gradient-descent-type MLP

- Define the gradient-descent-type MLP mapping $F_\xi : \mathbb{R}^d \to \mathbb{R}^d$ by (Sherry et al., 2024)

$$F_\xi(x) := x - \tau W^\top \mathrm{ReLU}(Wx + b)$$

- Learnable parameters : $\xi = (W, b, \tau)$.

Lemma 1 (Sherry et al., 2024)

Assume that

$$\tau \in [0, 2/\|W\|_2^2].$$

Then the map $F_\xi : (\mathbb{R}^d, \|\cdot\|_2) \to (\mathbb{R}^d, \|\cdot\|_2)$ *is* 1*-Lipschitz continuous.*

In-context Map

- Define the in-context attention map by (Castin et al., 2024)

$$\mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \ni (\mu, x) \mapsto \int \frac{\exp\Big(\langle Qx, Ky \rangle\Big)}{\int \exp\Big(\langle Qx, Kz \rangle\Big)\, d\mu(z)} Vy \, d\mu(y).$$

- The standard discrete attention corresponds to the case where

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Gradient-descent-type In-context Attention

- Introduce a gradient-descent-type in-context attention map $\Gamma_\theta : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\Gamma_\theta(\mu, x) := x - \eta \int \frac{e^{\langle x, Ay \rangle}}{\int e^{\langle x, Az \rangle} d\mu(z)} Ay d\mu(y).$$

- Learnable parameters : $\theta = (A, \eta)$.

Lemma 2

Let $\Omega \subset \mathbb{R}^d$ be a compact set. Assume that

$$\eta \in \left[0, \frac{2}{\sup_{y \in \Omega} \|Ay\|_2^2} \right]$$

Then the map $\Gamma_\theta(\mu, \cdot) : (\Omega, \|\cdot\|_2) \to (\mathbb{R}^d, \|\cdot\|_2)$ is 1-Lipschitz continuous for each $\mu \in \mathcal{P}(\Omega)$.

Gradient-descent-type In-context Transformer

- Gradient-descent-type in-context Transformer is the composition of the maps F_ξ and Γ_θ.

- Set

$$\mathcal{C}_1(\mathcal{P}(\Omega)\times\Omega,\mathbb{R}) := \Big\{\Lambda^* : \mathcal{P}(\Omega)\times\Omega\to\mathbb{R} :$$
$$\Lambda^*(\mu,\cdot) : (\Omega,\|\cdot\|_2)\to(\mathbb{R},|\cdot|) \text{ is 1-Lipschitz}$$
$$\Lambda^*(\cdot,x) : (\mathcal{P}(\Omega),W_1)\to(\mathbb{R},|\cdot|) \text{ is 1-Lipschitz}\Big\}.$$

Main theorem

Theorem 3 (Furuya, Murari, Schönlieb, 2026 (Informal))

Let $\Omega \subset \mathbb{R}^d$ be a compact set. Then, for any $\varepsilon \in (0,1)$ and any $\Lambda^* \in \mathcal{C}_1(\mathcal{P}(\Omega) \times \Omega, \mathbb{R})$, there exists a gradient-descent-type in-context Transformer $\Lambda \in \mathcal{C}_1(\mathcal{P}(\Omega) \times \Omega, \mathbb{R})$ such that

$$\sup_{(\mu,x)\in\mathcal{P}(\Omega)\times\Omega} |\Lambda(\mu, x) - \Lambda^*(\mu, x)| \leq \varepsilon.$$

- **Proof idea:** Apply the Restricted Stone–Weierstrass theorem.

Restricted Stone–Weierstrass Theorem

Lemma 4 (Restricted Stone–Weierstrass Theorem)

Let (X, d_X) be a compact metric space with at least two points, and let

$$L \subset \mathcal{C}_1(X, \mathbb{R})$$

Assume that

- *L is a lattice, i.e., $\max\{f, g\}, \min\{f, g\} \in L$ for $f, g \in L$.*
- *L has the Lipschitz point-separation property, i.e., for $x \neq y \in X$ and $a, b \in \mathbb{R}$ satisfying*

$$|a - b| < d_X(x, y),$$

there exists a function $f \in L$ such that $f(x) = a$ and $f(y) = b$.

Then L is dense in $\mathcal{C}_1(X, \mathbb{R})$ with respect to the uniform norm.

Proof Strategy

- **Lattice property**

  The gradient-descent-type Transformer class is stable under

$$(f, g) \mapsto \max\{f, g\}, \qquad (f, g) \mapsto \min\{f, g\}.$$

- **Lipschitz point-separation property**

  The gradient-descent-type Transformer class contains sufficiently many "1-Lipschitz affine-like separators" that separate arbitrary measures

$$\mu \neq \mu'.$$

Future Works

- Future work 1: The step size

$$\eta_\ell \in \left[0, \frac{2}{\sup_{y \in \Omega_\ell} \|Ay\|_2^2}\right]$$

  depends on the layerwise input domain Ω_ℓ, which is generally difficult to estimate precisely. Develop more practical and certifiable architectures.

- Future work 2: Our analysis focuses on scalar-valued targets due to the reliance on the Stone–Weierstrass theorem. Extend the theory to vector-valued targets by developing new approximation mechanisms.