The Strong Lottery Ticket Hypothesis for Multi-Head Attention Mechanisms

Hikari Otsuka[1,†], **Daiki Chijiwa**[2], **Yasuyuki Okoshi**[1],
Daichi Fujiki[1], **Susumu Takeuchi**[2], **Masato Motomura**[1]

[1]Institute of Science Tokyo   [2]NTT, Inc.
[†]otsuka.hikari@artic.iir.isct.ac.jp

Abstract

The strong lottery ticket hypothesis (SLTH) conjectures that high-performing subnetworks, called strong lottery tickets (SLTs), are hidden in randomly initialized neural networks. Although recent theoretical studies have established the SLTH across various neural architectures, the SLTH for transformer architectures still lacks theoretical understanding. In particular, the current theory of the SLTH does not yet account for the multi-head attention (MHA) mechanism, a core component of transformers. To address this gap, we introduce a theoretical analysis of the existence of SLTs within MHAs. We prove that, if a randomly initialized MHA of H heads and input dimension d has the hidden dimension $O(d \log(H d^{3/2}))$ for the key and value, it contains an SLT that approximates an arbitrary MHA with the same input dimension with high probability. Furthermore, by leveraging this theory for MHAs, we extend the SLTH to transformers without normalization layers. We empirically validate our theoretical findings, demonstrating that the approximation error between the SLT within a source model (MHA and transformer) and an approximate target counterpart decreases exponentially by increasing the hidden dimension of the source model.
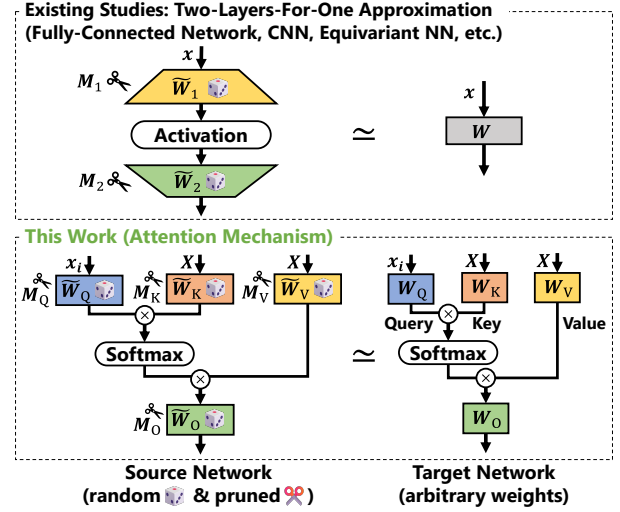
Figure 1: Comparison of the approximation techniques in conventional theories of the SLTH (top) and in our attention-specific approach (bottom). This work demonstrates that an arbitrary attention mechanism can be approximated by pruning a randomly initialized one.

1 Introduction

The *lottery ticket hypothesis* (Frankle and Carbin 2019)—overparameterized networks contain subnetworks that achieve comparable accuracy to fully trained networks even if trained in isolation—presented new possibilities for compact and high-performing models inherent in recent deep neural networks. Later, a stronger claim, which is formally defined as the *strong lottery ticket hypothesis* (SLTH), was proposed (Ramanujan et al. 2020; Malach et al. 2020): overparameterized networks contain subnetworks (called *strong lottery tickets* (SLTs)) that achieve comparable accuracy to the trained dense network even without any training. Whether such subnetworks exist is a fascinating question in itself, and studying them can bring us closer to understanding the principles behind overparameterized models.

The rigorous proof for the SLTH was firstly established in fully-connected networks. Early studies showed that a randomly-weighted fully-connected network of sufficient width (a *source network*) contains an SLT, which approximates an arbitrary fully-connected network with half the depth (a *target network*) (Malach et al. 2020; Orseau, Hutter, and Rivasplata 2020; Pensia et al. 2020). These theories are built on the foundational argument called a *two-layers-for-one approximation*: a two-layer source network with random weight matrices contains an SLT that approximates a single-layer target network with an arbitrary weight matrix (the top panel of Figure 1). Following this finding, subsequent studies have succeeded in proving the existence of SLTs in more complex networks, such as convolutional and equivariant networks (da Cunha, Natale, and Viennot 2022; Burkholz 2022a; Ferbach et al. 2023).

However, the theoretical foundation of the SLTH for *transformers*, which form the basis of modern language models, remains unexplored—due to a transformer-specific component, *an attention mechanism*. As shown in the bottom panel of Figure 1 (right side), one of the distinctive structures in transformers is the inner product between two vectors called *query* and *key*, obtained as linear projections of given inputs. This structure fundamentally differs from

the conventional components of non-transformer architectures for which the SLTH has been established (the top panel of Figure 1); thus, it remains a mystery whether transformers contain SLTs under existing theoretical insights. This gap motivates our key research question: *does an attention mechanism—an essential component of transformers—contain an SLT?*

In this work, we prove the existence of SLTs within attention mechanisms, extending the SLTH to transformers. More precisely, we prove a suitably pruned source attention mechanism with random weights can approximate any target attention mechanism with arbitrary weights:

Theorem 1 (informal). *Given inputs of length T, a suitably pruned randomly-initialized attention mechanism of the input dimension d and hidden dimension $n = O(d \log(d^{3/2}/\epsilon))$ can approximate an arbitrary attention mechanism of the same input dimension with an approximation error ϵ, with high probability.*

Our key idea is to reinterpret the inner product between the query and key vectors in the attention mechanism as a (linear) neural network weighted by the query and key projection matrices. Then, we can view the source and target inner products as neural networks with different numbers of layers: the source one has two layers with query and key projection matrices as its weights, while the target one has a single layer with a weight matrix obtained by merging these two projections. This reinterpretation makes it possible to apply a variant of the two-layers-for-one approximation, leading to the SLT existence within attention mechanisms (Theorem 3). Note that, as can be seen by comparing the top and bottom panels of Figure 1, our arguments do not require additional layers in the MHA for approximation, in contrast to the previous two-layers-for-one argument for fully-connected networks. By exploiting this theorem, we further establish the SLTH for transformers without normalization layers: a randomly-initialized transformer has an SLT that approximates an arbitrary transformer with similar structures (Theorem 6).

We also empirically validate our theory and confirm its implications. Specifically, we show that 1) the approximation error between the source and target attentions (or, more generally, source and target transformers) decays exponentially as the hidden dimension increases; and 2) this approximation error does not diverge even when the input length T increases. Also, based on our theoretical arguments, we derive a new, practical weight initialization scheme, leading to better SLTs in our experiments.

Our contributions are summarized as follows:

- We provide the first theoretical proof that SLTs exist within attention mechanisms and transformers by reinterpreting the inner product in attention mechanisms.
- We then empirically validate our theory under conditions that are close to our theoretical assumptions. More precisely, we carefully designed a synthetic experiment to observe how the hidden dimension or input length affects the approximation error of SLTs.
- Furthermore, we demonstrate that our theory not only explains the empirical results, but also provides a new in-

sight into a weight initialization for finding better SLTs in practical settings.

Notation: In this paper, scalars, vectors, and matrices are denoted by lowercase, bold lowercase, and bold uppercase letters, respectively. We use the norm of matrices and vectors $\| \cdot \|$ as the spectral norm unless otherwise specified by subscripts. We denote the uniform distribution on $[a, b]$ by $U[a, b]$. "$\odot$" represents an element-wise multiplication (i.e., the Hadamard product). The superscript (i) denotes the layer index, and we write $\{x^{(i)}\}_{i=1}^{H}$ to denote the set of elements $x^{(i)}$ indexed by i from 1 to H.

2 Preliminaries

This section reviews the prior theoretical studies on the strong lottery ticket hypothesis (SLTH) and the formulation of multi-head attention (MHA) mechanisms.

2.1 Strong Lottery Ticket Hypothesis

The strong lottery ticket hypothesis (SLTH) conjectured that a randomly-initialized network inherently contains subnetworks (strong lottery tickets (SLTs)) that achieve high accuracy comparable to trained dense networks, without any weight updates (Ramanujan et al. 2020; Malach et al. 2020). The first theoretical result of the SLTH was given by Malach et al. (2020). They proved the existence of SLTs in a fully-connected ReLU network. Subsequent studies relaxed the requirements for source networks to contain SLTs that approximate some target network (Orseau, Hutter, and Rivasplata 2020; Pensia et al. 2020; Burkholz 2022b). In particular, Pensia et al. (2020) introduced a subset-sum approximation technique (Lueker 1998) into the SLTH context and concluded that the logarithmic overparameterization of the source network to a given target is approximately optimal:

Lemma 2. *Given $\boldsymbol{x} \in \mathbb{R}^{d_1}$, $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$, $\tilde{\boldsymbol{W}}_1 \in \mathbb{R}^{n \times d_1}$, and $\tilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_2 \times n}$, we define the target and pruned source fully-connected networks as*

$$\mathrm{F_T}(\boldsymbol{x}) := \boldsymbol{W}\boldsymbol{x},$$
$$\mathrm{F_S}(\boldsymbol{x}) := (\tilde{\boldsymbol{W}}_2 \odot \boldsymbol{M}_2)\mathrm{ReLU}((\tilde{\boldsymbol{W}}_1 \odot \boldsymbol{M}_1)\boldsymbol{x}),$$

where $\boldsymbol{M}_1 \in \{0, 1\}^{n \times d_1}$ and $\boldsymbol{M}_2 \in \{0, 1\}^{d_2 \times n}$ are binary pruning masks. Assume that $\|\boldsymbol{W}\| \leq 1$, $\|\boldsymbol{x}\| \leq 1$, and each entry of $\tilde{\boldsymbol{W}}_1$ and $\tilde{\boldsymbol{W}}_2$ is drawn i.i.d. from $U[-1, 1]$. Also, for $0 < \epsilon < 1$, suppose that the hidden dimension n satisfies $n \geq d_1 C \log(2d_1 d_2/\epsilon)$, where $C > 0$ is some universal constant. Then, with probability at least $1 - \epsilon$, there exists a choice of binary pruning masks $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ such that

$$\|\mathrm{F_T}(\boldsymbol{x}) - \mathrm{F_S}(\boldsymbol{x})\| \leq \epsilon.$$

This approach, which approximates a single weight matrix by pruning two randomly initialized matrices (the top panel of Figure 1), is called the two-layers-for-one approximation, and is now the theoretical foundation of the SLTH for more complex architectures and problems (da Cunha, Natale, and Viennot 2022; Burkholz 2022a; Ferbach et al. 2023; Natale et al. 2024; Otsuka et al. 2025).
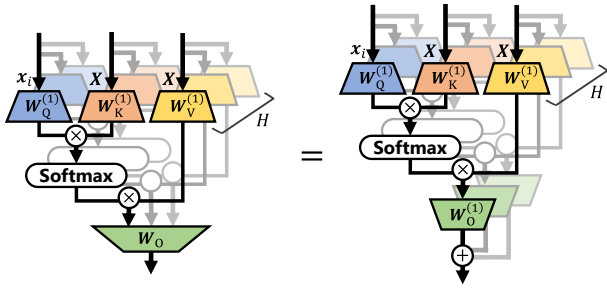
Figure 2: The structure of an MHA. By partitioning the output projection, the final result can be interpreted as the sum of outputs from all heads.

2.2 Multi-head Attention Mechanisms

Let $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times d_1}$ be a sequence of T input vector embeddings. For each embedding $\boldsymbol{x}_i$, we define a binary attention mask $\boldsymbol{a}_i \in \{0, 1\}^\top$, where $\boldsymbol{a}_{i,j} = 1$ indicates that the i-th embedding attends to the j-th one. We assume that each embedding attends to at least one other (i.e., $\|\boldsymbol{a}_i\|_1 \geq 1$). Given such inputs, a multi-head attention (MHA) mechanism (Vaswani et al. 2017) is defined as a function that computes their pair-wise relationships at each of the H attention heads (the left panel of Figure 2). For the i-th embedding $\boldsymbol{x}_i$, the MHA is of the following form:

$$\mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \{\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)}, \boldsymbol{W}_{\mathrm{V}}^{(j)}\}_{j=1}^H, \boldsymbol{W}_{\mathrm{O}})$$
$$:= \left[\mathrm{head}_i^{(1)}, \ldots, \mathrm{head}_i^{(H)}\right] \boldsymbol{W}_{\mathrm{O}} \in \mathbb{R}^{1 \times d_2},$$
$$\mathrm{head}_i^{(j)} := \sigma\left(\frac{\boldsymbol{q}_i^{(j)} \boldsymbol{K}^{(j)\top}}{\sqrt{d_{\mathrm{K}}}}; \boldsymbol{a}_i\right) \boldsymbol{V}^{(j)},$$
$$\boldsymbol{q}_i^{(j)} := \boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{Q}}^{(j)}, \quad \boldsymbol{K}^{(j)} := \boldsymbol{X} \boldsymbol{W}_{\mathrm{K}}^{(j)}, \quad \boldsymbol{V}^{(j)} := \boldsymbol{X} \boldsymbol{W}_{\mathrm{V}}^{(j)},$$
$$\sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)_j := \frac{a_{i,j} \exp(x_{i,j})}{\sum_{k=1}^T a_{i,k} \exp(x_{i,k})}.$$

Here, we define $\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)} \in \mathbb{R}^{d_1 \times d_{\mathrm{K}}}, \boldsymbol{W}_{\mathrm{V}}^{(j)} \in \mathbb{R}^{d_1 \times d_{\mathrm{V}}}$, and $\boldsymbol{W}_{\mathrm{O}} \in \mathbb{R}^{H d_{\mathrm{V}} \times d_2}$ as single-layer projections for the *query* $\boldsymbol{q}_i^{(j)}$, *key* $\boldsymbol{K}^{(j)}$, *value* $\boldsymbol{V}^{(j)}$, and output of the MHA, respectively. The softmax function with the attention mask is defined as $\sigma(\cdot)$. As shown in the right panel of Figure 2, by partitioning the output weight matrix $\boldsymbol{W}_{\mathrm{O}}$ into

$$\boldsymbol{W}_{\mathrm{O}} := [\boldsymbol{W}_{\mathrm{O}}^{(1)\top}, \ldots, \boldsymbol{W}_{\mathrm{O}}^{(H)\top}]^\top, \quad \boldsymbol{W}_{\mathrm{O}}^{(j)} \in \mathbb{R}^{d_{\mathrm{V}} \times d_2},$$

the form of $\mathrm{Attn}(\cdot)$ can be represented as follows:

$$\mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \{\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)}, \boldsymbol{W}_{\mathrm{V}}^{(j)}\}_{j=1}^H, \boldsymbol{W}_{\mathrm{O}})$$
$$= \mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)})$$
$$= \sum_{j=1}^H \mathrm{head}_i^{(j)} \boldsymbol{W}_{\mathrm{O}}^{(j)}.$$

We denote the set of all weights as

$$\boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)} := \{\boldsymbol{W}_{\mathrm{Q}}^{(j)}, \boldsymbol{W}_{\mathrm{K}}^{(j)}, \boldsymbol{W}_{\mathrm{V}}^{(j)}, \boldsymbol{W}_{\mathrm{O}}^{(j)}\}_{j=1}^H.$$

3 Strong Lottery Ticket Hypothesis for Transformers

This section analyzes the existence of SLTs within multi-head attention (MHA) mechanisms and extends it to the transformer architecture without normalization layers. For a detailed proof, see Section A.

3.1 Setups

We consider two MHAs: a target MHA $\mathrm{Attn_T}(\cdot)$ with arbitrary (tuned) weights, and a pruned source MHA $\mathrm{Attn_S}(\cdot)$ with randomly-initialized weights, denoted as follows:

$$\mathrm{Attn_T}(\boldsymbol{x}_i) = \mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)}), \tag{1}$$

$$\mathrm{Attn_S}(\boldsymbol{x}_i) = \mathrm{Attn}(\boldsymbol{x}_i; \boldsymbol{X}, (\tilde{\boldsymbol{W}} \odot \boldsymbol{M})_{\mathrm{Q:O}}^{(1:H)}). \tag{2}$$

Here, similarly to the weight set $\boldsymbol{W}_{\mathrm{Q:O}}^{(1:H)}$, we define the set of pruned random weights as

$$(\tilde{\boldsymbol{W}} \odot \boldsymbol{M})_{\mathrm{Q:O}}^{(1:H)} := \{\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)}, \tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)},$$
$$\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)}, \tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}\}_{j=1}^H,$$

where $\tilde{\boldsymbol{W}}_{\mathrm{Q}}^{(j)}, \tilde{\boldsymbol{W}}_{\mathrm{K}}^{(j)} \in \mathbb{R}^{d_1 \times n_{\mathrm{K}}}, \tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \in \mathbb{R}^{d_1 \times n_{\mathrm{V}}}$, and $\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \in \mathbb{R}^{n_{\mathrm{V}} \times d_2}$ are the randomly-weighted query, key, value, and output projections of the j-th head in $\mathrm{Attn_S}(\cdot)$, respectively. Also, $\boldsymbol{M}_{\mathrm{Q}}^{(j)}, \boldsymbol{M}_{\mathrm{K}}^{(j)}, \boldsymbol{M}_{\mathrm{V}}^{(j)}$, and $\boldsymbol{M}_{\mathrm{O}}^{(j)}$ are their corresponding binary pruning masks. Note that the target and source MHAs have different key and value hidden dimensions: d_{K} and d_{V} for the target, and n_{K} and n_{V} for the source. We assume that $\alpha \geq \max(\sqrt{d_1}, \sqrt{d_2})$ for the inputs, and $\|\boldsymbol{W}_{\mathrm{Q}}^{(j)}\|, \|\boldsymbol{W}_{\mathrm{K}}^{(j)}\|, \|\boldsymbol{W}_{\mathrm{V}}^{(j)}\|, \|\boldsymbol{W}_{\mathrm{O}}^{(j)}\| \leq 1$ for the j-th head of the target MHA. The source MHA is initialized such that each entry of $\tilde{\boldsymbol{W}}_{\mathrm{Q}}$ and $\tilde{\boldsymbol{W}}_{\mathrm{K}}$ is drawn i.i.d. from $U[-n_{\mathrm{K}}^{1/4}, n_{\mathrm{K}}^{1/4}]$, and each entry of $\tilde{\boldsymbol{W}}_{\mathrm{V}}$ and $\tilde{\boldsymbol{W}}_{\mathrm{O}}$ is drawn i.i.d. from $U[-1, 1]$.

3.2 The Existence of SLTs Within an MHA

Now, we prove the following SLT existence theorem:

Theorem 3. *Let* $\mathrm{Attn_T}(\cdot)$ *and* $\mathrm{Attn_S}(\cdot)$ *be as defined in Equations* (1) *and* (2). *Then, with probability at least* $1 - \epsilon$, *there exists a choice of binary pruning masks* $\boldsymbol{M}_{\mathrm{Q}}^{(j)}, \boldsymbol{M}_{\mathrm{K}}^{(j)}, \boldsymbol{M}_{\mathrm{V}}^{(j)}, \boldsymbol{M}_{\mathrm{O}}^{(j)}$ *that satisfy*

$$\max_{i \in [T]} \|\mathrm{Attn_S}(\boldsymbol{x}_i) - \mathrm{Attn_T}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the source hidden dimensions satisfy

$$n_{\mathrm{K}} \geq d_1 C \log\left(\frac{8 H \alpha^3 d_1^{3/2}}{\epsilon}\right),$$

$$n_{\mathrm{V}} \geq d_1 C \log\left(\frac{2 H \alpha d_1 \sqrt{d_2}}{\epsilon}\right),$$

for some universal constant $C > 0$.

Figure 3 shows an overview of our proof. To prove Theorem 3, we begin by focusing on the part before the softmax, the target and source inner products for the j-th head:

$$\frac{1}{\sqrt{d_K}} \boldsymbol{q}^{(j)} \boldsymbol{K}^{(j)\top} = \frac{1}{\sqrt{d_K}} (\boldsymbol{x}_i^\top \boldsymbol{W}_Q^{(j)})(\boldsymbol{X} \boldsymbol{W}_K^{(j)})^\top, \quad (3)$$

$$\frac{1}{\sqrt{n_K}} (\boldsymbol{x}_i^\top (\tilde{\boldsymbol{W}}_Q^{(j)} \odot \boldsymbol{M}_Q^{(j)}))(\boldsymbol{X}(\tilde{\boldsymbol{W}}_K^{(j)} \odot \boldsymbol{M}_K^{(j)}))^\top. \quad (4)$$

Since the only difference lies in the projection matrices, we consider the problem of pruning the source projections $\tilde{\boldsymbol{W}}_Q^{(j)}$ and $\tilde{\boldsymbol{W}}_K^{(j)}$ to approximate the target projections $\boldsymbol{W}_Q^{(j)}$ and $\boldsymbol{W}_K^{(j)}$. A naive idea might be to approximate each target projection independently. In this case, a single source random matrix must approximate each target matrix. However, pruning a single random matrix cannot generally approximate arbitrary ones; thus, this approach is infeasible. To overcome this limitation, we revisit the structure of the target inner product. By closely examining the formulation of the target inner product (Equation (3)), we observe that the query and (transposed) key projections appear adjacently and can be merged into a single joint projection (the right panel of Figure 3):

$$\frac{1}{\sqrt{d_K}} (\boldsymbol{x}_i^\top \boldsymbol{W}_Q^{(j)})(\boldsymbol{X} \boldsymbol{W}_K^{(j)})^\top = \boldsymbol{x}_i^\top \boldsymbol{W}_{QK}^{(j)} \boldsymbol{X}^\top,$$

$$\boldsymbol{W}_{QK}^{(j)} := \frac{1}{\sqrt{d_k}} \boldsymbol{W}_Q^{(j)} (\boldsymbol{W}_K^{(j)})^\top. \quad (5)$$

This reformulation enables us to reinterpret the original problem—not as approximating two target matrices—but as approximating a single merged projection matrix.

We now approximate this merged matrix $\boldsymbol{W}_{QK}^{(j)}$ by pruning the two source projections. On the source side (Equation (4)) as well, the query and key projections are adjacent. Thus, the source inner product can be viewed as a computation that first calculates the query and key projections (the left panel of Figure 3):

$$\frac{1}{\sqrt{n_K}} (\boldsymbol{x}_i^\top (\tilde{\boldsymbol{W}}_Q^{(j)} \odot \boldsymbol{M}_Q^{(j)}))(\boldsymbol{X}(\tilde{\boldsymbol{W}}_K^{(j)} \odot \boldsymbol{M}_K^{(j)}))^\top$$

$$= \boldsymbol{x}_i^\top \left((\tilde{\boldsymbol{W}}_Q^{\prime(j)} \odot \boldsymbol{M}_Q^{(j)})(\tilde{\boldsymbol{W}}_K^{\prime(j)} \odot \boldsymbol{M}_K^{(j)})^\top \right) \boldsymbol{X}^\top,$$

$$\tilde{\boldsymbol{W}}_Q^{\prime(j)} := \frac{1}{n_K^{1/4}} \tilde{\boldsymbol{W}}_Q^{(j)}, \quad \tilde{\boldsymbol{W}}_K^{\prime(j)} := \frac{1}{n_K^{1/4}} \tilde{\boldsymbol{W}}_K^{(j)},$$

where each entry of $\tilde{\boldsymbol{W}}_Q^{\prime(j)}$ and $\tilde{\boldsymbol{W}}_K^{\prime(j)}$ is drawn i.i.d. from $U[-1, 1]$ as per our assumption. Therefore, the task reduces to selecting masks $\boldsymbol{M}_Q^{(j)}$ and $\boldsymbol{M}_K^{(j)}$ such that the source matrix product $(\tilde{\boldsymbol{W}}_Q^{\prime(j)} \odot \boldsymbol{M}_Q^{(j)})(\tilde{\boldsymbol{W}}_K^{\prime(j)} \odot \boldsymbol{M}_K^{(j)})^\top$ closely approximates the target $\boldsymbol{W}_{QK}^{(j)}$. This allows us to draw an analogy to the conventional theoretical results of the SLTH, particularly the two-layers-for-one approximation (Lemma 2). We therefore establish and apply a variant of Lemma 2, which guarantees the existence of binary pruning masks that achieve such an approximation (the bottom panel of Figure 3):
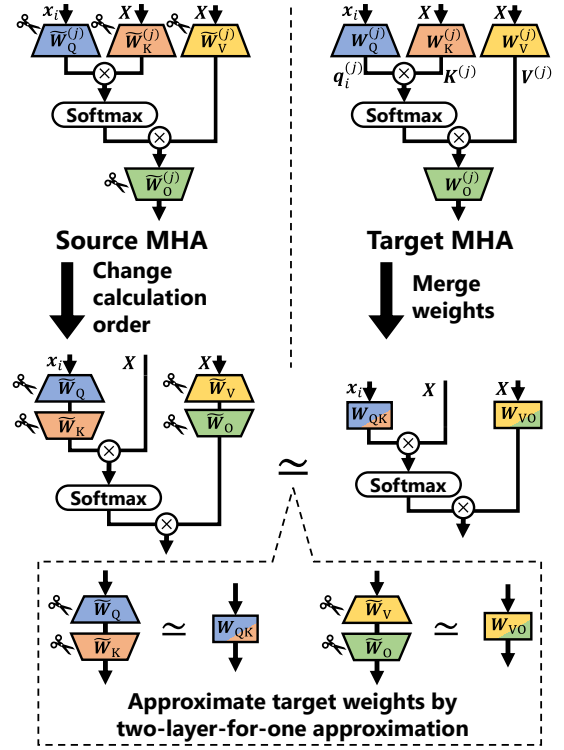

Figure 3: The diagram of our proof. By merging the target projections and changing the calculation order of the source MHA, we can apply a variant of the two-layers-for-one approximation technique and approximate the target MHA while keeping the original source and target structures.

Lemma 4. *Let $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ be a target matrix with $\|\boldsymbol{W}\| \leq 1$, and $\tilde{\boldsymbol{W}}_1 \in \mathbb{R}^{n \times d_1}$ and $\tilde{\boldsymbol{W}}_2 \in \mathbb{R}^{d_2 \times n}$ be source matrices whose entries are drawn i.i.d. from $U[-1, 1]$. Suppose that $n \geq d_1 C \log(d_1 d_2 / \epsilon)$ for some universal constant $C > 0$. Then, with probability at least $1 - \epsilon$, there exists a choice of binary pruning masks $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ such that*

$$\left\| \boldsymbol{W} - (\tilde{\boldsymbol{W}}_2 \odot \boldsymbol{M}_2)(\tilde{\boldsymbol{W}}_1 \odot \boldsymbol{M}_1) \right\|_{\max} \leq \frac{\epsilon}{d_1 d_2}.$$

We now turn to the components after the softmax function: the value and output projections. Similar to the query and key case, the value and output projections appear adjacently and can also be merged into a single composite transformation. Thus, we aim to approximate the target merged matrix $\boldsymbol{W}_{VO}^{(j)} := \boldsymbol{W}_V^{(j)} \boldsymbol{W}_O^{(j)}$. This approximation follows the same principle as before: we leverage the matrix product on the source side $(\tilde{\boldsymbol{W}}_V^{(j)} \odot \boldsymbol{M}_V^{(j)})(\tilde{\boldsymbol{W}}_O^{(j)} \odot \boldsymbol{M}_O^{(j)})$ to approximate the merged matrix $\boldsymbol{W}_{VO}^{(j)}$. Lemma 4 ensures that, with high probability, this approximation is successful via appropriately chosen binary pruning masks $\boldsymbol{M}_V^{(j)}$ and $\boldsymbol{M}_O^{(j)}$.

Assuming that all weights in the target MHA are approximated by the above procedure, we next analyze the error of the entire attention mechanism by investigating the behavior of the softmax. As a natural idea, one might consider exploiting the 1-Lipschitz continuity of the softmax (Gao and

Pavel 2017), which enables internal errors to propagate linearly to the output. However, since the MHA subsequently multiplies the softmax output and the input matrix $\boldsymbol{X}$, applying Lipschitz continuity results in a loose upper bound of the error between MHAs: as $\|\boldsymbol{X}\|$ can grow with T in the worst case, the bound depends on the input length T.

In contrast to this general approach, we provide a more precise analysis. In our setting, thanks to the accurate weight approximation technique mentioned earlier, the internal error of softmax is guaranteed to be finite and small. Leveraging this property, we analyze the softmax output and $\boldsymbol{X}$ simultaneously to obtain a T-independent bound as follows:

Lemma 5. *Let* $\boldsymbol{\epsilon} \in \mathbb{R}^{d_1}$ *be an error vector with* $\|\boldsymbol{\epsilon}\|_{\max} \leq \epsilon_{\max}$ *for some* $0 \leq \epsilon_{\max} \leq 1/2$. *Then,*

$$\max_{i \in [T]} \|\sigma(\boldsymbol{x}_i; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i + \boldsymbol{\epsilon}; \boldsymbol{a}_i)\boldsymbol{X}\| \leq 4\sqrt{d_1}\alpha\epsilon_{\max}.$$

Since this lemma provides a bound independent of $\boldsymbol{a}_i$, our theory holds for models with arbitrary attention masks, including encoder (Devlin et al. 2019) and decoder models (Radford et al. 2019). By applying these above analyses to each attention head, we complete the proof of Theorem 3. For the full proof, see Section A.3. We also empirically validate two main theoretical findings in Section 4.2: the accurate approximation of the target MHA becomes feasible with larger source hidden dimensions, and the the approximation error remains independent of the input length T.

Proof Sketch of Theorem 3: First, for each attention head, we reformulate the problem by merging the four original target projection matrices into two merged matrices: one combining the query and key projections, and the other the value and output projections. Applying Lemma 4 to these merged matrices enables us to prune each source head to produce an inner product that closely approximates the target one. Next, using Lemma 5, we bound how errors in approximating the query and key matrices propagate through the softmax operation. Lemma 5 ensures that the approximation error of the softmax depends on the approximation accuracy of the query-key projections and does not scale with the input length T. Thus, provided the source hidden dimensions n_K and n_V are sufficiently large, there exists a choice of binary masks for the source MHA which approximate the target MHA within an error ϵ. Also, by suitably setting lower bounds on n_K and n_V, a union bound guarantees that the approximation succeeds across all heads with probability at least $1 - \epsilon$.

3.3 The Existence of SLTs Within a Transformer

By leveraging our main theorem, we now extend the SLTH to transformers. We consider a transformer without the normalization layers for the original definition (Vaswani et al. 2017). The target transformer of B blocks are of the following form:

$$\mathrm{Tf}_{\mathrm{T}}(\boldsymbol{x}_i) := \mathrm{Blk}_{\mathrm{T}}^{(B)}(\mathrm{Blk}_{\mathrm{T}}^{(B-1)} \ldots \mathrm{Blk}_{\mathrm{T}}^{(1)}(\boldsymbol{x}_i)),$$
$$\mathrm{Blk}_{\mathrm{T}}^{(b)}(\boldsymbol{x}_i^{(b)}) := \mathrm{F}_{\mathrm{T}}^{(b)}(\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i^{(b)})^\top + \boldsymbol{x}_i^{(b)})$$
$$+ \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i^{(b)})^\top + \boldsymbol{x}_i^{(b)},$$

where $\mathrm{Blk}_{\mathrm{T}}^{(b)}$ is a b-th target block, and $\boldsymbol{x}_i^{(b)} \in \mathbb{R}^d$ is the i-th input embedding of the b-th target block. We employ single-layer projection $\mathrm{F}_{\mathrm{T}}^{(b)}(\cdot)$ for the fully-connected network of each target block. Similarly, we define the pruned source transformer as follows:

$$\mathrm{Tf}_{\mathrm{S}}(\boldsymbol{x}_i) := \mathrm{Blk}_{\mathrm{S}}^{(B)}(\mathrm{Blk}_{\mathrm{S}}^{(B-1)} \ldots \mathrm{Blk}_{\mathrm{S}}^{(1)}(\boldsymbol{x}_i)),$$
$$\mathrm{Blk}_{\mathrm{S}}^{(b)}(\boldsymbol{x}_i'^{(b)}) := \mathrm{F}_{\mathrm{S}}^{(b)}(\mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i'^{(b)})^\top + \boldsymbol{x}_i'^{(b)})$$
$$+ \mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i'^{(b)})^\top + \boldsymbol{x}_i'^{(b)},$$

where $\mathrm{Blk}_{\mathrm{S}}^{(b)}(\cdot)$ is a b-th source block and $\boldsymbol{x}_i'^{(b)}$ is the i-th input embedding of the b-th source block. We set the hidden dimension of $\mathrm{F}_{\mathrm{S}}^{(b)}(\cdot)$ as $n_{\mathrm{FC}}^{(b)}$ and assume the hidden dimensions of the MHA in the b-th source block as a same value $n_{\mathrm{MHA}}^{(b)}$ for simplicity. Then, we prove the following theorem:

Theorem 6. *Assume* $B \geq 2$. *Then, with probability at least* $1 - \epsilon$ *for* $0 < \epsilon < 1$, *there exists a choice of binary pruning masks that satisfies*

$$\|\mathrm{Tf}_{\mathrm{S}}(\boldsymbol{x}_i) - \mathrm{Tf}_{\mathrm{T}}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the hidden dimensions of b-th *source MHA and fully-connected network satisfy*

$$n_{\mathrm{MHA}}^{(b)} \geq d_1 C \log\left(\frac{c_1^{f_1(b,B)} H^{f_2(b,B)} d_1^{f_3(b,B)}}{\epsilon}\right),$$
$$n_{\mathrm{FC}}^{(b)} \geq d_1 C \log\left(\frac{c_2^{g_1(b,B)} H^{g_2(b,B)} d_1^{g_3(b,B)}}{\epsilon}\right),$$

for universal constants $C > 0$ *and* $c_1, c_2 > 0$ *including* α. *Here,* $f_1, f_2, f_3, g_1, g_2, g_3$ *are quadratic forms of* b *and* B.

Proof Sketch of Theorem 6: From the existing work (Lemma 2) and Theorem 3, we already know that an MHA and FFN contain SLTs with high probability if each module has a large hidden dimension; thus, by determining the lower bound of the hidden dimension of each module based on the error propagation from the input to output, we can prove that there exists an SLT, which approximates the output of an target transformer to an error of ϵ, within a randomly initialized transformer. By the union bound, the probability that all approximations hold simultaneously is at least $1 - \epsilon$.

For simplicity, this theorem uses target and source fully-connected networks as a single-layer and two-layer ReLU networks F_{T} and F_{S} in Lemma 2. It can be generalized to an L-layer target fully-connected network by applying the multi-layer approximation by Pensia et al. (2020). We show that theorem and its proof in Section A.5.

4 Experimental Results

This section empirically validates our SLTH theorems.

4.1 Experimental Settings

To empirically validate the approximation guarantees established by our SLTH theorems, we evaluate the approximation error on a synthetic dataset for angular velocity estimation. The input consists of a sequence of two-dimensional
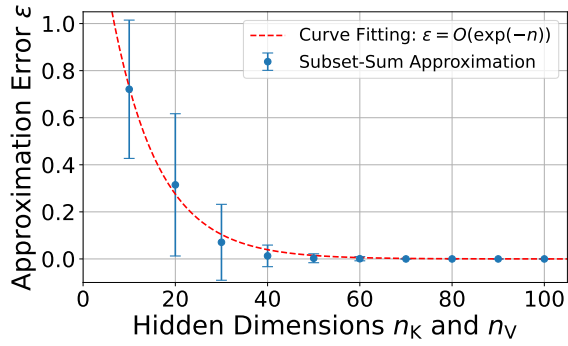
Figure 4: The approximation error ϵ of SLTs within a source MHA for the hidden dimensions $n_K = n_V$. This result shows that the error ϵ satisfies $\epsilon = O(\exp(-n))$, consistently with Theorem 3.

Figure 5: The approximation error ϵ of SLTs within an MHA for the sequence length T. This result suggests that the error ϵ does not diverge as T increases, as implied by Theorem 3.

Figure 6: The approximation error ϵ of SLTs within a randomly initialized transformer for the source hidden dimensions $n_{MHA} = n_{FC}$. This result suggests that error accumulates as the number of blocks increases, while each error holds $\epsilon = O(\exp(-n_{MHA}))$, consistently with Theorem 3.

vectors arranged on the unit circle with a fixed angular velocity. A regression token is used to estimate this velocity, and the source model uses the same regression token as the target model to ensure input consistency. The source and target models are both implemented as either single-head attention mechanisms or single-head transformers as defined in Section 3.3. Both models are initialized according to our theoretical setup: the entries of the query and key projection weights are drawn i.i.d. from $U[-n_K^{1/4}, n_K^{1/4}]$, and those of the value and output projection weights from $U[-1, 1]$. To identify SLTs that approximate the target network, we implement the weight approximation technique described in Lemma 4, which is based on the subset-sum approximation of Pensia et al. (2020). The target MHA is approximated using 100 randomly initialized source MHAs, and we report the mean and standard deviation of the approximation error.

We also investigate whether our theoretical insights generalize to practical settings. In this setting, we search for SLTs by the `edge-popup` algorithm (Ramanujan et al. 2020), which finds accurate subnetworks by backpropagation, instead of learning weights. We train models from the GPT-2 family (mini[1], small, and medium) (Radford et al. 2019) on the WikiText-103 dataset (Merity et al. 2017). The weights of these models are initialized based on the GPT-2 initialization scheme. For each model, we repeat training three times with different random seeds and report the mean and standard deviation of the final performance. See Section B for further details on experimental settings.

4.2 Verification of Main Theorems

We empirically verify our theoretical results by pruning a source network to approximate the target network.

Varying the Hidden Dimensions: We validate Theorem 3 by showing that increasing the hidden dimensions leads to an exponential decrease in approximation error. When we fit the empirical results to the function $\epsilon = \gamma \exp(-\delta n_K)$, we obtain $\epsilon = 0.8 \exp(-0.06 n_K)$, which closely matches

---

[1]A 4-layer GPT-2. For details, see the following repository: https://huggingface.co/erwanf/gpt2-mini (Wolf et al. 2020)
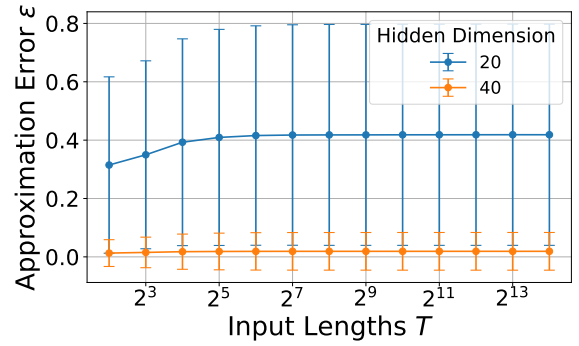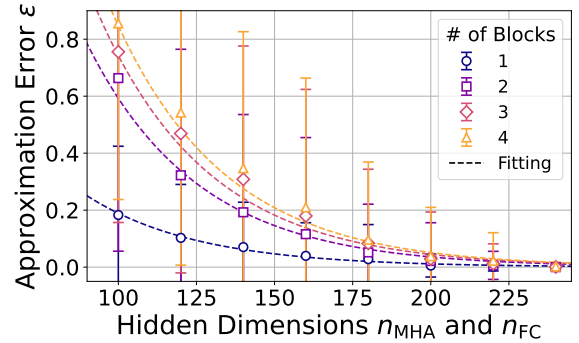
the observations. This finding supports our theoretical claim: given a target MHA, each source hidden dimension requires $O(\log(1/\epsilon))$ for the existence of SLTs.

Varying the Sequence Length: Theorem 3 also implies that the existence of SLTs in MHAs is independent of the input length T. In other words, with sufficiently large hidden dimensions, the approximation error has an upper bound that does not depend on T. Figure 5 empirically supports this argument: even as T increases, the error remains bounded, and the bound decreases with larger hidden dimensions.

Varying the Number of Blocks: To validate Theorem 6, we analyze how the approximation error behaves across different numbers of transformer blocks. We set $n_{MHA} = n_{FFN}$ and use an untrained target model to be close to our theoretical assumptions. As in the MHA experiment, we fit an exponential decay $\epsilon = \gamma \exp(-\delta n_K)$ to the error of each block, using the same decay rate δ obtained from the first block, as predicted by Theorem 6. Figure 6 shows that, consistent with our theoretical implication, the approximation error decreases rapidly with increasing hidden dimensions for all numbers of blocks. Despite fitting only the coefficient
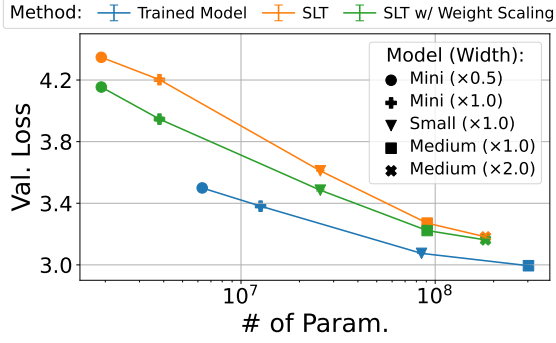
Figure 7: Loss comparison between SLTs with and without query and key weight scaling. By introducing a scale based on our theoretical assumptions, we can obtain better SLTs.
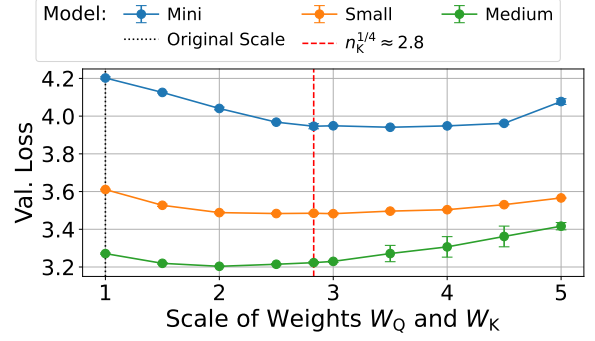

Figure 8: Loss comparison with respect to the weight scaling factor applied to query and key weights. Interestingly, in all models, the loss reaches its minimum near the weight scaling of our theoretical assumptions.

γ per block, the shared δ provides curves that closely match the empirical results, supporting our theoretical claim that only the scale factor varies across blocks.

4.3 Behavior of SLTs in Practical Settings

In the theoretical analysis, we employ a non-conventional initialization strategy: the query and key projection weights are initialized from $U[-n_K^{1/4}, n_K^{1/4}]$, scaled by a factor of $n_K^{1/4}$ compared to the value and output weights, which are initialized from $U[-1, 1]$. This weight scaling was introduced to facilitate the application of the weight approximation lemma in our analysis, and played an important role in establishing our theory. Its theoretical contribution motivates the following question: *does this scaled initialization strategy also benefit SLTs in realistic scenarios?* We empirically evaluate SLTs using the GPT-2 architectures and the WikiText-103 dataset. Figure 7 compares the validation loss of SLTs with and without scaling the query and key weights by $n_K^{1/4} \simeq 2.8$, with respect to the number of nonzero parameters. We observe that SLTs with the weight scaling tend to exhibit lower loss, approaching the performance of trained models. Interestingly, this specific scaling factor $n_K^{1/4}$ is nearly optimal for finding better SLTs: as shown in Figure 8, increasing the scale from 1 gradually decreases the loss up to a certain point, but further increasing it beyond $n_K^{1/4}$ results in increased loss. In all models, the lowest loss is consistently achieved around this scaling factor $n_K^{1/4}$. These findings suggest that our initialization strategy actually helps to ensure the existence of better SLTs within the practical transformer models.

5 Related Work

Strong Lottery Tickets: Zhou et al. (2019) and Ramanujan et al. (2020) empirically found the subnetworks that achieve high accuracy without any weight training. The existence of such high-performing subnetworks has been called the strong lottery ticket hypothesis (SLTH), and its theoretical proof was firstly provided in fully-connected ReLU networks (Malach et al. 2020; Orseau, Hutter, and Rivasplata

2020; Pensia et al. 2020; Burkholz 2022b).

Based on these pioneering studies, the SLTH has been extended in three main directions. The first direction involves introducing additional flexibility for relaxing the overparameterization of the source network (Chijiwa et al. 2021; Xiong, Liao, and Kyrillidis 2023). The second direction, in contrast, imposes additional constraints on the source network (Gadhikar, Mukherjee, and Burkholz 2023; Otsuka et al. 2025; Natale et al. 2024). The third direction extends the SLTH to various architectures (Diffenderfer and Kailkhura 2021; Burkholz 2022b; Fischer and Burkholz 2021; da Cunha, Natale, and Viennot 2022; Da Cunha and d'Amore 2023; Burkholz 2022a; Ferbach et al. 2023). Our work contributes to this third direction by proving the SLTH for attention mechanisms and transformers.

Randomly Weighted Transformers: Several studies have empirically investigated the capabilities of randomly weighted transformers. Shen et al. (2021a) demonstrated that a transformer with a few randomly weighted layers achieves accuracy comparable to fully trained models on translation and language understanding tasks. Zhong and Andreas (2024) found that randomly weighted transformers can solve toy tasks with high accuracy as the hidden dimension increases. Some studies empirically showed the existence of SLTs within randomly weighted transformers (Shen et al. 2021b; Ito et al. 2025). Our analysis provides theoretical support for these empirical results about the SLT existence. Furthermore, it provides a theoretical explanation for the improved performance of randomly weighted transformers as the hidden dimension increases, particularly when the pruning is used for optimization.

6 Conclusion

This work investigated the existence of SLTs within a multi-head attention (MHA) mechanism. We extended the existing theory of the SLTH to MHAs and proved that, if the source MHA has logarithmically large hidden dimensions, it contains an SLT that approximates an arbitrary MHA with high probability. Our proof revealed that, for the SLTH in

MHAs, additional layers are not required for approximation, in contrast to the existing theories that rely on approximating a single-layer structure by a two-layer one. Furthermore, by exploiting our findings, we established the theory of the SLTH for transformers without normalization layers. We empirically validated our theory and confirmed that the results are consistent with the theoretical implications. Interestingly, our theoretical implication, which provides an appropriate weight scale for initializing query and key projection weights, contributed to improving the performance of SLTs in practical settings. Our results not only extend SLTH to transformers, but also indicate a new research direction in the SLTH for practical transformer models. We hope these findings will lead to a fundamental understanding of overparameterized models.

Acknowledgments

References

Burkholz, R. 2022a. Convolutional and residual networks provably contain lottery tickets. In *International Conference on Machine Learning*, 2414–2433. PMLR.

Burkholz, R. 2022b. Most activation functions can win the lottery without excessive depth. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Chijiwa, D.; Yamaguchi, S.; Ida, Y.; Umakoshi, K.; and Inoue, T. 2021. Pruning randomly initialized neural networks with iterative randomization. *Advances in neural information processing systems*, 34: 4503–4513.

Da Cunha, A.; and d'Amore, F. 2023. Polynomially overparameterized convolutional neural networks contain structured strong winning lottery tickets. *Advances in Neural Information Processing Systems*, 36: 25929–25957.

da Cunha, A.; Natale, E.; and Viennot, L. 2022. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *ICLR 2022-10th International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Diffenderfer, J.; and Kailkhura, B. 2021. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. In *International Conference on Learning Representations*.

Ferbach, D.; Tsirigotis, C.; Gidel, G.; and Bose, J. 2023. A general framework for proving the equivariant strong lottery ticket hypothesis. In *The Eleventh International Conference on Learning Representations*.

Fischer, J.; and Burkholz, R. 2021. Towards strong pruning for lottery tickets with non-zero biases. *arXiv preprint arXiv:2110.11150*.

Frankle, J.; and Carbin, M. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Gadhikar, A. H.; Mukherjee, S.; and Burkholz, R. 2023. Why random pruning is all we need to start sparse. In *International Conference on Machine Learning*, 10542–10570. PMLR.

Gao, B.; and Pavel, L. 2017. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR.

Gurobi Optimization, LLC. 2024. Gurobi optimizer reference manual.

Ito, H.; Yan, J.; Otsuka, H.; Kawamura, K.; Motomura, M.; Chu, T. V.; and Fujiki, D. 2025. Uncovering strong lottery tickets in graph transformers: A path to memory efficient and robust graph learning. *Transactions on Machine Learning Research*.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lueker, G. S. 1998. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures & Algorithms*, 12(1): 51–62.

Malach, E.; Yehudai, G.; Shalev-Shwartz, S.; and Shamir, O. 2020. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, 6682–6691. PMLR.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Natale, E.; Ferre', D.; Giambartolomei, G.; Giroire, F.; and Mallmann-Trenn, F. 2024. On the sparsity of the strong lottery ticket hypothesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Orseau, L.; Hutter, M.; and Rivasplata, O. 2020. Logarithmic pruning is all you need. *Advances in Neural Information Processing Systems*, 33: 2925–2934.

Otsuka, H.; Chijiwa, D.; García-Arias, Á. L.; Okoshi, Y.; Kawamura, K.; Chu, T. V.; Fujiki, D.; Takeuchi, S.; and Motomura, M. 2025. Partially frozen random networks contain compact strong lottery tickets. *Transactions on Machine Learning Research*.

Pensia, A.; Rajput, S.; Nagle, A.; Vishwakarma, H.; and Papailiopoulos, D. 2020. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in neural information processing systems*, 33: 2599–2610.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ramanujan, V.; Wortsman, M.; Kembhavi, A.; Farhadi, A.; and Rastegari, M. 2020. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11893–11902.

Shen, S.; Baevski, A.; Morcos, A.; Keutzer, K.; Auli, M.; and Kiela, D. 2021a. Reservoir transformers. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4294–4309. Online: Association for Computational Linguistics.

Shen, S.; Yao, Z.; Kiela, D.; Keutzer, K.; and Mahoney, M. 2021b. What's hidden in a one-layer randomly weighted transformer? In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2914–2921. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17: 261–272.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. HuggingFace's transformers: State-of-the-art natural language processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Xiong, Z.; Liao, F.; and Kyrillidis, A. 2023. Strong lottery ticket hypothesis with ε-perturbation. In *International Conference on Artificial Intelligence and Statistics*, 6879–6902. PMLR.

Zhong, Z.; and Andreas, J. 2024. Algorithmic capabilities of random transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhou, H.; Lan, J.; Liu, R.; and Yosinski, J. 2019. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in neural information processing systems*, 32.

A Proofs of Main Theorems

This section presents the detailed proofs of the main theorems in the manuscript. We first introduce two lemmas: one for approximating a target weight matrix by pruning two random weight matrices, and another for bounding the effect of perturbations in the softmax function. These lemmas are then used to establish the SLTH for attention mechanisms. Then, leveraging the theory of the SLTH for attention mechanisms, we prove the existence of SLTs in transformer blocks and transformers without normalization layers.

A.1 Weight Approximation

Pensia et al. (2020) have shown that a two-layer fully-connected ReLU network can approximate arbitrary matrices with high probability. Our problem setting can be viewed as a simplified version of their construction, in which the ReLU nonlinearity is omitted. We follow their proof strategy and simplify it to the linear (non-activated) case.

Lemma 7. *Let $W \in \mathbb{R}^{d_2 \times d_1}$ be a target matrix with entries in $[-1, 1]$. Let $\tilde{W}_1 \in \mathbb{R}^{n \times d_1}$ and $\tilde{W}_2 \in \mathbb{R}^{d_2 \times n}$ be source random matrices whose entries are drawn i.i.d. from $U[-1, 1]$. For any $0 < \epsilon < 1$, suppose that $n \geq d_1 C \log(\frac{d_1 d_2}{\epsilon})$ for some universal constant $C > 0$. Then, with probability at least $1 - \epsilon$, there exists a choice of binary masks $M_1 \in \{0, 1\}^{n \times d_1}$ and $M_2 \in \{0, 1\}^{d_2 \times n}$ such that*

$$\left\| W - (\tilde{W}_2 \odot M_2)(\tilde{W}_1 \odot M_1) \right\|_{\max} \leq \frac{\epsilon}{d_1 d_2}.$$

Proof. Firstly, we structurally prune the random weight matrix $\tilde{W}_1$ by the pruning mask M_1:

$$\tilde{W}_1 \odot M_1 = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{d_1} \end{bmatrix}, \tag{6}$$

where $u_i \in \mathbb{R}^{n'}$. Next, we decompose $\tilde{W}_2 \odot M_2$ as follows:

$$\tilde{W}_2 \odot M_2 = \begin{bmatrix} (v_{1,1} \odot m_{1,1})^\top & (v_{1,2} \odot m_{1,2})^\top & \cdots & (v_{1,d_1} \odot m_{1,d_1})^\top \\ (v_{2,1} \odot m_{2,1})^\top & (v_{2,2} \odot m_{2,2})^\top & \cdots & (v_{2,d_1} \odot m_{2,d_1})^\top \\ \vdots & \vdots & \ddots & \vdots \\ (v_{d_2,1} \odot m_{d_2,1})^\top & (v_{d_2,2} \odot m_{d_2,2})^\top & \cdots & (v_{d_2,d_1} \odot m_{d_2,d_1})^\top \end{bmatrix}, \tag{7}$$

where $v_{i,j} \in \mathbb{R}^{n'}$ and $m_{i,j} \in \{0, 1\}^{n'}$. These operations enable us to rewrite the product of Equations (6) and (7) as follows:

$$(\tilde{W}_2 \odot M_2)(\tilde{W}_1 \odot M_1) = \begin{bmatrix} (v_{1,1} \odot m_{1,1})^\top u_1 & (v_{1,2} \odot m_{1,2})^\top u_2 & \cdots & (v_{1,d_1} \odot m_{1,d_1})^\top u_{d_1} \\ (v_{2,1} \odot m_{2,1})^\top u_1 & (v_{2,2} \odot m_{2,2})^\top u_2 & \cdots & (v_{2,d_1} \odot m_{2,d_1})^\top u_{d_1} \\ \vdots & \vdots & \ddots & \vdots \\ (v_{d_2,1} \odot m_{d_2,1})^\top u_1 & (v_{d_2,2} \odot m_{d_2,2})^\top u_2 & \cdots & (v_{d_2,d_1} \odot m_{d_2,d_1})^\top u_{d_1} \end{bmatrix} \tag{8}$$

We focus on the (i, j)-th entry of Equation (8). This entry can be rewritten as a subset sum of element-wise products between the vectors $v_{i,j}$ and u_j:

$$(v_{i,j} \odot m_{i,j})^\top u_j = \sum_{k=1}^{n'} m_{i,j,k} v_{i,j,k} u_{j,k}.$$

Here, each $m_{i,j,k}$ determines whether the corresponding product $v_{i,j,k} u_{j,k}$ is included in the subset sum. We aim to approximate the (i, j)-th entry of the target weight matrix W with the subset sum $\sum_{k=1}^{n'} m_{i,j,k} v_{i,j,k} u_{j,k}$ by appropriately choosing the binary mask $m_{i,j}$. Since each entry of $v_{i,j,k}$ and $u_{j,k}$ is drawn i.i.d. from $U[-1, 1]$, each product $v_{i,j,k} u_{j,k}$ can be viewed as drawn from the distribution including some uniform distribution; thus, we can apply Corollary 3.3 of Lueker (1998), which states that if $n' \geq C \log \left(\frac{d_1 d_2}{\epsilon} \right)$, then with probability at least $1 - \frac{\epsilon}{d_1 d_2}$, there exists a binary mask vector $m_{i,j}$ such that the subset sum $\sum_{k=1}^{n'} m_{i,j,k} v_{i,j,k} u_{j,k}$ approximates the (i, j)-th entry of W within an error of $\frac{\epsilon}{d_1 d_2}$. By the union bound, the probability that all entries of the weight matrix W are simultaneously approximated is at least $1 - \epsilon$:

$$1 - \sum_{i=1}^{d_2} \sum_{j=1}^{d_1} \frac{\epsilon}{d_1 d_2} = 1 - \epsilon.$$

Therefore, if $n = d_1 n' \geq d_1 C \log\left(\frac{d_1 d_2}{\epsilon}\right)$, then with probability at least $1 - \epsilon$, the following inequality holds:

$$\left\| W - (\tilde{W}_2 \odot M_2)(\tilde{W}_1 \odot M_1) \right\|_{\max} \leq \frac{\epsilon}{d_1 d_2}.$$

$\square$

A.2 Spectral Norm of Softmax Difference

In addition to approximating target weights, we need to analyze the stability of the softmax output under small input perturbations, with respect to the spectral norm of the resulting attention-weighted output.

Lemma 8. *Given $\epsilon \in \mathbb{R}^{d_1}$ as a perturbation vector such that $\|\epsilon\|_{\max} \leq \epsilon_{\max}$ for some $\epsilon_{\max} \geq 0$, we have*

$$\|\sigma(x_i; a_i)X - \sigma(x_i + \epsilon; a_i)X\| \leq \sqrt{d_1}\alpha\left(\exp(2\epsilon_{\max}) - 1\right).$$

Proof. Let $p_i := \sigma(x_i; a_i)$ and $p'_i := \sigma(x_i + \epsilon; a_i)$. Then, for each coordinate j, we have

$$p'_{i,j} = p_{i,j} \frac{\exp(\epsilon_j)}{Z},$$

$$Z = \sum_{k=1}^{T} p_{i,k} \exp(\epsilon_k).$$

By the assumption $\|\epsilon\|_{\max} \leq \epsilon_{\max}$, we have the following bound:

$$\left| 1 - \frac{\exp(\epsilon_j)}{Z} \right| \leq \exp(2\epsilon_{\max}) - 1. \tag{9}$$

Now, we can bound the spectral norm for the i-th input embedding:

$$
\begin{aligned}
\|p_i X - p'_i X\| &\leq \sqrt{d_1} \cdot \|p_i X - p'_i X\|_{\max} \\
&\leq \sqrt{d_1} \cdot \max_{j \in [d_1]} \left| \sum_{k=1}^{T} (p_{i,k} - p'_{i,k}) x_{k,j} \right| \\
&\leq \sqrt{d_1} \cdot \max_{j \in [d_1]} \sum_{k=1}^{T} |x_{k,j}| \cdot |p_{i,k} - p'_{i,k}| \\
&\leq \sqrt{d_1} \cdot \alpha \sum_{k=1}^{T} |p_{i,k} - p'_{i,k}| \\
&\leq \sqrt{d_1} \cdot \alpha \sum_{k=1}^{T} p_{i,k} \left| 1 - \frac{\exp(\epsilon_k)}{Z} \right| \\
&\leq \sqrt{d_1} \cdot \alpha \left(\exp(2\epsilon_{\max}) - 1\right) \sum_{k=1}^{T} p_{i,k} \qquad \text{(Using Equation (9))} \\
&= \sqrt{d_1} \cdot \alpha \left(\exp(2\epsilon_{\max}) - 1\right).
\end{aligned}
$$

This upper bound is independent of i; thus, the upper bound of $\max_{i \in [T]} \|pX - p'X\|$ is same as the final upper bound. $\square$

A.3 SLT Existence within Attention Mechanisms

By leveraging these two lemmas, we prove the following theorem:

Theorem 9. *Let $\mathrm{Attn}_S(\cdot)$ and $\mathrm{Attn}_T(\cdot)$ be as defined in Equations (1) and (2). Assume $\alpha \geq \max(\sqrt{d_1}, \sqrt{d_2})$ for the inputs. Then, with probability at least $1 - \epsilon$, there exists a choice of binary masks $M_Q^{(j)}, M_K^{(j)}, M_V^{(j)}, M_O^{(j)}$ that satisfy*

$$\max_{i \in [T]} \|\mathrm{Attn}_S(x_i) - \mathrm{Attn}_T(x_i)\| \leq \epsilon,$$

if the source dimensions satisfy

$$n_1 \geq d_1 C \log\left(\frac{8H\alpha^3 d_1^{3/2}}{\epsilon}\right),$$

$$n_2 \geq d_1 C \log\left(\frac{2H\alpha d_1 \sqrt{d_2}}{\epsilon}\right),$$

for some universal constant $C > 0$.

Proof. We prove the theorem in three steps.

Step 1: Weight Merging. We begin by merging the weight matrices of the target and source MHAs. The target MHA weights are merged as

$$\boldsymbol{W}_{\text{QK}}^{(j)} := \frac{1}{\sqrt{d_{\text{K}}}} \boldsymbol{W}_{\text{Q}}^{(j)} (\boldsymbol{W}_{\text{K}}^{(j)})^{\top}, \tag{10}$$

$$\boldsymbol{W}_{\text{VO}}^{(j)} := \boldsymbol{W}_{\text{V}}^{(j)} \boldsymbol{W}_{\text{O}}^{(j)}. \tag{11}$$

This operation (Equations (10) and (11)) enables us to represent each head of $\text{Attn}_{\text{T}}(\cdot)$ as

$$\text{Attn}_{\text{T}}(\boldsymbol{x}_i; \boldsymbol{X}, \boldsymbol{W}_{\text{Q:O}}^{(1:H)}) = \sum_{j=1}^{H} \sigma\left(\boldsymbol{x}_i^{\top} \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{X}^{\top}; \boldsymbol{a}_i\right) \boldsymbol{X} \boldsymbol{W}_{\text{VO}}^{(j)}.$$

From the assumption on the target weights, we have the following norm bounds:

$$\|\boldsymbol{W}_{\text{QK}}^{(j)}\| \leq 1/\sqrt{d_{\text{K}}}, \tag{12}$$

$$\|\boldsymbol{W}_{\text{VO}}^{(j)}\| \leq 1. \tag{13}$$

For the source MHA, we incorporate the scaling factor $1/\sqrt{n_{\text{K}}}$ into the query and key weight matrices:

$$\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} := \frac{1}{d_{\text{K}}^{1/4}} \tilde{\boldsymbol{W}}_{\text{Q}}^{(j)},$$

$$\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} := \frac{1}{d_{\text{K}}^{1/4}} \tilde{\boldsymbol{W}}_{\text{K}}^{(j)}.$$

Assuming that each entry of $\tilde{\boldsymbol{W}}_{\text{Q}}^{(j)}$ and $\tilde{\boldsymbol{W}}_{\text{K}}^{(j)}$ is drawn i.i.d. from $U[-d_{\text{K}}^{1/4}, d_{\text{K}}^{1/4}]$, each entry of the scaled matrices $\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)}$ and $\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)}$ is drawn i.i.d. from $U[-1, 1]$.

Step 2: Weight Approximation. From Lemma 7, for any $0 < \epsilon < 1$, if

$$n_{\text{K}} \geq d_1 C \log\left(\frac{8H\alpha^3 d_1^{3/2}}{\epsilon}\right),$$

then with probability at least $1 - \frac{\epsilon}{8H\alpha^3\sqrt{d_1}}$, there exists a choice of binary masks $\boldsymbol{M}_{\text{Q}}^{(j)}$ and $\boldsymbol{M}_{\text{K}}^{(j)}$ such that

$$\left\|\boldsymbol{W}_{\text{QK}}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} \odot \boldsymbol{M}_{\text{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} \odot \boldsymbol{M}_{\text{K}}^{(j)}\right)^{\top}\right\|_{\max} \leq \frac{\epsilon}{8H\alpha^3 d_1^{3/2}}. \tag{14}$$

This inequality Equation (14) implies a bound on the softmax input:

$$\left\|\boldsymbol{x}_i^{\top} \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{X}^{\top} - \boldsymbol{x}_i^{\top}\left(\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} \odot \boldsymbol{M}_{\text{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} \odot \boldsymbol{M}_{\text{K}}^{(j)}\right)^{\top} \boldsymbol{X}^{\top}\right\|_{\infty}$$

$$= \max_{k \in [T]} \left|\boldsymbol{x}_i^{\top} \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}_k - \boldsymbol{x}_i^{\top}\left(\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} \odot \boldsymbol{M}_{\text{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} \odot \boldsymbol{M}_{\text{K}}^{(j)}\right)^{\top} \boldsymbol{x}_k\right|$$

$$\leq \alpha^2 \left\|\boldsymbol{W}_{\text{QK}}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} \odot \boldsymbol{M}_{\text{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} \odot \boldsymbol{M}_{\text{K}}^{(j)}\right)^{\top}\right\|$$

$$\leq \alpha^2 d_1 \left\|\boldsymbol{W}_{\text{QK}}^{(j)} - \left(\tilde{\boldsymbol{W}}_{\text{Q}}^{\prime(j)} \odot \boldsymbol{M}_{\text{Q}}^{(j)}\right)\left(\tilde{\boldsymbol{W}}_{\text{K}}^{\prime(j)} \odot \boldsymbol{M}_{\text{K}}^{(j)}\right)^{\top}\right\|_{\max}$$

$$\leq \alpha^2 d_1 \frac{\epsilon}{8H\alpha^3 d_1^{3/2}}$$

$$= \frac{\epsilon}{8H\alpha\sqrt{d_1}}.$$

Let

$$\boldsymbol{p}_i^{(j)} := \sigma\left(\boldsymbol{x}_i^{\top} \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{X}^{\top}; \boldsymbol{a}_i\right),$$

$$\boldsymbol{p}_i'^{(j)} := \sigma\left(\boldsymbol{x}_i^\top (\tilde{\boldsymbol{W}}_{\mathrm{Q}}'^{(j)} \odot \boldsymbol{M}_{\mathrm{Q}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{K}}'^{(j)} \odot \boldsymbol{M}_{\mathrm{K}}^{(j)})^\top \boldsymbol{X}^\top; \boldsymbol{a}_i\right).$$

Applying Lemma 8, we obtain

$$\left\|\boldsymbol{p}_i^{(j)}\boldsymbol{X} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}\right\| \le \sqrt{d_1}\alpha\left(\exp\left(\frac{\epsilon}{4H\alpha\sqrt{d_1}}\right) - 1\right)$$

$$\le \frac{\epsilon}{2H}. \qquad\qquad \left(\text{Using } 0 < \tfrac{\epsilon}{4H\alpha\sqrt{d_1}} < 1\right)$$

For the value and output weights, from Lemma 7, if

$$n_{\mathrm{V}} \ge d_1 C \log\left(\frac{2H\alpha d_1\sqrt{d_2}}{\epsilon_2}\right),$$

then with probability at least $1 - \frac{\sqrt{d_2}\epsilon}{2H\alpha}$, there exists a choice of binary pruning masks $\boldsymbol{M}_{\mathrm{V}}^{(j)}$ and $\boldsymbol{M}_{\mathrm{O}}^{(j)}$ such that

$$\left\|\boldsymbol{W}_{\mathrm{VO}}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\|_{\max} \le \frac{\epsilon}{2H\alpha d_1\sqrt{d_2}}. \tag{15}$$

Step 3: Total Error Analysis.   We now bound the difference between the outputs of the source and target MHAs:

$$\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i)\| = \left\|\sum_{j=1}^{H}\left(\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right)\right\|$$

$$\le \sum_{j=1}^{H}\left\|\left(\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right)\right\|.$$

We apply the triangle inequality:

$$\left\|\boldsymbol{p}_i^{(j)}\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)} - \boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\|$$

$$\le \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)}\right\| + \left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\boldsymbol{W}_{\mathrm{VO}}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\|.$$

For the first term, by using $\|\boldsymbol{W}_{\mathrm{VO}}^{(j)}\| \le 1$ (Equation (13)), we obtain

$$\left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)}\right\| \le \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\right\|\left\|\boldsymbol{W}_{\mathrm{VO}}^{(j)}\right\|$$

$$\le \left\|(\boldsymbol{p}_i^{(j)} - \boldsymbol{p}_i'^{(j)})\boldsymbol{X}\right\|$$

$$\le \frac{\epsilon}{2H}. \tag{16}$$

For the second term, we obtain the following result by using Equation (15):

$$\left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}(\boldsymbol{W}_{\mathrm{VO}}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)}))\right\|$$

$$\le \sqrt{d_1}\left\|\boldsymbol{p}_i'^{(j)}\boldsymbol{X}\right\|_\infty \sqrt{d_1 d_2}\left\|\boldsymbol{W}_{\mathrm{VO}}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\|_{\max}$$

$$\le d_1\sqrt{d_2}\alpha\left\|\boldsymbol{W}_{\mathrm{VO}}^{(j)} - (\tilde{\boldsymbol{W}}_{\mathrm{V}}^{(j)} \odot \boldsymbol{M}_{\mathrm{V}}^{(j)})(\tilde{\boldsymbol{W}}_{\mathrm{O}}^{(j)} \odot \boldsymbol{M}_{\mathrm{O}}^{(j)})\right\|_{\max}$$

$$\le d_1\sqrt{d_2}\alpha\frac{\epsilon}{2H\alpha d_1\sqrt{d_2}} \qquad\qquad \text{(Using Equation (15))}$$

$$= \frac{\epsilon}{2H}. \tag{17}$$

These results of Equations (16) and (17) do not depend on the input index i; thus, adding the two terms across H heads gives

$$\max_{i\in[T]}\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{S}}(\boldsymbol{x}_i)\| \le \sum_{j=1}^{H}\left(\frac{\epsilon}{2H} + \frac{\epsilon}{2H}\right) = \epsilon.$$

Finally, using the union bound and the assumption $\alpha \ge \max(\sqrt{d_1}, \sqrt{d_2})$, the probability that all approximations hold is

$$1 - \frac{\epsilon}{8H\alpha^3\sqrt{d_1}} - \frac{\sqrt{d_2}\epsilon}{2H\alpha} \ge 1 - \epsilon.$$

$\square$

A.4 SLT Existence Within Transformer Blocks

By combining the SLT existence theorem for attention mechanisms (Theorem 9) and for multi-layer fully-connected ReLU networks (FC) proven by Pensia et al. (2020) (Theorem 10), we prove the SLT existence theorem for transformer blocks.

Theorem 10 (Theorem 1 in Pensia et al. (2020)). *Let*

$$\mathrm{F_T}\left(\boldsymbol{x}_i\right) = \boldsymbol{W}_L \mathrm{ReLU}(\boldsymbol{W}_{L-1} \ldots \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{x}_i))$$

be a target FC with L layers. Assume that each weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{d_{l+1} \times d_l}$ satisfies $|\boldsymbol{W}_l| \leq 1$ for all $l = 1, \ldots, L$. Consider a pruned source FC with 2L layers defined as

$$\mathrm{F_S}\left(\boldsymbol{x}_i\right) = \left(\tilde{\boldsymbol{W}}_{2L} \odot \boldsymbol{M}_{2L}\right) \mathrm{ReLU}\left(\left(\tilde{\boldsymbol{W}}_{2L-1} \odot \boldsymbol{M}_{2L-1}\right) \ldots \mathrm{ReLU}\left(\left(\tilde{\boldsymbol{W}}_{2L-1} \odot \boldsymbol{M}_{2L-1}\right) \boldsymbol{x}_i\right)\right),$$

where $\tilde{\boldsymbol{W}}_{2l-1} \in \mathbb{R}^{n_l \times d_l}$ and $\tilde{\boldsymbol{W}}_{2l} \in \mathbb{R}^{d_{l+1} \times n_l}$ for $l = 1, \ldots, L$, and each entry of $\tilde{\boldsymbol{W}}_l$ is drawn i.i.d. from $U[-1, 1]$. Then, with probability at least $1 - \epsilon$ for any $0 < \epsilon < 1$, there exists a choice of binary pruning masks $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{2L}$ that holds the following inequality:

$$\|\mathrm{F_S}(\boldsymbol{x}_i) - \mathrm{F_T}(\boldsymbol{x}_i)\| \leq \exp\left(\frac{\alpha\epsilon}{2}\right) - 1,$$

if each source dimension n_l satisfies

$$n_l \geq d_l C \log \frac{4L d_l d_{l+1}}{\epsilon},$$

for some universal constant $C > 0$.

We now state the main result for transformer blocks, which follows from combining the two SLT existence theorems.

Theorem 11. *Let*

$$\mathrm{Blk_T}(\boldsymbol{x}_i) = \mathrm{F_T}(\mathrm{Attn_T}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) + \mathrm{Attn_T}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i$$

be a target transformer block of an MHA $\mathrm{Attn_T}(\cdot)$ and FC $\mathrm{F_T}(\cdot)$ with L layers. For simplicity, we assume each layer of target FC dimensions is all d_1. Let

$$\mathrm{Blk_S}(\boldsymbol{x}_i) = \mathrm{F_S}(\mathrm{Attn_S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) + \mathrm{Attn_S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i$$

be a pruned random source transformer block of a pruned random MHA $\mathrm{Attn_S}(\cdot)$ and FC $\mathrm{F_S}(\cdot)$ with 2L layers. Assume that the input dimension of each even layer of the source FC is n_{FC}, and the input dimension of each odd layer is d_1. Furthermore, for simplicity, we assume key and value dimensions of $\mathrm{Attn_S}(\cdot)$ are the same dimension n_{MHA}. Then, with probability at least $1 - \epsilon$ for $0 < \epsilon < 1$, there exists a choice of binary masks that satisfies

$$\max_{i \in [T]} \|\mathrm{Blk_S}(\boldsymbol{x}_i) - \mathrm{Blk_T}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the hidden dimensions of the source MHA and FC satisfy

$$n_{\mathrm{MHA}} \geq d_1 C \log\left(\frac{32\alpha^3 H d_1^{\frac{3}{2}}}{\epsilon}\right),$$

$$n_{\mathrm{FC}} \geq d_1 C \log\left(\frac{24\alpha L H d_1^{\frac{5}{2}}}{\epsilon}\right),$$

for some universal constant $C > 0$.

Proof. Our proof strategy is first to apply the SLT existence theorem to the attention mechanism, and then apply the result for FCs. From Theorem 9, with probability at least $1 - \frac{\epsilon}{4}$, there exists a choice of binary masks so that $\mathrm{Attn_S}(\cdot)$ satisfies

$$\|\mathrm{Attn_S}(\boldsymbol{x}_i) - \mathrm{Attn_T}(\boldsymbol{x}_i)\| \leq \frac{\epsilon}{4}. \tag{18}$$

This inequality Equation (18) implies

$$\|\mathrm{Attn_S}(\boldsymbol{x}_i) - \mathrm{Attn_T}(\boldsymbol{x}_i)\| \leq \frac{\epsilon}{4},$$

$$\implies \|\mathrm{Attn_S}(\boldsymbol{x}_i)\| \leq \frac{\epsilon}{4} + \|\mathrm{Attn_T}(\boldsymbol{x}_i)\| \leq \frac{\epsilon}{4} + \alpha H \sqrt{d_1}.$$

Therefore, the norm of the input vector of the source FC satisfies

$$\|\mathrm{Attn_S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}\| \leq \|\mathrm{Attn_S}(\boldsymbol{x}_i)\| + \alpha$$

$$\leq \frac{\epsilon}{4} + \alpha(H\sqrt{d_1} + 1)$$
$$\leq 3\alpha H\sqrt{d_1}. \tag{19}$$

Assume that this upper bound of Equation (19) holds. Now, applying Theorem 10 to the source FC, if

$$n_{\text{FC}} \geq d_1 C \log\left(\frac{4Ld_1^2 \cdot 6\alpha H\sqrt{d_1}}{\epsilon}\right)$$
$$= d_1 C \log\left(\frac{24\alpha LHd_1^{\frac{5}{2}}}{\epsilon}\right),$$

with probability at least $1 - \frac{\epsilon}{6\alpha H\sqrt{d_1}}$, there exists a choice of binary pruning masks so that $\text{F}_\text{S}(\cdot)$ satisfies

$$\|\text{F}_\text{S}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) - \text{F}_\text{T}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i)\|$$
$$\leq \exp\left(\frac{3\alpha H\sqrt{d_1}\epsilon}{2 \cdot 6\alpha H\sqrt{d_1}}\right) - 1$$
$$= \exp\left(\frac{\epsilon}{4}\right) - 1.$$

Finally, we bound the total error between the source and target transformer blocks:

$$\max_{i\in[T]} \|\text{Blk}_\text{S}(\boldsymbol{x}_i) - \text{Blk}_\text{T}(\boldsymbol{x}_i)\| = \max_{i\in[T]} \|\text{F}_\text{S}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) + \text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x} - \text{F}_\text{T}(\text{Attn}_\text{T}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) - \text{Attn}_\text{T}(\boldsymbol{x}_i)^\top - \boldsymbol{x}\|$$
$$\leq \max_{i\in[T]}\Big(\|\text{F}_\text{S}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) - \text{F}_\text{T}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i)\|$$
$$+ \|\text{F}_\text{T}(\text{Attn}_\text{S}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) - \text{F}_\text{T}(\text{Attn}_\text{T}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i)\| + \|\text{Attn}_\text{S}(\boldsymbol{x}_i) - \text{Attn}_\text{T}(\boldsymbol{x}_i)\|\Big)$$
$$\leq \exp\left(\frac{\epsilon}{4}\right) - 1 + \max_{i\in[T]} 2\|\text{Attn}_\text{S}(\boldsymbol{x}_i) - \text{Attn}_\text{T}(\boldsymbol{x}_i)\|$$
$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}$$
$$= \epsilon.$$

From a union bound, the probability that this approximation holds is at least $1 - \epsilon$:

$$1 - \frac{\epsilon}{4} - \frac{\epsilon}{6\alpha H\sqrt{d_1}} \geq 1 - \epsilon.$$

$\square$

A.5 SLT Existence Within Transformers Without Normalization Layers

By exploiting the SLT existence theorem for transformer blocks (Theorem 11), we prove the SLT existence theorem for transformers without normalization layers.

We firstly prove the two lemmas used in the proof of the theorem.

Lemma 12. *Let* $\boldsymbol{X}' = [\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_T]^\top$ *be a perturbed input matrix, which satisfies* $\max_{i\in[T]} \|\boldsymbol{x}_i - \boldsymbol{x}'_i\| \leq \epsilon_{\max}$ *Then, an arbitrary target MHA* $\text{Attn}_\text{T}(\cdot)$ *holds the following inequality:*

$$\|\text{Attn}_\text{T}(\boldsymbol{x}_i) - \text{Attn}_\text{T}(\boldsymbol{x}'_i)\| \leq H\sqrt{d_1}(\alpha(\exp(4\alpha\epsilon_{\max}) - 1) + \epsilon_{\max}).$$

Proof. We begin by analyzing the upper bound of differences for different inputs:

$$\|\boldsymbol{x}_i^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{X}^\top - \boldsymbol{x}_i'^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{X}'^\top\|_{\max} = \max_{k\in[T]} |\boldsymbol{x}_i^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}_k - \boldsymbol{x}_i'^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}'_k|$$
$$\leq \max_{k\in[T]} \left(|\boldsymbol{x}_i^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}_k - \boldsymbol{x}_i'^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}_k| + |\boldsymbol{x}_i'^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}_k - \boldsymbol{x}_i'^\top \boldsymbol{W}_{\text{QK}}^{(j)} \boldsymbol{x}'_k|\right)$$
$$\leq \max_{k\in[T]} \left(\|\boldsymbol{x}_i - \boldsymbol{x}'_i\|\|\boldsymbol{W}_{\text{QK}}^{(j)}\|\|\boldsymbol{x}_k\| + \|\boldsymbol{x}'_i\|\|\boldsymbol{W}_{\text{QK}}^{(j)}\|\|\boldsymbol{x}_k - \boldsymbol{x}'_k\|\right)$$
$$\leq \alpha\epsilon_{\max} + \alpha\epsilon_{\max}$$
$$= 2\alpha\epsilon_{\max}.$$

Applying Lemma 8, the following inequality holds:

$$\|\sigma(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i)\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)} - \sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X}'\boldsymbol{W}_{\mathrm{VO}}^{(j)}\|$$

$$\leq \|\sigma(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X}\| + \|\sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X}'\|$$

$$= \|\sigma(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i)\boldsymbol{X} - \sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X}\| + \|\sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)(\boldsymbol{X} - \boldsymbol{X}')\|$$

$$\leq \sqrt{d_1}\alpha\left(\exp(4\alpha\epsilon_{\max}) - 1\right) + \sqrt{d_1}\epsilon_{\max}.$$

Then, we have the following bound:

$$\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}')\| = \|\sum_{j=1}^{H} \sigma(\boldsymbol{x}_i^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}^\top; \boldsymbol{a}_i)\boldsymbol{X}\boldsymbol{W}_{\mathrm{VO}}^{(j)} - \sum_{j=1}^{H} \sigma(\boldsymbol{x}_i'^\top \boldsymbol{W}_{\mathrm{QK}}^{(j)} \boldsymbol{X}'^\top; \boldsymbol{a}_i)\boldsymbol{X}'\boldsymbol{W}_{\mathrm{VO}}^{(j)}\|$$

$$\leq H\sqrt{d_1}(\alpha\left(\exp(4\alpha\epsilon_{\max}) - 1\right) + \epsilon_{\max}).$$

$\square$

Lemma 13. *An arbitrary target Attention block* $\mathrm{Blk}_{\mathrm{T}}(\cdot)$ *holds the following inequality:*

$$\|\mathrm{Blk}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Blk}_{\mathrm{T}}(\boldsymbol{x}_i')\| \leq H\sqrt{d_1}(\alpha\left(\exp(4\alpha\epsilon_{\max}) - 1\right) + 2\epsilon_{\max}).$$

Proof. From Lemma 12, we have the upper bound as follows:

$$\|\mathrm{Blk}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Blk}_{\mathrm{T}}(\boldsymbol{x}_i')\| = \|\mathrm{F}_{\mathrm{T}}(\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) + \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i - \mathrm{F}_{\mathrm{T}}(\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i')^\top + \boldsymbol{x}_i') - \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i')^\top - \boldsymbol{x}_i'\|$$

$$\leq \|\mathrm{F}_{\mathrm{T}}(\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i)^\top + \boldsymbol{x}_i) - \mathrm{F}_{\mathrm{T}}(\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i')^\top + \boldsymbol{x}_i')\| + \|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i')\| + \|\boldsymbol{x}_i - \boldsymbol{x}_i'\|$$

$$\leq 2\|\mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i) - \mathrm{Attn}_{\mathrm{T}}(\boldsymbol{x}_i')\| + 2\|\boldsymbol{x}_i - \boldsymbol{x}_i'\|$$

$$\leq 2H\sqrt{d_1}(\alpha(\exp(4\alpha\epsilon_{\max}) - 1) + \epsilon_{\max}) + 2\epsilon_{\max}$$

$$\leq 2H\sqrt{d_1}(\alpha(\exp(4\alpha\epsilon_{\max}) - 1) + 2\epsilon_{\max})$$

$\square$

Theorem 14. *Assume* $B \geq 2$, *and let*

$$\mathrm{Tf}_{\mathrm{T}}(\boldsymbol{x}_i) := \mathrm{Blk}_{\mathrm{T}}^{(B)}(\mathrm{Blk}_{\mathrm{T}}^{(B-1)} \ldots \mathrm{Blk}_{\mathrm{T}}^{(1)}(\boldsymbol{x}_i))$$

be a target transformer with B *blocks. Let*

$$\mathrm{Tf}_{\mathrm{S}}(\boldsymbol{x}_i) := \mathrm{Blk}_{\mathrm{S}}^{(B)}(\mathrm{Blk}_{\mathrm{S}}^{(B-1)} \ldots \mathrm{Blk}_{\mathrm{S}}^{(1)}(\boldsymbol{x}_i))$$

be a pruned random transformer with B *layers. Then, with probability at least* $1 - \epsilon$ *for* $0 < \epsilon < 1$, *there exists a choice of binary masks that satisfies*

$$\max_{i \in [T]} \|\mathrm{Tf}_{\mathrm{S}}(\boldsymbol{x}_i) - \mathrm{Tf}_{\mathrm{T}}(\boldsymbol{x}_i)\| \leq \epsilon,$$

if the hidden dimensions of the b-th source MHA and FC satisfy

$$n_{\mathrm{MHA}}^{(b)} \geq d_1 C \log\left(\frac{c_1^{f_1(b,B)} H^{f_2(b,B)} d_1^{f_3(b,B)}}{\epsilon}\right)$$

$$n_{\mathrm{FC}}^{(b)} \geq d_1 C \log\left(\frac{c_2^{g_1(b,B)} L H^{g_2(b,B)} d_1^{g_3(b,B)}}{\epsilon}\right)$$

for some universal constant $C > 0$ *and constants* $c_1, c_2 > 0$ *including* α. *Here,* f_1, f_2, f_3 *and* g_1, g_2, g_3 *are quadratic functions of* B, b.

Proof. We analyze the approximation errors in each block sequentially and identify the accumulated error in the last block.

Notation for the Proof: Let $\boldsymbol{x}_i^{(b)}$ be an input vector to the b-th target block:

$$\boldsymbol{x}_i^{(b)} = \begin{cases} \boldsymbol{x}_i & \text{if } b = 1, \\ \mathrm{Blk}_\mathrm{T}^{(b-1)}(\boldsymbol{x}_i^{(b-1)}) & \text{if } 2 \le b \le B. \end{cases}$$

Then, the final output of the target transformer is $\mathrm{Tf}_\mathrm{T}(\boldsymbol{x}_i) = \mathrm{Blk}_\mathrm{T}(\boldsymbol{x}_i^{(B)})$. The spectral norm of these input vectors is

$$\|\boldsymbol{x}_i^{(b)}\| = \begin{cases} \|\boldsymbol{x}_i\| = \alpha & \\ \quad =: \beta_1 & \text{if } b = 1, \\ \|\mathrm{Blk}_\mathrm{T}^{(b-1)}(\boldsymbol{x}_i^{(b-1)})\| & \\ \quad \le 2(H\sqrt{d_1}+1)\|\boldsymbol{x}_i^{(b-1)}\| & \\ \quad \le \alpha(2(H\sqrt{d_1}+1))^{b-1} & \text{if } 2 \le b \le B. \\ \quad \le \alpha(4H\sqrt{d_1})^{b-1} & \\ \quad =: \beta_b & \end{cases}$$

Similarly, let $\boldsymbol{x}_i^{\prime(b)}$ be an input vector to the b-th source block:

$$\boldsymbol{x}_i^{\prime(b)} = \begin{cases} \boldsymbol{x}_i & \text{if } b = 1, \\ \mathrm{Blk}_\mathrm{S}^{(b-1)}(\boldsymbol{x}_i^{\prime(b-1)}) & \text{if } 2 \le b \le B. \end{cases}$$

Then, the final output of the source transformer is $\mathrm{Tf}_\mathrm{S}(\boldsymbol{x}_i) = \mathrm{Blk}_\mathrm{S}(\boldsymbol{x}_i^{\prime(B)})$.

First Block Error: From Theorem 11, if

$$n_{\mathrm{MHA}}^{(1)} \ge d_1 C \log\left(\frac{32\beta_1^3 H d_1^{\frac{3}{2}} 2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2}{\epsilon} \right),$$

$$n_{\mathrm{FC}}^{(1)} \ge d_1 C \log\left(\frac{24\beta_1 L H d_1^{\frac{5}{2}} 2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2}{\epsilon} \right),$$

then with probability at least $1 - \frac{\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2}$, the following inequality holds independently of the input index i:

$$\|\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{\prime(2)}\| = \|\mathrm{Blk}_\mathrm{S}^{(1)}(\boldsymbol{x}_i) - \mathrm{Blk}_\mathrm{T}^{(1)}(\boldsymbol{x}_i)\|$$
$$\le \frac{\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2}. \tag{20}$$

This inequality Equation (20) implies the upper bound of $\|\boldsymbol{x}_i^{\prime(2)}\|$:

$$\|\boldsymbol{x}_i^{(2)} - \boldsymbol{x}_i^{\prime(2)}\| \le \frac{\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2}$$
$$\implies \|\boldsymbol{x}_i^{\prime(2)}\| \le \frac{\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2} + \|\boldsymbol{x}_i^{(2)}\| \qquad \text{(From the triangle inequality.)}$$
$$\le \frac{\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2} + \beta_2$$
$$\le 2\beta_2. \tag{21}$$

Second Block Error: We assume the approximation of the first block is successful (i.e., Equation (21) holds). Then, from Lemma 13, the following bound holds:

$$\|\mathrm{Blk}_\mathrm{T}^{(2)}(\boldsymbol{x}_i^{(2)}) - \mathrm{Blk}_\mathrm{T}^{(2)}(\boldsymbol{x}_i^{\prime(2)})\|$$
$$\le H\sqrt{d_1}\left(\beta_2\left(\exp\left(\frac{4\beta_2\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2} \right) - 1 \right) + \frac{2\epsilon}{2^{B-1} \prod_{j=2}^{B} 16H\sqrt{d_1}\beta_j^2} \right)$$

$$= H\sqrt{d_1}\left(\beta_2\exp\left(\frac{1}{4H\sqrt{d_1}\beta_2}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}\right) - \beta_2 + \frac{1}{8H\sqrt{d_1}\beta_2}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}\right)$$

$$\leq H\sqrt{d_1}\left(\frac{1}{2H\sqrt{d_1}}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} + \frac{1}{8H\sqrt{d_1}\beta_2}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}\right) \quad (\exp(x)\leq 2x+1 \text{ if } 0\leq x\leq 1.)$$

$$= \frac{1}{2}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} + \frac{1}{8\beta_2}\frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}$$

$$\leq \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}.$$

From Theorem 11, if

$$n_{\mathrm{MHA}}^{(2)} \geq d_1 C\log\left(\frac{32(2\beta_2)^3 H d_1^{\frac{3}{2}}2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}{\epsilon}\right)$$

$$n_{\mathrm{FC}}^{(2)} \geq d_1 C\log\left(\frac{24(2\beta_2)LH d_1^{\frac{5}{2}}2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}{\epsilon}\right),$$

then with probability at least $1 - \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}$, the following inequality holds:

$$\|\mathrm{Blk}_{\mathrm{S}}^{(2)}(\boldsymbol{x}_i'^{(2)}) - \mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i'^{(2)})\| \leq \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}.$$

Therefore, we have

$$\|\boldsymbol{x}_i^{(3)} - \boldsymbol{x}_3'^{(3)}\| = \|\mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i^{(2)}) - \mathrm{Blk}_{\mathrm{S}}^{(2)}(\boldsymbol{x}_i'^{(2)})\|$$

$$= \|\mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i^{(2)}) - \mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i'^{(2)}) + \mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i'^{(2)}) - \mathrm{Blk}_{\mathrm{S}}^{(2)}(\boldsymbol{x}_i'^{(2)})\|$$

$$\leq \|\mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i^{(2)}) - \mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i'^{(2)})\| + \|\mathrm{Blk}_{\mathrm{T}}^{(2)}(\boldsymbol{x}_i'^{(2)}) - \mathrm{Blk}_{\mathrm{S}}^{(2)}(\boldsymbol{x}_i'^{(2)})\|$$

$$\leq \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} + \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}$$

$$= \frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}. \tag{22}$$

This inequality Equation (22) implies the upper bound of $\|\boldsymbol{x}_i'^{(3)}\|$:

$$\|\boldsymbol{x}_i^{(3)} - \boldsymbol{x}_i'^{(3)}\| \leq \frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}$$

$$\implies \|\boldsymbol{x}_i'^{(3)}\| \leq \frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} + \|\boldsymbol{x}_i^{(3)}\| \qquad \text{(From the triangle inequality.)}$$

$$\leq \frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} + \beta_3$$

$$\leq 2\beta_3. \tag{23}$$

Third Block Error: We assume the approximation of second block is succeessful (i.e., Equation (23) holds). Then, from Lemma 13, the following bound holds:

$$\|\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i^{(3)}) - \mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})\|$$

$$\leq H\sqrt{d_1}\left(\beta_3\left(\exp\left(4\beta_3\frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}\right) - 1\right) + 2\frac{\epsilon}{2^{B-2}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2}\right)$$

$$= H\sqrt{d_1}\left(\beta_3\left(\exp\left(\frac{1}{4H\sqrt{d_1}\beta_3}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}\right) - 1\right) + \frac{1}{8H\sqrt{d_1}\beta_3}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}\right)$$

$$\leq H\sqrt{d_1}\left(\frac{1}{2H\sqrt{d_1}}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}+\frac{1}{8H\sqrt{d_1}\beta_3}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}\right)\quad(\exp(x)\leq 2x+1\text{ if }0\leq x\leq 1.)$$

$$=\frac{1}{2}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}+\frac{1}{8\beta_3}\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}$$

$$\leq\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}.$$

From Theorem 11, if

$$n_{\mathrm{MHA}}^{(3)}\geq d_1 C\log\left(\frac{32(2\beta_3)^3 Hd_1^{\frac{3}{2}}2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}{\epsilon}\right)$$

$$n_{\mathrm{FC}}^{(3)}\geq d_1 C\log\left(\frac{24(2\beta_3)LHd_1^{\frac{5}{2}}2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}{\epsilon}\right),$$

then with probability at least $1-\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}$, the following inequality holds:

$$\|\mathrm{Blk}_{\mathrm{S}}^{(3)}(\boldsymbol{x}_i'^{(3)})-\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})\|\leq\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}.$$

Therefore, we have

$$\|\boldsymbol{x}_i^{(4)}-\boldsymbol{x}_3'^{(4)}\|$$
$$=\|\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i^{(3)})-\mathrm{Blk}_{\mathrm{S}}^{(3)}(\boldsymbol{x}_i'^{(3)})\| \tag{24}$$
$$=\|\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i^{(3)})-\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})+\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})-\mathrm{Blk}_{\mathrm{S}}^{(3)}(\boldsymbol{x}_i'^{(3)})\|$$
$$\leq\|\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i^{(3)})-\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})\|+\|\mathrm{Blk}_{\mathrm{T}}^{(3)}(\boldsymbol{x}_i'^{(3)})-\mathrm{Blk}_{\mathrm{S}}^{(3)}(\boldsymbol{x}_i'^{(3)})\|$$
$$\leq\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}+\frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}$$
$$=\frac{\epsilon}{2^{B-3}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}. \tag{25}$$

This inequality Equation (25) implies the upper bound of $\|\boldsymbol{x}_i'^{(4)}\|$:

$$\|\boldsymbol{x}_i^{(4)}-\boldsymbol{x}_i'^{(4)}\|\leq\frac{\epsilon}{2^{B-3}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}$$
$$\implies\|\boldsymbol{x}_i'^{(4)}\|\leq\frac{\epsilon}{2^{B-3}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}+\|\boldsymbol{x}_i^{(4)}\|\quad\text{(From the triangle inequality.)}$$
$$\leq\frac{\epsilon}{2^{B-3}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2}+\beta_4$$
$$\leq 2\beta_4. \tag{26}$$

$(B-1)$**-th Block Error:** By repeating the same proof procedure as above in each block, we can propagate the error to $(B-1)$-th block. We assume all first-to-$(B-2)$-th block approximations are successful. Then, from Lemma 13, the following bound holds:

$$\|\mathrm{Blk}_{\mathrm{T}}^{(B-1)}(\boldsymbol{x}_i^{(B-1)})-\mathrm{Blk}_{\mathrm{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\|$$

$$\leq H\sqrt{d_1}\left(\beta_{B-1}\left(\exp\left(4\beta_{B-1}\frac{\epsilon}{2^2\prod_{j=B-1}^{B}16H\sqrt{d_1}\beta_j^2}\right)-1\right)+2\frac{\epsilon}{2^2\prod_{j=B-1}^{B}16H\sqrt{d_1}\beta_j^2}\right)$$

$$=H\sqrt{d_1}\left(\beta_{B-1}\left(\exp\left(\frac{1}{4H\sqrt{d_1}\beta_{B-1}}\frac{\epsilon}{2^2\cdot 16H\sqrt{d_1}\beta_B^2}\right)-1\right)+\frac{1}{8H\sqrt{d_1}\beta_{B-1}}\frac{\epsilon}{2^2\cdot 16H\sqrt{d_1}\beta_B^2}\right)$$

$$\leq H\sqrt{d_1}\left(\frac{1}{2H\sqrt{d_1}}\frac{\epsilon}{2^2\cdot 16H\sqrt{d_1}\beta_B^2}+\frac{1}{8H\sqrt{d_1}\beta_{B-1}}\frac{\epsilon}{2^2\cdot 16H\sqrt{d_1}\beta_B^2}\right)\quad(\exp(x)\leq 2x+1\text{ if }0\leq x\leq 1.)$$

$$= \frac{1}{2} \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2} + \frac{1}{8\beta_{B-1}} \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2}$$

$$\leq \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2}.$$

From Theorem 11, if

$$n_{\text{MHA}}^{(B-1)} \geq d_1 C \log \left(\frac{32(2\beta_{B-1})^3 H d_1^{\frac{3}{2}} 2^2 \cdot 16H\sqrt{d_1}\beta_B^2}{\epsilon} \right)$$

$$n_{\text{FC}}^{(B-1)} \geq d_1 C \log \left(\frac{24(2\beta_{B-1}) L H d_1^{\frac{5}{2}} 2^2 \cdot 16H\sqrt{d_1}\beta_B^2}{\epsilon} \right),$$

then with probability at least $1 - \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2}$, the following inequality holds independently of the input index i:

$$\|\text{Blk}_{\text{S}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)}) - \text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\| \leq \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_j^2}.$$

Therefore, we have the following inequality:

$$\|\boldsymbol{x}_i^{(B)} - \boldsymbol{x}_3'^{(B)}\| = \|\text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i^{(B-1)}) - \text{Blk}_{\text{S}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\|$$

$$= \|\text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i^{(B-1)}) - \text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)}) + \text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)}) - \text{Blk}_{\text{S}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\|$$

$$\leq \|\text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i^{(B-1)}) - \text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\| + \|\text{Blk}_{\text{T}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)}) - \text{Blk}_{\text{S}}^{(B-1)}(\boldsymbol{x}_i'^{(B-1)})\|$$

$$\leq \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2} + \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_B^2}$$

$$= \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2}. \tag{27}$$

This inequality Equation (27) implies the upper bound of $\|\boldsymbol{x}_i'^{(B)}\|$:

$$\|\boldsymbol{x}_i^{(B)} - \boldsymbol{x}_i'^{(B)}\| \leq \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2}$$

$$\implies \|\boldsymbol{x}_i'^{(B)}\| \leq \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2} + \|\boldsymbol{x}_i^{(B)}\| \qquad \text{(From the triangle inequality.)}$$

$$\leq \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2} + \beta_B$$

$$\leq 2\beta_B. \tag{28}$$

Final Block Error: We assume all first-to $(B-1)$-th block approximations are successful. Then, from Lemma 13, the following bound holds:

$$\|\text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i^{(B)}) - \text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)})\|$$

$$\leq H\sqrt{d_1} \left(\beta_B \left(\exp \left(4\beta_B \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2} \right) - 1 \right) + 2 \frac{\epsilon}{2 \cdot 16H\sqrt{d_1}\beta_B^2} \right)$$

$$= H\sqrt{d_1} \left(\beta_B \left(\exp \left(\frac{1}{8H\sqrt{d_1}\beta_B} \epsilon \right) - 1 \right) + \frac{1}{16H\sqrt{d_1}\beta_B} \epsilon \right)$$

$$\leq H\sqrt{d_1} \left(\frac{1}{4H\sqrt{d_1}} \epsilon + \frac{1}{16H\sqrt{d_1}\beta_B} \epsilon \right) \qquad (\exp(x) \leq 2x + 1 \text{ if } 0 \leq x \leq 1.)$$

$$= \frac{1}{4}\epsilon + \frac{1}{16\beta_B} \epsilon$$

$$\leq \frac{\epsilon}{2}.$$

From Theorem 11, if

$$n_{\text{MHA}}^{(B)} \geq d_1 C \log \left(\frac{32(2\beta_B)^3 H d_1^{\frac{3}{2}} \cdot 2}{\epsilon} \right),$$

$$n_{\text{FC}}^{(B)} \geq d_1 C \log\left(\frac{24(2\beta_B)LHd_1^{\frac{5}{2}} \cdot 2}{\epsilon}\right),$$

then with probability at least $1 - \frac{\epsilon}{2}$, the following inequality holds independently of the input index i:

$$\|\text{Blk}_{\text{S}}^{(B)}(\boldsymbol{x}_i'^{(B)}) - \text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)})\| \leq \frac{\epsilon}{2}.$$

Therefore, we finally obtain the following inequality:

$$
\begin{aligned}
\|\text{Tf}_{\text{T}}(\boldsymbol{x}_i) - \text{Tf}_{\text{S}}(\boldsymbol{x}_i)\| &= \|\text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i^{(B)}) - \text{Blk}_{\text{S}}^{(B)}(\boldsymbol{x}_i'^{(B)})\| \\
&= \|\text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i^{(B)}) - \text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)}) + \text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)}) - \text{Blk}_{\text{S}}^{(B)}(\boldsymbol{x}_i'^{(B)})\| \\
&\leq \|\text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i^{(B)}) - \text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)})\| + \|\text{Blk}_{\text{T}}^{(B)}(\boldsymbol{x}_i'^{(B)}) - \text{Blk}_{\text{S}}^{(B)}(\boldsymbol{x}_i'^{(B)})\| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon.
\end{aligned}
$$

Success Probability of the Approximation: By the union bound, the probability that $\|\text{Tf}_{\text{T}}(\boldsymbol{x}_i) - \text{Tf}_{\text{S}}(\boldsymbol{x}_i)\| \leq \epsilon$ holds is at least $1 - \epsilon$:

$$
1 - \frac{\epsilon}{2^{B-1}\prod_{j=2}^{B}16H\sqrt{d_1}\beta_j^2} - \frac{\epsilon}{2^{B-1}\prod_{j=3}^{B}16H\sqrt{d_1}\beta_j^2} - \frac{\epsilon}{2^{B-2}\prod_{j=4}^{B}16H\sqrt{d_1}\beta_j^2} - \cdots - \frac{\epsilon}{2^2 \cdot 16H\sqrt{d_1}\beta_j^2} - \frac{\epsilon}{2}
$$

$$
= 1 - \frac{2 + \sum_{k=2}^{B}\prod_{j=2}^{k}32H\sqrt{d_1}\beta_j^2}{2\prod_{j=2}^{B}32H\sqrt{d_1}\beta_j^2}\epsilon
$$

$$
= 1 - \frac{1}{2\prod_{j=2}^{B}32H\sqrt{d_1}\beta_j^2}\epsilon - \frac{1 + \sum_{k=2}^{B}\prod_{j=2}^{k}32H\sqrt{d_1}\beta_j^2}{2\prod_{j=2}^{B}32H\sqrt{d_1}\beta_j^2}\epsilon
$$

$$
\geq 1 - \frac{1}{64}\epsilon - \frac{1 + \sum_{k=2}^{B}\prod_{j=2}^{k}32H\sqrt{d_1}\beta_j^2}{2\prod_{j=2}^{B}32H\sqrt{d_1}\beta_j^2}\epsilon
$$

$$
= 1 - \frac{1}{64}\epsilon - \frac{1 + \sum_{k=2}^{B}(32H\alpha^2\sqrt{d_1})^{k-1}\prod_{j=2}^{k}(16H^2d_1)^{j-1}}{2(32H\alpha^2\sqrt{d_1})^{B-1}\prod_{j=2}^{k}(16H^2d_1)^{j-1}}\epsilon
$$

$$
= 1 - \frac{1}{64}\epsilon - \frac{1 + \sum_{k=2}^{B}(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}k})^{k-1}}{2(32H\alpha^2\sqrt{d_1})^{B-1}(16H^2d_1)^{\frac{1}{2}B(B-1)}}\epsilon
$$

$$
\geq 1 - \frac{1}{64}\epsilon - \frac{\sum_{k=1}^{B}(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B})^{k-1}}{2(32H\alpha^2\sqrt{d_1})^{B-1}(16H^2d_1)^{\frac{1}{2}B(B-1)}}\epsilon
$$

$$
\geq 1 - \frac{1}{64}\epsilon - \frac{(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B})^B - 1}{(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B} - 1)} \cdot \frac{1}{2(32H\alpha^2\sqrt{d_1})^{B-1}(16H^2d_1)^{\frac{1}{2}B(B-1)}}\epsilon
$$

$$
\geq 1 - \frac{1}{64}\epsilon - \frac{(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B})^B}{(32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B} - 1)} \cdot \frac{1}{2(32H\alpha^2\sqrt{d_1})^{B-1}(16H^2d_1)^{\frac{1}{2}B(B-1)}}\epsilon
$$

$$
= 1 - \frac{1}{64}\epsilon - \frac{1}{2\left(1 - \frac{1}{32H\alpha^2\sqrt{d_1}(16H^2d_1)^{\frac{1}{2}B}}\right)}\epsilon
$$

$$
\geq 1 - \frac{1}{64}\epsilon - \frac{1}{2\left(1 - \frac{1}{32}\right)}\epsilon
$$

$$
= 1 - \frac{1}{64}\epsilon - \frac{16}{32}\epsilon
$$

$$
= 1 - \frac{1055}{1984}\epsilon
$$

$$
\geq 1 - \epsilon.
$$

$\square$

B Experimental Details

This section describes the detailed experimental settings. All experiments can be verified with four NVIDIA H100 SXM5 94GB GPUs.

B.1 Synthetic Data Experiment

We construct a synthetic dataset for angular velocity estimation, where each input sequence consists of T two-dimensional vectors $x_1, ..., x_T$ such that $x_t = (\cos(\omega t + \theta_0), \sin(\omega t + \theta_0))$ for some angular velocity $\omega \in [-\pi, \pi]$ and initial phase $\theta_0 \in [0, \pi]$. The task is to estimate ω given the full sequence. Each sequence includes a special regression token—similar to the CLS token in BERT (Devlin et al. 2019)—at the beginning, and the model is trained to predict angular velocity by the regression token initialized to zero. We generate $10,000$ samples each for training, validation, and test sets, and input sequence lengths vary from 4 to 256 during training. We experiment with MHAs and transformers. In the MHA experiment, both the source and target MHAs are configured as single-head attention modules, with input and output dimensions of 2 and 1, respectively. The networks are trained using the AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 1024 and a learning rate of 0.1. Each target MHA is trained for 25 epochs with weight decay set to 0.01. In the transformer experiment, both the source and target models follow the construction described in Section 3.3. Each MHA has a single attention head, and both its input and output dimensions are set to 2. The same regression token is used for both the source and target models to ensure that the approximation quality reflects differences in the behavior of the models rather than token-level discrepancies. The query and key dimensions of the target models are set to 8. Target networks are initialized according to the assumptions of our theoretical results. Specifically, entries of the query and key projection matrices are drawn i.i.d. from $U[-n_K^{1/4}, n_K^{1/4}]$, and those of the value and output projection matrices from $U[-1, 1]$. The weights in fully-connected networks are also initialized with $U[-1, 1]$. Source networks are initialized with Xavier uniform distribution (Glorot and Bengio 2010). To identify SLTs, we use the weight approximation method in Lemma 4, based on the subset-sum approximation technique of Pensia et al. (2020). For each target network, we generate 100 source networks with random initialization and solve the associated subset-sum problem using Gurobi's mixed-integer programming solver (Gurobi Optimization, LLC 2024). In the experiments varying the hidden dimension, the input length is fixed at 4. We report the mean and standard deviation of the approximation error over these 100 candidates. We also fit exponential decay curves to the approximation error using SciPy (Virtanen et al. 2020).

B.2 Language Modeling Experiment

We further evaluate our theoretical framework in a practical language modeling setting. Here, we search for SLTs using the edge-popup algorithm (Ramanujan et al. 2020), which searches for accurate subnetworks by assigning scores to each connection and retaining only the top-$k\%$ entries during training. We set this k as 30. We train models from the GPT-2 family (Radford et al. 2019; Wolf et al. 2020) on the WikiText-103 dataset (Merity et al. 2017), using a maximum sequence length of 1024. The weights of these models are initialized based on the GPT-2 initialization scheme: they are drawn i.i.d. from a normal distribution with mean 0 and standard deviation 0.02. For the output projection in MHAs and the second layer of the fully-connected ReLU networks, the standard deviation is further scaled by $(2b)^{-1/2}$, where b is the number of transformer blocks. We train the models for 50 epochs, with 227 steps per epoch. The AdamW optimizer is used with an initial learning rate of 0.0001, which is decayed to 0.00001 via a cosine annealing scheduler (Loshchilov and Hutter 2017). A linear learning rate warm-up is applied during the first epoch. For each model size, we repeat training with three different random seeds and report the mean and standard deviation of the best performance.