Spectrum-Adaptive Generalization Bounds for Trained Deep Transformers

Mana Sakai[1,2] **Masaaki Imaizumi**[1,2,3]

[1]The University of Tokyo
[2]RIKEN Center for Advanced Intelligence Project
[3]Kyoto University

mana.sakai.77@gmail.com, imaizumi@g.ecc.u-tokyo.ac.jp

Abstract

Understanding why trained Transformers generalize well is a fundamental problem in modern machine learning theory, and complexity-based generalization bounds provide a principled way to study this question. While existing norm-based bounds for Transformers remove the explicit polynomial dependence on the hidden dimension, they typically impose fixed norm constraints specified a priori and can exhibit unfavorable exponential dependence on depth. In this paper, we derive spectrum-adaptive post hoc generalization bounds for multi-layer Transformers. Under layerwise spectral norm control, the bounds are expressed in terms of layerwise Schatten quantities of the query-key, value, and feedforward weight matrices. Since the Schatten indices need not be fixed a priori and can instead be selected after training, separately for each matrix type and layer, the bounds adaptively trade off spectral complexity against the dimension- and depth-dependent factors according to the learned singular-value profiles. Empirical comparisons of BERT-adapted proxies for the leading complexity factors suggest that the proxies induced by our bounds grow more slowly with depth and hidden dimension than the corresponding norm-based proxies. Overall, our results provide a complexity-based perspective on how the spectral structure of trained Transformers is reflected in generalization analyses.

1 Introduction

Transformers have become a central architecture of modern machine learning since their introduction by Vaswani et al. (2017). Their success across language (Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2023), vision (Dosovitskiy et al., 2021), and many other modalities has made it increasingly important to understand why heavily overparameterized Transformers can still generalize well. Generalization bounds offer one principled route toward this goal, because they can reveal which features of a learned model are statistically relevant.

For generalization analysis of Transformers, Edelman et al. (2022) established norm-based generalization gap bounds for multi-layer Transformers. A notable feature of their bounds is that the dependence on both the hidden dimension and the token length appears only through logarithmic factors. Trauger and Tewari (2024) sharpened this work by removing the explicit token-length dependence. More recently, Li et al. (2026) used the offset Rademacher complexity to derive excess risk bounds with an $O(1/n)$ rate, where n is the training sample size.

At the same time, these advances expose two difficulties that arise when existing norm-based bounds are used to assess trained deep Transformers. First, the improvement in explicit dimension dependence can come with layerwise propagation factors that can scale as C^L in the worst case, where L is the depth and C is determined by Lipschitz constants and spectral norm bounds across layers.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Leading complexity factor B in simplified generalization gap bounds of the form $\tilde{O}(B/\sqrt{n})$. All weights are $N \times N$ and satisfy spectral norm constraints $\|W\|_2 = O(1)$. Here L is the depth, n is the sample size, and C denotes the worst-case layerwise propagation factor appearing in existing norm-based bounds. For our post hoc bounds, the displayed expression is the common-p specialization of Theorem 3.1; the theorem itself allows a separate p for every matrix type and layer. While our bounds and the bounds of Edelman et al. (2022) depend on the token length logarithmically, the bounds of Trauger and Tewari (2024) are independent of the token length.

	Bound	Assumption	Post Hoc
Ours (Theorem 3.1)	$\inf_p(\|W\|_{\mathrm{s},p}^p)^{\frac{1}{p+2}} C^{\frac{Lp}{p+2}} L^{\frac{2p+2}{p+2}} N^{\frac{p+1}{p+2}}$		$\surd$
Edelman et al. (2022)	$C_{2,1} C^L L^{\frac{3}{2}}$	$\|W\|_{2,1} \le C_{2,1}$	
Trauger and Tewari (2024)	$C_{1,1} C^L L^{\frac{3}{2}}$	$\|W\|_{1,1} \le C_{1,1}$	

This issue is benign for shallow networks or when the propagation constant is at most one, but it can quickly make the bounds loose for deep models. Second, the fixed norm radii appearing in existing bounds are not naturally dimension-independent for large matrices. In particular, the bounds of Edelman et al. (2022) and Trauger and Tewari (2024) are controlled by mixed $(2,1)$- and $(1,1)$-norm radii, respectively, where the mixed (α,β)-norm is defined by $\|W\|_{\alpha,\beta} = [\sum_j (\sum_i |W_{ij}|^\alpha)^{\frac{\beta}{\alpha}}]^{\frac{1}{\beta}}$. Such quantities can remain small for matrices that are sparse in the corresponding coordinatewise sense, but they typically grow with the hidden dimension.

A further motivation comes from the observed heterogeneity of trained Transformer weights. Recent compression and spectral analyses suggest that different sublayers can have markedly different singular-value profiles: attention-related matrices often exhibit stronger low-rank structure or more compressible spectra than feedforward-related matrices (Li et al., 2024; Yuan et al., 2023). This suggests that a useful complexity measure should not impose a single global norm description on all layers and matrix types. Rather, it should be able to adapt to the spectral profile of each weight.

This point is especially important for post hoc generalization bounds. In post hoc bounds, the high-probability event is specified independently of the trained weights, while the admissible complexity parameters may be chosen after the weights are observed. Such guarantees are useful when the relevant structure of the trained model is not known a priori. In the context of DNNs and CNNs, Ledent et al. (2025) derived generalization bounds based on Schatten (quasi) norms that hold simultaneously over all choices of layerwise Schatten indices, so that these indices can be selected after training. Here, for a matrix W and a Schatten index $p \in (0,2]$, the Schatten p (quasi) norm is defined by $\|W\|_{\mathrm{s},p} = (\sum_i \sigma_i(W)^p)^{1/p}$, where $\sigma_i(W)$ is the i-th singular value. For Transformers, post hoc adaptivity is particularly natural because the spectral profiles of trained matrices may differ substantially across layers.

In this paper, we derive post hoc generalization gap bounds for multi-layer Transformers that can adaptively reflect the spectral structure of each learned weight. Specifically, our bounds are based on spectral quantities of the weights: besides the Schatten quantity $\|W\|_{\mathrm{s},p}^p$ ($p \in [0,2]$), with the convention $\|W\|_{\mathrm{s},0}^0 = \mathrm{rank}(W)$, we also use the spectral norm $\|W\|_2 = \sigma_1(W)$. Table 1 summarizes a specialized version of our result along with existing norm-based bounds. A notable feature of our bounds is that the complexity measure, through the Schatten index p, can be selected after training.

These bounds have several consequences. First, they provide a concrete interpolation between rank-based and norm-based regimes. When $p = 0$, the Schatten quantity is the rank; when $p = 2$, it is the squared Frobenius norm. Intermediate values of p describe a soft-rank measure through singular-value decay. Second, this interpolation directly controls the trade-off among spectral complexity, hidden dimension, and depth. Smaller values of p reduce the layerwise propagation factor $C^{Lp/(p+2)}$, the polynomial depth factor $L^{(2p+2)/(p+2)}$, and the hidden dimension $N^{(p+1)/(p+2)}$, whereas larger values of p provide a more Frobenius-like description. Because our bounds take an infimum over p after training, they automatically select the most favorable balance for each learned matrix. Third, using BERT Miniatures checkpoints of Turc et al. (2019), we find that the leading complexity factors suggested by our bounds grow substantially more slowly than the corresponding norm-based complexity factors of Edelman et al. (2022) as the depth or hidden dimension increases.

To obtain these spectrum-adaptive bounds, we derive covering number bounds that retain the dependence on the Schatten indices throughout the Transformer composition. Technically, we develop covering number bounds under layerwise spectral norm and Schatten-quantity constraints, building on the recently developed parametric interpolation (Ledent and Alves, 2024; Ledent et al., 2025). This technique decomposes each weight matrix into a low-rank leading part and a Frobenius-controlled tail, thereby allowing the bounds to interpolate between rank-based and norm-based regimes.

The contributions of this paper are summarized as follows.

- We prove post hoc generalization gap bounds for multi-layer Transformers under layerwise spectral norm control, where the complexity parameters can be selected separately for every layer and matrix type after training (Theorem 3.1).
- We extend parametric interpolation to matrix-valued function classes (Theorem 4.1) and use it as the basic building block for Transformer generalization bounds.
- We theoretically compare our bounds with existing norm-based bounds in representative regimes, including Frobenius norm or rank constraints, and show improved dependence on depth (Section 5; Table 2).
- We evaluate BERT-adapted leading complexity factors on BERT Miniatures checkpoints of Turc et al. (2019) and observe that our proxies grow more slowly than the corresponding norm-based proxies as the depth or hidden dimension increases (Section 6; Figure 1).

Related work is deferred to Appendix A.

2 Preliminaries

2.1 Notation

We use the following notation throughout the paper. We use $O(\cdot)$ and $\tilde{O}(\cdot)$ in the standard asymptotic sense, where $\tilde{O}(\cdot)$ suppresses polylogarithmic factors. We write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leq Cb$. For a vector v, we write v_i for its i-th entry. For a matrix W, we write W_{ij} for its (i, j)-th entry, and write $W_{i\cdot}$ and $W_{\cdot j}$ for its i-th row and j-th column, respectively. We denote by $\sigma_i(W)$ the i-th largest singular value of W.

We use the following matrix norms. First, $\|W\|_2 = \sigma_1(W)$ is the spectral norm. For $p \in (0, 2]$, the Schatten p (quasi) norm $\|W\|_{\mathrm{s},p}$ is defined by $\|W\|_{\mathrm{s},p} = (\sum_i \sigma_i(W)^p)^{1/p}$. We refer to p as the Schatten index. For the endpoint $p = 0$, we use the convention $\|W\|_{\mathrm{s},0}^0 := \mathrm{rank}(W)$. Finally, for $\alpha, \beta \geq 1$, we define the mixed (α, β)-norm by $\|W\|_{\alpha,\beta} = [\sum_j (\sum_i |W_{ij}|^\alpha)^{\beta/\alpha}]^{1/\beta}$.

2.2 Transformers

We specify the simplified Transformer architecture analyzed in this paper. Let $X \in \mathbb{R}^{T\times N}$ be an input matrix, where T is the token length and N is the hidden dimension. The model is obtained by composing Transformer blocks, each consisting of a single-head attention mechanism followed by a feedforward map and rowwise normalization.

Transformer head Let $W^{QK}, W^V \in \mathbb{R}^{N\times N}$ denote the combined query-key weight matrix and the value weight matrix, respectively. Define a Transformer head $f_{\mathrm{head}}(\,\cdot\,; W^{QK}, W^V) : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$ parameterized by W^{QK}, W^V by

$$f_{\mathrm{head}}(X; W^{QK}, W^V) = \mathrm{SoftMax}\left(XW^{QK}X^\top\right) XW^V, \tag{1}$$

where $\mathrm{SoftMax} : \mathbb{R}^{T\times T} \to \mathbb{R}^{T\times T}$ is applied rowwise:

$$(\mathrm{SoftMax}(Z))_{st} = \frac{\exp(Z_{st})}{\sum_{j=1}^T \exp(Z_{sj})} \qquad (s, t \in [T]).$$

Transformer block Following Edelman et al. (2022) and Trauger and Tewari (2024), we consider a normalized Transformer block. Let $\phi : \mathbb{R}^N \to \mathbb{R}^N$ be a fixed activation function in the feedforward layer, and apply it rowwise to matrices by setting $(\phi(Z))_{t\cdot} = \phi(Z_{t\cdot})$ for $Z \in \mathbb{R}^{T\times N}$. Let $\Pi_{\mathrm{norm}} : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$ be the rowwise projection onto the unit ball defined by $(\Pi_{\mathrm{norm}}(Z))_{t\cdot} = Z_{t\cdot}/(1 \vee \|Z_{t\cdot}\|_2)$, which keeps the row norms at most one. We define the Transformer block $f_{\mathrm{block}}(\,\cdot\,; W^{QK}, W^V, W^M) : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$ parameterized by $W^{QK}, W^V, W^M \in \mathbb{R}^{N\times N}$ as

$$f_{\mathrm{block}}(X; W^{QK}, W^V, W^M) = \Pi_{\mathrm{norm}}(\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X; W^{QK}, W^V)))W^M). \tag{2}$$

Multi-layer Transformer We introduce a multi-layer Transformer obtained by composing L Transformer blocks. For each $\ell \in [L]$, we write $W^{(\ell)} = (W^{QK,(\ell)}, W^{V,(\ell)}, W^{M,(\ell)})$ to denote the weight matrices for the ℓ-th Transformer block. Then we define

$$f_{\mathrm{tf}}^{(1)}(X; W^{(1)}) = f_{\mathrm{block}}(X; W^{QK,(1)}, W^{V,(1)}, W^{M,(1)}),$$

and, for $\ell \geq 2$,

$$f_{\mathrm{tf}}^{(\ell)}(X; W^{(1:\ell)}) = f_{\mathrm{block}}(f_{\mathrm{tf}}^{(\ell-1)}(X; W^{(1:\ell-1)}); W^{(\ell)}), \tag{3}$$

where $W^{(1:\ell)} = (W^{(1)}, \ldots, W^{(\ell)})$ denotes the weights corresponding to the first ℓ layers.

Scalar output of a Transformer We convert the matrix-valued output of the final Transformer layer into a scalar prediction by using a special classification token. Let $t_{\mathrm{CLS}} \in [T]$ denote the index of this token, and assume that the input $X \in \mathbb{R}^{T\times N}$ has already been augmented with the CLS token. For a matrix $Z \in \mathbb{R}^{T\times N}$, we write $[Z]_{[\mathrm{CLS}]} := (Z_{t_{\mathrm{CLS}}\cdot})^\top \in \mathbb{R}^N$ to denote the hidden representation at the CLS position. We then apply a trainable readout vector $w \in \mathbb{R}^N$ to the representation $[f_{\mathrm{tf}}^{(L)}(X; W^{(1:L)})]_{[\mathrm{CLS}]}$ and define

$$f_{\mathrm{out}}(X; W^{(1:L)}, w) = w^\top [f_{\mathrm{tf}}^{(L)}(X; W^{(1:L)})]_{[\mathrm{CLS}]}. \tag{4}$$

Class of Transformer outputs For each $\star \in \{QK, V, M\}$ and $\ell \in [L]$, define the class of parameters as

$$\mathcal{W}^{\star,(\ell)} = \{W \in \mathbb{R}^{N\times N} \mid \|W\|_2 \leq C_2^{\star,(\ell)}\}. \tag{5}$$

Set $\mathcal{W}^{(\ell)} = \mathcal{W}^{QK,(\ell)} \times \mathcal{W}^{V,(\ell)} \times \mathcal{W}^{M,(\ell)}$ and $\mathcal{W}^{(1:L)} = \mathcal{W}^{(1)} \times \cdots \times \mathcal{W}^{(L)}$. Then the scalar-output Transformer class is

$$\mathcal{F}_{\mathrm{out}} = \{f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) : \mathbb{R}^{T\times N} \to \mathbb{R} \mid W^{(1:L)} \in \mathcal{W}^{(1:L)}, w \in \mathbb{R}^N, \|w\|_2 \leq C_2^{\mathrm{out}}\}. \tag{6}$$

2.3 Generalization gap

Let $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a $B_{\mathcal{L}}$-bounded loss function that is $L_{\mathcal{L}}$-Lipschitz in its first argument. Suppose $\{(X_i, Y_i)\}_{i\in[n]}$ is an i.i.d. sample drawn from distribution $\mathcal{D}$ on $\mathbb{R}^{T\times N} \times \mathbb{R}$. For each $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}$, define the population risk and the empirical risk by

$$\mathcal{R}(f_{\mathrm{out}}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathcal{L}(f_{\mathrm{out}}(X), Y)], \qquad \hat{\mathcal{R}}_n(f_{\mathrm{out}}) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(f_{\mathrm{out}}(X_i), Y_i),$$

respectively. Then, the generalization gap is defined by

$$\mathrm{GAP}(f_{\mathrm{out}}) = |\mathcal{R}(f_{\mathrm{out}}) - \hat{\mathcal{R}}_n(f_{\mathrm{out}})|.$$

In this paper, we derive generalization bounds in which the high-probability event holds uniformly over a prescribed family of complexity parameters. In our setting, the post hoc parameters are the Schatten indices $\boldsymbol{p} = (p^{\star,(\ell)})_{\star\in\{QK,V,M\},\ell\in[L]} \in [0,2]^{3L}$ associated with the parameter matrices. Accordingly, our bounds take the following form: for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$\mathrm{GAP}(f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w)) \leq B(W^{(1:L)}, w, \boldsymbol{p}, \delta) \text{ for all } f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}},\ \boldsymbol{p} \in [0,2]^{3L}.$$

Therefore, the Schatten indices may be chosen after observing the trained weights.

3 Spectrum-adaptive post hoc generalization bounds for Transformers

In this section, we state our spectrum-adaptive post hoc generalization bounds for multi-layer Transformers. The model class is defined in Eq. (6). Throughout this paper, we impose the following regularity assumption.

Assumption 3.1. *The activation function $\phi : \mathbb{R}^N \to \mathbb{R}^N$ satisfies $\phi(0) = 0$ and is L_ϕ-Lipschitz with respect to the Euclidean norm; that is, $\|\phi(x) - \phi(y)\|_2 \le L_\phi \|x - y\|_2$ holds for all $x, y \in \mathbb{R}^N$.*

For example, the ReLU activation satisfies Assumption 3.1 with $L_\phi = 1$.

To isolate the leading scaling, the simplified bounds below treat the following quantities as $\Theta(1)$: the loss constants $L_{\mathcal{L}}$ and $B_{\mathcal{L}}$, the readout norm constraint C_2^{out}, the layerwise spectral norm constraints $C_2^{\star,(\ell)}$, the activation Lipschitz constant L_ϕ, and the input-norm constants appearing in our full statement. We also suppress polylogarithmic factors, except for those involving n, T, and the post hoc uniformity penalty.

Theorem 3.1 (Spectrum-adaptive post hoc bounds (Theorem D.11, simplified))**.** *Suppose $n \ge 3$ holds. Suppose further that there exists a universal constant $c_0 > 0$ such that $\|W^{\star,(\ell)}\|_2 \ge \exp[-c_0(L + \log(N))]$ holds for every nonzero $W^{\star,(\ell)}$.*[1] *Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, it holds simultaneously for all $f_{\mathrm{out}}(\cdot; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}$ satisfying the above condition that*

$$\begin{aligned}
&\mathrm{GAP}(f_{\mathrm{out}}) \\
&\lesssim \sqrt{\frac{\log(nT)}{n}} \inf_{\boldsymbol{p} \in [0,2]^{3L}} \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} (\|W^{\star,(\ell)}\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}})^{\frac{1}{p^{\star,(\ell)}+2}} C^{\frac{Lp^{\star,(\ell)}}{p^{\star,(\ell)}+2}} L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}} \\
&\quad + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta) + L\log(L + \log(N))}{n}},
\end{aligned} \tag{7}$$

where the infimum is taken over a vector of Schatten indices $\boldsymbol{p} \in [0,2]^{3L}$, and $C > 0$ is a constant depending on the spectral norm bounds $C_2^{\star,(\ell)}$ and on L_ϕ.

The three terms on the right-hand side of Eq. (7) have distinct origins. The first term is the main complexity term, obtained by covering the Transformer body, namely the query-key, value, and feedforward components across layers. The second term comes from covering the final readout vector w. The last term is the confidence and post hoc uniformity penalty. The $\log(1/\delta)$ part is the usual concentration term, while the additional $L\log(L + \log(N))$ factor is used to make the guarantee uniform over the post hoc choices of the Schatten indices.

The significance of Theorem 3.1 is that the Schatten indices become post hoc, weight-adaptive complexity parameters, rather than modeling assumptions fixed in advance. Since the high-probability event holds uniformly over all admissible Schatten indices, the index $p^{\star,(\ell)}$ can be selected after training, separately for each matrix type and each layer. Smaller values of $p^{\star,(\ell)}$ exploit rank-like spectral structure and reduce the per-matrix summand factors $C^{Lp^{\star,(\ell)}/(p^{\star,(\ell)}+2)}$, $L^{p^{\star,(\ell)}/(p^{\star,(\ell)}+2)}$, and $N^{(p^{\star,(\ell)}+1)/(p^{\star,(\ell)}+2)}$. In contrast, larger values of $p^{\star,(\ell)}$ yield a more Frobenius-like description. Thus, the infimum in Eq. (7) balances spectral complexity, hidden dimension, depth, and layerwise propagation on a matrixwise and layerwise basis. Consequently, even in regimes where existing norm-based bounds may become loose because of depth accumulation or large fixed norm constants, Theorem 3.1 can select a more favorable trade-off among depth, hidden dimension, and spectral complexity.

We note that the remaining logarithmic dependence on the token length T reflects a geometric trade-off. Trauger and Tewari (2024) remove explicit token-length dependence by using mixed $(1,1)$-norm constraints and an ℓ_1-type covering argument. In contrast, our bounds are based on SVD/Schatten quantities and use an ℓ_2-type covering argument over the empirical token rows, which introduces the factor $\log(nT)$.

[1] This condition is imposed only to state the post hoc bounds in the simplified form of Eq. (7): it rules out exponentially small nonzero weights so that the additional terms introduced by the post hoc discretization can be absorbed into the displayed logarithmic penalty. In the full statement, these scale-dependent terms are kept explicit.

4 Proof outline

We outline the proof of Theorem 3.1. Define the $2 \to \infty$ norm of a matrix by $\|W\|_{2\to\infty} = \max_i \|W_{i\cdot}\|_2$. We use empirical covering numbers with respect to this metric. Fix inputs $\{X_i\}_{i\in[n]}$ and a class $\mathcal{F}$ of matrix-valued functions. Then the covering number $\mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]})$ is the minimum cardinality of a finite set $\mathcal{C} \subset \mathcal{F}$ such that, for every $f \in \mathcal{F}$, there exists $\tilde{f} \in \mathcal{C}$ satisfying $\max_{i\in[n]} \|f(X_i) - \tilde{f}(X_i)\|_{2\to\infty} \le \epsilon$. For scalar-valued classes, we use the same notation with the metric $|\cdot|$.

4.1 Parametric interpolation for matrix-valued linear maps

The following theorem is the basic building block of our analysis. It gives covering number bounds for linear matrix-valued function classes under spectral norm and Schatten-quantity constraints, and therefore provides a flexible way to interpolate between rank-based and norm-based complexity measures.

Theorem 4.1 (Matrix-valued parametric interpolation (Theorem C.8, simplified))**.** *Fix an arbitrary $p \in [0, 2]$. Consider a class of matrix-valued functions $\mathcal{F} = \{f : \mathbb{R}^{d\times\ell} \to \mathbb{R}^{d\times m} \mid f(X) = XW,\ W \in \mathbb{R}^{\ell\times m},\ \|W\|_{\mathrm{s},p}^p \le C_\mathrm{s},\ \|W\|_2 \le C_2\}$. Then, for any $\epsilon > 0$, we have*

$$\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \lesssim \left(\frac{[C_\mathrm{s}(\ell+m)]^2(\min\{\ell,m\}m)^p}{\epsilon^{2p}}\right)^{\frac{1}{p+2}} \log(nd),$$

where logarithmic factors other than $\log(nd)$ are suppressed.

Proof idea. The proof is based on a parametric interpolation argument. For any W in the parameter set, write its singular value decomposition as $W = \sum_j \sigma_j u_j v_j^\top$ and fix a threshold $\tau > 0$. We decompose $W = W_1 + W_2$, where W_1 contains the singular components with $\sigma_j > \tau$ and W_2 is the remaining tail. The Schatten constraint implies $\mathrm{rank}(W_1) \le C_\mathrm{s}/\tau^p$, while the tail satisfies $\|W_2\|_{\mathrm{s},2} \le \tau\sqrt{\min\{\ell, m\}}$. Hence the leading part W_1 can be covered as a low-rank, spectral-norm-bounded matrix class, whereas the tail W_2 can be covered as a Frobenius-norm-bounded linear class. These two covers yield two competing entropy terms of the form $(\ell + m)C_\mathrm{s}/\tau^p$ and $\tau^2 \min\{\ell, m\}m/\epsilon^2$ up to logarithmic factors. Optimizing the threshold τ balances these two terms and gives the exponent $1/(p + 2)$ in the theorem. □

This argument follows the same principle as the vector-valued parametric interpolation argument of Ledent et al. (2025). The matrix-valued setting, however, requires additional care because the functions take values in $(\mathbb{R}^{d\times m}, \|\cdot\|_{2\to\infty})$. In particular, the Frobenius-tail cover must be converted into uniform rowwise output control over the sample $\{X_i\}_{i\in[n]}$.

The logarithmic dependence on T in Theorem 3.1 arises from this matrix-valued linear covering step. Theorem 4.1 contains $\log(nd)$, which gives $\log(nT)$ when applied to Transformer layers. This reflects the Schatten covering route, unlike the ℓ_1-based argument of Trauger and Tewari (2024).

4.2 Fixed-index generalization gap bounds

We next explain how Theorem 4.1 is lifted to multi-layer Transformers. First, fix the Schatten indices $\boldsymbol{p} \in [0, 2]^{3L}$ and the corresponding Schatten-quantity radii $\boldsymbol{C}_\mathrm{s} = (C_\mathrm{s}^{\star,(\ell)})_{\star\in\{QK,V,M\},\ell\in[L]} \in (0, \infty)^{3L}$ in advance. Define the constrained parameter class by

$$\mathcal{W}^{(1:L)}(\boldsymbol{p}, \boldsymbol{C}_\mathrm{s}) = \mathcal{W}^{(1)}(\boldsymbol{p}^{(1)}, \boldsymbol{C}_\mathrm{s}^{(1)}) \times \cdots \times \mathcal{W}^{(L)}(\boldsymbol{p}^{(L)}, \boldsymbol{C}_\mathrm{s}^{(L)}),$$
$$\mathcal{W}^{(\ell)}(\boldsymbol{p}^{(\ell)}, \boldsymbol{C}_\mathrm{s}^{(\ell)}) = \mathcal{W}^{QK,(\ell)}(p^{QK,(\ell)}, C_\mathrm{s}^{QK,(\ell)}) \times \mathcal{W}^{V,(\ell)}(p^{V,(\ell)}, C_\mathrm{s}^{V,(\ell)}) \times \mathcal{W}^{M,(\ell)}(p^{M,(\ell)}, C_\mathrm{s}^{M,(\ell)}),$$
$$\mathcal{W}^{\star,(\ell)}(p^{\star,(\ell)}, C_\mathrm{s}^{\star,(\ell)}) = \{W \in \mathcal{W}^{\star,(\ell)} \mid \|W\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}} \le C_\mathrm{s}^{\star,(\ell)}\}.$$

The corresponding scalar output class is

$$\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_\mathrm{s})} = \{f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) : \mathbb{R}^{T\times N} \to \mathbb{R} \mid W^{(1:L)} \in \mathcal{W}^{(1:L)}(\boldsymbol{p}, \boldsymbol{C}_\mathrm{s}), w \in \mathbb{R}^N, \|w\|_2 \le C_2^{\mathrm{out}}\}.$$

Theorem 4.1 is applied to the linear maps arising from the query-key, value, and feedforward matrices in each Transformer block. The resulting covers are composed layer by layer: the Lipschitz properties of the softmax map, the rowwise normalization, and the activation function allow the local approximation errors to be propagated through the Transformer in the $2 \to \infty$ norm, and the final readout layer converts the matrix-valued cover into a scalar output cover. Combining the Transformer covering number bounds with a Dudley-type entropy integral and the standard Lipschitz-loss generalization inequality yields the fixed-index generalization bounds in the following two theorems.

Theorem 4.2 (Fixed-index generalization bounds (Theorem D.8, simplified))**.** *Suppose $n \geq 3$ holds. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds simultaneously for all $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ that*

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \sqrt{\frac{\log(nT)}{n}} \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{1}{p^{\star,(\ell)}+2}} C^{\frac{Lp^{\star,(\ell)}}{p^{\star,(\ell)}+2}} L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}} + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

where $C > 0$ is the same constant that appeared in Theorem 3.1.

Theorem 4.3 (Common-p bounds (Theorem D.9, simplified))**.** *Consider the case $\boldsymbol{p} = (p, \ldots, p)$. Suppose $n \geq 3$ holds. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds simultaneously for all $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ that*

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \sqrt{\frac{\log(nT)}{n}} C^{\frac{Lp}{p+2}} \left(\sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{2}{3p+2}} \right)^{\frac{3p+2}{2(p+2)}} N^{\frac{p+1}{p+2}} + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

where $C > 0$ is the same constant that appeared in Theorem 3.1.

Theorem 4.3 is not merely the result of substituting a common value of p into Theorem 4.2. When all matrices share a common Schatten index, the proof can optimize the covering radii across layers and matrix types when we derive the covering entropy bounds. This optimized allocation leads to the aggregate term in Theorem 4.3 and improves the polynomial dependence on the depth L compared to the naive balanced allocation underlying Theorem 4.2.

4.3 Post hoc selection of the Schatten indices

The fixed-index bounds in Theorem 4.2 assume that the Schatten indices and the corresponding radii are specified in advance. To obtain the spectrum-adaptive bounds stated in Theorem 3.1, we remove this restriction by allowing the Schatten indices to be chosen after observing the trained weights. The key point is that the high-probability event can be made uniform over all admissible choices of the indices.

Technically, we first discretize the interval $[0,2]$ for each matrix type and layer, and combine this discretization with a dyadic peeling argument over the realized Schatten-quantity radii. A weighted union bound then gives simultaneous guarantees over all shells and grid points. Finally, each continuous index is rounded upward to the grid, and the resulting changes in the bounds are controlled using the assumption of Theorem 3.1. This allows us to pass from the finite grid to the full continuum $[0,2]^{3L}$. A detailed argument is deferred to Appendix D.7.

5 Examples

The main result of this paper is Theorem 3.1, which provides post hoc bounds that allow the Schatten indices to be chosen after training, separately for each matrix type and each layer. In this section, however, we use the common-p bounds in Theorem 4.3 in order to compare our bounds with existing bounds under matched constraints. Specializing the common Schatten index p, Theorem 4.3 yields Frobenius-type, rank-type, and spectral-norm-only regimes, which correspond to the three columns

Table 2: Comparison of the leading complexity factor B in the generalization gap bounds $\mathrm{GAP}(f_{\mathrm{out}}) = \tilde{O}(B/\sqrt{n})$ under the matched constraints. The three columns correspond, respectively, to Frobenius norm constraints, rank constraints, and spectral norm constraints only. All other settings remain the same as in Table 1. For the norm-based baselines, the displayed rates follow from the inequalities $\|W\|_{2,1} \le \sqrt{N}\|W\|_F \le \sqrt{N\,\mathrm{rank}(W)}\|W\|_2$ and $\|W\|_{1,1} \le N\|W\|_F \le N\sqrt{\mathrm{rank}(W)}\|W\|_2$ for every $W \in \mathbb{R}^{N\times N}$. We note that, although these inequalities are sharp in the worst case (think of the case, for example, $W = N^{-1}\mathbf{1}_N\mathbf{1}_N^\top$), they may be loose for particular trained weight matrices. Accordingly, the comparisons in the table do not imply pointwise dominance over the original mixed-norm bounds.

	$\|W\|_F \le C_F$	$\mathrm{rank}(W) \le r$	$\|W\|_2 = O(1)$ only
Ours	$\sqrt{C_F}C^{\frac{L}{2}}LN^{\frac{3}{4}}$	$\sqrt{rLN}$	$\sqrt{L}N$
Edelman et al. (2022)	$C_FC^LL^{\frac{3}{2}}\sqrt{N}$	$C^LL^{\frac{3}{2}}\sqrt{rN}$	$C^LL^{\frac{3}{2}}N$
Trauger and Tewari (2024)	$C_FC^LL^{\frac{3}{2}}N$	$C^LL^{\frac{3}{2}}\sqrt{r}N$	$C^LL^{\frac{3}{2}}N^{\frac{3}{2}}$

of Table 2. These examples show how our spectral complexity improves the depth dependence, and can also improve the hidden-dimension dependence in some of the regimes.

Example 5.1 (Frobenius and spectral norm constraints)**.** Set $p^{\star,(\ell)} = 2$ for all $\star \in \{QK, V, M\}$ and $\ell \in [L]$. This corresponds to the parameter class with simultaneous Frobenius and spectral norm constraints, $\mathcal{W}^{\star,(\ell)}(2, (C_F^{\star,(\ell)})^2) = \{W \in \mathbb{R}^{N\times N} \mid \|W\|_F \le C_F^{\star,(\ell)},\ \|W\|_2 \le C_2^{\star,(\ell)}\}$. Here, $C_F^{\star,(\ell)}$ denotes the Frobenius norm radius, so that $C_{\mathrm{s}}^{\star,(\ell)} = (C_F^{\star,(\ell)})^2$. Then, Theorem 4.3 gives

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim C^{\frac{L}{2}}\left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}\sqrt{C_F^{\star,(\ell)}}\right)N^{\frac{3}{4}}\sqrt{\frac{\log(nT)}{n}} + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

As summarized in Table 2, if the Frobenius radii are uniformly bounded by C_F, our bounds improve the dependence on the layerwise propagation factor, depth, and the Frobenius radius relative to existing norm-based bounds. This improvement comes with a larger hidden-dimension factor relative to Edelman et al. (2022), increasing from $N^{\frac{1}{2}}$ to $N^{\frac{3}{4}}$.

Example 5.2 (Rank and spectral norm constraints)**.** Set $p^{\star,(\ell)} = 0$ for all $\star \in \{QK, V, M\}$ and $\ell \in [L]$. This corresponds to imposing rank constraints together with spectral norm constraints on the parameter class, which can be written as $\mathcal{W}^{\star,(\ell)}(0, r^{\star,(\ell)}) = \{W \in \mathbb{R}^{N\times N} \mid \mathrm{rank}(W) \le r^{\star,(\ell)},\ \|W\|_2 \le C_2^{\star,(\ell)}\}$. Here, $r^{\star,(\ell)}$ denotes the rank bound, rather than $C_{\mathrm{s}}^{\star,(\ell)}$. Theorem 4.3 gives

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}} r^{\star,(\ell)}\right)^{\frac{1}{2}}\sqrt{\frac{N\log(nT)}{n}} + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

When the ranks are uniformly bounded by r, Table 2 shows that our bounds remove the layerwise propagation factor C^L and reduce the polynomial depth dependence from $L^{\frac{3}{2}}$ to $\sqrt{L}$, compared to the norm-based baselines.

Example 5.3 (Spectral norm constraints only)**.** Finally, consider the parameter class in Eq. (5), where only the layerwise spectral norm constraints are imposed. Since every $N \times N$ matrix has rank at most N, Example 5.2 can be applied with $r^{\star,(\ell)} = N$. This yields

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \frac{N\sqrt{L\log(nT)}}{\sqrt{n}} + \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

As shown in Table 2, our bounds improve the depth dependence from the existing $C^LL^{\frac{3}{2}}$ scaling to $\sqrt{L}$.

6 Empirical comparison of BERT-adapted proxies

We use the publicly released BERT Miniatures checkpoints of Turc et al. (2019) to compare how the leading complexity factors suggested by our theory scale for trained Transformer weights.[2] These

[2]We do not train any new models; all quantities are computed from the released checkpoints.

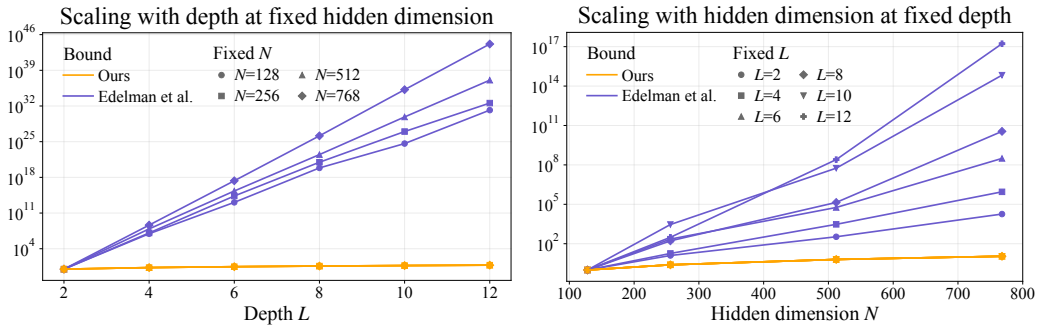

Figure 1: Comparison of the normalized BERT-adapted leading-factor proxies for our spectrum-adaptive bounds and the norm-based bounds of Edelman et al. (2022). Each curve is normalized by its own value at the smallest value of the varying architectural parameter: $L = 2$ at the corresponding fixed N in the left panel, and $N = 128$ at the corresponding fixed L in the right panel. **Left:** scaling with the depth L at fixed hidden dimension N. **Right:** scaling with the hidden dimension N at fixed depth L. In both regimes, our proxies grow more slowly than the proxies based on Edelman et al. (2022).

checkpoints are based on the BERT encoder architecture (Devlin et al., 2019), rather than the simplified single-head Transformer model analyzed in this paper. We therefore do not interpret the resulting quantities as numerical evaluations of the theorems for BERT. Instead, we extract the leading complexity factors from our bounds and from the norm-based bounds of Edelman et al. (2022), adapt these factors to the BERT architecture, and compare the resulting BERT-adapted proxies. These proxies are computed directly from the realized Schatten quantities and mixed $(2, 1)$-norms of the checkpoint weights, rather than from the generic norm inequalities used to derive the matched-constraint comparisons in Table 2. The precise construction of these proxies is given in Appendix F.

In Figure 1, each curve is normalized by its own value at the smallest value of the varying architectural parameter: $L = 2$ at the corresponding fixed N in the left panel, and $N = 128$ at the corresponding fixed L in the right panel. The relevant comparison is therefore the relative growth rate as L or N increases, rather than the absolute vertical scale. The figure shows that our post hoc proxies increase more slowly than the norm-based proxies both as the depth L increases at fixed hidden dimension N and as the hidden dimension N increases at fixed depth L.

Notably, for every BERT Miniatures checkpoint considered, the optimized proxy selects the rank-based endpoint $p = 0$ for every matrix type, layer, and attention head. Thus, while our theory permits a separate post hoc choice of Schatten index for each matrix, these experiments do not demonstrate the empirical selection of intermediate indices $p \in (0, 2)$. The endpoint selection is partly explained by the BERT adaptation: the headwise query-key and value-output matrices have rank at most the head dimension $d_h = 64$. For the feedforward matrices, which exhibit slower singular-value decay, the potential benefit of a positive Schatten index is outweighed by the corresponding depth- and dimension-dependent factors. Consequently, Figure 1 primarily illustrates improved scaling of the leading complexity proxies at the rank-based endpoint, rather than empirical adaptivity across Schatten indices.[3] Further diagnostics are provided in Appendix F.3.

Finally, the empirical comparison should be interpreted with several limitations. Since both bounds involve unspecified multiplicative constants and suppressed logarithmic factors, the normalized proxies do not establish pointwise dominance of the complete generalization bounds. Moreover, the absolute bounds are likely to be numerically vacuous at the scales considered here, particularly because the constants arising from spectral norm and Lipschitz control may be substantial. The released checkpoints also do not provide the corresponding empirical training losses needed to evaluate observed generalization gaps. Therefore, our experiments assess only the relative scaling of leading complexity proxies, rather than the numerical tightness or non-vacuity of the generalization bounds.

[3] Appendix E provides a controlled example with polynomially decaying singular values in which an intermediate Schatten index $p \in (0, 2)$ is optimal.

7 Conclusion

We derived spectrum-adaptive post hoc generalization bounds for a simplified class of multi-layer Transformers under layerwise spectral norm and readout norm constraints. The main feature of the bounds is that the Schatten indices can be selected after training, separately for each layer and each matrix type. This allows the complexity measures to adapt to the learned singular-value profiles of the query-key, value, and feedforward matrices, rather than imposing a single fixed complexity measure across the whole network. The resulting bounds interpolate between rank-based and norm-based regimes, and show improved dependence on depth in representative spectral settings. Our empirical comparison using BERT Miniatures checkpoints shows that the BERT-adapted proxies derived from our bounds grow more slowly with depth and hidden dimension than the corresponding norm-based proxies.

Several limitations remain. First, our analysis relies on layerwise spectral norm and readout norm constraints and considers a simplified Transformer architecture that omits components of modern Transformers, including residual connections, multi-head attention, and LayerNorm/RMSNorm. Extending the framework to these components, particularly normalization layers, requires additional control of approximation errors and their propagation across layers. Second, the absolute generalization bounds may be numerically vacuous in practical settings. Finally, our analysis is agnostic to the training process, and uses the learned weights only through post hoc spectral quantities. A more complete explanation of Transformer generalization may require incorporating information about the optimization trajectory, implicit regularization, and the evolution of representations during training. Furthermore, our bounds do not exploit potentially low-dimensional structure in intermediate activations. Combining such data-dependent structure with the spectral properties of the learned weights could lead to sharper and more explanatory bounds.

Acknowledgements

Mana Sakai was supported by RIKEN Junior Research Associate Program. Masaaki Imaizumi was supported by JSPS KAKENHI (Grant No. 24K02904), JST CREST (Grant No. JPMJCR21D2), JST FOREST (Grant No. JPMJFR216I), JST ASPIRE (Grant No. JPMJAP2329), JST ERATO (Grant No. JPMJER2601) and JST BOOST (Grant No. JPMJBY24A9).

References

Arora, S., R. Ge, B. Neyshabur, and Y. Zhang (2018, 10–15 Jul). Stronger generalization bounds for deep nets via a compression approach. In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263. PMLR.

Bartlett, P. L., D. J. Foster, and M. J. Telgarsky (2017). Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*(Nov), 463–482.

Baykal, C., L. Liebenwein, I. Gilitschenski, D. Feldman, and D. Rus (2019). Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *International Conference on Learning Representations*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel (2023). PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research 24*(240), 1–113.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (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)*, pp. 4171–4186.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Dudley, R. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Dziugaite, G. K. and D. M. Roy (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Edelman, B. L., S. Goel, S. Kakade, and C. Zhang (2022, 17–23 Jul). Inductive biases and variable creation in self-attention mechanisms. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR.

Fu, H., T. Guo, Y. Bai, and S. Mei (2023). What can a single attention layer learn? a study through the random features lens. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Advances in Neural Information Processing Systems*, Volume 36, pp. 11912–11951. Curran Associates, Inc.

Golowich, N., A. Rakhlin, and O. Shamir (2018, 06–09 Jul). Size-independent sample complexity of neural networks. In S. Bubeck, V. Perchet, and P. Rigollet (Eds.), *Proceedings of the 31st Conference On Learning Theory*, Volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR.

Huang, X., A. Yang, S. Bhattamishra, Y. Sarrof, A. Krebs, H. Zhou, P. Nakkiran, and M. Hahn (2025). A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*.

Ledent, A. and R. Alves (2024, 21–27 Jul). Generalization analysis of deep non-linear matrix completion. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 26290–26360. PMLR.

Ledent, A., R. Alves, and Y. Lei (2025). Generalization bounds for rank-sparse neural networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (Eds.), *Advances in Neural Information Processing Systems*, Volume 38, pp. 147927–147996. Curran Associates, Inc.

Li, G., Y. Tang, and W. Zhang (2024, 21–27 Jul). LoRAP: Transformer sub-layers deserve differentiated structured compression for large language models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 28657–28672. PMLR.

Li, Y., T. Hu, Z. Lian, W. Tian, Y. Peng, H. Zhang, and Z. Li (2026). Sharper generalization bounds for transformer. *arXiv preprint arXiv:2603.21541*.

Li, Y., M. E. Ildiz, D. Papailiopoulos, and S. Oymak (2023, 23–29 Jul). Transformers as algorithms: Generalization and stability in in-context learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR.

Mwigo, B. and A. Dasgupta (2026). Generalization bound for a shallow transformer trained using gradient descent. *Transactions on Machine Learning Research*.

Neyshabur, B., S. Bhojanapalli, and N. Srebro (2018). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*.

Neyshabur, B., R. Tomioka, and N. Srebro (2015, 03–06 Jul). Norm-based capacity control in neural networks. In P. Grünwald, E. Hazan, and S. Kale (Eds.), *Proceedings of The 28th Conference on Learning Theory*, Volume 40 of *Proceedings of Machine Learning Research*, Paris, France, pp. 1376–1401. PMLR.

Pinto, A., A. Rangamani, and T. A. Poggio (2025, 24–27 Feb). On generalization bounds for neural networks with low rank layers. In G. Kamath and P.-L. Loh (Eds.), *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, Volume 272 of *Proceedings of Machine Learning Research*, pp. 921–936. PMLR.

Suzuki, T., H. Abe, and T. Nishimura (2020). Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*.

Trauger, J. and A. Tewari (2024, 02–04 May). Sequence length independent norm-based generalization bounds for transformers. In S. Dasgupta, S. Mandt, and Y. Li (Eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR.

Truong, L. V. (2024). On rank-dependent generalisation error bounds for transformers. *arXiv preprint arXiv:2410.11500*.

Turc, I., M.-W. Chang, K. Lee, and K. Toutanova (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wei, C., Y. Chen, and T. Ma (2022). Statistically meaningful approximation: a case study on approximating Turing machines with transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 12071–12083. Curran Associates, Inc.

Yuan, Z., Y. Shang, Y. Song, D. Yang, Q. Wu, Y. Yan, and G. Sun (2023). ASVD: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research 2*(Mar), 527–550.

Zhang, Y., B. Liu, Q. Cai, L. Wang, and Z. Wang (2022). An analysis of attention via the lens of exchangeability and latent variable models. *arXiv preprint arXiv:2212.14852*.

Zhou, W., V. Veitch, M. Austern, R. P. Adams, and P. Orbanz (2019). Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*.

A Additional related work

A.1 Generalization bounds for Transformers

The closest line of work studies complexity-based generalization bounds for Transformers. Edelman et al. (2022) derive covering number and Rademacher complexity bounds for multi-layer self-attention networks under norm constraints. Their bounds show that bounded-norm self-attention can have only logarithmic dependence on the token length and the hidden dimension, but the resulting complexity is controlled by fixed mixed $(2,1)$-norm radii and by layerwise propagation factors. Trauger and Tewari (2024) further develop this covering number approach and obtain norm-based bounds that are independent of the input token length under mixed $(1,1)$-norm constraints. These results provide the main norm-based baselines for our analysis. In contrast, our bounds use spectral complexity measures of the learned weight matrices and allow the Schatten indices to be chosen after training, separately for each layer and matrix type.

Several recent works refine or complement this complexity-based view. Zhang et al. (2022) analyze attention through exchangeability and latent-variable models and obtain token-length-independent generalization guarantees under exchangeability assumptions on the input tokens. This differs from our setting, where no exchangeability assumption is imposed and the token length enters only through logarithmic factors. A closely related work is Truong (2024), which develops rank-dependent covering number bounds for linear function classes and applies them to single-layer Transformers. Their framework includes both prescribed column-subspace constraints and rank constraints, with the full-rank setting as a special case. Our focus is on post hoc selection of a separate Schatten index for each layer and matrix type. Li et al. (2026) use offset Rademacher complexity to derive fast excess risk bounds for Transformers from suitable covering number bounds. Our covering number bounds are complementary to this direction: in principle, they can be combined with the offset Rademacher complexity as in Li et al. (2026), whereas the present paper focuses on post hoc generalization gap bounds.

Other theoretical studies focus on more specialized Transformer settings. Fu et al. (2023) analyze a random-feature attention model with randomly sampled and frozen query-key parameters and trainable value parameters, and obtain excess risk bounds. Li et al. (2023) study in-context learning by viewing a Transformer as implementing an algorithm at inference time, and relate generalization to stability in multitask learning. Wei et al. (2022) study statistically meaningful approximation of Turing machines by Transformers and derive sample-complexity guarantees in that framework. Huang et al. (2025) develop a formal theory of length generalization for causal Transformers with learnable absolute positional encodings. Mwigo and Dasgupta (2026) analyze a shallow Transformer trained by gradient descent under a bounded-drift regime in which the parameters remain close to initialization. These works identify important mechanisms in particular tasks, architectures, or training regimes, whereas our bounds are training-agnostic, apply to multi-layer Transformers, and are expressed directly in terms of the spectra of the trained weights.

A.2 Norm-based and spectral generalization bounds for neural networks

Early norm-based generalization bounds for DNNs include Neyshabur et al. (2015), Bartlett et al. (2017), and Golowich et al. (2018). This line of work explains how large networks can admit dimension-favorable bounds when their learned weights have controlled norms. The Transformer bounds of Edelman et al. (2022) and Trauger and Tewari (2024) can be viewed as architecture-specific continuations of this program. Our contribution follows the same complexity-based philosophy, but it replaces fixed coordinatewise norm constraints by spectral quantities that can be evaluated and optimized post hoc.

A related approach derives generalization guarantees through compressibility. Compression-based bounds show that networks admitting short descriptions or accurate compressed approximations can generalize well (Arora et al., 2018; Baykal et al., 2019; Suzuki et al., 2020). In particular, Suzuki et al. (2020) convert compression guarantees into bounds for the original non-compressed network, making compressibility itself a statistical complexity measure. This viewpoint is close in spirit to ours, because fast singular-value decay can be interpreted as a form of spectral compressibility. However, our bounds do not require constructing an explicit compressed network; instead, the complexity is expressed directly through Schatten quantities of the trained weights.

PAC-Bayesian methods provide a complementary approach to generalization analysis. Dziugaite and Roy (2017) obtain non-vacuous certificates by optimizing a posterior over stochastic neural-network predictors, while Zhou et al. (2019) combine PAC-Bayesian analysis with compression to certify compressed networks.[4] These approaches differ from our direct generalization-gap bound for a given deterministic Transformer. Nevertheless, Neyshabur et al. (2018) show that PAC-Bayesian guarantees can also be transferred to the original deterministic predictor by controlling changes in network outputs under random weight perturbations. Exploring whether such methods can incorporate the learned spectral structure of Transformer weights while retaining post hoc adaptivity would be an interesting direction for future work.

Low-rank and spectral structures have also been studied more directly in neural-network generalization. Pinto et al. (2025) derive Gaussian-complexity bounds for networks with low-rank layers and show how low-rank constraints can mitigate the accumulation of dimension-dependent factors across depth. Ledent et al. (2025) obtain post hoc generalization bounds using Schatten (quasi) norms. Their analysis is technically close to ours and provides the main inspiration for the spectrum-adaptive aspect of our bounds. We extend this idea to Transformer architectures, where the proof must control matrix-valued function classes under the $\|\cdot\|_{2\to\infty}$ metric.

B Mathematical preliminaries

B.1 Matrix norms

For a matrix W, we use the notation summarized in Table 3. Here, $\sigma_i(W)$ denotes the i-th largest singular value of W, and $W_{i\cdot}$ denotes the i-th row of W.

Table 3: Matrix norm notation used throughout the paper.

Notation	Definition	Description
$\|W\|_F$	$(\sum_{i,j} W_{ij}^2)^{\frac{1}{2}}$	Frobenius norm
$\|W\|_{\alpha\to\beta}$	$\sup\{\|Wx\|_\beta : \|x\|_\alpha \le 1\}$	Induced norm from ℓ_α to ℓ_β
$\|W\|_\alpha$	$\|W\|_{\alpha\to\alpha}$	Induced norm on ℓ_α
$\|W\|_2$	$\sigma_1(W)$	Spectral norm
$\|W\|_{2\to\infty}$	$\max_i \|W_{i\cdot}\|_2$	Maximum row Euclidean norm
$\|W\|_{\alpha,\beta}$	$[\sum_j (\sum_i \|W_{ij}\|^\alpha)^{\frac{\beta}{\alpha}}]^{\frac{1}{\beta}}$	Mixed (α,β)-norm

For $\alpha, \beta \ge 1$, the induced norm from ℓ_α to ℓ_β is denoted by $\|W\|_{\alpha\to\beta}$. When $\alpha = \beta$, we simply write $\|W\|_\alpha := \|W\|_{\alpha\to\alpha}$. In particular, $\|W\|_2 = \sigma_1(W)$ is the spectral norm. The identity $\|W\|_{2\to\infty} = \max_i \|W_{i\cdot}\|_2$ is used throughout the paper.

We also use the following Schatten p (quasi) norm.

Definition B.1. For $p \in (0, 2]$, the Schatten p (quasi) norm $\|W\|_{\mathrm{s},p}$ of a matrix W is defined by $\|W\|_{\mathrm{s},p} = (\sum_i \sigma_i(W)^p)^{\frac{1}{p}}$. We refer to p as the Schatten index. For $p = 0$, we use the convention $\|W\|_{\mathrm{s},0}^0 := \lim_{p\downarrow 0} \|W\|_{\mathrm{s},p}^p = \mathrm{rank}(W)$.

B.2 Some basic results

The following results are some well-known matrix norm inequalities.

Lemma B.1. $\|AB\|_{2\to\infty} \le \|A\|_\infty \|B\|_{2\to\infty}$.

Lemma B.2. $\|AB\|_{2\to\infty} \le \|A\|_{2\to\infty} \|B\|_2$.

Lemma B.3. *(i)* $\|AB\|_F \le \|A\|_F \|B\|_2$.

(ii) $\|AB\|_F \le \|A\|_2 \|B\|_F$.

Lemma B.4. *For any $a \in \mathbb{R}^\ell$ and $B \in \mathbb{R}^{\ell\times m}$, we have*

$$\|a^\top B\|_2 \le \|a\|_1 \|B\|_{2\to\infty}.$$

[4] Numerical comparisons with the certificates of Dziugaite and Roy (2017) and Zhou et al. (2019) would require additional posterior optimization or compression for each model, whereas our experiments evaluate complexity proxies directly from released checkpoints. We therefore leave such comparisons for future work.

Proof. For $a = 0$, the claim is immediate. For $a \neq 0$, we have

$$\|a^\top B\|_2 = \|B^\top a\|_2 = \left\|B^\top \frac{a}{\|a\|_1}\right\|_2 \|a\|_1 \leq \|B^\top\|_{1\to 2}\|a\|_1 = \|B\|_{2\to\infty}\|a\|_1.$$

Here, the last equality follows from the duality relation $\|A\|_{2\to\infty} = \|A^\top\|_{1\to 2}$. □

We also state some useful properties of the softmax function and the rowwise projection operator.

Lemma B.5 (Edelman et al. (2022), Lemma A.9)**.** *Suppose Π_{norm} is the rowwise projection operator onto the unit ball. Then, for any $Z, Z' \in \mathbb{R}^{T\times N}$, we have*

$$\|\Pi_{\mathrm{norm}}(Z) - \Pi_{\mathrm{norm}}(Z')\|_{2\to\infty} \leq \|Z - Z'\|_{2\to\infty}.$$

Lemma B.6. *For any $G \in \mathbb{R}^{T\times T}$, the softmax function satisfies*

$$\|\mathrm{SoftMax}(G)\|_\infty = \max_{j\in[T]} \sum_{k=1}^{T} \mathrm{SoftMax}_k(G_{j\cdot}) = 1.$$

Lemma B.7 (Edelman et al. (2022), Corollary A.7)**.** *For any $x, y \in \mathbb{R}^T$, it holds that*

$$\|\mathrm{SoftMax}(x) - \mathrm{SoftMax}(y)\|_1 \leq 2\|x - y\|_\infty.$$

The following lemma is used when we optimize the allocation of covering radii in the covering entropy bounds of Transformers.

Lemma B.8. *Fix $a_i, b_i, c, \nu > 0$. The unique solution to the optimization problem*

$$\min_{z_1,\ldots,z_m>0} \sum_{i=1}^{m} a_i z_i^{-\nu} \qquad \text{subject to} \quad \sum_{i=1}^{m} b_i z_i = c$$

is given by

$$z_i^* = \frac{c a_i^{\frac{1}{\nu+1}} b_i^{-\frac{1}{\nu+1}}}{\sum_{j=1}^{m} a_j^{\frac{1}{\nu+1}} b_j^{\frac{\nu}{\nu+1}}} \qquad (i \in [m]).$$

Moreover, the minimum value is

$$\sum_{i=1}^{m} a_i (z_i^*)^{-\nu} = \frac{1}{c^\nu} \left(\sum_{i=1}^{m} a_i^{\frac{1}{\nu+1}} b_i^{\frac{\nu}{\nu+1}}\right)^{\nu+1}.$$

Proof. Since

$$\frac{\partial^2}{\partial z_i^2}(a_i z_i^{-\nu}) = \nu(\nu+1) a_i z_i^{-\nu-2} > 0$$

holds for each $i \in [m]$, the objective function is strictly convex on $(0,\infty)^m$. Hence, over the affine constraint set, any feasible point satisfying the first-order condition is the unique global minimizer. Consider the Lagrangian

$$\mathcal{L}(z_1, \ldots, z_m, \lambda) = \sum_{i=1}^{m} a_i z_i^{-\nu} + \lambda \left(\sum_{i=1}^{m} b_i z_i - c\right).$$

The first-order condition with respect to z_i is

$$\frac{\partial \mathcal{L}}{\partial z_i} = -\nu a_i z_i^{-\nu-1} + \lambda b_i = 0,$$

which implies

$$z_i = \left(\frac{\nu a_i}{\lambda b_i}\right)^{\frac{1}{\nu+1}} \qquad (i \in [m]).$$

Substituting this into the constraint yields

$$\sum_{i=1}^{m} b_i \left(\frac{\nu a_i}{\lambda b_i}\right)^{\frac{1}{\nu+1}} = \left(\frac{\nu}{\lambda}\right)^{\frac{1}{\nu+1}} \sum_{i=1}^{m} a_i^{\frac{1}{\nu+1}} b_i^{\frac{\nu}{\nu+1}} = c.$$

Hence, we require

$$\left(\frac{\nu}{\lambda}\right)^{\frac{1}{\nu+1}} = \frac{c}{\sum_{j=1}^m a_j^{\frac{1}{\nu+1}} b_j^{\frac{\nu}{\nu+1}}}.$$

Combining the above results, we obtain

$$z_i^* = \frac{c a_i^{\frac{1}{\nu+1}} b_i^{-\frac{1}{\nu+1}}}{\sum_{j=1}^m a_j^{\frac{1}{\nu+1}} b_j^{\frac{\nu}{\nu+1}}} \qquad (i \in [m]).$$

Finally, the corresponding minimum value is given by

$$\sum_{i=1}^m a_i (z_i^*)^{-\nu} = \left(\frac{\sum_{j=1}^m a_j^{\frac{1}{\nu+1}} b_j^{\frac{\nu}{\nu+1}}}{c}\right)^{\nu} \sum_{i=1}^m a_i^{\frac{1}{\nu+1}} b_i^{\frac{\nu}{\nu+1}} = \frac{1}{c^\nu}\left(\sum_{i=1}^m a_i^{\frac{1}{\nu+1}} b_i^{\frac{\nu}{\nu+1}}\right)^{\nu+1}.$$

□

B.2.1 Covering numbers and generalization gap

In this paper, we define covering numbers as follows.

Definition B.2 (Covering number of a general metric space)**.** Suppose $(\mathcal{A}, \rho)$ is a metric space. $\mathcal{C} \subset \mathcal{A}$ is called an ϵ-cover of $\mathcal{A}$ if for every $a \in \mathcal{A}$, there exists $a' \in \mathcal{C}$ such that $\rho(a, a') \le \epsilon$. The ϵ-covering number $\mathcal{N}(\mathcal{A}, \rho, \epsilon)$ is the minimum cardinality of any ϵ-cover $\mathcal{C}$ of $\mathcal{A}$.

Definition B.3 (Covering number of a function space)**.** Suppose $\mathcal{F}$ is a class of maps from $\mathcal{X}$ to $\mathcal{Y}$, where $\mathcal{Y}$ is equipped with a metric ρ. For inputs $\{x_i\}_{i\in[n]} \subset \mathcal{X}$, the covering number $\mathcal{N}_\infty(\mathcal{F}, \rho, \epsilon; \{x_i\}_{i\in[n]})$ is the minimum cardinality of any ϵ-cover $\mathcal{C} \subset \mathcal{F}$ such that for all $f \in \mathcal{F}$, there exists $f' \in \mathcal{C}$ satisfying

$$\max_{i\in[n]} \rho(f(x_i), f'(x_i)) \le \epsilon.$$

To convert the entropy bounds into generalization bounds, we recall a standard route through empirical Rademacher complexity.

Definition B.4 (Empirical Rademacher complexity)**.** Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. For inputs $\{x_i\}_{i\in[n]} \subset \mathcal{X}$, define the empirical Rademacher complexity by

$$\hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \,\middle|\, \{x_i\}_{i\in[n]}\right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a vector of independent Rademacher random variables.

Lemma B.9 (Dudley-type entropy integral; cf. Edelman et al. (2022), Lemma A.2; Dudley (1967))**.** *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Suppose $|f(x)| \le A$ holds for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$. Then, we have*

$$\hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) \lesssim \inf_{\alpha>0}\left(\alpha + \int_\alpha^A \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, |\cdot|, \epsilon; \{x_i\}_{i\in[n]})}{n}} d\epsilon\right).$$

Lemma B.10. *Fix an integer $J \ge 1$, constants $C_i > 0$ $(i \in [J])$, $\nu_i \in [0, 2)$ $(i \in [J-1])$, and $\nu_J = 2$. Suppose $\log \mathcal{N}_\infty(\mathcal{F}, |\cdot|, \epsilon; \{x_i\}_{i\in[n]}) \lesssim \sum_{i=1}^J C_i \epsilon^{-\nu_i}$ holds for all $\epsilon \in (0, A]$. Then, we have*

$$\hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) \lesssim \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{J-1} \frac{A^{1-\nu_i/2}}{1-\nu_i/2}\sqrt{C_i} + \left[1 + \log\left(1 + \frac{A\sqrt{n}}{\sqrt{C_J}}\right)\right]\sqrt{C_J}\right).$$

Proof. Using $(\sum_{i=1}^J a_i)^{1/2} \le \sum_{i=1}^J \sqrt{a_i}$ for $a_i \ge 0$, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) &\lesssim \inf_{0<\alpha\le A}\left(\alpha + \int_\alpha^A \sqrt{\frac{\sum_{i=1}^J C_i \epsilon^{-\nu_i}}{n}} d\epsilon\right) \\
&\le \inf_{0<\alpha\le A}\left(\alpha + \sum_{i=1}^J \sqrt{\frac{C_i}{n}} \int_\alpha^A \epsilon^{-\nu_i/2} d\epsilon\right).
\end{aligned}$$

Note that each integral can be bounded as

$$\int_\alpha^A \epsilon^{-\nu_i/2} d\epsilon = \frac{A^{1-\nu_i/2} - \alpha^{1-\nu_i/2}}{1-\nu_i/2} \le \frac{A^{1-\nu_i/2}}{1-\nu_i/2} \qquad (i \in [J-1]),$$
$$\int_\alpha^A \epsilon^{-\nu_J/2} d\epsilon = \int_\alpha^A \epsilon^{-1} d\epsilon = \log(A/\alpha).$$

Hence, we have

$$\hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) \lesssim \inf_{0<\alpha\le A} \left[\alpha + \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{J-1} \frac{A^{1-\nu_i/2}}{1-\nu_i/2} \sqrt{C_i} + \sqrt{C_J} \log(A/\alpha) \right) \right].$$

Taking $\alpha = \min\{A, \sqrt{C_J/n}\}$ proves the claim. □

Lemma B.11 (Bartlett and Mendelson (2002)). *Let $\mathcal{D}$ be a probability distribution on $\mathcal{X} \times \mathbb{R}$, and let $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a $B_\mathcal{L}$-bounded loss function that is $L_\mathcal{L}$-Lipschitz in its first argument. For each $f \in \mathcal{F}$, define the population risk and empirical risk by*

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathcal{L}(f(X), Y)] \qquad \textit{and} \qquad \hat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(f(x_i), Y_i),$$

respectively, where $\{(x_i, Y_i)\}_{i\in[n]}$ are i.i.d. samples from $\mathcal{D}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the bound

$$|\mathcal{R}(f) - \hat{\mathcal{R}}_n(f)| \lesssim L_\mathcal{L} \hat{\mathfrak{R}}_n(\mathcal{F}; \{x_i\}_{i\in[n]}) + B_\mathcal{L} \sqrt{\frac{\log(1/\delta)}{n}},$$

holds simultaneously for all $f \in \mathcal{F}$.

C Metric entropy bounds

C.1 Basic covering number bounds for matrices

We first collect elementary covering number bounds for matrix classes that will be used as ingredients in the later interpolation argument.

Fact C.1 (Vershynin (2018), Corollary 4.2.13). *Let $\mathcal{B} = \left\{a \in \mathbb{R}^\ell \mid \|a\|_2 \le 1\right\}$ be the Euclidean unit ball. Then, we have*

$$\ell \log \frac{1}{\epsilon} \le \log \mathcal{N}(\mathcal{B}, \|\cdot\|_2, \epsilon) \le \ell \log\left(\frac{2}{\epsilon} + 1\right).$$

By scaling, the same result implies that, for $\mathcal{A} = \left\{a \in \mathbb{R}^\ell \mid \|a\|_2 \le C\right\}$,

$$\ell \log \frac{C}{\epsilon} \le \log \mathcal{N}(\mathcal{A}, \|\cdot\|_2, \epsilon) \le \ell \log\left(\frac{2C}{\epsilon} + 1\right).$$

Corollary C.1. *Define*

$$\mathcal{W} = \left\{W \in \mathbb{R}^{\ell\times m} \mid \|W\|_F \le C\right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Then, we have

$$\ell m \log \frac{C}{\epsilon} \le \log \mathcal{N}(\mathcal{W}, \|\cdot\|_F, \epsilon) \le \ell m \log\left(\frac{2C}{\epsilon} + 1\right).$$

Proposition C.2 (Ledent et al. (2025), Proposition F.1 (modified)). *Define*

$$\mathcal{W} = \left\{W \in \mathbb{R}^{\ell\times m} \mid \operatorname{rank}(W) \le r,\ \|W\|_2 \le C\right\},$$

where $\|\cdot\|_2$ denotes the operator norm and $r \le \min\{\ell, m\}$. Then, we have

$$\log \mathcal{N}(\mathcal{W}, \|\cdot\|_F, \epsilon) \le (\ell + m) r \log\left(\frac{8C\sqrt{r}}{\epsilon} + 1\right).$$

Proof. For $W \in \mathcal{W}$, we can write $W = W_1 W_2^\top$ with $W_1 \in \mathbb{R}^{\ell \times r}$ and $W_2 \in \mathbb{R}^{m \times r}$ satisfying

$$\|W_1\|_2 \le \sqrt{C}, \quad \|W_2\|_2 \le \sqrt{C}.$$

It follows that for any $W, W' \in \mathcal{W}$,

$$\|W - W'\|_F = \|W_1 W_2^\top - W_1' W_2'^\top\|_F \le \|W_1 (W_2 - W_2')^\top\|_F + \|(W_1 - W_1') W_2'^\top\|_F$$
$$\le \|W_1\|_2 \|W_2 - W_2'\|_F + \|W_1 - W_1'\|_F \|W_2'\|_2 \le \sqrt{C} \left(\|W_1 - W_1'\|_F + \|W_2 - W_2'\|_F\right),$$

where the first inequality follows from the triangle inequality and the second inequality follows from Lemma B.3. Set

$$\mathcal{W}_1 = \{W \in \mathbb{R}^{\ell \times r} \mid \|W\|_2 \le \sqrt{C}\}, \quad \mathcal{W}_2 = \{W \in \mathbb{R}^{m \times r} \mid \|W\|_2 \le \sqrt{C}\}.$$

Then, by Lemma C.3, we have

$$\mathcal{N}\left(\mathcal{W}_1, \|\cdot\|_F, \frac{\epsilon}{2\sqrt{C}}\right) \le \left(\frac{8C\sqrt{r}}{\epsilon} + 1\right)^{\ell r}, \quad \mathcal{N}\left(\mathcal{W}_2, \|\cdot\|_F, \frac{\epsilon}{2\sqrt{C}}\right) \le \left(\frac{8C\sqrt{r}}{\epsilon} + 1\right)^{mr}.$$

Thus, we have

$$\mathcal{N}\left(\mathcal{W}, \|\cdot\|_F, \epsilon\right) \le \mathcal{N}\left(\mathcal{W}_1, \|\cdot\|_F, \frac{\epsilon}{2\sqrt{C}}\right) \mathcal{N}\left(\mathcal{W}_2, \|\cdot\|_F, \frac{\epsilon}{2\sqrt{C}}\right) \le \left(\frac{8C\sqrt{r}}{\epsilon} + 1\right)^{(\ell+m)r}.$$

□

Lemma C.3. *Define*

$$\mathcal{W} = \{W \in \mathbb{R}^{\ell \times m} \mid \|W\|_2 \le C, \ \mathrm{rank}(W) \le r\},$$

where $\|\cdot\|_2$ denotes the operator norm. Then, it holds that

$$\log \mathcal{N}\left(\mathcal{W}, \|\cdot\|_F, \epsilon\right) \le \ell m \log\left(\frac{4C\sqrt{r}}{\epsilon} + 1\right).$$

Proof. For any $W \in \mathcal{W}$, we have

$$\|W\|_F = \left(\sum_{i=1}^{\mathrm{rank}(W)} \sigma_i^2(W)\right)^{1/2} \le C\sqrt{r}.$$

Hence, $\mathcal{W}$ is contained in the Frobenius ball $\mathcal{B}_F = \{W \in \mathbb{R}^{\ell \times m} \mid \|W\|_F \le C\sqrt{r}\}$. By Corollary C.1, there exists an $\epsilon/2$-cover $\bar{\mathcal{C}}$ of $\mathcal{B}_F$ with respect to the Frobenius norm such that

$$|\bar{\mathcal{C}}| \le \left(\frac{4C\sqrt{r}}{\epsilon} + 1\right)^{\ell m}.$$

Since $\mathcal{W} \subset \mathcal{B}_F$, the same set $\bar{\mathcal{C}}$ is an external $\epsilon/2$-cover of $\mathcal{W}$ with respect to the Frobenius norm. Applying Lemma C.4 to this external cover yields a proper ϵ-cover $\mathcal{C} \subset \mathcal{W}$ satisfying $|\mathcal{C}| \le |\bar{\mathcal{C}}|$. Therefore, we have

$$\mathcal{N}\left(\mathcal{W}, \|\cdot\|_F, \epsilon\right) \le \left(\frac{4C\sqrt{r}}{\epsilon} + 1\right)^{\ell m}.$$

□

Lemma C.4 (Properization of external empirical covers)**.** *Let $\mathcal{F}$ be a function class. Suppose that $\bar{\mathcal{C}}$ is a finite set of functions, not necessarily contained in $\mathcal{F}$, such that for every $f \in \mathcal{F}$ there exists $\bar{f} \in \bar{\mathcal{C}}$ satisfying $\max_{i \in [n]} \rho(f(x_i), \bar{f}(x_i)) \le \epsilon$. Then there exists a proper 2ϵ-cover $\mathcal{C} \subset \mathcal{F}$ of $\mathcal{F}$ with $|\mathcal{C}| \le |\bar{\mathcal{C}}|$.*

Proof. Discard every $\bar{f} \in \bar{\mathcal{C}}$ for which $\{f \in \mathcal{F} \mid \max_{i \in [n]} \rho(f(x_i), \bar{f}(x_i)) \le \epsilon\}$ is empty. For each remaining $\bar{f}$, choose one representative $f_{\bar{f}} \in \mathcal{F}$ satisfying $\max_{i \in [n]} \rho(f_{\bar{f}}(x_i), \bar{f}(x_i)) \le \epsilon$, and set $\mathcal{C} = \{f_{\bar{f}} \mid \bar{f} \in \bar{\mathcal{C}}\}$. Then, for any $f \in \mathcal{F}$, choosing $\bar{f} \in \bar{\mathcal{C}}$ with $\max_{i \in [n]} \rho(f(x_i), \bar{f}(x_i)) \le \epsilon$ gives

$$\max_{i \in [n]} \rho(f(x_i), f_{\bar{f}}(x_i)) \le \max_{i \in [n]} \rho(f(x_i), \bar{f}(x_i)) + \max_{i \in [n]} \rho(\bar{f}(x_i), f_{\bar{f}}(x_i)) \le 2\epsilon.$$

Thus $\mathcal{C}$ is a proper 2ϵ-cover of $\mathcal{F}$. □

Remark C.1. Since $\|W\|_{2\to\infty} \le \|W\|_2 \le \|W\|_F$ holds for any matrix W, we have

$$\mathcal{N}\left(\mathcal{W}, \|\cdot\|_{2\to\infty}, \epsilon\right) \le \mathcal{N}\left(\mathcal{W}, \|\cdot\|_2, \epsilon\right) \le \mathcal{N}\left(\mathcal{W}, \|\cdot\|_F, \epsilon\right).$$

Lemma C.5 (Vershynin (2018), Lemma 4.4.1)**.** *Let $\mathcal{N}_m \subset \mathcal{S}^{m-1}$ be an ϵ-net of the Euclidean unit sphere. Then, for any $W \in \mathbb{R}^{\ell\times m}$ and $\epsilon \in [0,1)$, we have*

$$\|W\|_2 \le \frac{1}{1-\epsilon} \sup_{x\in\mathcal{N}_m} \|Wx\|_2.$$

C.2 Entropy bounds for linear functions

We next derive entropy bounds for linear function classes evaluated on a fixed sample. These results convert norm constraints on the weight matrix into uniform covering bounds for the induced matrix-valued maps under the $\|\cdot\|_{2\to\infty}$ metric.

Proposition C.6 (Zhang (2002), Theorem 4)**.** *Suppose $p \in [2,\infty)$ and $q \in [1,2]$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Define a class of linear functions by*

$$\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = x^\top w,\ w \in \mathbb{R}^d,\ \|w\|_q \le a,\ \|x\|_p \le b\}.$$

Then, for any $\epsilon > 0$, we have

$$\log \mathcal{N}_\infty(\mathcal{F}, |\cdot|, \epsilon; \{x_i\}_{i\in[n]}) \le 36(p-1)\frac{a^2b^2}{\epsilon^2} \log\left(2\left\lceil \frac{4ab}{\epsilon}+2\right\rceil n+1\right).$$

Corollary C.7. *Define a class of matrix-valued functions by*

$$\mathcal{F} = \{f : \mathbb{R}^{d\times\ell} \to \mathbb{R}^{d\times m} \mid f(X) = XW,\ W \in \mathcal{W}\},$$

where

$$\mathcal{W} = \{W \in \mathbb{R}^{\ell\times m} \mid \|W\|_F \le a\}.$$

Suppose $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ holds. Then, for any $\epsilon > 0$ and $\delta \in (0,1)$, we have

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \\
&\le \frac{36a^2B_{n,(2\to\infty)}^2}{(1-\delta)^2\epsilon^2} \log\left(2\left\lceil \frac{4aB_{n,(2\to\infty)}}{(1-\delta)\epsilon}+2\right\rceil \left(1+\frac{2}{\delta}\right)^m nd+1\right).
\end{aligned}$$

In particular, setting $\delta = 1/2$ gives

$$\begin{aligned}
\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) &\le \frac{144a^2B_{n,(2\to\infty)}^2}{\epsilon^2} \log\left(4\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon}+2\right\rceil 5^m nd\right) \\
&\le \frac{144a^2B_{n,(2\to\infty)}^2 m}{\epsilon^2} \log\left(20\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon}+2\right\rceil nd\right).
\end{aligned}$$

Proof. Fix $\delta \in (0,1)$. Let $X_{i,j\cdot}$ denote the j-th row of X_i. By Fact C.1, one may choose a δ-net $\mathcal{N}_m \subset \mathcal{S}^{m-1}$ of the Euclidean sphere satisfying $|\mathcal{N}_m| \le (1+2/\delta)^m$. For such $\mathcal{N}_m$, we can apply Lemma C.5 to obtain

$$\begin{aligned}
&\|X_i(W-W')\|_{2\to\infty} = \max_{j\in[d]} \|X_{i,j\cdot}(W-W')\|_2 \\
&\le \frac{1}{1-\delta} \max_{j\in[d]} \max_{u\in\mathcal{N}_m} |X_{i,j\cdot}(W-W')u| = \frac{1}{1-\delta} \max_{j\in[d]} \max_{u\in\mathcal{N}_m} \left|\langle W-W', (X_{i,j\cdot})^\top u^\top\rangle_F\right| \\
&= \frac{1}{1-\delta} \max_{j\in[d]} \max_{u\in\mathcal{N}_m} \left|\mathrm{vec}(W-W')^\top \mathrm{vec}((X_{i,j\cdot})^\top u^\top)\right|
\end{aligned}$$

for any $W, W' \in \mathcal{W}$ and $i \in [n]$. Define the finite set of vectors

$$\mathcal{Y} = \{\mathrm{vec}((X_{i,j\cdot})^\top u^\top) \mid i \in [n], j \in [d], u \in \mathcal{N}_m\}.$$

Then, any vector in $\mathcal{Y}$ is bounded in norm as

$$\|\mathrm{vec}((X_{i,j\cdot})^\top u^\top)\|_2 = \|(X_{i,j\cdot})^\top u^\top\|_F = \|X_{i,j\cdot}\|_2\|u\|_2 = \|X_{i,j\cdot}\|_2 \le \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}.$$

Thus, by applying Proposition C.6 with $p = q = 2$ and the linear function class

$$\tilde{\mathcal{F}} = \{f : \mathbb{R}^{\ell m} \to \mathbb{R} \mid f(y) = w^\top y,\ w \in \mathbb{R}^{\ell m},\ \|w\|_2 \le a\},$$

we obtain

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \\
&\le \log \mathcal{N}_\infty(\tilde{\mathcal{F}}, |\cdot|, (1-\delta)\epsilon; \mathcal{Y}) \\
&\le 36\frac{a^2 B_{n,(2\to\infty)}^2}{(1-\delta)^2\epsilon^2} \log\left(2\left\lceil \frac{4aB_{n,(2\to\infty)}}{(1-\delta)\epsilon} + 2\right\rceil nd|\mathcal{N}_m| + 1\right) \\
&\le 36\frac{a^2 B_{n,(2\to\infty)}^2}{(1-\delta)^2\epsilon^2} \log\left(2\left\lceil \frac{4aB_{n,(2\to\infty)}}{(1-\delta)\epsilon} + 2\right\rceil nd\left(1+\frac{2}{\delta}\right)^m + 1\right).
\end{aligned}$$

Furthermore, setting $\delta = 1/2$ gives

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \\
&\le \frac{144a^2 B_{n,(2\to\infty)}^2}{\epsilon^2} \log\left(2\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon} + 2\right\rceil 5^m nd + 1\right) \\
&\le \frac{144a^2 B_{n,(2\to\infty)}^2}{\epsilon^2} \log\left(4\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon} + 2\right\rceil 5^m nd\right) \\
&= \frac{144a^2 B_{n,(2\to\infty)}^2}{\epsilon^2} \left(m\log 5 + \log\left(4\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon} + 2\right\rceil nd\right)\right) \\
&\le \frac{144a^2 B_{n,(2\to\infty)}^2 m}{\epsilon^2} \log\left(20\left\lceil \frac{8aB_{n,(2\to\infty)}}{\epsilon} + 2\right\rceil nd\right).
\end{aligned}$$

□

C.3 Parametric interpolation

We now prove a matrix-valued version of the parametric interpolation argument of Ledent et al. (2025). The main additional point is that the output metric is $\|\cdot\|_{2\to\infty}$ evaluated on the sample $\{X_i\}_{i\in[n]}$; hence, the Frobenius-controlled tail must be converted into uniform rowwise output bounds.

Theorem C.8. *Fix an arbitrary $p \in [0, 2]$. Consider a class of matrix-valued functions*

$$\mathcal{F} = \{f : \mathbb{R}^{d\times\ell} \to \mathbb{R}^{d\times m} \mid f(X) = XW,\ W \in \mathcal{W}\},$$

where the parameter set is defined by

$$\mathcal{W} = \{W \in \mathbb{R}^{\ell\times m} \mid \|W\|_{\mathrm{s},p}^p \le C_{\mathrm{s}},\ \|W\|_2 \le C_2\}.$$

Suppose $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ holds. Then, for any $\epsilon > 0$, we have

$$\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \lesssim \left(\frac{[C_{\mathrm{s}}(\ell+m)]^2 (B_{n,(2\to\infty)}^2 \min\{\ell, m\}m)^p}{\epsilon^{2p}}\right)^{\frac{1}{p+2}} \log C_{2\to\infty},$$

where $C_{2\to\infty}$ is given by

$$\begin{aligned}
C_{2\to\infty} = &\left(32C_{\mathrm{s}}^{\frac{1}{p+2}} C_2 \left(\frac{B_{n,(2\to\infty)}}{\epsilon}\right)^{\frac{2p+2}{p+2}} \left(\frac{\min\{\ell, m\}m}{\ell+m}\right)^{\frac{p}{2(p+2)}} + 1\right) \\
&\times \left[\left(640\left(\frac{C_{\mathrm{s}} B_{n,(2\to\infty)}^p \min\{\ell, m\}^{p/2}(\ell+m)}{m\epsilon^p}\right)^{\frac{1}{p+2}} + 60\right) nd\right].
\end{aligned}$$

Proof. Fix $p \in [0,2]$ and use the convention $\|W\|_{\mathrm{s},0}^0 = \mathrm{rank}(W)$ when $p = 0$. For any $W \in \mathcal{W}$, consider its singular value decomposition

$$W = \sum_{i=1}^{\min\{\ell,m\}} \sigma_i u_i v_i^\top, \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq 0.$$

Fix $\tau > 0$ and define $r = \max\{i \mid \sigma_i > \tau\}$, with the convention $r = 0$ if this set is empty. Consider the decomposition of W as $W = W_1 + W_2$, where

$$W_1 = \sum_{i=1}^{r} \sigma_i u_i v_i^\top, \qquad W_2 = W - W_1.$$

It holds that $r\tau^p \leq \sum_{i=1}^r \sigma_i^p \leq \sum_{i=1}^{\min\{\ell,m\}} \sigma_i^p = \|W\|_{\mathrm{s},p}^p \leq C_{\mathrm{s}}$, which implies

$$\mathrm{rank}(W_1) \leq r \leq \frac{C_{\mathrm{s}}}{\tau^p}.$$

Therefore, we have $W_1 \in \mathcal{W}_1 = \{W \in \mathcal{W} \mid \mathrm{rank}(W) \leq C_{\mathrm{s}}/\tau^p\}$.[5] Furthermore, for any $W, W' \in \mathcal{W}_1$, we have

$$\max_{i\in[n]} \|X_i(W - W')\|_{2\to\infty} \leq \max_{i\in[n]} \|X_i\|_{2\to\infty} \|W - W'\|_2 \leq B_{n,(2\to\infty)} \|W - W'\|_2.$$

Define the function class $\mathcal{F}_1 = \{f : \mathbb{R}^{d\times\ell} \to \mathbb{R}^{d\times m} \mid f(X) = XW,\ W \in \mathcal{W}_1\}$. By applying Proposition C.2 and Remark C.1, there exists a $\epsilon/4$-cover $\mathcal{C}_{\mathcal{F}_1}$ of $\mathcal{F}_1$ on $\{X_i\}_{i\in[n]}$ that satisfies

$$\log|\mathcal{C}_{\mathcal{F}_1}| \leq \log \mathcal{N}\left(\mathcal{W}_1, \|\cdot\|_2, \frac{\epsilon}{4B_{n,(2\to\infty)}}\right) \leq (\ell+m)\frac{C_{\mathrm{s}}}{\tau^p} \log\left(\frac{32 C_{\mathrm{s}}^{\frac{1}{2}} C_2 B_{n,(2\to\infty)}}{\tau^{\frac{p}{2}}\epsilon} + 1\right).$$

On the other hand, since

$$\|W_2\|_F^2 = \sum_{i=r+1}^{\min\{\ell,m\}} \sigma_i^2 \leq \sum_{i=r+1}^{\min\{\ell,m\}} \tau^2 \leq \tau^2 \min\{\ell, m\}$$

holds, we have $W_2 \in \{W \in \mathcal{W} \mid \|W\|_F \leq \tau \min\{\ell,m\}^{1/2}\}$. By Corollary C.7, there exists a $\epsilon/4$-cover $\mathcal{C}_{\mathcal{F}_2}$ of $\mathcal{F}_2$ on $\{X_i\}_{i\in[n]}$ that satisfies

$$\log|\mathcal{C}_{\mathcal{F}_2}| \lesssim \frac{\tau^2 \min\{\ell,m\} B_{n,(2\to\infty)}^2 m}{\epsilon^2} \log\left[\left(\frac{640\tau \min\{\ell,m\}^{1/2} B_{n,(2\to\infty)}}{\epsilon} + 60\right) nd\right].$$

Set $\bar{\mathcal{C}}_{\mathcal{F}} = \{f_1 + f_2 \mid f_1 \in \mathcal{C}_{\mathcal{F}_1}, f_2 \in \mathcal{C}_{\mathcal{F}_2}\}$. Then, for any $f = f_1 + f_2 \in \mathcal{F}$ with $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, there exist $\tilde{f}_1 \in \mathcal{C}_{\mathcal{F}_1}$ and $\tilde{f}_2 \in \mathcal{C}_{\mathcal{F}_2}$ that satisfy

$$\max_{i\in[n]} \|f_1(X_i) - \tilde{f}_1(X_i)\|_{2\to\infty} \leq \frac{\epsilon}{4}, \qquad \max_{i\in[n]} \|f_2(X_i) - \tilde{f}_2(X_i)\|_{2\to\infty} \leq \frac{\epsilon}{4}.$$

Hence, it holds for $\tilde{f} = \tilde{f}_1 + \tilde{f}_2$ that

$$\begin{aligned}
&\max_{i\in[n]} \|f(X_i) - \tilde{f}(X_i)\|_{2\to\infty} \leq \max_{i\in[n]} \|f_1(X_i) - \tilde{f}_1(X_i)\|_{2\to\infty} + \max_{i\in[n]} \|f_2(X_i) - \tilde{f}_2(X_i)\|_{2\to\infty} \\
&\leq \frac{\epsilon}{2},
\end{aligned}$$

which implies that $\bar{\mathcal{C}}_{\mathcal{F}}$ is a (possibly improper) $\epsilon/2$-cover of $\mathcal{F}$ on $\{X_i\}_{i\in[n]}$. Finally, by Lemma C.4, we have

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \\
&\leq \log|\bar{\mathcal{C}}_{\mathcal{F}}| \leq \log|\bar{\mathcal{C}}_{\mathcal{F}_1}| + \log|\bar{\mathcal{C}}_{\mathcal{F}_2}| \\
&\lesssim (\ell+m)\frac{C_{\mathrm{s}}}{\tau^p} \log\left(\frac{32 C_{\mathrm{s}}^{\frac{1}{2}} C_2 B_{n,(2\to\infty)}}{\tau^{\frac{p}{2}}\epsilon} + 1\right) \\
&\quad + \frac{\tau^2 \min\{\ell,m\} B_{n,(2\to\infty)}^2 m}{\epsilon^2} \log\left[\left(\frac{640\tau \min\{\ell,m\}^{1/2} B_{n,(2\to\infty)}}{\epsilon} + 60\right) nd\right].
\end{aligned}$$

[5] Note that this argument holds for both $p > 0$ and $p = 0$.

Choosing $\tau = \left(\frac{C_{\mathrm{s}}(\ell+m)\epsilon^2}{B^2_{n,(2\to\infty)}\min\{\ell,m\}m}\right)^{\frac{1}{p+2}}$, we have

$$\log \mathcal{N}_\infty(\mathcal{F}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \lesssim \left(\frac{[C_{\mathrm{s}}(\ell+m)]^2 (B^2_{n,(2\to\infty)}\min\{\ell,m\}m)^p}{\epsilon^{2p}}\right)^{\frac{1}{p+2}} \log C_{2\to\infty}.$$

□

C.4 Entropy bounds for the class of composite functions

We finally record a simple composition rule for covering numbers. This lemma allows us to construct a cover of a composed function class by combining a cover of the inner class with sample-dependent covers of the outer class, while keeping track of the Lipschitz propagation of approximation errors.

Lemma C.9. *Let (S_1, d_1) and (S_2, d_2) be metric spaces. Let $\mathcal{F}$ be a class of functions from S to (S_1, d_1), and let $\mathcal{G}$ be a class of functions from (S_1, d_1) to (S_2, d_2). Suppose that there exists a constant $L_{\mathcal{G}} > 0$ such that, for every $g \in \mathcal{G}$ and every $y, y' \in \{f(x_i) \mid f \in \mathcal{F}, i \in [n]\}$,*

$$d_2(g(y), g(y')) \le L_{\mathcal{G}} d_1(y, y')$$

holds. Suppose $\mathcal{C}_{\mathcal{F}}$ is an $\epsilon_{\mathcal{F}}$-cover of $\mathcal{F}$ on $\{x_i\}_{i\in[n]}$. For each $\tilde{f} \in \mathcal{C}_{\mathcal{F}}$, suppose $\mathcal{C}_{\mathcal{G}}(\tilde{f})$ is an $\epsilon_{\mathcal{G}}$-cover of $\mathcal{G}$ on $\{\tilde{f}(x_i)\}_{i\in[n]}$. Then, $\mathcal{C} = \{\tilde{g} \circ \tilde{f} \mid \tilde{f} \in \mathcal{C}_{\mathcal{F}}, \tilde{g} \in \mathcal{C}_{\mathcal{G}}(\tilde{f})\}$ is an $(L_{\mathcal{G}}\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{G}})$-cover of $\mathcal{G} \circ \mathcal{F}$.

Proof. For any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, there exist $\tilde{f} \in \mathcal{C}_{\mathcal{F}}$ and $\tilde{g} \in \mathcal{C}_{\mathcal{G}}(\tilde{f})$ that satisfy

$$d_1(f(x_i), \tilde{f}(x_i)) \le \epsilon_{\mathcal{F}}, \qquad d_2(g(\tilde{f}(x_i)), \tilde{g}(\tilde{f}(x_i))) \le \epsilon_{\mathcal{G}}$$

for every $i \in [n]$. Therefore, by the triangle inequality and the Lipschitz continuity of g, we have

$$\begin{aligned} d_2(g \circ f(x_i), \tilde{g} \circ \tilde{f}(x_i)) &\le d_2(g \circ f(x_i), g \circ \tilde{f}(x_i)) + d_2(g \circ \tilde{f}(x_i), \tilde{g} \circ \tilde{f}(x_i)) \\ &\le L_{\mathcal{G}} d_1(f(x_i), \tilde{f}(x_i)) + d_2(g(\tilde{f}(x_i)), \tilde{g}(\tilde{f}(x_i))) \\ &\le L_{\mathcal{G}}\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{G}}. \end{aligned}$$

Thus, $\mathcal{C}$ is an $(L_{\mathcal{G}}\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{G}})$-cover of $\mathcal{G} \circ \mathcal{F}$. □

D Proofs of the main results

Throughout this section, the $O(\cdot)$ notation suppresses all logarithmic factors except those involving n and T.

D.1 Covering number bounds for Transformer heads

Consider a Transformer head $f_{\mathrm{head}}(\,\cdot\,; W^{QK}, W^V) : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$ defined in Eq. (1). We define the function class of the Transformer head as

$$\mathcal{F}_{\mathrm{head}} = \{f_{\mathrm{head}}(\,\cdot\,; W^{QK}, W^V) \mid W^{QK} \in \mathcal{W}^{QK}(p^{QK}, C_{\mathrm{s}}^{QK}),\ W^V \in \mathcal{W}^V(p^V, C_{\mathrm{s}}^V)\},$$

where, for fixed $p^{QK}, p^V \in [0, 2]$, the parameter sets $\mathcal{W}^{QK}, \mathcal{W}^V$ are defined by

$$\begin{aligned} \mathcal{W}^{QK}(p^{QK}, C_{\mathrm{s}}^{QK}) &= \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{\mathrm{s},p^{QK}}^{p^{QK}} \le C_{\mathrm{s}}^{QK},\ \|W\|_2 \le C_2^{QK}\}, \\ \mathcal{W}^V(p^V, C_{\mathrm{s}}^V) &= \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{\mathrm{s},p^V}^{p^V} \le C_{\mathrm{s}}^V,\ \|W\|_2 \le C_2^V\}. \end{aligned} \tag{8}$$

Proposition D.1. *Suppose $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ holds. Then, for any $\epsilon^{QK}, \epsilon^V > 0$, we have*

$$\begin{aligned} &\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{head}}, \|\cdot\|_{2\to\infty}, 2C_2^V B^2_{n,(2\to\infty)}\epsilon^{QK} + \epsilon^V; \{X_i\}_{i\in[n]}) \\ &\lesssim \left[\left(\frac{(C_{\mathrm{s}}^{QK})^2 B^{2p^{QK}}_{n,(2\to\infty)} N^{p^{QK}}}{(\epsilon^{QK})^{2p^{QK}}}\right)^{\frac{1}{p^{QK}+2}} + \left(\frac{(C_{\mathrm{s}}^V)^2 B^{2p^V}_{n,(2\to\infty)} N^{p^V}}{(\epsilon^V)^{2p^V}}\right)^{\frac{1}{p^V+2}}\right] N\log(nT), \end{aligned}$$

where we omit the logarithmic factors except $\log(nT)$*. Therefore, for any* $\epsilon > 0$*, setting* $\epsilon^{QK} = \epsilon/(4C_2^V B_{n,(2\to\infty)}^2)$ *and* $\epsilon^V = \epsilon/2$ *yields*

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{head}}, \|\cdot\|_{2\to\infty}, \epsilon; \{X_i\}_{i\in[n]}) \\
&\lesssim \left[\left(\frac{(C_{\mathrm{s}}^{QK})^2 (C_2^V)^{2p^{QK}} B_{n,(2\to\infty)}^{6p^{QK}} N^{p^{QK}}}{\epsilon^{2p^{QK}}} \right)^{\frac{1}{p^{QK}+2}} + \left(\frac{(C_{\mathrm{s}}^V)^2 B_{n,(2\to\infty)}^{2p^V} N^{p^V}}{\epsilon^{2p^V}} \right)^{\frac{1}{p^V+2}} \right] \\
&\quad \times N \log(nT).
\end{aligned}$$

Proof. Define linear function classes $\mathcal{F}_{QK}$ and $\mathcal{F}_V$ by

$$\begin{aligned}
\mathcal{F}_{QK} &= \{f : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N} \mid f(X) = XW,\ W \in \mathcal{W}^{QK}\}, \\
\mathcal{F}_V &= \{f : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N} \mid f(X) = XW,\ W \in \mathcal{W}^{V}\}.
\end{aligned}$$

Suppose $\mathcal{C}_{\mathcal{F}_{QK}}$ is an ϵ^{QK}-cover of $\mathcal{F}_{QK}$, and $\mathcal{C}_{\mathcal{F}_V}$ is an ϵ^V-cover of $\mathcal{F}_V$. By Theorem C.8, we can take $\mathcal{C}_{\mathcal{F}_{QK}}$ and $\mathcal{C}_{\mathcal{F}_V}$ with cardinalities satisfying

$$\begin{aligned}
\log|\mathcal{C}_{\mathcal{F}_{QK}}| &\lesssim \left(\frac{(C_{\mathrm{s}}^{QK})^2 B_{n,(2\to\infty)}^{2p^{QK}} N^{p^{QK}}}{(\epsilon^{QK})^{2p^{QK}}} \right)^{\frac{1}{p^{QK}+2}} N\log(nT), \\
\log|\mathcal{C}_{\mathcal{F}_V}| &\lesssim \left(\frac{(C_{\mathrm{s}}^{V})^2 B_{n,(2\to\infty)}^{2p^{V}} N^{p^{V}}}{(\epsilon^{V})^{2p^{V}}} \right)^{\frac{1}{p^{V}+2}} N\log(nT).
\end{aligned}$$

We denote the corresponding parameter sets of $\mathcal{C}_{\mathcal{F}_{QK}}$ and $\mathcal{C}_{\mathcal{F}_V}$ by $\mathcal{C}_{\mathcal{W}^{QK}}$ and $\mathcal{C}_{\mathcal{W}^V}$, respectively. We show that

$$\mathcal{C}_{\mathcal{F}_{\mathrm{head}}} = \{f_{\mathrm{head}}(\,\cdot\,; \tilde{W}^{QK}, \tilde{W}^V) \mid \tilde{W}^{QK} \in \mathcal{C}_{\mathcal{W}^{QK}}, \tilde{W}^V \in \mathcal{C}_{\mathcal{W}^V}\}$$

is a $(2C_2^V B_{n,(2\to\infty)}^2 \epsilon^{QK} + \epsilon^V)$-cover of $\mathcal{F}_{\mathrm{head}}$ on $\{X_i\}_{i\in[n]}$.

By the definition of $\mathcal{C}_{\mathcal{F}_{QK}}$ and $\mathcal{C}_{\mathcal{F}_V}$, for any $W^{QK} \in \mathcal{W}^{QK}$ and $W^V \in \mathcal{W}^V$, there exist $\tilde{W}^{QK} \in \mathcal{C}_{\mathcal{W}^{QK}}$ and $\tilde{W}^V \in \mathcal{C}_{\mathcal{W}^V}$ that satisfy

$$\max_{i\in[n]} \|X_i(W^{QK} - \tilde{W}^{QK})\|_{2\to\infty} \le \epsilon^{QK}, \qquad \max_{i\in[n]} \|X_i(W^V - \tilde{W}^V)\|_{2\to\infty} \le \epsilon^V.$$

For any $X \in \mathbb{R}^{T\times N}$, we first bound the distance between $f_{\mathrm{head}}(X; W^{QK}, W^V)$ and $f_{\mathrm{head}}(X; \tilde{W}^{QK}, \tilde{W}^V)$ as

$$\begin{aligned}
&\|f_{\mathrm{head}}(X; W^{QK}, W^V) - f_{\mathrm{head}}(X; \tilde{W}^{QK}, \tilde{W}^V)\|_{2\to\infty} \\
&= \|\mathrm{SoftMax}(XW^{QK}X^\top)XW^V - \mathrm{SoftMax}(X\tilde{W}^{QK}X^\top)X\tilde{W}^V\|_{2\to\infty} \\
&\le \|\mathrm{SoftMax}(XW^{QK}X^\top)X(W^V - \tilde{W}^V)\|_{2\to\infty} \\
&\quad + \|(\mathrm{SoftMax}(XW^{QK}X^\top) - \mathrm{SoftMax}(X\tilde{W}^{QK}X^\top))X\tilde{W}^V\|_{2\to\infty}.
\end{aligned}$$

The first term on the right-hand side can be bounded by

$$\begin{aligned}
&\|\mathrm{SoftMax}(XW^{QK}X^\top)X(W^V - \tilde{W}^V)\|_{2\to\infty} \\
&\le \|\mathrm{SoftMax}(XW^{QK}X^\top)\|_\infty \|X(W^V - \tilde{W}^V)\|_{2\to\infty} \le \|X(W^V - \tilde{W}^V)\|_{2\to\infty},
\end{aligned}$$

where the last inequality follows from Lemma B.6. For the second term, it holds that

$$\begin{aligned}
&\|(\mathrm{SoftMax}(XW^{QK}X^\top) - \mathrm{SoftMax}(X\tilde{W}^{QK}X^\top))X\tilde{W}^V\|_{2\to\infty} \\
&= \max_{t\in[T]} \|(\mathrm{SoftMax}_{t\cdot}(XW^{QK}X^\top) - \mathrm{SoftMax}_{t\cdot}(X\tilde{W}^{QK}X^\top))X\tilde{W}^V\|_2 \\
&\le \max_{t\in[T]} \|(\mathrm{SoftMax}_{t\cdot}(XW^{QK}X^\top) - \mathrm{SoftMax}_{t\cdot}(X\tilde{W}^{QK}X^\top))^\top\|_1 \|X\tilde{W}^V\|_{2\to\infty} \\
&\le 2\max_{t\in[T]} \|((XW^{QK}X^\top)_{t\cdot} - (X\tilde{W}^{QK}X^\top)_{t\cdot})^\top\|_\infty \|X\|_{2\to\infty} \|\tilde{W}^V\|_2,
\end{aligned}$$

where the first inequality follows from Lemma B.4 and the second inequality follows from Lemmas B.7 and B.2. Using Lemma B.2 once again, we further bound as

$$\begin{aligned}
&\max_{t\in[T]} \|((XW^{QK}X^\top)_{t\cdot} - (X\tilde{W}^{QK}X^\top)_{t\cdot})^\top\|_\infty = \max_{t\in[T]}\max_{s\in[T]} |X_{t\cdot}(W^{QK}-\tilde{W}^{QK})(X_{s\cdot})^\top| \\
&= \max_{s\in[T]} \|X(W^{QK}-\tilde{W}^{QK})(X_{s\cdot})^\top\|_{2\to\infty} \\
&\le \max_{s\in[T]} \|X(W^{QK}-\tilde{W}^{QK})\|_{2\to\infty}\|(X_{s\cdot})^\top\|_2 = \|X(W^{QK}-\tilde{W}^{QK})\|_{2\to\infty}\|X\|_{2\to\infty},
\end{aligned}$$

where the last equality uses $\|v\|_2 = \|v^\top\|_2$ for any vector v. Combining the results, we have

$$\begin{aligned}
&\max_{i\in[n]} \|f_{\mathrm{head}}(X_i; W^{QK}, W^V) - f_{\mathrm{head}}(X_i; \tilde{W}^{QK}, \tilde{W}^V)\|_{2\to\infty} \\
&\le \max_{i\in[n]} \|X_i(W^V-\tilde{W}^V)\|_{2\to\infty} + 2\max_{i\in[n]} \|X_i(W^{QK}-\tilde{W}^{QK})\|_{2\to\infty}\|X_i\|_{2\to\infty}^2\|\tilde{W}^V\|_2 \\
&\le \epsilon^V + 2\epsilon^{QK}B_{n,(2\to\infty)}^2 C_2^V,
\end{aligned}$$

which implies that $\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}$ is a $(2C_2^V B_{n,(2\to\infty)}^2\epsilon^{QK}+\epsilon^V)$-cover of $\mathcal{F}_{\mathrm{head}}$ on $\{X_i\}_{i\in[n]}$. Thus, we obtain the entropy bound

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{head}}, \|\cdot\|_{2\to\infty}, 2C_2^V B_{n,(2\to\infty)}^2\epsilon^{QK}+\epsilon^V; \{X_i\}_{i\in[n]}) \\
&\le \log|\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}| \\
&\le \log|\mathcal{C}_{\mathcal{F}_{QK}}| + \log|\mathcal{C}_{\mathcal{F}_V}| \\
&\lesssim \left[\left(\frac{(C_{\mathrm{s}}^{QK})^2 B_{n,(2\to\infty)}^{2p^{QK}} N^{p^{QK}}}{(\epsilon^{QK})^{2p^{QK}}}\right)^{\frac{1}{p^{QK}+2}} + \left(\frac{(C_{\mathrm{s}}^{V})^2 B_{n,(2\to\infty)}^{2p^{V}} N^{p^{V}}}{(\epsilon^{V})^{2p^{V}}}\right)^{\frac{1}{p^{V}+2}}\right] N\log(nT).
\end{aligned}$$

□

D.2 Covering number bounds for Transformer blocks

We define the Transformer block $f_{\mathrm{block}}(\,\cdot\,; W^{QK}, W^V, W^M) : \mathbb{R}^{T\times N}\to\mathbb{R}^{T\times N}$ by Eq. (2). Define $\mathcal{W}^{QK}(p^{QK}, C_{\mathrm{s}}^{QK})$ and $\mathcal{W}^V(p^V, C_{\mathrm{s}}^V)$ by Eq. (8), and define

$$\mathcal{W}^M(p^M, C_{\mathrm{s}}^M) = \{W\in\mathbb{R}^{N\times N} \mid \|W\|_{\mathrm{s},p^M}^{p^M} \le C_{\mathrm{s}}^M,\ \|W\|_2 \le C_2^M\}.$$

We introduce the class of Transformer blocks as

$$\mathcal{F}_{\mathrm{block}} = \{f_{\mathrm{block}}(\,\cdot\,; W^{QK}, W^V, W^M) \mid W^\star\in\mathcal{W}^\star(p^\star, C_{\mathrm{s}}^\star),\ \star\in\{QK, V, M\}\}.$$

We first note that by the Lipschitz property of the activation function $\phi:\mathbb{R}^N\to\mathbb{R}^N$ in Assumption 3.1, the same Lipschitz property holds for the rowwise extension of ϕ with respect to the $\|\cdot\|_{2\to\infty}$ norm. Indeed, for all $Z, Z'\in\mathbb{R}^{T\times N}$, it holds that

$$\|\phi(Z)-\phi(Z')\|_{2\to\infty} = \max_{t\in[T]}\|\phi(Z_{t\cdot})-\phi(Z'_{t\cdot})\|_2 \le L_\phi \max_{t\in[T]}\|Z_{t\cdot}-Z'_{t\cdot}\|_2 = L_\phi\|Z-Z'\|_{2\to\infty}. \quad (9)$$

Proposition D.2. *Suppose* $\max_{i\in[n]}\|X_i\|_{2\to\infty}\le B_{n,(2\to\infty)}$ *holds. Then, for any* $\epsilon^{QK}, \epsilon^V, \epsilon^M > 0$*, we have*

$$\begin{aligned}
&\log\mathcal{N}_\infty(\mathcal{F}_{\mathrm{block}}, \|\cdot\|_{2\to\infty}, 2L_\phi C_2^V C_2^M B_{n,(2\to\infty)}^2\epsilon^{QK} + L_\phi C_2^M\epsilon^V + \epsilon^M; \{X_i\}_{i\in[n]}) \\
&\lesssim (\Upsilon^{QK}(\epsilon^{QK}) + \Upsilon^V(\epsilon^V) + \Upsilon^M(\epsilon^M))N\log(nT),
\end{aligned}$$

where

$$\Upsilon^\star(\epsilon^\star) = \left(\frac{(C_{\mathrm{s}}^\star)^2 B_{n,(2\to\infty)}^{2p^\star} N^{p^\star}}{(\epsilon^\star)^{2p^\star}}\right)^{\frac{1}{p^\star+2}}, \qquad \Upsilon^M(\epsilon^M) = \left(\frac{(C_{\mathrm{s}}^M)^2 L_\phi^{2p^M} N^{p^M}}{(\epsilon^M)^{2p^M}}\right)^{\frac{1}{p^M+2}}$$

with $\star\in\{QK, V\}$.

Proof. Define $\epsilon_{\mathrm{head}} = 2C_2^V B_{n,(2\to\infty)}^2 \epsilon^{QK} + \epsilon^V$, and suppose $\mathcal{C}_{\mathcal{F}_{\mathrm{head}}} \subset \mathcal{F}_{\mathrm{head}}$ is an ϵ_{head}-cover of $\mathcal{F}_{\mathrm{head}}$ on $\{X_i\}_{i\in[n]}$. Note that by Proposition D.1, we can choose $\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}$ that satisfies

$$\log|\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}| \lesssim [\Upsilon^{QK}(\epsilon^{QK}) + \Upsilon^V(\epsilon^V)]N\log(nT).$$

Consider a linear function class

$$\mathcal{F}_M = \{f : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N} \mid f(X) = XW^M,\ W^M \in \mathcal{W}^M\},$$

and define $\mathcal{C}_{\mathcal{F}_M}(\tilde{f}_{\mathrm{head}})$ as an ϵ^M-cover of $\mathcal{F}_M$ on $\{\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))\}_{i\in[n]}$. Note that the input $\{\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))\}_{i\in[n]}$ has the norm bound

$$\begin{aligned}
&\max_{i\in[n]} \|\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))\|_{2\to\infty} = \max_{i\in[n]} \|\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))) - \phi(0)\|_{2\to\infty} \\
&\le L_\phi \max_{i\in[n]} \|\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))\|_{2\to\infty} \le L_\phi,
\end{aligned}$$

thanks to the normalization Π_{norm}. Thus, by Theorem C.8, we can choose such a cover with cardinality

$$\log|\mathcal{C}_{\mathcal{F}_M}(\tilde{f}_{\mathrm{head}})| \lesssim \Upsilon^M(\epsilon^M)N\log(nT).$$

Define

$$\mathcal{C}_{\mathcal{F}_{\mathrm{block}}} = \{X \mapsto \Pi_{\mathrm{norm}} \circ \tilde{f}_M \circ \phi \circ \Pi_{\mathrm{norm}} \circ \tilde{f}_{\mathrm{head}}(X) \mid \tilde{f}_{\mathrm{head}} \in \mathcal{C}_{\mathcal{F}_{\mathrm{head}}}, \tilde{f}_M \in \mathcal{C}_{\mathcal{F}_M}(\tilde{f}_{\mathrm{head}})\}.$$

Then, for any $f_{\mathrm{block}} = \Pi_{\mathrm{norm}} \circ f_M \circ \phi \circ \Pi_{\mathrm{norm}} \circ f_{\mathrm{head}} \in \mathcal{F}_{\mathrm{block}}$, there exist $\tilde{f}_{\mathrm{head}} \in \mathcal{C}_{\mathcal{F}_{\mathrm{head}}}$ and $\tilde{f}_M \in \mathcal{C}_{\mathcal{F}_M}(\tilde{f}_{\mathrm{head}})$ that satisfy

$$\begin{aligned}
&\|f_{\mathrm{head}}(X_i) - \tilde{f}_{\mathrm{head}}(X_i)\|_{2\to\infty} \le \epsilon_{\mathrm{head}}, \\
&\|f_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))) - \tilde{f}_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))))\|_{2\to\infty} \le \epsilon^M.
\end{aligned}$$

By Lemma B.5 and Eq. (9), we have

$$\begin{aligned}
&\|f_{\mathrm{block}}(X_i) - \tilde{f}_{\mathrm{block}}(X_i)\|_{2\to\infty} \\
&= \|\Pi_{\mathrm{norm}} \circ f_M \circ \phi \circ \Pi_{\mathrm{norm}} \circ f_{\mathrm{head}}(X_i) - \Pi_{\mathrm{norm}} \circ \tilde{f}_M \circ \phi \circ \Pi_{\mathrm{norm}} \circ \tilde{f}_{\mathrm{head}}(X_i)\|_{2\to\infty} \\
&\le \|f_M(\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X_i)))) - \tilde{f}_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))))\|_{2\to\infty}.
\end{aligned}$$

By a computation similar to the proof of Lemma C.9, we can bound the norm on the right-hand side as

$$\begin{aligned}
&\|f_M(\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X_i)))) - \tilde{f}_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))))\|_{2\to\infty} \\
&\le \|f_M(\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X_i)))) - f_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))))\|_{2\to\infty} \\
&\quad + \|f_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))) - \tilde{f}_M(\phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i))))\|_{2\to\infty} \\
&\le \|[\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X_i))) - \phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))]W^M\|_{2\to\infty} + \epsilon^M.
\end{aligned}$$

By Lemma B.2 and Assumption 3.1, we further bound the first term as

$$\begin{aligned}
&\|[\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(X_i))) - \phi(\Pi_{\mathrm{norm}}(\tilde{f}_{\mathrm{head}}(X_i)))]W^M\|_{2\to\infty} \\
&\le L_\phi \|f_{\mathrm{head}}(X_i) - \tilde{f}_{\mathrm{head}}(X_i)\|_{2\to\infty} \|W^M\|_2 \\
&\le L_\phi C_2^M \epsilon_{\mathrm{head}}.
\end{aligned}$$

Hence, we have

$$\|f_{\mathrm{block}}(X_i) - \tilde{f}_{\mathrm{block}}(X_i)\|_{2\to\infty} \le L_\phi C_2^M \epsilon_{\mathrm{head}} + \epsilon^M,$$

which implies that $\mathcal{C}_{\mathcal{F}_{\mathrm{block}}}$ is an $L_\phi C_2^M \epsilon_{\mathrm{head}} + \epsilon^M$-cover of $\mathcal{F}_{\mathrm{block}}$ on $\{X_i\}_{i\in[n]}$. Finally, we obtain the entropy bound

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{block}}, \|\cdot\|_{2\to\infty}, 2L_\phi C_2^V C_2^M B_{n,(2\to\infty)}^2 \epsilon^{QK} + L_\phi C_2^M \epsilon^V + \epsilon^M; \{X_i\}_{i\in[n]}) \\
&\le \log|\mathcal{C}_{\mathcal{F}_{\mathrm{block}}}| \\
&\le \log|\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}| + \sup_{\tilde{f}_{\mathrm{head}}\in\mathcal{C}_{\mathcal{F}_{\mathrm{head}}}} \log|\mathcal{C}_{\mathcal{F}_M}(\tilde{f}_{\mathrm{head}})| \\
&\lesssim (\Upsilon^{QK}(\epsilon^{QK}) + \Upsilon^V(\epsilon^V) + \Upsilon^M(\epsilon^M))N\log(nT).
\end{aligned}$$

□

D.3 Covering number bounds for multi-layer Transformers

We next consider a multi-layer Transformer in Eq. (3). For each $\star \in \{QK, V, M\}$ and $\ell \in [L]$, set

$$\mathcal{W}^{\star,(\ell)}(p^{\star,(\ell)}, C_{\mathrm{s}}^{\star,(\ell)}) = \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}} \le C_{\mathrm{s}}^{\star,(\ell)},\ \|W\|_2 \le C_2^{\star,(\ell)}\}.$$

We also define

$$\mathcal{W}^{(\ell)}(\boldsymbol{p}^{(\ell)}, \boldsymbol{C}_{\mathrm{s}}^{(\ell)}) = \mathcal{W}^{QK,(\ell)}(p^{QK,(\ell)}, C_{\mathrm{s}}^{QK,(\ell)}) \times \mathcal{W}^{V,(\ell)}(p^{V,(\ell)}, C_{\mathrm{s}}^{V,(\ell)}) \times \mathcal{W}^{M,(\ell)}(p^{M,(\ell)}, C_{\mathrm{s}}^{M,(\ell)})$$

and

$$\mathcal{W}^{(1:\ell)}(\boldsymbol{p}^{(1:\ell)}, \boldsymbol{C}_{\mathrm{s}}^{(1:\ell)}) = \mathcal{W}^{(1)}(\boldsymbol{p}^{(1)}, \boldsymbol{C}_{\mathrm{s}}^{(1)}) \times \cdots \times \mathcal{W}^{(\ell)}(\boldsymbol{p}^{(\ell)}, \boldsymbol{C}_{\mathrm{s}}^{(\ell)}) \qquad (\ell = 2, \dots, L).$$

We then define the class of ℓ-layer Transformers by

$$\mathcal{F}_{\mathrm{tf}}^{(\ell)} = \{f : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N} \mid f(X) = f_{\mathrm{tf}}^{(\ell)}(X; W^{(1:\ell)}),\ W^{(1:\ell)} \in \mathcal{W}^{(1:\ell)}(\boldsymbol{p}^{(1:\ell)}, \boldsymbol{C}_{\mathrm{s}}^{(1:\ell)})\}.$$

For each $\ell \in [L]$, we also define

$$\mathcal{F}_{\mathrm{block}}^{(\ell)} = \{f : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N} \mid f(X) = f_{\mathrm{block}}(X; W^{(\ell)}), W^{(\ell)} \in \mathcal{W}^{(\ell)}(\boldsymbol{p}^{(\ell)}, \boldsymbol{C}_{\mathrm{s}}^{(\ell)})\}.$$

Lemma D.3. *For every $\ell \in \{2, \dots, L\}$, consider any $f \in \mathcal{F}_{\mathrm{block}}^{(\ell)}$. Suppose $Z, Z' \in \mathbb{R}^{T\times N}$ satisfy*

$$\|Z\|_{2\to\infty} \le B_{2\to\infty}, \qquad \|Z'\|_{2\to\infty} \le B_{2\to\infty}.$$

Then, we have

$$\|f(Z) - f(Z')\|_{2\to\infty} \le L_\phi C_2^{M,(\ell)} C_2^{V,(\ell)} (1 + 4C_2^{QK,(\ell)} B_{2\to\infty}^2) \|Z - Z'\|_{2\to\infty}.$$

Proof. Take any $f(\cdot) = f_{\mathrm{block}}(\ \cdot\ ; W^{QK,(\ell)}, W^{V,(\ell)}, W^{M,(\ell)}) \in \mathcal{F}_{\mathrm{block}}^{(\ell)}$. Then, by Lemmas B.5, B.2, and Eq. (9), we have

$$\begin{aligned}
&\|f(Z) - f(Z')\|_{2\to\infty} \\
&\le \|\phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(Z; W^{QK,(\ell)}, W^{V,(\ell)}))) - \phi(\Pi_{\mathrm{norm}}(f_{\mathrm{head}}(Z'; W^{QK,(\ell)}, W^{V,(\ell)})))\|_{2\to\infty} C_2^{M,(\ell)} \\
&\le L_\phi C_2^{M,(\ell)} \|f_{\mathrm{head}}(Z; W^{QK,(\ell)}, W^{V,(\ell)}) - f_{\mathrm{head}}(Z'; W^{QK,(\ell)}, W^{V,(\ell)})\|_{2\to\infty}.
\end{aligned}$$

Hence it suffices to bound the distance between the head outputs. Write

$$A = \mathrm{SoftMax}(ZW^{QK,(\ell)}Z^\top), \qquad A' = \mathrm{SoftMax}(Z'W^{QK,(\ell)}Z'^\top).$$

Then, we have

$$\begin{aligned}
&\|f_{\mathrm{head}}(Z; W^{QK,(\ell)}, W^{V,(\ell)}) - f_{\mathrm{head}}(Z'; W^{QK,(\ell)}, W^{V,(\ell)})\|_{2\to\infty} \\
&= \|AZW^{V,(\ell)} - A'Z'W^{V,(\ell)}\|_{2\to\infty} \\
&\le \|A(Z - Z')W^{V,(\ell)}\|_{2\to\infty} + \|(A - A')Z'W^{V,(\ell)}\|_{2\to\infty}. \qquad (10)
\end{aligned}$$

For the first term, using Lemmas B.6, B.1 and B.2, we have

$$\|A(Z - Z')W^{V,(\ell)}\|_{2\to\infty} \le \|(Z - Z')W^{V,(\ell)}\|_{2\to\infty} \le C_2^{V,(\ell)} \|Z - Z'\|_{2\to\infty}. \qquad (11)$$

For the second term, we first note that by Lemmas B.4 and B.2, we have

$$\begin{aligned}
\|(A - A')Z'W^{V,(\ell)}\|_{2\to\infty} &= \max_{t\in[T]} \|(A_{t\cdot} - A'_{t\cdot})Z'W^{V,(\ell)}\|_2 \\
&\le \max_{t\in[T]} \|(A_{t\cdot} - A'_{t\cdot})^\top\|_1 \|Z'W^{V,(\ell)}\|_{2\to\infty} \\
&\le \max_{t\in[T]} \|(A_{t\cdot} - A'_{t\cdot})^\top\|_1 \|Z'\|_{2\to\infty} \|W^{V,(\ell)}\|_2 \\
&\le \max_{t\in[T]} \|(A_{t\cdot} - A'_{t\cdot})^\top\|_1 B_{2\to\infty} C_2^{V,(\ell)}.
\end{aligned}$$

It remains to bound the row difference of the attention matrices. The rows $A_{t\cdot}$ and $A'_{t\cdot}$ are written as

$$A_{t\cdot} = \mathrm{SoftMax}((ZW^{QK,(\ell)}Z^\top)_{t\cdot}), \qquad A'_{t\cdot} = \mathrm{SoftMax}((Z'W^{QK,(\ell)}Z'^\top)_{t\cdot}).$$

Thus, Lemma B.7 yields

$$\begin{aligned}\|(A_{t\cdot} - A'_{t\cdot})^\top\|_1 &\le 2\|((ZW^{QK,(\ell)}Z^\top)_{t\cdot} - (Z'W^{QK,(\ell)}Z'^\top)_{t\cdot})^\top\|_\infty \\ &= 2\max_{s\in[T]} |Z_{t\cdot}W^{QK,(\ell)}(Z_{s\cdot})^\top - Z'_{t\cdot}W^{QK,(\ell)}(Z'_{s\cdot})^\top|.\end{aligned}$$

For each $s, t \in [T]$, we decompose

$$\begin{aligned}&|Z_{t\cdot}W^{QK,(\ell)}(Z_{s\cdot})^\top - Z'_{t\cdot}W^{QK,(\ell)}(Z'_{s\cdot})^\top| \\ &\le |Z_{t\cdot}W^{QK,(\ell)}\left(Z_{s\cdot} - Z'_{s\cdot}\right)^\top| + |\left(Z_{t\cdot} - Z'_{t\cdot}\right)W^{QK,(\ell)}(Z'_{s\cdot})^\top| \\ &\le \|Z_{t\cdot}\|_2\|W^{QK,(\ell)}\left(Z_{s\cdot} - Z'_{s\cdot}\right)^\top\|_2 + \|Z_{t\cdot} - Z'_{t\cdot}\|_2\|W^{QK,(\ell)}(Z'_{s\cdot})^\top\|_2 \\ &\le \|Z_{t\cdot}\|_2\|W^{QK,(\ell)}\|_2\|Z_{s\cdot} - Z'_{s\cdot}\|_2 + \|Z_{t\cdot} - Z'_{t\cdot}\|_2\|W^{QK,(\ell)}\|_2\|Z'_{s\cdot}\|_2 \\ &\le B_{2\to\infty}C_2^{QK,(\ell)}(\|Z_{s\cdot} - Z'_{s\cdot}\|_2 + \|Z_{t\cdot} - Z'_{t\cdot}\|_2).\end{aligned}$$

Substituting these results into the previous inequality yields

$$\begin{aligned}\|(A - A')Z'W^{V,(\ell)}\|_{2\to\infty} &\le 2C_2^{QK,(\ell)}C_2^{V,(\ell)}B_{2\to\infty}^2 \max_{t\in[T]}\max_{s\in[T]}(\|Z_{s\cdot} - Z'_{s\cdot}\|_2 + \|Z_{t\cdot} - Z'_{t\cdot}\|_2) \\ &\le 4C_2^{QK,(\ell)}C_2^{V,(\ell)}B_{2\to\infty}^2\|Z - Z'\|_{2\to\infty}.\end{aligned}$$

Combining this with Eq. (10) and (11), we have

$$\begin{aligned}&\|f_{\mathrm{head}}(Z; W^{QK,(\ell)}, W^{V,(\ell)}) - f_{\mathrm{head}}(Z'; W^{QK,(\ell)}, W^{V,(\ell)})\|_{2\to\infty} \\ &\le C_2^{V,(\ell)}(1 + 4C_2^{QK,(\ell)}B_{2\to\infty}^2)\|Z - Z'\|_{2\to\infty},\end{aligned}$$

which implies

$$\|f(Z) - f(Z')\|_{2\to\infty} \le L_\phi C_2^{M,(\ell)}C_2^{V,(\ell)}(1 + 4C_2^{QK,(\ell)}B_{2\to\infty}^2)\|Z - Z'\|_{2\to\infty}$$

as desired. □

Proposition D.4. *Suppose* $\max_{i\in[n]}\|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ *holds. For each* $\ell \in [L]$*, fix arbitrary* $\epsilon^{QK,(\ell)}, \epsilon^{V,(\ell)}, \epsilon^{M,(\ell)} > 0$*, and define*

$$\epsilon^{(\ell)} = 2L_\phi C_2^{V,(\ell)}C_2^{M,(\ell)}B_{n,(2\to\infty)}^{2\delta_{\ell=1}}\epsilon^{QK,(\ell)} + L_\phi C_2^{M,(\ell)}\epsilon^{V,(\ell)} + \epsilon^{M,(\ell)},$$

where $\delta_{\ell=1}$ *is an indicator function that equals* 1 *if* $\ell = 1$ *and* 0 *otherwise. We also define*

$$\Upsilon^{\star,(\ell)}(\epsilon^{\star,(\ell)}) = \left(\frac{(C_{\mathrm{s}}^{\star,(\ell)})^2 B_{n,(2\to\infty)}^{2\delta_{\ell=1}p^{\star,(\ell)}} N^{p^{\star,(\ell)}}}{(\epsilon^{\star,(\ell)})^{2p^{\star,(\ell)}}}\right)^{\frac{1}{p^{\star,(\ell)}+2}},$$

$$\Upsilon^{M,(\ell)}(\epsilon^{M,(\ell)}) = \left(\frac{(C_{\mathrm{s}}^{M,(\ell)})^2 L_\phi^{2p^{M,(\ell)}} N^{p^{M,(\ell)}}}{(\epsilon^{M,(\ell)})^{2p^{M,(\ell)}}}\right)^{\frac{1}{p^{M,(\ell)}+2}}$$

with $\star \in \{QK, V\}$*. Then, we have*

$$\begin{aligned}&\log\mathcal{N}_\infty(\mathcal{F}_{\mathrm{tf}}^{(L)}, \|\cdot\|_{2\to\infty}, \eta^{(L)}; \{X_i\}_{i\in[n]}) \\ &\lesssim \sum_{k=1}^{L}\left(\Upsilon^{QK,(k)}(\epsilon^{QK,(k)}) + \Upsilon^{V,(k)}(\epsilon^{V,(k)}) + \Upsilon^{M,(k)}(\epsilon^{M,(k)})\right)N\log(nT),\end{aligned}$$

where $\eta^{(L)}$ *is defined by*

$$\eta^{(L)} = \sum_{j=1}^{L-1}L_\phi^{L-j}\epsilon^{(j)}\prod_{k=j+1}^{L}C_2^{M,(k)}C_2^{V,(k)}(1 + 4C_2^{QK,(k)}) + \epsilon^{(L)}.$$

Proof. First, note that for all $\ell \in [L]$ and $f_{\mathrm{tf}}^{(\ell)} \in \mathcal{F}_{\mathrm{tf}}^{(\ell)}$, we have

$$\max_{i\in[n]} \|f_{\mathrm{tf}}^{(\ell)}(X_i)\|_{2\to\infty} \le 1. \tag{12}$$

Set

$$\begin{aligned}
&\eta^{(1)} = \epsilon^{(1)},\\
&\eta^{(\ell)} = \sum_{j=1}^{\ell-1} L_\phi^{\ell-j}\epsilon^{(j)} \prod_{k=j+1}^{\ell} C_2^{M,(k)} C_2^{V,(k)}(1+4C_2^{QK,(k)}) + \epsilon^{(\ell)} \quad (\ell = 2,\dots,L).
\end{aligned}$$

We show by induction that for each $\ell \in [L]$, there exists a proper $\eta^{(\ell)}$-cover $\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}$ of $\mathcal{F}_{\mathrm{tf}}^{(\ell)}$ on $\{X_i\}_{i\in[n]}$ with cardinality

$$\log|\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}| \lesssim \sum_{k=1}^{\ell}\left(\Upsilon^{QK,(k)}(\epsilon^{QK,(k)}) + \Upsilon^{V,(k)}(\epsilon^{V,(k)}) + \Upsilon^{M,(k)}(\epsilon^{M,(k)})\right) N\log(nT). \tag{13}$$

For the base case $\ell = 1$, the claim follows from Proposition D.2. Now, suppose the claim holds for some $\ell \in [L-1]$. We show that it also holds for $\ell+1$.

Suppose $\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}$ is a $\eta^{(\ell)}$-cover of $\mathcal{F}_{\mathrm{tf}}^{(\ell)}$ satisfying Eq. (13). For each $\tilde{f}_{\mathrm{tf}}^{(\ell)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}$, let $\mathcal{C}_{\mathcal{F}_{\mathrm{block}}^{(\ell+1)}}(\tilde{f}_{\mathrm{tf}}^{(\ell)})$ be an $\epsilon^{(\ell+1)}$-cover of $\mathcal{F}_{\mathrm{block}}^{(\ell+1)}$ on $\{\tilde{f}_{\mathrm{tf}}^{(\ell)}(X_i)\}_{i\in[n]}$. Note that by Proposition D.2 and Eq. (12), we can choose such a cover with cardinality

$$\begin{aligned}
&\log|\mathcal{C}_{\mathcal{F}_{\mathrm{block}}^{(\ell+1)}}(\tilde{f}_{\mathrm{tf}}^{(\ell)})|\\
&\lesssim (\Upsilon^{QK,(\ell+1)}(\epsilon^{QK,(\ell+1)}) + \Upsilon^{V,(\ell+1)}(\epsilon^{V,(\ell+1)}) + \Upsilon^{M,(\ell+1)}(\epsilon^{M,(\ell+1)}))N\log(nT).
\end{aligned}$$

By Lemma D.3 and Eq. (12), we have for any $i \in [n]$, $\ell \in [L-1]$, $f_{\mathrm{block}}^{(\ell+1)} \in \mathcal{F}_{\mathrm{block}}^{(\ell+1)}$ and $f_{\mathrm{tf}}^{(\ell)}, \tilde{f}_{\mathrm{tf}}^{(\ell)} \in \mathcal{F}_{\mathrm{tf}}^{(\ell)}$ that

$$\begin{aligned}
&\|f_{\mathrm{block}}^{(\ell+1)}(f_{\mathrm{tf}}^{(\ell)}(X_i)) - f_{\mathrm{block}}^{(\ell+1)}(\tilde{f}_{\mathrm{tf}}^{(\ell)}(X_i))\|_{2\to\infty}\\
&\le L_\phi C_2^{M,(\ell+1)} C_2^{V,(\ell+1)}(1+4C_2^{QK,(\ell+1)})\|f_{\mathrm{tf}}^{(\ell)}(X_i) - \tilde{f}_{\mathrm{tf}}^{(\ell)}(X_i)\|_{2\to\infty}.
\end{aligned}$$

The recursive relation

$$\eta^{(\ell+1)} = L_\phi C_2^{M,(\ell+1)} C_2^{V,(\ell+1)}(1+4C_2^{QK,(\ell+1)})\eta^{(\ell)} + \epsilon^{(\ell+1)}.$$

holds. Therefore, Lemma C.9 implies that

$$\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell+1)}} = \{\tilde{f}_{\mathrm{block}}^{(\ell+1)} \circ \tilde{f}_{\mathrm{tf}}^{(\ell)} \mid \tilde{f}_{\mathrm{tf}}^{(\ell)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}, \tilde{f}_{\mathrm{block}}^{(\ell+1)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{block}}^{(\ell+1)}}(\tilde{f}_{\mathrm{tf}}^{(\ell)})\}$$

is an $\eta^{(\ell+1)}$-cover of $\mathcal{F}_{\mathrm{tf}}^{(\ell+1)}$. Moreover, we can bound the cardinality of $\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell+1)}}$ as

$$\begin{aligned}
&\log|\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell+1)}}|\\
&\le \log|\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}| + \sup_{\tilde{f}_{\mathrm{tf}}^{(\ell)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(\ell)}}} \log|\mathcal{C}_{\mathcal{F}_{\mathrm{block}}^{(\ell+1)}}(\tilde{f}_{\mathrm{tf}}^{(\ell)})|\\
&\lesssim \sum_{k=1}^{\ell+1}\left(\Upsilon^{QK,(k)}(\epsilon^{QK,(k)}) + \Upsilon^{V,(k)}(\epsilon^{V,(k)}) + \Upsilon^{M,(k)}(\epsilon^{M,(k)})\right) N\log(nT).
\end{aligned}$$

□

D.4 Covering number bounds for the scalar outputs of multi-layer Transformers

For a vector of Schatten indices

$$\boldsymbol{p} = (p^{QK,(1)}, p^{V,(1)}, p^{M,(1)}, \dots, p^{QK,(L)}, p^{V,(L)}, p^{M,(L)}) \in [0,2]^{3L}$$

and a vector of Schatten-quantity radii

$$\boldsymbol{C}_{\mathrm{s}} = (C_{\mathrm{s}}^{QK,(1)}, C_{\mathrm{s}}^{V,(1)}, C_{\mathrm{s}}^{M,(1)}, \ldots, C_{\mathrm{s}}^{QK,(L)}, C_{\mathrm{s}}^{V,(L)}, C_{\mathrm{s}}^{M,(L)}) \in (0,\infty)^{3L},$$

consider the class of scalar outputs

$$\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})} = \{f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) : \mathbb{R}^{T\times N} \to \mathbb{R} \mid W^{(1:L)} \in \mathcal{W}^{(1:L)}(\boldsymbol{p}, \boldsymbol{C}_{\mathrm{s}}),\ w \in \mathbb{R}^N,\ \|w\|_2 \le C_2^{\mathrm{out}}\},$$

where $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w)$ is defined in Eq. (4). For $\ell \in [L]$, we also define

$$\alpha^{(\ell)} = \prod_{k=\ell+1}^{L} L_\phi C_2^{M,(k)} C_2^{V,(k)} (1 + 4C_2^{QK,(k)}), \tag{14}$$

where we adopt the convention that an empty product is equal to 1.

Proposition D.5. *Assume the same setting as in Proposition D.4 for the localized class $\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$. Then, for any $\epsilon^{\mathrm{out}} > 0$, it holds that*

$$\begin{aligned}
&\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}, |\cdot|, C_2^{\mathrm{out}}\eta^{(L)} + \epsilon^{\mathrm{out}}; \{X_i\}_{i\in[n]}) \\
&\lesssim \sum_{k=1}^{L} \left(\Upsilon^{QK,(k)}(\epsilon^{QK,(k)}) + \Upsilon^{V,(k)}(\epsilon^{V,(k)}) + \Upsilon^{M,(k)}(\epsilon^{M,(k)})\right) N\log(nT) + \left(\frac{C_2^{\mathrm{out}}}{\epsilon^{\mathrm{out}}}\right)^2 \log(n).
\end{aligned}$$

Proof. Proposition D.4 implies that there exists an $\eta^{(L)}$-cover $\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}$ of $\mathcal{F}_{\mathrm{tf}}^{(L)}$ on $\{X_i\}_{i\in[n]}$ that satisfies

$$\log|\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}| \lesssim \sum_{k=1}^{L} \left(\Upsilon^{QK,(k)}(\epsilon^{QK,(k)}) + \Upsilon^{V,(k)}(\epsilon^{V,(k)}) + \Upsilon^{M,(k)}(\epsilon^{M,(k)})\right) N\log(nT).$$

Moreover, for every $\tilde{f}_{\mathrm{tf}}^{(L)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}$, it holds that

$$\max_{i\in[n]} \|[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}\|_2 \le \max_{i\in[n]} \|\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)\|_{2\to\infty} \le 1.$$

Fix $\tilde{f}_{\mathrm{tf}}^{(L)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}$. Consider the linear function class

$$\mathcal{F}_w = \{f : \mathbb{R}^N \to \mathbb{R} \mid f(x) = w^\top x, w \in \mathbb{R}^N, \|w\|_2 \le C_2^{\mathrm{out}}\}.$$

Applying Proposition C.6, we obtain an ϵ^{out}-cover $\mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\mathrm{tf}}^{(L)})$ of $\mathcal{F}_w$ on $\{[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}\}_{i\in[n]}$ satisfying

$$\log|\mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\mathrm{tf}}^{(L)})| \lesssim \left(\frac{C_2^{\mathrm{out}}}{\epsilon^{\mathrm{out}}}\right)^2 \log(n).$$

Now define

$$\mathcal{C}_{\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}} = \{f : \mathbb{R}^{T\times N} \to \mathbb{R} \mid f(X) = \tilde{w}^\top[\tilde{f}_{\mathrm{tf}}^{(L)}(X)]_{[\mathrm{CLS}]},\ \tilde{f}_{\mathrm{tf}}^{(L)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}},\ \tilde{w}^\top(\cdot) \in \mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\mathrm{tf}}^{(L)})\}.$$

Take any $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$. By the definition of $\mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}$, there exists $\tilde{f}_{\mathrm{tf}}^{(L)} \in \mathcal{C}_{\mathcal{F}_{\mathrm{tf}}^{(L)}}$ that satisfies

$$\max_{i\in[n]} \|f_{\mathrm{tf}}^{(L)}(X_i; W^{(1:L)}) - \tilde{f}_{\mathrm{tf}}^{(L)}(X_i)\|_{2\to\infty} \le \eta^{(L)}.$$

Also, by the definition of $\mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\mathrm{tf}}^{(L)})$, there exists $\tilde{w}^\top(\cdot) \in \mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\mathrm{tf}}^{(L)})$ that satisfies

$$\max_{i\in[n]} |(w-\tilde{w})^\top[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}| \le \epsilon^{\mathrm{out}}.$$

For such $\tilde{f}_{\mathrm{tf}}^{(L)}$ and $\tilde{w}$, we have

$$\begin{aligned}
&|f_{\mathrm{out}}(X_i; W^{(1:L)}, w) - \tilde{w}^\top[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}| \\
&\le |w^\top([f_{\mathrm{tf}}^{(L)}(X_i; W^{(1:L)})]_{[\mathrm{CLS}]} - [\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]})| + |(w-\tilde{w})^\top[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}| \\
&\le \|w\|_2 \|[f_{\mathrm{tf}}^{(L)}(X_i; W^{(1:L)})]_{[\mathrm{CLS}]} - [\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}\|_2 + |(w-\tilde{w})^\top[\tilde{f}_{\mathrm{tf}}^{(L)}(X_i)]_{[\mathrm{CLS}]}| \\
&\le C_2^{\mathrm{out}} \|f_{\mathrm{tf}}^{(L)}(X_i; W^{(1:L)}) - \tilde{f}_{\mathrm{tf}}^{(L)}(X_i)\|_{2\to\infty} + \epsilon^{\mathrm{out}} \\
&\le C_2^{\mathrm{out}}\eta^{(L)} + \epsilon^{\mathrm{out}}.
\end{aligned}$$

Therefore, $\mathcal{C}_{\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})}}$ is a $(C_2^{\rm out}\eta^{(L)}+\epsilon^{\rm out})$-cover of $\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})}$. Finally, its cardinality is bounded by

$$\log|\mathcal{C}_{\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})}}| \le \log|\mathcal{C}_{\mathcal{F}_{\rm tf}^{(L)}}| + \sup_{\tilde{f}_{\rm tf}^{(L)}\in\mathcal{C}_{\mathcal{F}_{\rm tf}^{(L)}}} \log|\mathcal{C}_{\mathcal{F}_w}(\tilde{f}_{\rm tf}^{(L)})|,$$

and substituting the above two entropy bounds proves the claim. □

D.5 ϵ-entropy bounds for the scalar outputs of multi-layer Transformers

We now derive the ϵ-entropy bounds for $\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})}$. For $\ell\in[L]$, define

$$\alpha^{(\ell)} = L_\phi^{L-\ell}\prod_{k=\ell+1}^{L} C_2^{M,(k)}C_2^{V,(k)}(1+4C_2^{QK,(k)}),$$

where we adopt the convention that an empty product is equal to 1. We also define

$$\beta^{QK,(\ell)} = L_\phi C_2^{\rm out}\alpha^{(\ell)}2C_2^{V,(\ell)}C_2^{M,(\ell)}B_{n,(2\to\infty)}^{2\delta_{\ell=1}},$$
$$\beta^{V,(\ell)} = L_\phi C_2^{\rm out}\alpha^{(\ell)}C_2^{M,(\ell)}, \qquad \beta^{M,(\ell)} = C_2^{\rm out}\alpha^{(\ell)}.$$

Then, we can write

$$C_2^{\rm out}\eta^{(L)} = \sum_{\ell=1}^{L}(\beta^{QK,(\ell)}\epsilon^{QK,(\ell)}+\beta^{V,(\ell)}\epsilon^{V,(\ell)}+\beta^{M,(\ell)}\epsilon^{M,(\ell)}).$$

Thus, we need to choose $\epsilon^{QK,(\ell)},\epsilon^{V,(\ell)},\epsilon^{M,(\ell)}$ for $\ell\in[L]$ and $\epsilon^{\rm out}>0$ such that

$$\epsilon = \sum_{\ell=1}^{L}(\beta^{QK,(\ell)}\epsilon^{QK,(\ell)}+\beta^{V,(\ell)}\epsilon^{V,(\ell)}+\beta^{M,(\ell)}\epsilon^{M,(\ell)})+\epsilon^{\rm out} \tag{15}$$

holds. Furthermore, setting

$$\Lambda^{\star,(\ell)} = ((C_{\rm s}^{\star,(\ell)})^2 B_{n,(2\to\infty)}^{2\delta_{\ell=1}p^{\star,(\ell)}} N^{p^{\star,(\ell)}})^{\frac{1}{p^{\star,(\ell)}+2}} \quad (\star\in\{QK,V\}),$$
$$\Lambda^{M,(\ell)} = ((C_{\rm s}^{M,(\ell)})^2 L_\phi^{2p^{M,(\ell)}} N^{p^{M,(\ell)}})^{\frac{1}{p^{M,(\ell)}+2}},$$
$$\nu^{\star,(\ell)} = \frac{2p^{\star,(\ell)}}{p^{\star,(\ell)}+2} \quad (\star\in\{QK,V,M\}),$$

for each $\ell\in[L]$, we can write

$$\Upsilon^{\star,(\ell)}(\epsilon^{\star,(\ell)}) = \Lambda^{\star,(\ell)}(\epsilon^{\star,(\ell)})^{-\nu^{\star,(\ell)}} \quad (\star\in\{QK,V,M\}).$$

Thus, by Proposition D.5, we obtain the general form

$$\log\mathcal{N}_\infty(\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})},|\cdot|,\epsilon;\{X_i\}_{i\in[n]}) \lesssim \sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}\frac{\Lambda^{\star,(\ell)}}{(\epsilon^{\star,(\ell)})^{\nu^{\star,(\ell)}}}N\log(nT)+\left(\frac{C_2^{\rm out}}{\epsilon^{\rm out}}\right)^2\log(n), \tag{16}$$

where $\{\epsilon^{QK,(\ell)},\epsilon^{V,(\ell)},\epsilon^{M,(\ell)}\}_{\ell\in[L]}$ and $\epsilon^{\rm out}>0$ satisfy Eq. (15).

We now state two entropy bounds for $\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})}$. The first allows for different $p^{QK,(\ell)},p^{V,(\ell)},p^{M,(\ell)}$ to vary across layers, whereas the second assumes a common p for all layers and all weight matrices.

Theorem D.6. *Suppose* $\max_{i\in[n]}\|X_i\|_{2\to\infty}\le B_{n,(2\to\infty)}$ *holds. Fix any* $\boldsymbol{p}\in[0,2]^{3L}$ *and* $\boldsymbol{C}_{\rm s}\in(0,\infty)^{3L}$. *For each* $\ell\in[L]$, *define*

$$\gamma^{QK,(\ell)} = 2C_2^{V,(\ell)}C_2^{M,(\ell)}B_{n,(2\to\infty)}^{3\delta_{\ell=1}}, \qquad \gamma^{V,(\ell)} = C_2^{M,(\ell)}B_{n,(2\to\infty)}^{\delta_{\ell=1}}, \qquad \gamma^{M,(\ell)}=1.$$

Then, for any $\epsilon>0$, *it holds that*

$$\log\mathcal{N}_\infty(\mathcal{F}_{\rm out}^{(\boldsymbol{p},\boldsymbol{C}_{\rm s})},|\cdot|,\epsilon;\{X_i\}_{i\in[n]})$$
$$\lesssim \sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}(C_{\rm s}^{\star,(\ell)})^{\frac{2}{p^{\star,(\ell)}+2}}\left(\frac{L_\phi C_2^{\rm out}\gamma^{\star,(\ell)}\alpha^{(\ell)}L}{\epsilon}\right)^{\frac{2p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}N^{1+\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}\log(nT)$$
$$+\left(\frac{C_2^{\rm out}}{\epsilon}\right)^2\log(n).$$

Proof. For any $\epsilon > 0$, set $\epsilon^{\mathrm{out}} = \epsilon/2$. For each $\ell \in [L]$, define

$$\epsilon^{QK,(\ell)} = \frac{\epsilon}{12LL_\phi C_2^{\mathrm{out}} \alpha^{(\ell)} C_2^{V,(\ell)} C_2^{M,(\ell)} B_{n,(2\to\infty)}^{2\delta_{\ell=1}}},$$
$$\epsilon^{V,(\ell)} = \frac{\epsilon}{6LL_\phi C_2^{\mathrm{out}} \alpha^{(\ell)} C_2^{M,(\ell)}}, \qquad \epsilon^{M,(\ell)} = \frac{\epsilon}{6LC_2^{\mathrm{out}} \alpha^{(\ell)}}.$$

Note that for each $\ell \in [L]$, we have

$$\beta^{QK,(\ell)}\epsilon^{QK,(\ell)} = \beta^{V,(\ell)}\epsilon^{V,(\ell)} = \beta^{M,(\ell)}\epsilon^{M,(\ell)} = \frac{\epsilon}{6L},$$

which shows that the within-layer allocation is balanced in the sense prescribed above. This also implies that Eq. (15) is satisfied. We now substitute this admissible choice into Eq. (16). We obtain

$$\frac{\Lambda^{\star,(\ell)}}{(\epsilon^{\star,(\ell)})^{\nu^{\star,(\ell)}}} = (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{2}{p^{\star,(\ell)}+2}} \left(\frac{6L_\phi C_2^{\mathrm{out}} \gamma^{\star,(\ell)} \alpha^{(\ell)} L}{\epsilon}\right)^{\frac{2p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}.$$

Combining these bounds with Eq. (16) proves the claim. □

Theorem D.7. *Suppose* $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ *holds. For some* $p \in [0,2]$*, suppose* $\boldsymbol{p} = (p,\dots,p)$ *holds. Fix any* $\boldsymbol{C}_{\mathrm{s}} \in (0,\infty)^{3L}$*. For each* $\ell \in [L]$*, define* $\alpha^{(\ell)}$ *by Eq.* (14) *and* $\Gamma^{(\ell)}$ *by*

$$\Gamma^{(\ell)} = \sum_{\star\in\{QK,V,M\}} (\gamma^{\star,(\ell)})^{\frac{2p}{3p+2}} (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{2}{3p+2}},$$

where $\gamma^{\star,(\ell)}$ *is defined in Theorem D.6. Then, for any* $\epsilon > 0$*, we have*

$$\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}, |\cdot|, \epsilon; \{X_i\}_{i\in[n]})$$
$$\lesssim \left(\frac{L_\phi C_2^{\mathrm{out}}}{\epsilon}\right)^{\frac{2p}{p+2}} \left(\sum_{\ell=1}^{L} (\alpha^{(\ell)})^{\frac{2p}{3p+2}} \Gamma^{(\ell)}\right)^{\frac{3p+2}{p+2}} N^{1+\frac{p}{p+2}} \log(nT) + \left(\frac{C_2^{\mathrm{out}}}{\epsilon}\right)^2 \log(n).$$

Proof. When $p^{\star,(\ell)} = p$ holds, we can rewrite

$$\Lambda^{\star,(\ell)} = ((C_{\mathrm{s}}^{\star,(\ell)})^2 B_{n,(2\to\infty)}^{2\delta_{\ell=1}p} N^p)^{\frac{1}{p+2}} \quad (\star \in \{QK,V\}), \qquad \Lambda^{M,(\ell)} = ((C_{\mathrm{s}}^{M,(\ell)})^2 L_\phi^{2p} N^p)^{\frac{1}{p+2}},$$
$$\nu^{\star,(\ell)} = \frac{2p}{p+2} = \nu \quad (\star \in \{QK,V,M\}).$$

For a fixed $\epsilon^{\mathrm{out}} < \epsilon$, we first consider the following optimization problem:

$$\min_{\{\epsilon^{\star,(\ell)}\}_{\star\in\{QK,V,M\},\ell\in[L]}} \sum_{\ell=1}^{L} \sum_{\star\in\{QK,V,M\}} \Lambda^{\star,(\ell)} (\epsilon^{\star,(\ell)})^{-\nu}$$
$$\text{subject to} \quad \sum_{\ell=1}^{L} \sum_{\star\in\{QK,V,M\}} \beta^{\star,(\ell)}\epsilon^{\star,(\ell)} = \epsilon - \epsilon^{\mathrm{out}}.$$

If we take $p = 0$, we have $\nu = 0$. Thus, the first term on the right-hand side of Eq. (16) reduces to $\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}} \Lambda^{\star,(\ell)} N \log(nT)$, which is independent of the allocation of $\{\epsilon^{\star,(\ell)}\}_{\ell\in[L],\star\in\{QK,V,M\}}$. Thus, setting $\epsilon^{\mathrm{out}} = \epsilon/2$ and choosing any $\{\epsilon^{\star,(\ell)}\}_{\ell\in[L],\star\in\{QK,V,M\}}$ such that Eq. (15) is satisfied, we obtain the desired bounds.

Now, we consider the case $p > 0$. By Lemma B.8, the optimal solution is given by

$$\epsilon^{\star,(\ell)} = \frac{(\epsilon - \epsilon^{\mathrm{out}})(\Lambda^{\star,(\ell)})^{\frac{1}{\nu+1}} (\beta^{\star,(\ell)})^{-\frac{1}{\nu+1}}}{\sum_{k=1}^{L}\sum_{\star\in\{QK,V,M\}} (\Lambda^{\star,(k)})^{\frac{1}{\nu+1}} (\beta^{\star,(k)})^{\frac{\nu}{\nu+1}}} \qquad (\star \in \{QK,V,M\}, \ell \in [L])$$

and the corresponding minimum value is

$$\frac{1}{(\epsilon - \epsilon^{\mathrm{out}})^\nu} \left(\sum_{\ell=1}^{L} \sum_{\star\in\{QK,V,M\}} (\Lambda^{\star,(\ell)})^{\frac{1}{\nu+1}} (\beta^{\star,(\ell)})^{\frac{\nu}{\nu+1}}\right)^{\nu+1}.$$

Noting that we can write

$$
\begin{aligned}
(\Lambda^{QK,(\ell)})^{\frac{1}{\nu+1}}(\beta^{QK,(\ell)})^{\frac{\nu}{\nu+1}} &= (2L_\phi C_2^{\mathrm{out}} C_2^{V,(\ell)} C_2^{M,(\ell)} \alpha^{(\ell)})^{\frac{2p}{3p+2}} (C_{\mathrm{s}}^{QK,(\ell)})^{\frac{2}{3p+2}} (B_{n,(2\to\infty)}^{\delta_{\ell=1}})^{\frac{6p}{3p+2}} N^{\frac{p}{3p+2}}, \\
(\Lambda^{V,(\ell)})^{\frac{1}{\nu+1}}(\beta^{V,(\ell)})^{\frac{\nu}{\nu+1}} &= (L_\phi C_2^{\mathrm{out}} C_2^{M,(\ell)} \alpha^{(\ell)})^{\frac{2p}{3p+2}} (C_{\mathrm{s}}^{V,(\ell)})^{\frac{2}{3p+2}} (B_{n,(2\to\infty)}^{\delta_{\ell=1}})^{\frac{2p}{3p+2}} N^{\frac{p}{3p+2}}, \\
(\Lambda^{M,(\ell)})^{\frac{1}{\nu+1}}(\beta^{M,(\ell)})^{\frac{\nu}{\nu+1}} &= (L_\phi C_2^{\mathrm{out}} \alpha^{(\ell)})^{\frac{2p}{3p+2}} (C_{\mathrm{s}}^{M,(\ell)})^{\frac{2}{3p+2}} N^{\frac{p}{3p+2}},
\end{aligned}
$$

we have

$$
\begin{aligned}
&\frac{1}{(\epsilon-\epsilon^{\mathrm{out}})^\nu}\left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}(\Lambda^{\star,(\ell)})^{\frac{1}{\nu+1}}(\beta^{\star,(\ell)})^{\frac{\nu}{\nu+1}}\right)^{\nu+1} \\
&= \left(\frac{L_\phi C_2^{\mathrm{out}}}{\epsilon-\epsilon^{\mathrm{out}}}\right)^{\frac{2p}{p+2}} \left(\sum_{\ell=1}^{L}(\alpha^{(\ell)})^{\frac{2p}{3p+2}}\Gamma^{(\ell)}\right)^{\frac{3p+2}{p+2}} N^{\frac{p}{p+2}}.
\end{aligned}
$$

Substituting $\epsilon^{\mathrm{out}} = \epsilon/2$ and combining the resulting bound with Eq. (16) completes the proof. □

D.6 Generalization gap bounds for Transformers

Theorem D.8. *Suppose* $\max_{i\in[n]}\|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ *holds. Fix any* $\boldsymbol{p}\in[0,2]^{3L}$ *and* $\boldsymbol{C}_{\mathrm{s}}\in(0,\infty)^{3L}$. *For each* $\ell\in[L]$ *and* $\star\in\{QK,V,M\}$, *define* $\Psi^{\star,(\ell)}$ *by*

$$
\Psi^{\star,(\ell)} = (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{1}{p^{\star,(\ell)}+2}}(L_\phi\gamma^{\star,(\ell)}\alpha^{(\ell)})^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}},
$$

where $\alpha^{(\ell)}$ *is defined in Eq.* (14) *and* $\gamma^{\star,(\ell)}$ *is defined in Theorem D.6. Suppose* $n\ge 3$ *holds. Then, for any* $\delta\in(0,1)$, *with probability at least* $1-\delta$, *it holds simultaneously for all* $f_{\mathrm{out}}\in\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ *that*

$$
\begin{aligned}
\mathrm{GAP}(f_{\mathrm{out}}) &\lesssim L_{\mathcal{L}}C_2^{\mathrm{out}}\sqrt{\frac{\log(nT)}{n}}\left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}\Psi^{\star,(\ell)}L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}}\right) \\
&\quad + L_{\mathcal{L}}C_2^{\mathrm{out}}\frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + B_{\mathcal{L}}\sqrt{\frac{\log(1/\delta)}{n}}.
\end{aligned}
$$

Proof. For each $\ell\in[L]$ and $\star\in\{QK,V,M\}$, define $\nu^{\star,(\ell)} = 2p^{\star,(\ell)}/(p^{\star,(\ell)}+2)\in[0,2)$. By Theorem D.6, for every $\epsilon\in(0,C_2^{\mathrm{out}}]$, we have

$$
\log\mathcal{N}_\infty(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})},|\cdot|,\epsilon;\{X_i\}_{i\in[n]}) \lesssim \sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}c^{\star,(\ell)}\epsilon^{-\nu^{\star,(\ell)}} + c^{\mathrm{out}}\epsilon^{-2},
$$

where $c^{\star,(\ell)}$ and c^{out} are defined by

$$
\begin{aligned}
c^{\star,(\ell)} &= (C_{\mathrm{s}}^{\star,(\ell)})^{\frac{2}{p^{\star,(\ell)}+2}}(L_\phi C_2^{\mathrm{out}}\gamma^{\star,(\ell)}\alpha^{(\ell)}L)^{\nu^{\star,(\ell)}}N^{1+\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}\log(nT), \\
c^{\mathrm{out}} &= (C_2^{\mathrm{out}})^2\log(n).
\end{aligned}
$$

Next, note that every $f\in\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ satisfies $|f(X)|\le C_2^{\mathrm{out}}$ for all $X\in\mathbb{R}^{T\times N}$, because $\|[f_{\mathrm{tf}}^{(L)}(X)]_{[\mathrm{CLS}]}\|_2\le 1$ holds by construction. Therefore, by Lemma B.10, we have

$$
\begin{aligned}
&\hat{\mathfrak{R}}_n(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})};\{X_i\}_{i\in[n]}) \\
&\lesssim \frac{1}{\sqrt{n}}\left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}(C_2^{\mathrm{out}})^{1-\nu^{\star,(\ell)}/2}\sqrt{c^{\star,(\ell)}} + \sqrt{c^{\mathrm{out}}}\left[1+\log\left(1+\frac{C_2^{\mathrm{out}}\sqrt{n}}{\sqrt{c^{\mathrm{out}}}}\right)\right]\right) \\
&\lesssim C_2^{\mathrm{out}}\sqrt{\frac{\log(nT)}{n}}\left(\sum_{\ell=1}^{L}\sum_{\star\in\{QK,V,M\}}\Psi^{\star,(\ell)}L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}}N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}}\right) + C_2^{\mathrm{out}}\frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}},
\end{aligned}
$$

where the last inequality follows from $1+\log(1+\sqrt{n/\log(n)})\lesssim\log(n)$. Finally, applying Lemma B.11 completes the proof. □

Theorem D.9. *Suppose* $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ *holds. For some* $p \in [0,2]$*, suppose* $\boldsymbol{p} = (p,\dots,p)$ *holds. Fix any* $\boldsymbol{C}_{\mathrm{s}} \in (0,\infty)^{3L}$*. Define*

$$\Xi^{(p)} = \left(\sum_{\ell=1}^{L} (\alpha^{(\ell)})^{\frac{2p}{3p+2}} \Gamma^{(\ell)}\right)^{\frac{3p+2}{2(p+2)}} N^{\frac{p+1}{p+2}},$$

where $\alpha^{(\ell)}$ *is defined in Eq.* (14) *and* $\Gamma^{(\ell)}$ *is defined in Theorem D.7. Suppose* $n \ge 3$ *holds. Then, for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, it holds simultaneously for all* $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ *that*

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \frac{L_{\mathcal{L}} C_2^{\mathrm{out}}}{\sqrt{n}} [L_\phi^{\frac{p}{p+2}} \Xi^{(p)} \sqrt{\log(nT)} + (\log(n))^{\frac{3}{2}}] + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta)}{n}}.$$

Proof. Define $\nu = 2p/(p+2) \in [0,2)$. By Theorem D.7, it holds for every $\epsilon \in (0, C_2^{\mathrm{out}}]$ that

$$\log \mathcal{N}_\infty(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}, |\cdot|, \epsilon; \{X_i\}_{i\in[n]}) \lesssim c_1 \epsilon^{-\nu} + c_2 \epsilon^{-2},$$

where

$$c_1 = (L_\phi C_2^{\mathrm{out}})^\nu (\Xi^{(p)})^2 \log(nT), \qquad c_2 = (C_2^{\mathrm{out}})^2 \log(n).$$

Next, note that every $f \in \mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}$ satisfies $|f(X)| \le C_2^{\mathrm{out}}$ for all $X \in \mathbb{R}^{T\times N}$, because $\|[f_{\mathrm{tf}}^{(L)}(X)]_{[\mathrm{CLS}]}\|_2 \le 1$ holds by construction. Hence we may apply Lemma B.10 as

$$\begin{aligned}
&\hat{\mathfrak{R}}_n(\mathcal{F}_{\mathrm{out}}^{(\boldsymbol{p},\boldsymbol{C}_{\mathrm{s}})}; \{X_i\}_{i\in[n]}) \\
&\lesssim (C_2^{\mathrm{out}})^{1-\nu/2} \sqrt{\frac{c_1}{n}} + \frac{\sqrt{c_2}}{\sqrt{n}} \left[1 + \log\left(1 + \frac{C_2^{\mathrm{out}}\sqrt{n}}{\sqrt{c_2}}\right)\right] \\
&\lesssim \frac{C_2^{\mathrm{out}}}{\sqrt{n}} \left(L_\phi^{\frac{p}{p+2}} \Xi^{(p)} \sqrt{\log(nT)} + (\log(n))^{\frac{3}{2}}\right),
\end{aligned}$$

where the last inequality follows from $1 + \log(1 + \sqrt{n/\log(n)}) \lesssim \log(n)$. Finally, applying Lemma B.11 completes the proof. □

D.7 Obtaining the post hoc generalization bounds

In this section, we consider the scalar-output Transformer class $\mathcal{F}_{\mathrm{out}}$ in Eq. (6). In what follows, we derive a result in which the high-probability generalization gap bounds are uniform over the admissible choices of Schatten indices $p^{\star,(\ell)}$, so the indices may be selected after the trained weights have been observed. For that purpose, we partition the possible values of the realized Schatten quantity into dyadic shells. Define

$$\mathcal{J} = \mathbb{Z} \cup \{\perp\}, \qquad R_\perp = 0, \qquad R_j = 2^j \quad (j \in \mathbb{Z}).$$

Here, the symbol $\perp$ is an auxiliary index that does not belong to $\mathbb{Z}$, and it is used to represent the zero-radius shell. Thus, $j = \perp$ corresponds to $\|W\|_{\mathrm{s},p}^p = 0$, whereas $j \in \mathbb{Z}$ corresponds to the positive dyadic shell $2^{j-1} < \|W\|_{\mathrm{s},p}^p \le 2^j$. Set

$$Z_{\mathcal{J}} = 1 + \sum_{j\in\mathbb{Z}} \frac{1}{(1+|j|)^2} = \frac{\pi^2}{3}, \qquad \omega_\perp = Z_{\mathcal{J}}^{-1}, \qquad \omega_j = \frac{Z_{\mathcal{J}}^{-1}}{(1+|j|)^2} \quad (j \in \mathbb{Z})$$

so that $\sum_{j\in\mathcal{J}} \omega_j = 1$ holds. For $p \in [0,2]$ and $W \in \mathbb{R}^{N\times N}$, define

$$\kappa_p(W) = \begin{cases} \perp, & \|W\|_{\mathrm{s},p}^p = 0, \\ \lceil \log_2 \|W\|_{\mathrm{s},p}^p \rceil, & \|W\|_{\mathrm{s},p}^p > 0. \end{cases}$$

Thus, if $\|W\|_{\mathrm{s},p}^p$ is nonzero, it belongs to the $\kappa_p(W)$-th dyadic shell as

$$2^{\kappa_p(W)-1} < \|W\|_{\mathrm{s},p}^p \le 2^{\kappa_p(W)}.$$

For an integer $m \geq 1$, define the grid

$$\mathcal{P}_m = \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{2m-1}{m}, 2\right\}^{3L}.$$

For $\boldsymbol{p} \in \mathcal{P}_m$, define the post hoc logarithmic penalty

$$\Omega_{\boldsymbol{p}}(W^{(1:L)}) = 3L\log(2m+1) + \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} \log(1/\omega_{\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)})}). \tag{17}$$

For $\boldsymbol{p} \in [0,2]^{3L}$ and weights $W^{(1:L)}$, define

$$\mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) = \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} (\|W^{\star,(\ell)}\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}})^{\frac{1}{p^{\star,(\ell)}+2}} (L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)} L)^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}}. \tag{18}$$

Here $\alpha^{(\ell)}$ is defined in Eq. (14), and $\gamma^{\star,(\ell)}$ is the quantity defined in Theorem D.6. When $p^{\star,(\ell)} = 0$, we interpret $\|W^{\star,(\ell)}\|_{\mathrm{s},0}^0$ as $\mathrm{rank}(W^{\star,(\ell)})$.

Lemma D.10 (Post hoc selection of the Schatten indices)**.** *Suppose* $\max_{i\in[n]} \|X_i\|_{2\to\infty} \leq B_{n,(2\to\infty)}$ *and* $n \geq 3$ *hold. Fix an integer* $m \geq 1$*. Then, for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, every* $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}$ *satisfies*

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \inf_{\boldsymbol{p}\in\mathcal{P}_m} \left(L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta) + \Omega_{\boldsymbol{p}}(W^{(1:L)})}{n}} \right) + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}}.$$

Proof. Write the index sets as $\mathcal{I} = \{QK, V, M\} \times [L]$. For $a = (\star, \ell) \in \mathcal{I}$, we write $W^a = W^{\star,(\ell)}$ and $p^a = p^{\star,(\ell)}$. For $\boldsymbol{p} \in \mathcal{P}_m$ and $\boldsymbol{j} = (j_a)_{a\in\mathcal{I}} \in \mathcal{J}^{3L}$, define

$$\mathcal{F}_{\boldsymbol{p},\boldsymbol{j}} = \{f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}} \mid \|W^a\|_{\mathrm{s},p^a}^{p^a} \leq R_{j_a}\ (a \in \mathcal{I})\}.$$

When $j_a = \perp$, this condition is interpreted as $\|W^a\|_{\mathrm{s},p^a}^{p^a} = 0$. For each $a = (\star, \ell) \in \mathcal{I}$, define

$$M_a = N(1 \vee (C_2^{\star,(\ell)})^2), \qquad \bar{R}_{j_a}^a = \begin{cases} 0, & j_a = \perp, \\ R_{j_a} \wedge M_a, & j_a \in \mathbb{Z}. \end{cases}$$

Since every element of $\mathcal{F}_{\mathrm{out}}$ satisfies $\|W^a\|_2 \leq C_2^{\star,(\ell)}$, we have $\|W^a\|_{\mathrm{s},p^a}^{p^a} \leq M_a$ for every $p^a \in [0,2]$. Therefore, under the spectral norm constraints already included in $\mathcal{F}_{\mathrm{out}}$, the class $\mathcal{F}_{\boldsymbol{p},\boldsymbol{j}}$ is equivalently described by the constraints

$$\|W^a\|_{\mathrm{s},p^a}^{p^a} \leq \bar{R}_{j_a}^a \qquad (a \in \mathcal{I}).$$

The localized fixed-radius version of Theorem D.8, applied to $\mathcal{F}_{\boldsymbol{p},\boldsymbol{j}}$ with Schatten-quantity radii $\bar{R}_{j_a}^a$, gives an event $E_{\boldsymbol{p},\boldsymbol{j}}$ such that

$$\mathrm{P}(E_{\boldsymbol{p},\boldsymbol{j}}) \geq 1 - \delta_{\boldsymbol{p},\boldsymbol{j}}, \qquad \delta_{\boldsymbol{p},\boldsymbol{j}} = \frac{\delta}{(2m+1)^{3L}} \prod_{a\in\mathcal{I}} \omega_{j_a},$$

and, on $E_{\boldsymbol{p},\boldsymbol{j}}$, the following bounds hold simultaneously for all $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{j}}$:

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \sum_{a=(\star,\ell)\in\mathcal{I}} (\bar{R}_{j_a}^a)^{\frac{1}{p^a+2}} (L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)} L)^{\frac{p^a}{p^a+2}} N^{\frac{p^a+1}{p^a+2}} + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta_{\boldsymbol{p},\boldsymbol{j}})}{n}}.$$

The logarithmic factors suppressed in this application are uniform over $\boldsymbol{p} \in \mathcal{P}_m$ and $\boldsymbol{j} \in \mathcal{J}^{3L}$, because the fixed-radius entropy bounds are used only with the truncated radii $\bar{R}^a_{j_a} \le M_a$. Thus any radius-dependent logarithmic factor is absorbed into the notation $\lesssim$. If $\bar{R}^a_{j_a} = 0$ for some a, the same conclusion follows by fixing the corresponding weight matrix to zero and applying the same proof with that coordinate omitted. Since $\bar{R}^a_{j_a} \le R_{j_a}$ holds, the preceding display implies the coarser bound

$$\begin{aligned}\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \sum_{a=(\star,\ell)\in\mathcal{I}} R_{j_a}^{\frac{1}{p^a+2}} (L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)} L)^{\frac{p^a}{p^a+2}} N^{\frac{p^a+1}{p^a+2}} \\ + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta_{\boldsymbol{p},\boldsymbol{j}})}{n}}. \end{aligned} \tag{19}$$

By the union bound, the event $E = \bigcap_{\boldsymbol{p}\in\mathcal{P}_m} \bigcap_{\boldsymbol{j}\in\mathcal{J}^{3L}} E_{\boldsymbol{p},\boldsymbol{j}}$ satisfies

$$\begin{aligned}&\mathrm{P}(E) = 1 - \mathrm{P}(E^C) = 1 - \mathrm{P}\left(\bigcup_{\boldsymbol{p}\in\mathcal{P}_m} \bigcup_{\boldsymbol{j}\in\mathcal{J}^{3L}} E^C_{\boldsymbol{p},\boldsymbol{j}} \right) \ge 1 - \sum_{\boldsymbol{p}\in\mathcal{P}_m} \sum_{\boldsymbol{j}\in\mathcal{J}^{3L}} \mathrm{P}(E^C_{\boldsymbol{p},\boldsymbol{j}}) \\ &\ge 1 - \sum_{\boldsymbol{p}\in\mathcal{P}_m} \sum_{\boldsymbol{j}\in\mathcal{J}^{3L}} \delta_{\boldsymbol{p},\boldsymbol{j}} = 1 - \delta,\end{aligned}$$

where the last equality follows from

$$\sum_{\boldsymbol{p}\in\mathcal{P}_m} \sum_{\boldsymbol{j}\in\mathcal{J}^{3L}} \delta_{\boldsymbol{p},\boldsymbol{j}} = \delta \sum_{\boldsymbol{j}\in\mathcal{J}^{3L}} \prod_{a\in\mathcal{I}} \omega_{j_a} = \delta \left(\sum_{j\in\mathcal{J}} \omega_j \right)^{3L} = \delta.$$

We now work on the event E. Fix an arbitrary $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}$ and an arbitrary $\boldsymbol{p} \in \mathcal{P}_m$. For each $a = (\star,\ell) \in \mathcal{I}$, set $j_a = \kappa_{p^a}(W^a)$. Then we have $f_{\mathrm{out}} \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{j}}$, and Eq. (19) is applicable. Moreover, by the definition of κ_{p^a}, we have

$$R_{j_a} \le 2 \|W^a\|^{p^a}_{\mathrm{s},p^a}$$

whenever $\|W^a\|^{p^a}_{\mathrm{s},p^a} > 0$, while both sides are zero when $\|W^a\|^{p^a}_{\mathrm{s},p^a} = 0$ and $j_a = \perp$. Since $p^a \in [0,2]$ holds, this implies

$$R_{j_a}^{\frac{1}{p^a+2}} \le 2^{\frac{1}{p^a+2}} (\|W^a\|^{p^a}_{\mathrm{s},p^a})^{\frac{1}{p^a+2}} \le \sqrt{2} (\|W^a\|^{p^a}_{\mathrm{s},p^a})^{\frac{1}{p^a+2}}.$$

The factor $\sqrt{2}$ is absorbed into the universal constant. We also have

$$\log(1/\delta_{\boldsymbol{p},\boldsymbol{j}}) = \log(1/\delta) + 3L \log(2m+1) + \sum_{a\in\mathcal{I}} \log\left(\frac{1}{\omega_{j_a}}\right) = \log(1/\delta) + \Omega_{\boldsymbol{p}}(W^{(1:L)}).$$

Substituting these two estimates into Eq. (19) gives

$$\begin{aligned}\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}} \\ + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta) + \Omega_{\boldsymbol{p}}(W^{(1:L)})}{n}}.\end{aligned}$$

This holds for every $\boldsymbol{p} \in \mathcal{P}_m$ on the same event E. Taking the infimum over $\boldsymbol{p} \in \mathcal{P}_m$ proves the claim. □

Theorem D.11 (Post hoc generalization bounds)**.** *Suppose* $\max_{i\in[n]} \|X_i\|_{2\to\infty} \le B_{n,(2\to\infty)}$ *and* $n \ge 3$ *hold. Suppose further that* $\gamma^{\star,(\ell)}\alpha^{(\ell)} > 0$ *holds for all* $\star \in \{QK, V, M\}$ *and* $\ell \in [L]$*, and* $L_\phi > 0$ *holds. For* $\boldsymbol{p} \in [0,2]^{3L}$ *and weights* $W^{(1:L)}$*, define* $\Omega_{\boldsymbol{p}}(W^{(1:L)})$ *and* $\mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)})$ *by Eq.* (17) *and Eq.* (18)*, respectively. Fix an integer* $m \ge 1$*, and define the coordinatewise upward*

grid projection $\Pi_m : [0,2]^{3L} \to \mathcal{P}_m$ *as follows. For* $t \in [0,2]$*, define* $\pi_m(t)$ *by* $\pi_m(t) = \lceil mt \rceil / m$ *and set*

$$\Pi_m(\boldsymbol{p}) = (\pi_m(p^{\star,(\ell)}))_{\star \in \{QK,V,M\}, \ell \in [L]}.$$

For weights $W^{(1:L)}$*, define*

$$\chi(W^{(1:L)}) = \max_{\ell \in [L], \star \in \{QK,V,M\} : W^{\star,(\ell)} \neq 0} (|\log \|W^{\star,(\ell)}\|_2| + |\log(L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)})|),$$

with the convention that the maximum over an empty set is zero. Then, for any $\delta \in (0,1)$*, with probability at least* $1-\delta$*, every* $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}$ *satisfies*

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \inf_{\boldsymbol{p} \in [0,2]^{3L}} \left[L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \exp\left(\frac{\chi(W^{(1:L)})}{m} \right) L^{\frac{1}{2m}} N^{\frac{1}{4m}} \mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) \right.$$
$$\left. + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta) + \Omega_{\Pi_m(\boldsymbol{p})}(W^{(1:L)})}{n}} \right] + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}}.$$

Proof. We first prove a deterministic rounding inequality. Fix $0 \le p \le q \le 2$ and a nonzero matrix $W \in \mathbb{R}^{N \times N}$. For $H > 0$, define

$$F_p(W; H) = (\|W\|_{\mathrm{s},p}^p)^{\frac{1}{p+2}} H^{\frac{p}{p+2}} N^{\frac{p+1}{p+2}}.$$

Set $\rho_p(W) = \|W\|_{\mathrm{s},p}^p / \|W\|_2^p$. Since the singular values normalized by $\|W\|_2$ belong to $[0,1]$ and the largest normalized singular value is one, we have

$$1 \le \rho_q(W) \le \rho_p(W) \le N. \tag{20}$$

Hence, it holds that

$$\begin{aligned} \log \frac{F_q(W;H)}{F_p(W;H)} &= \frac{q \log \|W\|_2 + \log \rho_q(W)}{q+2} - \frac{p \log \|W\|_2 + \log \rho_p(W)}{p+2} \\ &\quad + \left(\frac{q}{q+2} - \frac{p}{p+2} \right) \log H + \left(\frac{q+1}{q+2} - \frac{p+1}{p+2} \right) \log N \\ &\le (q-p) \left(\frac{|\log \|W\|_2|}{2} + \frac{|\log H|}{2} + \frac{\log N}{4} \right). \end{aligned}$$

Therefore, we have

$$F_q(W;H) \le \exp\left((q-p) \left(\frac{|\log \|W\|_2|}{2} + \frac{|\log H|}{2} + \frac{\log N}{4} \right) \right) F_p(W;H).$$

The same inequality is trivial when $W = 0$, because both sides are zero.

Now fix $\boldsymbol{p} \in [0,2]^{3L}$ and set $\tilde{\boldsymbol{p}} = \Pi_m(\boldsymbol{p})$. For every coordinate $(\star, \ell)$, we have

$$\tilde{p}^{\star,(\ell)} - p^{\star,(\ell)} = \frac{\lceil m p^{\star,(\ell)} \rceil}{m} - p^{\star,(\ell)} \in \left[0, \frac{1}{m} \right].$$

Applying the preceding deterministic inequality with $H = L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)} L$ to each summand in $\mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)})$ gives

$$\mathfrak{B}_{\Pi_m(\boldsymbol{p})}(W^{(1:L)}) \le \exp\left(\frac{\chi(W^{(1:L)})}{2m} \right) L^{\frac{1}{2m}} N^{\frac{1}{4m}} \mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}). \tag{21}$$

We next apply Lemma D.10. On an event of probability at least $1-\delta$, every $f_{\mathrm{out}}(\,\cdot\,; W^{(1:L)}, w) \in \mathcal{F}_{\mathrm{out}}$ satisfies

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim \inf_{\boldsymbol{q} \in \mathcal{P}_m} \left(L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \mathfrak{B}_{\boldsymbol{q}}(W^{(1:L)}) + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta) + \Omega_{\boldsymbol{q}}(W^{(1:L)})}{n}} \right)$$
$$+ L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}}.$$

Since $\Pi_m(\boldsymbol{p}) \in \mathcal{P}_m$ for every $\boldsymbol{p} \in [0,2]^{3L}$, we may upper bound the finite-grid infimum by evaluating it at $\boldsymbol{q} = \Pi_m(\boldsymbol{p})$. Using Eq. (21), we obtain, for every $\boldsymbol{p} \in [0,2]^{3L}$,

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} \sqrt{\frac{\log(nT)}{n}} \exp\left(\frac{\chi(W^{(1:L)})}{m}\right) L^{\frac{1}{2m}} N^{\frac{1}{4m}} \mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta) + \Omega_{\Pi_m(\boldsymbol{p})}(W^{(1:L)})}{n}} + L_{\mathcal{L}} C_2^{\mathrm{out}} \frac{(\log(n))^{\frac{3}{2}}}{\sqrt{n}}.$$

Taking the infimum over $\boldsymbol{p} \in [0,2]^{3L}$ proves the claim. □

Remark D.1 (Derivation of Theorem 3.1)**.** We derive Theorem 3.1 from Theorem D.11. Choose $m = \lceil L + \log(N) \rceil$. Since $\alpha^{(\ell)}$ is a product of at most $L - \ell$ fixed spectral norm and Lipschitz factors, there exists a constant $C > 0$ such that $L_\phi \gamma^{\star,(\ell)} \alpha^{(\ell)} \le C^L$ holds for all $\star$ and ℓ. Hence, we have

$$\mathfrak{B}_{\boldsymbol{p}}(W^{(1:L)}) \lesssim \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} (\|W^{\star,(\ell)}\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}})^{\frac{1}{p^{\star,(\ell)}+2}} C^{\frac{Lp^{\star,(\ell)}}{p^{\star,(\ell)}+2}} L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}}.$$

We now simplify the two additional terms, $\exp(\chi(W^{(1:L)})/m) L^{\frac{1}{2m}} N^{\frac{1}{4m}}$ and $\Omega_{\Pi_m(\boldsymbol{p})}(W^{(1:L)})$, introduced by the post hoc argument. First, we control the rounding factor $\exp(\chi(W^{(1:L)})/m) L^{\frac{1}{2m}} N^{\frac{1}{4m}}$. Suppose that the trained weights are not exponentially small, in the sense that there exists a universal constant $c_0 > 0$ that satisfies $\|W^{\star,(\ell)}\|_2 \ge \exp[-c_0(L + \log(N))]$ for every nonzero $W^{\star,(\ell)}$. Under the simplified spectral norm assumptions, the quantity $\chi(W^{(1:L)})$ appearing in Theorem D.11 satisfies $\chi(W^{(1:L)}) \lesssim L + \log(N)$. Consequently, our choice of m gives

$$\exp\left(\frac{\chi(W^{(1:L)})}{m}\right) L^{\frac{1}{2m}} N^{\frac{1}{4m}} \lesssim L^{\frac{1}{2m}} N^{\frac{1}{4m}} = \exp\left(\frac{\log(L)}{2m} + \frac{\log(N)}{4m}\right) \le \exp\left(\frac{1}{2} + \frac{1}{4}\right) \lesssim 1.$$

It remains to simplify the post hoc logarithmic penalty $\Omega_{\Pi_m(\boldsymbol{p})}(W^{(1:L)})$. Note that for $j \in \mathbb{Z}$, we have $\log(1/\omega_j) = \log(Z_{\mathcal{J}}) + 2\log(1 + |j|)$. For a nonzero matrix W and $p \in [0,2]$, we have $\|W\|_2^p \le \|W\|_{\mathrm{s},p}^p \le N \|W\|_2^p$ (see Eq. (20)). Together with the upper spectral norm constraint and the lower bound on nonzero spectral norms, this implies

$$|\kappa_p(W)| = |\lceil \log_2 \|W\|_{\mathrm{s},p}^p \rceil| \lesssim L + \log(N).$$

Thus, if $W^{\star,(\ell)} \neq 0$ and $\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)}) \neq 0$ hold, we have

$$\log(1/\omega_{\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)})}) \lesssim \log(|\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)})|) \lesssim \log(L + \log(N)).$$

On the other hand, if either $\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)}) = 0$ or $W^{\star,(\ell)} = 0$ holds, we have by definition that $\log(1/\omega_{\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)})}) = \log(Z_{\mathcal{J}})$. Hence, uniformly over $p^{\star,(\ell)} \in [0,2]$ and $W^{\star,(\ell)}$, we have

$$\log(1/\omega_{\kappa_{p^{\star,(\ell)}}(W^{\star,(\ell)})}) \lesssim \log(L + \log(N)).$$

Therefore, we obtain

$$\Omega_{\Pi_m(\boldsymbol{p})}(W^{(1:L)}) \lesssim L \log(L + \log(N))$$

uniformly over $\boldsymbol{p} \in [0,2]^{3L}$.

E Choice of $p^{\star,(\ell)}$ under polynomial singular value decay

In this section, we illustrate how the post hoc parameter $p^{\star,(\ell)}$ in Theorem 3.1 can be selected from the observed singular-value profile. For each $\star \in \{QK, V, M\}$, $\ell \in [L]$ and $p^{\star,(\ell)} \in [0,2]$, define

$$\tilde{\mathfrak{B}}^{\star,(\ell)}(p^{\star,(\ell)}) = (\|W^{\star,(\ell)}\|_{\mathrm{s},p^{\star,(\ell)}}^{p^{\star,(\ell)}})^{\frac{1}{p^{\star,(\ell)}+2}} C^{\frac{Lp^{\star,(\ell)}}{p^{\star,(\ell)}+2}} L^{\frac{p^{\star,(\ell)}}{p^{\star,(\ell)}+2}} N^{\frac{p^{\star,(\ell)}+1}{p^{\star,(\ell)}+2}}.$$

Then, the first term on the right-hand side of Eq. (7) is expressed as

$$\sqrt{\frac{\log(nT)}{n}} \sum_{\ell=1}^{L} \sum_{\star \in \{QK,V,M\}} \inf_{p^{\star,(\ell)} \in [0,2]} \tilde{\mathfrak{B}}^{\star,(\ell)}(p^{\star,(\ell)}).$$

Thus, it suffices to choose $p^{\star,(\ell)} \in [0,2]$ to minimize $\tilde{\mathfrak{B}}^{\star,(\ell)}(p^{\star,(\ell)})$.

Fix $\star \in \{QK, V, M\}$ and $\ell \in [L]$. Suppose that the singular values of $W^{\star,(\ell)}$ satisfy the polynomial decay condition

$$\sigma_i(W^{\star,(\ell)}) \asymp S^{\star,(\ell)} i^{-\tau^{\star,(\ell)}} \qquad (i \in [N]),$$

with dimension-independent $S^{\star,(\ell)}$ and $\tau^{\star,(\ell)} \geq 1/2$. For notational simplicity, we write $p = p^{\star,(\ell)}$, $\tau = \tau^{\star,(\ell)}$, $S = S^{\star,(\ell)}$, and $A = C^L L$. By the singular value decay assumption, we have

$$\|W^{\star,(\ell)}\|_{\mathrm{s},p}^p = \begin{cases} \sum_{i=1}^N \sigma_i(W^{\star,(\ell)})^p \lesssim S^p \sum_{i=1}^N i^{-\tau p}, & \text{if } p > 0, \\ \mathrm{rank}(W^{\star,(\ell)}) \leq N, & \text{if } p = 0. \end{cases}$$

Next, note that for any $q > 0$ and $N \geq 2$, we have

$$\sum_{i=1}^N i^{-q} \leq \begin{cases} N^{1-q}/(1-q), & \text{if } q < 1, \\ \log(N) + 1 \lesssim \log(N), & \text{if } q = 1, \\ q/(q-1), & \text{if } q > 1. \end{cases}$$

Therefore, we have

$$\tilde{\mathfrak{B}}^{\star,(\ell)}(p) \lesssim \begin{cases} N & \text{if } p = 0, \\ (1-\tau p)^{-\frac{1}{p+2}} A^{\frac{p}{p+2}} N^{1-\frac{\tau p}{p+2}}, & \text{if } 0 < p < 1/\tau, \\ A^{\frac{p}{p+2}} N^{\frac{p+1}{p+2}} (\log(N))^{\frac{1}{p+2}}, & \text{if } p = 1/\tau, \\ (\tau p/(\tau p - 1))^{\frac{1}{p+2}} A^{\frac{p}{p+2}} N^{\frac{p+1}{p+2}}, & \text{if } p > 1/\tau. \end{cases} . \tag{22}$$

To obtain an explicit polynomial-order description, suppose that $A \asymp N^\kappa$ holds for some $\kappa \geq 0$. For example, this includes the case $L = O(1)$ with $\kappa = 0$, or the case $C = 1$, $L \asymp N^\kappa$. For simplicity, we assume $N \gg 1$, and we only optimize in the polynomial order in N and suppress constants and logarithmic factors.

If $0 < p < 1/\tau$, the leading polynomial order in N of the right-hand side of Eq. (22) is determined by $1 + (\kappa - \tau)p/(p+2)$. This value decreases in p if $\kappa < \tau$ and increases in p if $\kappa > \tau$. On the other hand, if $p > 1/\tau$, the leading polynomial order in N is determined by $(p + 1 + \kappa p)/(p+2)$, which increases in p. Therefore, if $\kappa < \tau$, the choice $p = 1/\tau$ is optimal and we have

$$\tilde{\mathfrak{B}}^{\star,(\ell)}\left(\frac{1}{\tau}\right) \lesssim A^{\frac{1}{2\tau+1}} N^{\frac{\tau+1}{2\tau+1}} (\log(N))^{\frac{\tau}{2\tau+1}},$$

whereas if $\kappa \geq \tau$, the optimal choice is $p = 0$ and we have $\tilde{\mathfrak{B}}^{\star,(\ell)}(0) \lesssim N$.

This calculation gives the following interpretation. When the singular values decay rapidly, the optimal admissible Schatten index is small, and the resulting bounds behave closer to rank-based bounds. In contrast, when the singular values decay slowly, the optimal index moves toward larger values, and the bounds become closer to norm-based bounds. Therefore, Theorem 3.1 can exploit heterogeneous spectral behavior across layers and matrix types by selecting $p^{\star,(\ell)}$ separately after training.

F Empirical comparison and discussion

In this section, we first discuss a practical limitation of existing norm-based bounds: their leading factors involve fixed norm radii whose dimension-independent scaling is difficult to justify for large learned weight matrices. Then, we compare the growth of the leading-factor proxies induced by our bounds and by existing norm-based bounds. For illustration, we use the BERT Miniatures checkpoints of Turc et al. (2019). Since these checkpoints use multi-head attention and two-layer feedforward sublayers, we construct BERT-adapted proxies from the leading polynomial and spectral factors of the theoretical generalization gap bounds. The code used for the empirical analyses in Section 6 and this appendix is available at `https://github.com/manasakai/spectrum-adaptive-transformers`.

F.1 Existing norm-based bounds

We first recall the two existing norm-based bounds. The bounds of Edelman et al. (2022) assume mixed $(2,1)$-norm constraints and obtain a logarithmic dependence on the token length and hidden dimension. The bounds of Trauger and Tewari (2024) replace these assumptions with mixed $(1,1)$-norm constraints and remove the explicit dependence on the token length.

Theorem F.1 (Edelman et al. (2022), Theorem A.17)**.** *Consider the parameter classes*

$$\begin{aligned}\mathcal{W}^{QK,(\ell)} &= \{W \in \mathbb{R}^{N\times N} \mid \|W^\top\|_{2,1} \le C_{2,1}^{QK,(\ell)}, \|W\|_2 \le C_2^{QK,(\ell)}\},\\ \mathcal{W}^{V,(\ell)} &= \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{2,1} \le C_{2,1}^{V,(\ell)}, \|W\|_2 \le C_2^{V,(\ell)}\},\\ \mathcal{W}^{M,(\ell)} &= \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{2,1} \le C_{2,1}^{M,(\ell)}, \|W\|_2 \le C_2^{M,(\ell)}\}.\end{aligned}$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds simultaneously for all $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}$ that

$$\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} \left(1 + \sum_{\ell=1}^{L} (\alpha^{(\ell)})^{\frac{2}{3}} \xi^{(\ell)}\right)^{\frac{3}{2}} \sqrt{\frac{\log(NnT)}{n}} + B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta)}{n}}$$

with $\xi^{(\ell)} = (C_{2,1}^{M,(\ell)})^{\frac{2}{3}} + (2L_\phi C_2^{M,(\ell)} C_2^{V,(\ell)} C_{2,1}^{QK,(\ell)})^{\frac{2}{3}} + (L_\phi C_2^{M,(\ell)} C_{2,1}^{V,(\ell)})^{\frac{2}{3}}$.

Theorem F.2 (Trauger and Tewari (2024), Corollary 4.2.1)**.** *Consider parameter classes of the form*

$$\mathcal{W}^{\star,(\ell)} = \{W \in \mathbb{R}^{N\times N} \mid \|W\|_{1,1} \le C_{1,1}, \|W\|_2 \le C_2^\star\} \quad (\star \in \{QK, V, M\}, \ell \in [L]).$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds simultaneously for all $f_{\mathrm{out}} \in \mathcal{F}_{\mathrm{out}}$ that

$$\begin{aligned}\mathrm{GAP}(f_{\mathrm{out}}) \lesssim L_{\mathcal{L}} C_2^{\mathrm{out}} C_{1,1} &\left(1 + (L_\phi C_2^V)^{\frac{2}{3}} + \sum_{\ell=1}^{L} (\alpha^{(\ell)})^{\frac{2}{3}} \upsilon^{(\ell)}\right)^{\frac{3}{2}} \sqrt{\frac{\log(2N^2+1)}{n}}\\ &+ B_{\mathcal{L}} \sqrt{\frac{\log(1/\delta)}{n}},\end{aligned}$$

where

$$\upsilon^{(\ell)} = \begin{cases} (2L_\phi C_2^M C_2^V B_{n,(2\to\infty)})^{\frac{2}{3}} & \text{if } \ell = 1\\ 1 + (2L_\phi C_2^M C_2^V)^{\frac{2}{3}} + (L_\phi C_2^V)^{\frac{2}{3}} & \text{if } \ell > 1.\end{cases}$$

The apparent dimension-favorable behavior of these bounds is tied to their norm assumptions. Recall that for every $W \in \mathbb{R}^{N\times N}$, we have

$$\|W\|_{2,1} \le \sqrt{N \operatorname{rank}(W)} \|W\|_2, \qquad \|W\|_{1,1} \le N\sqrt{\operatorname{rank}(W)} \|W\|_2.$$

Thus, treating $C_{2,1}$ or $C_{1,1}$ as independent of N is substantially stronger than imposing only spectral norm constraints. This issue is visible in trained BERT weights: the mixed norms increase with the hidden dimension N, as shown in Figure 2.

We focus on the bounds of Edelman et al. (2022) in the numerical comparison. The present experiment suppresses constants and logarithmic factors and compares only the leading polynomial growth of BERT-adapted proxies. The main advantage of Trauger and Tewari (2024) is the removal of explicit token-length dependence, but this advantage is not visible in this polynomial-only comparison. Moreover, their bounds are based on mixed $(1,1)$-norm radii, which can grow more severely with the hidden dimension than the mixed $(2,1)$-norm radii used in Edelman et al. (2022). Thus, Edelman-type proxies provide a direct baseline for assessing the polynomial scaling behavior of our post hoc bounds. This does not diminish the importance of Trauger and Tewari (2024) in settings where the token-length dependence is the central quantity of interest.

F.2 BERT checkpoints and construction of proxies

We used the publicly released BERT Miniatures checkpoints of Turc et al. (2019), hosted on Hugging Face under the Google organization. The checkpoints are identified by the model names `bert_uncased_L-{2,4,6,8,10,12}_H-{128,256,512,768}_A-{2,4,8,12}`. We accessed the

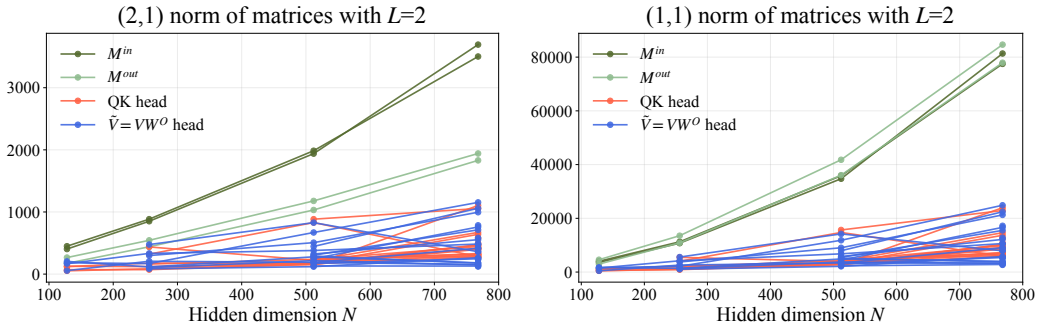

Figure 2: Scaling of the mixed $(2,1)$- and $(1,1)$-norms of trained BERT weights at depth $L = 2$. The observed growth with the hidden dimension N suggests that treating the corresponding mixed-norm radii as dimension-independent is not well supported by these checkpoints.

checkpoints through their public Hugging Face URLs, and did not modify or redistribute the checkpoint files. In this model family, the number of self-attention heads is $A_h = N/d_h$ with $d_h = 64$, and the intermediate dimension of the feedforward sublayer is $I = 4N$. The grid is

$$L \in \{2, 4, 6, 8, 10, 12\}, \qquad N \in \{128, 256, 512, 768\}.$$

There are several architectural differences between these checkpoints and the theoretical model in the main text. Since the theory is formulated in terms of single-head weight matrices, the BERT-adapted proxies below modify only the corresponding weight-matrix components. Specifically, we adapt the theoretical query-key, value, and feedforward matrices to four BERT-specific features: multi-head attention, headwise query-key products, the value-output projection structure, and the two-layer feedforward sublayer.

First, BERT uses multi-head attention. We therefore treat each head $h \in [A_h]$ as a local single-head component. Second, the query-key matrix in our theory is represented by the headwise product

$$W^{QK,(\ell,h)} = (W^{Q,(\ell,h)})^\top W^{K,(\ell,h)} \in \mathbb{R}^{N\times N}.$$

Since the query and key matrices in these checkpoints have dimensions $W^{Q,(\ell,h)}, W^{K,(\ell,h)} \in \mathbb{R}^{d_h\times N}$, the matrix $W^{QK,(\ell,h)}$ has rank at most $d_h = 64$. Third, our single-head theoretical model contains a single value matrix $W^{V,(\ell)} \in \mathbb{R}^{N\times N}$ and does not include a separate output projection for each attention head. In BERT, for head h, we write the value projection and the corresponding slice of the output projection in row-vector orientation as $W^{V,(\ell,h)} \in \mathbb{R}^{N\times d_h}$ and $W^{O,(\ell,h)} \in \mathbb{R}^{d_h\times N}$. We therefore use the composed headwise value-output matrix $W^{\tilde{V},(\ell,h)} := W^{V,(\ell,h)}W^{O,(\ell,h)} \in \mathbb{R}^{N\times N}$, whose rank is at most $d_h = 64$. Fourth, the feedforward sublayer is a two-layer map

$$X \mapsto \phi_{\mathrm{GELU}}(XW^{M,(\ell,\mathrm{in})} + b^{M,(\ell,\mathrm{in})})W^{M,(\ell,\mathrm{out})} + b^{M,(\ell,\mathrm{out})},$$

so we use $W^{M,(\ell,\mathrm{in})} \in \mathbb{R}^{N\times I}$ and $W^{M,(\ell,\mathrm{out})} \in \mathbb{R}^{I\times N}$ separately in the proxy. We also use $L_{\mathrm{GELU}} \le 1.13$ for the Lipschitz constant of the elementwise GELU map.

These definitions should not be interpreted as a formal extension of our generalization bounds to the full BERT encoder architecture. They do not incorporate residual connections, LayerNorm, bias parameters, or token and positional embeddings. These components may affect the effective Lipschitz constants, norm propagation, and parameter complexity of literal BERT generalization bounds, and incorporating them would require a separate covering analysis. Thus, the quantities defined in this appendix should be understood as BERT-adapted leading-factor diagnostics, whose purpose is to compare how the leading spectral and mixed-norm complexity factors scale across checkpoints.

Following Lemma D.10, we isolate the leading polynomial factor of our post hoc bounds. We suppress absolute constants, logarithmic factors, and confidence terms, because the experiment is intended to compare scaling across checkpoints. To adapt the single-head expression in Lemma D.10 to BERT, we first replace the propagation factor $\alpha^{(\ell)}$ by

$$\tilde{\alpha}^{(\ell)} = \prod_{k=\ell+1}^{L} L_\phi C_2^{M,(k,\mathrm{in})} C_2^{M,(k,\mathrm{out})} \sum_{h=1}^{A_h} C_2^{\tilde{V},(k,h)}(1 + 4C_2^{QK,(k,h)}).$$

We also replace the local factors $\gamma^{\star,(\ell)}$ by

$$\tilde{\gamma}^{QK,(\ell,h)} = 2C_2^{\tilde{V},(\ell,h)} C_2^{M,(\ell,\mathrm{out})} C_2^{M,(\ell,\mathrm{in})}, \qquad \tilde{\gamma}^{V,(\ell,h)} = C_2^{M,(\ell,\mathrm{out})} C_2^{M,(\ell,\mathrm{in})},$$
$$\tilde{\gamma}^{M,(\ell,\mathrm{in})} = C_2^{M,(\ell,\mathrm{out})}, \qquad \tilde{\gamma}^{M,(\ell,\mathrm{out})} = 1.$$

With $m = \lceil L + \log(N) \rceil$, we define the BERT-adapted proxies for the polynomial part of our bounds by $\tilde{O}(B_{\mathrm{ours}}/\sqrt{n})$, where

$$B_{\mathrm{ours}} = \inf_{\boldsymbol{p} \in \mathcal{P}_m} \mathfrak{B}_{\mathrm{BERT}}(\boldsymbol{p}) + \sqrt{L},$$
$$\mathfrak{B}_{\mathrm{BERT}}(\boldsymbol{p}) = \sum_{\ell=1}^{L} \left(\sum_{h=1}^{A_h} \mathfrak{B}_{QK}^{(\ell,h)} + \sum_{h=1}^{A_h} \mathfrak{B}_{\tilde{V}}^{(\ell,h)} + \mathfrak{B}_{M}^{(\ell,\mathrm{in})} + \mathfrak{B}_{M}^{(\ell,\mathrm{out})} \right),$$

and for $\star \in \{QK, \tilde{V}, M\}$, with $a \in [A_h]$ when $\star \in \{QK, \tilde{V}\}$ and $a \in \{\mathrm{in}, \mathrm{out}\}$ when $\star = M$,

$$\mathfrak{B}_{\star}^{(\ell,a)} = (\|W^{\star,(\ell,a)}\|_{\mathrm{s},p^{\star,(\ell,a)}}^{p^{\star,(\ell,a)}})^{\frac{1}{p^{\star,(\ell,a)}+2}} (\tilde{\gamma}^{\star,(\ell,a)} \tilde{\alpha}^{(\ell)} L)^{\frac{p^{\star,(\ell,a)}}{p^{\star,(\ell,a)}+2}} N^{\frac{p^{\star,(\ell,a)}+1}{p^{\star,(\ell,a)}+2}}. \tag{23}$$

Here $\mathcal{P}_m$ is the corresponding grid of Schatten indices for all headwise attention matrices and feedforward matrices.

For the bounds of Edelman et al. (2022), we analogously retain only the leading polynomial factor and define the BERT-adapted proxies by $\tilde{O}(B_{\mathrm{Edelman}}/\sqrt{n})$, where

$$B_{\mathrm{Edelman}} = \left(1 + \sum_{\ell=1}^{L} (\tilde{\alpha}^{(\ell)})^{\frac{2}{3}} \tilde{\xi}^{(\ell)} \right)^{\frac{3}{2}},$$
$$\tilde{\xi}^{(\ell)} = \sum_{h=1}^{A_h} (C_2^{M,(\ell,\mathrm{out})} C_2^{M,(\ell,\mathrm{in})} C_2^{\tilde{V},(\ell,h)} C_{2,1}^{QK,(\ell,h)})^{\frac{2}{3}} + \sum_{h=1}^{A_h} (C_2^{M,(\ell,\mathrm{out})} C_2^{M,(\ell,\mathrm{in})} C_{2,1}^{\tilde{V},(\ell,h)})^{\frac{2}{3}}$$
$$+ (C_2^{M,(\ell,\mathrm{out})} C_{2,1}^{M,(\ell,\mathrm{in})})^{\frac{2}{3}} + (C_{2,1}^{M,(\ell,\mathrm{out})})^{\frac{2}{3}}.$$

One may worry that the feedforward matrices in the BERT checkpoints have intermediate dimension $I = 4N$, and hence that the mixed $(2,1)$-norms used in the Edelman-type proxies are penalized by the larger rectangular shape. This effect should be interpreted with some care. Since the expansion ratio $I/N = 4$ is fixed for all checkpoints, replacing an $N \times N$ feedforward matrix by the two rectangular matrices $W^{M,(\ell,\mathrm{in})} \in \mathbb{R}^{N \times 4N}$ and $W^{M,(\ell,\mathrm{out})} \in \mathbb{R}^{4N \times N}$ does not by itself change the polynomial exponent in N; it changes only fixed aspect-ratio constants, together with the actual scaling of the trained weights. Indeed, if $W^{M,(\ell,\mathrm{in})}$ is partitioned into four $N \times N$ column blocks, then its mixed $(2,1)$-norm is exactly the sum of the mixed $(2,1)$-norms of these four blocks. Similarly, if $W^{M,(\ell,\mathrm{out})}$ is partitioned into four $N \times N$ row blocks, then its mixed $(2,1)$-norm lies between one half of the sum and the full sum of the blockwise mixed $(2,1)$-norms.[6] Thus the intermediate dimension contributes a fixed-width factor rather than a new N-dependent exponent. The remaining growth of the Edelman-type proxies with N is therefore not merely an artifact of using $I = 4N$, but reflects the behavior of the mixed norms of the trained feedforward weights.

F.3 Comparison and interpretation

Figure 1 plots the normalized version of the generalization gap proxies B_{ours} and B_{Edelman} as L or N varies. The resulting curves show that our proxies grow more slowly than the proxies obtained from Edelman et al. (2022) both when the depth is varied at fixed hidden dimension and when the hidden dimension is varied at fixed depth.

We next examine how the post hoc choice of Schatten indices produces this behavior. In all BERT Miniatures checkpoints considered here, the minimum of $\mathfrak{B}_{\mathrm{BERT}}(\boldsymbol{p})$ is attained at $p = 0$ for every matrix type, layer, and head. This common optimizer should not be interpreted as saying that all

[6] In general, if we partition $W \in \mathbb{R}^{I \times N}$ into four $N \times N$ row blocks as $W^\top = [W_1^\top \; W_2^\top \; W_3^\top \; W_4^\top]$ with $W_b \in \mathbb{R}^{N \times N}$, we have $2^{-1} \sum_{b=1}^{4} \|W_b\|_{2,1} \le \|W\|_{2,1} \le \sum_{b=1}^{4} \|W_b\|_{2,1}$.

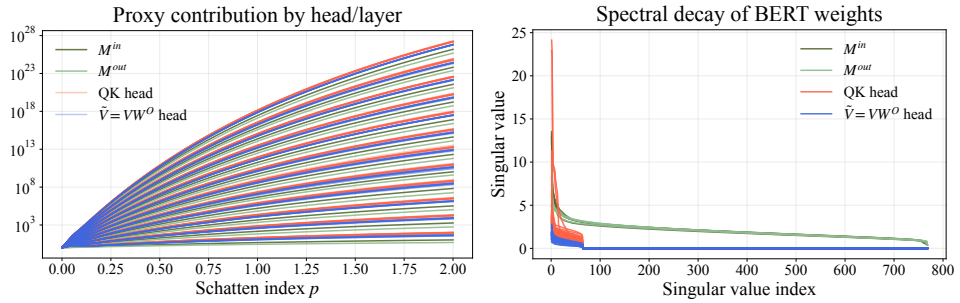

Figure 3: Diagnostics for the post hoc Schatten-index choice for each weight matrix at hidden dimension $N = 768$ and depth $L = 12$. **Left:** Matrixwise proxy contributions as a function of the Schatten index p, normalized by the value at $p = 0$. The increase of all curves as p moves away from zero explains why the proxies select the rank-based endpoint. **Right:** Singular-value spectra of BERT weight matrices. It shows that this common endpoint selection does not come from identical spectra: the headwise attention-related matrices are exactly low-rank with rank at most $d_h = 64$, whereas the feedforward matrices have slower spectral decay. Together, the two panels show that the selected indices reflect a bound-dependent trade-off between spectral structure and architectural factors.

trained weight matrices have the same spectral structure. Rather, it means that, for the contribution term in Eq. (23), the rank-based endpoint gives the most favorable balance among the Schatten term, the layerwise propagation factor, the depth, and the hidden dimension.

The reason is visible directly from Eq. (23). Increasing p replaces the rank-like term by a more norm-like Schatten quantity, but it also increases the architectural factor

$$(\tilde{\gamma}^{\star,(\ell,a)}\tilde{\alpha}^{(\ell)}L)^{\frac{p^{\star,(\ell,a)}}{p^{\star,(\ell,a)}+2}} N^{\frac{p^{\star,(\ell,a)}+1}{p^{\star,(\ell,a)}+2}}.$$

Therefore, positive Schatten indices can improve the proxies only when the reduction in the Schatten term is large enough to compensate for the additional depth- and dimension-dependent factors. In the grid of BERT Miniatures checkpoints used here, this compensation does not occur, so the rank-based endpoint remains optimal.

The left panel of Figure 3 gives a matrixwise view of the endpoint selection. After normalization by the value at $p = 0$, every contribution increases as p moves away from zero. The curves are organized mainly by layer rather than by head: within a fixed layer, the headwise curves for W^{QK} and $W^{\tilde{V}}$ remain close to one another, whereas the separation across layers is much larger. The higher blocks correspond to earlier layers and the lower blocks to later layers, which is consistent with the stronger effect of the propagation factor $\tilde{\alpha}^{(\ell)}$ in earlier layers. Within a fixed layer, the rise away from $p = 0$ is qualitatively strongest for query-key matrices, then for value-related matrices, and then for feedforward matrices.

The right panel of Figure 3 explains why this common optimizer still hides substantial matrix-type heterogeneity. The attention-related matrices W^{QK} and $W^{\tilde{V}}$ are headwise objects and, by construction, have rank at most $d_h = 64$. Hence, for the BERT checkpoints considered here, these matrices are exactly low-rank relative to the hidden dimension. This makes $p = 0$ particularly natural for the attention-related terms. Moreover, the singular values of W^{QK} tend to be larger than those of $W^{\tilde{V}}$, so using positive Schatten indices is relatively more expensive for the query-key proxies.

The feedforward matrices $W^{M,\text{in}}$ and $W^{M,\text{out}}$ behave differently. They do not have the same exact low-rank structure, and their spectral decay is slower. Thus, for these matrices, positive values of p could in principle exploit norm-based spectral information from the singular-value profile. However, in the present proxies, that possible spectral gain is outweighed by the corresponding increase in the propagation, depth, and dimension factors. Overall, Figure 3 shows that the post hoc Schatten indices are not merely diagnostics of spectral decay, but are bound-dependent complexity parameters that balance spectral structure against architectural scaling factors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: spectrum-adaptive post hoc generalization bounds for multi-layer Transformers, a matrix-valued parametric interpolation argument, comparisons with existing norm-based bounds, and proxy experiments using BERT Miniatures checkpoints; see Sections 1, 3, 5, and 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion: the analysis is training-agnostic, uses the learned weights only through post hoc spectral quantities, and does not exploit possible low-dimensional activation structure. It also explains that the BERT experiment is a proxy comparison rather than a literal numerical evaluation of the theorems; see Sections 6 and 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the relevant assumptions for the main theoretical results, including the Transformer model, the bounded/Lipschitz loss setup, layerwise spectral norm control, and the Lipschitz activation assumption in Sections 2 and 3. Proof sketches are given in Section 4, and the complete technical arguments are provided in Appendices B–D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The empirical part is a deterministic proxy comparison rather than a training experiment. Section 6 and Appendix F specify the use of BERT Miniatures checkpoints, the depth/hidden-dimension grid, the architectural adaptations from the theoretical model to BERT, and the construction and normalization of the proxy quantities.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The BERT Miniatures checkpoints used in the experiments are publicly available and cited, and Appendix F specifies the checkpoint grid and the BERT-adapted proxy construction. The supplementary material includes simulation code and instructions for reproducing the figures.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies that no new model is trained for the reported experiments; instead, the experiments compute proxy quantities from publicly released BERT Miniatures checkpoints. Appendix F gives the checkpoint grid, the treatment of multi-head attention and feedforward layers, the value and query-key proxy definitions, and the normalization used for the plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The reported empirical results are deterministic calculations of normalized leading-complexity proxy quantities on fixed public checkpoints, rather than stochastic training or test-set evaluations. Therefore, error bars or confidence intervals are not directly applicable to the main empirical comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments only compute deterministic proxy quantities from publicly available checkpoints and generate the corresponding plots. They do not involve model training or fine-tuning. The full experiment can be run on a standard laptop CPU in a few minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The work is primarily theoretical, and its empirical analysis is limited to evaluating of generalization bounds using public checkpoints. It does not involve private data, human subjects, or deployment of a potentially harmful system.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is foundational theory for understanding generalization in Transformers and we do not identify a direct path from the stated contributions to concrete societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new dataset, pretrained model, or other asset with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses the publicly released BERT Miniatures checkpoints of Turc et al. (2019) and cites the original work in Appendix F. The checkpoints hosted under the Google organization on Hugging Face are listed with the Apache-2.0 license. We use the checkpoints only to extract trained weight matrices for computing deterministic proxy quantities, and do not redistribute modified checkpoint files.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Simulation code is submitted and documented in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.