Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input

Shokichi Takakura [1 2] **Taiji Suzuki** [1 2]

Abstract

Despite the great success of Transformer networks in various applications such as natural language processing and computer vision, their theoretical aspects are not well understood. In this paper, we study the approximation and estimation ability of Transformers as sequence-to-sequence functions with infinite dimensional inputs. Although inputs and outputs are both infinite dimensional, we show that when the target function has anisotropic smoothness, Transformers can avoid the curse of dimensionality due to their feature extraction ability and parameter sharing property. In addition, we show that even if the smoothness changes depending on each input, Transformers can estimate the importance of features for each input and extract important features dynamically. Then, we proved that Transformers achieve similar convergence rate as in the case of the fixed smoothness. Our theoretical results support the practical success of Transformers for high dimensional data.

1. Introduction

Transformer networks, first proposed in Vaswani et al. (2017), empirically show high performance in various fields including natural language processing (Vaswani et al., 2017), computer vision (Dosovitskiy et al., 2021) and audio processing (Dong et al., 2018), where the dimensionality of inputs is relatively high. However, despite the growing interest in Transformer models, their theoretical properties are still unclear.

Aside from the Transformer architecture, there exists a line of work which studied the approximation and estimation ability of fully connected neural networks (FNN) for certain function spaces such as Hölder spaces (Schmidt-Hieber, 2020) and Besov spaces (Suzuki, 2018). For example, Schmidt-Hieber (2020) showed that FNNs with ReLU activation can achieve the near minimax optimal rate of the estimation error for composite functions in Hölder spaces. Although deep learning can achieve the near optimal rate for several function classes, the convergence rate is often strongly affected by the dimensionality of inputs. Some researches (Nakada & Imaizumi, 2022; Chen et al., 2022) considered the settings where the data are distributed on a low dimensional manifold and showed that deep neural networks can avoid the curse of dimensionality. However, this assumption is relatively strong since the low dimensionality of the data manifold is easily destroyed by noise injection. Then, Suzuki (2018) showed that even if the data manifold is not low dimensional, deep neural networks can avoid the curse of dimensionality under the assumption that the target function has anisotropic smoothness. Moreover, Okumoto & Suzuki (2022) showed that (dilated) convolutional neural networks (CNN) can avoid the curse of dimensionality even if inputs are infinite dimensional.

Although the learnability of FNNs and CNNs has been intensively studied, that of Transformer networks is not well understood. There are some researches (Edelman et al., 2022; Gurevych et al., 2022) which studied the learning ability of Transformers. Edelman et al. (2022) evaluated the capacity of Transformer networks and derived the sample complexity to learn sparse Boolean functions. Since they investigated discrete inputs, the smoothness of the target function was not considered. Gurevych et al. (2022) studied binary classification tasks and proved that Transformer networks can avoid the curse of dimensionality when a posteriori probability is represented by the hierarchical composition model with Hölder smoothness. However, these studies have some limitations. First, the previous works considered the fixed length input, although Transformer networks can be applied to sequences of any length due to the parameter sharing property, even if the length is infinite. Indeed, Transformers are often applied to very high dimensional data such as images and languages, and the learnability of Transformers for such extremely high dimensional data is still unclear. Second, their analysis is limited to the single output setting.

[1]Department of Mathematical Informatics, the University of Tokyo, Tokyo, Japan [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Shokichi Takakura <masayoshi361@g.ecc.u-tokyo.ac.jp>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

In some applications such as question-answering, it is necessary to learn a function that maps an input sequence to an output sequence. Transformer networks can be applied to such situations and achieve great practical success as represented by BERT (Devlin et al., 2019) although output is also high dimensional. Finally, in these studies, the intrinsic structure of target functions does not depend on each input and the pattern of attention weights does not change. This is in contrast to the dynamic nature of self-attention matrices observed in practice (Likhosherstov et al., 2021).

In this paper, we consider the non-parametric regression problems and study the approximation and estimation ability of Transformers for *sequence-to-sequence* functions with *infinite* dimensional inputs. In the high dimensional setting, the dependence of the target function on inputs varies depending on the direction. For example, in image classification, the target function is more dependent on foreground features than background features. To deal with such situations, we consider *direction-dependent* smoothness. Then, we derive the convergence rate of errors for mixed and anisotropic smooth functions and show that Transformer networks can avoid the curse of dimensionality. In addition, we consider the setting where the position of important features changes depending on each input and show that Transformer networks can avoid the curse of dimensionality, which reveals the *dynamical* feature extraction ability of self-attention mechanism. Our contribution can be summarized as follows.

- We derive the convergence rate of approximation and estimation error for shift-equivariant functions with the mixed or anisotropic smoothness. We show that the errors are dependent only on the smoothness of the target function and independent of the input and output dimension. This means that Transformers can avoid the curse of dimensionality even if the dimensionality of inputs and outputs is infinite.
- We consider the situation where the smoothness of each coordinate, which corresponds to the importance of each feature, changes depending on inputs and derive the similar convergence rate to the case of the fixed smoothness.

1.1. Other Related Works

Yun et al. (2020); Zaheer et al. (2020) proved that Transformers with learnable positional encodings are universal approximators of continuous sequence-to-sequence functions with compact support, but the results suffer from the curse of dimensionality. This is unavoidable as mentioned in Yun et al. (2020). To derive meaningful convergence results, it is necessary to restrict the function class. From this perspective, Edelman et al. (2022) investigated sparse Boolean functions and Gurevych et al. (2022) studied the hierarchical composition model. However, our analysis imposes smoothness structure on target functions more directly compared to these studies.

The mixed and anisotropic smooth function spaces which we consider in this study are extensions of the functions investigated in Okumoto & Suzuki (2022). In addition, the function spaces can be seen as an infinite dimensional counterpart of the mixed Besov space (Schmeisser, 1987) and anisotropic Besov space (Nikol'skii, 1975). From the deep learning perspective, the approximation and estimation error of FNNs for the mixed Besov space and anisotropic Besov space was analyzed in Suzuki (2018) and Suzuki & Nitanda (2021), respectively. However, it is not trivial to extend these results to Transformer architecture and multiple output setting.

In this paper, we also consider the piecewise γ-smooth function class. This function class is inspired by the piecewise smooth functions, which was investigated in Petersen & Voigtlaender (2018); Imaizumi & Fukumizu (2019), but these studies did not consider anisotropic smoothness and Transformer networks.

There are some studies that investigated the theoretical properties of Transformer networks from different perspective than ours. Jelassi et al. (2022) analyzed simplified Vision transformers and showed that they can learn the spatial structure via gradient descent. Zhang et al. (2022) studied the self-attention mechanism from the perspective of exchangeability and proved that Transformer networks can learn desirable representation of input tokens. Pérez et al. (2019) showed the Turing completeness of Transformers and Wei et al. (2021) introduced the notion of *statistically meaningful approximation* and gave the sample complexity to approximate Boolean circuits and Turing machines. Likhosherstov et al. (2021) showed that a self-attention module with fixed parameters can approximate any sparse matrix by designing an input appropriately.

1.2. Notations

Here, we prepare the notations. For $l \in \mathbb{N}$, let $[l]$ be the set $\{1, \dots, l\}$ and for $l, r \in \mathbb{Z}$ $(l \leq r)$, let $[l : r]$ be the set $\{l, \dots, r\}$. For a set $\mathbb{S} \subset \mathbb{R}$ and $d \in \mathbb{N}$, let

$$\mathbb{S}^{d\times\infty} := \left\{[\dots, s_{-1}, s_0, s_1, \dots, s_i, \dots] \mid s_i \in \mathbb{S}^d\right\},$$
$$\mathbb{S}_0^{d\times\infty} := \left\{s \in (\mathbb{S} \cup \{0\})^{d\times\infty} \mid |\mathrm{supp}(s)| < \infty\right\},$$

where $\mathrm{supp}(s)$ is defined as $\{(i, j) \in [d] \times \mathbb{Z} \mid s_{i,j} \neq 0\}$. Similary, $\mathbb{S}^{d\times[l:r]}$ denotes the set $\{[s_l, \dots s_r] \mid s_j \in \mathbb{S}\}$. For $X = [\dots, x_0, x_1, \dots] \in \mathbb{R}^{d\times\infty}$, $X[l : r]$ denotes $[x_l, \dots, x_r] \in \mathbb{R}^{d\times[l:r]}$. For $s \in \mathbb{R}_0^{d\times\infty}$, let $2^s := 2^{\sum_{i\in[d], j\in\mathbb{Z}} s_{ij}}$. For $X \in \mathbb{R}^{d\times\infty}$, $\|X\|_\infty$ denotes $\sup_{i\in[d], j\in\mathbb{Z}} |X_{i,j}|$ and for $x \in \mathbb{R}^l$, $\|x\|_1$ denotes $\sum_{i=1}^l |x_i|$.

For $F : \Omega \to \mathbb{R}^l$, let $\|F\|_\infty := \sup_{X\in\Omega} \|F(X)\|_\infty$. For the probability measure P_X on Ω and $p > 0$, the norm $\|\cdot\|_{p,P_X}$ is defined by

$$\|f\|_{p,P_X} = \left(\int_\Omega \|f(X)\|_p^p \,\mathrm{d}P_X\right)^{1/p}.$$

For a matrix A, let $\|A\|_0 = |\{(i,j) \mid A_{ij} \neq 0\}|$. For $j \in \mathbb{Z}$, we define the shift operator $\Sigma_j : \mathbb{R}^{d\times\infty} \to \mathbb{R}^{d\times\infty}$ by $(\Sigma_j(X))_i = x_{i+j}$ for $X = [\dots, x_0, \dots, x_i, \dots] \in \mathbb{R}^{d\times\infty}$. For a normed space $\mathcal{F}$, we define $U(\mathcal{F})$ by $U(\mathcal{F}) := \{f \in \mathcal{F} \mid \|f\|_\mathcal{F} \leq 1\}$, where $\|\cdot\|_\mathcal{F}$ is the norm of $\mathcal{F}$.

2. Problem Settings

2.1. Non-parametric Regression Problems

In this paper, we consider non-parametric regression problems with infinite dimensional inputs. We regard an input $X \in [0,1]^{d\times\infty}$ as a bidirectional sequence of tokens $\{x_i\}_{i=-\infty}^{\infty}$ ($x_i \in \mathbb{R}^d$). For example, each token x_i corresponds to a word vector in natural language processing and an image patch in image processing (Dosovitskiy et al., 2021). Let P_X be a probability measure on $([0,1]^{d\times\infty}, \mathcal{B}([0,1]^{d\times\infty}))$. We write Ω for the support of P_X. We assume that P_X is shift-invariant. That is, for any $i \in \mathbb{Z}$ and $B \in \mathcal{B}([0,1]^{d\times\infty})$, $P_X(B) = P_X(\{\Sigma_i(X) \mid X \in B\})$. In the non-parametric regression, we observe n i.i.d. pairs of inputs $X^{(i)} \sim P_X$ and outputs $Y^{(i)} \in \mathbb{R}^\infty$. We assume that there exists a true function $F^\circ : \Omega \to \mathbb{R}^\infty$, and outputs $Y^{(i)}$ is given by

$$Y^{(i)} := F^\circ(X^{(i)}) + \xi^{(i)},$$

where the noise $\xi_j^{(i)}$ follows the normal distribution $N(0,\sigma^2)$ ($\sigma > 0$) independently. We also assume that $\{\xi^{(i)}\}_{i=1}^n$ are independent of $\{X^{(i)}\}_{i=1}^n$. Note that unlike Okumoto & Suzuki (2022), we do *not* assume that the Radon-Nikodym derivative $\frac{\mathrm{d}P_X}{\mathrm{d}\lambda}$ for the uniform distribution λ on $([0,1]^{d\times\infty}, \mathcal{B}([0,1]^{d\times\infty}))$ satisfies $\left\|\frac{\mathrm{d}P_X}{\mathrm{d}\lambda}\right\|_\infty < \infty$.

Based on the observed data $\mathcal{D}^n := \left\{(X^{(i)}, Y^{(i)})\right\}_{i=1}^n$, we compute an estimator $\hat{F}$ which takes its value in the class of Transformer networks. To evaluate the statistical performance of an estimator $\hat{F}$, we consider the mean squared error

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

where the expectation is taken with respect to the training data $\mathcal{D}^n$. Here, to avoid the convergence argument, we consider a finite number of outputs $(Y_l^{(i)}, \dots, Y_r^{(i)})$, but we show later that the convergence rate of estimation error does not depend on l and r.

In this paper, we consider an empirical risk minimization (ERM) estimator, which is defined as a minimizer of the following minimization problem:

$$\min_{F\in\mathcal{T}} \sum_{i=1}^{n}\sum_{j=l}^{r}\left(F(X^{(i)})_j - Y_j^{(i)}\right)^2,$$

where $\mathcal{T}$ is supposed to be the set of Transformer networks defined in Eq. (1). Note that an ERM estimator $\hat{F}$ is a random variable which depends on the training dataset $\mathcal{D}^n$. In practice, it is difficult to solve the problem due to the non-convexity of the objective function. Some studies (Huang et al., 2020; Jelassi et al., 2022) investigated the optimization aspect of Transformers, but we do not pursue this direction in this study.

2.2. Transformer Architecture

Transformer architecture has three main components:

(i) (position-wise) FNN layer.

(ii) Self-attention layer.

(iii) Embedding layer.

(i) First, we introduce FNN layers. An FNN with depth L and width W is defined as

$$f(x) := (A_L\eta(\cdot) + b_L) \circ \cdots \circ (A_1 x + b_1),$$

where $A_i \in \mathbb{R}^{d_{i+1}\times d_i}, b_i \in \mathbb{R}^{d_{i+1}}$, $\max_i d_i \leq W$, and ReLU activation function $\eta(x) = \max\{x, 0\}$ is operated in an element-wise manner. Then, we define the class of FNN with depth L, width W, norm bound B and sparsity S by

$$\Psi(L,W,S,B) := \Big\{f \mid \max_i\{\|A_i\|_\infty, \|b_i\|_\infty\} \leq B, \sum_{i=1}^{L}\|A_i\|_0 + \|b_i\|_0 \leq S\Big\}.$$

(ii) Next, we define the self-attention layer. In this paper, we consider the sliding window attention, which is used in some practical architectures such as Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020). To focus on local context, the sliding window attention restricts the receptive field to the input around each token. Let D be the embedding dimension, H be the number of head, and U be the window size. Then, self-attention layer g with parameters $K_h \in \mathbb{R}^{D'\times D}, Q_h \in \mathbb{R}^{D'\times D}, V_h \in \mathbb{R}^{D\times D}$ ($D' \leq D, h = 1, \dots, H$) is defined by

$$g(X)_i := x_i + \sum_{h=1}^{H} V_h X[i-U : i+U] A_h,$$

where

$$A_h := \mathrm{Softmax}((K^{(h)}X[i-U:i+U])^\top (Q^{(h)}x_i)), \in \mathbb{R}^{[i-U:i+U]}.$$

Here, $\mathrm{Softmax} : \mathbb{R}^l \to \mathbb{R}^l$ is defined by

$$\mathrm{Softmax}(x) = \left[\frac{e^{x_1}}{\sum_{j\in[l]} e^{x_j}}, \dots, \frac{e^{x_l}}{\sum_{j\in[l]} e^{x_j}}\right]^\top.$$

Then, we define the class of self-attention layers with the window size $U \in \mathbb{N}$, the embedding dimension D, the number of head H, and the norm bound B by

$$\mathcal{A}(U, D, H, B) := \left\{ g \mid \max_h \{\|K_h\|_\infty, \|Q_h\|_\infty, \|V_h\|_\infty\} \le B \right\}.$$

(iii) Finally, we define the embedding layer. For embedding dimension D, an embedding layer is defined as

$$\mathrm{Enc}_P(X) = EX + P,$$

where $E \in \mathbb{R}^{D\times d}$, $P = [p_i]_{i=-\infty}^{\infty} \in \mathbb{R}^{D\times\infty}$. Here, P is called a positional encoding. Since position-wise FNN and self-attention layers are permutation equivariant, a positional encoding is often added to break the equivariance when positional information is important. Sometimes, a learnable positional encoding is used, but we consider that P is fixed since P is infinite dimensional in our setting. Relative positional encoding (Shaw et al., 2018) is another way to encode the positional information, but it requires extra trainable parameters. Therefore, we consider the absolute positional encoding in this paper.

We define the class of transformers by

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

where FNN is applied column-wise. Thanks to the parameter sharing property, $F \in \mathcal{T}$ can represent a function from $[0,1]^{d\times\infty}$ to $\mathbb{R}^\infty$ even though it has a finite number of parameters. In order to derive the estimation error for a model class $\mathcal{T}$, it is convenient to assume that there exists a constant $R > 0$ such that $\|f\|_\infty \le R$ for any $f \in \mathcal{F}$ since this assumption ensure the sub-Gaussianity of $f(X^{(i)})$. To ensure this property, we define the class of (clipped) Transformer networks by

$$\mathcal{T}_R := \left\{ \tilde{F} = \mathrm{clip}_R \circ F \mid F \in \mathcal{T} \right\}, \tag{1}$$

where $\mathrm{clip}_R(x) := R \wedge (x \vee -R)$ is applied element-wise. Note that clip_R can be realized by ReLU units.

For simplicity, we consider a modified version of the original architecture in Vaswani et al. (2017). That is, we consider the multilayer FNNs without skip connection instead of the single layer FNNs with skip connection. However, our argument can be applied to the original architecture with a slight modification. See Appendix B for details.

3. Function Spaces

In this paper, we assume the true function F° is shift-equivariant. A function $F : \Omega \to \mathbb{R}^{d'\times\infty}$ is called shift-equivariant if F satisfies

$$F(\Sigma_j(X)) = \Sigma_j(F(X)),$$

for any $j \in \mathbb{Z}$ and $X \in \Omega$. Such equivariance appears in various applications such as natural language processing, audio processing, and time-series analysis. We also assume that $F_i(X) := (F(X))_i$ is included in a certain function class which is defined in this section.

3.1. Anisotropic and Mixed Smoothness

First, we introduce the γ-smooth function class. This is an extension of the function class in Okumoto & Suzuki (2022) to the situation where the inputs are bidirectional sequences of tokens. For $r \in \mathbb{Z}_0^{d\times\infty}$, we define $\psi_{r_{ij}} : [0,1] \to \mathbb{R}$ by

$$\psi_{r_{ij}}(x) := \begin{cases} \sqrt{2}\cos(2\pi|r_{ij}|x) & (r_{ij} < 0), \\ 1 & (r_{ij} = 0), \\ \sqrt{2}\sin(2\pi|r_{ij}|x) & (r_{ij} > 0), \end{cases}$$

and $\psi_r : [0,1]^{d\times\infty} \to \mathbb{R}$ by $\psi_r(X) = \prod_{i=1}\prod_{j=1}\psi_{r_{ij}}(X_{ij})$. Since $\{\psi_r\}_{r\in\mathbb{Z}_0^{d\times\infty}}$ is a complete orthonormal system of $L^2([0,1]^{d\times\infty})$, any $f \in L^2([0,1]^{d\times\infty})$ can be expanded as $f = \sum_{r\in\mathbb{Z}_0^{d\times\infty}} \langle f, \psi_r\rangle \psi_r$. For $s \in \mathbb{N}_0^{d\times\infty}$, define $\delta_s(f)$ as

$$\delta_s(f) = \sum_{r\in\mathbb{Z}_0^{d\times\infty}, \lfloor 2^{s_{ij}-1}\rfloor \le r_{ij} < 2^{s_{ij}}} \langle f, \psi_r\rangle \psi_r.$$

This quantity represents the frequency component of f with frequency $|r_{ij}| \sim 2^{s_{ij}}$ for each coordinate. Then, we define the γ-smooth function class as follows.

Definition 3.1 (γ-smooth function class)**.** For a given $\gamma : \mathbb{N}_0^{d\times\infty} \to \mathbb{R}$ which is monotonically non-decreasing with respect to each coordinate and $p \ge 2, \theta \ge 1$, we define the γ-smooth function space as follows:

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

where the norm $\|f\|_{\mathcal{F}^{\gamma}_{p,\theta}}$ is defined as

$$\|f\|_{\mathcal{F}^{\gamma}_{p,\theta}} := \left(\sum_{s \in \mathbb{N}_0^{d\times\infty}} 2^{\theta\gamma(s)} \|\delta_s(f)\|_{p,P_X}^{\theta} \right)^{1/\theta}.$$

We also define the finite dimensional version of γ-smooth function space $\mathcal{F}^{\gamma}_{p,\theta}([0,1]^{d\times l})$ for $l \in \mathbb{N}$ in the same way.

Since $\delta_s(f)$ represents the frequency component of f with frequency $|r_{ij}| \sim 2^{s_{ij}}$ and weight $2^{\gamma(s)}$ is imposed on $\|\delta_s(f)\|_p$, γ controls the amplitude of each frequency component.

As a special case of γ, we consider the mixed and anisotropic smoothness.

Definition 3.2 (Mixed and anisotropic smoothness)**.** For $a \in \mathbb{R}^{d\times\infty}_{>0}$, mixed smoothness and anisotropic smoothness is defined as follows:

- mixed smoothness:

$$\gamma(s) = \langle a, s \rangle .$$

- anisotropic smoothness:

$$\gamma(s) = \max\{a_{ij}s_{ij} \mid i \in [d], j \in \mathbb{Z}\}.$$

The parameter a represents the smoothness for the coordinate $X_{i,j}$. That is, if a_{ij} is large, the function is smooth with respect to the variable $X_{i,j}$. In other words, small a_{ij} implies that the function is not smooth towards the coordinate (i,j) and X_{ij} is an *important feature*.

When $d = 1$, $p = \theta = 2$, and P_X is the uniform distribution on $[0,1]^l$, as shown in Okumoto & Suzuki (2022), the anisotropic smooth function space $\mathcal{F}^{\gamma}_{p,\theta}([0,1]^l)$ includes the anisotropic Sobolev space: $\mathcal{W}^a_2 := \left\{ f \in L^2([0,1]^l) \mid \sum_{i=1}^l \left\| \frac{\partial^{a_i} f}{\partial x_i^{a_i}} \right\|_2^2 < \infty \right\}$. Isotropic Sobolev spaces are a special case of anisotropic Sobolev spaces with $a_i = a_1$ ($\forall i \in [l]$). In that sense, $\mathcal{F}^{\gamma}_{p,\theta}([0,1]^{d\times\infty})$ is an extension of the finite dimensional Sobolev space.

Here, we define some quantities regarding the smoothness parameter a. Let $\bar{a} = \{\bar{a}_i\}_{i=1}^{\infty}$ be the sorted sequence in the ascending order. That is, $\bar{a} = [a_{i_1,j_1}, \dots, a_{i_k,j_k}, \dots]$ satisfies $a_{i_k,j_k} \le a_{i_{k+1},j_{k+1}}$ for any $k \in \mathbb{N}$. Then, weak l^{α}-norm for $\alpha > 0$ is defined by $\|a\|_{wl^{\alpha}} := \sup_j j^{\alpha}\bar{a}_j^{-1}$,and $\tilde{a}$ is defined by $\tilde{a} := \left(\sum_{i=1}^{\infty} \bar{a}_i^{-1}\right)^{-1}$.To simplify the notation, we define $a^{\dagger} = \bar{a}_1$ for the mixed smoothness and $a^{\dagger} = \tilde{a}$ for the anisotropic smoothness.

3.2. Piecewise Anisotropic and Mixed Smoothness

The mixed and anisotropic smooth functions represent the situations where the smoothness depends on the direction. However, the smoothness does not depend on each input. That is, the position of important tokens is fixed for *any input*. This is not the case in practical situations. For example, in natural language processing, the positions of important words should change if a meaningless word is inserted in the input sequence. Therefore, it is natural to assume that the smoothness for each coordinate changes depending on each input. To consider such situations, we define a novel function class called *piecewise* γ-smooth function class.

Definition 3.3 (Piecewise γ-smooth function class)**.** For an index set Λ, let $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ be a disjoint partition of Ω. That is, $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ satisfies

$$\Omega = \bigcup_{\lambda\in\Lambda} \Omega_\lambda,\ \Omega_\lambda \cap \Omega_{\lambda'} = \emptyset \quad (\lambda \ne \lambda').$$

For $V \in \mathbb{N}$ and a set of bijections $\{\pi_\lambda\}_{\lambda\in\Lambda}$ between $[2V+1]$ and $[-V:V]$, define $\Pi_\lambda : \mathbb{R}^{d\times[-V:V]} \to \mathbb{R}^{d\times(2V+1)}$ and $\Pi : \Omega \to \mathbb{R}^{d\times(2V+1)}$ by

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

Then, for $p \ge 2, \theta \ge 1$ and $\gamma : \mathbb{N}_0^{d\times\infty} \to \mathbb{R}$, the function class with piecewise γ-smoothness is defined as follows:

$$\mathcal{P}^{\gamma}_{p,\theta}(\Omega) := \{g = f \circ \Pi \mid f \in \mathcal{F}^{\gamma}_{p,\theta}([0,1]^{d\times(2V+1)}), \|g\|_{\mathcal{P}^{\gamma}_{p,\theta}} < \infty\},$$

where the norm $\|g\|_{\mathcal{P}^{\gamma}_{p,\theta}}$ is defined by

$$\|g\|_{\mathcal{P}^{\gamma}_{p,\theta}} := \left(\sum_{s\in\mathbb{N}_0^{d\times[-V:V]}} 2^{\theta\gamma(s)} \|\delta_s(f)\circ\Pi\|_{p,P_X}^{\theta} \right)^{1/\theta}.$$

On each domain Ω_λ, a piecewise γ-smooth function g can be seen as the restriction of a certain γ-smooth function g_λ. In addition, considering the mixed or anisotropic smoothness, the smoothness parameter of g_λ is a permutation of the original smoothness parameter a. Therefore, the relatively smooth directions of g change depending on each input. This situation is shown in Fig. 2.

In this paper, we assume that there exists an *importance function*, defined as follows.

Definition 3.4 (importance function)**.** A function $\mu : \Omega \to \mathbb{R}^{\infty}$ is called an *importance function* for $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ if μ satisfies

$$\Omega_\lambda = \{X \in \Omega \mid \mu(X)_{\pi_\lambda(1)} > \dots > \mu(X)_{\pi_\lambda(2V+1)}\}.$$

Here, we briefly explain the intuition behind the definition. For $X \in \Omega$, let $X' = \Pi(X)$. Assume that x'_i is more

important than x'_{i+1}. From the definition of Π, we have $x'_i = x_{\pi_\lambda(i)}$ when $X \in \Omega_\lambda$. Therefore, the token $x_{\pi_\lambda(i)}$ is more important than $x_{\pi_\lambda(i+1)}$. The definition of the importance function reflects this relationship. We also assume that an importance function μ is *well-separated*. That is, μ satisfies

$$\mu(X)_{\pi_\lambda(i)} \geq \mu(X)_{\pi_\lambda(i+1)} + ci^{-\beta}, \tag{2}$$

for any $X \in \Omega_\lambda$, where $c, \beta > 0$ is a constant. This implies that the probability that X satisfies $\mu(X)_i \simeq \mu(X)_j$ $(i \neq j)$ is zero. Similar assumption can be found in the analysis of infinite dimensional PCA (Hall & Horowitz, 2007). We will assume later that μ has mixed or anisotropic smoothness.

4. Approximation Error Analysis

In this section, we study the approximation ability of Transformers in the case that the target function has (piecewise) anisotropic or mixed smoothness. Our analysis shows that Transformer networks can approximate shift-equivariant functions under appropriate assumptions even if inputs and outputs are infinite dimensional. This is in contrast to the analysis for any continuous functions (Yun et al., 2020), where the number of parameters increases exponentially with respect to the input dimensionality.

4.1. Mixed and Anisotropic Smoothness

First, we derive the approximation error of Transformer networks for the mixed and anisotropic smoothness. In our analysis, we assume the following. Similar assumption can be found in Okumoto & Suzuki (2022).

Assumption 4.1. The true function F° is shift-equivariant and satisfies

$$F_0^\circ \in U(\mathcal{F}^\gamma_{p,\theta}),\ \|F_0\|_\infty \leq R,$$

where $R > 0$ is a constant and γ is mixed or anisotropic smoothness. In addition, the smoothness parameter a satisfies $\|a\|_{wl^\alpha} \leq 1$ for some $0 < \alpha < \infty$ and $a_{ij} = \Omega(\log(|j|+1))$. For the mixed smoothness, we also assume $\bar{a}_1 < \bar{a}_2$.

Note that the assumption implies that F_i° $(i \neq 0)$ also have mixed or anisotropic smoothness due to the shift-equivariance. The weak l^α norm condition implies the sparsity. Since the smoothness should increase in polynomial order if $\|a\|_{wl^\alpha} \leq 1$, most of a_{ij}s should be large, which means there exist few important features in an input. This assumption is partially supported by the fact that attention weights are often sparse in practice (Likhosherstov et al., 2021). On the other hand, the condition $a_{ij} = \Omega(\log(|j|+1))$ implies the locality. That is, a_{ij} should be large if $|j| \gg 1$, and thus x_j is not important. This is introduced due to the importance of local context in some applications such as natural language processing.

Under this assumption, the approximation error of Transformer networks is evaluated as follows.

Theorem 4.2. *Suppose that the target function F° satisfies Assumption 4.1. Then, for any $T > 0$, there exists a transformer network $\hat{F} \in \mathcal{T}(M, U, D, H, L, W, S, B)$ such that*

$$\left\|\hat{F}_i - F_i^\circ\right\|_{2,P_X} \lesssim 2^{-T}$$

for any $i \in \mathbb{Z}$, where $\phi = \frac{1}{2U_1+1}$, and

$$\begin{aligned} M &= 1,\ \log U_1 \sim T,\ D \sim T^{1/\alpha},\ H \sim T^{1/\alpha}, \\ L &\sim \max\left\{T^{2/\alpha}, T^2\right\},\ W \sim T^{1/\alpha} 2^{T/a^\dagger}, \\ \log B &\sim \max\left\{T^{1/\alpha}, T\right\}, \\ S &\sim T^{2/\alpha} \max\left\{T^{2/\alpha}, T^2\right\} 2^{T/a^\dagger} \\ p_i &= [0, \ldots, 0, \cos(i\phi), \sin(i\phi)]^\top. \end{aligned} \tag{3}$$

The proof can be found in Appendix E. The results show that even though inputs and outputs are infinite dimensional, the approximation error can be bounded by $N^{-a^\dagger}$ ignoring poly-log factor, where N denotes the number of parameters, since the total number of parameters is bounded by $N \lesssim M(S + HD^2) \sim 2^{T/a^\dagger}$ ignoring poly-log factor. This is in contrast to FNNs, where the number of parameters should increase at least linearly with the input and output length. The *parameter sharing* property and *feature extraction* ability of Transformers play an essential role in the proof. For anisotropic smoothnesss, the result can be seen as an extention of the result of FNN for anisotropic Besov space (Suzuki & Nitanda, 2021) to infinite dimensional input and sequence-to-sequence setting.

In addition, positional encoding is an important factor in extracting local context with limited interaction among tokens compared to FNN and CNN. Due to the shift-equivariance, Transformer networks should extract important tokens for each output by relative position. Since we use absolute positional encoding, it is not trivial to show that Transformers have such capability. Indeed, existing works (Edelman et al., 2022; Gurevych et al., 2022) used absolute position to focus tokens and their analysis cannot be applied to multiple output setting with shift-equivariance. To overcome this issue, we adopt the *sinusoidal positional encoding* since the shift operation $p_i \to p_{i+j}$ can be represented by linear transformation, as mentioned in Vaswani et al. (2017). This allows the self-attention mechanism to attend by relative position as shown in Fig. 1. In addition, the size of the positional encoding in (Edelman et al., 2022; Gurevych et al., 2022) depends on the size of the receptive field. For example, Gurevych et al. (2022) used the standard basis as positional encoding and the size of the positoinal encoding grows linearly with respect to the input length. On the other hand,

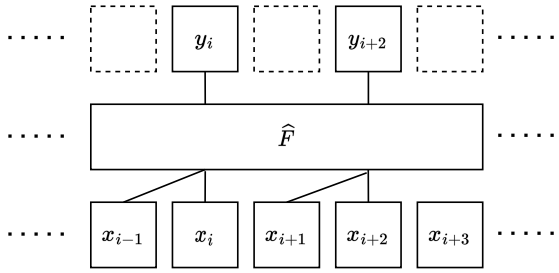

Figure 1. The self-attention mechanism can attend by relative position. In this diagram, each token attend to the previous token and itself.

we show that fixed length positional encoding is enough for feature extraction by adjusting the scale appropriately inside the self-attention mechanism.

Remark 4.3. The results in Theorem 4.2 can be extended to the 2D input setting like image processing by modifying the positional encoding as $p_{ij} = [0, \dots, 0, \cos(i\phi), \sin(i\phi), \cos(j\phi), \sin(j\phi)]$, where i, j represent the row index and column index of the token x_{ij}, respectively. That is, Transformer can approximate both vertically and horizontally shift-equivariant functions with 2D inputs under appropriate assumptions. Similarly, the other results in this paper can be extended to the 2D input setting. Therefore, to some extent, our analysis explains the practical success of Transformers in the field of image processing (Dosovitskiy et al., 2021).

4.2. Piecewise Smoothness

Next, we derive the approximation error for the piecewise mixed and anisotropic smoothness, where the smoothness depends on each input. For the piecewise smoothness, we assume the following.

Assumption 4.4. The true function F° is shift-equivariant and satisfies

$$F_0^\circ \in U(\mathcal{P}_{p,\theta}^\gamma),\ \|F_0^\circ\|_\infty \le R,$$

where R is a constant, γ is mixed or anisotropic smoothness, and the smoothness parameters a satisfies $a_{ij} = \Omega(j^\alpha)$ and $\|a\|_{wl^\alpha} \le 1$ for some $0 < \alpha < \infty$. For the mixed smoothness, we also assume $\bar{a}_1 < \bar{a}_2$. In addition, we assume the importance function μ satisfies Assumption 4.1 with $p = \infty, R = 1$.

For piecewise γ-smooth functions, it is necessary to extract important features depending on each input as shown in Fig. 2. Since Π is not a linear operator, linear dimension reduction methods such as PCA and linear convolutional layers cannot approximate Π. However, due to the dynamical feature extraction ability of self-attention mechanism, Transformers can approximate Π and achieve similar approximation error results as in the fixed smoothness setting.

Theorem 4.5. *For $d' = O(T^{2(\beta+1)/\alpha}\log V)$, let $\{u_i\}_{i=1}^{2V+1} \subset \mathbb{R}^{d'}$ be approximately orthonormal vectors which satisfies*

$$|\langle u_i, u_j\rangle| \le \varepsilon,\ \langle u_i, u_i\rangle = 1.$$

In addition, We define u_i for $i \notin [2V+1]$ by $u_i := u_{i \bmod (2V+1)}$. Suppose that the target function F° satisfies Assumption 4.4.

Then, for any $T > 0$, there exists a transformer $\hat{F} \in \mathcal{T}(M, U, D, H, L, W, S, B)$ such that

$$\left\|\hat{F}_i - F_i^\circ\right\|_{2,P_X} \lesssim 2^{-T}$$

for any $i \in \mathbb{Z}$, where

$$\begin{aligned} M &\sim T^{1/\alpha},\ D \sim T^{2(\beta+1)/\alpha}\log V, \\ \log U_1 &\sim \log T,\ U_i = V\ (i \ge 2),\ H \sim (\log T)^{1/\alpha}, \\ L &\sim \max\left\{T^{2/\alpha}, T^2\right\},\ W \sim T^{1/\alpha}2^{T/a^\dagger}, \\ \log B &\sim \max\left\{T^{1/\alpha}, T, \log\log V\right\} \\ S &\sim T^{2/\alpha}\max\left\{T^{2/\alpha}, T^2\right\}2^{T/a^\dagger}, \\ p_i &= [0, \dots, 0, 1, \cos(i\phi), \sin(i\phi), u_i^\top]^\top. \end{aligned} \tag{4}$$

See Appendix F for the proof. This theorem implies that multilayer Transformer networks can approximate mixed and anisotropic smooth functions even if inputs and outputs are infinite dimensional, and the smoothness structure depends on each input. In addition, the convergence rate is $N^{-a^\dagger}$, which is the same as in Theorem 4.2. This can be realized by the *dynamical* feature extraction ability of the self-attention mechanism. See Fig. 2 for an illustration of the feature extraction mechanism. The construction consists of two phases. First, the first layer of the Transformer network approximates the importance function μ. Next, the Transformer network selects important tokens by the self-attention mechanism based on the estimated importance. Thanks to the softmax operation in the self-attention mechanism, Transformers can extract the *most* important token. However, it is difficult to extract the second (and subsequent) most important tokens. To overcome this issue, we developed a novel approximately orthonormal basis coding technique for the positional encoding to memorize already extracted tokens. Approximately orthonormal vectors can be constructed via random sampling, as shown in Lemma C.3.

In this setting, the attention weights *dynamically* change depending on each input. This is in contrast to the existing works (Edelman et al., 2022; Okumoto & Suzuki, 2022), which studied the feature extraction ability of Transformer networks and CNNs, respectively. Our result matches the

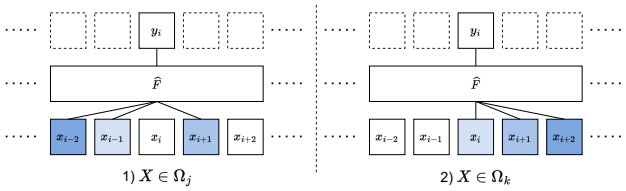

Figure 2. For piecewise γ-smoothness, the position of important tokens depends on each input. We show important tokens in darker color. In the case of $X \in \Omega_j$, the most important token to y_i is x_{i-2} and in the case of $X \in \Omega_k$, x_{i+2} is the most important. The self-attention mechanism can switch its attention (represented by black lines) depending on the importance of tokens.

empirical findings (Likhosherstov et al., 2021) and Theorem 4.5 theoretically supports the *dynamical* feature extraction ability of the self-attention mechanism by considering the novel function class.

5. Estimation Error Analysis

In this section, we show that Transformer networks can achieve polynomial estimation error rate and avoid the curse of dimensionality.

To evaluate the variance of estimators, the covering number is often used to capture the complexity of model classes.

Definition 5.1 (Covering Number)**.** For a normed space $\mathcal{F}$ with a norm $\|\cdot\|$, the δ-covering number is defined as

$$\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|) := \inf \{n \in \mathbb{N} \mid \exists (f_1, \dots, f_n) \in \mathcal{F}, \\ \forall f \in \mathcal{F}, \exists i \in [n], \|f_i - f\| \le \delta\}.$$

For non-parametric regression problems with infite dimensional inputs and outputs, the estimation error of an ERM estimator is evaluated as follows.

Theorem 5.2. *For a given class $\mathcal{F}$ of functions from $[0,1]^{d \times \infty}$ to $\mathbb{R}^\infty$, let $\hat{F} \in \mathcal{F}$ be an ERM estimator which minimizes the empirical cost. Suppose that there exists a constant $R > 0$ such that $\|F^\circ\|_\infty \le R$, $\|F\|_\infty \le R$ for any $F \in \mathcal{F}$, and $\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty) \ge 3$. Then, for any $0 < \delta < 1$, it holds that*

$$R_{l,r}(\hat{F}, F^\circ) \le 4 \inf_{F \in \mathcal{F}} \frac{1}{r-l+1} \sum_{i=l}^{r} \|F_i - F_i^\circ\|_{2,P_X}^2 \\ + C((R^2 + \sigma^2) \frac{\log \mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty)}{n} + (R + \sigma)\delta),$$

where $C > 0$ is a global constant.

This theorem is a direct extension of Lemma 4 in Schmidt-Hieber (2020) and Theorem 2.6 in Hayakawa & Suzuki (2020) to the multiple output setting. The proof can be found in Appendix G.

By carefully evaluating the complexity of the class of Transformer networks, we have the following bound on the log covering number of Transformer networks.

Theorem 5.3. *For given hyperparameters M, U, D, H, L, W, S, B, assume that $B \ge 1$ and $\|P\|_\infty \le B$. Then, we have the following log covering number bound:*

$$\log \mathcal{N}(\mathcal{T}(M, U, D, H, L, W, S, B), \delta, \|\cdot\|_\infty) \\ \lesssim M^3 L (S + HD^2) \log \left(\frac{DHLWB}{\delta} \right).$$

See Appendix H for the proof. Interestingly, the covering number bound does not depend on the dimensionality of inputs and outpus, and the width of the sliding window. This is because the number of parameters does not depend on these quantities due to the parameter sharing property and the magnitude of the hidden states are independent of the window size since the attention weights A are normalized as $\|A\|_1 = 1$. The parameter sharing property, on the other hand, leads to the limited interaction among tokens. This makes approximation analysis difficult, and thus it is necessary to design positional encoding carefully as mentioned in Section 4.

Combining above results, we have the following estimation error bound for the mixed and anisotropic smoothness.

Theorem 5.4. *Suppose that Assumption 4.1 holds. Let $\hat{F}$ be an ERM estimator in $\mathcal{T}_R(M, U, D, H, L, W, S, B)$, where M, U, D, H, L, W, S, B is defined as (3) and $T = \frac{a^\dagger}{2a^\dagger + 1} \log n$. Then, for any $l, r \in \mathbb{Z}$, we have*

$$R_{l,r}(\hat{F}, F) \lesssim n^{-\frac{2a^\dagger}{2a^\dagger + 1}} (\log n)^{2/\alpha + 2 + \max\{4/\alpha, 4\}}.$$

This can be shown by letting $\delta = 1/n$ in Theorem 5.2. See Appendix I for the proof. The results show that the convergence rate of the estimation error does not depend on the input dimensionality and the output size if the smoothness of the target function has sparse structure. This implies that Transformers can avoid the curse of dimensionality. When $d = 1$, this convergence rate matches that for CNNs (Okumoto & Suzuki, 2022) for single output setting. For anisotropic smoothness, this rate also matches, up to poly-log order, that of FNNs in the finite dimensional setting, which is known to be minimax optimal (Suzuki & Nitanda, 2021). That is, Transformers can achieve near-optimal rate in a minimax sense.

In addition, the Transformer architecture including the positional encoding P does not depend directly on the smoothness structure a. This implies that Transformer networks can find important features and select them *adaptively* to the smoothness of the target function by learning the intrinsic structure of the target function.

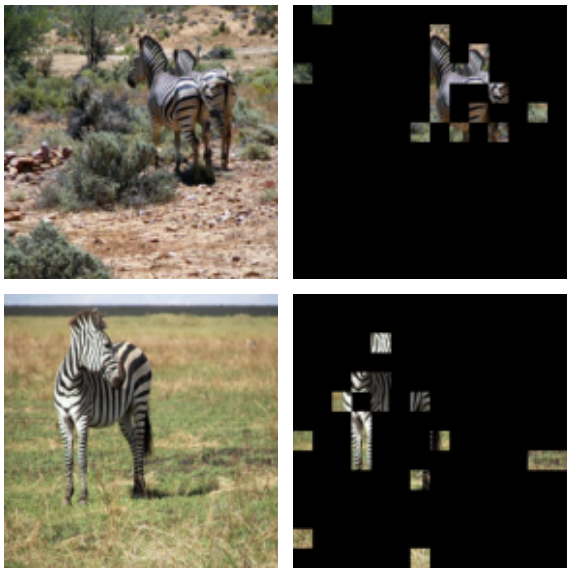

Figure 3. Two zebra images (left) and the corresponding images with 180 / 196 patches masked (right).

For the piecewise smoothness, we have the following estimation error bound.

Theorem 5.5. *Suppose that Assumption 4.4 holds. Let $\hat{F}$ be an ERM estimator in $\mathcal{T}_R(M, U, D, H, L, W, S, B)$, where M, U, D, H, L, W, S, B is defined as (4) and $T = \frac{a^\dagger}{2a^\dagger+1} \log n$. Then, for any $l, r \in \mathbb{Z}$, we have*

$$R_{l,r}(\hat{F}, F) \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}} (\log n)^{5/\alpha+2+\max\{4/\alpha,4\}} (\log V)^3.$$

See Appendix J for the proof. This convergence rate is the same as in Theorem 5.4 up to poly-log order if $V = \mathrm{poly}(n)$. This means that Transformers can avoid the curse of dimensionality even if the smoothness architecture depends on each input.

In addition, the Transformer architecture does not depend on the partition $\{\Omega_\lambda\}_{\lambda \in \Lambda}$ and the importance function μ. This means that Transformers can adapt to the intrinsic structure of the target function and realize the *dynamical* feature extraction according to the importance of tokens. This fact supports the practical success of Transformer networks in a wide range of applications with various structures.

6. Numerical Experiments

The assumptions in this paper essentially impose the sparsity of important features. To verify this, we conducted some numerical experiments using masked images as inputs. In this experiment, we prepared a pre-trained model (ViT-Base model (Dosovitskiy et al., 2021)) and two images of zebras in Fig. 3 from the validation set of ImageNet-1k (Russakovsky et al., 2015). We divided each image into 14×14 tokens and masked each token in turn. At each step, a token to be masked is selected using a greedy algorithm to maximize the predicted probability of the correct class by the pre-trained model. Since masking informative features strongly affects the predicted probability, important features will remain unmasked near the end of the procedure.

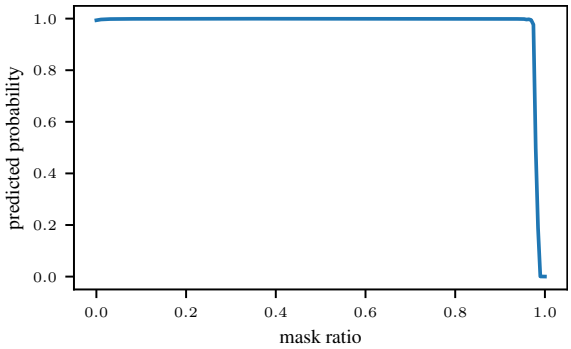

Figure 4. The predicted probability of the correct class for the top left image in Fig. 3. The predicted probability remains high even when most of the images are masked.

As shown in Fig. 4, the predicted probability remains high and the model can classify the image correctly even if about 90% of the input is masked. This means that a small fragment of the image is important for prediction. In addition, Fig. 3 demonstrates that the patterns of unmasked, i.e., important features in the two images differ significantly. This implies important features change depending on each input and the piecewise smoothness describes more practical situations than the fixed smoothness.

7. Conclusion

In this study, we have investigated the learnability of Transformer networks for sequence-to-sequence functions with infinite dimensional inputs. We have shown that Transformer networks can achieve a polynomial order convergence rate of estimation error when the smoothness of the target function has sparse structure. In addition, we have considered the situations where the smoothness depends on each input. Then, we have shown that Transformer networks can avoid the curse of dimensionality by switching the focus of attention based on input. We believe that our theoretical analysis provides a new insight into the nature of Transformer architecture.

Acknowledgements

ST was partially supported by Fujitsu Ltd. TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST.

References

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer, 2020. arXiv:2004.05150 [cs].

Chen, M., Jiang, H., Liao, W., and Zhao, T. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Dong, L., Xu, S., and Xu, B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive Biases and Variable Creation in Self-Attention Mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

Gurevych, I., Kohler, M., and Şahin, G. G. On the rate of convergence of a classifier based on a Transformer encoder. *IEEE Transactions on Information Theory*, 2022.

Hall, P. and Horowitz, J. L. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1), 2007.

Hayakawa, S. and Suzuki, T. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

Huang, X. S., Perez, F., Ba, J., and Volkovs, M. Improving Transformer Optimization Through Better Initialization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4475–4483. PMLR, 2020.

Imaizumi, M. and Fukumizu, K. Deep Neural Networks Learn Non-Smooth Functions Effectively. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 869–878. PMLR, 2019.

Jelassi, S., Sander, M. E., and Li, Y. Vision Transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*, 2022.

Lafferty, J., Liu, H., and Wasserman, L. Concentration on measure, 2008. URL http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.

Likhosherstov, V., Choromanski, K., and Weller, A. On the Expressive Power of Self-Attention Matrices, 2021. arXiv:2106.03764 [cs].

Nakada, R. and Imaizumi, M. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *The Journal of Machine Learning Research*, 21(1):174:7018–174:7055, 2022.

Nikol'skii, S. M. *Approximation of functions of several variables and imbedding theorems*, volume 205. Springer-Verlag Berlin Heidelberg, 1975.

Okumoto, S. and Suzuki, T. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022.

Petersen, P. and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Pérez, J., Marinković, J., and Barceló, P. On the Turing Completeness of Modern Neural Network Architectures. In *International Conference on Learning Representations*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Schmeisser, H. J. An unconditional basis in periodic spaces with dominating mixed smoothness properties. *Analysis Mathematica*, 13(2):153–168, 1987.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 2020.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-Attention with Relative Position Representations, 2018. arXiv:1803.02155 [cs].

Suzuki, T. 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*, 2018.

Suzuki, T. and Nitanda, A. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Wei, C., Chen, Y., and Ma, T. Statistically Meaningful Approximation: a Case Study on Approximating Turing Machines with Transformers, 2021. arXiv:2107.13163 [cs, stat].

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020.

Zhang, Y., Liu, B., Cai, Q., Wang, L., and Wang, Z. An Analysis of Attention via the Lens of Exchangeability and Latent Variable Models, 2022. arXiv:2212.14852 [cs].

A. Notation List

Table 1. Notation list

Notation	Definition
n	sample size
d	token dimension
$(X^{(i)}, y^{(i)})$	i-th observation
$\mathcal{D}^n = \left\{(X^{(i)}, y^{(i)})\right\}_{i=1}^n$	training data
P_X	data distribution
Ω	support of P_X
$\{\Omega_\lambda\}_{\lambda\in\Lambda}$	disjoint partition of Ω
$\mathcal{F}_{p,\theta}^{\gamma}$	γ-smooth function class
$\mathcal{G}_{p,\theta}^{\gamma}$	piecewise γ-smooth function class
σ	noise variance
F°	true function
a	smoothness parameter
$\bar{a} = [a_{i_1,j_1}, \dots, a_{i_k,j_k}, \dots]$	smoothness parameter sorted in the ascending order
$\tilde{a}$	$\left(\sum_{i=1}^\infty \bar{a}_i^{-1}\right)^{-1}$
$a^\dagger$	$\bar{a}_1$ for the mixed smoothness and $\tilde{a}$ for the anisotropic smoothness
μ	importance function
η	ReLU activation function
$I(T,\gamma)$	$\left\{(i,j) \mid \exists s \in \mathbb{N}_0^{d\times\infty}, s_{ij} \neq 0, \gamma(s) < T\right\}$
$I_j(T,\gamma)$	$\{i \mid (i,j) \in I(T,\gamma)\}$
$e_i \in \mathbb{R}^{[l:r]}$ $(i \in [l:r])$	standard basis of $\mathbb{R}^{[l:r]}$
$\delta_{i,j} \in \mathbb{R}^{[l_1:r_1]\times[l_2:r_2]}$ $(i \in [l_1, r_1],\ j \in [l_2 : r_2])$	$\delta_{i,j} := e_i e_j^\top$
$x \overset{\epsilon}{\approx} y$ $(x, y \in \mathbb{R}^d)$	$\|x - y\|_2 \lesssim \epsilon$
1_A	1 if A is true and 0 if A is false

B. Extension to the Original Architecture

In this paper, we consider the multilayer FNN without skip connection while the FNN in the original architecture (Vaswani et al., 2017) uses single hidden layer with skip connection. However, our argument can be applied to the original structure with slight modifications as follows.

- (Skip connection) Since there exists an FNN $f : \mathbb{R}^d \to \mathbb{R}^d$ with one hidden layer of width $2d$ which works as an identity map, we can cancel out the skip connection using f. This modification does not change the order of hyperparameters, and we can obtain the same convergence rate as the original one.

- (One hidden layer) By letting $V_h = O$ in a self-attention layer, the self-attention layer behaves as an identity map due to the skip connection in the self-attention layer. Therefore, L-layer Transformer blocks (one block consists of an FNN layer with one hidden layer and a self-attention layer) can represent an L-layer FNN. Then, we can obtain the same convergence rate as the original one up to poly-log order.

C. Auxiliary Lemmas

Lemma C.1. *For $\theta \in \mathbb{R}^d$, assume that there exist an index $i^* \in [d]$ and $\delta > 0$ such that $\theta_{i^*} \geq \theta_i + \delta$ for any $i \neq i^*$. Then, we have*

$$\|\mathrm{Softmax}(\theta) - e_{i^*}\|_1 \leq 2de^{-\delta}.$$

Proof. For $i \neq i^*$, the assumption $\theta_{i^*} \geq \theta_i + \delta$ yields that

$$0 \leq \mathrm{Softmax}(\theta)_i = \frac{e^{\theta_i}}{\sum_{j=1}^d e^{\theta_j}} \leq e^{\theta_i - \theta_{i^*}} \leq e^{-\delta}.$$

For i^*, we have

$$0 \leq 1 - \mathrm{Softmax}(\theta)_{i^*} = \sum_{i \neq i^*} \mathrm{Softmax}(\theta)_i \leq (d-1)e^{-\delta}.$$

Therefore,

$$\|\mathrm{Softmax}(\theta) - e_{i^*}\|_1 = |1 - \mathrm{Softmax}(\theta)_{i^*}| + \sum_{i \neq i^*} |\mathrm{Softmax}(\theta)_i| \leq 2de^{-\delta},$$

which completes the proof. □

Lemma C.2. *For any $\theta, \theta' \in \mathbb{R}^d$, we have*

$$\|\mathrm{Softmax}(\theta) - \mathrm{Softmax}(\theta')\|_1 \leq 2\|\theta - \theta'\|_\infty.$$

Proof. See Corollary A.7 in Edelman et al. (2022). □

Lemma C.3. *For any $l \in \mathbb{N}$ and $\varepsilon > 0$, there exists ε-approximately orthonormal vectors $\{u_i\}_{i=1}^l \subset \mathbb{R}^d$ ($d = O\left(\frac{\log l}{\varepsilon^2}\right)$) which satisfies*

$$\begin{aligned} |\langle u_i, u_j \rangle| &\leq \varepsilon, \\ \langle u_i, u_i \rangle &= 1, \end{aligned} \tag{5}$$

for any $i \neq j$.

Proof. Let each component of $u_i \in \mathbb{R}^d$ independently follows Bernoulli distribution $\Pr((u_i)_j = \pm 1/\sqrt{d}) = 1/2$. Then, we have $\|u_i\| = 1$. Since $\langle u_i, u_j \rangle$ can be seen as the sum of independent Bernoulli random variables, a Chernoff bound implies

$$\Pr(|\langle u_i, u_j \rangle| > \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2}d\right).$$

By letting $d = \frac{4\log 2l}{\varepsilon^2}$, we have

$$\Pr(|\langle u_i, u_j \rangle| > \varepsilon) \leq 2\exp(-2\log 2l) = \frac{1}{2l^2}.$$

Since there exists at most $l^2/2$ pairs (i, j), from the union bound, the probability that $|\langle u_i, u_j \rangle| \leq \varepsilon$ for all (i, j) such that $i \neq j$ is greater than $1 - \frac{l^2}{2} \cdot \frac{1}{2l^2} = 3/4$. This means there exist approximately orthonormal vectors which satisfy Eq. (5). □

Lemma C.4. *For $B \geq 1$, let $f \in \Psi(L, W, B, S)$, and $g \in \mathcal{A}(U, D, H, B)$. Then, for any $r \geq 1$, $x, x' \in \mathbb{R}^d$ and $X, X' \in [-r, r]^{d \times \infty}$, we have*

$$\begin{aligned} \|f(x) - f(x')\|_\infty &\leq (BW)^L \|x - x'\|_\infty \leq (6HDBW)^{4L} r^2 \|x - x'\|_\infty, \\ \|g(X) - g(X')\|_\infty &\leq 6HB^3D^4r^2\|X - X'\|_\infty \leq (6HDBW)^{4L} r^2 \|X - X'\|_\infty. \end{aligned} \tag{6}$$

Proof. Since the ReLU activation function η is 1-Lipschitz continuous and $\|W_i x + b - (W_i x' + b)\|_\infty \leq BW\|x - x'\|_\infty$, f is $(BW)^L$-Lipschitz continuous. This implies Eq. (6).

Let $\bar{X} := X[-U:U]$ and $A_h := \mathrm{Softmax}((K_h\bar{X})^\top(Q_h x_i))$. We define $\bar{X}', A'_h$ in the same way as $\bar{X}, A_h$. Then, we have

$$\begin{aligned}\|g(X)_i - g(X')_i\|_\infty &\le \|x_i - x'_i\|_\infty + \sum_{h=1}^{H} \left\|V_h\bar{X}A_h - V\bar{X}'A'_h\right\|_\infty \\ &\le \|x_i - x'_i\|_\infty + \sum_{h=1}^{H} \left\|V_h\bar{X}A_h - V\bar{X}A'_h\right\|_\infty + \left\|V\bar{X}A'_h - V\bar{X}'A'_h\right\|_\infty.\end{aligned}$$

From Lemma C.2, we have

$$\begin{aligned}\|A_h - A'_h\|_1 &\le 2\left\|(K_h\bar{X})^\top(Q_h x_i) - (K\bar{X}')^\top(Q_h x'_i)\right\|_\infty \\ &\le 2\left\|(K_h\bar{X})^\top(Q_h x_i) - (K\bar{X})^\top(Q_h x'_i)\right\|_\infty + 2\left\|(K_h\bar{X})^\top(Q_h x'_i) - (K_h\bar{X}')^\top(Q_h x'_i)\right\|_\infty \\ &\le 4rB^2D^3\|X_i - X'_i\|_\infty.\end{aligned}$$

Here, we used the fact that $\|K_h\|_\infty, \|Q_h\|_\infty, \|V_h\|_\infty \le B$ and $\|X\|_\infty \le r$. Then, it holds that

$$\begin{aligned}\left\|V_h\bar{X}A_h - V_h\bar{X}A'_h\right\|_\infty &\le BDr\|A_h - A'_h\|_1 \\ &\le 4B^3D^4r^2\|X_i - X'_i\|_\infty, \\ \left\|V_h\bar{X}A'_h - V_h\bar{X}'A'_h\right\|_\infty &\le BD\|X - X'\|_\infty.\end{aligned}$$

Since $B \ge 1, r \ge 1$, we have

$$\begin{aligned}\|g(X)_i - g(X')_i\|_\infty &\le \left(1 + 4B^3D^4r^2 + BD\right)\|X - X'\|_\infty \\ &\le 6HB^3D^4r^2\|X - X'\|_\infty,\end{aligned}$$

which completes the proof. □

Lemma C.5. *Let $f \in \Psi(L, W, S, B)$ and $g \in \mathcal{A}(U, D, H, B)$ for $B \ge 1$. For any $r \ge 1$, X ($\|X\|_\infty \le r$), and x ($\|x\|_\infty \le r$), we have*

$$\begin{aligned}\|f(x)\|_\infty &\le (2BW)^L r \le (6HDBW)^L r, \\ \|g(X)\|_\infty &\le 2HBDr \le (6HDBW)^L r.\end{aligned}$$

Proof. For the FNN layer f, since $\|\eta(z)\|_\infty \le \|z\|_\infty$ and

$$\begin{aligned}\|Wz + b\|_\infty &\le BWr' + B \\ &\le 2BWr'\end{aligned}$$

for $\|z\|_\infty \le r'$ ($r' \ge 1$), we have

$$\|f(X)\|_\infty \le (2BW)^L r$$

by induction.

For attention layer g, define $\bar{X}$ and A_h as in the proof of Lemma C.4. Then, we have

$$\begin{aligned}\|g_i(X)\|_\infty &\le \|x_i\|_\infty + \sum_{i=1}^{H} \left\|V\bar{X}A_h\right\|_\infty \\ &\le \|X\|_\infty + HBD\|X\|_\infty\|A_h\|_1 \\ &\le 2HBDr,\end{aligned}$$

since $\|A_h\|_1 = 1$. This completes the proof. □

Lemma C.6. *Define $f, \tilde{f} \in \Psi(L, W, B, S)$ by*

$$f(x) := (A_L \cdot + b_L) \circ \cdots \circ (A_1 x + b_1)$$
$$\tilde{f}(x) := (\tilde{A}_L \cdot + \tilde{b}_L) \circ \cdots \circ (\tilde{A}_1 x + \tilde{b}_1),$$

where $\left\|A_i - \tilde{A}_i\right\|_\infty \leq \delta$, $\left\|b_i - \tilde{b}_i\right\|_\infty \leq \delta$ for a given $\delta > 0$. For any $r \geq 1$ and x ($\|x\|_\infty \leq r$), we have

$$\left\|f(x) - \tilde{f}(x)\right\|_\infty \leq 2(2BW)^L \delta r \leq (6HDBW)^{4L} \delta r^3.$$

In addition, define $g, \hat{g} \in \mathcal{A}(U, D, H, B, S)$ by

$$g(X)_i := x_i + \sum_{i=1}^{H} V_h X[-U:U] \operatorname{Softmax}((K_h X[-U:U])^\top (Q_h x_i))$$
$$\tilde{g}(X)_i := x_i + \sum_{i=1}^{H} \tilde{V}_h X[-U:U] \operatorname{Softmax}((\tilde{K}_h X[-U:U])^\top (\tilde{Q}_h x_i)),$$

where $\left\|K_h - \tilde{K}_h\right\|_\infty \leq \delta$, $\left\|Q_h - \tilde{Q}_h\right\|_\infty \leq \delta$, and $\left\|V_h - \tilde{V}_h\right\|_\infty \leq \delta$ for a given $\delta > 0$. For any $r \geq 1$ and X ($\|X\|_\infty \leq r$), we have

$$\|g(X) - \tilde{g}(X)\|_\infty \leq 5HB^2D^4r^3\delta \leq (6HDBW)^{4L}\delta r^3.$$

Proof. By the same argument as Lemma 3 in Suzuki (2018), we have

$$\left\|f(x) - \tilde{f}(x)\right\|_\infty \leq 2(2BW)^L r\delta.$$

Define $\bar{X}, A_h, \tilde{A}_h$ as in Lemma C.4. From Lemma C.2, we have

$$\begin{aligned}
\left\|A_h - \tilde{A}_h\right\|_1 &\leq 2\left\|(K_h\bar{X})^\top(Q_h x_i) - (\tilde{K}_h\bar{X})^\top(\tilde{Q}_h x_i)\right\|_\infty \\
&\leq 2\left\|(K_h\bar{X})^\top(Q_h x_i) - (K_h\bar{X})^\top(\tilde{Q}_h x_i)\right\|_\infty + 2\left\|(K_h\bar{X})^\top(\tilde{Q}_h x_i) - (\tilde{K}_h\bar{X})^\top(\tilde{Q}_h x_i)\right\|_\infty \\
&\leq 4(BD^3r^2\delta).
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\|g(X)_i - \tilde{g}(X)_i\|_\infty &\leq \sum_{i=1}^{H} \left\|V_h\bar{X}A_h - \tilde{V}_h\bar{X}\tilde{A}_h\right\|_\infty \\
&\leq \sum_{i=1}^{H} \left\|V_h\bar{X}(A_h - \tilde{A}_h)\right\|_\infty + \left\|(V_h - \tilde{V}_h)\bar{X}\tilde{A}_h\right\|_\infty \\
&\leq \sum_{i=1}^{H} \|V_h\bar{X}\|_\infty \left\|A_h - \tilde{A}_h\right\|_1 + \left\|(V_h - \tilde{V}_h)\bar{X}\right\|_\infty \\
&\leq H(4(B^2D^4r^3\delta) + D\delta r) \\
&\leq 5HB^2D^4r^3\delta,
\end{aligned}$$

which completes the proof. □

Lemma C.7. *Assume that positive and monotonically non-decreasing sequences $\bar{a} = \{\bar{a}_i\}_{i=1}^\infty$ and $\bar{a}' = \{\bar{a}'_i\}_{i=1}^\infty$ satisfies $\bar{a}_1 \geq 1 = \bar{a}'_1$ and*

$$\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-(\bar{a}_i - \bar{a}_1)}} < \infty,$$
$$\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-\beta(\bar{a}_i - \bar{a}'_i)}} < \infty,$$

for a positive constant β. Then, we have

$$\sum_{s\in\mathbb{N}_0^\infty:\langle \bar{a}',s\rangle\geq T} 2^{-\beta\langle \bar{a},s\rangle} \leq (1-2^{-\beta})^{-1}\left(\prod_{i=2}^{\infty}\frac{1}{1-2^{-\beta(\bar{a}_i-\bar{a}'_i)}}\right)2^{-\beta T},$$
$$\sum_{s\in\mathbb{N}_0^\infty:\langle \bar{a},s\rangle< T} 2^{s} \leq 8\left(\prod_{i=2}^{\infty}\frac{1}{1-2^{-(\bar{a}_i-\bar{a}_1)}}\right)2^{T}.$$

Proof. See Lemma 18 in Okumoto & Suzuki (2022). □

D. Approximation ability of FNN

In this section, we show that an FNN can approximate (piecewise) γ-smooth function if important features are extracted properly. For a general γ-smooth function class, we have the following approximation error bound.

Lemma D.1. *For $\gamma : \mathbb{N}_0^{d\times\infty} \to \mathbb{R}_{>0}$, let*

$$\begin{aligned}
G(T,\gamma) &:= \sum_{s\in\mathbb{N}_0^{d\times\infty}:\gamma(s)<T} 2^s,\\
f_{\max}(T,\gamma) &:= \max_{s\in\mathbb{N}_0^{d\times\infty}:\gamma(s)<T}\ \max_{i\in[d],j\in\mathbb{Z}} s_{ij},\\
I(T,\gamma) &:= \left\{(i,j)\mid \exists s\in\mathbb{N}_0^{d\times\infty}, s_{ij}\neq 0, \gamma(s)<T\right\},\\
d_{\max}(T,\gamma) &:= |I(T,\gamma)|.
\end{aligned}$$

Assume that γ' satisfies $\gamma'(s)<\gamma(s)$ and the target function $f\in\mathcal{F}_{p,\theta}^{\gamma}([0,1]^{d\times\infty})(p\geq 2,\theta\geq 1)$ satisfies $\|f\|_\infty\leq R$ for a constant $R>0$.

For given $T>0$, let

$$(d_{\max}, f_{\max}, G) := \begin{cases}(d_{\max}(T,\gamma), f_{\max}(T,\gamma), G(T,\gamma)), & (\theta=1),\\ (d_{\max}(T,\gamma'), f_{\max}(T,\gamma'), G(T,\gamma')), & (\theta>1),\end{cases}$$

and

$$\begin{aligned}
L &:= 2K\max\left\{d_{\max}^2, T^2, (\log G)^2, \log f_{\max}\right\},\\
W &:= 21 d_{\max} G,\\
S &:= 1764 K d_{\max}^2 \max\left\{d_{\max}^2, T^2, (\log G)^2, \log f_{\max}\right\} G,\\
B &:= 2^{d_{\max}/2} K',
\end{aligned} \tag{7}$$

for positive constants K, K' which depend on only R. Then, there exists an FNN $\hat{f}_T\in\Psi(L,W,B,S)$ such that

$$\left\|f-\hat{f}_T\circ\Gamma\right\|_{2,P_X} \lesssim \begin{cases}2^{-T}\|f\|_{\mathcal{F}_{p,\theta}^{\gamma}}, & (\theta=1),\\ \left(\sum_{T\leq\gamma(s)} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta} 2^{-T}\|f\|_{\mathcal{F}_{p,\theta}^{\gamma}}, & (\theta>1),\end{cases}$$

where $\Gamma:\mathbb{R}^{d\times\infty}\to\mathbb{R}^{d_{\max}}$ is a feature extractor defined by

$$\Gamma(X) := [X_{i_1,j_1},\ldots,X_{i_{d_{\max}},j_{d_{\max}}}] \tag{8}$$

for $X\in\mathbb{R}^{d\times\infty}$, and $I(T,\gamma)=\{(i_1,j_1),\ldots,(i_{d_{\max}},j_{d_{\max}})\}$.

Proof. Define f_T by $f_T=\sum_{\gamma(s)<T}\delta_s(f)$ for $\theta=1$ and $f_T=\sum_{\gamma'(s)<T}\delta_s(f)$ for $\theta>1$. Then, for a given $\hat{f}\in\Psi(L,W,S,B)$, we have

$$\left\|f-\hat{f}\circ\Gamma\right\|_{2,P_X} \leq \|f-f_T\|_{2,P_X} + \left\|f_T-\hat{f}\circ\Gamma\right\|_{2,P_X}.$$

First, we evaluate $\|f - f_T\|_{2,P_X}$. By the Cauchy-Schwartz inequality, we have, for $p > 2$,

$$\begin{aligned}\|\delta_s\|_{2,P_X}^2 &= \int |\delta_s(f)|^2 \,\mathrm{d}P_X \\ &\le \left(\int \left(|\delta_s(f)|^2\right)^{p/2} \mathrm{d}P_X\right)^{2/p} \left(\int 1 \,\mathrm{d}P_X\right)^{1-2/p} \\ &= \|\delta_s(f)\|_{p,P_X}^2.\end{aligned}$$

Therefore, for any $p \ge 2$, it holds that

$$\|\delta_s(f)\|_{2,P_X} \le \|\delta_s(f)\|_{p,P_X}.$$

In the case of $\theta = 1$, we have

$$\begin{aligned}\|f - f_T\|_{2,P_X} &= \left\|\sum_{\gamma(s)\ge T} \delta_s(f)\right\|_{2,P_X} \\ &\le \sum_{\gamma(s)\ge T} \|\delta_s(f)\|_{2,P_X} \\ &\le \sum_{\gamma(s)\ge T} \|\delta_s(f)\|_{p,P_X} \\ &= \sum_{\gamma(s)\ge T} 2^{\gamma(s)} 2^{-\gamma(s)} \|\delta_s(f)\|_{p,P_X} \\ &\le 2^{-T} \sum_{\gamma(s)\ge T} 2^{\gamma(s)} \|\delta_s(f)\|_{p,P_X} \\ &\le 2^{-T} \|f\|_{\mathcal{F}_{p,\theta}^{\gamma}}.\end{aligned}$$

In the case of $\theta > 1$, we have

$$\begin{aligned}\|f - f_T\|_{2,P_X} &\le \sum_{\gamma'(s)\ge T} \|\delta_s(f)\|_{p,P_X} \\ &= \sum_{\gamma'(s)\ge T} 2^{-\gamma'(s)} 2^{\gamma'(s)-\gamma(s)} 2^{\gamma(s)} \|\delta_s(f)\|_{p,P_X} \\ &\le 2^{-T} \sum_{\gamma'(s)\ge T} 2^{\gamma'(s)-\gamma(s)} 2^{\gamma(s)} \|\delta_s(f)\|_{p,P_X} \\ &\le 2^{-T} \left(\sum_{\gamma'(s)\ge T} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta} \left(\sum_{\gamma'(s)\ge T} 2^{\theta\gamma(s)} \|\delta_s(f)\|_{p,P_X}^{\theta}\right)^{1/\theta} \\ &\le 2^{-T} \left(\sum_{\gamma'(s)\ge T} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta} \|f\|_{\mathcal{F}_{p,\theta}^{\gamma}}.\end{aligned}$$

For the third inequality, we used Hölder inequality. Combining these two cases, we obtain

$$\|f - f_T\|_{2,P_X} \le \begin{cases} 2^{-T}\|f\|_{p,\theta}^{\gamma} & (\theta = 1), \\ \left(\sum_{\gamma'(s)\ge T} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta} 2^{-T}\|f\|_{p,\theta}^{\gamma} & (\theta > 1). \end{cases} \tag{9}$$

From Lemma 17 in (Okumoto & Suzuki, 2022), there exists an FNN $\hat{f}_T \in \Psi(L, W, S, B)$ such that

$$\left\|f_T - \hat{f}_T \circ \Gamma\right\|_{\infty} \le 2^{-T}. \tag{10}$$

Combining Eq. (9) and (10), we have

$$\left\|f-\hat{f}_T\circ\Gamma\right\|_{2,P_X}\lesssim\begin{cases}2^{-T}\|f\|_{p,\theta}^{\gamma} & (\theta=1),\\ \left(\sum_{\gamma'(s)\geq T}2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta}2^{-T}\|f\|_{p,\theta}^{\gamma} & (\theta>1).\end{cases}$$

This completes the proof. □

In addition, for a general piecewise γ-smooth function class, we have the following approximation error bound.

Lemma D.2. *For $\gamma:\mathbb{N}_0^{d\times(2V+1)}\to\mathbb{R}_{>0}$, let*

$$\begin{aligned}G(T,\gamma)&:=\sum_{s\in\mathbb{N}_0^{d\times(2V+1)}:\gamma(s)<T}2^s,\\ f_{\max}(T,\gamma)&:=\max_{s\in\mathbb{N}_0^{d\times(2V+1)}:\gamma(s)<T}\max_{i\in[d],j\in[2V+1]}s_{ij},\\ I(T,\gamma)&:=\left\{(i,j)\mid s\in\mathbb{N}_0^{d\times(2V+1)},s_{ij}\neq0,\gamma(s)<T\right\},\\ d_{\max}(T,\gamma)&:=|I(T,\gamma)|.\end{aligned}$$

Assume that γ' satisfies $\gamma'(s)<\gamma(s)$ and the target function $f\in\mathcal{P}_{p,\theta}^{\gamma}(p\geq2,\theta\geq1)$ satisfies $\|f\|_\infty\leq R$ for a constant $R>0$.

For given $T>0$, let

$$(d_{\max},f_{\max},G):=\begin{cases}(d_{\max}(T,\gamma),f_{\max}(T,\gamma),G(T,\gamma)), & (\theta=1),\\ (d_{\max}(T,\gamma'),f_{\max}(T,\gamma'),G(T,\gamma')), & (\theta>1),\end{cases}$$

and define L,W,S,B by Eq. (7). Then, there exists an FNN $\hat{f}_T\in\Psi(L,W,B,S)$ such that

$$\left\|f-\hat{f}_T\circ\Gamma\circ\Pi\right\|_{2,P_X}\lesssim\begin{cases}2^{-T}\|f\|_{\mathcal{P}_{p,\theta}^{\gamma}}, & (\theta=1),\\ \left(\sum_{T\leq\gamma(s)}2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta}2^{-T}\|f\|_{\mathcal{P}_{p,\theta}^{\gamma}}, & (\theta>1),\end{cases}$$

where $\Gamma:\mathbb{R}^{d\times(2V+1)}\to\mathbb{R}^{d_{\max}}$ is a feature extractor defined by

$$\Gamma(X):=[X_{i_1,j_1},\ldots,X_{i_{d_{\max}},j_{d_{\max}}}]\tag{11}$$

for $X\in\mathbb{R}^{d\times(2V+1)}$, and $I(T,\gamma)=\{(i_1,j_1),\ldots,(i_{d_{\max}},j_{d_{\max}})\}$.

Proof. From the definition of $\mathcal{P}_{p,\theta}^{\gamma}$, there exist $f'\in\mathcal{F}_{p,\theta}^{\gamma}([0,1]^{d\times[2V+1]})$ such that $f=f'\circ\Pi$. Define f_T by $f_T:=\sum_{\gamma(s)<T}\delta_s(f')$ for $\theta=1$ and $f_T:=\sum_{\gamma'(s)<T}\delta_s(f')$ for $\theta>1$. By the same argument as in the case of γ-smoothness, we have

$$\|f-f_T\circ\Pi\|_{2,P_X}\leq2^{-T}\|f\|_{\mathcal{P}_{p,\theta}^{\gamma}}$$

for $\theta=1$, and

$$\|f-f_T\circ\Pi\|_{2,P_X}\leq2^{-T}\left(\sum_{\gamma'(s)\geq T}2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))}\right)^{1-1/\theta}\|f\|_{\mathcal{P}_{p,\theta}^{\gamma}}$$

for $\theta>1$. By the same argument as in Lemma 17 in (Okumoto & Suzuki, 2022), there exists an FNN $\hat{f}_T\in\Psi(L,W,S,B)$ such that

$$\left\|f_T-\hat{f}_T\circ\Gamma\right\|_\infty\leq2^{-T}.$$

This implies

$$\left\| f_T \circ \Pi - \hat{f}_T \circ \Gamma \circ \Pi \right\|_\infty \leq 2^{-T}.$$

Therefore, we have

$$\begin{aligned}\left\| f - \hat{f}_T \circ \Gamma \circ \Pi \right\|_{2,P_X} &\leq \| f - f_T \circ \Pi \|_{2,P_X} + \left\| f_T \circ \Pi - \hat{f}_T \circ \Gamma \circ \Pi \right\|_\infty \\ &\lesssim \begin{cases} 2^{-T} \| f \|_{\mathcal{P}^\gamma_{p,\theta}}, & (\theta = 1), \\ \left(\sum_{T \leq \gamma(s)} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))} \right)^{1-1/\theta} 2^{-T} \| f \|_{\mathcal{P}^\gamma_{p,\theta}}, & (\theta > 1), \end{cases}\end{aligned}$$

which completes the proof. □

Next, we evaluate the approximation ability of FNN for the (piecewise) mixed and anisotropic smoothness when the smoothness parameter a satisfies $\|a\|_{wl^\alpha} \leq 1$.

Theorem D.3. *Suppose that the target functions $f \in \mathcal{F}^\gamma_{p,\theta}$ and $g \in \mathcal{P}^\gamma_{p,\theta}$ satisfy $\|f\|_\infty \leq R$ and $\|g\|_\infty \leq R$, where $R > 0$ and γ is the mixed or anisotropic smoothness and the smoothness parameter a satisfies $\|a\|_{wl^\alpha} \leq 1$. For any $T > 0$, there exist FNNs $\hat{f}_T, \hat{g}_T \in \Psi(L, W, S, B)$ such that*

$$\begin{aligned}\left\| \hat{f}_T \circ \Gamma - f \right\|_{2,P_X} &\lesssim 2^{-T}, \\ \| \hat{g}_T \circ \Gamma \circ \Pi - g \|_{2,P_X} &\lesssim 2^{-T},\end{aligned}$$

where

$$\begin{aligned}&L \sim \max\left\{ T^{2/\alpha}, T^2 \right\}, W \sim T^{1/\alpha} 2^{T/a^\dagger}, \\ &S \sim T^{2/\alpha} \max\left\{ T^{2/\alpha}, T^2 \right\} 2^{T/a^\dagger}, \log B \sim T^{1/\alpha}.\end{aligned}$$

The result for the fixed smoothness case is similar to Theorem 7 in Okumoto & Suzuki (2022), but we omit the case $1 \leq p \leq 2$ since we relax the assumption $\left\| \frac{\mathrm{d}P_X}{\mathrm{d}\lambda} \right\|_\infty < \infty$, which is imposed in the previous work.

Proof. We show the result for the mixed smoothness and anisotropic smoothness separately.

Mixed smoothness Let us consider the case $\theta = 1$. Since $\|a\|_{wl^\alpha} \leq 1$, $\bar{a}_j \geq j^\alpha$ for any $j \in \mathbb{N}$. Therefore, we see that $d_{\max} \sim T^{1/\alpha}$ and $f_{\max} \sim T$. In addition, from Lemma C.7, we have

$$G(T, \gamma) = \sum_{\langle a/\bar{a}_1, s \rangle < T/\bar{a}_1} 2^s \lesssim 2^{T/\bar{a}_1}.$$

Therefore, from Lemmas D.1 and D.2, there exists $\hat{f}_T, \hat{g}_T \in \Psi(L, W, S, B)$ such that

$$\begin{aligned}\left\| \hat{f}_T \circ \Gamma - f \right\|_{2,P_X} &\lesssim 2^{-T}, \\ \| \hat{g}_T \circ \Gamma \circ \Pi - g \|_{2,P_X} &\lesssim 2^{-T},\end{aligned}$$

where L, W, S, B is defined in Eq. (7).

In the case of $\theta > 1$, let $\bar{a}'_1 = a_1/2$ and $\bar{a}'_i = \bar{a}_i/u$ $(i \geq 2)$ for $2 < u < 2 + \frac{2\delta}{\bar{a}_1}$, where $\delta = \bar{a}_1 - \bar{a}_2 > 0$. Then, $\bar{a}'$ is a positive monotonically increasing sequence. Since $\bar{a}_i \geq i^\alpha$, we have, for any $c > 0$,

$$\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-c\left(\bar{a}_i/\bar{a}_1 - 2\bar{a}'_i/\bar{a}_1\right)}} < \infty.$$

Therefore, by adjusting the scale of $\bar{a}, \bar{a}'$, Lemma C.7 yields that

$$\sum_{\langle a', s\rangle \geq T} 2^{\frac{\theta}{\theta-1}(\gamma'(s)-\gamma(s))} \lesssim 2^{-\frac{\theta}{\theta-1}T},$$
$$G(T, \gamma') \lesssim 2^{T/\bar{a}'_1} = 2^{2T/\bar{a}_1}.$$

Therefore, from Lemma D.1, there exists $\hat{f}_T, \hat{g}_T \in \Psi(L, W, S, B)$ such that

$$\left\|\hat{f}_T \circ \Gamma - f\right\|_{2,P_X} \lesssim \left(2^{-\frac{\theta}{\theta-1}T}\right)^{1-1/\theta} 2^{-T} = 2^{-2T},$$
$$\|\hat{g}_T \circ \Gamma \circ \Pi - g\|_{2,P_X} \lesssim \left(2^{-\frac{\theta}{\theta-1}T}\right)^{1-1/\theta} 2^{-T} = 2^{-2T}.$$

By replacing $2T \leftarrow T$, we obtain the result.

Anisotropic smoothness For $s \in \mathbb{N}_0^\infty$, define $\bar{s}$ as the sequence rearranged in the same way as $\bar{a}$. Since $\gamma(s) \leq T$ is equivalent to $\bar{s}_i \leq T/\bar{a}_i$ for any $i \in \mathbb{N}$, we have

$$\sum_{\gamma(s)<T} 2^s \leq \prod_{i=1}^{\infty} \left(\sum_{\bar{s}_i=0}^{\lceil T/\bar{a}_i \rceil} 2^{\bar{s}_i} \right) \leq 2^{\sum_{i=1}^{\infty} \lceil T/\bar{a}_i \rceil} \sim 2^{T/\tilde{a}}.$$

Then, by the same argument as in the case of mixed smoothness, we obtain the result. □

E. Proof of Theorem 4.2

First, we construct the embedding layer Enc_P. Let D be $d + d_{\max} + 2$ and the embedding matrix $E \in \mathbb{R}^{D\times d}$ be the matrix such that $Ex = [x_1, \dots, x_d, 0, \dots, 0]^\top \in \mathbb{R}^D$ for any $x \in \mathbb{R}^d$. Note that $\|E\|_\infty = 1$. Define $Z^{(m)}$ by

$$\begin{aligned} Z^{(0)} &= \left\{z_j^{(0)}\right\}_{j=-\infty}^{\infty} := \mathrm{Enc}_P(X), \\ Z^{(m)} &= \left\{z_j^{(m)}\right\}_{j=-\infty}^{\infty} := g_m \circ f_{m-1} \circ g_{m-1} \circ \dots f_1 \circ g_1 \circ \mathrm{Enc}_P(X) \quad (j = 1, \dots, M). \end{aligned} \tag{12}$$

Then, the embedded token $z_j^{(0)}$ is given by $[X_{1,j}, \dots, X_{d,j}, 0, \dots, 0, \cos(j\phi), \sin(j\phi)]^\top$.

Next, we construct the attention layer g_1. Here, for each self-attention head, define the parameters $Q_h, K_h, V_h (h = 1, \dots, d_{\max})$ by

$$\begin{aligned} Q_h &:= \chi \begin{bmatrix} 0 & \dots & 0 & \cos(j_h\phi) & -\sin(j_h\phi) \\ 0 & \dots & 0 & \sin(j_h\phi) & \cos(j_h\phi) \end{bmatrix}, \\ K_h &:= \begin{bmatrix} 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix}, \\ V_h &:= \delta_{d+h, i_h}, \end{aligned}$$

where χ is a constant and $I(T, \gamma) = \{(i_h, j_h)\}_{h=1}^{d_{\max}}$. Then, the key, query, and value vectors for a given head h are given by

$$\begin{aligned} q_i &:= Q_h z_i^{(0)} = \chi[\cos((i + j_h)\phi), \sin((i + j_h)\phi)]^\top, \\ k_i &:= K_h z_i^{(0)} = [\cos(i\phi), \sin(i\phi)]^\top, \\ v_i &:= V_h z_i^{(0)} = X_{i_h, i} e_{d+h}. \end{aligned}$$

From the assumption $a_{ij} = \Omega(\log(|j| + 1))$, there exists a window size U such that $\log U \sim T$ and $j \leq U$ if $a_{ij} \leq T$. That is, $j \in [-U, U]$ if $(i, j) \in I(T, \gamma)$. Define $\tilde{Z} = [\dots, \tilde{z}_0, \dots, \tilde{z}_i, \dots]$ by

$$\tilde{z}_j := z_j^{(0)} + \sum_{h=1}^{H} V_h Z^{(0)}[j - U : j + U] e_{j_h} = z_j^{(0)} + \sum_{h=1}^{H} V_h z_{j_h}^{(0)}.$$

Intuitively, $\tilde{z}_j$ corresponds to $z_j^{(1)}$ in the situation where the softmax operation in the attention layer is replaced by *hardmax* operation. Then, we have

$$\tilde{z}_j = [X_{1,j}, \dots, X_{d,j}, X_{i_1,j+j_1}, \dots, X_{i_{d_{\max}},j+j_{d_{\max}}}, \cos(j\phi), \sin(j\phi)]^\top.$$

Note that $\tilde{z}_j$ contains important features $(X_{i_1,j+j_1}, \dots, X_{i_{d_{\max}},j+j_{d_{\max}}})$ for j-th output. For a little while, we focus on the 0-th token. Let $s^{(h)} := (K_h Z^{(0)}[-U:U])^\top (Q_h z_0^{(0)})$. Then, we have

$$s_i^{(h)} = k_i^\top q_0 = \chi(\cos(i\phi)\cos((j_h\phi)) + \sin(i\phi)\sin(j_h\phi)) = \chi\cos((j_h - i)\phi),$$

which implies $s_{j_h}^{(h)} \geq s_j^{(h)} + \chi/U^2 (\forall j \neq j_h)$. From Lemma C.1, we have

$$\begin{aligned}
\left\| z_0^{(1)} - \tilde{z}_0 \right\|_\infty &= \left\| \sum_{h=1}^{H} V_h Z^{(0)}[-U:U]\Big(e_{j_h} - \mathrm{Softmax}(s^{(h)})\Big) \right\|_\infty \\
&\leq \sum_{h=1}^{H} \left\| V_h Z^{(0)}[-U:U] \right\|_\infty \left\| \Big(e_{j_h} - \mathrm{Softmax}(s^{(h)})\Big) \right\|_1 \\
&\leq 2HUe^{-\chi/U^2},
\end{aligned}$$

where the last inequality holds because $\left\| V^{(h)} z_j^{(0)} \right\|_\infty = \left\| Z_{i_h,j}^{(0)} e_{d+h} \right\|_\infty \leq 1$ for any $j \in [-U:U]$. Similarly, we have $\left\| z_j^{(1)} - \tilde{z}_j \right\|_\infty \leq 2HUe^{-\chi/U^2}$ for any $j \in \mathbb{Z}$. Let $C \in \mathbb{R}^{d_{\max} \times D}$ be the matrix such that for any $x \in \mathbb{R}^D$, $Cx = [x_{d+1}, \dots, x_{d+H}]^\top$. Then, we have $C\tilde{z}_j = [X_{i_1,j+j_1}, \dots, X_{i_H,j+j_H}]^\top = \Gamma \circ \Sigma_j(X)$, where Γ is defined in Eq. (8). Next, we construct the FNN layer f_1. From Theorem D.3, there exists an FNN $\hat{f} \in \Psi(L, W, S, B')$ such that

$$\left\| \hat{f} \circ \Gamma - F_0^\circ \right\|_{2,P_X} \lesssim 2^{-T}.$$

where $\log B' \sim T^{1/\alpha}$. From the shift-equivariance of F°, we have

$$F_i^\circ(X) = F_0^\circ(\Sigma_i(X)), \tag{13}$$

Therefore, we have

$$\begin{aligned}
\left\| \hat{f} \circ \Gamma \circ \Sigma_i - F_i^\circ \right\|_{2,P_X} &= \left(\int_\Omega \left\| \hat{f} \circ \Gamma \circ \Sigma_i(X) - F_i^\circ(X) \right\|_2^2 \mathrm{d}P_X \right)^{1/2} \\
&= \left(\int_\Omega \left\| \hat{f} \circ \Gamma(\Sigma_i(X)) - F_0^\circ(\Sigma_i(X)) \right\|_2^2 \mathrm{d}P_X \right)^{1/2} \\
&= \left(\int_\Omega \left\| \hat{f} \circ \Gamma(X) - F_0^\circ(X) \right\|_2^2 \mathrm{d}P_X \right)^{1/2} \\
&\lesssim 2^{-T},
\end{aligned}$$

for any $i \in \mathbb{Z}$. For the third equality, we used the shift-invariance of P_X. Let $f_1 = \hat{f} \circ C$ and $\hat{F} = f_1 \circ g_1 \circ \mathrm{Enc}_P$. Since $f_1 \in \Psi(L, W, S, B')$ and f_1 is $(B'W)^L$-Lipschitz continuous with respect to $\|\cdot\|_\infty$ norm, we have, for $X \in [0,1]^{d \times \infty}$,

$$\begin{aligned}
\left| \hat{f} \circ \Gamma \circ \Sigma_i(X) - \hat{F}_i(X) \right| &= \left| f_1(\tilde{z}_i) - f_1(z_i^{(1)}) \right| \\
&\leq (B'W)^L \cdot 2HUe^{-\chi/U^2}.
\end{aligned}$$

By putting $\chi = U^2 \log\big(2HU(B'W)^L 2^T\big)$, we have

$$\left| \hat{f} \circ \Gamma \circ \Sigma_i(X) - \hat{F}_i(X) \right| \leq 2^{-T}.$$

Therefore, it holds that

$$\begin{aligned}\left\|F_i^\circ - \hat{F}_i\right\|_{2,P_X} &\le \left\|F_i^\circ - \hat{f}\circ\Gamma\circ\Sigma_i\right\|_{2,P_X} + \left\|\hat{f}\circ\Gamma\circ\Sigma_i - \hat{F}_i\right\|_\infty \\ &\lesssim 2^{-T}.\end{aligned}$$

The scaling factor χ can be evaluated as follows:

$$\begin{aligned}\log\chi &= \log\left[U^2\log\big(2HU(B'W)^L 2^T\big)\right] \\ &\sim T.\end{aligned}$$

Therefore, $g_1 \in \mathcal{A}(U,D,H,B)$ and $\hat{F} \in \mathcal{T}(M,U,D,H,L,W,S,B)$, which completes the proof. □

F. Proof of Theorem 4.5

For $T > 0$, define

$$\begin{aligned}I_j(T,\gamma) &:= \{i \mid (i,j) \in I(T,\gamma)\} = \left\{i_1^{(j)},\dots,i_{|I_j|}^{(j)}\right\}, \\ r_{\max}(T,\gamma) &:= \max\{j \mid |I_j(T,\gamma)| \neq 0\}.\end{aligned}$$

Note that $r_{\max} \sim T^{1/\alpha}$ since $a_{ij} = \Omega(j^\alpha)$. Let $\{u_i\}_{i=1}^{2V+1}$ be ε_1-approximately orthonormal vectors such that

$$\begin{aligned}\|u_i\|_2 &= 1, \\ \langle u_i, u_j\rangle &\overset{\varepsilon_1}{\approx} 1_{i=j},\end{aligned}$$

where $\varepsilon_1 > 0$ and $d' = O(\frac{\log V}{\varepsilon_1^2})$. In addition, we define $u_i = u_{i \mod 2V+1}$ for $i \notin [2V+1]$. Note that for any $i,j,k \in \mathbb{N}$ such that $i,j \in [k-V : k+V]$, $\langle u_i, u_j\rangle \overset{\varepsilon_1}{\approx} 1_{i=j}$ holds.

Let $M = r_{\max}+1$, $\phi = 2\pi/(2U_1+1)$, $D = d + d_{\max} + 2d' + 4$, $E = \sum_{i=1}^d \delta_{ii} \in \mathbb{R}^{D\times d}$, and

$$p_i = [0,\dots,0,1,\cos(i\phi),\sin(i\phi),u_i^\top]^\top \in \mathbb{R}^d.$$

Then, Enc_P is defined as

$$\mathrm{Enc}_P(X)_i := Ex_i + p_i = [x_i^\top, 0,\dots,0,1,\cos(i\phi),\sin(i\phi),u_i^\top]^\top.$$

Let $Z^{(0)} = \mathrm{Enc}_P(X)$, $Z^{(m)} = f_m \circ g_m(Z^{(m-1)})$ $(m = 1,\dots,M)$. By the same argument as in Theorem 4.2 with $T \sim \log 1/\varepsilon_1$, there exists an FNN f_1 and an attention layer g_1 such that

$$\begin{aligned}z_i^{(1)} &= x_i^{(1)} + y_i, \\ x_i^{(1)} &= \tilde{x}_i^{(1)} := [x_i^\top, 0,\dots,0,u_i^\top,0,\dots,0]^\top, \\ y_i &:= [0,\dots,0,\hat{\mu}_i(X),1,0,\dots,0]^\top, \\ \tilde{y}_i &:= [0,\dots,0,\mu_i(X),1,0,\dots,0]^\top, \\ \|\hat{\mu}-\mu\|_\infty &\lesssim \varepsilon_1.\end{aligned}$$

Note that Theorem 4.2 holds for $\|\cdot\|_{2,P_X}$ but it can be easily extended to $\|\cdot\|_\infty$ sicne $p = \infty$ is assumed. We also define $\tilde{z}_i^{(1)} = \tilde{x}_i^{(1)} + \tilde{y}_i$.

In the following, we fix $\lambda \in \Lambda$ and $X \in \Omega_\lambda$. For fixed λ and X, we denote π_λ by π and $\mu(X)_j$ by μ_j for simplicity. Let $r_i(m) = \pi_{\lambda_i}(m) + i$, where $\Sigma_i(X) \in \Omega_{\lambda_i}$. Since $\eta(x) - \eta(-x) = x$, there exists an FNN $f \in \Psi(2,2d,4d,1)$ such that $f(x) = x$ for any $x \in \mathbb{R}^d$. We set $f_m = f$ for $m = 2,\dots,M-1$. In addition, we set the parameters for i-th heads of g_m $(i = 2,\dots H,\ m = 2,\dots,M)$ by zero matrix. For the first head of g_m, we define the parameters U_m, K_m, Q_m, V_m by

$U_m = V$ and

$$K_{m+1} := \delta_{1,d+d_{\max}+d'+1} + \sum_{i=1}^{d'} \delta_{i+1,D-2d'+i},$$
$$Q_{m+1} := \chi\left(\delta_{1,d+d_{\max}+d'+2} - (2 + cr_{\max}^{-\beta}) \sum_{i=1}^{d'} \delta_{i+1,D-d'+i}\right),$$
$$V_{m+1} := \sum_{i=1}^{|I_m|} \delta_{i+\sum_{m'=1}^{m-1}|I_{(m')}|, i_i^{m')}} + \sum_{i=1}^{d'} \delta_{D-d'+i,D-2d'+1}.$$

For $m = 1, \dots, M-1$, let

$$\tilde{x}_i^{(m+1)} := \tilde{x}_i^{(m)} + V_{m+1}\tilde{X}^{(m)}[i - U_m : i + U_m]e_{r_i(m)},$$
$$\tilde{z}_i^{(m+1)} := \tilde{z}_i^{(m)} + V_{m+1}\tilde{Z}^{(m)}[i - U_m : i + U_m]e_{r_i(m)}.$$

Note that $\tilde{z}_i^{(m)} = \tilde{x}_i^{(m)} + \tilde{y}_i$ and $z_i^{(m)} = x_i^{(m)} + y_i$ since $V_{m+1}\tilde{y}_i = V_{m+1}y_i = 0$ for any m. Since

$$V_{m+1}\tilde{x}_i^{(m)} = [0, \dots, 0, X_{i_1^{(m)},i}, \dots, X_{i_{|I_m|}^{(m)},i}, 0, \dots, 0, u_i^\top]^\top,$$

we have

$$\tilde{x}_i^{(m+1)} = [x_i^\top, X_{i_1^{(1)},r_i(1)}, \dots, X_{i_{|I_m|}^{(m)},r_i(m)}, 0, \dots, 0, u_i^\top, w_i^{(m+1)}],$$

where $w_i^{(m+1)} := \sum_{m'=1}^{m} u_{r_i(m')}$. Let $C := \sum_{i=1}^{d_{\max}} \delta_{i,d+i} \in \mathbb{R}^{d_{\max}\times D}$. Then,

$$C\tilde{z}_k^{(M)} = [X_{i_1^{(1)},r_k(1)}, \dots, X_{i_{|I_{r_{\max}}|}^{(r_{\max})},r_k(r_{\max})}]^\top = \Gamma \circ \Pi \circ \Sigma_k, \tag{14}$$

since $[\Pi \circ \Sigma_k(X)]_{i,j} = [\Sigma_k(X)]_{i,\pi_{\lambda_k}(j)} = X_{i,\pi_{\lambda_k}(j)+k}$.

Next, we show that for any $\varepsilon_2 = O(3^{-r_{\max}}\varepsilon_1/D)$, $\left\|x_i^{(M)} - \tilde{x}_i^{(M)}\right\|_\infty \le 3^{r_{\max}}\varepsilon_2$ by setting χ appropriately. For $m = 1$, we have $\left\|x_i^{(m)} - \tilde{x}_i^{(m)}\right\|_\infty = 0$. Assume that $\left\|x_i^{(m)} - \tilde{x}_i^{(m)}\right\|_\infty = 3^{m-1}\varepsilon_2$ for some $m = 1, \dots, M-1$. The key and query vectors,

$$k_i := K_{m+1}z_i^{(m)},\ \tilde{k}_i := K_{m+1}\tilde{z}_i^{(m)} = [\mu_i, u_i]^\top,$$
$$q_i := Q_{m+1}z_i^{(m)},\ \tilde{q}_i := Q_{m+1}\tilde{z}_i^{(m)} = \chi[1, -(2 + cr_{\max}^{-\beta})w_i^{(m)}]^\top,$$

satisfy

$$\left\|k_i - \tilde{k}_i\right\|_2 \lesssim \left\|z_i^{(m)} - \tilde{z}_i^{(m)}\right\|_2 \le \|y_i - \tilde{y}_i\|_2 + D\left\|x_i^{(m)} - \tilde{x}_i^{(m)}\right\|_\infty \lesssim (\varepsilon_1 + 3^{(m-1)}D\varepsilon_2) \sim \varepsilon_1,$$
$$\|q_i - \tilde{q}_i\|_2 \lesssim \chi\left\|z_i^{(m)} - \tilde{z}_i^{(m)}\right\|_2 \lesssim \chi\varepsilon_1,$$
$$\|k_i\|_2 \lesssim 1,\ \|\tilde{q}_i\|_2 \lesssim \chi r_{\max}.$$

Therefore, for $m' \in [2V+1]$, we have

$$\begin{aligned} k_{r_i(m')}^\top q_i &\overset{\chi\varepsilon_1}{\sim} k_{r_i(m')}^\top \tilde{q}_i \\ &\overset{\chi r_{\max}\varepsilon_1}{\sim} \tilde{k}_{r_i(m')}^\top \tilde{q}_i \\ &= \chi(\mu_{r_i(m')} - (2 + cr_{\max}^{-\beta})u_{r_i(m')}^\top w_i^{(m)}) \\ &\overset{\chi r_{\max}\varepsilon_1}{\sim} \chi(\mu_{r_i(m')} - (2 + cr_{\max}^{-\beta})1_{m'\in[m-1]}), \end{aligned}$$

since $u_{r_i(m')}^\top w_i^{(m)} = \sum_{k=1}^{m-1} u_{r_i(m')}^\top u_{r_i(k)} \overset{r_{\max}\varepsilon_1}{\approx} 1_{m'\in[m-1]}$. For $m' < m$, we have

$$\begin{aligned} k_{r_i(m)}^\top q_i - k_{r_i(m')}^\top q_i &\overset{\chi r_{\max}\varepsilon_1}{\approx} \chi(\mu_{r_i(m')} - \mu_{r_i(m)} + 2 + cr_{\max}^{-\beta}), \\ &\geq \chi c r_{\max}^{-\beta}, \end{aligned}$$

since $|\mu_i| \leq 1$. For $m' > m$, we have

$$\begin{aligned} k_{r_i(m)}^\top q_i - k_{r_i(m')}^\top q_i &\overset{\chi r_{\max}\varepsilon_1}{\approx} \chi(\mu_{r_i(m')} - \mu_{r_i(m)}), \\ &\geq \chi c r_{\max}^{-\beta}. \end{aligned}$$

This is because for $m' > m \leq r_{\max}$,

$$[\mu \circ \Sigma_i(X)]_{\pi_{\lambda_i}(j_m)} - [\mu \circ \Sigma_i(X)]_{\pi_{\lambda_i}(j_{m'})} \geq cm^{-\beta} \geq cr_{\max}^{-\beta},$$

since μ is well-separated, and

$$\mu_{r_i(m')} = [\mu \circ \Sigma_i(X)]_{\pi_{\lambda_i}(j_{m'})},$$

which yields

$$\mu_{r_i(m)} - \mu_{r_i(m')} \geq cr_{\max}^{-\beta}.$$

Therefore, we have, for any $m' \neq m$,

$$k_{r_i(m)}^\top q_i - k_{r_i(m')}^\top q_i \geq \chi c r_{\max}^{-\beta}/2,$$

by letting $\varepsilon_1 \sim r_{\max}^{-\beta-1}$.

Let

$$s_i := \mathrm{Softmax}([k_{i-V}^\top q_i, \ldots, k_{i+V}^\top q_i]).$$

From Lemma C.1, we have

$$\begin{aligned} \left\| s_i - e_{r_i(m)} \right\|_1 &\leq (4V+2)e^{-\chi \frac{cr_{\max}^{-\beta}}{2}} \\ &\leq \varepsilon_2 / r_{\max}, \end{aligned}$$

by letting $\chi = \frac{2r_{\max}^\beta}{c} \log[(4V+2)r_{\max}/\varepsilon_2]$. Therefore, it holds that

$$\begin{aligned} \left\| x_i^{(m+1)} - \tilde{x}_i^{(m+1)} \right\|_\infty &\leq \left\| x_i^{(m)} - \tilde{x}_i^{(m)} \right\|_\infty + \left\| V_{m+1}\tilde{X}^{(m)}[i-V, i+V]e_{r_i(m)} - V_{m+1}X^{(m)}[i-V, i+V]s_i \right\|_\infty, \\ &\leq 3^{m-1}\varepsilon_2 + \left\| \tilde{X}^{(m)} \right\|_\infty \left\| e_{r_i(m)} - s_i \right\|_1 + \left\| X^{(m)} - \tilde{X}^{(m)} \right\|_\infty \|s_i\|_1, \\ &\leq 3^{m-1}\varepsilon_2 + \varepsilon_2 + 3^{m-1}\varepsilon_2 \leq 3^m \varepsilon_2, \end{aligned}$$

since $\|s_i\|_1 = 1$ and $\left\| \tilde{X}^{(m)} \right\|_\infty \leq r_{\max}$. By induction, we obtain $\left\| x_i^{(M)} - \tilde{x}_i^{(M)} \right\|_\infty \leq 3^{r_{\max}}\varepsilon_2$.

From Lemma D.3, there exists $\hat{f}_T \in \Psi(L, W, S, B')$ such that $\log B' \sim T^{1/\alpha}$ and

$$\left\| \hat{f}_T \circ \Gamma \circ \Pi - F_0^\circ \right\|_{2,P_X} \lesssim 2^{-T}.$$

By the same argument as in Theorem 4.2, we have

$$\begin{aligned} \left\| \hat{f}_T \circ \Gamma \circ \Pi \circ \Sigma_i - F_i^\circ \right\|_{2,P_X} &= \left\| \hat{f}_T \circ \Gamma \circ \Pi - F_0^\circ \right\|_{2,P_X} \\ &\lesssim 2^{-T} \end{aligned}$$

for any i. Let $f_M = \hat{f}_T \circ C \in \Psi(L, W, S, B')$. Then, Eq. (14) yields

$$f_M(\tilde{z}_i^{(M)}) = \hat{f}_T \circ \Gamma \circ \Pi \circ \Sigma_i(X).$$

Define $\hat{F} = f_M \circ g_M \circ \dots f_1 \circ g_1 \circ \mathrm{Enc}_P$. Since f_M is $(B'W)^L$-Lipschitz continuous, we have

$$\begin{aligned}\left|\hat{f}_T \circ \Gamma \circ \Pi \circ \Sigma_i(X) - \hat{F}_i(X)\right| &\le (B'W)^L \left\|z_i^{(M)} - \tilde{z}_i^{(M)}\right\|_\infty \\ &\le (B'W)^L 3^{r_{\max}} \varepsilon_2,\end{aligned}$$

for any $X \in \Omega$. Therefore, by letting $\varepsilon_2 = 3^{-r_{\max}}(B'W)^{-L}2^{-T}$, we have

$$\begin{aligned}\left\|F_i^\circ - \hat{F}_i\right\|_{2,P_X} &\le \left\|F_i^\circ - \hat{f}_T \circ \Gamma \circ \Pi \circ \Sigma_i\right\|_{2,P_X} + \left\|\hat{f}_T \circ \Gamma \circ \Pi \circ \Sigma_i - \hat{F}_i\right\|_\infty \\ &\lesssim 2^{-T}.\end{aligned}$$

Here, we have

$$\begin{aligned}\log \chi &= \log\left(\frac{2r_{\max}^\beta}{c}\log\big((4V+2)r_{\max}3^{r_{\max}}(B'W)^L 2^T\big)\right) \sim \log T + \log\log V, \\ D &= d + d_{\max} + 2d' + 2 \sim T^{2(\beta+1)/\alpha}\log V, \\ H &\sim (\log 1/\varepsilon_1)^{1/\alpha} \sim (\log T)^{1/\alpha}, \\ \log U_1 &\sim \log(1/\varepsilon_1) \sim \log T.\end{aligned}$$

Thus, $g_i \in \mathcal{A}(U_i, D, H, B)$, $f_i \in \Psi(L, W, S, B)$, and $\hat{F} \in \mathcal{T}(M, U, D, H, L, W, S, B)$, which completes the proof.

G. Proof of Theorem 5.2

To simplify the notation, let $F^\circ(X) \leftarrow F^\circ(X)[l:r]$, $\xi^{(i)} \leftarrow \xi^{(i)}[l:r]$, $Y^{(i)} \leftarrow Y^{(i)}[l:r]$, $l \leftarrow r - l + 1$, and $N \leftarrow \mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty)$. Define

$$\begin{aligned}\hat{R} &:= \mathbb{E}\left[\frac{1}{nl}\sum_{i=1}^n \left\|\hat{F}(X^{(i)}) - F^\circ(X^{(i)})\right\|_2^2\right], \\ \mathcal{D} &:= \left|\hat{R} - R(\hat{F}, F^\circ)\right|.\end{aligned}$$

Then, we have

$$R(\hat{F}, F^\circ) \le \hat{R} + \mathcal{D}.$$

First, we evaluate $\mathcal{D}$. Let G_δ be a minimal δ-covering of $\mathcal{F}$ in L^∞ norm such that $|G_\delta| = N$. Then, there exists a random variable $J \in [N]$ such that $\left\|\hat{F} - F_J\right\|_\infty \le \delta$. Define

$$g_j(X, X') = \frac{1}{l}\Big\{\|F_j(X) - F^\circ(X)\|_2^2 - \|F_j(X') - F^\circ(X')\|_2^2\Big\},$$

and we have

$$\begin{aligned}\left|\left\|\hat{F}(X) - F^\circ(X)\right\|_2^2 - \|F_J(X) - F^\circ(X)\|_2^2\right| &= \left\langle \hat{F}(X) - F_J(X), \hat{F}(X) + F_J(X) - 2F^\circ(X)\right\rangle \\ &\le \left\|\hat{F}(X) - F_J(X)\right\|_2 \left\|\hat{F}(X) + F_J(X) - 2F^\circ(X)\right\|_2 \\ &\le 4lR\delta. \qquad (15)\end{aligned}$$

For the last inequality, we use $\left\|\hat{F}(X) - F_J(X)\right\|_2 \le \sqrt{l}\left\|\hat{F}(X) - F_J(X)\right\|_\infty \le \sqrt{l}\delta$ and $\|F^\circ\|_\infty \le R, \|F\|_\infty \le R$ for any $F \in \mathcal{F}$.

Let $\tilde{X}^{(1)}, \dots, \tilde{X}^{(n)}$ be i.i.d. random variables independent of $(X^{(i)}, Y^{(i)})$. Then,

$$R(\hat{F}, F^\circ) = \frac{1}{nl} \sum_{i=1}^{n} \mathbb{E}\left[\left\| \hat{F}(\tilde{X}^{(i)}) - F^\circ(\tilde{X}^{(i)}) \right\|_2^2 \right]$$

holds and we have

$$\begin{aligned}
\mathcal{D} &= \left| \frac{1}{nl} \sum_{i=1}^{n} \left(\mathbb{E}\left[\left\| \hat{F}(X^{(i)}) - F^\circ(X^{(i)}) \right\|_2^2 \right] - \mathbb{E}\left[\left\| \hat{F}(\tilde{X}^{(i)}) - F^\circ(\tilde{X}^{(i)}) \right\|_2^2 \right] \right) \right| \\
&\leq \frac{1}{nl} \mathbb{E}\left[\left| \sum_{i=1}^{n} \left\| \hat{F}(X^{(i)}) - F^\circ(X^{(i)}) \right\|_2^2 - \left\| \hat{F}(\tilde{X}^{(i)}) - F^\circ(\tilde{X}^{(i)}) \right\|_2^2 \right| \right] \\
&\leq \frac{1}{n} \mathbb{E}\left[\left| \sum_{i=1}^{n} g_J(X^{(i)}, \tilde{X}^{(i)}) \right| \right] + 8R\delta.
\end{aligned}$$

Here, we used Eq. (15). Let $r_i = \max\left\{ A, l^{-1/2} \| F_j - F^\circ \|_2 \right\}$ and

$$T := \max_j \sum_{i=1}^{n} \frac{g_j(X^{(i)}, \tilde{X}^{(i)})}{r_j}$$

for $A > 0$, which is determined later. Then, we have

$$\begin{aligned}
\mathcal{D} &\leq \frac{1}{n} \mathbb{E}[r_J T] + 8R\delta, \\
&\leq \frac{1}{n} \sqrt{\mathbb{E}[r_J^2] \mathbb{E}[T^2]} + 8R\delta, \\
&\leq \frac{1}{2} \mathbb{E}[r_J^2] + \frac{1}{2n^2} \mathbb{E}[T^2] + 8R\delta. \qquad (16)
\end{aligned}$$

Here we use Cauchy-Schwarz inequality and the AM-GM inequality. From the definition of r_J and Eq. (15), it holds that

$$\begin{aligned}
\mathbb{E}[r_J^2] &\leq A^2 + \frac{1}{l} \mathbb{E}\left[\| F_J - F^\circ \|_2^2 \right] \\
&\leq A^2 + \frac{1}{l} \mathbb{E}\left[\left\| \hat{F} - F^\circ \right\|_2^2 \right] + 4R\delta. \qquad (17)
\end{aligned}$$

Since $X^{(1)}, \dots, X^{(n)}, \tilde{X}^{(1)}, \dots, \tilde{X}^{(n)}$ are independent of each other, we have

$$\begin{aligned}
\mathbb{V}\left[\sum_{i=1}^{n} \frac{g_j(X^{(i)}, \tilde{X}^{(i)})}{r_j} \right] &= \sum_{i=1}^{n} \mathbb{V}\left[\frac{\| F_j(X^{(i)}) - F^\circ(X^{(i)}) \|_2^2 - \left\| F_j(\tilde{X}^{(i)}) - F^\circ(\tilde{X}^{(i)}) \right\|_2^2}{r_j} \right] \\
&\leq \frac{2}{l^2 r_j^2} \sum_{i=1}^{n} \mathbb{E}\left[\left\| F_j(X^{(i)}) - F^\circ(X^{(i)}) \right\|_2^4 \right] \\
&\leq \frac{8R^2}{l r_j^2} \sum_{i=1}^{n} \mathbb{E}\left[\left\| F_j(X^{(i)}) - F^\circ(X^{(i)}) \right\|_2^2 \right] \\
&\leq 8nR^2, \\
\left| \frac{g_j(X^{(i)}, \tilde{X}^{(i)})}{r_j} \right| &\leq \left| \frac{g_j(X^{(i)}, \tilde{X}^{(i)})}{r_j} \right| \\
&\leq \frac{4R^2}{r_j},
\end{aligned}$$

where $\mathbb{V}[\cdot]$ denotes the variance of a random variable. Using Bernstein's ineqautlity and the union bound, we have, for any $t > 0$,

$$\begin{aligned}\Pr(T^2 \geq t) &= \Pr(T \geq \sqrt{t}) \\ &\leq 2N \exp\left\{-\frac{t}{2R^2(8n + \frac{4\sqrt{t}}{3r})}\right\} \\ &\leq 2N \exp\left\{-\frac{t}{32nR^2}\right\} + 2N \exp\left\{-\frac{3r\sqrt{t}}{16R^2}\right\},\end{aligned}$$

where $r = \min_{j\in[N]}\{r_j\}$. Then, for any $t_0 > 0$, we have

$$\begin{aligned}\mathbb{E}[T^2] &= \int_0^\infty \Pr(T^2 \geq t)\,\mathrm{d}t \\ &\leq t_0 + \int_{t_0}^\infty \Pr(T^2 \geq t)\,\mathrm{d}t \\ &\leq t_0 + 2N\int_{t_0}^\infty \exp\left[-\frac{t}{32nR^2}\right]\mathrm{d}t + 2N\int_{t_0}^\infty \exp\left[\frac{3r\sqrt{t}}{4R^2}\right]\mathrm{d}t\,. \end{aligned} \tag{18}$$

The integrals in Eq. (18) can be evaluated as follows:

$$\begin{aligned}\int_{t_0}^\infty \exp\left(-\frac{t}{32nR^2}\right)\mathrm{d}t &= \left[-32R^2\exp\left(-\frac{t}{32nR^2}\right)\right]_{t_0}^\infty \\ &= 32nR^2\exp\left(-\frac{t_0}{32nR^2}\right), \\ \int_{t_0}^\infty \exp\left(-\frac{3r}{16R^2}\right)\mathrm{d}t &= \left[-\frac{2(a\sqrt{t}+1)}{a^2}\exp\left(-a\sqrt{t}\right)\right]_{t_0}^\infty \quad (a = \frac{3r}{16R^2}) \\ &= \left(\frac{512R^4}{9r^2} + \frac{32R^2\sqrt{t_0}}{3r}\right)\exp\left(-\frac{3r\sqrt{t_0}}{16R^2}\right).\end{aligned}$$

Let $A = \frac{\sqrt{t_0}}{6n}$. Then, we have $r \geq A = \frac{\sqrt{t_0}}{6n}$ and

$$\mathbb{E}[T^2] \leq t_0 + 2N\left(32nR^2 + 64nR^2 + \frac{2048n^2R^4}{t_0}\right)\exp\left(-\frac{t_0}{32nR^2}\right).$$

Here, we determine $t_0 = 32nR^2\log N$. Then, we have

$$\mathbb{E}[T^2] \leq 32nR^2\left(\log N + 6 + \frac{4}{\log N}\right). \tag{19}$$

Combining (16), (17), (19), $A^2 = \frac{8R^2\log N}{9n}$, and $\log N \geq 1$, $\mathcal{D}$ can be evaluated as follows:

$$\begin{aligned}\mathcal{D} &\leq \frac{1}{2}\mathbb{E}[r_J^2] + \frac{1}{2n^2}\mathbb{E}[T^2] + 8R\delta \\ &\leq \frac{1}{2}A^2 + \frac{1}{2}\mathbb{E}\left[\frac{1}{l}\left\|\hat{F} - F^\circ\right\|_2^2\right] + \frac{1}{2n^2}\mathbb{E}[T^2] + 10R\delta \\ &\leq \frac{1}{2}A^2 + \frac{1}{2}\mathbb{E}\left[\frac{1}{l}\left\|\hat{F} - F^\circ\right\|_2^2\right] + \frac{16R^2}{n}\left(\log N + 6 + \frac{4}{\log N}\right) + 10R\delta \\ &\leq \frac{1}{2}R(\hat{F}, F^\circ) + \frac{4R^2}{n}\left(\frac{37}{9}\log N + 40\right) + 10R\delta.\end{aligned} \tag{20}$$

Next, we evaluate $\hat{R}$. Since $\hat{F}$ is an empirical risk minimizer, it holds that

$$\mathbb{E}\left[\frac{1}{nl}\sum_{i=1}^n \left\|\hat{F}(X^{(i)}) - Y^{(i)}\right\|_2^2\right] \leq \mathbb{E}\left[\frac{1}{nl}\sum_{i=1}^n \left\|F(X^{(i)}) - Y^{(i)}\right\|_2^2\right],$$

for any $F \in \mathcal{F}$. Substituting $Y^{(i)} = F^\circ(X^{(i)}) + \xi^{(i)}$, we have

$$\begin{aligned}
0 &\le \mathbb{E}\left[\frac{1}{nl}\sum_{i=1}^n \left\|F(X^{(i)}) - Y^{(i)}\right\|_2^2\right] - \mathbb{E}\left[\frac{1}{nl}\sum_{i=1}^n \left\|\hat{F}(X^{(i)}) - Y^{(i)}\right\|_2^2\right] \\
&= \mathbb{E}\left[\frac{1}{l}\left\|F(X^{(i)}) - F^\circ(X^{(i)})\right\|_2^2\right] - \mathbb{E}\left[\frac{1}{l}\left\|\hat{F}(X^{(i)}) - F^\circ(X^{(i)})\right\|_2^2\right] + \frac{2}{nl}\sum_{i=1}^n \mathbb{E}\left[\left\langle \xi^{(i)}, \hat{F}(X^{(i)})\right\rangle\right] \\
&= \frac{1}{l}\|F - F^\circ\|_2^2 + \frac{2}{nl}\sum_{i=1}^n \mathbb{E}\left[\left\langle \xi^{(i)}, \hat{F}(X^{(i)})\right\rangle\right] - \hat{R}.
\end{aligned}$$

Therefore, we have

$$\hat{R} \le \frac{1}{l}\|F - F^\circ\|_2^2 + \frac{2}{nl}\sum_{i=1}^n \mathbb{E}\left[\left\langle \xi(i), \hat{F}(X^{(i)})\right\rangle\right].$$

For the second term, we have

$$\begin{aligned}
\frac{2}{nl}\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, \hat{F}(X^{(i)})\right\rangle\right] &= \frac{2}{nl}\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, \hat{F}(X^{(i)}) - F^\circ(X^{(i)})\right\rangle\right] \\
&= \frac{2}{nl}\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, \hat{F}(X^{(i)}) - F_J(X^{(i)})\right\rangle\right] + \frac{2}{nl}\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, F_J(X^{(i)}) - F^\circ(X^{(i)})\right\rangle\right].
\end{aligned}$$

By the Cauchy-Schwartz ineqaulity, we have

$$\begin{aligned}
\frac{2}{nl}\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, \hat{F}(X^{(i)}) - F_J(X^{(i)})\right\rangle\right] &\le \frac{2}{nl}\mathbb{E}\left[\left(\sum_{i=1}^n \left\|\xi^{(i)}\right\|_2^2\right)^{1/2}\left(\sum_{i=1}^n \left\|\hat{F}(X^{(i)}) - F_J(X^{(i)})\right\|_2^2\right)^{1/2}\right] \\
&\le \frac{2\delta}{(nl)^{1/2}}\mathbb{E}\left[\left(\sum_{i=1}^n \left\|\xi^{(i)}\right\|_2^2\right)^{1/2}\right] \\
&\le \frac{2\delta}{(nl)^{1/2}}\mathbb{E}\left[\sum_{i=1}^n \left\|\xi^{(i)}\right\|_2^2\right]^{1/2} \\
&= 2\delta\sigma.
\end{aligned}$$

Define random variables $\varepsilon_1, \ldots, \varepsilon_N$ as

$$\varepsilon_j := \frac{\sum_{i=1}^n \left\langle \xi^{(i)}, F_j(X^{(i)}) - F^\circ(X^{(i)})\right\rangle}{\left(\sum_{i=1}^n \left\|F_j(X^{(i)} - F^\circ(X^{(i)}))\right\|_2^2\right)^{1/2}}.$$

If the denominator is zero, we define $\varepsilon_j = 0$. Then, we have

$$\begin{aligned}
\frac{2}{nl}\left|\mathbb{E}\left[\sum_{i=1}^n \left\langle \xi^{(i)}, F_J(X^{(i)}) - F^\circ(X^{(i)})\right\rangle\right]\right| &= \frac{2}{nl}\left|\mathbb{E}\left[\left(\sum_{i=1}^n \|F_J(X_i) - F^\circ(X_i)\|_2^2\right)^{1/2}\varepsilon_J\right]\right| \\
&\le \frac{2}{\sqrt{nl}}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|F_J(X_i) - F^\circ(X_i)\|_2^2\right]^{1/2}\mathbb{E}\left[\varepsilon_J^2\right]^{1/2} \\
&\le \frac{2}{\sqrt{n}}\sqrt{\hat{R} + 4R\delta}\,\mathbb{E}\left[\max_j \varepsilon_j^2\right]^{1/2} \\
&\le \frac{1}{2}(\hat{R} + 4R\delta) + \frac{2}{n}\mathbb{E}\left[\max_j \varepsilon_j^2\right].
\end{aligned}$$

Since each ε_j follows $N(0,\sigma^2)$ for given X^i, by the same argument as Theorem 7.47 in Lafferty et al. (2008), we have

$$\mathbb{E}[\max_j \varepsilon_j^2] \le 4\sigma^2 \log\left(\sqrt{2}N\right) \le 4\sigma^2(\log N + 1).$$

Therefore, it holds that

$$\hat{R} \le \frac{1}{l}\|F - F^\circ\|_2^2 + 2\delta\sigma + \frac{1}{2}(4F\delta + \hat{R}) + \frac{8}{n}\sigma^2(\log N + 1)$$

and then,

$$\hat{R} \le 2\frac{1}{l}\|F - F^\circ\|_2^2 + 4(R+\sigma)\delta + \frac{16}{n}\sigma^2(\log N + 1) \tag{21}$$

holds.

Combining Eq. (20) and (21), we have

$$\begin{aligned} R(\hat{F}, F) &\le \hat{R} + \mathcal{D} \\ &\le \frac{2}{l}\|F - F^\circ\|_2^2 + 4(R+\sigma)\delta + \frac{16}{n}\sigma^2(\log N + 1) + \frac{1}{2}R(\hat{F}, F) + \frac{4R^2}{n}\left(\frac{37}{9}\log N + 40\right) + 10R\delta, \end{aligned}$$

and thus,

$$R(\hat{F}, F) \le \frac{4}{l}\|F - F^\circ\|_2^2 + 8(R+\sigma)\delta + \frac{32}{n}\sigma^2(\log N + 1) + \frac{8R^2}{n}\left(\frac{37}{9}\log N + 40\right) + 20R\delta.$$

Since F is arbitrary, it holds that

$$R(\hat{F}, F) \le 4\inf_{F\in\mathcal{F}} \frac{1}{l}\|F - F^\circ\|_2^2 + 8(R+\sigma)\delta + \frac{32}{n}\sigma^2(\log N + 1) + \frac{8R^2}{n}\left(\frac{37}{9}\log N + 40\right) + 20R\delta,$$

which completes the proof. □

H. Proof of Theorem 5.3

For a Transformer $F \in \mathcal{T}(M,U,D,H,L,W,B,S,P)$, let θ_F be a vector of all the parameters of F. Suppose that $F, \tilde{F} \in \mathcal{T}(M,U,D,H,L,W,B,S,P)$ satisfies $\|\theta_F - \theta_{\tilde{F}}\|_\infty \le \delta$ for $\delta > 0$. That is, for any parameter θ in F, the corresponding parameter $\tilde{\theta}$ in $\tilde{F}$ satisfies $\left|\theta - \tilde{\theta}\right| \le \delta$. Transformer networks F and $\tilde{F}$ can be expressed in the form:

$$\begin{aligned} F(X) &= h_{2M} \circ \cdots \circ h_1 \circ (EX + P), \\ \tilde{F}(X) &= \tilde{h}_{2M} \circ \cdots \circ \tilde{h}_1 \circ (\tilde{E}X + P), \end{aligned}$$

where $h_i, \tilde{h}_i \in \Psi(L,W,B,S)$ if i is even, and $h_i, \tilde{h}_i \in \mathcal{A}(U_{(i+1)/2}, D, H, B)$ if i is odd. For fixed $X \in [0,1]^{d\times\infty}$, it holds that

$$\begin{aligned} \left\|F(X) - \tilde{F}(X)\right\|_\infty &\le \left\|h_{2M} \circ \dots h_1(EX + P) - h_{2M} \circ \dots h_1(\tilde{E}X + P)\right\|_\infty \\ &+ \sum_{m=1}^{2M} \left\|h_{2M} \circ \cdots \circ h_m \circ \tilde{h}_{m-1} \circ \cdots \circ \tilde{h}_1(\tilde{E}X + P) - h_{2M} \circ \cdots \circ h_{m+1} \circ \tilde{h}_m \circ \cdots \circ \tilde{h}_1(\tilde{E}X + P)\right\|_\infty. \end{aligned} \tag{22}$$

Since $\|P\|_\infty$ is assumed to be less than B, we have $\|EX + P\|_\infty, \left\|\tilde{E}X + P\right\|_\infty \le 2BD$. By applying Lemma C.5 repeatedly, we have

$$\begin{aligned} \left\|h_m \circ \tilde{h}_{m-1} \circ \cdots \circ \tilde{h}_1 \circ \mathrm{Enc}_P(X)\right\|_\infty &\le (6HDBW)^{2LM} \cdot 2BD \le (6HDBW)^{3LM}, \\ \left\|\tilde{h}_m \circ \cdots \circ \tilde{h}_1 \circ \mathrm{Enc}_P(X)\right\|_\infty &\le (6HDBW)^{2LM} \cdot 2BD \le (6HDBW)^{3LM}. \end{aligned}$$

In addition, Lemma C.4 yields that

$$\|h_m(X) - h_m(X')\|_\infty \le (6HDMW)^{4L+6LM}\|X - X'\|_\infty \le (6HDMW)^{10LM}\|X - X'\|_\infty \tag{23}$$

for $\|X\|_\infty, \|X'\|_\infty \le (6HDMW)^{3LM}$. Therefore, for the first term in Eq. (22), we have

$$\begin{aligned}\left\|h_{2M} \circ \dots h_1(EX + P) - h_{2M} \circ \dots h_1(\tilde{E}X + P)\right\|_\infty &\le (6HDBW)^{20LM^2}\left\|EX + P - (\tilde{E}X + P)\right\|_\infty \\ &\le (6HDBW)^{20LM^2} D\delta \le (6HDBW)^{21LM^2}\delta.\end{aligned}$$

For any $1 \le m \le 2M$, we have

$$\left\|h_{2M} \circ \dots \circ h_m \circ \dots \tilde{h}_1 \circ \tilde{\mathrm{Enc}}_P - h_{2M} \circ \dots \circ \tilde{h}_m \circ \dots \tilde{h}_1 \circ \tilde{\mathrm{Enc}}_P\right\|_\infty \le (6HDMW)^{20M^2L}\left\|h_m(Z) - \tilde{h}_m(Z)\right\|_\infty,$$

where $Z = \tilde{h}_{m-1} \circ \tilde{h}_1 \circ \tilde{\mathrm{Enc}}_P(X)$. Since $\|Z\| \le (6HDBW)^{3LM}$, Lemma C.6 implies that

$$\left\|h_m(Z) - \tilde{h}_m(Z)\right\| \le (6HDBW)^{4L+9LM}\delta \le (6HDBW)^{13LM}\delta.$$

Thus, we have

$$\begin{aligned}\left\|h_{2M} \circ \dots \circ h_m \circ \dots \tilde{h}_1 \circ \tilde{\mathrm{Enc}}_P - h_{2M} \circ \dots \circ \tilde{h}_m \circ \dots \tilde{h}_1 \circ \tilde{\mathrm{Enc}}_P\right\|_\infty &\le (6HDMW)^{20M^2L}(6HDBW)^{13LM}\delta \\ &\le (6HDMW)^{33M^2L}\delta.\end{aligned}$$

Then, we have

$$\begin{aligned}\left\|F(X) - \hat{F}(X)\right\|_\infty &\le (6HDBW)^{21LM^2}\delta + 2M(6HDMW)^{33M^2L}\delta \\ &\le (6HDMW)^{34M^2L}\delta\end{aligned}$$

Here, the number of non-zero components in θ_F is bounded by $M(S+3HD^2)+D^2$, where MS for FNN layers, $3MHD^2$ for attention layers, and $Dd \le D^2$ for an embedding layer. Therefore, if we fix the sparsity pattern, the covering number is bounded by

$$\left(\frac{(6HDMW)^{38M^2L}}{\delta}\right)^{M(S+3HD^2)+D^2}.$$

Since the total number of parameters is bounded by $M(L(W^2+W)+3HD^2)+D^2 \le 4M(LW^2+HD^2)$, the number of configurations of the sparsity pattern is bounded by

$$\binom{4M(LW^2+HD^2)}{M(S+3HD^2)+D^2} \le \left(4M(LW^2+HD^2)\right)^{M(S+3HD^2)+D^2}.$$

Therefore, the covering number of Transformer networks is bounded by

$$\left(4M(LW^2+HD^2)\right)^{M(S+3HD^2)+D^2}\left(\frac{(6HDMW)^{34M^2L}}{\delta}\right)^{M(S+3HD^2)+D^2} \le \left(\frac{(6HDMWL)^{36M^2L}}{\delta}\right)^{M(S+3HD^2)+D^2},$$

which completes the proof. □

I. Proof of Theorem 5.4

From Theorem 4.2, there exists a Transformer $\mathcal{T}_R(M, U, D, H, L, W, S, B)$ such that $\left\|\hat{F}_i - F_i^\circ\right\|_{2,P_X} \lesssim 2^{-T}$ for any $i \in \mathbb{Z}$, where

$$\begin{aligned}
M &= 1, \\
\log U_1 &\sim T, \\
D &\sim T^{1/\alpha}, \\
H &\sim T^{1/\alpha}, \\
L &\sim \max\left\{T^{2/\alpha}, T^2\right\}, \\
W &\sim T^{1/\alpha} 2^{T/a^\dagger}, \\
S &\sim T^{2/\alpha} \max\left\{T^{2/\alpha}, T^2\right\} 2^{T/a^\dagger}, \\
\log B &\sim \max\left\{T^{1/\alpha}, T\right\},
\end{aligned}$$

because $\|F\|_\infty \leq R$. Therefore, the bias of the estimator $\hat{F} \in \mathcal{T}_R$ can be evaluated as follows:

$$\inf_{F' \in \mathcal{T}_R} \frac{1}{r-l+1} \sum_{i=l}^{r} \|F_i' - F_i^\circ\|_{2,P_X}^2 \leq 2^{-2T}.$$

From Lemma 5.3, the log covering number $\log \mathcal{N}(\mathcal{T}_R, \delta, \|\cdot\|_\infty)$ is evaluated as follows:

$$\log \mathcal{N}(\mathcal{T}_R, \delta, \|\cdot\|_\infty) \leq \log \mathcal{N}(\mathcal{T}, \delta, \|\cdot\|_\infty) \lesssim 2^{T/a^\dagger} T^{2/\alpha+1} \max\left\{T^{4/\alpha}, T^4\right\} \log \frac{T}{\delta}.$$

Therefore, from Lemma 5.4, the ERM estimator $\hat{F}$ satisfies

$$R_{l,r}(\hat{F}, F^\circ) \lesssim 2^{-2T} + \frac{2^{T/a^\dagger} T^{2/\alpha+1} \max\left\{T^{4/\alpha}, T^4\right\} \log(T/\delta)}{n} + \delta.$$

By letting $T = \frac{a^\dagger}{2a^\dagger+1} \log n$ and $\delta = 1/n$, we have

$$R_{l,r}(\hat{F}, F^\circ) \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}} (\log n)^{2/\alpha+2} \max\left\{(\log n)^{4/\alpha}, (\log n)^4\right\}.$$

□

J. Proof of Theorem 5.5

Let

$$\begin{aligned}
M &= T^{1/\alpha}, \\
\log U_i &\lesssim \max\{\log T, \log V\}, \\
D &\sim T^{2(\beta+1)/\alpha} \log V, \\
H &\sim (\log T)^{1/\alpha}, \\
L &\sim \max\left\{T^{2/\alpha}, T^2\right\}, \\
W &\sim T^{1/\alpha} 2^{T/a^\dagger}, \\
S &\sim T^{2/\alpha} \max\left\{T^{2/\alpha}, T^2\right\} 2^{T/a^\dagger}, \\
\log B &\sim \max\left\{T^{1/\alpha}, T, \log\log V\right\}.
\end{aligned}$$

Then, by the same argument as in Theorem 5.4, it holds that

$$\log \mathcal{N}(\mathcal{T}_R, \delta, \|\cdot\|_\infty) \lesssim T^{5/\alpha+1} \max\left\{T^{4/\alpha}, T^4\right\} 2^{T/a^\dagger} (\log V)^2 \log\left(\frac{T \log V}{\delta}\right).$$

By letting $T = \frac{a^\dagger}{2a^\dagger+1} \log n$ and $\delta = 1/n$, we have

$$R_{l,r}(\hat{F}, F^\circ) \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}} (\log n)^{5/\alpha+2} \max\left\{(\log n)^{4/\alpha}, (\log n)^4\right\} (\log V)^3.$$

□