
Understanding the Expressive Power and Mechanisms of Transformer for Sequence Modeling

Mingze Wang

School of Mathematical Sciences, Peking University, Beijing, China
mingzewang@stu.pku.edu.cn

Weinan E †

Center for Machine Learning Research and School of Mathematical Sciences, Peking University, Beijing, China
AI for Science Institute, Beijing, China
weinan@math.pku.edu.cn

Abstract

We conduct a systematic study of the approximation properties of Transformer for sequence modeling with long, sparse and complicated memory. We investigate the mechanisms through which different components of Transformer, such as the dot-product self-attention, positional encoding and feed-forward layer, affect its expressive power, and we study their combined effects through establishing explicit approximation rates. Our study reveals the roles of critical parameters in the Transformer, such as the number of layers and the number of attention heads. These theoretical insights are validated experimentally and offer natural suggestions for alternative architectures.

1 Introduction

In recent years, Transformer networks (Vaswani et al., 2017) have emerged as foundational models, setting new benchmarks across various domains, including natural language processing (NLP), computer vision (CV), and protein folding. Despite their impressive practical achievements, the underlying mechanisms and theoretical foundations of Transformer networks remain largely elusive.

Transformer networks encompass various components, posing challenges to their comprehensive understanding. A typical Transformer comprises multiple layers, each consisting of a multi-head self-attention (Attn) sub-layer and a feed-forward network (FFN) sub-layer, integrated with residual blocks. FFN is a two-layer nonlinear network, while Attn includes dot-product (DP) and positional encoding (PE). To get a better understanding of how Transformer works in practice, we need to study several key issues. These include:

- (i) *How do the key hyper-parameters, for example, the number of layers, the number of Attn heads and the with of FFN layers, affect the performance of the Transformer network?*
- (ii) *How do the Attn and FFN layers contribute differently to the overall performance?*
- (iii) *How does DP attention work, and is the DP structure necessary?*
- (iv) *How efficient is PE in modeling long-range correlations?*

Extensive empirical research on Transformer components has led to the proposal of numerous alternatives to the current structure of Transformer. For example, several relative positional encodings (RPE) (Shaw et al., 2018; Raffel et al., 2020; Su et al., 2024; Press et al., 2022) have been proposed

to substitute the original absolute positional encoding (APE), yielding superior performance in challenging tasks like length generalization (Ontanón et al., 2022; Csordás et al., 2021; Anil et al., 2022). Additionally, the necessity of the computationally expensive DP in Attn layers has been widely questioned, and researchers proposed numerous alternatives of DP that show considerable efficacy in specific tasks (Kitaev et al., 2020; Wang et al., 2020; Choromanski et al., 2020; Tay et al., 2021; Allen-Zhu and Li, 2023). Nonetheless, these explorations have not yielded a satisfactory theoretical understanding of the mechanisms of these components.

In this work, we investigate the expressive power of Transformer and the underlying mechanisms of its components for sequence modeling. **Our contributions** are summarized as follows:

We categorize three types of sequence modeling tasks with varying complexity, which are relevant to a broad spectrum of application areas. *Task I: Modeling fixed, long but sparse memories.* This is relevant to sparse Boolean functions and the traditional n -gram model in NLP. *Task II: Modeling adaptive, long but sparse memories.* This is relevant to multi-step reasoning tasks as well as various NLP tasks such as dependency parsing, sentiment analysis, and continuation writing. *Task III: Modeling essentially sparse memories.* Examples include feature representation in CV and wavelet analysis in classical signal processing.

For these sequence modeling tasks, we theoretically investigate the expressive power of Transformer and its variants, establishing explicit approximation rates. Our meticulous analysis provides *theoretical insights* into the underlying mechanisms of Transformer components. Specifically,

- **The distinct roles of the number of layers, the number of Attn heads, and the width of FFN layers.** Deeper Transformer are capable of handling memories with more intricate interrelationships, such as nested relationships (Thm 4.4). In contrast, for memories lacking such interrelationships, single-layer Transformer with sufficient number of Attn heads and FFN width should suffice (Thm 4.1). This is quite intuitive: If the content of the next token relies on a few previous tokens in an independent way, we can treat each such dependence by a separate attention head. There is no need for many layers. Additionally, increasing the depth can also alleviate the reliance on the number of heads and width (Prop 4.5).
- **The different roles of Attn layers and FFN layers.** Our results consistently suggest that: FFN layers are tasked with approximating nonlinear memory functions and the readout function, while Attn layers are responsible for extracting the tokens from these memory locations.
- **The functionality and necessity of DP.** For the relatively simple Task I, DP is not necessary and can be omitted (Thm 3.1). However, for the more complex Task II, the cooperation between DP and RPE provides the needed interaction between the temporal space and the token space, crucial for the extraction of adaptive memories (Thm 4.1 and 4.4). Additionally, for Task II, while the nonlinearity provided by DP is necessary (Prop 4.2), a computationally efficient alternative to DP exists, as we show in Prop 4.3.
- **The efficiency of RPE in modeling long-range correlations.** Our results consistently suggest that the primary role of RPE is to approximate the memory kernels. Specifically, for Task III, we demonstrate that Transformer with suitable RPE can handle heavy-tailed memories, thus overcoming the Curse of Memory faced by recurrent neural networks (Thm 5.1). Moreover, our findings give theoretical support to the choice of RPE in practice.

Finally, we conduct experiments to validate our theoretical insights.

2 Preliminaries

Basic notations. We use bold-faced letters for vectors or matrices and lowercase letters for scalars, e.g. $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and $\mathbf{W} = (W_{ij})_{m \times n} \in \mathbb{R}^{m \times n}$. The standard Euclidean inner product between two vectors is denoted by $\langle \cdot, \cdot \rangle$, and the l_p norm of a vector is represented by $\|\cdot\|_p$.

We employ standard big-O notations \mathcal{O} , Ω , Θ to hide absolute positive constants and use $\tilde{\mathcal{O}}$, $\tilde{\Omega}$, $\tilde{\Theta}$ to further hide logarithmic constants. For any positive integer n , let $[n] = \{1, \dots, n\}$. Denote by $\mathbb{I}\{E\}$ the indicator function for an event E . Denote by $a \vee b = \max\{a, b\}$ for real number a, b .

2.1 Sequence modeling with long but sparse memories

Sequence modeling. For convenience, we consider input sequences of infinite length ($t \in \mathbb{Z}$). It is important to note, however, that our theoretical framework can be adapted to finite-length input sequences by masking distant tokens. Formally, the output sequence $\mathbf{Y} = (\mathbf{y}_t)_{t \in \mathbb{Z}} \in \mathbb{R}^{c \times \mathbb{Z}}$ is generated from the input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \in \mathcal{X} \subset \mathbb{R}^{d \times \mathbb{Z}}$ via an unknown mapping $\mathbf{H}(\cdot)$ dependent on the input sequence up to the prediction time, and this can be expressed as:

$$\mathbf{y}_t = \mathbf{H}_t(\mathbf{X}) = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \dots), \quad t \in \mathbb{Z}. \quad (1)$$

Our objective is to learn the mapping $\mathbf{H}(\cdot)$. Additionally, we define the norm $\|\mathbf{H}\| := \sup_{t \in \mathbb{Z}} \sup_{\mathbf{X} \in \mathcal{X}} \|\mathbf{H}_t(\mathbf{X})\|$. Without loss of generality, we assume $\|\mathbf{x}_t\|_2 \leq 1$ for any $\mathbf{X} \in \mathcal{X}$ and set the output dimension $c = 1$ for simplicity.

Long but sparse memories. To model such sequences, we define three types of memories: fixed, long but sparse memories; adaptive, long but sparse memories; and essentially sparse memories. These memory types are prevalent in sequence modeling tasks across diverse domains such as NLP, CV, signal processing, and sparse function representation. In Section 3, 4, and 5, we will formally define these different types and investigate Transformer’s capacity to model them.

2.2 Transformer architecture

Transformer network. Transformer (Vaswani et al., 2017) is a network architecture designed for processing sequences and generating predictions. Given an input sequence \mathbf{X} , Transformer executes the following steps. Initially, each d -dimensional (dim) input token is transformed into a D -dim vector through an embedding mapping such as $\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E$, where $\mathbf{W}_E \in \mathbb{R}^{D \times d}$, $\mathbf{b}_E \in \mathbb{R}^D$. Subsequently, a typical L -layer Transformer with residual block operates according to the formulation:

$$\begin{aligned} \mathbf{X}^{(l-\frac{1}{2})} &= \mathbf{X}^{(l-1)} + \text{Attn}^{(l)}(\mathbf{X}^{(l-1)}), \quad l \in [L]; \\ \mathbf{X}^{(l)} &= \mathbf{X}^{(l-\frac{1}{2})} + \text{FFN}^{(l)}(\mathbf{X}^{(l-\frac{1}{2})}), \quad l \in [L]. \end{aligned} \quad (2)$$

At the l -th layer, $\text{FFN}^{(l)}(\cdot)$ denotes a standard (point-wise) two-layer ReLU networks with m neurons: for a given input $\mathbf{x} \in \mathbb{R}^D$, $\text{FFN}^{(l)}(\mathbf{x}) = \sum_{k=1}^m \mathbf{a}_k^{(l)} \sigma(\mathbf{b}_k^{(l)\top} \mathbf{x} + c_k^{(l)})$, where $\sigma(\cdot)$ is the activation function such as ReLU. Additionally, in the final (L -th) FFN layer, the residual block is omitted, commonly referred to as the readout function. Moreover, $\text{Attn}^{(l)}(\cdot)$ refers to a multi-head self-attention, as elaborated below.

Multi-head self-attention. Our focus lies on standard dot-product Attn, denoted as $\text{Attn}^{(l)}(\cdot)$ and consisting of H heads. When applied to an input sequence \mathbf{X} , Attn operates as follows:

$$\text{Attn}^{(l)}(\mathbf{X}) = \mathbf{W}_O^{(l)} \sum_{h=1}^H \mathbf{W}_V^{(l,h)} \mathbf{X} \text{softmax}_c \left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{X}, \mathbf{W}_K^{(l,h)} \mathbf{X} \right\rangle + \mathbf{R}^{(l,h)} \right). \quad (3)$$

Here, the parameters $\mathbf{W}_Q^{(l,h)}$, $\mathbf{W}_K^{(l,h)}$, $\mathbf{W}_V^{(l,h)}$, $\mathbf{W}_O^{(l,h)}$ correspond to the query, key, value, output matrices of the (l, h) -th head, respectively. softmax_c represents taking softmax normalization across column. Furthermore, $\mathbf{R}^{(l,h)} \in \mathbb{R}^{\mathbb{Z} \times \mathbb{Z}}$ denotes the relative positional encoding matrix, which satisfies $R_{t,s}^{(l,h)} = -\infty$ for $t < s$ in the next-token prediction paradigm. Consequently, the t -th output of Attn is expressed as:

$$\text{Attn}_t^{(l)}(\mathbf{X}) = \mathbf{W}_O^{(l)} \sum_{h=1}^H \sum_{s=0}^{+\infty} \frac{\mathbf{W}_V^{(l,h)} \mathbf{x}_{t-s} \exp \left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-s} \right\rangle + R_{t,t-s}^{(l,h)} \right)}{\sum_{j=0}^{+\infty} \exp \left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-j} \right\rangle + R_{t,t-j}^{(l,h)} \right)}.$$

Logarithmic and Power relative positional encoding. As highlighted in Section A, among various types of RPEs, the RPEs used in T5 and KERPLE(log) demonstrate superior performance over Alibi, significantly outperforming other RPEs and APEs in the length generalization task (Kazemnejad et al., 2023; Chi et al., 2022). This finding motivates our focus on the T5-type, KERPLE(log), and Alibi-type RPEs throughout this paper. All of these RPE matrices are Toeplitz, with the form of

$R_{t,s} = r(t-s)$. Notably, for T5 and KERPLE(log), $r(t-s)$ undergoes an initial linear decrease followed by a logarithmic decrease as the relative distance $t-s$ increases (Please refer to Section G.1 for more details). In contrast, for Alibi, $r(t-s)$ decreases linearly. Inspired by these discussions, we examine the following RPEs with different decay rates:

$$\phi_{\log}(z) = \begin{cases} -\log z, & z \geq 1 \\ -\infty, & \text{otherwise} \end{cases}; \quad \phi_{\text{lin}}(z) = \begin{cases} -z, & z \geq 0 \\ -\infty, & \text{otherwise} \end{cases}.$$

We will study Transformer with ϕ_{type} RPE ($\text{type} \in \{\log, \text{lin}\}$). Specifically, the RPE in the (l, h) -th head (3) is as follows:

$$R_{t,s}^{(l,h)} := p^{(l,h)} \phi_{\text{type}}(t-s), \quad (4)$$

where $p^{(l,h)} \in \mathbb{R}_+$ is a trainable parameter.

Remark 2.1. For standard Transformer (2) incorporating Attn (3) with RPE (4), the parameters are: the embedding matrix \mathbf{W}_E ; $\mathbf{a}_k^{(l)}, \mathbf{b}_k^{(l)}, \mathbf{c}_k^{(l)}$ in the FFN layers; $\mathbf{W}_Q^{(l,h)}, \mathbf{W}_K^{(l,h)}, \mathbf{W}_V^{(l,h)}, p^{(l,h)}, \mathbf{W}_O^{(l)}$ in the Attn layers. Notably, the number of parameters is independent of the sequence length, thus enabling the model to handle input sequences of arbitrary length.

Remark 2.2. In the subsequent sections, we will analyze Transformer and its variants. For the sake of brevity, some shorthand notations are introduced here. For examples, Transformer (2) using $\phi_{\log}/\phi_{\text{lin}}$ RPE (4) is referred to as ‘‘Transformer with log/lin-RPE’’; Transformer with $\mathbf{W}_Q^{(l,h)}, \mathbf{W}_K^{(l,h)} = \mathbf{0}$ is called ‘‘dot-product-free Transformer’’.

2.3 Expressive power via approximation theory

This paper delves into the expressive power of Transformer through the lens of approximation theory, with a specific focus on establishing explicit approximation rates for Transformers in modeling long but sparse memories.

Approximation rates v.s. universal approximation. In approximation theory, results are generally categorized into two types: universal approximation (density-type) and approximation rates (Jackson-type) (Jackson, 1930). Universal approximation investigates whether the hypothesis class is dense in the target class. Although this property is fundamental, it does not offer detailed insights into approximation efficiency. In contrast, approximation rates go deeper, emphasizing the efficiency of the approximation. A typical example within this framework is the approximation theory of two-layer neural networks (2NNs).

Barron space of 2NNs. The well-known universal approximation result for 2NNs asserts that 2NNs can approximate any continuous function (Barron, 1992; 1993; 1994). Nonetheless, this result lacks a characterization of the approximation efficiency, i.e., how many neurons are needed to achieve a certain approximation accuracy? This gap was addressed by the Barron space theory (E et al., 2019; 2021; Ma et al., 2020). It is established that for any function within Barron space $f \in \mathcal{B}$ (Appendix G.2), 2NNs with m neurons (denoted by \mathcal{H}_m) can approximate them efficiently, at a rate of $\inf_{f_m \in \mathcal{H}_m} \|f - f_m\| \leq \mathcal{O}(\|f\|_{\mathcal{B}} / \sqrt{m})$, remarkably independent of the input dimension d , thus avoiding the *Curse of Dimensionality* (Bellman, 1966; Bach, 2017).

3 Fixed, long but M -sparse memories

3.1 Problem formulation

Fixed, long but M -sparse memories. In this section, we investigate a fundamental category of long but sparse memories. Our focus is on scenarios where the positions of the sparse memories remain fixed and are independent of the tokens. The target function is represented by:

$$y_t = f(\mathbf{x}_t, \mathbf{x}_{t-T_1}, \dots, \mathbf{x}_{t-T_M}), \quad (5)$$

where $1 \leq T_1 < \dots < T_M < +\infty$ signify the fixed positions of the memories. Despite the memories being fixed (token-independent) and sparse (finite M), the task can still be complex due to the potentially long-range memories (T_1, \dots, T_M can be large enough).

Examples. (I) For Boolean inputs, (5) aligns with *sparse Boolean functions*, also studied in (Edelman et al., 2022; Bhattamishra et al., 2022). Notably, Bhattamishra et al. (2022) observed that Transformers outperform LSTMs in learning sparse parities. (II) Selecting the simplest case of $T_i = i$ in (5) corresponds to the traditional *n-gram model*, which consists of short and sparse memories.

Target class. We focus on target functions described in (5). The readout function f is considered within the standard Barron space \mathcal{B} , i.e., which can be effectively approximated by 2NNs. Moreover, we assume that f is Lipschitz, denoted by $f \in \mathcal{L}$. Thus, we can focus more on investigating the memory extraction power of Transformer. Formally, we define the target class for modeling fixed, long but M -sparse memories as:

$$\mathcal{H}^{\text{Fix}} := \{\mathbf{H} : \mathbf{H}_t(\mathbf{X}) = (5), \text{ where } 1 \leq T_1 < \dots < T_M < +\infty, f \in \mathcal{B} \cap \mathcal{L}\}. \quad (6)$$

Transformer hypothesis class. As mentioned in Section 1, one of our main aims is to study the necessity and roles of different components in Transformer, such as DP and RPE. This section focuses on the “simplest” one-layer Transformer and investigates whether it can effectively model this task. Formally, our hypothesis class includes all one-layer *DP-free* Transformers, configured with H Attn heads and FFN width m :

$$\mathcal{TF}_{(1,H,m)}^{\text{DPF,type}} := \{\mathbf{TF} : \mathbf{TF} \text{ is a 1-layer, } H\text{-head, } m\text{-width dot-product-free Transformer with type-RPE}\}. \quad (7)$$

3.2 Theoretical results and insights

Theorem 3.1 (Approximation rate). *For any target $\mathbf{H} \in \mathcal{H}^{\text{Fix}}$ (6), rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,type}}$ (7) and a constant $C(n)$ such that*

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

$$\text{where } \mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right) \text{ and } \mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M e^{0.01T_i}\right)^{n+1}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M T_i^{1.01}\right)^{n+1}\right), & \text{type} = \text{log} \end{cases}.$$

Theorem 3.1 establishes the *approximation rate* of one-layer DP-free Transformer for modeling fixed, long but sparse memories. Here, the model complexity is governed by the number of Attn heads H and the width of FFN layers m , while the target complexity arises from the lengths of the memories T_1, \dots, T_M and the complexity of the readout function f . The approximation error comprises two components: the error in the FFN component \mathcal{E}_{FFN} and the error in the Attn component $\mathcal{E}_{\text{Attn}}(\text{type})$. The error \mathcal{E}_{FFN} aligns with classical results, showcasing its effectiveness in approximating Barron functions. On the other hand, $\mathcal{E}_{\text{Attn}}(\text{type})$ hinges on the capacity of the Attn block for modeling long-range memories. Specifically, with increasing memory length, the necessary number of Attn heads grows at a small exponential rate for lin-RPE and at a polynomial rate for log-RPE.

The proof of Theorem 3.1 is deferred to Appendix B. We can draw some insights from Theorem 3.1 and its proof.

Different roles of the Attn layer and the FFN layer. The Attn and FFN layers fulfill distinct roles in this task. Specifically, the FFN layer efficiently approximates the nonlinear readout function f , while the Attn layer is responsible for extracting the token \mathbf{x}_{t-T_i} by approximating the memory kernel $\mathbb{I}\{\cdot = T_i\}$. These components together enable effective modeling of fixed, long, but sparse memories.

Non-necessity of DP. Theorem 3.1 suggests that the DP component in Attn is not necessary and can be omitted for modeling fixed, long but sparse memories. This is due to the relative simplicity of modeling fixed memory kernels. In a more complex scenario in Section 4, the role of the dot-product becomes important. In contrast to Edelman et al. (2022), which utilizes the property of DP to prove that Transformer can model sparse Boolean functions, our result reveals that one-layer Transformer can successfully tackle the same task *even without the dot product in the attention layer*.

Effect of RPE types on expressivity. Our result indicates that the type of the RPE used in the Attn layer subtly influences the Transformer’s ability to model long-range memories. As the range of the memory increases, the required head number grows at a slightly exponential rate for lin-RPE

and at a polynomial rate for log-RPE. The subtle difference is attributed to the relative simplicity of approximating the memory kernel $\mathbb{I}\{\cdot = T_i\}$. We will explore a more complex task in Section 5, where the impact of different types of RPE becomes even more pronounced.

4 K -Adaptive, long but M -sparse memories

4.1 Problem formulation

In this section, we delve into a more complex modeling scenario closely aligned with typical language processing tasks.

K -Adaptive, long but M -sparse memories. This section investigates the scenario where the positions of the sparse memories are “adaptive”, meaning they depend on the input tokens. The target function is formulated as:

$$y_t = f(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}), \quad (8)$$

where the positions of the memory tokens t_1, \dots, t_M follow a nested relationship:

$$\begin{aligned} t_1 &= g_1(\mathbf{x}_t); t_2 = g_2(\mathbf{x}_t, \mathbf{x}_{t-t_1}); \dots; t_{K+1} = g_{K+1}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_K}); \\ \dots; t_M &= g_M(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_K}). \end{aligned}$$

Here, M denotes the number of memory tokens, and K measures the nesting complexity in the memory structure. We assume that memory functions g_i generate positive integers for the input tokens, and there exist maximum values T_i such that $g_i \leq T_i$. In this adaptive framework, each position of the memory token depends on multiple input tokens and is nested within other memory structures, leading to potential influence of later memory tokens by the earlier ones.

To facilitate understanding, we first consider a warm-up case, i.e., $K = 0$ in (8). In this case, the positions of memories only depend on the current token, without interaction with each other. It can be represented as:

$$y_t = f(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}), \quad (9)$$

where $t_i = g(x_i), i \in [M]$.

Target class. The target classes for modeling adaptive, long but sparse memories in both warm-up and general cases are as follows:

$$\mathcal{H}_{(1,M)}^{\text{Adap}} := \{\mathbf{H} : \mathbf{H}_t(\mathbf{X}) = (9), \text{ where } g_i \in \mathcal{B}, 1 \leq g_i \leq T_i, i \in [M]; f \in \mathcal{B} \cap \mathcal{L}\}. \quad (10)$$

$$\mathcal{H}_{(K,M)}^{\text{Adap}} := \{\mathbf{H} : \mathbf{H}_t(\mathbf{X}) = (8), \text{ where } g_i \in \mathcal{B}, 1 \leq g_i \leq T_i, i \in [M]; f \in \mathcal{B} \cap \mathcal{L}\}. \quad (11)$$

Examples. Adaptive memories are commonly encountered in practical scenarios. (I) *Adaptive sparse Boolean functions*, e.g., $y_t = x_t \cdot x_{t-g(x_t)} \cdot x_{t-g(x_t-g(x_t))}$, where $\mathbf{X} \in \{\pm 1\}^{\mathbb{Z}}$, $g(x) = 1$ for $x = 1$ and $g(x) = 2$ for $x = -1$. This fits within our framework (8) with $K = M = 2$. (II) *Multi-step reasoning*, e.g., modeling the K -adaptive, long, but K -sparse memories contains a complicated K -step reasoning task, which require the sequential search following the rule $((\dots((x_t \mapsto x_{t-t_1}) \mapsto x_{t-t_2} \dots)) \mapsto x_{t-t_{K-1}}) \mapsto x_{t-t_K}$. (III) In *NLP tasks* like dependency parsing, part-of-speech tagging, sentiment analysis, or continuation writing, the positions of relevant prefix tokens usually depend on the context itself, and can vary depending the content. Additionally, the nested structure is a fundamental characteristic of natural language (Hawkins, 2021).

Transformer hypothesis class. Some previous works Yun et al. (2019); Kim et al. (2022) treated the softmax with normalization as an approximation of hardmax, suggesting the potential importance of the normalization. In contrast, in this section, we remove the normalization in the denominator of softmax and investigate its ability for sequence modeling. Additionally, to address the discreteness of time and memory values, we consider Transformer with specific precision, as detailed in Appendix C. The precision technique is widely used in LLM training (Kalamkar et al., 2019), such as BF16. Formally, the hypothesis class is defined as follows, encompassing all normalization-free L -layer Transformer, configured with H Attn heads and FFN width m and using type-RPE and specific precision.

$$\mathcal{T}\mathcal{F}_{(L,H,m)}^{\text{type}} := \{\mathbf{TF} : \mathbf{TF} \text{ is an } L\text{-layer, } H\text{-head, } m\text{-width} \\ \text{Transformer with type-RPE and specific precision}\}. \quad (12)$$

4.2 Theoretical results and insights: The warm-up case

Theorem 4.1 (Approximation rate, warm-up case). *For any target $\mathbf{H} \in \mathcal{H}_{(1,M)}^{\text{Adap}}$ (8), rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists a two-layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(2,H,m)}^{\text{type}}$ (12) and a constant $C(n)$*

such that: if the width satisfies $m \geq \begin{cases} \tilde{\Omega}(\sum_{i=1}^M \|g_i\|_{\mathcal{B}}^2) & , \text{type} = \text{lin} \\ \tilde{\Omega}(\sum_{i=1}^M \|\log g_i\|_{\mathcal{B}}^2 T_i^2) & , \text{type} = \text{log} \end{cases}$, then the following approximation rate holds:

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

$$\text{where } \mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right) \text{ and } \mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M e^{0.01T_i}\right)^{n+1}\right) & , \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M T_i^{1.01}\right)^{n+1}\right) & , \text{type} = \text{log} \end{cases}.$$

In Theorem 4.1, we present the *approximation rate* of two-layer Transformer for the warm-up case: modeling 1-adaptive, long but M -sparse memories. This theorem reveals that the approximation error comprises two distinct components: the error in the FFN component \mathcal{E}_{FFN} and the error in the Attn component $\mathcal{E}_{\text{Attn}}(\text{type})$. A critical difference from 3.1 is the presence of the condition related to the width m of FFN layers. This term arises from using the FFN layer to approximate the memory function g_i . Owing to the discreteness of memory g_i and the implementation of rounding operations, the approximation within rounding accuracy all achieves zero error after rounding, while it can not get correct rounding beyond this accuracy. In contrast, the error \mathcal{E}_{FFN} is caused by using FFN to approximate the readout function f , the same as \mathcal{E}_{FFN} in Theorem 3.1.

The proof of Theorem 4.1 can be found in Appendix C.1. Theorem 4.1 and its proof offer several critical insights into the underlying mechanism of Transformer.

Distinct roles of Attn layers and FFN layers. Our proof elucidates that the FFN layers are tasked with approximating the readout function f and memory functions g_i , while the Attn layers are responsible for the extraction of the adaptive memories. It is essential to clarify the difference between “approximating memory functions” and “memory extraction”. The former refers to utilizing some function to estimate the memory function g_i , whereas the latter pertains to extracting the token $\mathbf{x}_{t-g_i(\mathbf{x}_t)}$ from the memory location.

Cooperation between DP and RPE. In the 2-nd Attn layer, the extraction of the memory functions is achieved through an interplay between DP and RPE. Specifically, this is done through a *nice interaction between the temporal space (provided by RPE) and the token space (provided by DP)*. Please refer to Appendix C.1 for more details.

Rethinking DP in Attn. Our proof highlights that the core mechanism of Attn is to provide a nice interaction between the temporal space and the token space through the cooperation of DP and RPE. This leads us to the following question: *Is DP in Attn necessary and replaceable?* The following two propositions provide some hints.

Proposition 4.2 (DP vs. DP-free (informal)). *There exists a target $\mathbf{H} \in \mathcal{H}_{(1,1)}^{\text{Adap}}$ (10) such that:*

(A) *For any $\epsilon > 0$, there exists a 1-layer Attn $\mathbf{Attn}^{\text{DP}}$ such that $\|\mathbf{H} - \mathbf{Attn}^{\text{DP}}\| \leq \epsilon$.*

(B) *For any 1-layer DP-free Attn $\mathbf{Attn}^{\text{DPF}}$, a uniform lower bound holds: $\|\mathbf{H} - \mathbf{Attn}^{\text{DPF}}\| \geq \frac{2}{3}$.*

Proposition 4.2 reveal a significant distinction in the expressiveness of two network types for modeling adaptive, long, but sparse memories. Specifically, 1-layer Attn with DP can effectively model this task, while 1-layer DP-free Attn provably fails. This finding underscores the essential role of DP in providing the necessary nonlinearity for Attn to model adaptive memories. The formal version of Proposition 4.2 and its proof can be found in Appendix C.2.

Proposition 4.3 (Substitute for DP (informal)). *There exists a substitute structure for DP, requiring only $\mathcal{O}(D)$ parameters (compared to $\mathcal{O}(D^2)$ in standard DP) that can effectively model $\mathbf{H} \in \mathcal{H}_{(1,M)}^{\text{Adap}}$ (10). Specifically, if we substitute DP with this structure, 1-layer Transformer can achieve the same approximation rate as stated in Section 4.1.*

Proposition 4.3 demonstrates the existence of a structurally simpler yet effective alternative to traditional DP for modeling (10). This alternative is proposed based on our insights into the role of Attn in facilitating the interaction between the temporal space and the token space. Specifically, we propose a more direct structure to achieve this interaction. More details are deferred to Appendix C.3.

4.3 Theoretical results and insights: The general case

Theorem 4.4 (Approximation rate, general case). *For any target $\mathbf{H} \in \mathcal{H}_{(K,M)}^{\text{Adap}}$, rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists an L -layer ($L = K + 1 + \mathbb{I}\{M \geq K + 1\}$) Transformer $\mathbf{TF} \in \mathcal{TF}_{(L,H,m)}^{\text{type}}$ (12) and a constant $C(n)$ such that: if the width satisfies if the width satisfies $m \geq \begin{cases} \tilde{\Omega}(\max_{i \in [K]} \vee \sum_{i=K+1}^M \|g_i\|_{\mathcal{B}}^2) & , \text{type} = \text{lin}, \\ \tilde{\Omega}(\max_{i \in [K]} \vee \sum_{i=K+1}^M \|\log g_i\|_{\mathcal{B}}^2 T_i^2) & , \text{type} = \text{log} \end{cases}$, then the following approximation rate holds:*

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}), \text{ where}$$

$$\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right), \mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l} + \left(\sum_{l=K+1}^M e^{0.01T_l}\right)^{2n+2}}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)} + \left(\sum_{l=K+1}^M T_l^{1.01}\right)^{2n+2}}\right), & \text{type} = \text{log} \end{cases}$$

In Theorem 4.4, we establish the *approximation rate* of deep Transformer for modeling K -adaptive, long but M -sparse memories. Similar to that in Theorem 4.1, the approximation error divides into two distinct terms. A key difference from Theorem 4.1 is the impact of the nested relationships among the memory functions on the required number of layers, Attn heads, and the width of FFN layers. The nested structure within the initial K memories mandates sequential processing in the first K layers one by one. If $M \geq K + 1$, then in the $K + 1$ -th layer, the remaining $M - K$ non-nested memory functions t_{K+1}, \dots, t_M are concurrently processed. The proof of Theorem 4.4 is deferred to Appendix D.1.

Distinct roles of the number of layers L , the number of Attn heads H , and the width of FFN layers m . Theorem 4.4 and its proof highlight the distinct roles of three key hyper-parameters of Transformer: L , H , and m . Deeper Transformer are capable of handling the memories with more intricate nested relationships, requiring a $K + 1$ layer network for a nesting complexity of K . In contrast, the number of heads and width needed is dictated by the individual complexity of memory functions themselves ($\|g_i\|_{\mathcal{B}}, \|\log g_i\|_{\mathcal{B}}, T_i$ for memory g_i), necessitating that each layer's Attn heads and FFN width are sufficient to capture the memory functions extracted in that layer. This understanding is quite intuitive: If the content of the next token relies on a few previous tokens in an independent way, we can treat each such dependence with a separate attention head. There is no need for many layers.

Mitigating required head and width with depth. Recalling Theorem 4.1, the memories lacking nested relationships can be efficiently approximated by 2-layer Transformer with a sufficient number of heads and width. The subsequent proposition further explores how increasing the depth of Transformer can influence its efficiency for modeling memories without nested relationships.

Proposition 4.5 (Deep network, warm-up case). *For any target $\mathbf{H} \in \mathcal{H}_{(1,M)}^{\text{Adap}}$ (8), rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists an $M + 1$ -layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(M+1,H,m)}^{\text{type}}$ (12) and a constant $C(n)$ such that: if the width satisfies $m \geq \begin{cases} \tilde{\Omega}(\max_{i \in [K]} \|g_i\|_{\mathcal{B}}^2) & , \text{type} = \text{lin}, \\ \tilde{\Omega}(\max_{i \in [K]} \|\log g_i\|_{\mathcal{B}}^2 T_i^2) & , \text{type} = \text{log} \end{cases}$, then the following approximation rate holds:*

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

$$\text{where } \mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right) \text{ and } \mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}\right), & \text{type} = \text{log} \end{cases}$$

Upon comparing Proposition 4.5 with Theorem 4.1, a notable distinction becomes evident between 2-layer and $M + 1$ -layer Transformer in terms of the requirement of the number of Attn heads and the width of FFN layers. Specifically, for 2-layer Transformer, the required width is proportionally linked

to the *sum* of all the memory functions’ complexity ($\|g_i\|_{\mathcal{B}}, \|\log g_i\|_{\mathcal{B}}, T_i$ for memory function g_i). In contrast, for $M + 1$ -layer Transformer, the required width correlates with the *maximum* complexity of the memory functions, much lower than that for 2-layer Transformer. Similarly, the required number of heads for $M + 1$ -layer Transformer is much fewer than that for 2-layer Transformer. Please refer to Appendix D.2 for a detailed comparison. The observation suggests that increased depth can significantly reduce the demands on the number of heads and the width. The underlying reason is that deep networks can distribute the memories across different layers for processing, with each layer focusing on approximating only a single memory function.

5 Essentially M -sparse memories

5.1 Problem formulation

In language tasks, each token possesses clear semantic meaning. As a result, the structure of the memory is sparse in the original space. This aligns well with our modeling assumptions discussed in Section 3 and 4. However, in other machine learning tasks, we may encounter situations where the input tokens lack distinct semantic meaning. This might happen in image processing or classical signal processing. In these situations, the memory structure could potentially be dense in the original space. Nonetheless, the memory structure might exhibit sparsity in some transformed domain. We call such memory structure “essentially sparse”. In this section, we study the situation in which the memory structure is long-ranged but essentially sparse. For simplicity, we consider the situation in which the positions of the memory kernels are fixed. The analysis can be easily extended to the situation with an adaptive memory structure.

Fixed, essentially M -sparse memory. Consider the following situation:

$$y_t = f((\mathbf{X} * \rho_1)(t), \dots, (\mathbf{X} * \rho_M)(t)), \quad (13)$$

where $\rho_1(\cdot), \dots, \rho_M(\cdot) \in \ell^1(\mathbb{N})$ serve as memory kernels, and $(\mathbf{X} * \rho_k)(t) = \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \rho_k(s)$ denotes the convolution of the inputs with kernel ρ_k .

Target class and Transformer hypothesis class. The target class for modeling essentially sparse memories is defined as:

$$\mathcal{H}^{\text{Ess}} := \{\mathbf{H} : \mathbf{H}_t(\mathbf{X}) = (13), \text{ where } \rho_1, \dots, \rho_M \in \ell^1(\mathbb{N}), f \in \mathcal{B} \cap \mathcal{L}\}. \quad (14)$$

For the hypothesis class, we consider one-layer dot-product-free Transformer with Attn head number H and FFN width m , as defined in (7).

Examples. Essentially sparse memories are prevalent in real-world scenarios:

(I) *Image Tasks.* In CV, a fundamental objective is identifying and representing meaningful “features”, such as ears, nose, etc. These features can often be modeled using convolution kernels, leading to a task in the form $y = f(\mathbf{X} * \rho_{\text{eye}}, \mathbf{X} * \rho_{\text{nose}}, \mathbf{X} * \rho_{\text{ear}})$. This is an extension of the task we discussed above, in which the kernel functions $\{\rho_j\}$ are data-dependent (“adaptive” in the terminology used in the previous section).

(II) *Signal processing.* In signal processing, it is commonly the case that the signals are highly sparse under Wavelet or Fourier transforms. For instance, let $\psi(\cdot)$ be a wavelet function and define $\psi_{a,b}(t) := \psi(\frac{t-b}{a})/\sqrt{|a|}$. Then we have $y = f(\mathbf{X} * \psi_{a_1,b_1}, \dots, \mathbf{X} * \psi_{a_M,b_M})$ where $(a_1, b_1), \dots, (a_M, b_M)$ might be data-dependent.

(III) *Mathematical calculation.* Consider algebraic operations where memory exhibits sparsity under specific linear transformations. For example, $y_t = 10x_t + x_{t-4}/(\sum_{s=0}^{100} w_s x_{t-10-s}) - \sum_{s=0}^{+\infty} v_s x_{t-100-s}$ can be represented in our framework as $y = f(\mathbf{X} * \rho_1, \dots, \mathbf{X} * \rho_4)$, where each ρ_i represents a specific linear transformation.

5.2 Theoretical results and insights

Theorem 5.1 (Approximation rates).

(A) Consider \mathcal{H}^{Ess} (14) with exponentially decayed memory kernels, i.e., there exists $\beta > 0$ such that

$\rho_1(t), \dots, \rho_M(t) = \mathcal{O}(e^{-\beta t})$. Then for any target $\mathbf{H} \in \mathcal{H}^{\text{Ess}}$, rate $n \in \lceil \lceil 99\beta \rceil \rceil$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer DP-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,lin}}$ (7) and a constant $C(n)$ such that

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}};$$

(B) Consider \mathcal{H}^{Ess} (14) with polynomially decayed memory kernels, i.e., there exists $\beta > 1$ such that $\rho_1(t), \dots, \rho_M(t) = \mathcal{O}(t^{-\beta})$. Then for any target $\mathbf{H} \in \mathcal{H}^{\text{Ess}}$, rate $n \in \lceil \lceil 0.99\beta \rceil - 1 \rceil$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer DP-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,log}}$ (7) and a constant $C(n)$ such that

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}};$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathbb{R}}}{\sqrt{m}}\right)$, $\mathcal{E}_{\text{Attn}} = \mathcal{O}\left(\frac{C(n)M^{n+1}}{H^n}\right)$.

Theorem 5.1 illustrates that one-layer DP-free Transformer with lin-RPE is effective in modeling essentially sparse memories with exponentially decayed kernels, and one-layer DP-free Transformer with log-RPE can efficiently model the memories with polynomially decayed kernels. A key difference between Theorem 5.1 and Theorem 3.1 lies in the memory kernels they address. In Theorem 5.1, the Attn layer should approximate general memory kernels $\rho_i(\cdot)$, instead of approximating indicator kernels $\mathbb{I}\{\cdot = T_i\}$ in Theorem 3.1. The proof of Theorem 5.1 can be found in Appendix E.

Overcoming the Curse of Memory (CoM). For recurrent neural networks (RNN), it was discovered (Li et al., 2021; 2022) that both approximation and optimization become exceedingly difficult when the target has long-term memory. This phenomenon is referred as the “curse of memory”, or “CoM”. It was shown in (Li et al., 2021; 2022) that RNN requires an exponentially large number of neurons to approximate targets with heavy-tailed memory kernels, such as the ones that exhibit polynomial decay. In contrast, Theorem 5.1 reveals that Transformer with log-RPE efficiently handles polynomial decaying memory kernels, requiring only a polynomial number of neurons for effective approximation. This finding theoretically elucidates the superior performance of T5’s RPE and KERPLE(log) in length generalization task in practice (Section G.1).

6 Experimental Validation

As summarized in Section 1, our theoretical analysis reveals novel insights into the expressive power and mechanisms of Transformer. To validate these insights, we conduct experiments ranging from simple toy models to more complex language model pre-training. Due to space constraints, detailed experimental validation and practical implications of our insights are presented in Appendix H.

7 Conclusion and Future Work

In this work, we investigate theoretically the expressive power and the mechanisms of Transformer for modeling long but sparse memories. Our analysis establishes explicit approximation rates and offers much-needed insights into the functionalities of the various components of Transformer. However, we still have a long way to go for a full theoretical understanding of Transformer. For instance, although we have investigated the mechanisms of Transformer in terms of expressive power, the evolution of the mechanisms during the training process remains elusive. Recent studies revealed that Transformer exhibits multi-phase learning dynamics (Boix-Adsera et al., 2023) and undergoes phase transitions (Olsson et al., 2022) during training, akin to the phenomenon of learning with increasing complexity in classical neural networks (Kalimeris et al., 2019; Xu et al., 2019; Rahaman et al., 2019; Abbe et al., 2023a; Wang and Ma, 2023). These and other issues will be studied in future work.

Acknowledgments

This work is supported in part by the National Key Basic Research Program of China (No. 2015CB856000). We thank Prof. Qianxiao Li, Prof. Lei Wu, Dr. Zhong Li, and Dr. Hongkang Yang for helpful discussions and anonymous reviewers for their valuable suggestions.

References

- Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023a. 10
- Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. *International Conference on Machine Learning*, 2023b. 17
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022. 17
- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023. 2, 17
- Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022. 2, 17
- Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017. 4, 64
- Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023. 17
- Andrew R Barron. Neural net approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, volume 1, pages 69–72, 1992. 4, 64
- Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993. 4, 64
- Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994. 4, 64
- Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966. 4, 64
- Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 17
- Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 17
- Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse boolean functions. *arXiv preprint arXiv:2211.12316*, 2022. 5
- Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. *arXiv preprint arXiv:2306.07042*, 2023. 10
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 17
- Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008. 17
- Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022. 3, 17
- Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 2, 17

- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. 17
- Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 2, 17
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *International Conference on Learning Representations*, 2019. 16
- David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. 17
- Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019. 4, 64
- Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pages 1–38, 2021. 4, 64
- Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022. 5, 16
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>. 17
- Guhao Feng, Yuntian Gu, Bohang Zhang, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *arXiv preprint arXiv:2305.15408*, 2023. 17
- W Nelson Francis and Henry Kucera. Brown corpus manual. *Letters to the Editor*, 5(2):7, 1979. 17
- Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022. 17
- Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *International Conference on Machine Learning*, 2023. 16
- Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. <http://Skylion007.github.io/OpenWebTextCorpus>, 2019. 66
- Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. 17
- Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Basic Books, 2021. 6
- Dunham Jackson. *The theory of approximation*, volume 11. American Mathematical Soc., 1930. 4, 65
- Haotian Jiang and Qianxiao Li. Approximation theory of transformer networks for sequence modeling. *arXiv preprint arXiv:2305.18475*, 2023. 17
- Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019. 6, 23

- Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019. 10
- Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 2023. 3, 17
- Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022. 6
- Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 2, 17
- Zhong Li, Jiequn Han, Qianxiao Li, and Weinan E. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. *International Conference on Learning Representations*, 2021. 10
- Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022. 10
- Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023. 17
- Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022. 17
- Chao Ma and Lexing Ying. Why self-attention is natural for sequence-to-sequence problems? a perspective from symmetries. *arXiv preprint arXiv:2210.06741*, 2022. 17
- Chao Ma, Stephan Wojtowytsch, Lei Wu, and Weinan E. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020. 4, 64, 65
- Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023. 17
- William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023a. 17
- William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023b. 17
- William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022. 17
- Yves Meyer. *Wavelets and Operators: Volume 1*. Cambridge university press, 1992. 17
- Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, 2003. 17
- Joakim Nivre and Mario Scholz. Deterministic dependency parsing of english text. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 64–70, 2004. 17
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. 10, 17

- Santiago Ontanón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. Making transformers solve compositional tasks. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 2, 17
- OpenAI. Gpt-4 technical report. <https://cdn.openai.com/papers/gpt-4.pdf>, 2023. 17
- Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is turing complete. *The Journal of Machine Learning Research*, 22(1):3463–3497, 2021. 16
- Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations*, 2022. 1, 17
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 17, 64
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 10
- Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 17
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 1, 17
- Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Do pretrained transformers really learn in-context by gradient descent? *arXiv preprint arXiv:2310.08540*, 2023. 17
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1, 17
- Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *International Conference on Machine Learning*, pages 10183–10192. PMLR, 2021. 2, 17
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 17
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3, 17, 66
- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023. 17
- Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Max Vladymyrov, Razvan Pascanu, et al. Uncovering mesa-optimization algorithms in transformers. *arXiv preprint arXiv:2309.05858*, 2023. 17
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 17
- Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of relu networks. *Advances in Neural Information Processing Systems*, 2023. 10
- Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2, 17

- Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022a. 16
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b. 17
- Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021. 17
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 17
- Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 2019. 10
- Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019. 6, 16
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020. 17
- Zhongwang Zhang, Zhiwei Wang, Junjie Yao, Zhangchen Zhou, Xiaolong Li, Zhi-Qin John Xu, et al. Anchor function: a type of benchmark functions for studying language models. *arXiv preprint arXiv:2401.08309*, 2024. 17

Appendix

A Detailed Related Works	16
B Proof of Section 3	18
B.1 Proof of Theorem 3.1	18
C Proof of Section 4.2	23
C.1 Proof of Theorem 4.1	23
C.2 Proof of Proposition 4.2	33
C.3 Proof of Proposition 4.3	35
D Proof of Section 4.3	38
D.1 Proof of Theorem 4.4	38
D.2 Proof of Proposition 4.5	48
E Proof of Section 5	50
E.1 Proof of Theorem 5.1	50
F Key Lemmas about Approximation	55
F.1 Approximation by the sum of exponential decay	55
F.2 Approximation by the sum of polynomial decay	59
G Some Background and Proof Preparation	64
G.1 T5’s relative positional encoding	64
G.2 Barron space theory	64
G.3 Useful approximation lemmas	65
H Experiments	66
H.1 Restatement of our theoretical insights	66
H.2 Experimental Validation	66
H.3 Practical Implications	69

A Detailed Related Works

Theoretical results of Transformer. We first review the expressive power results of Transformer. [Yun et al. \(2019\)](#) first proved the universal approximation property (UAP) of Transformer, highlighting the crucial role of PE in breaking permutation invariance. [Edelman et al. \(2022\)](#) demonstrated that Transformer can approximate fixed sparse functions. [Dehghani et al. \(2019\)](#); [Pérez et al. \(2021\)](#); [Wei et al. \(2022a\)](#) explored the Turing-completeness of infinite-precision and finite-precision Transformer. [Giannou et al. \(2023\)](#) showed that looped Transformer can implement practical computer programs.

Jiang and Li (2023) provided explicit approximation rates for Transformer in sequences modeling with inherent graph structures. Liu et al. (2022) found that Transformer can execute finite-state automata. Ma and Ying (2022) asserted the natural suitability of Attn for achieving permutation equivariance. Besides these affirmative results, several studies characterized the expressivity limitation of Transformers, particularly in modeling formal languages or simulating circuits (Hahn, 2020; Weiss et al., 2021; Bhattamishra et al., 2020; Merrill and Sabharwal, 2023b; Merrill et al., 2022). Additionally Feng et al. (2023); Merrill and Sabharwal (2023a) examined the expressivity of Transformer using Chain of Thought prompting (Wei et al., 2022b). Moreover, some studies showed that the in-context learning ability of Transformer is attainable by simulating gradient-based iterations across various layers (Garg et al., 2022; Akyürek et al., 2022; von Oswald et al., 2023; Von Oswald et al., 2023; Mahankali et al., 2023; Bai et al., 2023; Shen et al., 2023). Besides, experimental studies also provide insights into the mechanisms of Transformer through induction head (Elhage et al., 2021; Olsson et al., 2022), information flow (Wang et al., 2023), anchor functions (Zhang et al., 2024), etc.

Positional encoding. One core component of Transformer is the PE, which facilitates the representation of input sequence order. Theoretically, Transformer without PE lacks UAP and is restricted to representing permutation-invariant functions. PE was first introduced in Vaswani et al. (2017). It has limitations in encoding unseen positions. To overcome this difficulty, Shaw et al. (2018) introduced RPE. Subsequent studies proposed various different RPE types. Notable examples include T5’s RPE (Raffel et al., 2020), Rotary RPE (Su et al., 2024) (utilized in PaLM (Chowdhery et al., 2023) and LLaMA (Touvron et al., 2023)), Alibi RPE (Press et al., 2022) (employed in BLOOM (Workshop et al., 2022)), and KERPLE (Chi et al., 2022). A prevailing belief is that RPEs can outperform APEs in the “length generalization task” (Ontanon et al., 2022; Csordás et al., 2021)– the ability to generalize from smaller training contexts to larger ones, a critical challenge for Large Language Models (Anil et al., 2022; Abbe et al., 2023b). However, Press et al. (2022) revealed that the commonly used Rotary RPE may exhibit suboptimal performance in this task. The recent work (Kazemnejad et al., 2023; Chi et al., 2022) conducted systematic experiments comparing the length generalization capabilities of Transformers with various RPEs and APEs, suggesting that *the RPEs used in T5 and KERPLE(log) demonstrate superior performance over other types.*

Rethinking dot-product. Another critical component of Transformer is the DP structure. Due to its quadratic cost as a function of the sequence length, the necessity of DP has always been questioned. Numerous variants of DP have been proposed, demonstrating competitive performance across diverse tasks. Representative examples include Longformer (Beltagy et al., 2020), Big Bird (Zaheer et al., 2020), Reformer (Kitaev et al., 2020), Linformer (Wang et al., 2020), Performer (Choromanski et al., 2020), Synthesizer (Tay et al., 2021), etc. In particular, a recent study (Allen-Zhu and Li, 2023) compared standard and DP-free Transformers in modeling “context-free grammar”. Their findings suggested that the presence of DP has a marginal impact on performance. *These evidences motivate us to rethink the necessity of DP in Transformer.*

Sparsity (Donoho, 2006; Candès and Wakin, 2008) has gained considerable attention in sequence modeling. In classical signal processing, there is a prevailing notion that valuable signals are extremely sparse. For example, when representing an image, one often finds that only a few wavelet coefficients hold significant values in wavelet space (Meyer, 1992). In NLP, the starting point of the traditional n -gram model (Shannon, 1948) is that the next token only relies on a few previous tokens. Such models, however, overlook long-range information, often resulting in suboptimal performance. For NLP tasks such as dependency parsing (Nivre and Scholz, 2004), sentiment analysis (Nasukawa and Yi, 2003), part-of-speech tagging (Francis and Kucera, 1979), and continuation writing (Brown et al., 2020; OpenAI, 2023), it is indeed often the case that only a limited subset of preceding information is crucial for accurate prediction. However, these relevant information can be quite distant. For instance, the resolution of a mystery novel may hinge on elements introduced at the outset. Moreover, for Transformer networks, extensive research into the visualization and interpretability has revealed that (i) the learned activation maps of FFN layers are extremely sparse (Li et al., 2023); (ii) the learned self-attention matrices exhibit notable sparsity, yet it does not closely resemble a diagonal configuration (Elhage et al., 2021). These observations suggest that the prediction of the next token is influenced by a small number of previous tokens which might be far away. Therefore, *being able to represent sparse but long-range dependence is important for sequence modeling.*

B Proof of Section 3

B.1 Proof of Theorem 3.1

In this subsection, we give the detailed proofs of fixed, long but sparse memory:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-T_1}, \dots, \mathbf{x}_{t-T_M}),$$

where $1 \leq T_1 < \dots < T_M < +\infty$ signify the fixed positions of the memories.

Theorem B.1 (Restatement of Theorem 3.1). *For any target $\mathbf{H} \in \mathcal{H}^{\text{Fix}}$ (6), rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,type}}$ (7) and a constant $C(n)$ such that*

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathbb{B}}}{\sqrt{m}}\right)$ and

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M e^{0.01T_i}\right)^{n+1}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M T_i^{1.01}\right)^{n+1}\right), & \text{type} = \text{log} \end{cases}.$$

Proof of Theorem B.1.

First, we choose the embedding dimension $D = (M+1)d$, and select the simple embedding $\mathbf{W}_E = (\mathbf{I}_{d \times d}, \mathbf{0})^\top \in \mathbb{R}^{D \times d}$, $\mathbf{b}_E = \mathbf{0} \in \mathbb{R}^D$.

Then for any input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}}$, the token after embedding satisfies:

$$\mathbf{x}_t^E = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = (\mathbf{x}_t^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^D.$$

For one-layer Dot-product-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,type}}$ with ϕ_{type} , the output token $\mathbf{TF}_t(\mathbf{X})$ of t -th input token \mathbf{x}_t satisfies:

$$\begin{aligned} \mathbf{x}_t^{(1/2)} &= \mathbf{x}_t^{(0)} + \mathbf{W}_O^{(1)} \sum_{h=1}^H \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}), \\ \mathbf{x}_t^{(1)} &= \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \end{aligned}$$

where

$$\text{Attn}_t^{(1,h)}(\mathbf{X}) = \mathbf{W}_V^{(1,h)} \sum_{s=0}^{+\infty} \frac{\mathbf{x}_{t-s} \exp(p^{(1,h)} \phi_{\text{type}}(s))}{\sum_{j=0}^{+\infty} \exp(p^{(1,h)} \phi_{\text{type}}(j))}.$$

This proof can be summarized as the following process:

$$\begin{aligned} & \dots \quad \mathbf{x}_t^E \quad \dots \\ \text{Step I. Attn layer} & \downarrow \\ & \dots \quad \mathbf{x}_t^{(1/2)} \approx (\mathbf{x}_t^\top, \mathbf{x}_{t-T_1}^\top, \dots, \mathbf{x}_{t-T_M}^\top)^\top \quad \dots \\ \text{Step II. FFN layer} & \downarrow \\ & \dots \quad \mathbf{x}_t^{(1)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-T_1}, \dots, \mathbf{x}_{t-T_M}) \quad \dots \end{aligned}$$

Now we give the formal proof.

Step I. Extract the memory locations by (Dot-product-free) Attn layer.

We consider to use H_k attention heads (from $\sum_{i=1}^{k-1} H_i + 1$ -th head to $\sum_{i=1}^k H_i$ -th head) to extract it, and it satisfies to $\sum_{k=1}^M H_k = H$.

For simplicity, we denote the following projection matrices:

$$\begin{aligned} \mathbf{P}^{(k)} &:= (\mathbf{0}_{d \times kd} \quad \mathbf{I}_{d \times d} \quad \mathbf{0}) \in \mathbb{R}^{d \times D}, \quad 1 \leq k \leq M. \\ \mathbf{P}_{\perp}^{(k)} &:= \begin{pmatrix} \mathbf{I}_{kd \times kd} & \mathbf{0}_{d \times d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{d \times d} & \mathbf{I}_{(M-k)d \times (M-k)d} \end{pmatrix} \in \mathbb{R}^{Md \times D}, \quad 1 \leq k \leq M. \end{aligned}$$

Now we consider the extraction of k -th memory \mathbf{x}_{t-T_k} ($1 \leq k \leq M$).

- **Case type = lin.**

By Lemma F.1, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and a function

$$\phi_k^{\text{exp}}(t) = \sum_{\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h e^{-\beta_h t}$$

such that $\beta_h > 0$ and

$$\|\mathbb{I}\{\cdot = T_k\} - \phi_k^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} = \sum_{s=0}^{+\infty} |\mathbb{I}\{s = T_k\} - \phi_k^{\text{exp}}(s)| \leq C(n) \frac{e^{0.01(n+1)T_k}}{H_k^n}.$$

Therefore, for these attention heads ($\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i$), we can choose

$$p^{(1,h)} = \beta_h, \quad \mathbf{W}_V^{(1,h)} = \alpha_h \left(\sum_{j=0}^{+\infty} \exp(-\beta_h j) \right) \boldsymbol{\delta}_{(k+1,1)}^{d \times d},$$

where $\boldsymbol{\delta}^{(k+1,1)} \in \mathbb{R}^{D \times D}$ means that: it equals to $\mathbf{I}_{d \times d}$ for the $(k+1, 1)$ -th $d \times d$ blocks, and $\mathbf{0}_{d \times d}$ for the other $d \times d$ blocks.

Then it holds that:

$$\sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-s} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^D.$$

This implies:

$$\begin{aligned} \mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) &= \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \mathbf{x}_{t-s}, \\ \mathbf{P}_{\perp}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) &= \mathbf{0}, \end{aligned}$$

moreover, the following estimate holds:

$$\begin{aligned} & \left\| \mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) - \mathbf{x}_{t-T_k} \right\|_2 \\ &= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \mathbf{x}_{t-s} - \mathbf{x}_{t-T_k} \right\|_2 \end{aligned}$$

$$\begin{aligned}
&= \left\| \sum_{s=0}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h s} - \mathbb{I}\{s = T_k\} \right) \mathbf{x}_{t-s} \right\|_2 \\
&\leq \sum_{s=0}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h s} - \mathbb{I}\{s = T_k\} \right| \\
&= \|\phi_k^{\text{exp}}(\cdot) - \mathbb{I}\{\cdot = T_k\}\|_{\ell_1(\mathbb{N})} \leq C(n) \frac{e^{0.01(n+1)T_k}}{H_k^n}.
\end{aligned}$$

- **Case type = log.**

By Lemma F.4, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and a function

$$\phi_k^{\text{poly}}(t) = \sum_{\sum_{i=1}^{k-1} H_i+1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h t^{-\beta_h},$$

such that $\beta_h > 1$ and

$$\left\| \mathbb{I}\{\cdot = T_k\} - \phi_k^{\text{poly}}(\cdot) \right\|_{\ell_1(\mathbb{N}_+)} = \sum_{s=1}^{+\infty} \left| \mathbb{I}\{s = T_k\} - \phi_k^{\text{poly}}(s) \right| \leq C(n) \frac{T_k^{1.01(n+1)} H_k^n}{\cdot}.$$

Therefore, for these attention heads $(\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i)$, we can choose

$$p^{(1,h)} = \beta_h, \quad \mathbf{W}_V^{(1,h)} = \alpha_h \left(\sum_{j=1}^{+\infty} j^{-\beta_h} \right) \boldsymbol{\delta}^{(k+1,1)},$$

where $\boldsymbol{\delta}^{(k+1,1)} \in \mathbb{R}^{D \times D}$ means that: it equals to $\mathbf{I}_{d \times d}$ for the $(k+1, 1)$ -th $d \times d$ blocks, and $\mathbf{0}_{d \times d}$ for the other $d \times d$ blocks.

Then it holds that:

$$\begin{aligned}
\mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) &= \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=1}^{+\infty} s^{-\beta_h} \mathbf{x}_{t-s}, \\
\mathbf{P}_\perp^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) &= \mathbf{0},
\end{aligned}$$

moreover, the following estimate holds:

$$\begin{aligned}
&\left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) - \mathbf{x}_{t-T_k} \right\|_2 \\
&= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=1}^{+\infty} s^{-\beta_h} \mathbf{x}_{t-s} - \mathbf{x}_{t-T_k} \right\|_2 \\
&= \left\| \sum_{s=1}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h s^{-\beta_h} - \mathbb{I}\{s = T_k\} \right) \mathbf{x}_{t-s} \right\|_2
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{s=1}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h s^{-\beta_h} - \mathbb{I}\{s = T_k\} \right| \\
&= \left\| \phi_k^{\text{poly}}(\cdot) - \mathbb{I}\{\cdot = T_k\} \right\|_{\ell_1(\mathbb{N}_+)} \leq \mathcal{O} \left(C(n) \frac{T_k^{1.01(n+1)}}{H_k^n} \right).
\end{aligned}$$

Then we combine the results for all $k \in [M]$ for these two cases. By choose $\mathbf{W}_O = \mathbf{I}_D$, we have:

$$\begin{aligned}
&\left\| \mathbf{x}_t^{(1/2)} - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-T_1} \\ \vdots \\ \mathbf{x}_{t-T_M} \end{pmatrix} \right\|_2 \\
&= \left\| \begin{pmatrix} \mathbf{x}_t \\ \mathbf{0}_d \\ \vdots \\ \mathbf{0}_d \end{pmatrix} + \sum_{h=1}^M \mathbf{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-T_1} \\ \vdots \\ \mathbf{x}_{t-T_M} \end{pmatrix} \right\|_2 = \left\| \sum_{h=1}^M \mathbf{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_d \\ \mathbf{x}_{t-T_1} \\ \vdots \\ \mathbf{x}_{t-T_M} \end{pmatrix} \right\|_2 \\
&= \left\| \sum_{k=1}^M \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-T_k} \\ \mathbf{0}_d \end{pmatrix} \right) \right\|_2 \\
&\leq \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-T_k} \\ \mathbf{0}_d \end{pmatrix} \right\|_2 \\
&= \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}) - \mathbf{x}_{t-T_k} \right\|_2 \\
&\leq \mathcal{E}_{\text{Attn}}(\text{type}) := \begin{cases} C(n) \sum_{k=1}^M \frac{e^{0.01(n+1)T_k}}{H_k^n}, & \text{type} = \text{lin} \\ C(n) \sum_{k=1}^M \frac{T_k^{1.01(n+1)}}{H_k^n}, & \text{type} = \text{log} \end{cases}.
\end{aligned}$$

Consequently, one detail is to assign the head number $\{H_k\}_{k=1}^M$ such that the error's sum $\mathcal{E}_{\text{Attn}}(\text{type})$ is as small as possible. Our way is solving the minimization problem:

$$\begin{aligned}
&\min_{H_1, \dots, H_M} : \mathcal{E}_{\text{Attn}}(\text{type}) \\
&\text{s.t. } \sum_{k=1}^M H_k = H,
\end{aligned}$$

which suggests that we should choose the head number:

$$\begin{aligned}
H_k &= \frac{e^{0.01T_k}}{\sum_{j=1}^M e^{0.01T_j}} H, \quad k \in [M], \quad \text{type} = \text{lin}; \\
H_k &= \frac{T_k^{1.01}}{\sum_{j=1}^M T_j^{1.01}} H, \quad k \in [M], \quad \text{type} = \text{log}.
\end{aligned}$$

Thus, we obtain the bound in Step I:

$$\mathcal{E}_{\text{Attn}}(\text{type}) \leq \begin{cases} \frac{C(n)}{H^n} \left(\sum_{k=1}^M e^{0.01T_k} \right)^{n+1}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \left(\sum_{k=1}^M T_k^{1.01} \right)^{n+1}, & \text{type} = \text{log} \end{cases}.$$

Furthermore, by choosing $\mathcal{E}_{\text{Attn}}(\text{type}) \leq 1$, it holds that

$$\left\| \mathbf{x}_t^{(1/2)} \right\|_{\infty} \leq \left\| \mathbf{x}_t^{(1/2)} - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-t_1} \\ \vdots \\ \mathbf{x}_{t-t_M} \end{pmatrix} \right\|_{\infty} + \left\| \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-t_1} \\ \vdots \\ \mathbf{x}_{t-t_M} \end{pmatrix} \right\|_{\infty} \leq \mathcal{E}_{\text{Attn}}(\text{type}) + 1 \leq 2.$$

Step II. Approximate the readout function by FFN layer.

In this step, we aim to approximate the function f using two-layer network. By Lemma G.6, there exists a two layer neural network with m neurons defined on \mathbb{R}^D

$$\text{FFN}^{(1)}(\mathbf{y}) = \sum_{k=1}^m a_k \sigma(\mathbf{b}_k^{\top} \mathbf{y} + c_k)$$

such that

$$\mathcal{E}_{\text{FFN}} := \left\| \text{FFN}^{(1)} - f \right\|_{L^{\infty}([-2,2]^D)} \leq \tilde{\mathcal{O}} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right).$$

The final bound.

For any t and $\mathbf{X} \in \mathcal{X}$, it holds that

$$\begin{aligned} & \left\| \mathbf{H}_t(\mathbf{X}) - \mathbf{x}_t^{(1)} \right\| = \left| f(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &= \left| f(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}) - f(\mathbf{x}_t^{(1/2)}) + f(\mathbf{x}_t^{(1/2)}) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &\leq \left| f(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}) - f(\mathbf{x}_t^{(1/2)}) \right| + \left| f(\mathbf{x}_t^{(1/2)}) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &\leq \|f\|_{\text{Lip}} \left\| (\mathbf{x}_t^{\top}, \mathbf{x}_{t-t_1}^{\top}, \dots, \mathbf{x}_{t-t_M}^{\top}) - \mathbf{x}_t^{(1/2)} \right\|_2 + \left\| f - \text{FFN}^{(1)} \right\|_{L^{\infty}([-2,2]^D)} \\ &\leq \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}}(\text{type}) + \mathcal{E}_{\text{FFN}}, \end{aligned}$$

where

$$\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right);$$

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O} \left(\frac{C(n)}{H^n} \left(\sum_{k=1}^M e^{0.01 T_k} \right)^{n+1} \right), & \text{type} = \text{lin} \\ \mathcal{O} \left(\frac{C(n)}{H^n} \left(\sum_{k=1}^M T_k^{1.01} \right)^{n+1} \right), & \text{type} = \text{log} \end{cases}.$$

Due to the arbitrariness of t and \mathbf{X} , the proof is completed. □

C Proof of Section 4.2

In this section, we give the detailed proofs of the approximation theory of Transformer for modeling the warm-up case of adaptive, long but sparse memory:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}),$$

where the adaptive memory satisfies to:

$$t_k = g_k(\mathbf{x}_t), \quad k \in [M].$$

Moreover, $g_k(\cdot)$ generate positive integers for the input tokens, and there exist maximum values T_k such that $1 \leq g_k(\mathbf{x}_t) \leq T_k$ holds for any \mathbf{x}_t and $k \in [M]$.

To tackle the discrete values of the time and the memory values $g_k(\mathbf{x}_t)$, a modified version of standard FFN, termed ‘‘FFN with precision’’, is considered. This approach ensures that the output of FFN undergoes a simple rounding operation. Notably, the precision technique is widely used in LLM training (Kalamkar et al., 2019), such as BF16. Specifically, for Transformer using RPE with type, we use the following FFN with precision:

$$\begin{aligned} \widetilde{\text{FFN}}(\mathbf{x}) &:= [\text{FFN}(\mathbf{x})], & \text{type} &= \text{lin}; \\ \widetilde{\text{FFN}}(\mathbf{x}) &:= \log[\exp(\text{FFN}(\mathbf{x}))], & \text{type} &= \text{log}, \end{aligned} \quad (15)$$

where $[\cdot]$ signifies rounding to the nearest integer, i.e., $[x] = \arg \min_{n \in \mathbb{Z}} |n - x|$ ($x \in \mathbb{R}$).

It is important to note that the rounding obtained by using the operator $\log[\exp(z)]$, used in (15), is *quite fine*, which is much finer than the vanilla rounding obtained by $[z]$. To elaborate, the following proposition is presented:

Proposition C.1. *For any $z \geq 1$, the following holds:*

$$(i) \quad |\log[\exp(z)] - z| \leq \frac{1}{2 \min\{e^z, [e^z]\}}; \quad (ii) \quad |[z] - z| \leq \frac{1}{2}.$$

C.1 Proof of Theorem 4.1

Theorem C.2 (Restatement of Theorem 4.1). *For any target $\mathbf{H} \in \mathcal{H}_{(1,M)}^{\text{Adap}}$ (8), rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists a two-layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(2,H,m)}^{\text{NF,type}}$ (12) and a constant $C(n)$ such that: if the width satisfies*

$$m \geq \begin{cases} \tilde{\Omega}\left(\sum_{i=1}^M \|g_i\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin} \\ \tilde{\Omega}\left(\sum_{i=1}^M \|\log g_i\|_{\mathcal{B}}^2 T_i^2\right), & \text{type} = \text{log} \end{cases},$$

then the following approximation rate holds:

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right)$ and

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M e^{0.01T_i}\right)^{n+1}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{i=1}^M T_i^{1.01}\right)^{n+1}\right), & \text{type} = \text{log} \end{cases}.$$

Proof of Theorem C.2.

First, we choose the embedding dimension $D = (M + 1)(d + 1)$, and select a simple embedding $\mathbf{W}_E = (\mathbf{I}_{d \times d}, \mathbf{0})^\top \in \mathbb{R}^{D \times d}$, $\mathbf{b}_E = \mathbf{0} \in \mathbb{R}^D$.

Then for any input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}}$, the token after embedding satisfies:

$$\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = (\mathbf{x}_t^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^D.$$

To tackle the discrete values of $g_m(\mathbf{x}_t)$, we utilize $\widetilde{\text{FFN}}$, FFN with precision (15). It ensures that the output of FFN undergoes a simple rounding operation.

Thus, for two-layer normalization-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(2,H,m)}^{\text{NF,type}}$ with ϕ_{type} , the output token $\mathbf{x}_t^{(2)}$ of t -th input token satisfies:

$$\begin{aligned}\mathbf{x}_t^{(1/2)} &= \mathbf{x}_t^{(0)} + \mathbf{W}_O^{(1)} \sum_{h=1}^H \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}), \\ \mathbf{x}_t^{(1)} &= \mathbf{x}_t^{(1/2)} + \widetilde{\text{FFN}}^{(1)}(\mathbf{x}_t^{(1/2)}), \\ \mathbf{x}_t^{(3/2)} &= \mathbf{x}_t^{(1)} + \mathbf{W}_O^{(2)} \sum_{h=1}^H \text{Attn}_t^{(2,h)}(\mathbf{X}^{(1)}), \\ \mathbf{x}_t^{(2)} &= \text{FFN}^{(2)}(\mathbf{x}_t^{(3/2)}),\end{aligned}$$

where

$$\text{Attn}_t^{(l,h)}(\mathbf{X}) = \mathbf{W}_V^{(l,h)} \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \exp\left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-s} \right\rangle + p^{(l,h)} \phi_{\text{type}}(s)\right).$$

This proof can be summarized as the following process:

- **Case type = lin.**

$$\begin{aligned}& \mathbf{x}_t^{(0)} \\ \text{Step I. 1-st Attn } & \downarrow \\ & \mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)} \\ \text{Step II. 1-st FFN } & \downarrow \\ & \mathbf{x}_t^{(1)} = (\mathbf{x}_t^\top, \mathbf{0}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top \\ \text{Step III. 2-st Attn } & \downarrow \\ & \mathbf{x}_t^{(3/2)} \approx (\mathbf{x}_t^\top, \mathbf{x}_{t-g_1(\mathbf{x}_t)}^\top, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top \\ \text{Step IV. 2-st FFN } & \downarrow \\ & \mathbf{x}_t^{(2)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-g_1(\mathbf{x}_t)}, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)})\end{aligned}$$

- **Case type = log.**

$$\begin{aligned}& \mathbf{x}_t^{(0)} \\ \text{Step I. 1-st Attn } & \downarrow \\ & \mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)} \\ \text{Step II. 1-st FFN } & \downarrow \\ & \mathbf{x}_t^{(1)} = (\mathbf{x}_t^\top, \mathbf{0}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top \\ \text{Step III. 2-st Attn } & \downarrow \\ & \mathbf{x}_t^{(3/2)} \approx (\mathbf{x}_t^\top, \mathbf{x}_{t-g_1(\mathbf{x}_t)}^\top, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top \\ \text{Step IV. 2-st FFN } & \downarrow \\ & \mathbf{x}_t^{(2)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-g_1(\mathbf{x}_t)}, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)})\end{aligned}$$

Now we give the formal proof.

Step I. Identity map.

For the first Attn layer, we only need to do the identity map by taking $\mathbf{W}_0^{(1)} = \mathbf{0}$. Then $\mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)}$.

Step II. Approximate the adaptive memory function by the first FFN layer.

- **Case type = lin.** Our main idea is that using the first FFN layer to express $(\mathbf{x}_t^\top, \mathbf{0}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top$ exactly.

First, we consider to approximate the r -th memory function $g_r(\mathbf{x})$ by standard FFN.

For any $r \in [M]$, by Lemma G.6, there exists a two-layer neural network with m_r neurons

$$f_{(1,r)}^{2\text{NN}}(\mathbf{x}) = \sum_{k=1}^{m_r} a_k^{(1,r)} \sigma \left(\mathbf{b}_k^{(1,r)\top} \mathbf{x} + c_k^{(1,r)} \right)$$

defined on \mathbb{R}^d such that

$$\left\| g_r - f_{(1,r)}^{2\text{NN}} \right\|_{L^\infty([-1,1]^D)} \leq \tilde{\mathcal{O}} \left(\frac{\|g_r\|_{\mathcal{B}}}{\sqrt{m_r}} \right).$$

Therefore, if we choose

$$\tilde{\mathcal{O}} \left(\frac{\|g_r\|_{\mathcal{B}}}{\sqrt{m_r}} \right) < \frac{1}{2},$$

the following holds:

$$\left| g_r(\mathbf{x}_t) - f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t) \right| \leq \left\| g_r - f_{(1,r)}^{2\text{NN}} \right\|_{L^\infty([-1,1]^d)} < \frac{1}{2},$$

Noticing $g_r(\mathbf{x}_t) \in \mathbb{N}_+$, we have $\left[f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t) \right] = g_r(\mathbf{x}_t)$, which implies:

$$\widetilde{f_{(1,r)}^{2\text{NN}}}(\mathbf{x}_t) = \left[f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t) \right] = g_r(\mathbf{x}_t).$$

Consequently, in order to construct the form $(\mathbf{0}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top \in \mathbb{R}^D$, we need to arrange the parameters $a_k^{(1,r)}$, $\mathbf{b}_k^{(1,r)}$, and $c_k^{(1,r)}$ ($k \in [m_r], r \in [M]$) appropriately.

Denote $\bar{\mathbf{b}}_k^{(1,r)} = (\mathbf{b}_k^{(1,r)\top}, \mathbf{0}^\top)^\top \in \mathbb{R}^D$ for $k \in [m_r], r \in [M]$. Consider the following two-layer neural network with $1 + \sum_{r=1}^M m_r$ neurons defined on \mathbb{R}^D :

$$\begin{aligned} \mathbf{FFN}^{(1)}(\mathbf{x}) &= \sum_{r=1}^M \sum_{1 + \sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} \mathbf{e}_{D-M+r-1} a_k^{(1,r)} \sigma \left(\bar{\mathbf{b}}_k^{(1,r)\top} \mathbf{x} + c_k^{(1,r)} \right) \\ &\quad + \mathbf{e}_D \cdot 1 \cdot \sigma(0 + 1). \end{aligned}$$

It is easy to verify that for any $\mathbf{x}_t^{(1/2)}$, it holds that

$$\begin{aligned} \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) &= \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(0)}) \\ &= \sum_{r=1}^M \sum_{1 + \sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} \mathbf{e}_{D-M+r-1} a_k^{(1,r)} \sigma \left(\bar{\mathbf{b}}_k^{(1,r)\top} \mathbf{x}_t^{(0)} + c_k^{(1,r)} \right) + \mathbf{e}_D \cdot 1 \cdot \sigma(0 + 1) \\ &= \sum_{r=1}^M \sum_{1 + \sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} \mathbf{e}_{D-M+r-1} a_k^{(1,r)} \sigma \left(\mathbf{b}_k^{(1,r)\top} \mathbf{x}_t + c_k^{(1,r)} \right) + \mathbf{e}_D \cdot 1 \cdot \sigma(0 + 1) \\ &= \sum_{r=1}^M \mathbf{e}_{D-M+r-1} f_r^{2\text{NN}}(\mathbf{x}_t) + \mathbf{e}_D \\ &= (\mathbf{0}_d^\top, f_{(1,1)}^{2\text{NN}}(\mathbf{x}_t), \dots, f_{(1,M)}^{2\text{NN}}(\mathbf{x}_t), 1)^\top \in \mathbb{R}^D. \end{aligned}$$

Moreover, it satisfies that

$$\widetilde{\mathbf{FFN}^{(1)}}(\mathbf{x}_t^{(1/2)}) = \left[\mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right]$$

$$\begin{aligned}
&= (\mathbf{0}_d^\top, [f_{(1,1)}^{2\text{NN}}(\mathbf{x}_t)], \dots, [f_{(1,M)}^{2\text{NN}}(\mathbf{x}_t)], 1)^\top \\
&= (\mathbf{0}_d^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top \in \mathbb{R}^D.
\end{aligned}$$

Thus, we have achieved our goal in this step:

$$\mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + \widetilde{\text{FFN}}^{(1)}(\mathbf{x}_t^{(1/2)}) = (\mathbf{x}_t^\top, \mathbf{0}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top.$$

- **Case type = log.** Our main idea is that using the first FFN layer to express $(\mathbf{x}_t^\top, \mathbf{0}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top$ exactly.

First, we consider to approximate the r -th memory function $\log g_r(\mathbf{x})$ by standard FFN.

For any $r \in [M]$, by Lemma G.6, there exists a two-layer neural network with m_r neurons

$$f_{(1,r)}^{2\text{NN}}(\mathbf{x}) = \sum_{k=1}^{m_r} a_k^{(1,r)} \sigma(\mathbf{b}_k^{(1,r)\top} \mathbf{x} + c_k^{(1,r)})$$

defined on \mathbb{R}^d such that

$$\left\| \log g_r - f_{(1,r)}^{2\text{NN}} \right\|_{L^\infty([-1,1]^D)} \leq \tilde{\mathcal{O}}\left(\frac{\|\log g_r\|_{\mathcal{B}}}{\sqrt{m_r}}\right).$$

Therefore, if we choose

$$\tilde{\mathcal{O}}\left(\frac{\|\log g_r\|_{\mathcal{B}}}{\sqrt{m_r}}\right) < \frac{1}{4T_r},$$

the following holds:

$$\left| \log g_r(\mathbf{x}_t) - f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t) \right| \leq \left\| g_r - f_{(1,r)}^{2\text{NN}} \right\|_{L^\infty([-1,1]^d)} < \frac{1}{4T_r},$$

which ensures

$$\begin{aligned}
&\left| \exp\left(f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t)\right) - g_r(\mathbf{x}_t) \right| = \left| \exp\left(f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t)\right) - \exp(\log(g_r(\mathbf{x}_t))) \right| \\
&\leq \exp\left(\max\left\{f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t), \log(g_r(\mathbf{x}_t))\right\}\right) \left| f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t) - \log(g_r(\mathbf{x}_t)) \right| \\
&\leq \exp\left(\log g_r(\mathbf{x}_t) + \frac{1}{4}\right) \frac{1}{4T_r} \\
&\leq e^{1/4} \cdot T_r \cdot \frac{1}{4T_r} < \frac{1}{2}.
\end{aligned}$$

Noticing $g_r(\mathbf{x}_t) \in \mathbb{N}_+$, we have $\left[\exp\left(f_{(1,r)}^{2\text{NN}}(\mathbf{x}_t)\right)\right] = g_r(\mathbf{x}_t)$, which implies:

$$\widetilde{f_{(1,r)}^{2\text{NN}}}(\mathbf{x}_t) = \log\left[\exp\left(f_{(1,r)}^{2\text{NN}}\right)\right] = \log g_r(\mathbf{x}_t).$$

Consequently, in order to construct the form $(\mathbf{0}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top$, we need to arrange the parameters $a_k^{(1,r)}$, $\mathbf{b}_k^{(1,r)}$, and $c_k^{(1,r)}$ ($k \in [m_r], r \in [M]$) appropriately.

Denote $\bar{\mathbf{b}}_k^{(1,r)} = (\mathbf{b}_k^{(1,r)\top}, \mathbf{0}^\top)^\top \in \mathbb{R}^D$ for $k \in [m_r], r \in [M]$. Consider the following two-layer neural network with $1 + \sum_{r=1}^M m_r$ neurons defined on \mathbb{R}^D :

$$\begin{aligned}
\text{FFN}^{(1)}(\mathbf{x}) &= \sum_{r=1}^M \sum_{1 + \sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} e_{D-M+r-1} a_k^{(1,r)} \sigma\left(\bar{\mathbf{b}}_k^{(1,r)\top} \mathbf{x} + c_k^{(1,r)}\right) \\
&\quad + e_D \cdot 1 \cdot \sigma(0 + \log 2).
\end{aligned}$$

It is easy to verify that for any $\mathbf{x}_t^{(1/2)}$, it holds that

$$\begin{aligned}
& \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) = \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(0)}) \\
&= \sum_{r=1}^M \sum_{1+\sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} \mathbf{e}_{D-M+r-1} a_k^{(1,r)} \sigma \left(\bar{\mathbf{b}}_k^{(1,r) \top} \mathbf{x}_t^{(0)} + c_k^{(1,r)} \right) + \mathbf{e}_D \cdot 1 \cdot \sigma(0 + \log 2) \\
&= \sum_{r=1}^M \sum_{1+\sum_{j=0}^{r-1} m_j \leq k \leq \sum_{j=0}^r m_j} \mathbf{e}_{D-M+r-1} a_k^{(1,r)} \sigma \left(\mathbf{b}_k^{(1,r) \top} \mathbf{x}_t + c_k^{(1,r)} \right) + \mathbf{e}_D \cdot 1 \cdot \sigma(0 + \log 2) \\
&= \sum_{r=1}^M \mathbf{e}_{D-M+r-1} f_r^{2\text{NN}}(\mathbf{x}_t) + \mathbf{e}_D \log 2 \\
&= (\mathbf{0}_d^\top, f_{(1,1)}^{2\text{NN}}(\mathbf{x}_t), \dots, f_{(1,M)}^{2\text{NN}}(\mathbf{x}_t), \log 2)^\top \in \mathbb{R}^D.
\end{aligned}$$

Moreover, it satisfies that

$$\begin{aligned}
& \widetilde{\mathbf{FFN}}^{(1)}(\mathbf{x}_t^{(1/2)}) = \log \left[\exp \left(\mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right) \right] \\
&= (\mathbf{0}_d^\top, \log \left[\exp \left(f_{(1,1)}^{2\text{NN}}(\mathbf{x}_t) \right) \right], \dots, \log \left[\exp \left(f_{(1,M)}^{2\text{NN}}(\mathbf{x}_t) \right) \right], \log 2, \mathbf{0}^\top)^\top \\
&= (\mathbf{0}_d^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top.
\end{aligned}$$

Thus, we have achieved our goal in this step:

$$\mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + \widetilde{\mathbf{FFN}}^{(1)}(\mathbf{x}_t^{(1/2)}) = (\mathbf{x}_t^\top, \mathbf{0}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top.$$

As established in the proof above, the width m must satisfy:

$$m \geq 1 + \sum_{r=1}^M m_r = \begin{cases} \tilde{\Omega} \left(\sum_{r=1}^M \|g_r\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\sum_{r=1}^M \|\log g_r\|_{\mathcal{B}}^2 T_r^2 \right), & \text{type} = \text{log} \end{cases}.$$

Step III. Extract the adaptive memories by the second Attn layer.

We consider to use H_k attention heads (from $\sum_{i=1}^{k-1} H_i + 1$ -th head to $\sum_{i=1}^k H_i$ -th head) to extract it, and it satisfies to $\sum_{k=1}^M H_k = H$.

For simplicity, we denote the following projection matrices in $\mathbb{R}^{D \times D}$:

$$\begin{aligned}
& \mathbf{P}^{(k)} := (\mathbf{0}_{d \times kd} \quad \mathbf{I}_{d \times d} \quad \mathbf{0}) \in \mathbb{R}^{d \times D}, \quad 1 \leq k \leq M; \\
& \mathbf{P}_{\perp}^{(k)} := \begin{pmatrix} \mathbf{I}_{kd \times kd} & \mathbf{0}_{d \times d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{d \times d} & \mathbf{I}_{(D-(k+1)d) \times (D-(k+1)d)} \end{pmatrix} \in \mathbb{R}^{(D-d) \times D}, \quad 1 \leq k \leq M; \\
& \mathbf{Q}^{(M)} := (\mathbf{I}_{(M+1)d \times (M+1)d} \quad \mathbf{0}) \in \mathbb{R}^{(M+1)d \times D}.
\end{aligned}$$

Now we consider the extraction of k -th adaptive memory $\mathbf{x}_{t-g_k(\mathbf{x}_t)}$ ($1 \leq k \leq M$).

- **Case type = lin.**

By Lemma F.2, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and a function

$$\begin{aligned}
\phi_k^{\text{exp}}(t; B) &= \sum_{\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h \exp(-\beta_h(t - B)) \\
&= \sum_{\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h \exp\left(\beta_h B - \beta_h t\right)
\end{aligned}$$

such that $\beta_h > 0$ and

$$\sup_{1 \leq B \leq T_k} \|\mathbb{I}\{\cdot = B\} - \phi_k^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T_k}}{H_k^n}.$$

Moreover, Noticing that $1 \leq g_k(\mathbf{x}_t) \leq T_k$ holds for any $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \in \mathcal{X}$, the following holds:

$$\begin{aligned} & \sup_{\mathbf{X}} \|\mathbb{I}\{\cdot = g_k(\mathbf{x}_t)\} - \phi_k^{\text{exp}}(\cdot; g_k(\mathbf{x}_t))\|_{\ell_1(\mathbb{N})} \\ & \leq \sup_{1 \leq B \leq T_k} \|\mathbb{I}\{\cdot = B\} - \phi_k^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T_k}}{H_k^n}. \end{aligned}$$

Therefore, for these attention heads $(\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i)$, we can choose:

$$\begin{aligned} p^{(2,h)} &= \beta_h, \quad \mathbf{W}_O^{(1)} = \mathbf{I}_{D \times D}, \quad \mathbf{W}_V^{(2,h)} = \alpha_h \delta_{(k+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}, \\ \mathbf{W}_Q^{(2,h)} &= \sqrt{\beta_h} \delta_{(D-M+k-1,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times D}, \quad \mathbf{W}_K^{(2,h)} = \sqrt{\beta_h} \delta_{(D,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times D}, \end{aligned}$$

where $\delta_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

Then it is easy to verify:

$$\left\langle \mathbf{W}_Q^{(2,h)} \mathbf{x}_t^{(1)}, \mathbf{W}_K^{(2,h)} \mathbf{x}_{t-s}^{(1)} \right\rangle + p^{(2,h)} \phi_{\text{lin}}(s) = -\beta_h (s - g_k(\mathbf{x}_t)),$$

which implies:

$$\sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(1)}) = \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h(s - g_k(\mathbf{x}_t))} \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-s} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^D,$$

Then it holds that:

$$\begin{aligned} \mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(0)}) &= \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h(s - g_k(\mathbf{x}_t))} \mathbf{x}_{t-s}, \\ \mathbf{P}_{\perp}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(0)}) &= \mathbf{0}, \end{aligned}$$

moreover, the following estimate holds:

$$\begin{aligned} & \left\| \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \mathbf{x}_{t-g_k(\mathbf{x}_t)} \right\|_2 \\ &= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h(s - g_k(\mathbf{x}_t))} \mathbf{x}_{t-s} - \mathbf{x}_{t-g_k(\mathbf{x}_t)} \right\|_2 \end{aligned}$$

$$\begin{aligned}
&= \left\| \sum_{s=0}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h(s-g_k(\mathbf{x}_t))} - \mathbb{I}\{s = g_k(\mathbf{x}_t)\} \right) \mathbf{x}_{t-s} \right\|_2 \\
&\leq \sum_{s=0}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h(s-g_k(\mathbf{x}_t))} - \mathbb{I}\{s = g_k(\mathbf{x}_t)\} \right| \\
&= \|\phi_k^{\text{exp}}(\cdot; g_k(\mathbf{x}_t)) - \mathbb{I}\{\cdot = g_k(\mathbf{x}_t)\}\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T_k}}{H_k^n}.
\end{aligned}$$

- **Case type = log.**

By Lemma F.5, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and a function

$$\begin{aligned}
\phi_k^{\text{poly}}(t; B) &= \sum_{\sum_{i=1}^{k-1} H_i+1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h (t/B)^{-\beta_h} \\
&= \sum_{\sum_{i=1}^{k-1} H_i+1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h \exp(\beta_h \log B - \beta_h \log t)
\end{aligned}$$

such that $\beta_h > 1$ and

$$\sup_{1 \leq B \leq T_k} \left\| \mathbb{I}\{\cdot = B\} - \phi_k^{\text{poly}}(\cdot; B) \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T_k^{1.01(n+1)}}{H_k^n}.$$

Moreover, Noticing that $1 \leq g_k(\mathbf{x}_t) \leq T_k$ holds for any $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \in \mathcal{X}$, the following holds:

$$\begin{aligned}
&\sup_{\mathbf{X}} \left\| \mathbb{I}\{\cdot = g_k(\mathbf{x}_t)\} - \phi_k^{\text{poly}}(\cdot; g_k(\mathbf{x}_t)) \right\|_{\ell_1(\mathbb{N}_+)} \\
&\leq \sup_{1 \leq B \leq T_k} \left\| \mathbb{I}\{\cdot = B\} - \phi_k^{\text{poly}}(\cdot; B) \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T_k^{1.01(n+1)}}{H_k^n}.
\end{aligned}$$

Therefore, for these attention heads $(\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i)$, we can choose:

$$\begin{aligned}
p^{(2,h)} &= \beta_h, \quad \mathbf{W}_O^{(1)} = \mathbf{I}_{D \times D}, \quad \mathbf{W}_V^{(2,h)} = \alpha_h \delta_{(k+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}, \\
\mathbf{W}_Q^{(2,h)} &= \sqrt{\beta_h} \delta_{(D-M+k-1,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times D}, \quad \mathbf{W}_K^{(2,h)} = \frac{\sqrt{\beta_h}}{\log 2} \delta_{(D,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times D},
\end{aligned}$$

where $\delta_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

Then it is easy to verify:

$$\left\langle \mathbf{W}_Q^{(2,h)} \mathbf{x}_t^{(1)}, \mathbf{W}_K^{(2,h)} \mathbf{x}_{t-s}^{(1)} \right\rangle + p^{(2,h)} \phi_{\log}(s) = -\beta_h \log(s/g_k(\mathbf{x}_t)),$$

which implies:

$$\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(2,h)}(\mathbf{X}^{(1)}) = \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} (s/g_k(\mathbf{x}_t))^{-\beta_h} \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-s} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^D,$$

Then it holds that:

$$\begin{aligned} \mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(0)}) &= \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} (s/g_k(\mathbf{x}_t))^{-\beta h} \mathbf{x}_{t-s}, \\ \mathbf{P}_{\perp}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(0)}) &= \mathbf{0}, \end{aligned}$$

moreover, the following estimate holds:

$$\begin{aligned} & \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \mathbf{x}_{t-g_k(\mathbf{x}_t)} \right\|_2 \\ &= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} (s/g_k(\mathbf{x}_t))^{-\beta h} \mathbf{x}_{t-s} - \mathbf{x}_{t-g_k(\mathbf{x}_t)} \right\|_2 \\ &= \left\| \sum_{s=0}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h (s/g_k(\mathbf{x}_t))^{-\beta h} - \mathbb{I}\{s = g_k(\mathbf{x}_t)\} \right) \mathbf{x}_{t-s} \right\|_2 \\ &\leq \sum_{s=0}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h (s/g_k(\mathbf{x}_t))^{-\beta h} - \mathbb{I}\{s = g_k(\mathbf{x}_t)\} \right| \\ &= \left\| \phi_k^{\text{poly}}(\cdot; g_k(\mathbf{x}_t)) - \mathbb{I}\{\cdot = g_k(\mathbf{x}_t)\} \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T_k^{1.01(n+1)}}{H_k^n}. \end{aligned}$$

Then we combine the estimate for all $k \in [M]$ for these two cases. It holds that

$$\begin{aligned} & \left\| \mathbf{Q}^{(M)} \mathbf{x}_t^{(3/2)} - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-g_1(\mathbf{x}_t)} \\ \vdots \\ \mathbf{x}_{t-g_M(\mathbf{x}_t)} \end{pmatrix} \right\|_2 \\ &= \left\| \begin{pmatrix} \mathbf{x}_t \\ \mathbf{0}_{Md} \end{pmatrix} + \sum_{h=1}^M \mathbf{Q}^{(M)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}^{(1)}) - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-g_1(\mathbf{x}_t)} \\ \vdots \\ \mathbf{x}_{t-g_M(\mathbf{x}_t)} \end{pmatrix} \right\|_2 \\ &= \left\| \sum_{h=1}^M \mathbf{Q}^{(M)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_d \\ \mathbf{x}_{t-g_1(\mathbf{x}_t)} \\ \vdots \\ \mathbf{x}_{t-g_M(\mathbf{x}_t)} \end{pmatrix} \right\|_2 \\ &= \left\| \sum_{k=1}^M \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Q}^{(M)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-g_k(\mathbf{x}_t)} \\ \mathbf{0} \end{pmatrix} \right) \right\|_2 \\ &\leq \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Q}^{(M)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{kd} \\ \mathbf{x}_{t-g_k(\mathbf{x}_t)} \\ \mathbf{0} \end{pmatrix} \right\|_2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \mathbf{Attn}_t^{(2,h)}(\mathbf{X}) - \mathbf{x}_{t-g_k(\mathbf{x}_t)} \right\|_2 \\
&\leq \mathcal{E}_{\text{Attn}}(\text{type}) := \begin{cases} \frac{C(n)e^{0.01(n+1)T_k}}{H_k^n}, & \text{type} = \text{lin} \\ \frac{C(n)T_k^{1.01(n+1)}}{H_k^n}, & \text{type} = \text{log} \end{cases}.
\end{aligned}$$

Similar to the proof of Theorem B.1, we choose the head number:

$$\begin{aligned}
H_k &= \frac{e^{0.01T_k}}{\sum_{j=1}^M e^{0.01T_j}} H, \quad k \in [M], \quad \text{type} = \text{lin}; \\
H_k &= \frac{T_k^{1.01}}{\sum_{j=1}^M T_j^{1.01}} H, \quad k \in [M], \quad \text{type} = \text{log}.
\end{aligned}$$

Thus, we obtain the final bound in Step III:

$$\mathcal{E}_{\text{Attn}}^{\text{Soft}}(\text{type}) \leq \begin{cases} \frac{C(n)}{H^n} \left(\sum_{k=1}^M e^{0.01T_k} \right)^{n+1}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \left(\sum_{k=1}^M T_k^{1.01} \right)^{n+1}, & \text{type} = \text{log} \end{cases}.$$

Furthermore, by choosing $\mathcal{E}_{\text{Attn}}(\text{type}) \leq 1$, it holds that

$$\begin{aligned}
\left\| \mathbf{Q}^{(M)} \mathbf{x}_t^{(3/2)} \right\|_\infty &\leq \left\| \mathbf{Q}^{(M)} \mathbf{x}_t^{(3/2)} - \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-g_1(\mathbf{x}_t)} \\ \vdots \\ \mathbf{x}_{t-g_M(\mathbf{x}_t)} \end{pmatrix} \right\|_\infty + \left\| \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-g_1(\mathbf{x}_t)} \\ \vdots \\ \mathbf{x}_{t-g_M(\mathbf{x}_t)} \end{pmatrix} \right\|_\infty \\
&\leq \mathcal{E}_{\text{Attn}}(\text{type}) + 1 \leq 2.
\end{aligned}$$

Step IV. Approximate the nonlinear function by 2-nd FFN layer.

In this step, we aim to approximate the function f using two-layer network. By Lemma G.6, there exists a two-layer neural network with m neurons

$$f_{(2)}^{2\text{NN}}(\mathbf{x}) = \sum_{k=1}^m a_k^{(2)} \sigma \left(\mathbf{b}_k^{(2)\top} \mathbf{x} + c_k^{(2)} \right)$$

defined on $\mathbb{R}^{(M+1)d}$ such that

$$\left\| f - f_{(2)}^{2\text{NN}} \right\|_{L^\infty([-2,2]^{(M+1)d})} \leq \tilde{\mathcal{O}} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right).$$

Denote $\bar{\mathbf{b}}_k^{(2)} = (\mathbf{b}_k^{(2)\top}, \mathbf{0}^\top)^\top \in \mathbb{R}^D$ for $k \in [m]$. And we consider the following two-layer neural network with m neurons defined on \mathbb{R}^D :

$$\text{FFN}^{(2)}(\mathbf{x}) := \sum_{k=1}^m a_k^{(2)} \sigma \left(\bar{\mathbf{b}}_k^{(2)\top} \mathbf{x} + c_k^{(2)} \right).$$

It is easy to verify:

$$\text{FFN}^{(2)}\left(\mathbf{x}_t^{(3/2)}\right) = f_{(2)}^{2\text{NN}}\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right).$$

Moreover,

$$\mathcal{E}_{\text{FFN}}^{(2)} := \left\| f - \text{FFN}^{(2)} \right\|_{L^\infty([-2,2]^{(M+1)d})} \leq \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right).$$

The final bound.

For any t and $\|\mathbf{X}\| \in \mathcal{X}$, it holds that

$$\begin{aligned} & \left\| \mathbf{H}_t(\mathbf{X}) - \mathbf{x}_t^{(2)} \right\| = \left| f\left(\mathbf{x}_t, \mathbf{x}_{t-g_1(\mathbf{x}_t)}, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)}\right) - \text{FFN}^{(2)}\left(\mathbf{x}_t^{(3/2)}\right) \right| \\ &= \left| f\left(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}\right) - f_{(2)}^{2\text{NN}}\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) \right| \\ &= \left| f\left(\mathbf{x}_t, \mathbf{x}_{t-g_1(\mathbf{x}_t)}, \dots, \mathbf{x}_{t-g_M(\mathbf{x}_t)}\right) - f\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) \right. \\ & \quad \left. + f\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) - f_{(2)}^{2\text{NN}}\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) \right| \\ &\leq \left| f\left(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}\right) - f\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) \right| + \left| f\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) - f_{(2)}^{2\text{NN}}\left(\mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)}\right) \right| \\ &\leq \|f\|_{\text{Lip}} \left\| \left(\mathbf{x}_t^\top, \mathbf{x}_{t-t_1}^\top, \dots, \mathbf{x}_{t-t_M}^\top\right)^\top - \mathbf{Q}^{(M)}\mathbf{x}_t^{(3/2)} \right\|_2 + \left\| f - f_{(2)}^{2\text{NN}} \right\|_{L^\infty([-2,2]^{(M+1)D})} \\ &\leq \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}} + \mathcal{E}_{\text{FFN}}, \end{aligned}$$

where

$$\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right);$$

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{k=1}^M e^{0.01T_k}\right)^{n+1}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \left(\sum_{k=1}^M T_k^{1.01}\right)^{n+1}\right), & \text{type} = \text{log} \end{cases}.$$

Moreover, recalling our proof in Step II, we further need the hard condition:

$$m \geq \begin{cases} \tilde{\Omega}\left(\sum_{r=1}^M \|g_r\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin} \\ \tilde{\Omega}\left(\sum_{r=1}^M \|\log g_r\|_{\mathcal{B}}^2 T_r^2\right), & \text{type} = \text{log} \end{cases}.$$

Due to the arbitrariness of t and \mathbf{X} , the proof is completed. \square

Remark C.3. The **core step** in this proof is **Step III**, where the extraction of the memory functions is achieved through a *nice interaction between the temporal space (provided by RPE) and the token space (provided by DP)*. Specifically, the memory functions $g_i(\mathbf{x}_t)$ (in token space) are mapped into the temporal space s , resulting in the form of $\mathbf{x}_{s-g_i(\mathbf{x}_t)}$ for DP Attn with lin-RPE or $\log(s/g_i(\mathbf{x}_t))$ for DP Attn with log-RPE.

C.2 Proof of Proposition 4.2

For Proposition 16, we denote the following one-layer Attn hypothesis class:

$$\begin{aligned} \mathcal{ATTN}_{(1,H)}^{\text{type}} &:= \{\text{Attn} : \mathbf{TF} \text{ is a 1-layer, } H\text{-head (normalization-free) Attn with type-RPE}\}; \\ \mathcal{ATTN}_{(1,H)}^{\text{DPF,type}} &:= \{\mathbf{TF} : \text{Attn is a 1-layer, } H\text{-head DP-free Attn with type-RPE}\}. \end{aligned} \quad (16)$$

Proposition C.4 (The formal version of Proposition 4.2). *Consider 1-layer Attn. Then, there exists a target $\mathbf{H} \in \mathcal{H}_{(1,1)}^{\text{adap}}$ (10) such that:*

(A) (Attn with DP) For any $\epsilon > 0$, there exists a 1-layer Attn $\text{Attn}^{\text{DP}} \in \mathcal{ATTN}_{(1,H)}^{\text{type}}$ such that

$$\left\| \mathbf{H} - \text{Attn}^{\text{DP}} \right\| \leq \epsilon.$$

(B) (Attn without DP) For any 1-layer DP-free Attn $\text{Attn}^{\text{DPF}} \in \mathcal{ATTN}_{(1,H)}^{\text{DPF,type}}$, a uniform lower bound holds:

$$\left\| \mathbf{H} - \text{Attn}^{\text{DPF}} \right\| \geq \frac{2}{3}.$$

Proof of Proposition C.4.

Consider the following target function $\mathbf{H} \in \mathcal{H}_{1,1}^{\text{Adap}}$. Let the input sequence $\mathbf{X} \in \mathcal{X} = \{-1, 0, 1\}^{\mathbb{Z}}$, and we consider the target

$$y_t = \mathbf{H}_t(\mathbf{X}) := x_{t-g(x_t)},$$

where the adaptive memory is

$$g(x) = \begin{cases} 0, & x = -1 \\ 1, & x = 0 \\ 2, & x = 1 \end{cases}.$$

Part (A). The Efficiency of Attn with DP.

First, we choose the embedding dimension $D = 2$, and select simple embedding $\mathbf{W}_E = \mathbf{b}_E = (1, 0)^\top \in \mathbb{R}^{2 \times 1}$. Then for any input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}}$, the token after embedding satisfies:

$$\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = (x_t, 1)^\top.$$

We consider one-layer normalization-free Self-attention with ϕ_{exp} , which has the form:

$$\text{Attn}_t^{\text{DP}}(\mathbf{X}) = \mathbf{W}_O \sum_{h=1}^H \mathbf{W}_V^{(1,h)} \sum_{s=0}^{+\infty} \mathbf{x}_{t-s}^{(0)} \exp\left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t^{(0)}, \mathbf{W}_K^{(1,h)} \mathbf{x}_{t-s}^{(0)} \right\rangle + p^{(1,h)} s\right).$$

By Lemma F.2 (for $n = 1$), there exists a constant $C > 0$ and a function

$$\phi^{\text{exp}}(t; B) = \sum_{h=1}^H \alpha_h \exp(-\beta_h(t - B)) = \sum_{h=1}^H \alpha_h \exp\left(\beta_h B - \beta_h t\right)$$

such that $\beta_h > 0$ and

$$\sup_{0 \leq B \leq 2} \|\mathbb{I}\{\cdot = B\} - \phi^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C e^{0.01 \cdot 2 \cdot 2}}{H} = \mathcal{O}\left(\frac{1}{H}\right).$$

Moreover, Noticing that $0 \leq g(x_t) \leq 2$ holds for any $\mathbf{X} = (x_t)_{t \in \mathbb{Z}} \in \mathcal{X}$, the following holds:

$$\begin{aligned} & \sup_{\mathbf{X}} \|\mathbb{I}\{\cdot = g(x_t)\} - \phi^{\text{exp}}(\cdot; g(x_t))\|_{\ell_1(\mathbb{N})} \\ & \leq \sup_{0 \leq B \leq 2} \|\mathbb{I}\{\cdot = B\} - \phi^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \mathcal{O}\left(\frac{1}{H}\right). \end{aligned}$$

Therefore, for attention heads ($1 \leq h \leq H$), we can choose:

$$\begin{aligned} p^{(1,h)} &= \beta_h, \quad \mathbf{W}_O^{(1)} = (1, 0)^\top \in \mathbb{R}^{2 \times 1}, \quad \mathbf{W}_V^{(1,h)} = \alpha_h \boldsymbol{\delta}_{(1,1)}^{(1 \times 1)} \in \mathbb{R}^{2 \times 2}, \\ \mathbf{W}_Q^{(1,h)} &= \sqrt{\beta_h} (1, 1)^\top \in \mathbb{R}^{2 \times 1}, \quad \mathbf{W}_K^{(1,h)} = \sqrt{\beta_h} (0, 1)^\top \in \mathbb{R}^{2 \times 1}, \end{aligned}$$

where $\boldsymbol{\delta}_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

Then it is easy to verify:

$$\left\langle \mathbf{W}_Q^{(1,h)} \mathbf{x}_t^{(0)}, \mathbf{W}_K^{(1,h)} \mathbf{x}_{t-s}^{(0)} \right\rangle + p^{(2,h)} s = -p^{(1,h)} \left(s - (x_t + 1) \right) = -p^{(1,h)} \left(s - g(x_t) \right).$$

Thus, the following estimate holds:

$$\begin{aligned} & \left\| \text{Attn}_t^{\text{DP}}(\mathbf{X}) - x_{t-g(x_t)} \right\|_2 \\ &= \left\| \sum_{h=1}^H \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h (s-g(x_t))} x_{t-s} - x_{t-g(x_t)} \right\|_2 \\ &= \left\| \sum_{s=0}^{+\infty} \left(\sum_{h=1}^H \alpha_h e^{-\beta_h (s-g(x_t))} - \mathbb{I}\{s = g(x_t)\} \right) x_{t-s} \right\|_2 \\ &\leq \sum_{s=0}^{+\infty} \left| \sum_{h=1}^H \alpha_h e^{-\beta_h (s-g(x_t))} - \mathbb{I}\{s = g(x_t)\} \right| \\ &= \|\phi^{\text{exp}}(\cdot; g(x_t)) - \mathbb{I}\{\cdot = g(x_t)\}\|_{\ell_1(\mathbb{N})} \leq \mathcal{O}\left(\frac{1}{H}\right). \end{aligned}$$

Due to the arbitrariness of t and \mathbf{X} , the proof is completed: for any $\epsilon > 0$, we only need to use $H = \Omega(1/\epsilon)$ heads to approximate it.

Part (B). The Hardness Result of Attn without DP.

We consider one-layer Dot-product-free Self-attention with ϕ_{exp} or ϕ_{log} . For any input \mathbf{X} , the corresponding output can be written as:

$$\text{Attn}_t^{\text{DPF}}(\mathbf{X}) = \sum_{s=0}^{+\infty} \rho_s (W_E x_{t-s} + b_E).$$

For simplicity, we denote:

$$x_t^{(0)} = W_E x_t + b_E,$$

then we have the following estimate:

$$\| \mathbf{H} - \text{Attn}^{\text{DPF}} \| = \sup_t \sup_{\mathbf{X}} |\mathbf{H}_t^{\text{DPF}}(\mathbf{X}) - \text{Attn}_t(\mathbf{X})| \geq \sup_{\mathbf{X}} |\mathbf{H}_0(\mathbf{X}) - \text{Attn}_0^{\text{DPF}}(\mathbf{X})|$$

$$\begin{aligned}
&= \sup_{(\dots, x_{-2}, x_{-1}, x_0)} \left| x_{-g(x_0)} - \sum_{s=0}^{+\infty} \rho_s x_{-s}^{(0)} \right| \\
&\stackrel{\text{fixing } x_{-s} = 0 \text{ for } s \geq 3}{\geq} \sup_{(x_{-2}, x_{-1}, x_0)} \left| x_{-g(x_0)} - \sum_{s=0}^2 \rho_s x_{-s}^{(0)} \right| \\
&\stackrel{x_0 \in \{-1, 0, 1\}}{\geq} \max \left\{ \begin{aligned} &\max_{(x_{-2}, x_{-1})} \left| -1 - \left(\rho_0(-W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right|, \\ &\max_{(x_{-2}, x_{-1})} \left| x_{-g(0)} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right|, \\ &\max_{(x_{-2}, x_{-1})} \left| x_{-g(1)} - \left(\rho_0(W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \end{aligned} \right\} \\
&= \max_{(x_{-2}, x_{-1})} \max \left\{ \begin{aligned} &\left| x_{-1} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right|, \\ &\left| -1 - \left(\rho_0(-W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right|, \\ &\left| x_{-2} - \left(\rho_0(W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \end{aligned} \right\} \\
&\stackrel{\max\{a,b,c\} \geq \frac{1}{3}(a+b+c)}{\geq} \frac{1}{3} \max_{(x_{-2}, x_{-1})} \left(\begin{aligned} &\left| x_{-1} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \\ &+ \left| -1 - \left(\rho_0(-W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \\ &+ \left| x_{-2} - \left(\rho_0(W_E + b_E) + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \end{aligned} \right) \\
&\stackrel{|a|+|b| \geq |a+b|}{\geq} \frac{1}{3} \max_{(x_{-2}, x_{-1})} \left(\begin{aligned} &\left| x_{-1} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \\ &+ \left| -1 + x_{-2} - 2 \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \end{aligned} \right) \\
&\geq \frac{1}{3} \max_{(x_{-2}, x_{-1})} \left(\begin{aligned} &\left| x_{-1} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \\ &+ \left| \frac{-1 + x_{-2}}{2} - \left(\rho_0 b_E + \rho_1 x_{-1}^{(0)} + \rho_2 x_{-2}^{(0)} \right) \right| \end{aligned} \right) \\
&\stackrel{|a|+|b| \geq |a-b|}{\geq} \frac{1}{3} \max_{(x_{-2}, x_{-1})} \left| x_{-1} + \frac{1}{2} - \frac{x_{-2}}{2} \right| = \frac{2}{3}.
\end{aligned}$$

□

C.3 Proof of Proposition 4.3

In this subsection, we propose a structurally simpler yet effective alternative to traditional Dot-product in Self-attention. This alternative is proposed based on our insights into the role of Attn in facilitating interaction between the temporal-space and the token-space. Specifically, we propose a more direct structure to achieve the interaction $\phi_{\text{type}}(s) - \phi_{\text{type}}(\mathbf{w}^\top \mathbf{x}_t)$.

Definition C.5 (TMX Transformer). We define the TMX (“ t minus \mathbf{x} ”) Transformer as follows. In standard Transformer (2), we modify the term

$$\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-s} \right\rangle + p^{(l,h)} \phi_{\text{type}}(s)$$

in Multi-head Dot-product Self-attention to the new formulation:

$$p^{(l,h)} \left(\phi_{\text{type}}(s) - \phi_{\text{type}}(\mathbf{w}^{(l,h)\top} \mathbf{x}_t) \right),$$

where the parameters $\mathbf{w}^{(l,h)} \in \mathbb{R}^{D \times 1}$.

Notice that in TMX Transformer, the revised term requires only $\mathcal{O}(D)$ parameters, much less than $\mathcal{O}(D^2)$ in standard Dot-product Transformer.

Consequently, we define the following TMX Transformer hypothesis class.

$$\mathcal{TF}_{(1,H,m)}^{\text{TMX,type}} := \left\{ \mathbf{TF} : \mathbf{TF} \text{ is a 1-layer, } H\text{-head, } m\text{-width} \right. \\ \left. (\text{normalization-free}) \text{ TMX Transformer with type-RPE} \right\}. \quad (17)$$

Proposition C.6 (The formal version of Proposition 4.3). *Under the same conditions in Theorem 4.1, there exists a two-layer TMX Transformer $\mathbf{TF} \in \mathcal{TF}_{(2,H,m)}^{\text{TMX,type}}$ (17) such that it can achieve the same approximation rate as standard Transformer presented in Theorem 4.1.*

Proof of Proposition C.6.

It is worth noting that TMX Transformer only replaces $\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-s} \right\rangle + p^{(l,h)} \phi_{\text{type}}(s)$ in standard Transformer with $p^{(l,h)} \left(\phi_{\text{type}}(s) - \phi_{\text{type}}(\mathbf{w}^{(l,h) \top} \mathbf{x}_t) \right)$. Therefore, the proof is highly similar to that of Theorem C.2. We only need to prove that TMX Attn can also achieve **Step I** and **Step III** in the proof of Theorem C.2.

Step I. Step I is trivial due to the same use of the residual block.

Step III. Extract the adaptive memories by the second Attn layer.

We still consider to use H_k attention heads (from $\sum_{i=1}^{k-1} H_i + 1$ -th head to $\sum_{i=1}^k H_i$ -th head) to extract it, and it satisfies to $\sum_{k=1}^M H_k = H$.

Now we consider the extraction of k -th adaptive memory $\mathbf{x}_{t-g_k(\mathbf{x}_t)}$ ($1 \leq k \leq M$).

- **Case type = lin.**

For the proof of standard Transformer (the proof of Theorem C.2), for the attention heads ($\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i$), we can construct specific $p^{(2,h)}$, $\mathbf{W}_Q^{(2,h)}$, $\mathbf{W}_K^{(2,h)}$, $\mathbf{W}_V^{(2,h)}$ such that

$$\left\langle \mathbf{W}_Q^{(2,h)} \mathbf{x}_t^{(1)}, \mathbf{W}_K^{(2,h)} \mathbf{x}_{t-s}^{(1)} \right\rangle + p^{(2,h)} \phi_{\text{lin}}(s) = -\beta_h (s - g_k(\mathbf{x}_t)).$$

In this proof, we only need to prove that we can also construct specific $\mathbf{w}^{(l,h)}$, $\mathbf{W}_V^{(2,h)}$ such that

$$p^{(2,h)} \left(\phi_{\text{lin}}(s) - \phi_{\text{lin}}(\mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)}) \right) = -\beta_h (s - g_k(\mathbf{x}_t)).$$

Recalling the proof of Theorem C.2,

$$\mathbf{x}_t^{(1)} = (\mathbf{x}_t^\top, \mathbf{0}^\top, g_1(\mathbf{x}_t), \dots, g_M(\mathbf{x}_t), 1)^\top \in \mathbb{R}^D.$$

Therefore, we can choose

$$p^{(2,h)} = \beta_h, \quad \mathbf{w}^{(2,h)} = \boldsymbol{\delta}_{(D-M+k-1,1)}^{(1 \times 1)} \in \mathbb{R}^D, \quad \mathbf{W}_V^{(2,h)} = \alpha_h \boldsymbol{\delta}_{(k+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}.$$

where $\boldsymbol{\delta}_{(p_1,p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

The the following holds:

$$p^{(2,h)} \left(\phi_{\text{lin}}(s) - \phi_{\text{lin}}(\mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)}) \right)$$

$$= -p^{(2,h)} \left(s - \mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)} \right) = -\beta_h \left(s - g_k(\mathbf{x}_t) \right).$$

- **Case type = log.**

For the proof of standard Transformer (the proof of Theorem C.2), for the attention heads $(\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i)$, we can construct specific $p^{(2,h)}$, $\mathbf{W}_Q^{(2,h)}$, $\mathbf{W}_K^{(2,h)}$, $\mathbf{W}_V^{(2,h)}$ such that

$$\left\langle \mathbf{W}_Q^{(2,h)} \mathbf{x}_t^{(1)}, \mathbf{W}_K^{(2,h)} \mathbf{x}_{t-s}^{(1)} \right\rangle + p^{(2,h)} \phi_{\log}(s) = -\beta_h \log \left(s/g_k(\mathbf{x}_t) \right).$$

In this proof, we only need to prove that we can also construct specific $\mathbf{w}^{(l,h)}$, $\mathbf{W}_V^{(2,h)}$ such that

$$p^{(2,h)} \left(\phi_{\log}(s) - \phi_{\log} \left(\mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)} \right) \right) = -\beta_h \log \left(s/g_k(\mathbf{x}_t) \right).$$

Recalling the proof of Theorem C.2,

$$\mathbf{x}_t^{(1)} = (\mathbf{x}_t^\top, \mathbf{0}^\top, \log g_1(\mathbf{x}_t), \dots, \log g_M(\mathbf{x}_t), \log 2)^\top.$$

Therefore, we can choose

$$p^{(2,h)} = \beta_h, \quad \mathbf{w}^{(2,h)} = \delta_{(D-M+k-1,1)}^{(1 \times 1)} \in \mathbb{R}^D, \quad \mathbf{W}_V^{(2,h)} = \alpha_h \delta_{(k+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}.$$

where $\delta_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

The following holds:

$$\begin{aligned} & p^{(2,h)} \left(\phi_{\log}(s) - \phi_{\log} \left(\mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)} \right) \right) \\ &= -p^{(2,h)} \log \left(s / \left(\mathbf{w}^{(2,h) \top} \mathbf{x}_t^{(1)} \right) \right) = -\beta_h \log \left(s/g_k(\mathbf{x}_t) \right). \end{aligned}$$

The rest of the proof is exactly the same as the proof of Theorem C.2, and we do not repeat it.

□

D Proof of Section 4.3

D.1 Proof of Theorem 4.4

In this subsection, we give the detailed proofs for the general case of K -adaptive, long but M -sparse memory:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M}),$$

where the adaptive memories satisfy:

$$\begin{aligned} t_1 &= g_1(\mathbf{x}_t); \\ t_2 &= g_2(\mathbf{x}_t, \mathbf{x}_{t-t_1}); \\ &\dots \\ t_{K+1} &= g_{K+1}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_K}); \\ &\dots \\ t_{K+2} &= g_{K+2}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_K}); \\ &\dots \\ t_M &= g_{K+1}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_K}), \end{aligned}$$

where $1 \leq t_k \leq T_k$ holds for any $k \in [M]$.

Theorem D.1 (Restatement of Theorem 4.4). *For any target $\mathbf{H} \in \mathcal{H}_{(K,M)}^{\text{Adap}}$, rate $n \in \mathbb{N}_+$, and $H, m \in \mathbb{N}_+$, there exists an L -layer ($L = K + 1 + \mathbb{I}\{M \geq K + 1\}$) Transformer $\mathbf{TF} \in \mathcal{TF}_{(L,H,m)}^{\text{NF,type}}$ (12) and a constant $C(n)$ such that: if the width satisfies*

$$m \geq \begin{cases} \tilde{\Omega}\left(\max_{i \in [K]} \vee \sum_{i=K+1}^M \|g_i\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin}, \\ \tilde{\Omega}\left(\max_{i \in [K]} \vee \sum_{i=K+1}^M \|\log g_i\|_{\mathcal{B}}^2 T_i^2\right), & \text{type} = \text{log} \end{cases},$$

then the following approximation rate holds:

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \mathcal{E}_{\text{Attn}}(\text{type}),$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right)$ and

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l} + \left(\sum_{l=K+1}^M e^{0.01T_l}\right)^{2n+2}}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)} + \left(\sum_{l=K+1}^M T_l^{1.01}\right)^{2n+2}}\right), & \text{type} = \text{log} \end{cases}.$$

Proof of Theorem D.1.

First, we choose the embedding dimension $D = (M + 1)(d + 1)$, and select the same embedding matrix $\mathbf{W}_E = (\mathbf{I}_d, \mathbf{0})^\top \in \mathbb{R}^{D \times d}$, $\mathbf{b}_E = \mathbf{0} \in \mathbb{R}^D$ as the proof of Theorem C.2. Moreover, we still use the network with precision $\widetilde{\text{FFN}}$ defined in Appendix C.1 to tackle the discrete values of memories.

Then for any input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}}$, the token after embedding satisfies:

$$\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = (\mathbf{x}_t^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^D.$$

Thus, for L -layer ($L = K + 1 + \mathbb{I}\{M \geq K + 1\}$) normalization-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(L,H,m)}^{\text{NF,type}}$ with ϕ_{type} , the output token $\mathbf{x}_t^{(K+1)}$ of t -th input token satisfies:

$$\mathbf{x}_t^{(l-1/2)} = \mathbf{x}_t^{(l)} + \mathbf{W}_O^{(l)} \sum_{h=1}^H \text{Attn}_t^{(l,h)}(\mathbf{X}^{(l-1)}), \quad 1 \leq l \leq L + 1$$

$$\begin{aligned}\mathbf{x}_t^{(l)} &= \mathbf{x}_t^{(l-1/2)} + \widetilde{\mathbf{FFN}}^{(l)}(\mathbf{x}_t^{(l-1/2)}), \quad 1 \leq l \leq L \\ \mathbf{x}_t^{(L+1)} &= \mathbf{FFN}^{(L+1)}(\mathbf{x}_t^{(L+1/2)}),\end{aligned}$$

where

$$\text{Attn}_t^{(l,h)}(\mathbf{X}) = \mathbf{W}_V^{(l,h)} \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \exp\left(\left\langle \mathbf{W}_Q^{(l,h)} \mathbf{x}_t, \mathbf{W}_K^{(l,h)} \mathbf{x}_{t-s} \right\rangle + p^{(l,h)} \phi_{\text{type}}(s)\right).$$

Since the proof of this theorem is similar to the proof of Theorem C.2, we mainly discuss the differences.

The proof can be summarized as the following process:

- **Case type = lin.**

- **Regime $M \geq K + 1$.**

$$\begin{aligned}& \mathbf{x}_t^{(0)} \\ \text{Step 1. 1-st Attn} & \downarrow \\ & \mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)} \\ \text{Step 2. 1-st FFN} & \downarrow \\ & \mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + (\mathbf{0}^\top, t_1, \mathbf{0}_{M-1}^\top, 1)^\top \\ \text{Step 3. 2-st Attn} & \downarrow \\ & \mathbf{x}_t^{(3/2)} \approx \mathbf{x}_t^{(1)} + (\mathbf{0}_d^\top, \mathbf{x}_{t-t_1}^\top, \mathbf{0}^\top)^\top \\ \text{Step 4. 2-st FFN} & \downarrow \\ & \mathbf{x}_t^{(2)} = \mathbf{x}_t^{(3/2)} + (\mathbf{0}^\top, t_2, \mathbf{0}_{M-1}^\top)^\top \\ \text{Step 5. 3-st Attn} & \downarrow \\ & \mathbf{x}_t^{(5/2)} \approx \mathbf{x}_t^{(2)} + (\mathbf{0}_{2d}^\top, \mathbf{x}_{t-t_2}^\top, \mathbf{0}^\top)^\top \\ & \dots \\ \text{Step } 2K + 1. K + 1\text{-st Attn} & \downarrow \\ & \mathbf{x}_t^{(K+1/2)} \approx \mathbf{x}_t^{(K)} + (\mathbf{0}_{Kd}^\top, \mathbf{x}_{t-t_K}^\top, \mathbf{0}^\top)^\top \\ \text{Step } 2K + 2. K + 1\text{-st FFN} & \downarrow \\ & \mathbf{x}_t^{(K+1)} = \mathbf{x}_t^{(K+1/2)} + (\mathbf{0}^\top, t_{K+1}, \dots, t_M, 0)^\top \\ \text{Step } 2K + 3. K + 2\text{-st Attn} & \downarrow \\ & \mathbf{x}_t^{(K+3/2)} \approx \mathbf{x}_t^{(K+1)} + (\mathbf{0}_{(K+1)d}^\top, \mathbf{x}_{t-t_{K+1}}^\top, \dots, \mathbf{x}_{t-t_M}^\top, 0)^\top \\ \text{Step } 2K + 4. K + 2\text{-st FFN} & \downarrow \\ & \mathbf{x}_t^{(K+2)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M})\end{aligned}$$

- **Regime $M = K$.**

$$\begin{aligned}& \mathbf{x}_t^{(0)} \\ \text{Step 1. 1-st Attn} & \downarrow \\ & \mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)} \\ \text{Step 2. 1-st FFN} & \downarrow \\ & \mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + (\mathbf{0}^\top, t_1, \mathbf{0}_{M-1}^\top, 1)^\top \\ \text{Step 3. 2-st Attn} & \downarrow \\ & \mathbf{x}_t^{(3/2)} \approx \mathbf{x}_t^{(1)} + (\mathbf{0}_d^\top, \mathbf{x}_{t-t_1}^\top, \mathbf{0}^\top)^\top\end{aligned}$$

Step 4. 2-st FFN ↓

$$\mathbf{x}_t^{(2)} = \mathbf{x}_t^{(3/2)} + (\mathbf{0}^\top, t_2, \mathbf{0}_{M-1}^\top)^\top$$

Step 5. 3-st Attn ↓

$$\mathbf{x}_t^{(5/2)} \approx \mathbf{x}_t^{(2)} + (\mathbf{0}_{2d}^\top, \mathbf{x}_{t-t_2}^\top, \mathbf{0}^\top)^\top$$

...

Step $2K + 1$. $K + 1$ -st Attn ↓

$$\mathbf{x}_t^{(K+1/2)} \approx \mathbf{x}_t^{(K)} + (\mathbf{0}_{Kd}^\top, \mathbf{x}_{t-t_K}^\top, \mathbf{0}^\top)^\top$$

Step $2K + 2$. $K + 1$ -st FFN ↓

$$\mathbf{x}_t^{(K+1)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M})$$

• **Case type = log.**

– **Regime $M \geq K + 1$.**

$$\mathbf{x}_t^{(0)}$$

Step 1. 1-st Attn ↓

$$\mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)}$$

Step 2. 1-st FFN ↓

$$\mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + (\mathbf{0}^\top, \log t_1, \mathbf{0}_{M-1}^\top, \log 2)^\top$$

Step 3. 2-st Attn ↓

$$\mathbf{x}_t^{(3/2)} \approx \mathbf{x}_t^{(1)} + (\mathbf{0}_d^\top, \mathbf{x}_{t-t_1}^\top, \mathbf{0}^\top)^\top$$

Step 4. 2-st FFN ↓

$$\mathbf{x}_t^{(2)} = \mathbf{x}_t^{(3/2)} + (\mathbf{0}^\top, \log t_2, \mathbf{0}_{M-1}^\top)^\top$$

Step 5. 3-st Attn ↓

$$\mathbf{x}_t^{(5/2)} \approx \mathbf{x}_t^{(2)} + (\mathbf{0}_{2d}^\top, \mathbf{x}_{t-t_2}^\top, \mathbf{0}^\top)^\top$$

...

Step $2K + 1$. $K + 1$ -st Attn ↓

$$\mathbf{x}_t^{(K+1/2)} \approx \mathbf{x}_t^{(K)} + (\mathbf{0}_{Kd}^\top, \mathbf{x}_{t-\log t_K}^\top, \mathbf{0}^\top)^\top$$

Step $2K + 2$. $K + 1$ -st FFN ↓

$$\mathbf{x}_t^{(K+1)} = \mathbf{x}_t^{(K+1/2)} + (\mathbf{0}^\top, \log t_{K+1}, \dots, \log t_M, 0)^\top$$

Step $2K + 3$. $K + 2$ -st Attn ↓

$$\mathbf{x}_t^{(K+3/2)} \approx \mathbf{x}_t^{(K+1)} + (\mathbf{0}_{(K+1)d}^\top, \mathbf{x}_{t-t_{K+1}}^\top, \dots, \mathbf{x}_{t-t_M}^\top, 0)^\top$$

Step $2K + 4$. $K + 2$ -st FFN ↓

$$\mathbf{x}_t^{(K+2)} \approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M})$$

– **Regime $M = K$.**

$$\mathbf{x}_t^{(0)}$$

Step 1. 1-st Attn ↓

$$\mathbf{x}_t^{(1/2)} = \mathbf{x}_t^{(0)}$$

Step 2. 1-st FFN ↓

$$\mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + (\mathbf{0}^\top, \log t_1, \mathbf{0}_{M-1}^\top, \log 2)^\top$$

Step 3. 2-st Attn ↓

$$\mathbf{x}_t^{(3/2)} \approx \mathbf{x}_t^{(1)} + (\mathbf{0}_d^\top, \mathbf{x}_{t-t_1}^\top, \mathbf{0}^\top)^\top$$

Step 4. 2-st FFN ↓

$$\begin{aligned}
\mathbf{x}_t^{(2)} &= \mathbf{x}_t^{(3/2)} + (\mathbf{0}^\top, \log t_2, \mathbf{0}_{M-1}^\top)^\top \\
\text{Step 5. 3-st Attn } &\downarrow \\
\mathbf{x}_t^{(5/2)} &\approx \mathbf{x}_t^{(2)} + (\mathbf{0}_{2d}^\top, \mathbf{x}_{t-t_2}^\top, \mathbf{0}^\top)^\top \\
&\dots \\
\text{Step } 2K + 1. K + 1\text{-st Attn } &\downarrow \\
\mathbf{x}_t^{(K+1/2)} &\approx \mathbf{x}_t^{(K)} + (\mathbf{0}_{Kd}^\top, \mathbf{x}_{t-t_K}^\top, \mathbf{0}^\top)^\top \\
\text{Step } 2K + 2. K + 1\text{-st FFN } &\downarrow \\
\mathbf{x}_t^{(K+1)} &\approx \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-t_1}, \dots, \mathbf{x}_{t-t_M})
\end{aligned}$$

For simplicity, we denote the following projection matrices:

$$\begin{aligned}
\mathbf{P}^{(k)} &:= (\mathbf{0}_{d \times kd} \quad \mathbf{I}_{d \times d} \quad \mathbf{0}) \in \mathbb{R}^{d \times D}, \quad 1 \leq k \leq M; \\
\mathbf{P}_\perp^{(k)} &:= \begin{pmatrix} \mathbf{I}_{kd \times kd} & \mathbf{0}_{d \times d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{d \times d} & \mathbf{I}_{(D-(k+1)d) \times (D-(k+1)d)} \end{pmatrix} \in \mathbb{R}^{(D-d) \times D}, \quad 1 \leq k \leq M; \\
\mathbf{Q}^{(k)} &:= (\mathbf{I}_{(k+1)d \times (k+1)d} \quad \mathbf{0}) \in \mathbb{R}^{(k+1)d \times D}, \quad 1 \leq k \leq M; \\
\mathbf{Q}_\perp^{(k)} &:= (\mathbf{0} \quad \mathbf{I}_{(D-(k+1)d) \times (D-(k+1)d)}) \in \mathbb{R}^{(D-(k+1)d) \times D}, \quad 1 \leq k \leq M; \\
\mathbf{R} &:= (\mathbf{0}_{(M-K)d \times (K+1)d} \quad \mathbf{I}_{(M-K)d \times (M-K)d} \quad \mathbf{0}) \in \mathbb{R}^{(M-K)d \times D}; \\
\mathbf{R}_\perp &:= \begin{pmatrix} \mathbf{I}_{(K+1)d \times (K+1)d} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{(M+1) \times (M+1)} \end{pmatrix} \in \mathbb{R}^{(D-(M-K)d) \times D}.
\end{aligned}$$

Step 1 is trivial due to the use of the residual block.

Step 2. In the same way as Step II in the proof of Theorem C.2, we obtain the conclusion in this step: If the width of FFN satisfies

$$m \geq \begin{cases} \tilde{\Omega} \left(\|g_1\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\|\log g_1\|_{\mathcal{B}}^2 T_1^2 \right), & \text{type} = \text{log} \end{cases},$$

then the following holds:

$$\mathbf{x}_t^{(1)} = \mathbf{x}_t^{(1/2)} + (\mathbf{0}^\top, t_1, \mathbf{0}_M^\top, 1)^\top.$$

Thus, (E1) holds for $l = 2$.

Step 3 ~ Step $2K + 1$.

• **Case type = lin.**

– **FFN layers.**

We use l -th ($2 \leq l \leq K$) FFN layer to express l -th memory t_l exactly.

By Lemma G.6, there exists a two-layer neural network with m neurons defined on \mathbb{R}^{ld}

$$f_{(l)}^{2\text{NN}}(\mathbf{x}) = \sum_{k=1}^m a_k^{(l)} \sigma(\mathbf{b}_k^{(l)\top} \mathbf{x} + c_k^{(l)})$$

such that

$$\left\| g_l - f_{(l)}^{2\text{NN}} \right\|_{L^\infty([-2, 2]^{ld})} \leq \tilde{O} \left(\frac{\|g_l\|_{\mathcal{B}}}{\sqrt{m}} \right).$$

For l -th FFN layer, we only need to arrange the parameters $a_k^{(l)}$, $\mathbf{b}_k^{(l)}$, and $c_k^{(l)}$ ($k \in [m], 2 \leq l \leq M$).

Denote $\bar{\mathbf{b}}_k^{(l)} = (\mathbf{b}_k^{(l)\top}, \mathbf{0}^\top)^\top \in \mathbb{R}^D$ for $k \in [m], 2 \leq l \leq K-1$. Consider the following two-layer neural network with m neurons defined on \mathbb{R}^D :

$$\mathbf{FFN}^{(l)}(\mathbf{x}) = \sum_{k=1}^m \mathbf{e}_{D-M+l-1} a_k^{(l)} \sigma \left(\bar{\mathbf{b}}_k^{(l)\top} \mathbf{x} + c_k^{(l)} \right).$$

It is easy to verify

$$\mathbf{FFN}^{(l)}(\mathbf{x}) = \left(\mathbf{0}^\top, f_{(l)}^{2\text{NN}} \left(\mathbf{Q}^{(l)} \mathbf{x} \right), \mathbf{0}_{D-M+l-1}^\top \right)^\top \in \mathbb{R}^D, \quad \forall \mathbf{x} \in \mathbb{R}^D.$$

Notice that if the width m satisfies

$$\tilde{O} \left(\frac{\|g_l\|_{\mathcal{B}}}{\sqrt{m}} \right) < \frac{1}{4},$$

and the input $\mathbf{x}_t^{(l-1/2)}$ satisfies

$$\|g_l\|_{\text{Lip}} \cdot \left\| (\mathbf{x}_t^\top, \dots, \mathbf{x}_{t-t_{l-1}}^\top)^\top - \mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)} \right\|_2 \leq \frac{1}{4}$$

the following holds:

$$\begin{aligned} & \left| g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) - f_{(l)}^{2\text{NN}} \left(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)} \right) \right| \\ & \leq \left| g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) - g_l(\mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)}) \right| \\ & \quad + \left| g_l(\mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)}) - f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)}) \right| \\ & \leq \|g_l\|_{\text{Lip}} \left\| (\mathbf{x}_t^\top, \dots, \mathbf{x}_{t-t_{l-1}}^\top)^\top - \mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)} \right\|_2 + \left\| g_l - f_{(l)}^{2\text{NN}} \right\|_{L^\infty} \\ & < \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \end{aligned}$$

Noticing $t_l = g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \in \mathbb{N}_+$, we have

$$\widetilde{f_{(l)}^{2\text{NN}}}(\mathbf{Q}^{(l-1)} \mathbf{x}_t^{(l-1/2)}) = t_l.$$

Thus, it holds that:

$$\widetilde{\mathbf{FFN}}^{(l)}(\mathbf{x}_t^{(l-1/2)}) = (\mathbf{0}^\top, t_l, \mathbf{0}_{D-M+l-1}^\top)^\top.$$

– Attn layers.

By Lemma F.2, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and K functions

$$\begin{aligned} \phi_l^{\text{exp}}(t; B) &= \sum_{h=1} \alpha_{l,h} \exp(-\beta_{l,h}(t-B)) \\ &= \sum_{h=1}^H \alpha_{l,h} \exp(\beta_{l,h}B - \beta_{l,h}t), \quad 1 \leq l \leq K \end{aligned}$$

such that $\beta_{l,h} > 0$ and

$$\sup_{1 \leq B \leq T_l} \|\mathbb{I}\{\cdot = B\} - \phi_l^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T_l}}{H^n}, \quad 1 \leq l \leq K.$$

Moreover, Noticing that $1 \leq g_l(\cdot) \leq T_l$ holds for any $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \in \mathcal{X}$ and $2 \leq l \leq K$, the following holds:

$$\begin{aligned} & \sup_{\mathbf{X}} \|\mathbb{I}\{\cdot = t_l\} - \phi_l^{\text{exp}}(\cdot; t_l)\|_{\ell_1(\mathbb{N})} \\ & \leq \sup_{1 \leq B \leq T_l} \|\mathbb{I}\{\cdot = B\} - \phi_l^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T_l}}{H^n}. \end{aligned}$$

Therefore, for the attention heads h ($h \in [H]$) in each layer l ($1 \leq l \leq K$), we can choose:

$$\begin{aligned} p^{(l+1,h)} &= \beta_{l,h}, \quad \mathbf{W}_O^{(l+1)} = \mathbf{I}, \quad \mathbf{W}_V^{(l+1,h)} = \alpha_{l,h} \boldsymbol{\delta}_{(l+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}, \\ \mathbf{W}_Q^{(l+1,h)} &= \sqrt{\beta_{l,h}} \boldsymbol{\delta}_{(D-M+l,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times (D/H)}, \\ \mathbf{W}_K^{(l+1,h)} &= \sqrt{\beta_{l,h}} \boldsymbol{\delta}_{(D,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times (D/H)}, \end{aligned}$$

where $\boldsymbol{\delta}_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

- **Case type = log.**

– **FFN layers.**

We use l -th ($2 \leq l \leq K$) FFN layer to express l -th memory t_l exactly.

By Lemma G.6, there exists a two-layer neural network with m neurons defined on \mathbb{R}^{ld}

$$f_{(l)}^{2\text{NN}}(\mathbf{x}) = \sum_{k=1}^m a_k^{(l)} \sigma(\mathbf{b}_k^{(l) \top} \mathbf{x} + c_k^{(l)})$$

such that

$$\left\| \log g_l - f_{(l)}^{2\text{NN}} \right\|_{L^\infty} \leq \tilde{\mathcal{O}} \left(\frac{\|\log g_l\|_{\mathcal{B}}}{\sqrt{m}} \right).$$

For l -th FFN layer, we only need to arrange the parameters $a_k^{(l)}$, $\mathbf{b}_k^{(l)}$, and $c_k^{(l)}$ ($k \in [m], 2 \leq l \leq M$).

Denote $\bar{\mathbf{b}}_k^{(l)} = (\mathbf{b}_k^{(l) \top}, \mathbf{0}^\top)^\top \in \mathbb{R}^D$ for $k \in [m], 2 \leq l \leq K-1$. We consider the following l -th layer 2NN with m neurons defined on \mathbb{R}^D :

$$\mathbf{FFN}^{(l)}(\mathbf{x}) = \sum_{k=1}^m e_{D-M+l-1} a_k^{(r)} \sigma(\bar{\mathbf{b}}_k^{(r) \top} \mathbf{x} + c_k^{(r)}).$$

It is easy to verify

$$\mathbf{FFN}^{(l)}(\mathbf{x}) = \left(\mathbf{0}^\top, f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}), \mathbf{0}_{D-M+l-1}^\top \right)^\top \in \mathbb{R}^D, \quad \forall \mathbf{x} \in \mathbb{R}^D.$$

Notice that if the width m satisfies

$$\tilde{\mathcal{O}} \left(\frac{\|\log g_l\|_{\mathcal{B}}}{\sqrt{m}} \right) < \frac{1}{8T_l},$$

and the input $\mathbf{x}_t^{(l-1/2)}$ satisfies

$$\|\log g_l\|_{\text{Lip}} \cdot \left\| (\mathbf{x}_t^\top, \dots, \mathbf{x}_{t-t_{l-1}}^\top)^\top - \mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)} \right\|_2 \leq \frac{1}{8T_l},$$

the following holds:

$$\begin{aligned} & \left| \log g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) - f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) \right| \\ & \leq \left| \log g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) - \log g_l(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) \right| \\ & \quad + \left| \log g_l(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) - f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) \right| \end{aligned}$$

$$\begin{aligned} &\leq \|\log g_l\|_{\text{Lip}} \left\| (\mathbf{x}_t^\top, \dots, \mathbf{x}_{t-t_{l-1}}^\top)^\top - \mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)} \right\|_2 + \left\| \log g_l - f_{(l)}^{2\text{NN}} \right\|_{L^\infty} \\ &< \frac{1}{8T_l} + \frac{1}{8T_l} = \frac{1}{4T_l}, \end{aligned}$$

which ensures

$$\begin{aligned} &\left| \exp \left(f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) \right) - g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right| \\ &= \left| \exp \left(f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) \right) - \exp \left(\log \left(g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right) \right) \right| \\ &\leq \exp \left(\max \left\{ f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}), \log \left(g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right) \right\} \right) \\ &\quad \cdot \left| f_{(l)}^{2\text{NN}}(\mathbf{Q}^{(l)} \mathbf{x}_t^{(l-1/2)}) - \log \left(g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right) \right| \\ &\leq \exp \left(\log \left(g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right) + \frac{1}{8} \right) \frac{1}{4T_r} \\ &\leq e^{1/8} \cdot T_r \cdot \frac{1}{4T_r} < \frac{1}{2}. \end{aligned}$$

Noticing $t_l = g_l(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \in \mathbb{N}_+$, we have

$$\widetilde{f_{(l)}^{2\text{NN}}}(\mathbf{Q}_l \mathbf{x}_t^{(l-1/2)}) = \log \left(\exp \left[\text{FFN}^{(l)}(\mathbf{x}_t, \dots, \mathbf{x}_{t-t_{l-1}}) \right] \right) = \log t_l.$$

Thus, it holds that:

$$\widetilde{\text{FFN}}^{(l)}(\mathbf{x}_t^{(l-1/2)}) = (\mathbf{0}^\top, \log t_l, \mathbf{0}_{D-M+l-1}^\top)^\top.$$

– **Attn layers.**

By Lemma F.5, for any rate $n \in \mathbb{N}_+$, there exists a constant $C(n)$ and K functions

$$\begin{aligned} \phi_l^{\text{poly}}(t; B) &= \sum_{h=1}^H \alpha_{l,h} (t/B)^{-\beta_{l,h}} \\ &= \sum_{h=1}^H \alpha_{l,h} \exp \left(-\beta_{l,h} \log(t/B) \right), \quad 1 \leq l \leq K \end{aligned}$$

such that $\beta_{l,h} > 0$ and

$$\sup_{1 \leq B \leq T_l} \left\| \mathbb{I}\{\cdot = B\} - \phi_l^{\text{poly}}(\cdot; B) \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T_l^{1.01(n+1)}}{H^n}, \quad 1 \leq l \leq K.$$

Moreover, Noticing that $1 \leq g_l(\cdot) \leq T_l$ holds for any $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \in \mathcal{X}$ and $1 \leq l \leq K$, the following holds:

$$\begin{aligned} &\sup_{\mathbf{X}} \left\| \mathbb{I}\{\cdot = t_l\} - \phi_l^{\text{poly}}(\cdot; t_l) \right\|_{\ell_1(\mathbb{N}_+)} \\ &\leq \sup_{1 \leq B \leq T_l} \left\| \mathbb{I}\{\cdot = B\} - \phi_l^{\text{poly}}(\cdot; B) \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T_l^{1.01(n+1)}}{H^n}. \end{aligned}$$

Therefore, for the attention heads h ($h \in [H]$) in each layer l ($1 \leq l \leq K$), we can choose:

$$\begin{aligned} p^{(l+1,h)} &= \beta_{l,h}, \quad \mathbf{W}_O^{(l+1)} = \mathbf{I}, \quad \mathbf{W}_V^{(l+1,h)} = \alpha_{l,h} \boldsymbol{\delta}_{(l+1,1)}^{(d \times d)} \in \mathbb{R}^{D \times D}, \\ \mathbf{W}_Q^{(l+1,h)} &= \sqrt{\beta_{l,h}} \boldsymbol{\delta}_{(D-M+l,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times (D/H)}, \\ \mathbf{W}_K^{(l+1,h)} &= \sqrt{\beta_{l,h}} \boldsymbol{\delta}_{(D,1)}^{(1 \times 1)} \in \mathbb{R}^{D \times (D/H)}, \end{aligned}$$

where $\boldsymbol{\delta}_{(p_1, p_2)}^{(r \times r)}$ means that: it equals to $\mathbf{I}_{r \times r}$ for the (p_1, p_2) -th $r \times r$ blocks, and $\mathbf{0}_{r \times r}$ for the other $r \times r$ blocks.

Similar to the estimate in Step II and Step III in the proof of Theorem C.2, it is easy to prove the following estimates by induction.

If the width satisfies

$$m \geq \begin{cases} \tilde{\Omega} \left(\max_{l \in [K]} \|g_l\|_{\mathcal{B}}^2 \right) = \tilde{\Omega} \left(\|g_K\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\max_{l \in [K]} \|\log g_l\|_{\mathcal{B}}^2 T_l^2 \right) = \tilde{\Omega} \left(\|\log g_K\|_{\mathcal{B}}^2 T_K^2 \right), & \text{type} = \text{log} \end{cases},$$

and the head number satisfies

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} e^{0.02(n+1)T_l}} \leq \frac{1}{4 \max_{l \in [K-1]} \|g_l\|_{\text{Lip}}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} T_l^{2.02(n+1)}} \leq \frac{1}{4 \max_{l \in [K-1]} \|\log g_l\|_{\text{Lip}}}, & \text{type} = \text{log} \end{cases},$$

then the following estimates hold:

- **(E1)** for any $2 \leq l \leq K$,

$$\mathbf{x}_t^{(l)} = \mathbf{x}_t^{(l-1/2)} + (\mathbf{0}^\top, t_l, \mathbf{0}_{M-l+1}^\top)^\top;$$

- **(E2)** for any $1 \leq l \leq K$,

$$\mathbf{P}_\perp^{(l)} \mathbf{x}_t^{(l+1/2)} = \mathbf{P}_\perp^{(l)} \mathbf{x}_t^{(l)};$$

- **(E3)** for any $1 \leq l \leq K$,

$$\left\| \mathbf{P}^{(l)} \left(\mathbf{x}_t^{(l+1/2)} - \left(\mathbf{x}_t^{(l)} + (\mathbf{0}_{ld}^\top, \mathbf{x}_{t-t_l}^\top, \mathbf{0}^\top)^\top \right) \right) \right\|_2 \leq \begin{cases} \frac{C(n)e^{0.01(n+1)T_l}}{H^n}, & \text{type} = \text{lin} \\ \frac{C(n)T_l^{1.01(n+1)}}{H^n}, & \text{type} = \text{log} \end{cases};$$

$$\left\| \mathbf{Q}^{(l)} \left(\mathbf{x}_t^{(l+1/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 \leq \begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{j=1}^l e^{0.02(n+1)T_j}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{j=1}^l T_j^{2.02(n+1)}}, & \text{type} = \text{log} \end{cases}.$$

The Remained Steps.

- **Regime** $M \geq K + 1$.

Step $2K + 2$ and $2K + 3$.

In the similar way as Step 3 ~ Step $2K - 1$ in this proof and Step II, Step III in the proof of Theorem C.2, it is easy to verify the following estimate.

If the width satisfies

$$m \geq \begin{cases} \tilde{\Omega} \left(\sum_{l=K+1}^M \|g_l\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\sum_{l=K+1}^M \|\log g_l\|_{\mathcal{B}}^2 T_l^2 \right), & \text{type} = \text{log} \end{cases},$$

and the head number satisfies

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}} \leq \frac{1}{4 \max_{l \in [K]} \|g_l\|_{\text{Lip}}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}} \leq \frac{1}{4 \max_{l \in [K]} \|\log g_l\|_{\text{Lip}}}, & \text{type} = \text{log} \end{cases},$$

then the following estimates hold:

$$\mathbf{x}_t^{(K+1)} = \mathbf{x}_t^{(K+1/2)} + (\mathbf{0}^\top, t_{K+1}, \dots, t_M, 0)^\top;$$

$$\mathbf{R}_\perp \mathbf{x}_t^{(K+3/2)} = \mathbf{R}_\perp \mathbf{x}_t^{(K+1)};$$

$$\begin{aligned} & \left\| \mathbf{R} \left(\mathbf{x}_t^{(K+3/2)} - \left(\mathbf{x}_t^{(K+1)} + (\mathbf{0}_{(K+1)D}^\top, \mathbf{x}_{t-t_{K+1}}^\top, \dots, \mathbf{x}_{t-t_M}^\top, \mathbf{0}^\top)^\top \right) \right) \right\|_2 \\ & \leq \begin{cases} C(n) \left(\frac{\sum_{l=K+1}^M e^{0.01T_l}}{H^{\frac{n}{n+1}}} \right)^{n+1}, & \text{type} = \text{lin} \\ C(n) \left(\frac{\sum_{l=K+1}^M T_l^{1.01}}{H^{\frac{n}{n+1}}} \right)^{n+1}, & \text{type} = \text{log} \end{cases}, \\ & \left\| \mathbf{Q}^{(M)} \left(\mathbf{x}_t^{(K+3/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 \\ & \leq \begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l} + \left(\sum_{l=K+1}^M e^{0.01T_l} \right)^{2n+2}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)} + \left(\sum_{l=K+1}^M T_l^{1.01} \right)^{2n+2}}, & \text{type} = \text{log} \end{cases}. \end{aligned}$$

Step $2K + 4$ and the final bound.

In the same way as Step IV in the proof of Theorem C.2, there exists FFN, such that the following estimate holds for any t and \mathbf{X} :

$$\begin{aligned} \left\| \mathbf{H}_t(\mathbf{X}) - \mathbf{x}_t^{(K+2)} \right\| & \leq \|f\|_{\text{Lip}} \cdot \left\| \mathbf{Q}^{(M)} \left(\mathbf{x}_t^{(K+3/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 + \mathcal{E}_{\text{FFN}} \\ & = \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}}(\text{type}) + \mathcal{E}_{\text{FFN}}, \end{aligned}$$

where

$$\mathcal{E}_{\text{FFN}} = \mathcal{O} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right),$$

$$\begin{aligned} \mathcal{E}_{\text{Attn}}(\text{type}) & = \left\| \mathbf{Q}^{(M)} \left(\mathbf{x}_t^{(K+3/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 \\ & = \begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l} + \left(\sum_{l=K+1}^M e^{0.01T_l} \right)^{2n+2}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)} + \left(\sum_{l=K+1}^M T_l^{1.01} \right)^{2n+2}}, & \text{type} = \text{log} \end{cases}. \end{aligned}$$

Recalling our analysis, we need the head number satisfies

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}} \leq \frac{1}{4 \max_{l \in [K]} \|g_l\|_{\text{Lip}}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}} \leq \frac{1}{4 \max_{l \in [K]} \|\log g_l\|_{\text{Lip}}}, & \text{type} = \text{log} \end{cases}.$$

Due to

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}} \leq \mathcal{E}_{\text{Attn}}(\text{type}), & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}} \leq \mathcal{E}_{\text{Attn}}(\text{type}), & \text{type} = \text{log} \end{cases},$$

when n is large enough, this condition holds naturally and do not affect the approximation rate.

Moreover, we need the following condition on the width:

$$m \geq \begin{cases} \tilde{\Omega} \left(\max_{l \in [K]} \vee \sum_{l=K+1}^M \|g_l\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\max_{l \in [K]} \vee \sum_{l=K+1}^M \|\log g_l\|_{\mathcal{B}}^2 T_l^2 \right), & \text{type} = \text{log} \end{cases},$$

- **Regime $M = K$.**

Step $2K + 2$ and the final bound.

In the same way as Step IV in the proof of Theorem C.2, there exists \mathbf{FFN} , such that the following estimate holds for any t and \mathbf{X} :

$$\begin{aligned} \left\| \mathbf{H}_t(\mathbf{X}) - \mathbf{x}_t^{(K+1)} \right\| &\leq \|f\|_{\text{Lip}} \cdot \left\| \mathbf{Q}^{(M)} \left(\mathbf{x}_t^{(K+1/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 + \mathcal{E}_{\text{FFN}} \\ &= \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}}(\text{type}) + \mathcal{E}_{\text{FFN}}, \end{aligned}$$

where

$$\mathcal{E}_{\text{FFN}} = \mathcal{O} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right),$$

$$\begin{aligned} \mathcal{E}_{\text{Attn}}(\text{type}) &= \left\| \mathbf{Q}_M \left(\mathbf{x}_t^{(K+1/2)} - \mathbf{x}_t^{(0)} \right) \right\|_2 \\ &= \begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}}, & \text{type} = \text{log} \end{cases}. \end{aligned}$$

Recalling our analysis, we need the head number satisfies

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} e^{0.02(n+1)T_l}} \leq \frac{1}{4 \max_{l \in [K-1]} \|g_l\|_{\text{Lip}}}, & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} T_l^{2.02(n+1)}} \leq \frac{1}{4 \max_{l \in [K-1]} \|\log g_l\|_{\text{Lip}}}, & \text{type} = \text{log} \end{cases}.$$

Due to

$$\begin{cases} \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} e^{0.02(n+1)T_l}} \leq \mathcal{E}_{\text{Attn}}(\text{type}), & \text{type} = \text{lin} \\ \frac{C(n)}{H^n} \sqrt{\sum_{l=1}^{K-1} T_l^{2.02(n+1)}} \leq \mathcal{E}_{\text{Attn}}(\text{type}), & \text{type} = \text{log} \end{cases},$$

when H is large enough, this condition holds naturally and do not affect the approximation rate.

Moreover, we need the following condition on the width:

$$m \geq \begin{cases} \tilde{\Omega} \left(\max_{l \in [K]} \|g_l\|_{\mathcal{B}}^2 \right), & \text{type} = \text{lin} \\ \tilde{\Omega} \left(\max_{l \in [K]} \|\log g_l\|_{\mathcal{B}}^2 T_l^2 \right), & \text{type} = \text{log} \end{cases},$$

Combining these two regimes, we complete our proof. □

D.2 Proof of Proposition 4.5

Proof of Proposition 4.5.

This proposition is a direct corollary of Theorem 4.4. It can be seen as a special case of $M = K$ in Theorem 4.4.

Therefore, under the same conditions, there exists a $K + 1$ -layer Transformer $\mathbf{TF} \in \mathcal{TF}_{(K+1,H,m)}^{\text{NF,type}}$ (12) and a constant $C(n)$ such that: if the width satisfies

$$m \geq \begin{cases} \tilde{\Omega}\left(\max_{i \in [K]} \|g_i\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin}, \\ \tilde{\Omega}\left(\max_{i \in [K]} \|\log g_i\|_{\mathcal{B}}^2 T_i^2\right), & \text{type} = \text{log} \end{cases},$$

then the following approximation rate holds:

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type}),$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right)$ and

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \begin{cases} \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{H^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}}\right), & \text{type} = \text{log} \end{cases}.$$

□

Comparison between Proposition 4.5 and Theorem 4.1.

We compare 2-layer Transformer and $M + 1$ -layer Transformer regarding the requirement of the number of heads and width.

- *The required width of FFN layers.*

- For 2-layer Transformer, the required width of FFN layers $m_{\text{need}}^{(2)}$ is proportionally linked to the **sum** of all the memory functions' complexity:

$$m_{\text{need}}^{(2)} = \begin{cases} \tilde{\Omega}\left(\sum_{i \in [K]} \|g_i\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin}, \\ \tilde{\Omega}\left(\sum_{i \in [K]} \|\log g_i\|_{\mathcal{B}}^2 T_i^2\right), & \text{type} = \text{log} \end{cases}.$$

- For $M + 1$ -layer Transformer, the required width of FFN layers $m_{\text{need}}^{(M+1)}$ correlates with the **maximum** complexity of the memory functions:

$$m_{\text{need}}^{(M+1)} = \begin{cases} \tilde{\Omega}\left(\max_{i \in [K]} \|g_i\|_{\mathcal{B}}^2\right), & \text{type} = \text{lin}, \\ \tilde{\Omega}\left(\max_{i \in [K]} \|\log g_i\|_{\mathcal{B}}^2 T_i^2\right), & \text{type} = \text{log} \end{cases}.$$

It is easy to see that:

$$\frac{m_{\text{need}}^{(M+1)}}{m_{\text{need}}^{(2)}} = \frac{\max\{a_1, \dots, a_M\}}{\sum_{k=1}^M a_k},$$

$$\max\{a_1, \dots, a_M\} \leq \sum_{k=1}^M a_k$$

- *The required number of Attn heads.* To achieve the same $\mathcal{E}_{\text{Attn}}(\text{type}) = \epsilon$,

– for 2-layer Transformer, the required number of Attn heads $H_{\text{need}}^{(2)}$ satisfies:

$$\epsilon = \begin{cases} \mathcal{O}\left(\frac{C(n)}{(H_{\text{need}}^{(2)})^n} \sqrt{\sum_{l=1}^K e^{0.02(n+1)T_l}}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{(H_{\text{need}}^{(2)})^n} \sqrt{\sum_{l=1}^K T_l^{2.02(n+1)}}\right), & \text{type} = \text{log} \end{cases}.$$

– for $M + 1$ -layer Transformer, the required number of Attn heads $H_{\text{need}}^{(M+1)}$ satisfies:

$$\epsilon = \begin{cases} \mathcal{O}\left(\frac{C(n)}{(H_{\text{need}}^{(M+1)})^n} \left(\sum_{i=1}^M e^{0.01T_i}\right)^{n+1}\right), & \text{type} = \text{lin} \\ \mathcal{O}\left(\frac{C(n)}{(H_{\text{need}}^{(M+1)})^n} \left(\sum_{i=1}^M T_i^{1.01}\right)^{n+1}\right), & \text{type} = \text{log} \end{cases}.$$

It is easy to see that:

$$\left(\frac{H_{\text{need}}^{(M+1)}}{H_{\text{need}}^{(2)}}\right)^{2n} = \frac{b_1^2 + \dots + b_M^2}{(b_1 + \dots + b_M)^2},$$

$$b_1^2 + \dots + b_M^2 \leq (b_1 + \dots + b_M)^2.$$

This finding suggests that increased depth can significantly reduce the demands on the number of heads and the width. The underlying reason is that deep networks can distribute memories across different layers for processing, with each layer focusing on approximating only a single memory function.

E Proof of Section 5

E.1 Proof of Theorem 5.1

In this subsection, we give the detailed proofs of the warm-up case of (fixed) essentially sparse memories as follows:

$$y_t = f((\mathbf{X} * \rho_1)(t), \dots, (\mathbf{X} * \rho_M)(t)),$$

where $\rho_1(\cdot), \dots, \rho_M(\cdot) \in \ell^1(\mathbb{N})$ serve as memory kernels, and $(\mathbf{X} * \rho_k)(t) = \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \rho_k(s)$ denotes the convolution of the inputs with kernel ρ_k .

Theorem E.1 (Restatement of Theorem 5.1).

(A) Consider \mathcal{H}^{Ess} (14) with exponentially decayed memory kernels, i.e., there exists $\beta > 0$ such that $\rho_1(t), \dots, \rho_M(t) = \mathcal{O}(e^{-\beta t})$. Then for any target $\mathbf{H} \in \mathcal{H}^{\text{Ess}}$, rate $n \in [[99\beta]]$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer DP-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,exp}}$ (7) and a constant $C(n)$ such that

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type});$$

(B) Consider \mathcal{H}^{Ess} (14) with polynomially decayed memory kernels, i.e., there exists $\beta > 1$ such that $\rho_1(t), \dots, \rho_M(t) = \mathcal{O}(t^{-\beta})$. Then for any target $\mathbf{H} \in \mathcal{H}^{\text{Ess}}$, rate $n \in [[0.99\beta] - 1]$, and $H, m \in \mathbb{N}_+$, there exists a 1-layer DP-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,poly}}$ (7) and a constant $C(n)$ such that

$$\|\mathbf{H} - \mathbf{TF}\| \leq \mathcal{E}_{\text{FFN}} + \|f\|_{\text{Lip}} \mathcal{E}_{\text{Attn}}(\text{type});$$

where $\mathcal{E}_{\text{FFN}} = \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathbb{B}}}{\sqrt{m}}\right)$ and

$$\mathcal{E}_{\text{Attn}}(\text{type}) = \mathcal{O}\left(\frac{C(n)M^{n+1}}{H^n}\right).$$

Proof of Theorem E.1.

The proof of this theorem is highly similar to the proof of Theorem B.1. The only difference is that the Attn layer needs to be used to approximate general memory kernel $\rho_k(\cdot)$ instead of simple $\mathbb{I}\{\cdot = T_k\}$. But for the completeness of the proof in this section, we still provide the detailed proof.

First, we choose the embedding dimension $D = Md$, and select the simple embedding $\mathbf{W}_E = (\mathbf{I}_{d \times d}, \mathbf{0})^\top \in \mathbb{R}^{D \times d}$, $\mathbf{b}_E = \mathbf{0} \in \mathbb{R}^D$.

For any input sequence $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}}$, the token after embedding satisfies:

$$\mathbf{x}_t^E = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = (\mathbf{x}_t^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^D.$$

Then for one-layer Dot-product-free Transformer $\mathbf{TF} \in \mathcal{TF}_{(1,H,m)}^{\text{DPF,type}}$ without residual blocks, the output token $\mathbf{TF}_t(\mathbf{X})$ of t -th input token \mathbf{x}_t satisfies:

$$\begin{aligned} \mathbf{x}_t^{(1/2)} &= \mathbf{W}_O^{(1)} \sum_{h=1}^H \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}), \\ \mathbf{x}_t^{(1)} &= \mathbf{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \end{aligned}$$

where

$$\mathbf{Attn}_t^{(1,h)}(\mathbf{X}) = \mathbf{W}_V^{(1,h)} \sum_{s=0}^{+\infty} \frac{\mathbf{x}_{t-s} \exp(p^{(1,h)} \phi_{\text{type}}(s))}{\sum_{j=0}^{+\infty} \exp(p^{(1,h)} \phi_{\text{type}}(j))}.$$

This proof can be summarized as the following process:

$$\dots \quad \mathbf{x}_t^E \quad \dots$$

Step I. Attn layer ↓

$$\cdots \mathbf{x}_t^{(1/2)} \approx ((\mathbf{X} * \rho_1)(t), \cdots, (\mathbf{X} * \rho_M)(t))^\top \cdots$$

Step II. FFN layer ↓

$$\cdots \mathbf{x}_t^{(1)} \approx \mathbf{f}((\mathbf{X} * \rho_1)(t), \cdots, (\mathbf{X} * \rho_M)(t)) \cdots$$

Now we give the formal proof.

Step I. Extract the memory locations by (Dot-product-free) Attn layer.

We consider to use H_k attention heads (from $\sum_{i=1}^{k-1} H_i + 1$ -th head to $\sum_{i=1}^k H_i$ -th head) to extract it, and it satisfies to $\sum_{k=1}^M H_k = H$.

$$\mathbf{P}^{(k)} := \begin{pmatrix} \mathbf{0}_{d \times (k-1)d} & \mathbf{I}_{d \times d} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{d \times D}, \quad 1 \leq k \leq M.$$

$$\mathbf{P}_\perp^{(k)} := \begin{pmatrix} \mathbf{I}_{(k-1)d \times (k-1)d} & \mathbf{0}_{d \times d} & \mathbf{0} \\ \mathbf{0}_{d \times d} & \mathbf{0}_{d \times d} & \mathbf{I}_{(M-k-1)d \times (M-k-1)d} \end{pmatrix} \in \mathbb{R}^{(M-1)d \times D}, \quad 1 \leq k \leq M.$$

Now we consider the extraction of k -th memory $(\mathbf{X} * \rho_k)(t)$ ($1 \leq k \leq M$).

- **Case (A). Approximating exponentially decayed memories by type = lin.**

Because there exists $\beta > 0$ such that $\rho_k(t) = \mathcal{O}(e^{-\beta t})$, by Lemma F.3, for any $n \in \llbracket 99\beta \rrbracket$ and $m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a function

$$\phi_k^{\text{exp}}(t) = \sum_{\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h e^{-\beta_h t}$$

such that $\beta_h > 0$ and

$$\|\rho_k(\cdot) - \phi_k^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} = \sum_{s=0}^{+\infty} |\rho_k(s) - \phi_k^{\text{exp}}(s)| \leq \frac{C(n)}{m^n}.$$

Therefore, for these attention heads ($\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i$), we can choose

$$p^{(1,h)} = \beta_h, \quad \mathbf{W}_V^{(1,h)} = \alpha_h \left(\sum_{j=0}^{+\infty} \exp(-\beta_h j) \right) \boldsymbol{\delta}_{(k,1)}^{d \times d},$$

where $\boldsymbol{\delta}_{(k,1)} \in \mathbb{R}^{D \times D}$ means that: it equals to $\mathbf{I}_{d \times d}$ for the $(k, 1)$ -th $d \times d$ blocks, and $\mathbf{0}_{d \times d}$ for the other $d \times d$ blocks.

Then it holds that:

$$\sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \begin{pmatrix} \mathbf{0}_{(k-1)d} \\ \mathbf{x}_{t-s} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^D,$$

This implies:

$$\mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \sum_{h=\sum_{i=1}^{k-1} H_i + 1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \mathbf{x}_{t-s},$$

$$\mathbf{P}_\perp^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \mathbf{0},$$

moreover, the following estimate holds:

$$\begin{aligned} & \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) - (\mathbf{X} * \rho_k)(t) \right\|_2 \\ &= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=0}^{+\infty} e^{-\beta_h s} \mathbf{x}_{t-s} - \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \rho_k(s) \right\|_2 \\ &= \left\| \sum_{s=0}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h s} - \mathbb{I}\{s = T_k\} \right) \mathbf{x}_{t-s} \right\|_2 \\ &\leq \sum_{s=0}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h e^{-\beta_h s} - \rho_k(s) \right| \\ &= \|\phi_k^{\text{exp}}(\cdot) - \rho_k(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)}{H_k^n}. \end{aligned}$$

- **Case (B). Approximating polynomially decayed memories by type = log.**

Because there exists $\beta > 0$ such that $\rho_k(t) = \mathcal{O}(t^{-\beta})$, by Lemma F.6, for any $n \in \llbracket [0.99\beta] - 1 \rrbracket$ and $m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a function

$$\phi_k^{\text{poly}}(t) = \sum_{\sum_{i=1}^{k-1} H_i+1 \leq h \leq \sum_{i=1}^k H_i} \alpha_h t^{-\beta_h}$$

such that $\beta_h > 1$ and

$$\left\| \rho_k(\cdot) - \phi_k^{\text{poly}}(\cdot) \right\|_{\ell_1(\mathbb{N}_+)} = \sum_{s=1}^{+\infty} \left| \rho_k(s) - \phi_k^{\text{poly}}(s) \right| \leq \frac{C(n)}{m^n}.$$

Therefore, for these attention heads ($\sum_{i=1}^{k-1} H_i + 1 \leq h \leq \sum_{i=1}^k H_i$), we can choose

$$p^{(1,h)} = \beta_h, \quad \mathbf{W}_V^{(1,h)} = \alpha_h \left(\sum_{j=1}^{+\infty} j^{-\beta_h} \right) \boldsymbol{\delta}^{(k,1)},$$

where $\boldsymbol{\delta}^{(k,1)} \in \mathbb{R}^{D \times D}$ means that: it equals to $\mathbf{I}_{d \times d}$ for the $(k, 1)$ -th $d \times d$ blocks, and $\mathbf{0}_{d \times d}$ for the other $d \times d$ blocks.

Then it holds that:

$$\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=1}^{+\infty} s^{-\beta_h} \begin{pmatrix} \mathbf{0}_{(k-1)d} \\ \mathbf{x}_{t-s} \\ \mathbf{0} \end{pmatrix},$$

This implies:

$$\mathbf{P}^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=1}^{+\infty} s^{-\beta_h} \mathbf{x}_{t-s},$$

$$\mathbf{P}_\perp^{(k)} \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) = \mathbf{0},$$

$$\begin{aligned} & \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) - (\mathbf{X} * \rho_k)(t) \right\|_2 \\ &= \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h \sum_{s=1}^{+\infty} s^{-\beta_h} \mathbf{x}_{t-s} - \sum_{s=0}^{+\infty} \mathbf{x}_{t-s} \rho_k(s) \right\|_2 \\ &= \left\| \sum_{s=1}^{+\infty} \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h s^{-\beta_h} - \rho_k(s) \right) \mathbf{x}_{t-s} \right\|_2 \\ &\leq \sum_{s=1}^{+\infty} \left| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \alpha_h s^{-\beta_h} - \rho_k(s) \right| \\ &= \left\| \phi_{H_k}^{\text{poly}}(\cdot) - \rho_k(\cdot) \right\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)}{H_k^n}. \end{aligned}$$

Then we combine the estimate for all $k \in [M]$ for these two cases. By choose $\mathbf{W}_O = \mathbf{I}_D$, we have:

$$\begin{aligned} & \left\| \mathbf{x}_t^{(1/2)} - \begin{pmatrix} (\mathbf{X} * \rho_1)(t) \\ \vdots \\ (\mathbf{X} * \rho_M)(t) \end{pmatrix} \right\|_2 \\ &= \left\| \sum_{k=1}^M \left(\sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{(k-1)d} \\ (\mathbf{X} * \rho_k)(t) \\ \mathbf{0}_d \end{pmatrix} \right) \right\|_2 \\ &\leq \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \text{Attn}_t^{(1,h)}(\mathbf{X}) - \begin{pmatrix} \mathbf{0}_{(k-1)d} \\ (\mathbf{X} * \rho_k)(t) \\ \mathbf{0}_d \end{pmatrix} \right\|_2 \\ &= \sum_{k=1}^M \left\| \sum_{h=\sum_{i=1}^{k-1} H_i+1}^{\sum_{i=1}^k H_i} \mathbf{P}^{(k)} \text{Attn}_t^{(1,h)}(\mathbf{X}^{(0)}) - (\mathbf{X} * \rho_k)(t) \right\|_2 \\ &\leq \mathcal{E}_{\text{Attn}} := \sum_{k=1}^M \frac{C(n)}{H_k^n}, \quad \text{for both Case (A) and Case (B)}. \end{aligned}$$

Consequently, one detail is to assign the head number $\{H_k\}_{k=1}^M$ such that the error's sum $\mathcal{E}_{\text{Attn}}(\text{type})$ is as small as possible. Here, we simply choose the same H_k :

$$H_k = \frac{H}{M}, \quad k \in [M].$$

Thus, we obtain the bound in Step I:

$$\mathcal{E}_{\text{Attn}} = \sum_{k=1}^M \frac{C(n)}{H_k^n} = \frac{C(n)M^{n+1}}{H^n}.$$

Furthermore, by choosing $\mathcal{E}_{\text{Attn}} \leq 1$, it holds that

$$\left\| \mathbf{x}_t^{(1/2)} \right\|_{\infty} \leq \left\| \mathbf{x}_t^{(1/2)} - \begin{pmatrix} (\mathbf{X} * \rho_1)(t) \\ \vdots \\ (\mathbf{X} * \rho_M)(t) \end{pmatrix} \right\|_{\infty} + \left\| \begin{pmatrix} (\mathbf{X} * \rho_1)(t) \\ \vdots \\ (\mathbf{X} * \rho_M)(t) \end{pmatrix} \right\|_{\infty} \leq \mathcal{E}_{\text{Attn}} + 1 \leq 2.$$

Step II. Approximate the readout function by FFN layer.

In this step, we aim to approximate the function f using two-layer network. By Lemma G.6, there exists a two-layer neural network with m neurons defined on \mathbb{R}^D

$$\text{FFN}^{(1)}(\mathbf{y}) = \sum_{k=1}^m a_k \sigma(\mathbf{b}_k^\top \mathbf{y} + c_k)$$

such that

$$\mathcal{E}_{\text{FFN}} := \left\| \text{FFN}^{(1)} - f \right\|_{L^\infty([-2,2]^D)} \leq \tilde{\mathcal{O}} \left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}} \right).$$

The final bound.

For any t and $\mathbf{X} \in \mathcal{X}$, it holds that

$$\begin{aligned} & \left\| \mathbf{H}_t(\mathbf{X}) - \mathbf{x}_t^{(1)} \right\| = \left| f((\mathbf{X} * \rho_1)(t), \dots, (\mathbf{X} * \rho_M)(t)) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &= \left| f((\mathbf{X} * \rho_1)(t), \dots, (\mathbf{X} * \rho_M)(t)) - f(\mathbf{x}_t^{(1/2)}) + f(\mathbf{x}_t^{(1/2)}) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &\leq \left| f((\mathbf{X} * \rho_1)(t), \dots, (\mathbf{X} * \rho_M)(t)) - f(\mathbf{x}_t^{(1/2)}) \right| + \left| f(\mathbf{x}_t^{(1/2)}) - \text{FFN}^{(1)}(\mathbf{x}_t^{(1/2)}) \right| \\ &\leq \|f\|_{\text{Lip}} \left\| ((\mathbf{X} * \rho_1)(t)^\top, \dots, (\mathbf{X} * \rho_M)(t)^\top)^\top - \mathbf{x}_t^{(1/2)} \right\|_2 + \left\| f - \text{FFN}^{(1)} \right\|_{L^\infty([-2,2]^D)} \\ &\leq \|f\|_{\text{Lip}} \cdot \mathcal{E}_{\text{Attn}} + \mathcal{E}_{\text{FFN}}, \end{aligned}$$

where

$$\mathcal{E}_{\text{FFN}} = \frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}; \quad \mathcal{E}_{\text{Attn}} = \frac{C(n)M^{n+1}}{H^n}, \quad \text{for both Case (A) and Case (B).}$$

Due to the arbitrariness of t and \mathbf{X} , the proof is completed. □

F Key Lemmas about Approximation

F.1 Approximation by the sum of exponential decay

Lemma F.1 (Exp decay, fixed Delta function). *For any $T \in \mathbb{N}_+$, $n, m \in \mathbb{N}_+$, there exists and absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{exp}}(t) = \sum_{k=1}^m \alpha_k e^{-\beta_k t}$ such that*

$$\|\mathbb{I}(\cdot = T) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)e^{0.01(n+1)T}}{m^n}.$$

where $\beta_k > 0$ holds for any $k \in [m]$.

Proof of Lemma F.1.

Let $\alpha, \gamma > 0$ be constants, and they will take specific values at the end of the proof.

First, recall the standard bump function on $[-1, 1]$:

$$\Psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & x \in (-1, 1) \\ 0, & \text{otherwise} \end{cases},$$

and we can define the following constants for $T \geq 1$:

$$\mu_T = e^{-\alpha T}, \quad \sigma_T = e^{-\alpha T} - e^{-\alpha(T+1)}.$$

Then we consider the following bump function $\Psi_T \in \mathcal{C}^\infty([0, 1])$:

$$\Psi_T(x) = \begin{cases} V_T \Psi\left(\frac{x-\mu_T}{\sigma_T}\right), & x \in (\mu_T - \sigma_T, \mu_T + \sigma_T) \\ 0, & \text{otherwise} \end{cases},$$

where V_T is a scaling constant such that $\Psi_T(e^{-\alpha T}) = e^{\gamma T}$.

First, we consider the approximation of Ψ_T on $[0, 1]$.

Notice that $\Psi_T \in \mathcal{C}^\infty([0, 1])$, and $\Psi_T^{(k)}(0) = 0$ for any $k \in \mathbb{N}$. For the standard bump function Ψ , for any $n \in \mathbb{N}_+$, there exists an absolute constant $M(n) > 0$ only depending on n , such that $\max_{0 \leq k \leq 10} \sup_{x \in [-1, 1]} |\Psi^{(k)}(x)| \leq M(n)$.

Notice that for any $k \in \mathbb{N}$ and $x \in [0, 1]$,

$$\Psi_T^{(k)}(x) = \frac{V_T}{\sigma_T^k} \Psi^{(k)}\left(\frac{x-\mu_T}{\sigma_T}\right).$$

Therefore, the following upper bound holds:

$$\begin{aligned} M_T(n) &= \max_{0 \leq k \leq n} \frac{V_T}{\sigma_T^k} M(n) = \frac{V_T}{\sigma_T^n} M(n) \\ &= \frac{e^{\gamma T} \cdot e}{(e^{-\alpha T} - e^{-\alpha(T+1)})^n} M(n) = \frac{M(n)e}{(1 - 1/e)^n} e^{(\gamma+n\alpha)T} := C(n, \alpha) e^{(\gamma+n\alpha)T}. \end{aligned}$$

By Lemma G.5, for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0, 1]} |\Psi_T(x) - Q_m(x)| \leq \frac{M_T(n)}{m^n} \leq \frac{C(n, \alpha) e^{(\gamma+n\alpha)T}}{m^n}.$$

Now we use the transform $x = e^{-\alpha t}$ on the function Ψ and consider

$$\Phi_T(t) := e^{-\gamma t} \Psi_T(e^{-\alpha t}), \quad t \in [0, +\infty).$$

It is easy to verify that Φ_T satisfies that

$$\Phi_T(t)|_{\mathbb{N}} = \mathbb{I}(t = T).$$

Moreover, we consider the function

$$P_m(t) := e^{-\gamma t} Q_m(e^{-\alpha t}), \quad t \in [0, +\infty).$$

Then by choosing $\alpha = \gamma = 0.01$, the following error estimate holds:

$$\begin{aligned} \|P_m(\cdot) - \mathbb{I}(\cdot = T)\|_{\ell_1(\mathbb{N})} &= \sum_{t=0}^{+\infty} |P_m(t) - \Phi_T(t)| \\ &= \sum_{t=0}^{+\infty} e^{-\gamma t} |Q_m(e^{-\alpha t}) - \Psi_T(e^{-\alpha t})| \leq \sum_{t=0}^{+\infty} e^{-\gamma t} \frac{M_T(n)}{m^n} \\ &\leq \frac{C(n, \alpha) e^{(\gamma+n\alpha)T}}{m^n} \sum_{t=0}^{+\infty} e^{-\gamma t} \leq \frac{C(n) e^{0.01(n+1)T}}{m^n} \frac{1}{1 - e^{-\gamma}} \\ &= \frac{\tilde{C}(n) e^{0.01(n+1)T}}{m^n}. \end{aligned}$$

Finally, notice that $P_m(t) = e^{-\gamma t} Q_m(e^{-\alpha t}) = \sum_{k=0}^{m-1} \alpha_k e^{-(0.01+0.01k)t}$, so we can select $\phi_m^{\text{exp}}(t) := P_m(t)$.

□

Lemma F.2 (Exp decay, adaptive Delta function). *For any $T \in \mathbb{N}$, $n, m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{exp}}(t; B) = \sum_{k=1}^m \alpha_k e^{-\beta_k(t-B)}$ such that*

$$\max_{1 \leq B \leq T} \|\mathbb{I}(\cdot = B) - \phi_m^{\text{exp}}(\cdot; B)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n) e^{0.01(n+1)T}}{m^n}.$$

where $\beta_k > 0$ holds for any $k \in [m]$.

Proof of Lemma F.2.

The **key point** of the proof is to note that the adaptability of B can be eliminated by the **translation operator** $t - B$.

First, recall our proof of Lemma F.1. For the same $\Psi_T(\cdot)$, for any $n, m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0, 1]} |\Psi_T(x) - Q_m(x)| \leq \frac{C(n) e^{0.01(n+1)T}}{m^n}.$$

Moreover, using the transform $x = e^{-0.01(t-B+T)}$ ($t \geq 0$) on the function Ψ and consider

$$\Phi_T(t; B) := e^{-0.01(t-B+T)} \Psi_T\left(e^{-0.01(t-B+T)}\right), \quad t \in [0, +\infty).$$

It is easy to verify that $\Phi_T(\cdot; B)$ satisfies that

$$\Phi_T(t; B)|_{\mathbb{N}} = \mathbb{I}(t = B).$$

And we consider the function

$$P_m(t; B) := e^{-0.01(t-B+T)} Q_m\left(e^{-0.01(t-B+T)}\right), \quad t \in [0, +\infty).$$

Then, for any $1 \leq B \leq T$, the following error estimate holds:

$$\begin{aligned}
& \|P_m(\cdot; B) - \mathbb{I}(\cdot = B)\|_{\ell_1(\mathbb{N})} = \sum_{t=0}^{+\infty} |P_m(t; B) - \Phi_T(t; B)| \\
& = \sum_{t=0}^{+\infty} e^{-0.01(t-B+T)} \left| Q_m \left(e^{-0.01(t-B+T)} \right) - \Psi_T \left(e^{-0.01(t-B+T)} \right) \right| \\
& \leq \sum_{t=0}^{+\infty} e^{-0.01t} \sup_{x \in [0,1]} |Q_m(x) - \Psi_T(x)| \\
& \leq \frac{C(n)e^{0.01(n+1)T}}{m^n} \sum_{t=0}^{+\infty} e^{-0.01t} = \frac{\tilde{C}(n)e^{0.01(n+1)T}}{m^n}.
\end{aligned}$$

Due to the arbitrariness of B , the proof is completed. \square

Lemma F.3 (Exp decay, fixed Delta function). *Consider a exponentially decayed memory $\rho(\cdot)$: there exists $\beta > 0$ such that $\rho(t) = \mathcal{O}(e^{-\beta t})$. Then for any $n \in [[99\beta]]$ and $m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{exp}}(t) = \sum_{k=1}^m \alpha_k e^{-\beta_k t}$ such that*

$$\|\rho(\cdot) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)}{m^n},$$

where $\beta_k > 0$ holds for any $k \in [m]$.

Proof of Lemma F.3.

There exists $C > 0$ such that $|\rho(t)| \leq C e^{-\beta t}$.

Let $\alpha, \gamma > 0$ be constants, and they will take specific values at the end of the proof.

First, recall the standard bump function on $[-1, 1]$:

$$\Psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & x \in (-1, 1) \\ 0, & \text{otherwise} \end{cases},$$

and we can define the following constants for $T \geq 1$:

$$\mu_T = e^{-\alpha T}, \quad \sigma_T = \frac{1}{2} \left(e^{-\alpha T} - e^{-\alpha(T+1)} \right),$$

and we consider the following bump function $\Psi_T \in \mathcal{C}^\infty([0, 1])$:

$$\Psi_T(x) = \begin{cases} V_T \Psi\left(\frac{x-\mu_T}{\sigma_T}\right), & x \in (\mu_T - \sigma_T, \mu_T + \sigma_T) \\ 0, & \text{otherwise} \end{cases},$$

where V_T is a scaling constant such that $\Psi_T(e^{-\alpha T}) = e^{\gamma T} \rho(T)$.

Consequently, we consider the sum of bump functions on $[0, 1]$:

$$\varphi(x) := \sum_{T=1}^{+\infty} \Psi_T(x).$$

It is easy to verify that $(\mu_{T_1} - \sigma_{T_1}, \mu_{T_1} + \sigma_{T_1}) \cap (\mu_{T_2} - \sigma_{T_2}, \mu_{T_2} + \sigma_{T_2}) = \emptyset$ for any $T_1 \neq T_2$ and

$$\varphi(x) = \begin{cases} \Psi_T(x), & \mu_T - \sigma_T \leq x \leq \mu_T + \sigma_T \\ 0, & \text{otherwise} \end{cases}.$$

First, we study the property of $\varphi(\cdot)$.

We denote the absolute constants $M_k = \sup_x |\varphi^{(k)}(x)|$. Notice that for any $k \in \mathbb{N}$,

$$\Psi_T^{(k)}(x) = \frac{V_T}{\sigma_T^k} \Psi^{(k)}\left(\frac{x - \mu_T}{\sigma_T}\right).$$

Therefore, it holds that

$$\begin{aligned} \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\varphi^{(k)}(x)| &= \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\Psi_T^{(k)}(x)| \\ &\leq \frac{V_T}{\sigma_T^k} M_k = \frac{e^{\gamma T} \rho(T)}{(e^{-\alpha T} - e^{-\alpha(T+1)})^k} M_k e \leq \frac{C M_k e}{(1 - e^{-\alpha})^k} e^{(\gamma + k\alpha - \beta)T}. \end{aligned}$$

Therefore, if $\beta \geq \gamma + k\alpha$, then the following uniform bounds holds:

$$\begin{aligned} \sup_{x \in (0,1]} |\varphi^{(k)}(x)| &= \sup_{T \geq 1} \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\varphi^{(k)}(x)| \\ &\leq \sup_{T \geq 1} \frac{C M_k e}{(1 - e^{-\alpha})^k} e^{(\gamma + k\alpha - \beta)T} \leq \frac{C M_k e}{(1 - e^{-\alpha})^k} := C(k, \alpha). \end{aligned}$$

Consequently, we consider the smoothness of Φ at $x = 0$.

Recalling the previous results, for any $x \in (0, 1]$, we have

$$\begin{aligned} \frac{|\varphi^{(k)}(x)|}{x} &\leq C(k, \alpha) \frac{e^{(\gamma + k\alpha - \beta)T}}{\mu_T - \sigma_T} = \frac{2C(k, \alpha)}{1 - e^{-\alpha}} e^{(\gamma + (k+1)\alpha - \beta)T}, \quad x \in (\mu_T - \sigma_T, \mu_T + \sigma_T); \\ \frac{|\varphi^{(k)}(x)|}{x} &= 0, \text{ otherwise} \end{aligned}$$

Thus, by induction, it is easy to verify that for any $i < \frac{\beta - \gamma}{\alpha}$ ($i \in \mathbb{N}$),

$$\varphi^{(i)}(0) = 0.$$

Therefore, for any $n < \frac{\beta - \gamma}{\alpha}$ ($n \in \mathbb{N}$), $\varphi^{(k)}(0) = 0$ holds for any $0 \leq k \leq n$. Moreover, there exists absolute constant $C(n, \alpha)$ such that:

$$\max_{0 \leq k \leq n} \sup_{x \in [0,1]} |\varphi^{(k)}(x)| \leq C(n, \alpha).$$

By Lemma G.5, for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0,1]} |\varphi(x) - Q_m(x)| \leq \frac{C(n, \alpha)}{m^n}.$$

Now we use the transform $x = e^{-\alpha t}$ ($t \geq 0$) on the function φ and consider

$$\Phi(t) := \frac{1}{e^{\gamma t}} \varphi\left(\frac{1}{e^{\alpha t}}\right), \quad t \in [0, +\infty).$$

It is easy to verify that Φ satisfies that

$$\Phi(t)|_{\mathbb{N}} = \rho(t)|_{\mathbb{N}}.$$

Moreover, we consider the function

$$P_m(t) := \frac{1}{e^{\gamma t}} Q_m\left(\frac{1}{e^{\alpha t}}\right), \quad t \in [0, +\infty).$$

Then for any $n < \frac{\beta-\gamma}{\alpha}$ ($n \in \mathbb{N}$), the following error estimate holds:

$$\begin{aligned} \|P_m(\cdot) - \rho(\cdot)\|_{\ell_1(\mathbb{N})} &= \sum_{t=0}^{+\infty} |P_m(t) - \Phi(t)| \\ &= \sum_{t=0}^{+\infty} e^{-\gamma t} |Q_m(e^{-\alpha t}) - \Psi_T(e^{-\alpha t})| \leq \frac{C(n, \alpha)}{m^n} \sum_{t=0}^{+\infty} e^{-\gamma t}. \end{aligned}$$

By choosing $\alpha = 5 \cdot 10^{-3}$ and $\gamma = 10^{-2}\beta$, it holds that $99\beta < \frac{\beta-\gamma}{2\alpha} = \frac{\beta-\gamma}{\alpha}$.

Thus, we obtain our result: for any $n \in [[99\beta]]$ ($\beta \geq 1/99$), the following error estimate holds:

$$\|P_m(\cdot) - \rho(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{C(n)}{m^n} \sum_{t=0}^{+\infty} e^{-\gamma t} = \frac{C(n)}{m^n} \frac{1}{1 - e^{-10^{-2}\beta}} = \frac{\tilde{C}(n)}{m^n}.$$

□

F.2 Approximation by the sum of polynomial decay

Lemma F.4 (Poly decay, fixed Delta function). *For any $T, n, m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{poly}}(t) = \sum_{k=1}^m \alpha_k t^{-\beta_k}$ such that*

$$\|\mathbb{I}(\cdot = T) - \phi_m^{\text{poly}}(\cdot)\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)T^{1.01(n+1)}}{m^n},$$

where $\beta_k > 1$ holds for any $k \in [m]$.

Proof of Lemma F.4.

Let $\alpha, \gamma > 0$ be constants, and they will take specific values at the end of the proof

First, recall the standard bump function on $[-1, 1]$:

$$\Psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & x \in (-1, 1) \\ 0, & \text{otherwise} \end{cases},$$

and we can define the following constants for $T \geq 1$:

$$\mu_T = \frac{1}{T^\alpha}, \quad \sigma_T = \frac{1}{T^\alpha} - \frac{1}{(T+1)^\alpha}.$$

Then we consider the following bump function $\Psi_T \in \mathcal{C}^\infty([0, 1])$:

$$\Psi_T(x) = \begin{cases} V_T \Psi\left(\frac{x-\mu_T}{\sigma_T}\right), & x \in (\mu_T - \sigma_T, \mu_T + \sigma_T) \\ 0, & \text{otherwise} \end{cases},$$

where V_T is a scaling constant such that $\Psi_T\left(\frac{1}{T^\alpha}\right) = T^{1+\gamma}$.

First, we consider the approximation of Ψ_T on $[0, 1]$.

Notice that $\Psi_T \in \mathcal{C}^\infty([0, 1])$, and $\Psi_T^{(k)}(0) = 0$ for any $k \in \mathbb{N}$. For the standard bump function Ψ , for any $n \in \mathbb{N}_+$, there exists an absolute constant $M(n) > 0$ only depending on n , such that

$$\max_{0 \leq k \leq n} \sup_{x \in [-1, 1]} |\Psi^{(k)}(x)| \leq M(n).$$

Notice that for any $k \in \mathbb{N}$ and $x \in [0, 1]$,

$$\Psi_T^{(k)}(x) = \frac{V_T}{\sigma_T^k} \Psi^{(k)}\left(\frac{x-\mu_T}{\sigma_T}\right).$$

Therefore, the following upper bound holds:

$$\begin{aligned} M_T(n) &= \max_{0 \leq k \leq n} \frac{V_T}{\sigma_T^k} M(n) = \frac{V_T}{\sigma_T^n} M(n) = \frac{T^{1+\gamma} e}{(1/T^\alpha - 1/(T+1)^\alpha)^n} M(n) \\ &\leq \frac{T^{1+\gamma} (T+1)^{n(1+\alpha)} M(n) e}{\alpha^n} \leq \frac{2^n M(n) e}{\alpha^n} T^{1+\gamma+n(1+\alpha)} := C(n, \alpha) T^{1+\gamma+n(1+\alpha)}. \end{aligned}$$

By Lemma G.5, for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0,1]} |\Psi_T(x) - Q_m(x)| \leq \frac{M_T(n)}{m^n} \leq \frac{C(n, \alpha) T^{1+\gamma+n(1+\alpha)}}{m^n}.$$

Now we use the transform $x = \frac{1}{t^\alpha}$ ($t \geq 1$) on the function Ψ and consider

$$\Phi_T(t) := \frac{1}{t^{1+\gamma}} \Psi_T\left(\frac{1}{t^\alpha}\right), \quad t \in [1, +\infty).$$

It is easy to verify that Φ_T satisfies that

$$\Phi_T(t)|_{\mathbb{N}_+} = \mathbb{I}(t = T).$$

Moreover, we consider the function

$$P_m(t) := \frac{1}{t^{1+\gamma}} Q_m\left(\frac{1}{t^\alpha}\right), \quad t \in [1, +\infty).$$

Then by choosing $\alpha = \gamma = 0.01$, the following error estimate holds:

$$\begin{aligned} \|P_m(\cdot) - \mathbb{I}(\cdot = T)\|_{\ell_1(\mathbb{N}_+)} &= \sum_{t=1}^{+\infty} |P_m(t) - \Phi_T(t)| \\ &= \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}} \left| Q_m\left(\frac{1}{t^\alpha}\right) - \Psi_T\left(\frac{1}{t^\alpha}\right) \right| \leq \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}} \frac{M_T(n)}{m^n} \\ &\leq \frac{C(n, \alpha) T^{1+\gamma+n(1+\alpha)}}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}} = \frac{C(n) T^{1.01(n+1)}}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1+0.01}} \\ &= \frac{\tilde{C}(n) T^{1.01(n+1)}}{m^n}. \end{aligned}$$

Finally, notice that $P_m(\cdot)$ satisfies to $P_m(t) = \frac{1}{t^{1+\gamma}} Q_m\left(\frac{1}{t^\alpha}\right) = \sum_{k=0}^{m-1} \alpha_k t^{-(1.01+0.01k)}$, so we can select $\phi_m^{\text{poly}}(t) := P_m(t)$.

□

Lemma F.5 (Poly decay, adaptive Delta function). *For any $T, n, m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{poly}}(t; B) = \sum_{k=1}^m \alpha_k (t/B)^{-\beta_k}$ such that*

$$\max_{1 \leq B \leq T} \|\mathbb{I}(\cdot = B) - \phi_m^{\text{poly}}(\cdot; B)\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n) T^{1.01(n+1)}}{m^n},$$

where $\beta_k > 1$ holds for any $k \in [m]$.

Proof of Lemma F.5.

The **key point** of the proof is to note that the adaptability of B can be eliminated by the **rescaling operator** t/B .

First, recall our proof of Lemma F.4. For the same $\Psi_T(\cdot)$, for any $n, m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0,1]} |\Psi_T(x) - Q_m(x)| \leq \frac{C(n)T^{1.01(n+1)}}{m^n}.$$

We use the transform $x = \frac{1}{t^{0.01}}$ ($t \geq 1$) on the function Ψ and consider

$$\Phi_T(t; B) := \left(\frac{B}{tT}\right)^{1.01} \Psi_T\left(\left(\frac{B}{tT}\right)^{0.01}\right), \quad t \in [1, +\infty).$$

It is easy to verify that $\Phi_T(\cdot; \cdot)$ satisfies that

$$\Phi_T(t; B)|_{\mathbb{N}_+} = \mathbb{I}(t = B).$$

And we consider the function

$$P_m(t; B) := \left(\frac{B}{tT}\right)^{1.01} Q_m\left(\left(\frac{B}{tT}\right)^{0.01}\right), \quad t \in [1, +\infty).$$

Then, for any $1 \leq B \leq T$, the following error estimate holds:

$$\begin{aligned} \|P_m(\cdot; B) - \mathbb{I}(\cdot = B)\|_{\ell_1(\mathbb{N}_+)} &= \sum_{t=1}^{+\infty} |P_m(t; B) - \Phi_T(t; B)| \\ &= \sum_{t=1}^{+\infty} \left(\frac{B}{tT}\right)^{1.01} \left| Q_m\left(\left(\frac{B}{tT}\right)^{0.01}\right) - \Psi_T\left(\left(\frac{B}{tT}\right)^{0.01}\right) \right| \\ &\leq \sum_{t=1}^{+\infty} \frac{1}{t^{1.01}} \sup_{x \in [0,1]} |Q_m(x) - \Psi_T(x)| \\ &\leq \frac{C(n)T^{1.01(n+1)}}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1.01}} = \frac{\tilde{C}(n)T^{1.01(n+1)}}{m^n}. \end{aligned}$$

Due to the arbitrariness of B , the proof is completed. □

Lemma F.6 (Poly decay, fixed Delta function). *Consider a polynomially decayed memory $\rho(\cdot)$: there exists $\beta > 1$ such that $\rho(t) = \mathcal{O}(t^{-\beta})$. Then for any $n \in [\lceil 0.99\beta \rceil - 1]$ and $m \in \mathbb{N}_+$, there exists an absolute constant $C(n)$ only depending on n and a $\phi_m^{\text{poly}}(t) = \sum_{k=1}^m \alpha_k t^{-\beta k}$ such that*

$$\|\rho(\cdot) - \phi_m^{\text{poly}}(\cdot)\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)}{m^n},$$

where $\beta_k > 1$ holds for any $k \in [m]$.

Proof of Lemma F.6.

There exists $C > 0$ such that $|\rho(t)| \leq C/t^\beta$.

Let $\alpha, \gamma > 0$ be constants, and they will take specific values at the end of the proof

First, recall the standard bump function on $[-1, 1]$:

$$\Psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & x \in (-1, 1) \\ 0, & \text{otherwise} \end{cases},$$

and we can define the following constants for $T \geq 1$:

$$\mu_T = \frac{1}{T^\alpha}, \quad \sigma_T = \frac{1}{2} \left(\frac{1}{T^\alpha} - \frac{1}{(T+1)^\alpha} \right),$$

and we consider the following bump function $\Psi_T \in \mathcal{C}^\infty([0, 1])$:

$$\Psi_T(x) = \begin{cases} V_T \Psi \left(\frac{x - \mu_T}{\sigma_T} \right), & x \in (\mu_T - \sigma_T, \mu_T + \sigma_T) \\ 0, & \text{otherwise} \end{cases},$$

where V_T is a scaling constant such that $\Psi_T(\frac{1}{T^\alpha}) = T^{1+\gamma} \rho(T)$.

Consequently, we consider the sum of bump functions on $[0, 1]$:

$$\varphi(x) := \sum_{T=1}^{+\infty} \Psi_T(x).$$

It is easy to verify that $(\mu_{T_1} - \sigma_{T_1}, \mu_{T_1} + \sigma_{T_1}) \cap (\mu_{T_2} - \sigma_{T_2}, \mu_{T_2} + \sigma_{T_2}) = \emptyset$ for any $T_1 \neq T_2$ and

$$\varphi(x) = \begin{cases} \Psi_T(x), & \mu_T - \sigma_T \leq x \leq \mu_T + \sigma_T \\ 0, & \text{otherwise} \end{cases}.$$

First, we study the property of $\varphi(\cdot)$.

We denote the absolute constants $M_k = \sup_x |\varphi^{(k)}(x)|$. Notice that for any $k \in \mathbb{N}$,

$$\Psi_T^{(k)}(x) = \frac{V_T}{\sigma_T^k} \Psi^{(k)} \left(\frac{x - \mu_T}{\sigma_T} \right).$$

Therefore, it holds that

$$\begin{aligned} \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\varphi^{(k)}(x)| &= \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\Psi_T^{(k)}(x)| \\ &\leq \frac{V_T}{\sigma_T^k} M_k = \frac{T^{1+\gamma} \rho(T)}{\left(\frac{1}{T^\alpha} - \frac{1}{(T+1)^\alpha} \right)^k} 2^k M_k e \\ &\leq \frac{(T+1)^{k(1+\alpha)} T^{1+\gamma-\beta} C 2^k M_k e}{\alpha^k} \leq \frac{2^{k(2+\alpha)} C M_k e}{\alpha^k} T^{1+\gamma+k(1+\alpha)-\beta}. \end{aligned}$$

Therefore, if $k \leq \frac{\beta-(1+\gamma)}{1+\alpha}$, the following uniform bounds hold:

$$\begin{aligned} \sup_{x \in (0,1]} |\varphi^{(k)}(x)| &= \sup_{T \geq 1} \sup_{x \in (\mu_T - \sigma_T, \mu_T + \sigma_T)} |\varphi^{(k)}(x)| \\ &\leq \sup_{T \geq 1} \frac{2^{k(2+\alpha)} C M_k e}{\alpha^k} T^{1+\gamma+k(1+\alpha)-\beta} \leq \frac{2^{k(2+\alpha)} C M_k e}{\alpha^k} := C(k, \alpha). \end{aligned}$$

Consequently, we consider the smoothness of Φ at $x = 0$.

Recalling the previous results, for any $x \in (0, 1]$, we have

$$\begin{aligned} \frac{|\varphi^{(k)}(x)|}{x} &\leq C(k, \alpha) \frac{T^{1+\gamma+k(1+\alpha)-\beta}}{\mu_T - \sigma_T} \leq \frac{C(k, \alpha) 2^{2+\alpha}}{\alpha} T^{1+\gamma+(k+1)(1+\alpha)-\beta}, \quad x \in (\mu_T - \sigma_T, \mu_T + \sigma_T); \\ \frac{|\varphi^{(k)}(x)|}{x} &= 0, \quad \text{otherwise} \end{aligned}$$

Thus, by induction, it is easy to verify that for any $i < \frac{\beta-(1+\gamma)}{1+\alpha}$ ($i \in \mathbb{N}$),

$$\varphi^{(i)}(0) = 0.$$

Therefore, for any $n < \frac{\beta-(1+\gamma)}{1+\alpha}$ ($n \in \mathbb{N}_+$), $\varphi^{(k)}(0) = 0$ holds for any $0 \leq k \leq n$. Moreover, the following uniform bound holds:

$$\max_{0 \leq k \leq n} \sup_{x \in [0,1]} |\varphi^{(k)}(x)| \leq C(n, \alpha).$$

By Lemma G.5, for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that

$$\sup_{x \in [0,1]} |\varphi(x) - Q_m(x)| \leq \frac{C(n, \alpha)}{m^n}.$$

Now we use the transform $x = \frac{1}{t^\alpha}$ ($t \geq 1$) on the function φ and consider

$$\Phi(t) := \frac{1}{t^{1+\gamma}} \varphi\left(\frac{1}{t^\alpha}\right), \quad t \in [1, +\infty).$$

It is easy to verify that Φ satisfies that

$$\Phi(t)|_{\mathbb{N}_+} = \rho(t)|_{\mathbb{N}_+}.$$

Moreover, we consider the function

$$P_m(t) := \frac{1}{t^{1+\gamma}} Q_m\left(\frac{1}{t^\alpha}\right), \quad t \in [1, +\infty).$$

Then for any $n < \frac{\beta-(1+\gamma)}{1+\alpha}$ ($n \in \mathbb{N}$), the following error estimate holds:

$$\begin{aligned} \|P_m(\cdot) - \rho(\cdot)\|_{\ell_1(\mathbb{N}_+)} &= \sum_{t=1}^{+\infty} |P_m(t) - \Phi(t)| \\ &= \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}} \left| Q_m\left(\frac{1}{t^\alpha}\right) - \Psi_T\left(\frac{1}{t^\alpha}\right) \right| \leq \frac{C(n, \alpha)}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}}. \end{aligned}$$

By choosing $\alpha = 10^{-2}$ and $\gamma = 10^{-4}\beta$, we have $0.99\beta - 1 = \frac{\beta-\gamma}{1+\alpha} - 1 = \frac{\beta-(1+\gamma+\alpha)}{1+\alpha} < \frac{\beta-(1+\gamma)}{1+\alpha}$. Thus, we obtain our result: for any $n \in \lceil [0.99\beta] - 1 \rceil$ ($\beta \geq 2/0.99$), the following error estimate holds:

$$\|P_m(\cdot) - \rho(\cdot)\|_{\ell_1(\mathbb{N}_+)} \leq \frac{C(n)}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1+\gamma}} \leq \frac{C(n)}{m^n} \sum_{t=1}^{+\infty} \frac{1}{t^{1+10^{-4}}} = \frac{\tilde{C}(n)}{m^n}.$$

□

G Some Background and Proof Preparation

G.1 T5’s relative positional encoding

The T5’s Relative Positional Encoding is primary focus of this study. Its standard form in practical applications (Raffel et al., 2020) adheres to $R_{t,s} = r(t - s)$, where

$$-r(n) = \begin{cases} n, & \text{if } n < \mathcal{B} \\ \mathcal{B} + \lfloor \mathcal{B} \cdot \frac{\log(n/\mathcal{B})}{\log(\mathcal{D}/\mathcal{B})} \rfloor, & \text{if } \mathcal{B} \leq n < \mathcal{D} . \\ 2\mathcal{B} - 1, & \text{if } n \geq \mathcal{D} \end{cases}$$

Here, \mathcal{D} is a large integer, signifying the longest distance of concern, while \mathcal{B} is a small integer. One can see that for $n < \mathcal{B}$, $r(\cdot)$ exhibits polynomial decay, whereas for $\mathcal{B} < n < \mathcal{D}$, $r(\cdot)$ demonstrates logarithmic decay. Consequently, the *overall decay rate* of $r(\cdot)$ is *logarithmic*.

The following Table further provides an example of standard T5’s Relative Positional Encoding.

Table 1: An example of standard T5’s Relative Positional Encoding

$t - s$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$-r(t - s)$	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9
$t - s$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	...
$-r(t - s)$	10	10	10	10	10	10	10	11	11	11	11	11	11	11	11	...

G.2 Barron space theory

The well-known universal approximation result for 2NNs asserts that 2NNs can approximate any continuous function (Barron, 1992; 1993; 1994). Nonetheless, this result lacks a characterization of the approximation efficiency, i.e., how many neurons are needed to achieve a certain approximation accuracy? This gap was addressed by the Barron space theory (E et al., 2019; 2021). It is established that for any function within Barron space $f \in \mathcal{B}$, 2NNs with m neurons (denoted by \mathcal{H}_m) can approximate them efficiently, at a rate of $\inf_{f_m \in \mathcal{H}_m} \|f - f_m\| \leq \mathcal{O}(\|f\|_{\mathcal{B}} / \sqrt{m})$, remarkably independent of the input dimension d , thus avoiding the *Curse of Dimensionality* (Bellman, 1966; Bach, 2017). Specifically, the Barron space is defined by:

Definition G.1 (Barron space (E et al., 2019; 2021; Ma et al., 2020)). Consider functions $f : X \rightarrow \mathbb{R}$ that admit the following representation:

$$f(\mathbf{x}) = \int_{\Omega} a \sigma(\mathbf{b}^{\top} \mathbf{x} + c) \rho(\mathrm{d}a, \mathrm{d}\mathbf{b}, \mathrm{d}c), \quad \mathbf{x} \in X.$$

For any $p \in [1, +\infty]$, we define the Barron norm:

$$\|f\|_{\mathcal{B}_p} := \inf_{\rho} \left(\mathbb{E}_{\rho} [|a|^p (\|\mathbf{b}\|_1 + |c|)^p] \right)^{1/p}.$$

Then the Barron space are defined as:

$$\mathcal{B}_p := \{f \in \mathcal{C} : \|f\|_{\mathcal{B}_p} < +\infty\}.$$

Proposition G.2. For any $p \in [1, +\infty]$, $\mathcal{B}_p = \mathcal{B}_{\infty}$ and $\|f\|_{\mathcal{B}_p} = \|f\|_{\mathcal{B}_{\infty}}$.

Remark G.3. From the Proposition above, the Barron spaces \mathcal{B}_p are equivalent for any $p \in [1, +\infty]$. Consequently, in this paper, we use \mathcal{B} and $\|\cdot\|_{\mathcal{B}}$ to denote the Barron space and Barron norm.

Remark G.4. For Barron space \mathcal{B} , both Direct and Inverse Approximation Theorems hold (E et al., 2021). In this paper, we mainly utilize the Direct Approximation Theorem, stated in Lemma G.6.

G.3 Useful approximation lemmas

Lemma G.5 (Jackson (1930)). *Let $f \in \mathcal{C}^n([0, 1])$ with $f(0) = f'(0) = \dots = f^{(n)}(0) = 0$. Then for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_m(x) = \sum_{k=0}^{m-1} \alpha_k x^k$ such that*

$$\|f - Q_m\|_{L^\infty([0,1])} \leq \frac{M(n)}{m^n},$$

where $M(n) = \max_{k \leq n} \|f^{(k)}\|_{L^\infty([0,1])}$.

Lemma G.6 (Ma et al. (2020)). *For any $f \in \mathcal{B}$ and $m \in \mathbb{N}$, there exists a two-layer ReLU neural network $f_m(\mathbf{x}) = \sum_{k=1}^m a_k \sigma(\mathbf{b}_k^\top \mathbf{x} + c_k)$ with m neurons such that*

$$\|f - f_m\|_{L^\infty([0,1]^d)} \leq \tilde{\mathcal{O}}\left(\frac{\|f\|_{\mathcal{B}}}{\sqrt{m}}\right).$$

H Experiments

H.1 Restatement of our theoretical insights

As detailed in Section 1, our theoretical analysis reveals the following novel insights into the expressive power and mechanisms of Transformer:

Insight (1). The distinct roles of the number of layers, the number of Attn heads, and the width of FFN layers. (1a) Deeper Transformers can handle tasks with memories with more intricate interrelationships, such as nested relationships (Type II). **(1b)** In contrast, for tasks with memories lacking such interrelationships (Type I), a single-layer Transformer with sufficient Attn heads and FFN width should suffice.

Insight (2). The different roles of Attn layers and FFN layers. (2a) FFN layers are tasked with approximating nonlinear memory functions and the readout function, **(2b)** while Attn layers are responsible for extracting the tokens from the memory locations.

Insight (3). The functionality and necessity of Dot-product (DP). (3a) For the relatively simple Task I, DP is not necessary and can be omitted. **(3b)** However, for the more complex Task II, DP provides necessary nonlinearity: the cooperation between DP and RPE provides the needed interaction between the temporal space and the token space.

Insight (4). The efficiency of Relative Positional Encoding (RPE) in modeling long-range correlations. The primary role of RPE is to approximate the memory kernels. **(4a)** Transformer with log-type RPE can handle heavy-tailed memories. **(4b)** Transformer with lin-type RPE can handle light-tailed memories.

H.2 Experimental Validation

To validation of our theoretical **insights (1a)~(4b)**, we conduct 8 experiments, from simple toy models to more complex LLM pre-training. The experiments are conducted on 1 A100.

H.2.1 Validation of Insight (1a)

Objective. As indicated in Section 4, numerous NLP tasks exhibit complex interrelationships among tokens and belong to our Task II. This experiment aims to verify our Insight (1a): for such tasks, increasing the number of layers L is more efficient than increasing the number of Attn heads H .

Setup. Specifically, we pretrain decoder-only Transformers (Vaswani et al., 2017) with different L and H on the OpenWebText dataset (Gokaslan and Cohen, 2019) for 10,000 iterations (approximately 1B tokens) on 1 A100, using cross-entropy loss and AdamW with the same hyperparameters. To ensure comparability, we meticulously balance the total number of parameters across both experimental setups.

Results and conclusion. The final validation losses are shown in Table 2. By comparing these two subtables, the benefits brought by increasing L are much greater than the benefits brought by increasing H ($0.802 > 0.136$), thereby corroborating our Insight (1a).

Table 2: Results of the experiment supporting Insight (1a).

$L = 1, H = 8$ (26M)	$L = 1, H = 12$ (29M)	$L = 1, H = 12$ (32M)
5.796 (baseline)	5.689 (\downarrow 0.107)	5.660 (\downarrow 0.136)
$L = 1, H = 8$ (26M)	$L = 4, H = 8$ (29M)	$L = 8, H = 8$ (32M)
5.796 (baseline)	5.374 (\downarrow 0.422)	4.994 (\downarrow 0.802)

H.2.2 Validation of Insight (1b)

Objective. As mentioned in Section 3, sparse Boolean functions have no interactions among the memories and belong to our Task I. This experiment aims to verify our Insight (1b): for such tasks, a single-layer Transformer equipped with a sufficient number of Attn heads H and FFN width m suffices, and there is no need to increase the number of layers L .

Setup. Specifically, we train single-layer DP-free Transformers with different H and m to learn a sparse Boolean target function f^* : $f^*(\mathbf{x}) := g^*(x_{48}, x_{56}, x_{99}) := \sum_{k=1}^{64} \text{ReLU}(\langle \mathbf{w}_k^*, (x_{48}, x_{56}, x_{99}) \rangle)$ for input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$, where \mathbf{w}_k^* are generated by $\mathbf{w}_k^* \sim N(0, I_3)$. Training proceeds for 10,000 iterations (1M samples) using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are shown in Table 3. As shown in this table, a single-layer Transformer equipped with a sufficient H (32) and m (256) is adequate for representing this sparse Boolean function. This empirical evidence supports our Insight (1b).

Table 3: Results of the experiment supporting Insight (1b).

$H = 2, m = 16$	$H = 8, m = 64$	$H = 32, m = 256$
0.21	0.04	0.01

H.2.3 Validation of Insight (2a)

Objective. This experiment aims to verify our Insight (2a): to learn a sparse Boolean function with a “complex” readout function and “simple” memories, increasing the FFN width m can significantly improve the performance, whereas increasing the number of Attn heads H brings almost no benefit.

Setup. Specifically, we train single-layer DP-free Transformers with different H and m to learn a sparse Boolean function with a “complex” readout function (g^*) and a “simple” single memory (x_{99}): $f^*(\mathbf{x}) := g^*(x_{99}) := \sum_{k=1}^{64} \text{ReLU}(\mathbf{w}_k^* \cdot x_{99})$ for any input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$, where \mathbf{w}_k^* are generated by $\mathbf{w}_k^* \sim N(0, 1)$. Training proceeds for 10,000 iterations (1M samples) using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are shown in Table 4. The tables indicate that, for learning a sparse Boolean function with a “complex” readout function and “simple” memories, increasing m can significantly improve the performance (0.49 \rightarrow 0.002), almost completing this task perfectly. Conversely, increasing H fails to yield substantial improvement. This empirical evidence supports our Insight (2a).

Table 4: Results of the experiment supporting Insight (2a).

	$m = 8$	$m = 64$	$m = 512$		$H = 8$	$H = 64$	$H = 512$
$H = 8$	0.49	0.006	0.002	$m = 8$	0.49	0.49	0.52

H.2.4 Validation of Insight (2b)

Objective. Contrasting with Experiment (2a), this experiment aims to verify our Insight (2b): for learning a sparse Boolean function with a “simple” readout function and “complex” memories, increasing the number of Attn headers H can substantially improve the performance while increasing FFN width m will offer almost no benefit.

Setup. Specifically, we train single-layer DP-free Transformers with different H and m to learn a sparse Boolean function with a “simple” linear readout function (g^*) and relatively “complex” memories (x_{48}, x_{56}, x_{99}): $f^*(\mathbf{x}) := g^*(x_{48}, x_{56}, x_{99}) := x_{48} + x_{56} + x_{99}$ for any input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$. Training processes for 10,000 iterations (1M samples), using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are presented in Table 5. The tables indicate that, for learning a sparse Boolean function with a “simple” readout function and “complex” memories, increasing m can significantly improve the performance (1.16 $\rightarrow 10^{-6}$), closely achieving task perfection. In contrast, increasing m brings almost no benefits. This empirical evidence supports our Insight 2(b).

H.2.5 Validation of Insight (3a)

Objective. As mentioned in Section 3, learning sparse Boolean functions has no interactions among the memories and belongs to our Task I. This experiment aims to verify our insight (3a): for such

Table 5: Results of the experiment supporting Insight (2b).

	$m = 2$	$m = 64$	$m = 256$		$H = 2$	$H = 64$	$H = 256$
$H = 2$	1.16	0.81	1.23	$m = 2$	1.16	<1e-6	<1e-6

tasks, a DP-free Transformer equipped with a sufficient number of Attn heads H and FFN width m is sufficiently capable. Moreover, there is no need to use DP structure in Attn.

Setup. Specifically, we train single-layer DP-free Transformers with different H and m and with DP or without DP to learn a sparse Boolean target function $f^*: f^*(\mathbf{x}) := g^*(x_{48}, x_{56}, x_{99}) := \sum_{k=1}^{64} \text{ReLU}(\langle \mathbf{w}_k^*, (x_{48}, x_{56}, x_{99}) \rangle)$ for input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$, where \mathbf{w}_k^* are generated by $\mathbf{w}_k^* \sim N(0, I_3)$. Training proceeds for 10,000 iterations (1M samples) using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are shown in Table 6. The findings illustrate that a DP-free Transformer equipped with a sufficient H (32) and m (256) is adept at accurately representing the given sparse Boolean function. Additionally, the Incorporation of the DP structure into the layers contributes marginally to performance enhancement. This substantiates our Insight (3a).

Table 6: Results of the experiment supporting Insight (3a).

	$H = 2, m = 16$	$H = 8, m = 64$	$H = 32, m = 256$
with DP	0.21	0.04	0.01
without DP	0.17	0.11	0.02

H.2.6 Validation of Insight (3b)

Objective. As indicated in Section 4, numerous NLP tasks exhibit complex interrelationships among tokens and belong to our Task II. This experiment aims to verify our Insight (3b): for such tasks, the utilization of DP structure in Attn layers is necessary.

Setup. Specifically, we pre-train Transformers with DP or without DP on the OpenWebText dataset for 10,000 iterations (approximately 1B tokens) on 1 A100, using cross-entropy loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are presented in Table 7. As shown in the table, for NLP pre-training tasks, Transformer incorporating DP structure is more efficient than Transformer without DP ($5.796 < 5.830, 5.374 < 5.486, 4.994 < 5.274$), thereby supporting our Insight 3(b).

Table 7: Results of the experiment supporting Insight (3b).

	$L = 1, H = 8$	$L = 4, H = 8$	$L = 8, H = 8$
with DP	5.796	5.374	4.994
without DP	5.830	5.486	5.274

H.2.7 Validation of Insight (4a)

Objective. This experiment aims to verify our Insight (4a): for learning Task III with heavy-tailed memories, Transformers with log-type RPE are efficient, whereas those with lin-type RPE fail.

Setup. Specifically, we train single-layer, FFN-free, DP-free Transformers with log-type RPE or lin-type RPE and varying numbers of Attn heads H . The target function involves a heavy-tailed memory kernel $\rho(t) = t^{-0.5}$: $f^*(\mathbf{x}) := \sum_{s=1}^{1000} x_s \rho(1000 - s)$ for any input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$. Training processes for 10,000 iterations (1M samples) using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are shown in Table 8. As shown in the table, to learn heavy-tailed memories, even single-head Transformer with log-type RPE can complete it

perfectly ($< 10^{-5}$). Conversely, Transformers employing lin-type RRE exhibit limited improvement even with up to 64 heads (0.19). This empirical evidence supports our Insight (4a).

Table 8: Results of the experiment supporting Insight (4a).

	$H = 1$	$H = 4$	$H = 16$	$H = 64$
type=log	$<1e-5$	$<1e-5$	$<1e-5$	$<1e-5$
type=lin	0.73	0.68	1.16	0.19

H.2.8 Validation of Insight (4b)

Objective. In contrast to Experiment (4a), this experiment aims to verify that for learning our Task III with light-tailed memories, Transformers with lin-type RPE are efficient, whereas those with log-type RPE fail.

Setup. Specifically, we train single-layer, FFN-free, DP-free Transformers with log-type RPE or lin-type RPE and varying numbers of Attn heads H . The target function involves a heavy-tailed memory kernel $\rho(t) = e^{-5t}$: $f^*(\mathbf{x}) := \sum_{s=1}^{1000} x_s \rho(1000 - s)$ for any input sequence $\mathbf{x} = (x_1, \dots, x_{1000}) \in \{\pm 1\}^{1000}$. Training processes for 10,000 iterations (1M samples) using squared loss and AdamW with the same hyperparameters.

Results and conclusion. The final validation losses are shown in Table 9. As shown in the table, to learn light-tailed memories, even single-head Transformer with lin-type RRE can complete it perfectly ($< 10^{-7}$). Conversely, Transformers employing log-type RRE exhibit limited improvement even with up to 64 heads (5.3×10^{-4}). This empirical evidence supports our Insight (4b).

Table 9: Results of the experiment supporting Insight (4b).

	$H = 1$	$H = 4$	$H = 16$	$H = 64$
type=log	9.1e-4	3.7e-3	2.6e-3	5.3e-4
type=lin	$<1e-7$	$<1e-7$	$<1e-7$	$<1e-7$

H.3 Practical Implications

Our theoretical insights and empirical evidence can directly lead to the following 8 practical implications, such as the strategic selection of Transformer hyperparameters for specific tasks.

- **Implication (1a).** (supported by Insight (1a) and Experiment (1a))
For sequence modeling tasks with complex interrelations between memories, such as many NLP applications, enhancing the number of layers L is more beneficial than increasing the number of Attn heads H and FFN width m .
- **Implication (1b).** (supported by Insight (1b) and Experiment (1b))
For simple sequence modeling tasks with almost no memory intercorrelation, such as the learning of sparse Boolean function, improving performance necessitates only sufficient H and m in a single-layer Transformer, without a need to increase L .
- **Implication (2a).** (supported by Insight (2a) and Experiment (2a))
For sequence modeling tasks with complex readout or memory functions, increasing m can significantly improve performance.
- **Implication (2b).** (supported by Insight (2b) and Experiment (2b))
For sequence modeling tasks with multiple memories, increasing H can markedly improve performance.
- **Implication (3a).** (supported by Insight (3a) and Experiment (3a))
For simple sequence modeling tasks with almost no memory correlations, such as learning sparse Boolean functions, omitting the DP structure in Attn layers can still perform well.
- **Implication (3b).** (supported by Insight (3b) and Experiment (3b))

For sequence modeling tasks with complex correlations between memories, such as many NLP tasks, preserving the DP structure in attention layers is crucial for achieving high performance due to its indispensable nonlinearity.

- **Implication (4a).** (supported by Insight (4a) and Experiment (4a))

For sequence modeling tasks with heavy-tailed memories, the employment of log-type RPE (such as T5's RPE and KERPLE (log)) is recommended over lin-type RPE (such as Alibi).

- **Implication (4b).** (supported by Insight (4b) and Experiment (4b))

For sequence modeling tasks with light-tailed memories, the employment of lin-type RPE (such as Alibi) is recommended over log-type RPE (such as T5's RPE and KERPLE (log)).

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We believe that the abstract and introduction reflect the contributions and scope of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: In Section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In Section 2, 3, 4, and 5; Appendix B, C, D, E, and F.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We believe that all of the experimental results are reproducible in our work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code or data of the experiments are simple and easy to reproduce following the description in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: the approximation error is deterministic and there is no need to consider the error bars here.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Section H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We have confirmed that the research is conducted with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.