Preprint 1–59

Higher-Order Transformer Derivative Estimates for Explicit Pathwise Learning Guarantees

Yannick Limmer LIMMERY@MATHS.OX.AC.UK
Department of Mathematics, Oxford-Man Institute, University of Oxford

Anastasis Kratsios KRATSIOA@MCMASTER.CA
Department of Mathematics, McMaster University, Vector Institute

Xuwei Yang YANGX212@MCMASTER.CA
Department of Mathematics, McMaster University

Raeid Saqur KRATSIOA@MCMASTER.CA
Department of Computer Science, University of Toronto, Vector Institute

Blanka Horvath HORVATH@MATHS.OX.AC.UK
Department of Mathematics, Oxford-Man Institute, University of Oxford

Abstract

An inherent challenge in computing fully-*explicit* generalization bounds for transformers involves obtaining covering number estimates for the given transformer class $\mathcal{T}$. Crude estimates rely on a uniform upper bound on the local-Lipschitz constants of transformers in $\mathcal{T}$, and finer estimates require an analysis of their higher-order partial derivatives. Unfortunately, these precise higher-order derivative estimates for (realistic) transformer models are not currently available in the literature as they are combinatorially delicate due to the intricate compositional structure of transformer blocks.

This paper fills this gap by precisely estimating all the higher-order derivatives of all orders for the transformer model. We consider realistic transformers with multiple (non-linearized) attention heads per block and layer normalization. We obtain fully-explicit estimates of all constants in terms of the number of attention heads, the depth and width of each transformer block, and the number of normalization layers. Further, we explicitly analyze the impact of various standard activation function choices (e.g. SWISH and GeLU). As an application, we obtain explicit *pathwise generalization bounds* for transformers on a *single trajectory of an exponentially-ergodic Markov process* valid at a fixed future time horizon. We conclude that real-world transformers can learn from N (non-i.i.d.) samples of a single Markov process's trajectory at a rate of $\mathcal{O}\big(\operatorname{polylog}(N)/\sqrt{N}\big)$.

1. Introduction

Transformers Vaswani et al. (2017) have become the main architectural building block in deep learning-based state-of-the-art foundation models Bommasani et al. (2021); Zhao et al. (2023); Wei et al. (2022). The undeniable success of these large language models (LLMs) in predicting on *non-i.i.d.* sequential data has inspired a deluge of research into the statistical properties of transformers, ranging from minimax (order) optimal guarantees transformers trained in context Kim et al. (2024), statistical analyses of their training dynamics under linearized/surrogate attention mechanisms Chen et al. (2024); Lu et al. (2024); Zhang et al. (2024b); Siyu et al. (2024); Kim and Suzuki (2024); Duraisamy (2024) and their infinite-width limits surrogates Zhang et al. (2024a), as well as PAC-Bayesian guarantees Mitarchuk et al. (2024). Though each result helps to explain a piece of the LLM puzzle, each has one of the following drawbacks which we address:

(i) **Non-Explicit Constants:** They only provide learning rates with unknown constants,

(ii) **Non-Sequential Data:** Their guarantees do not apply to non-i.i.d. (sequential) data,

(iii) **Unrealistic Transformers:** They rely on linearized attention surrogates, a single attention head, omitting key normalization layers, or non-standard activation functions.

This paper fills this gap by deriving statistical guarantees for transformers which do not suffer from the limitations in *(i)-(iii)*. We guarantee (Theorem 3) that transformers trained on a single time-series trajectory can generalize at future moments in time, with explicit constants. We, therefore, consider the learning problem where the user is supplied with N paired samples $(X_1, Y_1), \dots, (X_N, Y_N)$, where each input $Y_n = f^\star(X_n)$ for a smooth (unknown) *target function* $f^\star : \mathbb{R}^{d\times M} \to \mathbb{R}^D$ is to be learned, depending on a history length M, and where the inputs are generated by a time-homogeneous Markov process $X_\cdot \overset{\text{def.}}{=} (X_n)_{n=1}^\infty$, or any process admitting a finite-dimensional Markovian lift in the sense of Cuchiero and Teichmann (2019). The assumption $Y_n = f^\star(X_n)$ results in only a mild loss of generality since if $X_\cdot$ is a discretized solution to a stochastic differential equation then $Y_n \approx$ signal+noise due to stochastic calculus considerations (see Appendix G).

The performance of any transformer model $\mathcal{T} : \mathbb{R}^{d\times M} \to \mathbb{R}^D$ is quantified via a smooth loss function $\ell : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. When $M = 1$, the generalization of such a $\mathcal{T}$ is measured by the gap between its *empirical risk* $\mathcal{R}^{(N)}$, computed from the single-path training data, and its *(true) t-future risk* $\mathcal{R}_t$ at a (possibly infinite) future time $N \leqslant t \leqslant \infty$ ($t \in \mathbb{N}_+$) defined by

$$\mathcal{R}_t(\mathcal{T}) \overset{\text{def.}}{=} \mathbb{E}\big[\ell(\mathcal{T}(X_t), f^\star(X_t))\big]$$

where $\mathcal{R}_t$ (resp. $\mathcal{R}_\infty$) is computed with respect to the distribution of X_t (resp. *stationary distribution* of $X_\cdot$). The time-t excess-risk $\mathcal{R}_t$, which is generally unobservable, is estimated by a single-path estimator known as the empirical risk computed using all the noisy samples observed thus far

$$\mathcal{R}^{(N)}(\mathcal{T}) \overset{\text{def.}}{=} \frac{1}{N}\sum_{n=1}^{N} \ell(\mathcal{T}(X_n), f^\star(X_n)).$$

Our objective is to obtain a statistical learning guarantee bounding the gap between the empirical risk and the t-future risk of transformer models trained on a single path.

Contributions. Our main *statistical guarantee* (Theorem 3) is a bound on the *future-generalization*, free of limitations (i)-(iii), valid at any given time $t \geqslant N$, of a transformer trained from N samples collected from an unknown transformation ($f^\star$) of any suitable unknown Markov process ($X_\cdot$). Fix a class of transformers $\mathcal{TC}$ for respective input and output dimensions d and D, i.e. determine the number of transformer blocks, the number of attention heads, channel sizes, and specify a constraints on its weights. Then, the first takeaway of our main result is that with probability at least $1-\delta$

$$\sup_{\mathcal{T}\in\mathcal{TC}} \left|\mathcal{R}_t(\mathcal{T}) - \mathcal{R}^{(N)}(\mathcal{T})\right| \in \mathcal{O}\left(\frac{\log(1/\delta) + \log(N)^{1/s}}{\sqrt{N}}\right) \qquad \text{(FutureGen)}$$

where $s > 0$ can be made arbitrarily large and $\mathcal{O}$ hides a constant which is challenging to *estimate accurately* due to the involved *combinatorics*, and is explicitly computed in (Theorem 2).

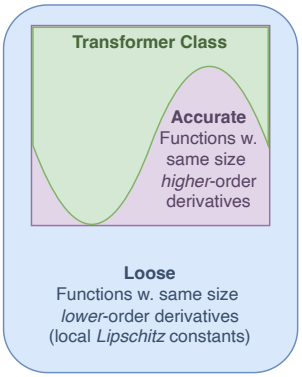

Figure 1: **Our Results:** Tight estimates on the constants under $\mathcal{O}$ in (FutureGen) (Theorem 3) is a consequence of our main result (Theorems 41). The latter accurately estimates the *radius* of the *smallest* C^s-ball containing the realistic transformer class $\mathcal{TC}$; this radius is precisely the in the largest s^{th}-order partial derivative of any transformer in the class $\mathcal{TC}$. **Why it works:** Our higher-order derivative-based analysis accurately approximates the transformer class and provides both *dimension-free* convergence *rates* and accurate constants. In contrast, low-order local-Lipschitz analysis provides loose generalization bounds by overestimating the size of the class $\mathcal{TC}$ via inclusion in a large class of functions with comparable first derivatives.

Our *primary contribution* (Theorems 2) is a *full analysis* of the higher order sensitivities/ derivatives of the transformer network; which in-turn yield explicit accuracy constants under the big $\mathcal{O}$ in our main statistical guarantees (Theorem 3). With this, our result provides a generalization bound applicable to transformers trained on non-i.i.d. data with *explicit constants*.

Insights from explicit constants: Our result enables us to analyse the effects of the number of attention heads, depth, and width of the transformed model, and weight and bias restriction, as well as on the activation functions used on the generalization of the transformer model in detail. This is because the explicit constants in our main results are clearly stated in terms of these quantities, for which we provide efficient code to compute. In particular, we are able to perform an in-depth analysis for the Swish Ramachandran et al. (2017), GeLU Hendrycks and Gimpel (2016), and the $\tanh$ activation functions. Our analysis aligns with the empirical evidence that the activation functions such as Swish provide superior performance than unconventional choices such as $\tanh$.

Further consequences: Although not explicitly addressed here, our results can also be applied to derive statistical guarantees for transformers using alternative frameworks (see Section 1.1 for an overview). This can, for instance, be achieved for i.i.d. data through chaining arguments relying on covering number estimates for balls in smoothness spaces (van der Vaart and Wellner, 2023, Theorem 2.71) containing the relevant transformer. The primary gap in the existing literature, which Theorems 2 fills, is the lack of partial derivative bounds necessary to compute the radii of the balls in these smoothness spaces, which contain the transformer architecture (as illustrated in Figure 1).

1.1. Related Work

Statistical Guarantees for Non-I.i.d. Data. There are several results in the literature addressing learning with non-i.i.d. data satisfying a mixing/ergodicity condition dating back, e.g. Yu (1994); Mohri and Rostamizadeh (2008, 2010); Kuznetsov and Mohri (2017); Simchowitz et al. (2018); Foster et al. (2020); Ziemann and Tu (2022), dating at least back to Yu (1994). However, *none* of these results provide *explicit* generalization bounds. Rather, they only provide convergence rates without constants for the class $\mathcal{TC}$ since they require covering number estimates, which would ultimately rely on the partial derivative bounds (Theorems 41 and 42).

This is because they either rely on bounding the Rademacher complexity of the transformer class, e.g. in applying Mohri and Rostamizadeh (2008), or they rely on computing the cardinality of delta nets Ziemann and Tu (2022), both of which necessitate the computation of the worst-case Lipschitz (or C^s norm) of any transformer in the hypothesis class using (van der Vaart and Wellner, 2023,

Theorem 2.7.4) and (Lorentz et al., 1996, Equation (15.1.8)). Since these higher-order derivatives where *previously unavailable* in the literature, then these results did not work out explicit constants for their generalization bounds. We also note the results of Simchowitz et al. (2018); Foster et al. (2020), which obtained statistical guarantees for non i.i.d. data under restrictive assumptions on the data-generating process, in addition to the aforementioned restrictions.

We highlight the martingale arguments e.g. Kontorovich (2014) and concentration of measure phenomena for martingale sums e.g. Bercu et al. (2015); Boucheron et al. (2013) techniques for deriving generalization bounds for models trained on non-i.i.d. data. Again, these results do not yield explicit constants for $\mathcal{TC}$ without estimating the derivatives (or at least local-Lipschitz constants) of the transformers therein, again requiring Theorems 41 and 42 for a completely explicit result.

Statistical Guarantees for LLMs Trained on I.i.d. Data. In addition to the aforementioned statistical guarantees for LLMs; the available statistical guarantees for transformers which the authors are aware of either concern in-context learning for linear transformers Zhang et al. (2024a); Garg et al. (2022), transformers Von Oswald et al. (2023); Akyürek et al. (2023) trained with gradient descent, or instance-dependent bounds Trauger and Tewari (2023) for general transformers. These results do not apply in contexts of time series analysis where each training sample is not independent of the others but is rather generated by some recursive stochastic process.

Ergodic Markov Processes and Mixing Times. We require that the data-generating Markov process has an exponentially contracting Markov kernel Kloeckner (2020). For Markov chains, i.e. finite-state space Markov processes, this means that the generator (Q-matrix) of the Markov chain has a *spectral gap*. These spectral gaps are actively studied in the Markov chain literature Mufa (1996); Kontoyiannis and Meyn (2012); Atchadé (2021); Paulin (2015); Kloeckner (2019) since these have a finite mixing time, meaning that the distribution of such Markov chains approaches their stationary limit after a large finite time has elapsed; i.e. they have well-behaved (approximate) mixing times Montenegro et al. (2006); Hsu et al. (2015); Wolfer and Kontorovich (2019); Zamanlooy (2024). We rely on actively-studied optimal transport-theoretic notions of mixing since it is easily verified, or already known, for most data-generating processes than more classical notions; e.g. Kuznetsov and Mohri (2017); Mohri and Rostamizadeh (2010).

We note that our generalization bounds rely on the concentration of measure arguments for the "smooth" optimal transport distances in Kloeckner (2020); Riekert (2022). The last step of our analysis refines the concentration of measure arguments of Hou et al. (2023b); Benitez et al. (2023); Kratsios et al. (2024) to the non-i.i.d. setting using exponentially ergodic Markov process theory.

2. Background and Preliminaries

This section overviews the necessary background for a self-contained formulation of our main results. This includes the definition of transformers and examples of data-generating processes treatable within our framework.

2.1. Admissible Data-Generating Processes

Our statistical guarantees are driven by Markovian arguments; we emphasize that these are not heavily restrictive since we can Markovianize processes in higher dimensional state spaces. Such state-space extensions, known as Markovian lifts in the literature Cuchiero and Teichmann (2020),

typically require tracking only a few recent movements of the data-generating process rather than solely its current state.

We operate in the following setting: Fix dimensions $d, D \in \mathbb{N}_+$, a finite memory $M \in \mathbb{N}_+$, and let $X_\cdot \overset{\text{def.}}{=} (X_n)_{n\in\mathbb{N}_0}$ be a stochastic process taking values in $\mathbb{R}^d$, such that the lifted process $X_\cdot^M \overset{\text{def.}}{=} (X^M_{[0\vee(n-M),\dots,0\vee n]})_{n\in\mathbb{N}_0}$ is Markovian on $\mathbb{R}^{Md}$. Let P be a Markov kernel on a non-empty Borel $\mathcal{X}^M \subseteq \mathbb{R}^{Md}$ with initial distribution $X_0 \sim \mu_0 \in \mathcal{P}(\mathbb{R}^{Md})$ given by $X_n^M \sim \mu_n \overset{\text{def.}}{=} P^n\mu_0 \overset{\text{def.}}{=} \mathbb{P}(X_n^M \in \cdot)$ and for each $x \in \mathbb{R}^{Md}$ and $n \in \mathbb{N}_+$, set $P^n(x,\cdot) \overset{\text{def.}}{=} \mathbb{P}(X_n^M \in \cdot | X_0^M = x)$. The process $X_\cdot^M$ is a *Markovian lift* of $X_\cdot$. Examples of processes with *finite-dimensional* Markovian lifts are ARIMA times-series models, see e.g. Cryer and Kellet (1991).

Assumption 1 (Bounded Trajectories) *There is a $c > 0$ such that $\mathbb{P}(\sup_{t\in\mathbb{N}} \|X_t\| \leqslant c) = 1$.*

Assumption 2 (Exponential Ergodicity) *There is a $\kappa \in (0,1)$ such that: for each $\mu, \nu \in \mathcal{P}(\mathbb{R}^{Md})$ and every $t \in \mathbb{N}_+$ one has $\mathcal{W}_1(P^t\mu, P^t\nu) \leqslant \kappa^t \mathcal{W}_1(\mu,\nu)$.*

2.1.1. EXAMPLES: PROJECTED SDES - FROM LANGEVIN DYNAMICS TO MARTINGALES

A broad class of non-i.i.d. data-generating processes satisfying our assumptions is a broad generalization of any Markov processes obtained by "projecting" the strong solution to a stochastic differential equation (SDE) with overdampened drift onto a compact convex subset of $\mathbb{R}^d$. The processes which we can project are vast generalizations of the forward process used in *denoising diffusion models*; see e.g. Song et al. (2020) whose convergence is by now well-understood; see e.g. Chen et al. (2023).

Example 1 (Projected SDEs with Overdampened Drift) *Consider a latent dimension $\bar{d} \in \mathbb{N}_+$, $\mu : \mathbb{R}^{\bar{d}} \to \mathbb{R}^{\bar{d}}$ be Lipschitz and the gradient of a strongly convex function; i.e. there is a $K > 0$ such that $(\mu(x) - \mu(y))^\top (x-y) \leqslant -K\|x-y\|^2$ for all $x, y \in \mathbb{R}^d$. For any $x \in \mathbb{R}^{\bar{d}}$ let $Z_\cdot^x \overset{\text{def.}}{=} (Z_t^x)_{t\geqslant 0}$ be the unique strong solution (which exists by (Da Prato, 2008, Theorem 8.2) since μ is Lipschitz)*

$$Z_t^x = x + \int_0^t \mu(Z_s^x)\, ds + \int_0^t W_s \tag{1}$$

where $W_\cdot$ is a $\bar{d}$-dimensional Brownian motion. Let $f : \mathbb{R}^{\bar{d}} \to \mathbb{R}^d$ be a bounded 1-Lipschitz function and consider the discrete-time Markov process $X_\cdot \overset{\text{def.}}{=} (X_n)_{n=0}^\infty$ on $\mathbb{R}^d$ given by

$$X_n^x \overset{\text{def.}}{=} f(Z_n^x).$$

As shown in Proposition 5, $X_\cdot$ satisfies Assumptions 1 and 2. The standard example of SDEs (1) are Langevin dynamics for a strictly convex potential $U : \mathbb{R}^d \to \mathbb{R}$. As shown in Bolley et al. (2012)

$$\mu(x) = -\nabla U(x)/2.$$

Example 2 (Projections of Diffusive Martingales) *Let $d \in \mathbb{N}_+$. Let $\sigma : \mathbb{R}^d \to P_d^+$ taking values in the cone P_d^+ of $d \times d$-dimensional positive definite matrices, be Lipschitz with the Fröbenius norm on $\mathbb{R}^{d\times d}$, and satisfy the uniform ellipticity condition: there exists a $\lambda > 0$ such that for every $x \in \mathbb{R}^d$ holds $s_{min}(\sigma(x)\sigma(x)^\top) \geqslant \lambda$, where $s_{min}(A)$ denotes the minimal singular values of a matrix A. The martingale $Z_\cdot$ (see (Da Prato, 2008, Proposition 6.15) for a proof of martingality) defined for each $t \geqslant 0$ by $Z_t = \int_0^t \sigma(Z_s)\, dW_s$ where $W_\cdot \overset{\text{def.}}{=} (W_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be any 1-Lipschitz bounded function. By Proposition 7, the data-generating Markov process $X_\cdot \overset{\text{def.}}{=} (X_n)_{n=0}^\infty$, defined for each $n \in \mathbb{N}_+$ by $X_n \overset{\text{def.}}{=} f(Z_n)$ satisfies both Assumptions 1 and 2.*

We have presented the simplest cases here; which is readily generalizable. By Lemma 6 to any Markov process exponentially ergodic $Z_\cdot$, not necessarily solving the simple dynamics (1), automatically yields examples of data-generating processes satisfying both Assumptions 1 and 2. We list some examples of such processes here: McKean-Vlasov type with relatively general, i.e. it can have non-constant law-dependent drift and diffusion coefficients (Wang, 2023, Corollary 4.4) (possibly with reflections), several SDEs is driven by a pure-jump Lévy process (Luo and Wang, 2019, Theorem 3.1). When considering reflected SDEs, where the reflections constrain the process to remain in a bounded convex domain, we do not need f to be bounded, as the processes themselves are already bounded. Further examples of such can be constructed using compact Riemannian sub-manifolds of $\mathbb{R}^d$ with suitable curvature bounds Ollivier (2009). Additional examples include Markov processes whose kernels satisfy a log-Sobolev inequality (see Appendix B.2 for details).

2.2. The Transformer Model

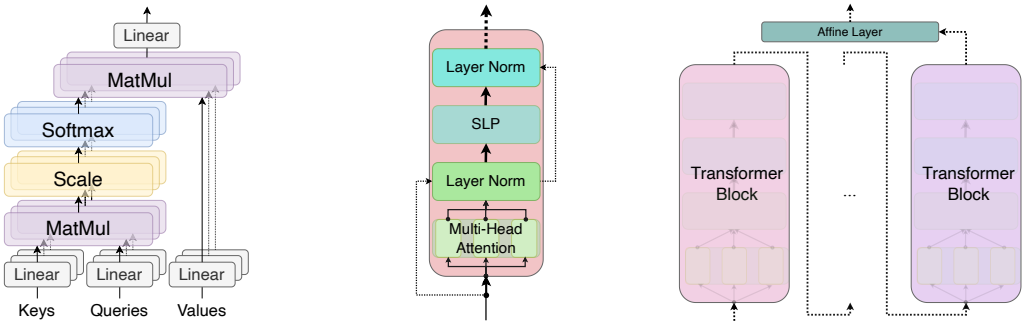

Figure 2: Multi-Head Att. (Def. 10), Transformer Block (Def. 11), and Transformer (Def. 12).

The overall structure of transformers is summarized in Figure 2, and we give an in-depth definition of all components with their respective dimensions in Appendix C, which is relevant for the details of the bound computation. On a high level, the most important aspects are:

Multi-Head Attention [$\mathcal{MH}$]. Consists of parallel application of the attention mechanism, described by the following steps.

(i) Inputs are used three-fold, as keys, queries, and values, all are transformed by distinct linear transformations.
(ii) Keys and queries are multiplied, scaled, and transformed by a softmax application.
(iii) This output is combined in a matrix multiplication with the values.

Transformer Block [$\mathcal{TB}$]. Here,

(i) input features are mapped to contexts via *multi-head attention mechanism*,
(ii) the output of the multi-head attention mechanism and the input features (via a skip connection) are normalized by a *layer-norm [$\mathcal{LN}$]* (see Appendix C),
(iii) the normalized contextual features are transformed non-linearly by a *single-layer perceptron [$\mathcal{PL}$]*, and
(iv) its outputs, together with the first set of normalized context (via another skip connection), are normalized by a final *layer-norm* and returned by the transformer block.

Transformer [$\mathcal{T}$]. Iteratively feed input features through a series of transformer blocks before processing their outputs with a (fully connected affine layer). We denote a class of transformers of a fixed architecture by $\mathcal{TC}$, with each parameter bounded to a predefined domain.

2.3. Setting

We consider smooth loss and target functions that are concentrated on a compact region, along with their derivatives. The growth rate of the C^s-norm (see Theorem 4 in Appendix A) of the loss function and its derivatives quantifies the degree of concentration. One easily verifies that any function in the Schwartz class satisfies this former of the following conditions, cf. Treves (2016).

Definition 1 (Polynomial Growth of Derivatives) *Let $d, D, M \in \mathbb{R}$. A smooth function $g : \mathbb{R}^{Md} \to \mathbb{R}^D$ is in the class $C^\infty_{poly:C,r}([0,1]^{Md}, \mathbb{R}^D)$ if $C, r \geqslant 0$ are such that $\|g\|_{C^s([0,1]^{Md})} \leqslant C\, s^r$ for each $s \in \mathbb{N}_+$. Here, $\|\cdot\|_{C^s}$ is the uniform Sobolev norm on the specified domain.*

In Appendix E.2 we show that, in one dimension, any real analytic function whose power series expansion at 0, has coefficients growing at an $\mathcal{O}((s+1)^r)$ rate belongs to $g \in C^\infty_{poly:C,r}([0,1], \mathbb{R})$. One can easily extend this argument to multiple dimensions to obtain further examples.

We consider an *realizable* PAC learning problem, determined by a smooth 1-Lipschitz *target function* $f^\star : \mathbb{R}^d \to \mathbb{R}^D$ which we would like to learn using a sequence of random observations $\big((X_t, Y_t)\big)_{t\in\mathbb{N}}$ as our training data. That is, for each $t \in \mathbb{N}_+$

$$Y_t \overset{\text{def.}}{=} f^\star(X_t)$$

We aim to learn f from a *single path*. The ability of a model to reliably recover the function $f : \mathbb{R}^d \to \mathbb{R}^D$ at time t, given the input X_t, is quantified by the *t-future risk*

$$\mathcal{R}_t(f) \overset{\text{def.}}{=} \mathbb{E}\big[\ell(f(X_t), f^\star(X_t))\big].$$

The time-t excess-risk $\mathcal{R}_t$, which is generally unobservable, is estimated by a single-path estimator known as the empirical risk computed using all the noisy samples to observed thus far

$$\mathcal{R}^{(N)}(f) \overset{\text{def.}}{=} \frac{1}{N}\sum_{n=1}^{N} \ell(f(X_t), f^\star(X_t)).$$

Our objective is to obtain a statistical learning guarantee on the quality of our estimate of the target function given by the *time t ($\in \mathbb{N}_+$) generalization gap* $\big|\mathcal{R}_t(f) - \mathcal{R}^{(N)}(f)\big|$.

We now summarize our setting and all parameters defining it, e.g. dimension, number of attention heads in the transformer, growth rate of the derivatives of the target and loss functions, etc.

Setting 2.1 (Standing Assumptions) *Consider a hypothesis class $\mathcal{TC}$. Fix $r_f, r_\ell, C_f, C_\ell \geqslant 0$, as well as a target function f^* and loss function ℓ with*

$$f^\star \in C^\infty_{poly:C_f,r_f}(\mathbb{R}^{Md}, \mathbb{R}^D) \quad \text{and} \quad \ell \in C^\infty_{poly:C_\ell,r_\ell}(\mathbb{R}^{2D}, \mathbb{R});$$

and suppose that Assumption 1 and either of Assumptions 2 and 3 hold.

3. Main Results

3.1. Analytic Results: Bounds on the Higher-Order Derivatives of Transformer

Our primary contribution is the computation of $C_{\ell,\mathcal{TC},K,s} \stackrel{\text{def.}}{=} \sup_{\mathcal{T}\in\mathcal{TC}} \|\ell(\mathcal{T}, f^*)\|_{C^s}$, which encodes the maximal size of the first s partial derivatives of any transformer in the class $\mathcal{TC}$. Thus, it describes the complexity of the class $\mathcal{TC}$ (e.g. int terms of number of attention heads, depth, width, etc...), the size of the compact set K, and the smoothness of the loss function and target functions.

We note that, any uniform generalization bound for smooth functions thus necessarily contains constants of the same order *hidden within the big* $\mathcal{O}$. See e.g. the entropy bound in (van der Vaart and Wellner, 2023, Theorem 2.71) which yields VC-dimension bounds via standard Dudley integral estimates in the i.i.d. case.

Critically, when the function class is defined by function composition, i.e. as in deep learning, then these maximal partial derivatives tend to grow factorially in s. This is a feature of the derivatives of composite functions in high dimensions as characterized by the multi-variate chain rule (i.e. the Faá di Bruno formula Faa di Bruno (1855); Constantine and Savits (1996)). The combinatorics of these partial derivatives is encoded by the coefficients in the well-studied bell-polynomials Bell (1934); Mihoubi (2008); Wang and Wang (2009) whose growth rate has been recently understood in Khorunzhiy (2022) and contains factors of the order of $\mathcal{O}(\left(\frac{2s}{e\ln s}(1+o(1))\right)^s)$.

Remark that, in the feedforward case, i.e. when no layernorms or multihead attention are used, then the $s=1$ case is bounded above by the well-studied path-norms; see e.g. Bartlett et al. (2017); Neyshabur et al. (2015), which are simply the product of the weight matrices of in the network and serve as a simple upper-bound for the largest Lipschitz constant (i.e. C^1 norm) of the class $\mathcal{TC}$. These constants are included as very specific cases of our constant bounds. This is why we present two versions: a weaker but simpler bound, as well as a more accurate but detailed bound.

Theorem 2 ($\mathcal{TC}$-bound in terms of $\mathcal{O}$) *In the case of a single transformer block $C_{\ell,\mathcal{TC},K,s}$ is of the order of*

$$\mathcal{O}\Big(\underbrace{C^{\ell,f^\star}}_{\text{Loss \& Target}} \underbrace{C^{\mathcal{LN}}_{K^{(3)}}(\leqslant s)^s C^{\mathcal{LN}}_{K^{(1)}}(\leqslant s)^{s^3}}_{\text{Layernorms}} \underbrace{C^{\mathcal{PL}}_{K^{(2)}}(\leqslant s)^{s^2}}_{\text{Perceptron}} \underbrace{\Big(1+C^{\mathcal{MH}}_{K}(\leqslant s)\Big)^{s^4}}_{\text{Multihead Attention}} \underbrace{D^{s^2}\, d^{2s^3}}_{\text{dimensions}} \underbrace{c_s^{s^s+s^3+s^4}}_{\text{Generic: } s\text{-th order Derivative}} \Big)$$

where the "generic higher-order derivative constant" is $c_s \stackrel{\text{def.}}{=} \frac{2s}{e\ln s}(1+o(1))$. Further,

$$C^{\ell,f^\star} = \mathcal{O}\big(C_f^s\, s^{r_\ell+2s^2}\big), \qquad C^{\mathcal{PL}}_K(\leqslant s) = \mathcal{O}\big(c^{\mathcal{PL}} + d_{\text{ff}}\|\sigma\|_s \tilde{c}_s^s (c^{\mathcal{PL}})^{s+1}\big),$$
$$C^{\mathcal{LN}}_K(\leqslant s) = \mathcal{O}\big(s^{(1+s)/2} c_s^s\big), \qquad C^{\mathcal{MH}}_K(\leqslant s) = \mathcal{O}\big(e^{-2s} M^2 (2d_{\text{in}} d_K \cdot c_s)^s (s\cdot c^{\mathcal{MH}})^{2s+2}\big).$$

Here $\tilde{c}_s \stackrel{\text{def.}}{=} s^{1/2}(n/e)^s c_s^s$; d_{in} is the input-dimension and d_K is the key-dimension of the multi-head attention $\mathcal{MH}$ (see Theorem 10 for details); d_{ff} is the width of the neural network $\mathcal{PL}$ (see Theorem 11 for details); $c^{\mathcal{PL}}$ as well as $c^{\mathcal{MH}}$ are parameter bounds on $\mathcal{PL}$ as well as $\mathcal{MH}$, respectively (see Theorem 42 for details); and $\|\sigma\|_s$ is the C^s-bound of the activation function used. If no layer norms, SLP, or multi head attention mechanisms are included in the class, then their respective terms in our order estimate should be taken to be 1.

Proof The result is a direct consequence of Theorems 42 and 46. The order of the bounds $C^{\mathcal{LN}}_K$, $C^{\mathcal{PL}}_K$, and $C^{\mathcal{MH}}_K$ are given by Theorems 28, 37 and 40. ■

Explicit bound computation. We further refined this result by deriving formulae that enable the precise calculation of these bounds. In order to enhance the accuracy of these estimates, we distinguished not only between different levels of derivatives but also between various types of derivatives. An exemplary improvement of the bound by this distinction can be seen on the right-hand side of Figure 3.

By *derivative type*, we are distinguishing the different multi-variate derivative operators, but regard those as equivalent that are identical after sorting. In a 2-dim setting, for instance, we would say the $(1, 2)$-derivative (once in the first component, twice in the second component) has the same type as the $(2, 1)$-derivative, but not the $(3, 0)$-derivative. The reason why we choose to consider an equivalence is that the size of the set $\mathcal{P}$ in Faà di Bruno's Formula (Theorem 15) grows too big in high dimensions for numeric evaluation.

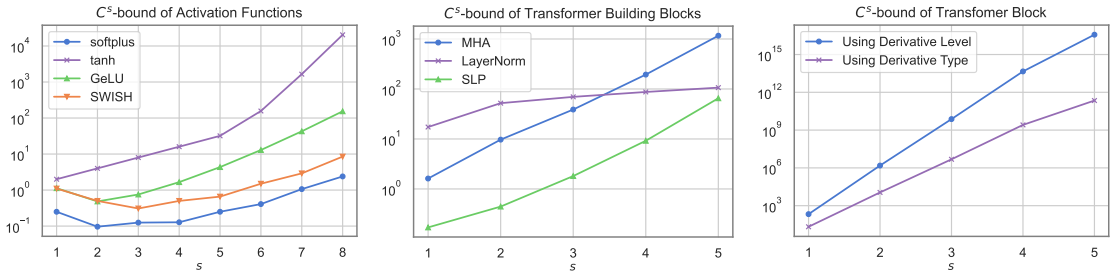

Figure 3: Effects of Transformer Components of FutureGen: (left to right.) The first figure shows the C^s bound of various activation functions according to results in Appendix F.4.3. The second illustrates C^s bounds for Multi-Head Attention (Theorem 10), single-layer perceptrons, and the layer norm. The third shows the C^s-bound of a transformer block (Theorem 11), distinguishing if the bound was computed level-specific (Theorem 18) or type specific (Theorem 19). The parameters used for the above plots are the base cases of Tables 1 to 5.

Since these results are fairly technical and verbose, we relegate them to Appendix F.4, see Theorem 41 for the analogue result to Theorem 42 and Theorems 26, 36 and 39 for tighter bounds on $C_K^{\mathcal{LN}}$, $C_K^{\mathcal{PL}}$, and $C_K^{\mathcal{MH}}$. Additionally, we provide software tools to efficiently compute the bounds of a given transformer architecture.[1]

3.2. Statistical Results: Bounds on Future Generalization for Non-I.i.d. Data

Having characterized an upper bound for $C_{\ell,\mathcal{TC},K,s}$, we may now state our statistical guarantee, which is a version of (FutureGen). This version provides insights on the future-generalization of transformers via 1) explicit constants and 2) explicit *phase transition times* above which the convergence rate in (FutureGen) accelerates by a polylogarithmic factor. We express these times of convergence rate acceleration using the following *convergence rate function*

$$\text{rate}_s(N) \overset{\text{def.}}{=} \begin{cases} \frac{\log(cN)^{Md-2s+s/(Md)}}{c_2 N^{s/(Md)}} & \text{if } Md > 2s \quad \textit{(initial phases)} \\ \frac{\log(cN)}{cN^{1/2}} & \text{if } Md = 2s \quad \textit{(critical phase)} \\ \frac{\log(cN)^{Md/(2s+1)}}{cN^{1/2}} & \text{if } Md < 2s \quad \textit{(eventual phases)} \end{cases} \tag{rate}$$

1. The source code to compute derivative bounds is available at https://anonymous.4open.science/r/transfomer-bounds-B476.

where $c \overset{\text{def.}}{=} 1-\kappa$, $c_2 \overset{\text{def.}}{=} c^{s/(Md)}$, and $0<\kappa<1$ are constants depending only on $X_{\cdot}$.

Our main statistical guarantee provides an infinite number of generalization bounds, which all converge, each with different rates and constants. The bounds with the fastest convergence have the largest constants and visa-versa. This allows us to *sort* this family of simultaneous bounds as a function of the sample size N; with the small constant bounds providing tighter generalization guarantees for small sample sizes N while the rapidly converging bounds provide tighter guarantees for large sample sizes N.

The small-sample bounds depend only on the first few derivatives of the transformers in our class, while the large-sample bounds rely on higher-order derivatives. The point at which one bound in our class surpasses another, becoming the tightest bound for a given sample size N, is characterized by a transition time, denoted τ_s, as illustrated in Figure 4. In this way, every single derivative order computed in Theorem 2 is simultaneously utilized to obtain precise statistical guarantees optimized to the sample size N.

Theorem 3 (Pathwise Generalization Bounds for Transformers) *In Setting 2.1, there exists $\kappa \in (0,1)$, depending only on $X_{\cdot}$, and $t_0 \in \mathbb{N}_0$; such that for each $t_0 \leqslant N \leqslant t \leqslant \infty$ and $\delta \in (0,1]$ the following holds with probability at-least $1-\delta$*

$$\sup_{\mathcal{T}\in\mathcal{TC}} \left|\mathcal{R}_{\max\{t,N\}}(\mathcal{T}) - \mathcal{R}^{(N)}(\mathcal{T})\right| \lesssim \sum_{s=0}^{\infty} I_{N\in[\tau_s,\tau_{s+1})}\, C_{\ell,\mathcal{TC},K,s}\left(\underbrace{\kappa^t}_{\text{(I)}} + \underbrace{\text{rate}_s(N)}_{\text{(II)}} + \underbrace{\frac{\sqrt{2\ln(1/\delta)}}{N^{1/2}}}_{\text{(III)}}\right)$$

with $I_{\cdot}$ as indicator function, $\text{rate}_s(N)$ as in (rate), the constant $C_{\ell,\mathcal{TC},K,s} \overset{\text{def.}}{=} \sup_{\mathcal{T}\in\mathcal{TC}} \|\ell(\mathcal{T},f^*)\|_{C^s}$, and the transition times $(\tau_s)_{s=0}^{\infty}$ are given iteratively by $\tau_0 \overset{\text{def.}}{=} 0$ and for each $s\in\mathbb{N}_+$

$$\tau_s \overset{\text{def.}}{=} \inf\left\{ t \geqslant \tau_{s-1} : C_{\ell,\mathcal{TC},K,s}(\kappa^t + \text{rate}_s(N) + \tfrac{\sqrt{\log(1/\delta)}}{\sqrt{N}}) \leqslant C_{\ell,\mathcal{TC},K,s-1}(\kappa^t + \text{rate}_{s-1}(N) + \tfrac{\sqrt{\log(1/\delta)}}{\sqrt{N}}) \right\}.$$

Furthermore, $c \overset{\text{def.}}{=} 1-\kappa$, $c_2 \overset{\text{def.}}{=} c^{s/(Md)}$, $\kappa^{\infty} \overset{\text{def.}}{=} \lim_{t\to\infty} \kappa^t = 0$, and $\lesssim$ hides an absolute constant.

Theorem 3 implies the order estimate in (FutureGen). This is because $C_{\ell,\mathcal{TC},K,s}$ is constant in N and $\text{rate}_s(N) < \text{rate}_{s-1}(N)$; thus, for every $s > 0$ the right-hand side our bound is eventually bounded by any $C_{\ell,\mathcal{TC},K,s}(\kappa^t + \sqrt{2\ln(1/\delta)}/N^{1/2} + \text{rate}_s(N))$ for N large enough. However, unlike the order estimate (FutureGen), Theorem 3 provides an explicit description of the actual size of the future-generalization gap in terms of three factors which we now interpret.

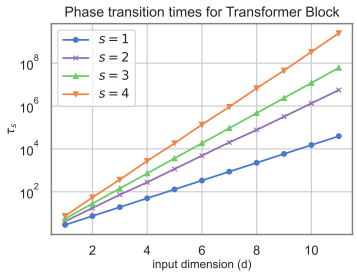

Figure 4: **Transition times:** (y-axis) when the future-generalization bound accelerates by a polylogarithmic factor (in N) for a single transformer block in terms of the input dimension d (x axis). See Section 3.1 for details on constants.

Term (I): Non-Stationarity. This term quantifies the rate at which the data-generating Markov process $X_{\cdot}$ becomes stationary. This term only depends on the time t and a constant $0<\kappa<1$ determined only by $X_{\cdot}$. The limiting case is described using the notational convention $\kappa^{\infty} \overset{\text{def.}}{=} \lim_{t\to\infty} \kappa^t = 0$.

Term (II): Model Complexity Term (Phase Transitions). This term captures the complexity of the transformer network in terms of the number of self-attention heads, depth, width, and the activation function used to define the class $\mathcal{TC}$. Each constant $C_1 \leqslant \ldots \leqslant C_s \leqslant \ldots$ collects the higher-order sensitivities (s^{th} order partial derivatives; where $s \in \mathbb{N}_+$) of the transformer model. Each $0 = \tau_0 \leqslant \tau_1 \leqslant \ldots \leqslant \tau_s \leqslant$ indicates the times at which there is a phase-transition in the convergence rate of the generalization bound accelerates. Once $t \geqslant \tau_s$, then the convergence rate of Term (II) accelerates, roughly speaking, by a reciprocal log-factor of $1/\log(N)$. Observe that the rate function is asymptotically equal to the rate function from the central limit theorem, as s tends to infinity; that is, $\lim_{s\to\infty} \text{rate}_s(N) = 1/(c\sqrt{N})$. The rate (rate) is the (optimal) rate at which the empirical measure generated by observations from a Markov process converges to its stationary distribution in 1-Wasserstein distance Kloeckner (2020); Riekert (2022). The polylogarithmic factor is removable if the data is i.i.d. Graf and Luschgy (2000); Dereich et al. (2013).

Term (III): Probabilistic Validity Term. This term captures the cost of the bound being valid with probability at least $1 - \delta$. The convergence rate of this term cannot be improved due to the central limit theorem. It is responsible for the overall convergence rate of our generalization bound being "stuck" at the optimal rate of $\mathcal{O}(1/\sqrt{N})$ from the central limit theorem; as the other two terms converge exponentially to 0.

4. Discussion: Implications of architecture choices.

Figure 3 illustrates the effect of various building blocks in the construction of a transformer (e.g. activation choice, multi-head attention (MHA), layernorms) through their effect on the constants in our generalization bounds. While Tables 1 to 5 contain more details, highlight here some key implications that architecture choices have on the bound:

I) Choice of Activation Function: We found (see Theorems 29, 31, 33 and 35) that the C^s-bounds of activation function may vary substantially, framing `softplus` and `swish` as the more regular, and `tanh` resulting in the highest bound. This is also illustrated in Figure 3. Note that the activation bound impacts the $\mathcal{PL}$-bound linearly and therefore effects the transformer-block bound of order s^2.

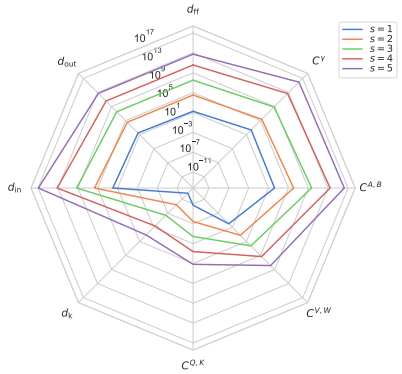

Figure 5: Absolute changes in C^s-bound for changes in architecture. Changes in dimensions ($d_\cdot$) are $\times 2$, while changes in parameter-bounds ($C^\cdot$) are $\times 10$, from the base parameters (see Tables 1 to 5).

II) Effects of Three Different Block-Types: Considering the three components – $\mathcal{MH}, \mathcal{LN}, \mathcal{PL}$ – that make up a transformer block, we observe that for low s the regularization by $\mathcal{LN}$ has the highest bound, but becomes less relevant with the exponential increase of the $\mathcal{MH}, \mathcal{PL}$-bounds for larger s. Exemplary values are shown in Figure 3.

III) Weight Size for MLP vs. Multi-Head Attention: As evident in Figure 5 (and Table 4), the parameter-bounds on $\mathcal{PL}$ (denoted by $C^{A,B}$) seem to have a more substantial impact on the bound than the parameter bounds of $\mathcal{MH}$. For the latter, bounds on key- and query-matrices ($C^{K,Q}$) seem to have bigger impacts for lower s than value- and aggregation-matrices ($C^{V,W}$) (see Theorem 10 for

details on notation), however show larger growth rates for larger s, as also shown in Table 5. The simple local-Lipschitz case of this result was obtained in Kim et al. (2021) (however, the complicated higher-order version of this result is completely new).

IV) Effect of Dimensions (Key, Input, etc...): Eventually, we can examine how various dimensions effect the bound. The input dimension (d_{in}) has a slightly higher impact than the output dimension (d_{out}). When it comes to choosing latent dimensions, scaling the hidden dimension of the $\mathcal{PL}$ (d_{ff}), has an effect similar to changes in the output dimension, and substantially higher comparing to the key-dimension d_K (see Theorems 10 and 11 in Appendix C for details and notation).

Consequentially, we show the effect on the phase-transition times $(\tau_t)_{t=0}^{\infty}$, defined in Theorem 3, dictating when the bound accelerates by a polylogarithmic factor in N.

Acknowledgments

AK acknowledges financial support from an NSERC Discovery Grant No. RGPIN-2023-04482 and their McMaster Startup Funds. RS is supported by Canada NSERC CGS-D Doctoral Grant. RS and AK acknowledge that resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute https://vectorinstitute.ai/partnerships/current-partners/.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral probability metrics pac-bayes bounds. *Advances in Neural Information Processing Systems*, 35:3123–3136, 2022.

Yves F Atchadé. Approximate spectral gaps for markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3):854–872, 2021.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2012. ISBN 978-3-642-25982-1. doi: 10.1007/978-3-642-25983-8. URL https://doi.org/10.1007/978-3-642-25983-8.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions, 2023.

Eric Temple Bell. Exponential polynomials. *Annals of Mathematics*, pages 258–277, 1934.

J Benitez, Takashi Furuya, Florian Faucher, Anastasis Kratsios, Xavier Tricoche, and Maarten V de Hoop. Out-of-distributional risk bounds for neural operators with applications to the helmholtz equation. *arXiv preprint arXiv:2301.11509*, 2023.

Bernard Bercu, Bernard Delyon, and Emmanuel Rio. *Concentration inequalities for sums and martingales*. SpringerBriefs in Mathematics. Springer, Cham, 2015. ISBN 978-3-319-22098-7; 978-3-319-22099-4. doi: 10.1007/978-3-319-22099-4. URL https://doi.org/10.1007/978-3-319-22099-4.

S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.*, 163(1):1–28, 1999. ISSN 0022-1236,1096-0783. doi: 10.1006/jfan.1998.3326. URL https://doi.org/10.1006/jfan.1998.3326.

Fran¸cois Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2012.07.007. URL https://doi.org/10.1016/j.jfa.2012.07.007.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities, 2013. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Khristo N. Boyadzhiev. Derivative polynomials for tanh, tan, sech and sec in explicit form. *Fibonacci Quarterly*, 45(4):291–303, 2007.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: biasing gradient descent into wide valleys*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124018, dec 2019. doi: 10.1088/1742-5468/ab39d9. URL https://dx.doi.org/10.1088/1742-5468/ab39d9.

Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-Sobolev inequalities for mixture distributions. *J. Funct. Anal.*, 281(11):Paper No. 109236, 17, 2021. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2021.109236. URL https://doi.org/10.1016/j.jfa.2021.109236.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. Characterizing overfitting in kernel ridgeless regression through the eigenspectrum, 2024a.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. A comprehensive analysis on the learning curve in kernel ridge regression, 2024b.

Samuel N. Cohen and Robert J. Elliott. *Stochastic calculus and applications*. Probability and its Applications. Springer, Cham, second edition, 2015. ISBN 978-1-4939-2866-8; 978-1-4939-2867-5. doi: 10.1007/978-1-4939-2867-5. URL https://doi.org/10.1007/978-1-4939-2867-5.

G. M. Constantine and T. H. Savits. A multivariate Faà di Bruno formula with applications. *Trans. Amer. Math. Soc.*, 348(2):503–520, 1996. ISSN 0002-9947. doi: 10.1090/S0002-9947-96-01501-2. URL https://doi.org/10.1090/S0002-9947-96-01501-2.

Jonathan D Cryer and Natalie Kellet. *Time series analysis*. Springer, 1991.

Christa Cuchiero and Josef Teichmann. Markovian lifts of positive semidefinite affine volterra-type processes. *Decisions in Economics and Finance*, 42:407–448, 2019.

Christa Cuchiero and Josef Teichmann. Generalized feller processes and markovian lifts of stochastic volterra processes: the affine case. *Journal of evolution equations*, 20(4):1301–1348, 2020.

Giuseppe Da Prato. *Introduction to stochastic analysis and Malliavin calculus*, volume 7. Edizioni della Normale, Pisa, second edition, 2008.

Mauri Aparecido de Oliveira and Ricardo Hirata Ikeda. Representation of the n-th derivative of the normal pdf using bernoulli numbers and gamma function. *Applied Mathematical Sciences*, 6(74): 3661–3673, 2012.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(4):1183–1203, 2013. ISSN 0246-0203,1778-7017. doi: 10.1214/12-AIHP489. URL https://doi.org/10.1214/12-AIHP489.

E. Dolera and E. Mainini. Lipschitz continuity of probability kernels in the optimal transport framework. *Ann. Inst. Henri Poincaré Probab. Stat.*, 59(4):1778–1812, 2023. ISSN 0246-0203,1778-7017. doi: 10.1214/23-aihp1389. URL https://doi.org/10.1214/23-aihp1389.

Karthik Duraisamy. Finite sample analysis and bounds of generalization error of gradient descent in in-context linear regression. *arXiv preprint arXiv:2405.02462*, 2024.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Francesco Faa di Bruno. Sullo sviluppo delle funzioni. *Annali di scienze matematiche e fisiche*, 6(1): 479–480, 1855.

Charles L. Fefferman. A sharp form of Whitney's extension theorem. *Ann. of Math. (2)*, 161 (1):509–577, 2005. ISSN 0003-486X,1939-8980. doi: 10.4007/annals.2005.161.509. URL https://doi.org/10.4007/annals.2005.161.509.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Nathael Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *Ann. Probab.*, 37(6):2480–2498, 2009. ISSN 0091-1798,2168-894X. doi: 10.1214/09-AOP470. URL https://doi.org/10.1214/09-AOP470.

Nathael Gozlan. Poincaré inequalities and dimension free concentration of measure. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(3):708–739, 2010. ISSN 0246-0203,1778-7017. doi: 10.1214/09-AIHP209. URL https://doi.org/10.1214/09-AIHP209.

Nathael Gozlan, Cyril Roberto, and Paul-Marie Samson. From dimension free concentration to the Poincaré inequality. *Calc. Var. Partial Differential Equations*, 52(3-4):899–925, 2015. ISSN 0944-2669,1432-0835. doi: 10.1007/s00526-014-0737-6. URL https://doi.org/10.1007/s00526-014-0737-6.

Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. ISBN 3-540-67394-6. doi: 10.1007/BFb0103945. URL https://doi.org/10.1007/BFb0103945.

Leonard Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975. ISSN 0002-9327,1080-6377. doi: 10.2307/2373688. URL https://doi.org/10.2307/2373688.

Michael Hardy. Combinatorics of partial derivatives. *Electron. J. Combin.*, 13(1):Research Paper 1, 13, 2006. doi: 10.37236/1027. URL https://doi.org/10.37236/1027.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GeLUs). *arXiv preprint arXiv:1606.08415*, 2016.

Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Andreas Krause, and Jonas Rothfuss. Instance-dependent generalization bounds via optimal transport. *J. Mach. Learn. Res.*, 24:Paper No. [349], 51, 2023a. ISSN 1532-4435,1533-7928.

Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Andreas Krause, and Jonas Rothfuss. Instance-dependent generalization bounds via optimal transport. *Journal of Machine Learning Research*, 24:1–50, 2023b.

Daniel J Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. *Advances in neural information processing systems*, 28, 2015.

James Inglis and Ioannis Papageorgiou. Log-Sobolev inequalities for infinite-dimensional Gibbs measures with non-quadratic interactions. *Markov Process. Related Fields*, 25(5):879–897, 2019. ISSN 1024-2953.

Oleksiy Khorunzhiy. On asymptotic properties of bell polynomials and concentration of vertex degree of large random graphs. *Journal of Theoretical Probability*, 35:1–32, 03 2022. doi: 10.1007/s10959-020-01025-w.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5562–5571. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *ICML 2024 Workshop on In-Context Learning*, 2024. URL https://openreview.net/forum?id=WjrKBQTWKp.

Benoît Kloeckner. Effective berry–esseen and concentration bounds for markov chains with a spectral gap. *The Annals of Applied Probability*, 29(3):1778–1807, 2019.

Benoît R. Kloeckner. Empirical measures: regularity is a counter-curse to dimensionality. *ESAIM Probab. Stat.*, 24:408–434, 2020. ISSN 1292-8100,1262-3318. doi: 10.1051/ps/2019025. URL https://doi.org/10.1051/ps/2019025.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 28–36, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/kontorovicha14.html.

Ioannis Kontoyiannis and Sean P Meyn. Geometric ergodicity and the spectral gap of non-reversible markov chains. *Probability Theory and Related Fields*, 154(1):327–339, 2012.

Anastasis Kratsios, A. Martina Neuman, and Gudmund Pammer. Tighter generalization bounds on digital computers via discrete optimal transport, 2024.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 2006.

Michel Ledoux, Ivan Nourdin, and Giovanni Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geom. Funct. Anal.*, 25(1):256–306, 2015. ISSN 1016-443X,1420-8970. doi: 10.1007/s00039-015-0312-0. URL https://doi.org/10.1007/s00039-015-0312-0.

George G. Lorentz, Manfred v. Golitschek, and Yuly Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-57028-4. doi: 10.1007/978-3-642-60932-9. URL https://doi.org/10.1007/978-3-642-60932-9. Advanced problems.

Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2024.

Dejun Luo and Jian Wang. Exponential convergence in lp-Wasserstein distance for diffusion processes without uniformly dissipative drift. *Math. Nachr.*, 289(14-15):1909–1926, 2016. ISSN 0025-584X,1522-2616. doi: 10.1002/mana.201500351. URL https://doi.org/10.1002/mana.201500351.

Dejun Luo and Jian Wang. Refined basic couplings and Wasserstein-type distances for SDEs with Lévy noises. *Stochastic Process. Appl.*, 129(9):3129–3173, 2019. ISSN 0304-4149,1879-209X. doi: 10.1016/j.spa.2018.09.003. URL https://doi.org/10.1016/j.spa.2018.09.003.

Miloud Mihoubi. Bell polynomials and binomial type sequences. *Discrete Mathematics*, 308(12): 2450–2459, 2008.

Ali A. Minai and Ronald D. Williams. On the derivatives of the sigmoid. *Neural Networks*, 6(6):845–853, 1993. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80129-7. URL https://www.sciencedirect.com/science/article/pii/S0893608005801297.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29, 2016.

Volodimir Mitarchuk, Clara Lacroce, Rémi Eyraud, Rémi Emonet, Amaury Habrard, and Guillaume Rabusseau. Length independent pac-bayes bounds for simple rnns. In *International Conference on Artificial Intelligence and Statistics*, pages 3547–3555. PMLR, 2024.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

Ravi Montenegro, Prasad Tetali, et al. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006.

Chen Mufa. Estimation of spectral gap for markov chains. *Acta Mathematica Sinica*, 12(4):337–360, 1996.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Christian Olivera and Ciprian Tudor. Density for solutions to stochastic differential equations with unbounded drift. *Brazilian Journal of Probability and Statistics*, 33(3):520–531, 2019.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864, 2009. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2008.11.001. URL https://doi.org/10.1016/j.jfa.2008.11.001.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:no. 79, 32, 2015. ISSN 1083-6489. doi: 10.1214/EJP.v20-4039. URL https://doi.org/10.1214/EJP.v20-4039.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Adrian Riekert. Convergence rates for empirical measures of Markov chains in dual and Wasserstein distances. *Statist. Probab. Lett.*, 189:Paper No. 109605, 8, 2022. ISSN 0167-7152,1879-2103. doi: 10.1016/j.spl.2022.109605. URL https://doi.org/10.1016/j.spl.2022.109605.

Walter Rudin. *Principles of mathematical analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, third edition, 1976.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

James B Simon, Madeline Dickens, Dhruva Karkada, and Michael Deweese. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *Transactions on Machine Learning Research*, 2023.

Chen Siyu, Sheen Heejune, Wang Tianhao, and Yang Zhuoran. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4573–4573. PMLR, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. *arXiv preprint arXiv:2310.13088*, 2023.

François Treves. *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25*, volume 25. Elsevier, 2016.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

A. W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes—with applications to statistics*. Springer Series in Statistics. Springer, Cham, second edition, 2023. ISBN 978-3-031-29038-1; 978-3-031-29040-4. doi: 10.1007/978-3-031-29040-4. URL https://doi.org/10.1007/978-3-031-29040-4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Feng-Yu Wang. Exponential ergodicity for singular reflecting McKean-Vlasov SDEs. *Stochastic Process. Appl.*, 160:265–293, 2023. ISSN 0304-4149,1879-209X. doi: 10.1016/j.spa.2023.03.009. URL https://doi.org/10.1016/j.spa.2023.03.009.

Weiping Wang and Tianming Wang. General identities on bell polynomials. *Computers & Mathematics with Applications*, 58(1):104–118, 2009.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682, 2022.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)*, 62:548–564, 1955. ISSN 0003-486X. doi: 10.2307/1970079. URL https://doi.org/10.2307/1970079.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. II. *Ann. of Math. (2)*, 65:203–207, 1957. ISSN 0003-486X. doi: 10.2307/1969956. URL https://doi.org/10.2307/1969956.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In *Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.

Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). *Advances in Neural Information Processing Systems*, 35:703–715, 2022.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Behnoosh Zamanlooy. On the mixing times of contractive markov kernels. *ArXiV*, 2024.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a.

Ruiqi Zhang, Jingfeng Wu, and Peter L Bartlett. In-context learning of a linear transformer block: benefits of the mlp component and one-step gd initialization. *arXiv preprint arXiv:2402.14951*, 2024b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *ArXiv preprint*, abs/2303.18223, 2023.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

David Zimmermann. Logarithmic Sobolev inequalities for mollified compactly supported measures. *J. Funct. Anal.*, 265(6):1064–1083, 2013. ISSN 0022-1236,1096-0783. doi: 10.1016/j.jfa.2013.05.029. URL https://doi.org/10.1016/j.jfa.2013.05.029.

Appendix Contents

Appendix A. Notation

In this section, we present the notation that will be employed throughout the appendix. This notation builds upon the framework established in the main body of the text, while incorporating additional levels of specificity. Given the technical nature of certain results discussed herein, a more detailed and precise formulation of the notation is necessary to ensure clarity and rigor in the statements that follow.

Notation 1 (Multi-index Notation) *We will fix the following multivariate notation.*

- *Multi-indices* $\alpha \overset{\text{def.}}{=} (\alpha_1,\dots,\alpha_k) \in \mathbb{N}^k, k \in \mathbb{N}$ *are denoted by Greek letters.*

- *The sum of entries is given by* $|\alpha| \overset{\text{def.}}{=} \sum_{i=1}^{k} \alpha_k$.
- *Its faculty is defined by* $\alpha! \overset{\text{def.}}{=} \prod_{i=1}^{k} \alpha_k!$,
- *We denote the derivative w.r.t.* α *by* $D^{\alpha} \overset{\text{def.}}{=} \partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_k^{\alpha_k}$ *if* $|\alpha| > 0$ *else* D^{α} *is the identity operator.*
- *For a vector* $x \in \mathbb{R}^k$*, we write* $x^{\alpha} \overset{\text{def.}}{=} \prod_{i=1}^{k} x_k^{\alpha_k}$.
- *We define the relation* $\alpha \prec \beta$ *for* $\beta \in \mathbb{N}^k$ *if one of the three following holds*
  - *(i)* $|\alpha| < |\beta|$*;*
  - *(ii)* $|\alpha| = |\beta|$*, and* $\alpha_1 < \beta_1$*; or*
  - *(iii)* $|\alpha| = |\beta|$*, and* $\alpha_i = \beta_i$ *for* $i \in \{1, \ldots, j-1\}$ *and* $\alpha_j < \beta_j$ *for* $j \in \{2, \ldots, k\}$.
- *Unit vectors* $e_i \in \{0,1\}^k$ *are defined by* $(e_i)_j = 0$ *for* $i \neq j$ *and* $(e_i)_i = 1$.

Definition 4 (C^s-norm) *For any* $s > 0$*, the norm* $\|\cdot\|_{C^s}$ *of a smooth function* $f : \mathbb{R}^d \to \mathbb{R}$ *is defined by*

$$\|f\|_{C^s} \overset{\text{def.}}{=} \max_{k=1,\ldots,s-1} \max_{\alpha \in \{1,\ldots,d\}^d} \left\| \frac{\partial^k f}{\partial x_{\alpha_1} \ldots \partial x_{\alpha_k}} \right\|_{\infty} + \max_{\alpha \in \{1,\ldots,d\}^{s-1}} \operatorname{Lip}\left(\frac{\partial^{s-1} f}{\partial x_{\alpha_1} \ldots \partial x_{\alpha_{s-1}}} \right).$$

We use the following notation to streamline the analytic challenges the tacking of C^s-norms.

Notation 2 (Order operator for multi-indeces) *Define the order operator* $\mathfrak{o}$ *for multi-indeces by*

$$\mathfrak{o} : \mathbb{N}^k \longrightarrow \mathbb{N}^k, \quad \alpha_1, \ldots, \alpha_k \longmapsto \alpha_{\tau_\alpha(1)}, \ldots, \alpha_{\tau_\alpha(k)},$$

where $\tau_\alpha : \{1, \ldots, k\} \to \{1, \ldots, k\}$ *s.t.* $\alpha_{\tau_\alpha(1)} \geqslant \ldots \geqslant \alpha_{\tau_\alpha(k)}$*. We write* $\alpha \sim \beta$ *if* $\mathfrak{o}(\alpha) = \mathfrak{o}(\beta)$ *for* $\alpha, \beta \in \mathbb{N}^k$*. Further, denote by* $\mathfrak{O}_n^k$ *the set* $\{\mathfrak{o}(\alpha) : \alpha \in \mathbb{N}^k, |\alpha| = n\}$ *and write* $\mathfrak{O}_{\leqslant n}^k \overset{\text{def.}}{=} \{\mathfrak{o}(\alpha) : \alpha \in \mathbb{N}^k, |\alpha| \leqslant n\}$*. Eventually, define* $N(\alpha) \overset{\text{def.}}{=} \#|\{\alpha' \in \mathbb{N}^k : \mathfrak{o}(\alpha') = \alpha\}|$.

We will use the following notation to tabulate the sizes of a C^s-norm.

Notation 3 (Derivatives) *Let* $k \in \mathbb{N}$*,* $K \in \mathbb{R}^k$ *be a set,* $f : K \to \mathbb{R}^m$ *a function and* $\alpha \in \mathfrak{O}_n^k$ *an ordered multi-index. Then,*

- *the uniform bound of* α*-like derivatives on* K *is given by*

$$C_K^f(\alpha) \overset{\text{def.}}{=} \max_{i \in \{1,\ldots,m\}} \max_{\gamma \sim \alpha} \|D^{\gamma} f_i\|_K,$$

- *we define the bound at / up to derivative level* n *by*

$$C_K^f(n) \overset{\text{def.}}{=} \max_{\alpha \in \mathfrak{O}_n^k} C_K^f(\alpha), \qquad C_K^f(\leqslant n) \overset{\text{def.}}{=} \max_{\alpha \in \mathfrak{O}_{\leqslant n}^k} C_K^f(\alpha),$$

- *we write* $\|K\| \overset{\text{def.}}{=} \sup_{x \in K} \|x\|$*, and*
- *the* ℓ^{∞}*-matrix norm of any* $n \times m$ *matrix* $A \in \mathbb{R}^{n \times m}$ *is abbreviated as*

$$C^A \overset{\text{def.}}{=} \max_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,m\}} |A_{i,j}|.$$

When segmenting, truncating, or manipulating time series we will using the following notation.

Notation 4 (Time Series Notation) *The following notation is when indexing paths of any time series.*

- Realized Path up to time t *is denoted by* $x_{\leqslant t} \overset{\text{def.}}{=} (x_s)_{s\in\mathbb{Z},\, s\leqslant t}$.
- Segment of a Path *Given a sequence* $x \in \mathbb{R}^{\mathbb{Z}}$ *and integers* $s \leqslant t$, *we denote* $x_{[s:t]} \overset{\text{def.}}{=} (x_i)_{i=s}^{t}$.

Lastly, we recorded some additional notations that were required throughout our manuscript.

Notation 5 (Miscellaneous) *We define:*

- N-Simplex. *For* $N \in \mathbb{N}$ *we write*

$$\Delta_N \overset{\text{def.}}{=} \{u \in [0,1]^N : \sum_{i=1}^{N} u_i = 1\}.$$

- Infinite powers*: For* $c \in (0,1)$, *we define*

$$c^{\infty} \overset{\text{def.}}{=} \lim_{t\to\infty} c^t = 0.$$

- Reshape operator*: For any* $F_1, F_2 \in \mathbb{N}_+$, *the operator is given by* $\operatorname{reshape}_{F_1\times F_2}$, *mapping any vector* $u \in \mathbb{R}^{F_1F_2}$ *to the* $F_1 \times F_2$ *matrix*

$$\operatorname{reshape}_{F_1\times F_2}(x)_{i,j} \overset{\text{def.}}{=} x_{(i-1)F_2+j}.$$

  We denote the inverse of the map $\operatorname{reshape}_{F_1\times F_2}$ *by* $\operatorname{vec}_{F_1,F_2} : \mathbb{R}^{F_1\times F_2} \to \mathbb{R}^{F_1F_2}$.
- Softmax operator*: For each* $F \in \mathbb{N}_+$ *and each* $x \in \mathbb{R}^F$,

$$\operatorname{softmax}(x) \overset{\text{def.}}{=} \operatorname{smax}(x) \overset{\text{def.}}{=} (\exp(x_i)/\textstyle\sum_{j=0}^{F-1}\exp(x_j))_{i=0}^{F-1}.$$

Appendix B. Examples of Data-Generating Processes Satisfying Assumptions 1 and 2

This section provides several examples of stochastic (data-generating) processes which satisfy our assumptions and are outside the i.i.d. restrictions.

B.1. Projected Exponentially Ergodic Latent Processes

Proposition 5 (Lipschitz-Transformed SDEs with Overdampened Drift) *In the setting of Example 1,* $\{(P^n(x,\cdot))_{n=0}^{\infty}\}_{x\in[0,1]^d}$ *satisfies both Assumptions 1 and 2.*

The proof of Proposition 5 uses the following lemma.

Lemma 6 (Enforcing Boundedness via 1-Lipschitz Maps Preserves Exponential Ergodicity) *Let* $\tilde{d}, d \in \mathbb{N}_+$ *and* $Z.$ *be a Markov process on* $\mathbb{R}^{\tilde{d}}$ *satisfying Assumption 2. Given any bounded Lipschitz function* $f : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^d$ *the Markov process* $X. \overset{\text{def.}}{=} (X_n)_{n=0}^{\infty}$ *in* $\mathbb{R}^d$, *defined for each* n *by* $X_n \overset{\text{def.}}{=} f(Z_n)$, *satisfies both Assumption 1 and 2.*

Proof [Proof of Lemma 6] Since f is bounded, then there exists some $r > 0$ such that $f(\mathbb{R}^d) \subset B_r^d \overset{\text{def.}}{=} \{u \in \mathbb{R}^d : \|u\| \leqslant\}$. For each $x \in \mathbb{N}_+$, let $P^n(x,\cdot) \overset{\text{def.}}{=} \mathbb{P}(X_t \in \cdot | X_0 = x) = \mathbb{P}(f(Z_t) \in \cdot | f(Z_0) = f(x)) = f_{\#}P^n(x,\cdot)$ then the Kantorovich duality, see e.g. (Villani, 2009, Theorem 5.10), implies that $f_{\#} : \mathcal{P}_1(\mathbb{R}^d) \to \mathcal{P}_1(B_r^d)$ is 1-Lipschitz; whence (2) imples that: for each $x \in [0,1]^d$ and every $n \in \mathbb{N}$ we have

$$\mathcal{W}_1\big(P^n(x,\cdot), P^n(y,\cdot)\big) \leqslant \operatorname{Lip}(f)\mathcal{W}_1\big(\tilde{P}^n(x,\cdot), \tilde{P}^n(y,\cdot)\big) \leqslant \kappa\,\|x-y\|. \tag{2}$$

Thus, Assumption 2 holds. Finally, we note that Assumption 1 holds since each $P^n(x,\cdot) \in \mathcal{P}_1(B_r^d)$. ■

Proof [Proof of Proposition 5] For any $\mu \in \mathcal{P}_1(\mathbb{R}^D)$ consider the unique strong solution (which exists by our Lipschitz assumption) For the following SDE (which is a Markov process)

$$Z_t^\mu = Z_0^\mu + \int_0^t \mu(Z_s^\mu)\, ds + \int_0^t W_s$$

where $W_\cdot \overset{\text{def.}}{=} (W_n)_{n=0}^\infty$ is a d-dimensional Brownian motion and Z_0^μ is distributed according to μ. For every $n \in \mathbb{N}_+$ let $\tilde{P}^n\mu \overset{\text{def.}}{=} \mathbb{P}(Z_n^\mu \in \cdot)$ and, for each $x \in \mathbb{R}^d$, let $\tilde{P}^n(x,\cdot) \overset{\text{def.}}{=} \tilde{P}^n\delta_x$. Then (Luo and Wang, 2016, Theorem 1.1) implies that: for all $n \in \mathbb{N}$ and each $\mu,\nu \in \mathcal{P}_1(\mathbb{R}^d)$ we have

$$\mathcal{W}_1\big(\tilde{P}^n\mu, \tilde{P}^n\nu\big) \leqslant \kappa\, \mathcal{W}_1(\mu,\nu) \tag{3}$$

where $\kappa = \exp(-K)$; note that $\kappa \in (0,1)$ since $K > 0$. That is, $(\tilde{P}^n)_{n=0}^\infty$ satisfies Assumption 2 upon taking $\mu = \delta_x$ and $\nu = \delta_y$, for any given $x, y \in \mathbb{R}^d$, since $\mathcal{W}_1(\delta_x,\delta_y) = \|x-y\|$, see e.g. (Villani, 2009, page 99 point 5) or note that the only coupling between δ_x and δ_y is the product measure $\delta_x \otimes \delta_y$).

$$\mathcal{W}_1\big(\tilde{P}^n(x,\cdot), \tilde{P}^n(x,\cdot)\big) \leqslant \kappa\, \|x-y\|.$$

Applying Lemma 6 yields the conclusion. ■

Proposition 7 *Consider the setting of Example 2. Then, the process $X_\cdot$ satisfies both Assumptions 1 and 2.*

Proof [Proof of Proposition 7] Under our assumptions σ satisfies (Wang, 2023, Assumption (A8) (1) and (A8) (3)). Therefore, the stochastic process $Z_\cdot \overset{\text{def.}}{=} (Z_t)_{t\geqslant 0}$ defined by

$$Z_t \overset{\text{def.}}{=} \int_0^t \sigma(Z_s)\, dW_s \tag{4}$$

where $W_\cdot$ is a d-dimensional Brownian motion, satisfies the conditions of (Wang, 2023, Corollary 4.4) from which we deduce that $Z_\cdot$ satisfies Assumption 2. Applying Lemma 6 yields the conclusion. ■

B.2. Markov Processes Satisfying a Log-Sobolev Inequalities

Our main result is equally valid under the assumption that the stationary distribution of the Markov chain and its kernels all satisfy a log-Sobolev inequality (LSI). Since their introduction, LSIs have been heavily studied Gross (1975); Ledoux et al. (2015); Zimmermann (2013); Inglis and Papageorgiou (2019); Chen et al. (2021) and have found numerous applications in differential privacy Minami et al. (2016); Ye and Shokri (2022), optimization Chaudhari et al. (2019), random matrix theory Wigner (1955, 1957), optimal transport Dolera and Mainini (2023), since they typically

imply Gozlan (2010); Gozlan et al. (2015) and effectively characterizes Gozlan (2009) dimension-free rate for concentration of measure. We define the *entropy functional* $\mathcal{H}_\mu$ associated to any Borel probability measure μ on $\mathbb{R}^d$ acts on smooth functions $g : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathbb{H}_\mu(g) \stackrel{\text{def.}}{=} \mathbb{E}_{X\sim\mu}\Big[g(X)\log\Big(\frac{g(X)}{\mathbb{E}_{Z\sim\mu}[g(Z)]}\Big)\Big].$$

The entropy functional can be used to express the log-Sobolev inequalities.

Definition 8 (Log-Sobolev Inequality) *A probability measure μ on $\mathbb{R}^d$ is said to satisfy a log-Sobolev inequality with constant $C > 0$ (LSI_C) if for every smooth function $g : \mathbb{R}^d \to \mathbb{R}$*

$$\mathbb{H}_\mu(g^2) \leqslant C\,\mathbb{E}_{X\sim\mu}[\|\nabla g(X)\|^2]$$

We require that the Markov process is time-homogeneous to admit a satisfactory measure. Further, we require that its Markov kernel and its stationary measure all satisfy LSI_C.

Assumption 3 (Satisfactions of the Log-Sobolev Inequality) *There exists a $C > 0$ such that $\bar{\mu}$, μ_0, and $P(x,\cdot)$ all satisfy LSI_C, for each $x \in \mathcal{X}$.*

Instead of the compact support Assumption 1 we may consider the following weaker condition.

Assumption 4 (Exponential Moments) *There exist $\lambda, \tilde{C} > 0$ and $\gamma \in (0,1)$ such that: for each $x \in \mathcal{X}$ we have $\mathbb{E}_{X\sim P(x,\cdot)}[e^{\lambda|X|}] \leqslant \gamma\, e^{\lambda|x|} + \tilde{C}$.*

Note that, Assumption 1 implies 4, but not conversely.

Several examples of Markov processes satisfying LSI inequalities are given in Ledoux (2006) and Gaussian processes satisfy the Exponential Moments Assumption.

Proposition 9 (Log-Sobolev Conditions and Exponential Moments Imply Assumption 2) *If Assumptions 4 and 3 hold then the process $X.$ satisfies Assumption 2.*

Proof [Proof of Proposition 9]

Under the log-Sobolev Assumption 3, (Bobkov and Götze, 1999, Theorem 1.3) can be applied to $\bar{\mu}$ and $P(x,\cdot)$ for each $x \in \mathcal{X}$, implying that the transport inequalities hold: for each $\nu \in \mathcal{P}(\mathcal{X})$ and each $\tilde{\mu} \in \{\bar{\mu}, \mu_0\} \cup \{P(x,\cdot)\}_{x\in\mathcal{X}}$

$$\mathcal{W}_1(\tilde{\mu},\nu)^2 \leqslant 2C^2\ \mathrm{KL}(\nu|\tilde{\mu}) \tag{5}$$

where we recall the definition of the Kullback–Leibler divergence $\mathrm{KL}(\nu|\mu) \stackrel{\text{def.}}{=} \mathbb{E}_{X\sim\nu}[\log(\frac{d\nu}{d\mu}(X))]$. Thus, (5) implies that the following exponential contractility property of the Markov kernel: there exists some $\kappa \in (0,1)$ such that for each $x, \tilde{x} \in \mathcal{X}$ and every $t \in \mathbb{N}_+$

$$\mathcal{W}_1\big(P^t(x,\cdot), P^t(\tilde{x},\cdot)\big) \leqslant \kappa^t\,\|x - \tilde{x}\|. \tag{6}$$

This completes the proof. ■

Appendix C. Transformer Definition Details

For any $F \in \mathbb{N}_+$, we will consider a weighted (parametric) variant of the *layer normalization* function of Ba et al. (2016), which permits a variable level of regularization. Our weighted *layer normalization* is defined by $\mathrm{LayerNorm} : \mathbb{R}^F \to \mathbb{R}^F$ defined for any $u \in \mathbb{R}^F$ by

$$\mathcal{LN}(u; \gamma, \beta, w) \overset{\text{def.}}{=} \gamma \frac{(u - \mu_u^w)}{\sqrt{1 + (\sigma_u^w)^2}} + \beta$$

where $\mu_u^w \overset{\text{def.}}{=} \sum_{i=1}^F \frac{w}{F} u_i$ and $(\sigma_u^w)^2 \overset{\text{def.}}{=} \sum_{i=1}^F \frac{w}{F} \|u_i - \mu_u\|^2$, $\mathrm{splus} \overset{\text{def.}}{=} \ln(1 + \exp(\cdot))$, parameters $\beta \in \mathbb{R}^F$ and $\gamma \in \mathbb{R}$, and the normalization strength parameter $w \in [0, 1]$ with $w = 1$ being the default choice. Here, we prohibit the layer norm from magnifying the size of its outputs when the layer-wise weighted variance σ_u^w is small.[2]

Definition 10 (Multi-Head Self-Attention) *Fix $d_{\text{in}} \in \mathbb{N}$. For $x \in \mathbb{R}^{M \times d_{\text{in}}}$, $Q, K \in \mathbb{R}^{d_K \times d_{\text{in}}}$, and $V \in \mathbb{R}^{d_V \times d_{\text{in}}}$, where we have key-dimension $d_K \in \mathbb{N}$ and value-dimension $d_V \in \mathbb{N}$; we define*

$$\mathrm{Att}(x; Q, K, V) \overset{\text{def.}}{=} \left(\sum_{j=0}^{M} \mathrm{softmax}\left(\left(\frac{\langle Q x_m, K x_i \rangle}{\sqrt{d_k}} \right)_{i=0}^{M} \right)_j V x_j \right)_{m=1}^{M} \in \mathbb{R}^{M \times d_V}.$$

For $H \in \mathbb{N}$, set $Q \overset{\text{def.}}{=} (Q^{(h)})_{h=1}^H, K \overset{\text{def.}}{=} (K^{(h)})_{h=1}^H \subseteq \mathbb{R}^{d_K \times d_{\text{in}}}$, $V \overset{\text{def.}}{=} (V^{(h)})_{h=1}^H \subseteq \mathbb{R}^{d_V \times d_{\text{in}}}$, and $W \overset{\text{def.}}{=} (W^{(h)})_{h=1}^H \subseteq \mathbb{R}^{d_{\text{in}} \times d_V}$. For $x \in \mathbb{R}^{M \times d_{\text{in}}}$, we define

$$\mathcal{MH}(x; Q, K, V, W) \overset{\text{def.}}{=} \left(\sum_{h=1}^{H} W^{(h)} \mathrm{Att}(x; Q^{(h)}, K^{(h)}, V^{(h)})_m \right)_{m=1}^{M} \in \mathbb{R}^{M \times d_{\text{in}}}.$$

Each transformer block takes a set of inputs and intersperses normalization via layer norms, contextual comparisons via multi-head attention mechanisms, and non-linear transformations via a single layer perceptron (SLP). We also allow the transformer block to extend or contract the length of the generated sequence.

Definition 11 (Transformer Block) *Fix a non-affine activation function $\sigma \in C^\infty(\mathbb{R})$. Fix a dimensional multi-index $d = (d_{\text{in}}, d_K, d_V, d_{\text{ff}}, d_{\text{out}}) \in \mathbb{N}^5$, a sequence length $M \in \mathbb{N}_+$, and a number of self-attention heads $H \in \mathbb{N}_+$. A transformer block is a permutation equivariant map $\mathcal{TB} : \mathbb{R}^{M \times d_{\text{in}}} \to \mathbb{R}^{M \times d_{\text{out}}}$ represented for each $x \in \mathbb{R}^{M \times d_{\text{in}}}$*

$$\begin{aligned} \mathcal{TB}(x) &\overset{\text{def.}}{=} \left(\mathcal{LN}\left(B^{(1)} x'_m + B^{(2)} \left(\sigma \bullet (A\, x'_m + a) \right); \gamma_2, \beta_2, w_2 \right) \right)_{m=1}^{M} \\ x' &\overset{\text{def.}}{=} \left(\mathcal{LN}\left(x_m + \mathcal{MH}(x; Q, K, V, W)_m; \gamma_1, \beta_1, w_1 \right) \right)_{m=1}^{M} \end{aligned} \tag{7}$$

for $\gamma_1, \gamma_2 \in \mathbb{R}$, $w_1, w_2 \in [0, 1]$, $\beta_1 \in \mathbb{R}^{d_{\text{in}}}$, $\beta_2 \in \mathbb{R}^{d_{\text{out}}}$, $A \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{in}}}$, $a \in \mathbb{R}^{d_{\text{ff}}}$, $B^{(1)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $B^{(2)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{ff}}}$, and Q, K, V, W as in Theorem 10. Above, we write $\bullet$ for a pointwise application.

The class of transformer blocks with representation (7) and bounds on $\gamma_1, \gamma_2, \beta_1, \beta_2, a, A, B^{(1)}, B^{(2)}, Q, K, V, W$ is denoted by $\mathcal{TBC}$.

2. Note that this formulation of the layer norm avoids division by 0 when the entries of u are identical.

A transformer concatenates several transformer blocks before passing their outputs to an affine layer and ultimately outputting its prediction.

Definition 12 (Transformers) *Fix depth $L \in \mathbb{N}_+$, memory $M \in \mathbb{N}$, width $W \in \mathbb{N}_+^5$, number of heads $H \in \mathbb{N}_+$, and input-output dimensions $D, d \in \mathbb{N}_+$. A transformer (network) is a map $\mathcal{T} : \mathbb{R}^{M\times D} \to \mathbb{R}^d$ with representation*

$$\mathcal{T}(x) = A\big(\operatorname{vec}_{1+M,d_{\text{out}}^L} \circ \mathcal{TB}_L \circ \cdots \circ \mathcal{TB}_1(x)\big) + b \tag{8}$$

where multi-indices $d^l = (d_{\text{in}}^l, d_K^l, d_V^l, d_{\text{ff}}^l, d_{\text{out}}^l) \leqslant W$ are such that $d_{\text{in}}^1 = D$, $d_{\text{in}}^{l+1} = d_{\text{out}}^l$ for each $l = 1, \ldots, L-1$, and where $H \overset{\text{def.}}{=} (H^l)_{l=1}^L$ are the number of self-attention heads, $C' \overset{\text{def.}}{=} (C^l)_{l=1}^L$ the parameter bounds, and for $l = 1, \ldots, L$ we have $\mathcal{TB}_l \in \mathcal{TBC}_l$, where $\mathcal{TBC}_l$ is a transformer block class with $d_{\text{in}} = d^l$, $M = M$, and $H = H^l$. Furthermore, $A \in \mathbb{R}^{d\times M d_{\text{out}}^L}$ and $b \in \mathbb{R}^d$.

The set of transformer networks with representation (8) and bounds on A, b is denoted by $\mathcal{TC}$.

Appendix D. Elucidation of Constants in Theorem 3

The aim of this section is to elucidate the magnitude of the constants appearing in Theorem 3. We aim of to make each of these concrete by numerically estimating them, which we report in a series of tables. Importantly, we see how subtle choices of the activation function used to define the transformer model can have dramatic consequences on the size of these constants, which could otherwise be hidden in big $\mathcal{O}$ notation.

Interestingly, in Tables 1 and 2, we see that the $\mathrm{softplus}$ activation function produces significantly tighter bounds than the $\tanh$ activation function through much smaller constants, and the GeLU and SWISH activation functions are a relatively comparable second-place.

The bounds depicted in Table 2 exhibits a notable trait of independence from both input dimension and the compactum they are defined on. Notably, the selection of latent dimensionality demonstrates a relatively minor influence in contrast to the pronounced impact of parameter bounds. This suggests that while adjusting the latent dimension may have some effect, the primary driver of the derivative bound lies within the constraints imposed on the parameters. Despite the seemingly conservative nature of the chosen parameter-bounds, it is important to acknowledge their alignment with the parameter ranges observed in trained transformer-models.

Note that the latter can be observed as well for Multi-Head attention (Table 5), however, we see that here the input dimension (composed of d_{in} and M) is of greater importance with respect to the derivative bound.

Table 1: C^s-bounds of activation functions based on numerical maximization of analytic derivatives in Appendix F.4.3.

Bound	softplus	GeLU	tanh	Swish
C^1	0.25	1.12	4.00	1.10
C^2	0.10	0.48	8.00	0.50
C^3	0.12	0.75	16.00	0.31
C^4	0.13	1.66	32.00	0.50
C^5	0.25	4.34	156.65	0.66
C^6	0.41	12.95	1651.32	1.50
C^7	1.06	42.77	20405.43	2.91
C^8	2.39	153.76	292561.95	8.50
C^9	7.75	594.17	4769038.09	21.76
C^{10}	22.25	2445.69	87148321.71	77.50

The bound of the layer-norm (see Table 3) seems to be particularly effected by the domain it is defined on, which can be problematic if it appears in later layers. An immediate solution is the usage of its parameter γ, a more drastic approach would be applications in combination with an upstream sigmoid activation.

Eventually, as also shown in Figure 3, we included in Tables 4 and 5 a comparison of using type-specific bounds (see Theorem 19) or level-specific bounds (Theorem 16) in the computation of the constants. This effect seems to become more evident with higher number of function compositions.

Table 2: Derivative Bounds of the Perceptron Layer by derivative level according to Theorem 39.

Parameters			Derivative Level				
σ	d_{ff}	$C^{\{A,B^{(1)},B^{(2)}\}}$	1	2	3	4	5
softmax	64	1.0	17.00	50.47	236.94	1.34E+03	1.33E+04
tanh			129.00	1.02E+03	8.96E+03	9.83E+04	1.34E+06
GeLU			73.25	237.41	1.25E+03	1.16E+04	1.81E+05
SWISH			71.39	236.78	870.75	5.33E+03	4.13E+04
	16		5.00	12.62	59.23	334.15	3.31E+03
	32		9.00	25.24	118.47	668.30	6.63E+03
	128		33.00	100.95	473.87	2.67E+03	2.65E+04
	256		65.00	201.90	947.75	5.35E+03	5.30E+04
		0.01	17.00	44.32	180.94	919.87	6.52E+03
		0.1	17.00	44.38	181.00	919.92	6.52E+03
		10.0	17.00	660.15	5.62E+04	4.17E+06	6.73E+08
		100.0	17.00	6.16E+04	5.60E+07	4.17E+10	6.73E+13

Table 3: Layer Norm

Parameters			Derivative Level				
k	$\|K\|$	γ	1	2	3	4	5
5	10.0	0.1	18.67	28.56	104.49	1.49E+03	4.93E+03
3			18.67	28.56	104.49	945.21	4.93E+03
10			18.67	28.56	104.49	1.49E+03	4.93E+03
20			18.67	28.56	104.49	1.49E+03	4.93E+03
	0.1		0.17	3.61	5.37	7.05	8.87
	1.0		1.73	5.20	6.95	8.71	10.64
	100.0		321.71	7.68E+03	7.88E+05	1.42E+08	4.39E+09
	1000.0		321.71	7.68E+03	7.88E+05	1.42E+08	4.39E+09
		0.01	1.73	2.07	2.24	2.42	2.59
		1.0	321.71	7.75E+03	7.91E+05	1.42E+08	4.44E+09

Table 4: Derivative Bounds of Transformer Block by derivative level according to Theorem 42.

Parameters				Derivative Level				
d_{in}	$C^{\{K,Q,V,W\}}$	$C^{\{A,B^{(1,2)}\}}$	γ	1	2	3	4	5
5	0.01	0.001	0.01	21.15	1.13E+04	4.81E+06	2.59E+09	2.22E+11
	— using derivative level —			212.70	1.53E+06	7.47E+09	4.55E+13	3.75E+16
10				111.32	4.51E+05	1.71E+09	1.45E+13	8.70E+16
20				1.29E+03	1.25E+08	3.47E+13	1.85E+19	2.20E+24
	0.001			21.15	1.13E+04	4.81E+06	2.59E+09	2.22E+11
	0.1			21.16	1.13E+04	4.83E+06	2.61E+09	2.32E+11
	1.0			22.30	4.64E+04	6.96E+08	1.70E+13	1.95E+17
		0.0001		5.05	126.27	1.12E+04	4.94E+06	6.87E+08
		0.01		182.17	1.12E+06	4.66E+09	2.43E+13	1.71E+16
		0.1		1.79E+03	1.12E+08	4.65E+12	2.42E+17	1.70E+21
			0.0001	0.21	108.21	4.44E+04	2.27E+07	1.58E+09
			0.001	2.09	1.09E+03	4.46E+05	2.29E+08	1.60E+10
			0.1	240.09	2.45E+05	7.31E+08	4.96E+12	8.79E+15

Table 5: Derivative Bounds of Multi-Head Attention by derivative level according to Theorem 28.

Parameters								Derivative Level				
d_{in}	M	d_K	C^K	C^Q	C^V	C^W	$\|K\|$	1	2	3	4	5
5	1	3	0.1	0.1	0.1	0.1	1.0	7.67	46.08	184.82	931.81	5.73E+03
— using derivative level —								7.67	46.15	186.90	1.01E+03	7.84E+03
10								7.82E+03	4.87E+04	8.14E+05	2.95E+08	2.38E+11
20								1.57E+04	1.09E+05	1.03E+07	9.25E+09	1.50E+13
	5							3.90E+03	2.37E+04	1.34E+05	9.96E+06	3.83E+09
	20							4.71E+06	2.85E+07	7.11E+08	2.66E+12	1.40E+16
		3						7.67	45.99	183.98	920.53	5.53E+03
		10						7.70	46.41	190.35	1.07E+03	1.03E+04
		20						7.75	47.52	226.02	2.87E+03	1.23E+05
			0.01					7.65	45.91	183.61	918.01	5.51E+03
			1.0					7.90	54.45	740.55	6.51E+04	9.33E+06
				0.01				7.65	45.91	183.61	918.01	5.51E+03
				1.0				7.90	54.45	740.55	6.51E+04	9.33E+06
					0.01			0.77	4.61	18.48	93.18	573.32
					1.0			76.75	460.80	1.85E+03	9.32E+03	5.73E+04
						0.01		0.77	4.61	18.48	93.18	573.32
						1.0		76.75	460.80	1.85E+03	9.32E+03	5.73E+04
			0.001	0.001	0.001	0.001		0.00	0.00	0.02	0.09	0.55
			0.01	0.01	0.01	0.01		0.08	0.46	1.84	9.18	55.08
			1.0	1.0	1.0	1.0		1.02E+03	8.06E+04	4.95E+07	5.70E+10	8.04E+13
							0.1	0.77	32.14	142.34	752.79	4.68E+03
							10.0	79.00	259.65	5.54E+03	5.73E+05	8.04E+07
							100.0	1.02E+03	7.66E+04	4.88E+07	5.63E+10	7.90E+13

Appendix E. Supporting Technical Results on the C^s-Norms of Smooth Functions

This section contains many of the technical tools on which we build our analysis. Most results concern smooth functions, especially their derivatives and those of compositions thereof. However, the first set of results concerns the integral probability metric d_s.

E.1. Integral Probability Metrics and Restriction to Compact Sets

Fix $d \in \mathbb{N}_+$ and a non-empty compact subset $K \subseteq \mathbb{R}^d$. Observe that any Borel probability measure μ on K can be canonically extended to a compactly supported Borel probability measure μ^+ on all of $\mathbb{R}^d$ via

$$\mu^+(B) \overset{\text{def.}}{=} \mu(B \cap K),$$

for any Borel subset B of $\mathbb{R}^d$; noting only that $B \cap K$ is Borel.

Let $\mathcal{P}(K)$ denote the set of Borel probability measures on K. Suppose that K is a regular compact set, i.e. the closure of its interior is itself. As usual, see Evans (2022), for any $s \in \mathbb{N}_+$, we denote the set of functions from the interior of K to $\mathbb{R}$ with s continuous partial derivatives thereon and with a continuous extension to K by $C^s(K)$. This space, is a Banach space when equipped with the (semi-)norm

$$\|f\|_{s:K} \overset{\text{def.}}{=} \max_{k=1,\dots,s-1} \max_{\alpha\in\{1,\dots,d\}^k} \sup_{u\in K} \left\| \frac{\partial^k f}{\partial x_{\alpha_1}\dots\partial x_{\alpha_k}}(u) \right\| + \max_{\alpha\in\{1,\dots,d\}^{s-1}} \operatorname{Lip}\left(\frac{\partial^{s-1} f}{\partial x_{\alpha_1}\dots\partial x_{\alpha_{s-1}}} \right).$$

We may define an associated integral probability metric $d_{s:K}$ on $\mathcal{P}(K)$ via

$$d_{s:K}(\mu,\nu) \overset{\text{def.}}{=} \sup_{f\in C^s(K)} \|\mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{X\sim\nu}[f(X)]\|$$

for any $\mu, \nu \in \mathcal{P}(K)$. The main purpose of this technical subsection is simply to reassure ourselves, and the reader, that quantities $d_{s:K}(\mu,\nu)$ and $d_s(\mu^+,\nu^+)$ are equal for any $\mu,\nu \in \mathcal{P}(K)$. Therefore, we may use them interchangeably.

Lemma 13 (Consistency of Smooth IMP Extension - Beyond Regular Compact Sets) *Fix $d, s \in \mathbb{N}_+$ and let K be a non-empty regular compact subset of $\mathbb{R}^d$. For any $\mu,\nu \in \mathcal{P}(K)$ the following holds*

$$d_{s:K}(\mu,\nu) = d_s(\mu^+,\nu^+).$$

Proof Let $\operatorname{int}(K)$ denote the interior of K, By the Whitney extension theorem, as formulated in (Fefferman, 2005, Theorem A), for any $f \in C^s(K)$ there exists a C^s-extension $F : \mathbb{R}^d \to \mathbb{R}$ of $f|_{\operatorname{int}(K)}$ to all of $\mathbb{R}^d$; i.e. $F|_{\operatorname{int}(K)} = f$ and $\in C^s(\mathbb{R}^d)$. Since any continuous function is uniformly continuous on a compact set, $\operatorname{int}(K)$ is dense in K, and since uniformly continuous functions are uniquely determined by their values on compact sets, then f coincides with F on all of K (not only on $\operatorname{int}(K)$).

For any $\mu \in \mathcal{P}(K)$, by definition of μ^+ we have that

$$\mathbb{E}_{X\sim\mu^+}[F(X)] = \mathbb{E}_{X\sim\mu^+}[F(X)I_{X\in K}] = \mathbb{E}_{X\sim\mu^+}[f(X)I_{X\in K}] = \mathbb{E}_{X\sim\mu}[f(X)].$$

Therefore, for any $\mu,\nu \in \mathcal{P}(K)$ we conclude that and each $f \in C^s(K)$ there exists some $F \in C^s(\mathbb{R}^d)$ such that

$$\mathbb{E}_{Y\sim\mu}[f(Y)] - \mathbb{E}_{Y\sim\nu}[f(Y)] = \mathbb{E}_{X\sim\mu^+}[F(X)] - \mathbb{E}_{X\sim\nu^+}[F(X)].$$

Consequentially, $d_{s:K}(\mu,\nu) \leqslant d_s(\mu^+,\nu^+)$. Conversely, since the restriction of any $g \in C^s(\mathbb{R}^d)$ to K belongs to $C^s(K)$ then the reverse inequality holds; namely, $d_{s:K}(\mu,\nu) \geqslant d_s(\mu^+,\nu^+)$. ■

By Lemma 13 we henceforth may interpret any such μ as its extension μ^+, without loss of generality.

E.2. Examples of Functions in The Classes $C^s_{poly:C,r}([0,1]^d,\mathbb{R})$ and $C^s_{exp:C,r}([0,1]^d,\mathbb{R})$

In several learning theory papers, especially in the kernel ridge regression literature e.g. Simon et al. (2023); Barzilai and Shamir (2023); Tsigler and Bartlett (2023); Simon et al. (2023); Cheng et al. (2024a,b), one often quantifies the *learnability* of a target function in terms of some sort of decay/growth rates of its coefficients in an appropriate expansion; e.g. the decay of its coefficients in an eigenbasis associated to a kernel. These decay/growth rates are often equivalent to the smoothness of a function[3]. Therefore, in a like spirit, we unpack the meaning of the smoothness condition in Assumption 1 which impacts the learning rates in Theorem 3 by giving examples of functions in the classes $C^s_{poly:C,r}([0,1]^d,\mathbb{R})$ and $C^s_{exp:C,r}([0,1]^d,\mathbb{R})$.

For brevity and transparency in our illustration, we consider the one-dimensional case. In particular, this shows that the class is far from being void.

Proposition 14 (Functions with Polynomially/Exponentially Growing C^s-Norms on $[0,1]$) *Fix $d \in \mathbb{N}_+$ and let K be a non-empty regular compact subset of $\mathbb{R}^d$. If $f : \mathbb{R} \to \mathbb{R}$ is real-analytic with power-series expansion at 0 given by*

$$f(x) = \sum_{i=0}^{\infty} \frac{\beta_i\, x^i}{i!},$$

and if there are $C, r > 0$ such that

(i) ***Polynomial Growth:*** *$|\beta_i| \leqslant C e^{i\,r}$ $(\forall i \in \mathbb{N})$, then $f \in C^\infty_{poly:C,r}([0,1],\mathbb{R})$; or*

(ii) ***Exponential Growth:*** *$|\beta_i| \leqslant C(1+i)^r$ $(\forall i \in \mathbb{N})$, then $f \in C^\infty_{poly:C,r}([0,1],\mathbb{R})$.*

Proof Since f is real-analytic we may consider its Maclaurin-Taylor series expansion which, coincides with $\sum_{i=0}^\infty \frac{\beta_i\, x^i}{i!}$; meaning that for each $i \in \mathbb{N}$ we have $\beta_i = \partial^i f(0)$. Therefore, standard analytic estimates and manipulations of the Maclaurin-Taylor series—see e.g. (Rudin, 1976, page 173)—yield

$$\begin{aligned}\max_{0\leqslant x\leqslant 1}\Big|\sum_{i=0}^{\infty}\frac{\beta_i\,x^i}{i!}-f(x)\Big| &\leqslant \frac{1}{(s+1)!}\sup_{0\leqslant x\leqslant 1}\Big|\big(\sum_{i=0}^{\infty}\frac{\beta_i\,x^i}{i!}\big)^{s+1}-\partial f^{s+1}(x)\Big| \\ &\leqslant \frac{1}{(s+1)!}\,\beta_s(s+1)!.\end{aligned} \tag{9}$$

If (i) holds, then the right-hand side of (9) is bounded from above by $C\,(s+1)^r$ and $f \in C^\infty_{poly:C,r}([0,1],\mathbb{R})$. If instead (ii) holds, then the right-hand side of (9) is bounded from above by $C\,e^{s\,r}$ implying $f \in C^\infty_{exp:C,r}([0,1],\mathbb{R})$. ■

3. See e.g. (Atkinson and Han, 2012, page 120-121) for an example between the decay rate of the Laplacian eigenspectrum characterize the smoothness of the functions in the RKHS of radially symmetric kernels.

Appendix F. Proof of Theorem 3

Before entering into the formal proof of Theorem 3, we overview its structure by overlooking its key steps and ideas on a high-level. The first step in deriving our generalization bounds is to quantify the regularity of the transformer model as a function of its depth, number of attention heads, and norm of its weight matrices. By *regularity*, we mean the number and size of the continuous partial derivatives admitted by the transformer. To quantify the size of the partial derivatives of the transformer we first remark that it is *smooth*; that is, it admits continuous partial derivatives of all orders (see Theorem 22).

We will uniformly bound the generalization capabilities of the class of transformers $\mathcal{T} \in \mathcal{TC}$ by instead uniformly bounding the generalization of any C^s functions on $\mathbb{R}^{Md}$ with C^s-norm at most equal to the largest C^s-norm in the class $\mathcal{TC}$. That is, we control the right-hand side of

$$\sup_{\mathcal{T}\in\mathcal{TC}} \left|\mathcal{R}_t(\mathcal{T}) - \mathcal{R}^{(N)}(\mathcal{T})\right| \leqslant \sup_{\hat{f}\in\mathcal{C}^s_R(\mathbb{R}^{Md})} \left|\mathcal{R}_t(\hat{f}) - \mathcal{R}^{(N)}(\hat{f})\right| \tag{10}$$

where $R = C_{\ell,\mathcal{TC},K,s}$ as defined in Theorem 3, describes the higher-order fluctuations of the "difference" between the target function $f^\star$ and any transformer $\mathcal{T} \in \mathcal{TC}$, as quantified by the loss function ℓ. Our first step is thus to bound R by upper-bounding maximal size of the s^{th} partial derivatives of all transformers $\mathcal{T} \in \mathcal{TC}$. Explicit bounds are computed in Theorem 41, and their order estimates (as a function of s) are given in Theorem 42. Combing these estimates with the maximal s^{th} partial derivatives of the loss and target function, via a Faá di Bruno-type formula (in Theorem 17 or Lemma 16), which is a multivariant higher-order chain rule, yields our estimate for R in (10).

Now that we have bounded R, appearing in the supremum term in (10), it remains to translate this into a generalization bound. We can do this by relating it to the so-called *smooth Wasserstein distance* d_s between the distribution of the Markov chain at time μ_t and its empirical distribution $\mu^{(N)} \overset{\text{def.}}{=} 1/N \sum_{n=1}^N \delta_{X_n}$ obtained by collecting samples up to time N. The *smooth Wasserstein distance* d_s, studied by Kloeckner (2020); Riekert (2022); Hou et al. (2023a), is the integral probability metric (IPM)-type distance quantifying the distance between any two Borel probability measures μ, ν on $\mathbb{R}^{Md}$ as the maximal distance which they can produced when tested on any function in $C^s_1(\mathbb{R}^{Md})$

$$d_s(\mu,\nu) \overset{\text{def.}}{=} \sup_{g\in\mathcal{C}^s_1(\mathbb{R}^{Md})} \mathbb{E}_{X\sim\mu}[g(X)] - \mathbb{E}_{Y\sim\nu}[g(Y)].$$

The right-hand side (RHS) of (10) can be expressed as R times the d_s distance between the (true) distribution μ_t of the process $X_\cdot$ at time t and the (empirical) distribution $\mu^{(N)}$ collected from samples

$$\text{RHS}\,(10) \leqslant \sup_{\hat{f}\in C^s_R(\mathbb{R}^{Md})} \|\ell(\hat{f}, f^\star)\|_{C^s}\, d_s(\mu_t, \mu^{(N)}). \tag{11}$$

The d_s distance between the process $X_\cdot$ at time t, i.e. μ_t, and the running empirical distribution $\mu^{(N)}$ can be accomplished in two steps. First, we *fast-forward time* and bound the d_s-distance between $\mu^{(N)}$ and the *stationary distribution* μ_∞ of the data-generating Markov chain $X_\cdot$ (at time $t = \infty$). We then *rewind time* and bound the d_s-distance between the stationary distribution μ_∞ and the distribution μ_t of the Markov process up to time t; by setting up the i.i.d. concentration of measure results of Kloeckner (2019); Riekert (2022). This last step is possible since our assumptions on $X_\cdot$ essentially guarantee that it has a finite (approximate) mixing time.

F.1. The Formal Proof

The proof of Theorem 3 will be largely broken down into two steps. First, we derive our concentration of measure result for the empirical mean compared to the true mean general of an arbitrary C^s function applied to a random input, where the C^s-norm of the C^s function is at most $R \geqslant 0$ (in Subsection F.3).

Next, (in Subsection F.3), we use the Faà di Bruno-type results in Section F.2 to bound the maximal C^s norm over the relevant class of transformer networks. We do this by first individually bounding each of the C^s-norms of its constituent pieces, namely the multi-head attention layers, the SLP blocks with smooth activation functions, and then ultimately, we bound the C^s-norms of the composition of transformer blocks using the earlier Faà di Bruno-type results.

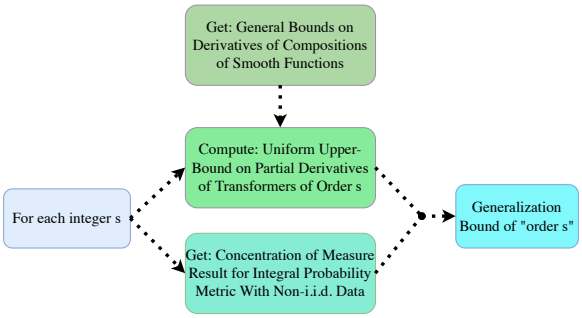

Figure 6: Workflow of the proof technique used to derive Theorem 3.

Our main result (Theorem 3) is then then obtained upon merging these two sets of estimates. The workflow which we use can be applied to derive generalization bounds for other machine learning, and is summarized in Figure 6.

F.2. Step 0 - Bounds on the C^s Regularity of Multivariate Composite Functions

In this section, we will derive a bound for the Sobolev norm of multivariate composite functions.

F.2.1. MULTIVARIATE FAÀ DI BRUNO FORMULA REVISITED

We begin by establishing notation and stating the multivariate Faà di Bruno formula from Constantine and Savits (1996).

Theorem 15 (Multivariate Faà di Bruno Formula, Constantine and Savits (1996)) *Let $n, m, k \in \mathbb{N}$, $\alpha \in \mathbb{N}^k$ with $|\alpha| = n$, and define*

$$h(x_1, \dots, x_k) \overset{\text{def.}}{=} f^{(1)}(g^{(1)}(x_1, \dots, x_k), \dots, g^{(m)}(x_1, \dots, x_k)).$$

Then, using the multivariate notation from Notation 1,

$$D^{\alpha} h(x) = \sum_{1 \leqslant |\beta| \leqslant n} (D^{\beta} f)(g(x)) \sum_{\eta, \zeta \in \mathcal{P}(\alpha, \beta)} \alpha! \prod_{j=1}^{n} \frac{[D^{\zeta^{(j)}} g(x)]^{\eta^{(j)}}}{\eta^{(j)}! (\zeta^{(j)}!)^{|\eta^{(j)}|}}.$$

where

$$\mathcal{P}(\alpha,\beta)=\Big\{\eta\overset{\text{def.}}{=}(\eta^{(1)},\dots,\eta^{(n)})\in(\mathbb{N}^m)^n,\zeta\overset{\text{def.}}{=}(\zeta^{(1)},\dots,\zeta^{(n)})\in(\mathbb{N}^k)^n:\\ \exists j\leqslant m:\eta^{(i)}=0,\zeta^{(i)}=0\,\textit{for}\,i<j,|\eta^{(i)}|>0\,\textit{for}\,i\geqslant j,\\ 0\prec\zeta^{(j)}\prec\ldots\prec\zeta^{(n)},\textstyle\sum_{i=1}^n\eta^{(i)}=\beta\,\textit{and}\,\sum_{i=1}^n|\eta^{(i)}|\zeta^{(i)}=\alpha\Big\}.$$

Proof See Constantine and Savits (1996). ■

F.2.2. UNIVERSAL BOUNDS

Theorem 16 *In the notation of Theorem 15, we have for a compact set $K\subseteq\mathbb{R}^k$ and an multi-index $\alpha\in\mathbb{N}^k$, $|\alpha|=n$,*

$$C_K^h(\alpha)\leqslant\max_{n'\in\{1,\ldots,n\}}C_{g[K]}^g(n')C_K^f(\leqslant n)^{n'}\sum_{1\leqslant|\beta|\leqslant n}\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}$$

where $C_K^h(\cdot)$, $C_{g[K]}^f(\cdot)$, $C_K^g(\cdot)$ are defined as in Notation 3.

Proof Using Theorem 15,

$$\begin{aligned}C_K^h(\alpha)&\leqslant\sum_{1\leqslant|\beta|\leqslant n}\|D^\beta f\|_{g[K]}\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{\prod_{i=1}^m\|(D^{\zeta^{(j)}}g)_i\|_K^{\eta_i^{(j)}}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}\\&\leqslant\sum_{1\leqslant|\beta|\leqslant n}C_{g[K]}^g(|\beta|)\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{\prod_{i=1}^m C_K^f(\leqslant n)^{\eta_i^{(j)}}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}\\&\leqslant\sum_{1\leqslant|\beta|\leqslant n}C_{g[K]}^g(|\beta|)C_K^f(\leqslant n)^{|\beta|}\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}\\&\leqslant\max_{n'\in\{1,\ldots,n\}}C_{g[K]}^g(n')C_K^f(\leqslant n)^{n'}\sum_{1\leqslant|\beta|\leqslant n}\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}.\end{aligned}$$

■

Next, we refine the strategy used in Hou et al. (2023b) to convert our uniform risk-bound to a concentration of measure problem. Once done, the remainder of the proof will be to obtain bounds on the rate at which this measure concentrates.

Lemma 17 *For $\alpha\in\{1,\cdots,k\}^n$, it satisfies that*

$$\sum_{1\leqslant|\beta|\leqslant n}\sum_{\eta,\zeta\in\mathcal{P}(\alpha,\beta)}\alpha!\prod_{j=1}^n\frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}=\Big[\frac{2m|\alpha|}{e\ln|\alpha|}(1+o(1))\Big]^{|\alpha|}$$

where $\mathcal{P}(\alpha,\beta)$ is as defined in Theorem 15.

Proof Consider functions

$$g^{(i)}(x) = g^{(i)}(x_1, \cdots, x_d) \overset{\text{def.}}{=} \exp\big(\sum_{j=1}^{d} x_j\big) : \mathbb{R}^d \to \mathbb{R}, \quad i = 1, ..., 2m,$$

$$f(g^{(1)}, \cdots, g^{(2m)}) \overset{\text{def.}}{=} \exp\big(\sum_{i=1}^{2m} g^{(i)}\big) : \mathbb{R}^{2m} \to \mathbb{R},$$

Since

$$\frac{\partial}{\partial g^{(i)}} f(g^{(1)}, \cdots, g^{(2m)}) = f(g^{(1)}, \cdots, g^{(2m)}),$$

it follows that

$$(D^\beta f)(g^{(1)}(x), \cdots, g^{(2m)}(x)) = f(g^{(1)}(x), \cdots, g^{(2m)}(x)), \quad \forall\, \beta \in \{1, \cdots, 2m\}^n.$$

Since

$$\begin{aligned}
&\frac{\partial}{\partial x^j} f(g^{(1)}(x_1, \cdots, x_k), \cdots, g^{(2m)}(x_1, \cdots, x_k)) \\
=&\sum_{i=1}^{2m} \frac{\partial}{\partial g^{(i)}} f(g^{(1)}(x_1, \cdots, x_k), \cdots, g^{(2m)}(x_1, \cdots, x_k)) \frac{\partial g^{(i)}(x_1, \cdots, x_k)}{\partial x_j} \\
=&\sum_{i=1}^{2m} f(g^{(1)}(x_1, \cdots, x_k), \cdots, g^{(2m)}(x_1, \cdots, x_k)) g^{(i)}(x_1, \cdots, x_k) \\
=&\sum_{i=1}^{2m} \frac{\partial}{\partial g^{(i)}} f(g^{(1)}(x), \cdots, g^{(2m)}(x)) \frac{\partial g^{(i)}(x)}{\partial x_j} \\
=&f(g^{(1)}(x), \cdots, g^{(2m)}(x)) \sum_{i=1}^{2m} g^{(i)}(x)
\end{aligned}$$

and

$$\frac{\partial g^{(i)}(x_1, \cdots, x_k)}{\partial x_j} = g^{(i)}(x_1, \cdots, x_k),$$

we can show by the Faà di Bruno formula that

$$\begin{aligned}
&D^\alpha f(g^{(1)}(x), \cdots, g^{(2m)}(x)) \\
&= D^\alpha \exp\big(\sum_{i=1}^{2m} g^{(i)}(x)\big) \\
&= \sum \frac{|\alpha|!}{\gamma_1!(1!)^{\gamma_1} \cdots \gamma_{|\alpha|}!(|\alpha|!)^{\gamma_{|\alpha|}}} \big(D^{\gamma_1 + \cdots + \gamma_{|\alpha|}} \exp\big)\big(\sum_{i=1}^{2m} g^{(i)}(x)\big) \prod_{j=1}^{|\alpha|} \Big[m^j \big(\sum_{i=1}^{2m} g^{(i)}(x)\big)\Big]^{\gamma_j},
\end{aligned}$$

where the summation on the right side of the last equality is over all $|\alpha|$-tuples $(\gamma_1, \cdots, \gamma_{|\alpha|}) \geqslant 0$ such that $1 \cdot \gamma_1 + 2 \cdot \gamma_2 + \cdots + |\alpha| \cdot \gamma_{|\alpha|} = |\alpha|$.

By the multivariate Faà di Bruno formula. For each $n = 1, ..., s-1$ fixed, and for each $\alpha \in \{1, \cdots, k\}^n$, we have

$$\begin{aligned}
&D^\alpha f(g^{(1)}(x), \cdots, g^{(2m)}(x)) \\
&= \sum_{1 \leqslant |\beta| \leqslant n} (D^\beta f)(g^{(1)}(x), \cdots, g^{(2m)}(x)) \sum_{\eta, \zeta \in \mathcal{P}(\alpha, \beta)} \alpha! \prod_{j=1}^{n} \frac{[D^{\zeta^{(j)}}(g^{(1)}(x), \cdots, g^{(2m)}(x))]^{\eta^{(j)}}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}.
\end{aligned}$$

Taking $x = (x_1, \cdots, x_k) = 0$, we have

$$\begin{aligned}
D^\alpha f(g^{(1)}(0), \cdots, g^{(2m)}(0)) &= \sum \frac{|\alpha|!}{\gamma_1!(1!)^{\gamma_1} \cdots \gamma_{|\alpha|}!(|\alpha|!)^{\gamma_{|\alpha|}}} \exp(2m) \prod_{j=1}^{|\alpha|} (2m)^{\gamma_j} \\
(m^\beta f)(g^{(1)}(0), \cdots, g^{(2m)}(0)) &= f(g^{(1)}(0), \cdots, g^{(2m)}(0)) = \exp(2m), \\
D^{\zeta^{(j)}}(g^{(1)}(x), \cdots, g^{(2m)}(x)) &= (1, \cdots, 1).
\end{aligned}$$

Substituting the above derivatives into the Faà di Bruno formula, we obtain

$$\begin{aligned}
\sum_{1 \leqslant |\beta| \leqslant n} \sum_{\eta, \zeta \in \mathcal{P}(\alpha, \beta)} \alpha! \prod_{j=1}^{k} \frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} &= \sum \frac{|\alpha|!}{\gamma_1!(1!)^{\gamma_1} \cdots \gamma_{|\alpha|}!(|\alpha|!)^{\gamma_{|\alpha|}}} \prod_{j=1}^{|\alpha|} (2m)^{\gamma_j} \\
&\leqslant (2m)^{|\alpha|} \sum \frac{|\alpha|!}{\gamma_1!(1!)^{\gamma_1} \cdots \gamma_{|\alpha|}!(|\alpha|!)^{\gamma_{|\alpha|}}} \\
&= (2m)^{|\alpha|} \Big(\frac{|\alpha|}{e \ln |\alpha|} \Big)^{|\alpha|} (1 + o(1))^{|\alpha|},
\end{aligned}$$

where the last equality follows from (Khorunzhiy, 2022, Theorem 2.1),

$$\sum \frac{|\alpha|!}{\gamma_1!(1!)^{\gamma_1} \cdots \gamma_{|\alpha|}!(|\alpha|!)^{\gamma_{|\alpha|}}} = \Big(\frac{|\alpha|}{e \ln |\alpha|} \Big)^{|\alpha|} (1 + o(1))^{|\alpha|}.$$

■

Corollary 18 (Level Specific C^s-Norm Bounds for Transformer Blocks) *In the notation of Theorem 16, it holds for $n \in \mathbb{N}, n > 1$ that*

$$C_K^h(\leqslant n) \leqslant \max_{n' \in \{1, \ldots, n\}} C_{g[K]}^g(n') C_K^f(\leqslant n)^{n'} \Big[\frac{2mn}{e \ln n} (1 + o(1)) \Big]^n.$$

and if $C_K^f(\leqslant n) \geqslant 1$,

$$C_K^h(\leqslant n) \leqslant C_{g[K]}^g(\leqslant n) C_K^f(\leqslant n)^n \Big[\frac{2mn}{e \ln n} (1 + o(1)) \Big]^n.$$

Proof Follows directly from Theorem 16 and Theorem 17. ■

F.2.3. BOUNDS IN DERIVATIVE TYPE

The goal of this section is to bound the derivative of composite functions by grouping with respect to $\sim$, defined in Notation 2.

Theorem 19 *In the notation of Theorem 15, we have for a compact set $K \subseteq \mathbb{R}^k$ and an ordered multi-index $\alpha \in \mathfrak{O}_n^k$*

$$C_K^h(\alpha) \leqslant \alpha! \sum_{\beta \in \mathfrak{O}_{\leqslant n}^m} N(\beta) C_{g[K]}^f(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,\beta)} \prod_{j=1}^n \frac{C_K^g(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}},$$

where $C_K^h(\cdot)$, $C_{g[K]}^f(\cdot)$, $C_K^g(\cdot)$ are defined as in Notation 3; and

$$\begin{aligned}\mathcal{P}'(\alpha,\beta) = \Big\{ &\eta \overset{\text{def.}}{=} (\eta^{(1)},\dots,\eta^{(n)}) \in (\mathbb{N}^m)^n, \zeta \overset{\text{def.}}{=} (\zeta^{(1)},\dots,\zeta^{(n)}) \in (\mathbb{N}^k)^n : \\ &\exists j \leqslant m : \eta^{(i)} = 0, \zeta^{(i)} = 0 \text{ for } i < j, |\eta^{(i)}| > 0 \text{ for } i \geqslant j, \\ &0 < \zeta^{(j)} \lhd \dots \lhd \zeta^{(n)}, \textstyle\sum_{i=1}^n \eta^{(i)} = \beta \text{ and } \sum_{i=1}^n |\eta^{(i)}| \zeta^{(i)} = \alpha \Big\},\end{aligned}$$

where $\alpha \lhd \beta$ for $\alpha, \beta \in \mathbb{N}^k$ if $|\alpha| \leqslant |\beta|$ and $\alpha \neq \beta$.

Proof We have for $\alpha \in \mathfrak{O}_n^k$

$$\begin{aligned} C_K^h(\alpha) &\leqslant \max_{\gamma \sim \alpha} \sum_{1 \leqslant |\beta| \leqslant n} \|D^\beta f\|_{g[K]} \sum_{\eta,\zeta \in \mathcal{P}(\gamma,\beta)} \alpha! \prod_{j=1}^n \frac{\prod_{i=1}^m \|(D^{\zeta^{(j)}} g)_i\|_K^{\eta_i^{(j)}}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \\ &\leqslant \sum_{1 \leqslant |\beta| \leqslant n} C_{g[K]}^f(\mathfrak{o}(\beta)) \max_{\gamma \sim \alpha} \sum_{\eta,\zeta \in \mathcal{P}(\gamma,\beta)} \alpha! \prod_{j=1}^n \frac{C_K^g(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}. \end{aligned}$$

Then

$$\left\{ \eta, (\mathfrak{o}(\zeta^{(1)}), ..., \mathfrak{o}(\zeta^{(n)})) \middle| (\eta,\zeta) \in \mathcal{P}'(\alpha,\beta) \right\}$$

is invariant in α with respect to $\sim$ and thus

$$C_K^h(\alpha) \leqslant \sum_{1 \leqslant |\beta| \leqslant n} C_{g[K]}^f(\mathfrak{o}(\beta)) \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,\beta)} \alpha! \prod_{j=1}^n \frac{C_K^g(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}.$$

Further, notice that

$$\left\{ ((|\eta^{(1)}|, \eta^{(1)}!), \dots, (|\eta^{(n)}|, \eta^{(n)}!)), \zeta \middle| (\eta,\zeta) \in \mathcal{P}'(\alpha,\beta) \right\}$$

is invariant in β with respect to $\sim$ and the assertion follows. ■

Corollary 20 *In the notation of Theorem 19, if f is affine-linear,*

$$C_K^h(\alpha) \leqslant m\alpha! C_{g[K]}^f(e_1) C_K^g(\alpha),$$

where $C_{g[K]}^f(e_1)$ is the maximum weight of the matrix representing f.

Proof Theorem 19 yields

$$C_K^h(\alpha) \leqslant m\alpha! C_{g[K]}^f(e_1) \sum_{\eta,\zeta\in\mathcal{P}'(\alpha,\beta)} \prod_{j=1}^{n} \frac{C_K^g(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}},$$

and since $\mathcal{P}'(\alpha, e_1) = \{(0, ..., 0, e_1), (0, ..., 0, \alpha)\}$ the result follows. ■

F.3. Step 1 - Concentration of Measure - Bounding the Right-Hand Side of (10)

We are now ready to derive our main concentration of measure results used to derive our risk-bound. This corresponds to bounding term (10) by controlling the integral probability term $d_s(\mu_t, \mu^{(N)})$ in (11), with high probability, where the randomness is due to the randomness of the empirical measure $\mu^{(N)}$.

We state the next bound in the case where the input space if $\mathbb{R}^d$. Note that the results hold for any other input dimension, such as Md, simply by relabeling $d \leftarrow Md$. Thus, it applies to the *finite-dimensional* Markovian lifts $X_\cdot^M$ of data-generating processes $X_\cdot$, where $M \in \mathbb{N}_+$, by relabeling. Therefore, for notational minimality, we chose to label the input dimension d and not dM.

Proposition 21 (Excess Risk-Bound) *Under Assumption 1 and either 2 or 3, let $f^\star : \mathbb{R}^d \to \mathbb{R}^D$, $\ell : \mathbb{R}^{2D} \to \mathbb{R}$, and $R, r > 0$ be a such that the composite map $\mathbb{R}^d \ni x \mapsto \ell(f^\star(x), f(x))$ belongs to $\mathcal{C}_R^s(\mathbb{R}^d)$ for all $f \in C_r^2(\mathbb{R}^d)$. Then, there exists some $\kappa \in (0,1)$ depending only on the Markov chain $X_\cdot$ and some $t_0 \in \mathbb{N}_0$ such that for each $t_0 \leqslant N < t \leqslant \infty$, each "rate-to-constant-tradeoff parameter" $s \in \mathbb{N}_+$, and every "confidence level" $\delta \in (0,1]$ the following*

$$\sup_{g\in\mathcal{C}_R^s(\mathbb{R}^d)} \frac{\left|\mathcal{R}_{\max\{t,N\}}(g) - \mathcal{R}^{(N)}(g)\right|}{R} \lesssim \kappa^t + \frac{\sqrt{2\ln(1/\delta)}}{N^{1/2}} + \begin{cases} \frac{\log\left(c\,N\right)^{d/(2s+1)}}{c\,N^{1/2}} & \text{if } d < 2s \\ \frac{\log\left(c\,N\right)}{c\,N^{1/2}} & \text{if } d = 2s \\ \frac{\log\left(c\,N\right)^{d-2s+(s/d)}}{c_2\,N^{s/d}} & \text{if } d = 2s \end{cases}$$

holds with probability at least $1-\delta$; where $0 < \kappa < 1$, and we use the notation $\kappa^\infty \overset{\text{def.}}{=} 0$.

Proof [Proof of Proposition 21] By hypothesis, $\tilde{f} \in C_r^s(\mathbb{R}^d)$ the induced map

$$\begin{aligned} f : \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto f(x) \overset{\text{def.}}{=} \ell\big(f^\star(x), \tilde{f}(x)\big) \end{aligned} \tag{12}$$

belongs to $C_R^s(\mathbb{R}^d)$.

Conversion to a Concentration of Measure Problem. Denote the empirical (random) measure associated with the samples $\{(X_n, Y_n)\}_{n=1}^N$ by $\mu^{(N)} \overset{\text{def.}}{=} \frac{1}{N}\sum_{n=1}^N \delta_{(X_n,Y_n)}$. Note that the generalization bound is 0 for any constant function; therefore, we consider the bound over $C_R^s(\mathbb{R}^d) \setminus \mathrm{Lip}_0$ where

Lip_0 denotes the set of constant functions from $\mathbb{R}^d$ to $\mathbb{R}$. Note the bijection between $C^s_R(\mathbb{R}^d) \setminus \mathrm{Lip}_0$ and $C^s_1(\mathbb{R}^d) \setminus \mathrm{Lip}_0$ given by $f \mapsto \frac{1}{\max\{1, \|f\|_{C^s(\mathbb{R}^d)}\}} f$. Therefore, we compute

$$\begin{aligned}
\left|\mathcal{R}_t(f) - \mathcal{R}^{(N)}(g)\right| &\leqslant \sup_{g \in \mathcal{C}^s_R(\mathbb{R}^d)} \left|\mathcal{R}_t(f) - \mathcal{R}^{(N)}(g)\right| \\
&\leqslant R \sup_{g \in \mathcal{C}^s_1(\mathbb{R}^d)} \left|\mathcal{R}_t(g) - \mathcal{R}^{(N)}(g)\right| \\
&\leqslant R\, d_{C^s}(\mu_{\max\{t,N\}}, \mu^{(N)}) \\
&\leqslant R \left(\underbrace{d_{C^s}(\mu_t, \bar{\mu})}_{\text{(IV)}} + \underbrace{d_{C^s}(\bar{\mu}, \mu^{(N)})}_{\text{(V)}} \right). \qquad (13)
\end{aligned}$$

Next, we bound terms (I) and (II).

Bounding Term (IV). If Assumption 2 holds then: for every $t \in \mathbb{N}_+$ each $x, \tilde{x} \in \mathcal{X}$ we have

$$\mathcal{W}_1\big(P^t(x,\cdot), P^t(\tilde{x},\cdot)\big) \leqslant \kappa^t\, \mathcal{W}_1(\delta_x, \delta_{\tilde{x}}) = \kappa^t \|x - \tilde{x}\|.$$

If, instead, we operate under the log-Sobolev Assumption 3, then (Bobkov and Götze, 1999, Theorem 1.3) can be applied to $\bar{\mu}$ and $P(x,\cdot)$ for each $x \in \mathcal{X}$, implying that the transport inequalities hold: for each $\nu \in \mathcal{P}(\mathcal{X})$ and each $\tilde{\mu} \in \{\bar{\mu}, \mu_0\} \cup \{P(x,\cdot)\}_{x \in \mathcal{X}}$

$$\mathcal{W}_1(\tilde{\mu}, \nu)^2 \leqslant 2C^2\ \mathrm{KL}(\nu|\tilde{\mu}) \qquad (14)$$

where we recall the definition of the Kullback–Leibler divergence $\mathrm{KL}(\nu|\mu) \overset{\text{def.}}{=} \mathbb{E}_{X\sim\nu}[\log(\frac{d\nu}{d\mu}(X))]$. Thus, (14) implies that the following exponential contractility property of the Markov kernel: there exists some $\kappa \in (0,1)$ such that for each $x, \tilde{x} \in \mathcal{X}$ and every $t \in \mathbb{N}_+$

$$\mathcal{W}_1\big(P^t(x,\cdot), P^t(\tilde{x},\cdot)\big) \leqslant \kappa^t \|x - \tilde{x}\|. \qquad (15)$$

Furthermore, (15) implies that the conditions for (Riekert, 2022, Theorem 1.5) are met; whence, for every $\varepsilon \geqslant 0$ and each $N \in \mathbb{N}$ the following holds with probability at-least $1 - \exp\left(\frac{-N\,\varepsilon^2(1-\kappa)^2}{2C^2}\right)$

$$\text{(IV)} = d_s(\bar{\mu}, \mu^{(N)}) \leqslant \mathbb{E}\big[d_s(\bar{\mu}, \mu^{(N)})\big] + \varepsilon, \qquad (16)$$

for some $C > 0$. Upon setting $\varepsilon \overset{\text{def.}}{=} \frac{C\sqrt{2\ln(1/\delta)}}{\sqrt{N\,(1-\kappa^2)}}$, (16) implies that: for every $N \in \mathbb{N}$ and each $\delta \in (0,1]$ the following holds with probability at-least $1 - \delta$

$$\text{(IV)} = d_s(\bar{\mu}, \mu^{(N)}) \leqslant \underbrace{\mathbb{E}\big[d_s(\bar{\mu}, \mu^{(N)})\big]}_{\text{(VI)}} + \frac{C\sqrt{2\ln(1/\delta)}}{\sqrt{N\,(1-\kappa^2)}}. \qquad (17)$$

It remains to bound the expectation term (VI) in (17) to bound term (IV).

Under the exponential moment assumption 4, we have that

$$\mathbb{E}_{X\sim P(x,\cdot)}[e^{\beta|X|} - 1] \leqslant \gamma\,(e^{\beta|x|} - 1) + (C - 1 + \gamma). \qquad (18)$$

Therefore (Riekert, 2022, Proposition 1.3), implies that $\sup_{t\in\mathbb{N}_0}\mathbb{E}[e^{\beta|X_t|}-1]<\infty$. Whence, (Riekert, 2022, Assumption 2) holds with Young function $\Phi(x)=\frac{1}{\max\{1,\sup_{t\in\mathbb{N}_+}\mathbb{E}[e^{\beta|X_t}-1]}(e^{\beta|X_t|}-1)$; namely, $\sup_{t\in\mathbb{N}_0}\mathbb{E}[\Phi(|X_t|)]\leqslant 1$. Consequentially, (Riekert, 2022, Theorem 1.1) applies from which we conclude that there is some $t_0\in\mathbb{N}_+$ such that for all $N\geqslant t_0$

$$\text{(VI)}=\mathbb{E}\big[d_s(\bar{\mu},\mu^{(N)})\big]\lesssim \log\big((1-\kappa)\,N\big)^s\begin{cases}\frac{\log\left((1-\kappa)\,N\right)^{d/(2s)+1}}{(1-\kappa)^{1/2}\,N^{1/2}} & \text{if } 1=d<2s\\ \frac{\log\left((1-\kappa)\,N\right)}{(1-\kappa)^{1/2}\,N^{1/2}} & \text{if } d=2s\\ \frac{\log\left((1-\kappa)\,N\right)^{d-2s+s/d}}{(1-\kappa)^{s/d}\,N^{s/d}} & \text{if } d=2s\end{cases}. \tag{19}$$

Combining the order estimate of (VI) in (19) with the estimate in (17) implies that: for every $N\geqslant t_0$ and each $\delta\in(0,1]$ we have

$$\text{(IV)}=d_s(\bar{\mu},\mu^{(N)})\lesssim\frac{\sqrt{2\ln(1/\delta)}}{N^{1/2}}+\begin{cases}\frac{\log\left(c\,N\right)^{d/(2s)+1}}{c\,N^{1/2}} & \text{if } 1=d<2s\\ \frac{\log\left(c\,N\right)}{c\,N^{1/2}} & \text{if } d=2s\\ \frac{\log\left(c\,N\right)^{d-2s+s/d}}{c_2\,N^{s/d}} & \text{if } d=2s\end{cases} \tag{20}$$

where $c\overset{\text{def.}}{=}(1-\kappa),c_2\overset{\text{def.}}{=}c^{s/d}\in(0,1)$, and $\lesssim$ suppresses the absolute constant $\max\{1,C\}>0$.

Bounding Term (V). Next, we bound (V) by computing

$$\begin{aligned}\text{(V)}=d_{C^s}(\mu_t,\bar{\mu})&\overset{\text{def.}}{=}\sup_{g\in\mathcal{C}_1^s(\mathbb{R}^d)}\mu_t[g]-\bar{\mu}[g]\\ &\leqslant\sup_{g\in\text{Lip}_1(\mathbb{R}^d)}\mu_t[g]-\bar{\mu}[g] &(21)\\ &=\mathcal{W}_1(\mu_t,\bar{\mu}) &(22)\\ &=\mathcal{W}_1(P^t\mu_0,\bar{\mu}) &(23)\\ &=\mathcal{W}_1(P^t\mu_0,P^t\bar{\mu}) &(24)\\ &\leqslant\kappa^t\,\mathcal{W}_1(\mu_0,\bar{\mu})\overset{\text{def.}}{=}\kappa^t\,C &(25)\end{aligned}$$

where (21) held by definition of the MMD d_{C^s} and by the inclusion of $\mathcal{C}_1^s(\mathbb{R}^d)\subset\text{Lip}_1(\mathbb{R}^d)$, (22) held by Kantorovich duality (see (Villani, 2009, Theorem 5.10)), (24) held since $\bar{\mu}$ is the stationary probability measure for the Markov chain $X_\cdot$, it is invariant to the action of the Markov kernel, and (25) followed from (Olivera and Tudor, 2019, Corollary 21) since we deduced the exponential contractility property (15) of the Markov kernel. Note that $C\overset{\text{def.}}{=}\mathcal{W}_1(\mu_0,\bar{\mu})$ is a constant depending only on the initial and stationary distributions of the Markov chain.

Conclusion. Incorporating the estimates for (V) and (IV) into the right-hand side of (13) implies that: for every $t,N\geqslant t_0$, $s\in\mathbb{N}_+$, and each $\delta\in(0,1]$ the following holds

$$\sup_{g\in\mathcal{C}_R^s(\mathbb{R}^d)}\frac{\left|\mathcal{R}_{\max\{t,N\}}(g)-\mathcal{R}^{(N)}(g)\right|}{R}\lesssim I_{t<\infty}\,\kappa^t+\frac{\sqrt{2\ln(1/\delta)}}{N^{1/2}}+\begin{cases}\frac{\log\left(c\,N\right)^{d/(2s+1)}}{c\,N^{1/2}} & \text{if } 1=d<2s\\ \frac{\log\left(c\,N\right)}{c\,N^{1/2}} & \text{if } d=2s\\ \frac{\log\left(c\,N\right)^{d-2s+(s/d)}}{c_2\,N^{s/d}} & \text{if } d=2s\end{cases}$$

with probability at-least $1-\delta$; where $c \overset{\text{def.}}{=} (1-\kappa)$ and $\kappa \in (0,1)$; where $I_{t<\infty}k^{\infty} \overset{\text{def.}}{=} 0$ if $t=\infty$. ■

F.4. Step 2 (A) - Bounding the C^s Regularity of Transformer Building Blocks

We begin by the following simple remark, that if the activation function used to defined the transformer is smooth, then so must the entire transformer model.

Proposition 22 (Transformers with Smooth Activation Functions are Smooth) *Fix $\mathcal{TC}$, as in Theorem 12, then every transformer $\mathcal{T} \in \mathcal{TC}$ is smooth.*

Proof [Theorem 22] The smoothness of Att follows directly from the smoothness of $\mathrm{softmax}$, which immediately implies smoothness of $\mathcal{MH}$ since the operators used for its definition are smooth. Furthermore, the $\mathcal{LN}$ is smooth due to its smooth and lower-bounded denominator and the activation function σ is smooth by definition, therefore we conclude that $\mathcal{TB} \in \mathcal{TBC}$ is smooth for every $\mathcal{TBC}$ as in Theorem 11 and we obtain smoothness of $\mathcal{T} \in \mathcal{TC}$ as a consequence. ■

F.4.1. THE SOFTMAX FUNCTION

Lemma 23 (Representation of higher-order $\mathrm{softmax}$ derivatives) *For $F \in \mathbb{N}$ and*

$$\mathrm{smax} : \mathbb{R}^F \to \mathbb{R}^F, \quad x \mapsto \left(\exp(x_i)/\textstyle\sum_{j=0}^{F-1}\exp(x_j)\right)_{i=1}^{F}.$$

there exists for any multi-index $\alpha \in \mathbb{N}^F$ and $m \in \{1,\dots,F\}$ indicators $(a_{i,j}^k)_{i,j\in I(\alpha)}^{k\in\{1,\dots,|\alpha|!\}} \subseteq \{0,1\}$ such that

$$\mathrm{smax}^{(\alpha)}(x_m) = \sum_{k=1}^{|\alpha|!} \mathrm{smax}(x_m) \prod_{i,j\in I(\alpha)} (a_{i,j}^k - \mathrm{smax}(x_j)), \tag{26}$$

where $I(\alpha) \overset{\text{def.}}{=} \{(i,j) : i = 1,\dots,F, j = 1,\dots,\alpha_i\}$.

Proof For $|\alpha| = 0$, we have $n \in \{1,\dots,F\}$ s.t. $\alpha_n = 1$, therfore

$$\mathrm{smax}^{(\alpha)}(x_m) = \frac{\partial\,\mathrm{smax}}{\partial x_n}(x_m) = \mathrm{smax}(x_m)\left(\delta_{mn} - \mathrm{smax}(x_n)\right),$$

which is of the form (26). Now, let $\alpha \in \mathbb{N}^F$ arbitrary, therefore, by defining $\alpha' \in \mathbb{N}^F$ by $\alpha_i' \overset{\text{def.}}{=} \alpha_i$ for $i \neq n$ and $\alpha_n' \overset{\text{def.}}{=} \alpha_n - 1$ (w.l.o.g. $\alpha_n > 0$). We have

$$\begin{aligned}\mathrm{smax}^{(\alpha)}(x_m) &= \frac{\partial\,\mathrm{smax}^{(\alpha')}}{\partial x_n}(x_m) \\ &= \frac{\partial}{\partial x_n}\sum_{k=1}^{|\alpha'|!} \mathrm{smax}(x_m) \prod_{i,j\in I(\alpha')} (a_{i,j}'^k - \mathrm{smax}(x_j)).\end{aligned}$$

Since for any k

$$\begin{aligned}
&\frac{\partial}{\partial x_n}\operatorname{smax}(x_m)\prod_{i,j\in I(\alpha')}(a'^k_{i,j}-\operatorname{smax}(x_j))\\
&=\operatorname{smax}(x_m)\left(\delta_{mn}-\operatorname{smax}(x_n)\right)\prod_{i,j\in I(\alpha')}(a'^k_{i,j}-\operatorname{smax}(x_j))\\
&\quad+\operatorname{smax}(x_m)\sum_{i',j'\in I(\alpha')}-\operatorname{smax}(x_{j'})\left(\delta_{j',n}-\operatorname{smax}(x_n)\right)\prod_{\substack{i,j\in I(\alpha)\\(i,j)\neq(i',j')}}(a'^k_{i,j}-\operatorname{smax}(x_j)),
\end{aligned}$$

we can define $(a^k_{i,j})^{k\in\{1,\dots,|\alpha'|+1\}}_{i,j\in I(\alpha)}\subseteq\{0,1\}$ such that

$$\frac{\partial}{\partial x_n}\operatorname{smax}(x_m)\prod_{i,j\in I(\alpha')}(a'^k_{i,j}-\operatorname{smax}(x_j))=\sum_{k=1}^{|\alpha|}\operatorname{smax}(x_m)\prod_{i,j\in I(\alpha)}(a^k_{i,j}-\operatorname{smax}(x_j)).$$

Since $|\alpha|!=|\alpha|\cdot|\alpha'|!$, this concludes the proof. ■

Lemma 24 (Bound of higher-order softmax **derivatives)** *With Notation 3, it holds for any set $K\in\mathbb{R}^k, k\in\mathbb{N}$ and any $\alpha\in\mathfrak{O}^k_{<\infty}$ that*

$$C^{\operatorname{smax}}(\alpha)\leqslant|\alpha|!.$$

Proof This is a direct consequence of the representation in Theorem 23 together with $\|\operatorname{smax}\|=1$. ■

F.4.2. THE MULTI-HEAD SELF-ATTENTION MECHANISM

Lemma 25 (Bound of Dot product) *In the notation of Theorem 10 and for $m\in\{1,\dots,M\}$*

$$\operatorname{dp}_m(\,\cdot\,;Q,K):\mathbb{R}^{Md_{\text{in}}}\longrightarrow\mathbb{R}^M,\quad x\longmapsto\langle Qx_m,Kx_j\rangle_{j=0}^M$$

we have using Notation 3

1. $C^{\operatorname{dp}_m}_K(e_1)\leqslant 2d_{\text{in}}d_K\|K\|C^QC^K$, *where* $C^Q\overset{\text{def.}}{=}\max_{i,i'\in\{1,\dots,d_K\}\times\{1,\dots,d_{\text{in}}\}}|Q_{i,i'}|$, C^K *analogously, and* $\|K\|\overset{\text{def.}}{=}\max_{x\in K}\|x\|$. *Additionally,*
2. $C^{\operatorname{dp}_m}_K(\alpha)\leqslant 2d_KC^QC^K$, *for* $|\alpha|=2$, *and*
3. $C^{\operatorname{dp}_m}_K(\alpha)=0$ *for* $|\alpha|>2$.

Since all bounds are not dependent on m we write C^{dp} short for C^{dp_m}.

Proof

1. Let $l = (l_1, l_2) \in \{1, \dots, M\} \times \{1, \dots, d_{\text{in}}\}$. Assume $l_1 = m$. If $j \neq m$, then

$$D^{e_l} \operatorname{dp}_m(x; Q, K)_j = D^{e_l} \sum_{i=1}^{d_K} (Kx_j)_i \sum_{i'=1}^{d_{\text{in}}} Q_{i,i'}(x_m)_{i'} = \sum_{i=1}^{d_K} \left(\sum_{i'=1}^{d_{\text{in}}} K_{i,i'}(x_j)_{i'} \right) Q_{i,l_2},$$

implying

$$\|D^{e_l} \operatorname{dp}_m(x; Q, K)\| \leqslant \|K\| \sum_{i=1}^{d_K} Q_{i,l_2} \sum_{i'=1}^{d_{\text{in}}} K_{i,i'} \leqslant d_{\text{in}} d_K \|K\| C^Q C^K. \tag{27}$$

If $j = m$,

$$\begin{aligned} D^{e_l} \operatorname{dp}_m(x; Q, K)_j &= D^{e_l} \sum_{i=1}^{d_K} \left(\sum_{i'=1}^{d_{\text{in}}} K_{i,i'}(x_m)_{i'} \right) \left(\sum_{i'=1}^{d_{\text{in}}} Q_{i,i'}(x_m)_{i'} \right) \\ &= \sum_{i=1}^{d_K} \left(K_{i,l_2} \sum_{i'=1}^{d_{\text{in}}} Q_{i,i'}(x_m)_{i'} + Q_{i,l_2} \sum_{i'=1}^{d_{\text{in}}} K_{i,i'}(x_m)_{i'} \right) \end{aligned}$$

therefore implying

$$\|D^{e_l} \operatorname{dp}_m(x; Q, K)\| \leqslant 2 d_{\text{in}} d_K \|K\| C^Q C^K.$$

If $l_1 \neq m$ then for $j \neq l_1$, $D^{e_l} \operatorname{dp}_m(x; Q, K)_j = 0$, for $j = l_1$

$$D^{e_l} \operatorname{dp}_m(x; Q, K)_j = D^{e_l} \sum_{i=1}^{d_K} (Qx_m)_i \sum_{i'=1}^{d_{\text{in}}} K_{i,i'}(x_j)_{i'} = \sum_{i=1}^{d_K} \left(\sum_{i'=1}^{d_{\text{in}}} Q_{i,i'}(x_m)_{i'} \right) K_{i,l_2},$$

and we obtain (27) analogously.

2. If $l_1 = m$ and $j \neq m$

$$D^{e_l} \left((x_m)_{l_2} \sum_{i=1}^{d_K} \left(\sum_{i'=1}^{d_{\text{in}}} k_{i,i'}(x_j)_{i'} \right) q_{i,l_2} \right) = 0,$$

implying $\|D^{2e_l} \operatorname{dp}_m(x; Q, K)\| \leqslant 0$, what analogously holds for $l_1 \neq m$. However, for $l_1 = m$ and $j = m$

$$D^{e_l} \left(\sum_{i=1}^{d_K} K_{i,l_2} \sum_{i'=1}^{d_{\text{in}}} Q_{i,i'}(x_m)_{i'} + Q_{i,l_2} \sum_{i'=1}^{d_{\text{in}}} K_{i,i'}(x_m)_{i'} \right) = \sum_{i=1}^{d_K} K_{i,l_2} Q_{i,l_2} + Q_{i,l_2} K_{i,l_2}$$

we have

$$\|D^{2e_l} \operatorname{dp}_m(x; Q, K)\| \leqslant 2 d_K C^Q C^K.$$

3. Let $l' = (l'_1, l'_2) \in \{1, \dots, M\} \times \{1, \dots, d_{\text{in}}\}$. Assume $l_1 = m$, $j \neq m$. If $l'_1 \neq j$, $D^{e_l + e_{l'}} \operatorname{dp}_m(x; Q, K)_j = 0$. For $l'_1 = j$ follows $D^{e_l + e_{l'}} \operatorname{dp}_m(x; Q, K)_j = \sum_{i=1}^{d_K} K_{i,l'_2} Q_{i,l_2}$. If $l_1 = m$, $j \neq m$, we have $D^{e_l + e_{l'}} \operatorname{dp}_m(x; Q, K)_j = 0$ in the case that $l'_1 \neq m$, and for $l'_1 \neq m$ we obtain

$$D^{e_l + e_{l'}} \operatorname{dp}_m(x; Q, K)_j = \sum_{i=1}^{d_K} K_{i,l_2} Q_{i,l'_2} + Q_{i,l_2} K_{i,l'_2}.$$

This means, we can use the bound

$$\|D^{e_l + e_{l'}} \operatorname{dp}_m(x; Q, K)\| \leqslant 2 d_K C^Q C^K.$$

■

Lemma 26 (Bound of Self-Attention for Derivative Type) *Using the notation of Notation 3, Theorem 10 and Theorem 25, it holds that*

$$C_K^{\mathrm{Att}}(\alpha) \leqslant d_{\mathrm{in}} M C^V \left(\|K\| C_K^{\mathrm{smax}\circ \mathrm{dp}}(\alpha) + \sum_{l=1}^{Md_{\mathrm{in}}} \alpha_l C_K^{\mathrm{smax}\circ \mathrm{dp}}(\alpha - e_l) \right)$$

where

$$C_K^{\mathrm{smax}\circ \mathrm{dp}}(\alpha) \leqslant \alpha! \sum_{\beta \in \mathfrak{O}_{\leqslant n}^M} N(\beta) C_{\mathrm{dp}[K]}^{\mathrm{smax}}(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,\beta)} \prod_{j=1}^{n} \frac{C_K^{\mathrm{dp}}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}! (\zeta^{(j)}!)^{|\eta^{(j)}|}}. \tag{28}$$

Proof Fix $\alpha \in \mathbb{N}^k$, and note that

$$\|D^\alpha \operatorname{Att}(x;Q,K,V)\| \leqslant \max_{m\in\{1,\ldots,M\}} \max_{i\in\{0,\ldots,d_V\}} \|D^\alpha \operatorname{Att}(x;Q,K,V)_{m,i}\|$$

and

$$\begin{aligned}
&\|D^\alpha \operatorname{Att}(x;Q,K,V)_{m,i}\| \\
&\quad \leqslant \sum_{j=1}^{M} \sum_{i'=0}^{d_{\mathrm{in}}} \|D^\alpha \operatorname{smax}\circ \operatorname{dp}(x;Q,K)_j V_{i,i'}(x_j)_{i'}\| \\
&\quad \leqslant d_{\mathrm{in}} M \max_{j\in\{1,\ldots,M\}} \max_{i'\in\{0,\ldots,d_{\mathrm{in}}\}} \|D^\alpha \operatorname{smax}\circ \operatorname{dp}(x;Q,K)_j V_{i,i'}(x_j)_{i'}\|.
\end{aligned}$$

Due to the extended Leibnitz rule Hardy (2006), we have

$$\begin{aligned}
&\|D^\alpha \operatorname{smax}\circ \operatorname{dp}(x;Q,K)_j V_{i,i'}(x_j)_{i'}\| \\
&\quad \leqslant \|D^\alpha \operatorname{smax}\circ \operatorname{dp}(x;Q,K)_j V_{i,i'}(x_j)_{i'}\| + \sum_{l=1}^{Md_{\mathrm{in}}} V_{i,i'}\alpha_l \|D^{\alpha-e_l} \operatorname{smax}\circ \operatorname{dp}(x;Q,K)_j\|.
\end{aligned}$$

Equation (28) follows directly from Theorem 19. ■

Corollary 27 (Bound of Self-Attention for Derivative Level) *Using the setting of Theorem 26, for $n \in \mathbb{N}$,*

$$C_K^{\mathrm{Att}}(n) \leqslant d_{\mathrm{in}} M C^V C_K^{\mathrm{smax}\circ \mathrm{dp}}(\leqslant n) \left(\|K\| + n d_{\mathrm{in}} M\right) \tag{29}$$

where

$$C_K^{\mathrm{smax}\circ \mathrm{dp}}(\leqslant n) \leqslant C_{\mathrm{dp}[K]}^{\mathrm{smax}}(\leqslant n) C_K^{\mathrm{dp}}(\leqslant n)^n \left[\frac{2nM}{e \ln n}(1+o(1))\right]^n. \tag{30}$$

Proof Equation (29) follows directly from Theorem 26; and (30) is a consequence of Theorem 18. ■

Corollary 28 (Bound of Multi-head Self-Attention) *In the notation of Theorem 10, Theorem 19 and Theorem 25 it holds that*

$$C_K^{\mathcal{MH}}(\alpha) \leqslant \alpha! d_V C^W C_K^{\mathrm{Att}}(\alpha)$$

where

$$C_K^{\mathrm{Att}} \overset{\text{def.}}{=} \max_{h\in\{1,\dots,H\}} C_K^{\mathrm{Att}(\,\cdot\,;Q^{(h)},K^{(h)},V^{(h)})}, \quad C^W \overset{\text{def.}}{=} \max_{h\in\{1,\dots,H\}} W^{(h)}.$$

In particular, we have the following order estimate

$$C_K^{\mathcal{MH}}(\leqslant n) \in \mathcal{O}\Big(M^2\|K\|\|W\|\|V\|(c_{d_{\mathrm{in}},d_K}\|K\|\|Q\|\|K\|)^n\ n^2\Big(\frac{n}{e}\Big)^{2n} C_n^n\Big).$$

Proof From Theorem 20 and Theorem 26 we directly obtain

$$\begin{aligned} C_K^{\mathcal{MH}}(\alpha) \leqslant & n! d_V C^W d_{\mathrm{in}} M C^V n! (2 d_{\mathrm{in}} d_K \|K\| C^Q C^K)^n \\ & \times (\|K\| + n d_{\mathrm{in}} M) \Big[\frac{2nM}{e \ln n}(1+o(1))\Big]^n. \end{aligned} \tag{31}$$

Applying Stirling's approximation, we have that

$$C_K^{\mathcal{MH}}(\alpha) \in \mathcal{O}\Big(M^2\|K\|\|W\|\|V\|(c_{d_{\mathrm{in}},d_K}\|K\|\|Q\|\|K\|)^n\ n^2\Big(\frac{n}{e}\Big)^{2n} C_n^n\Big), \tag{32}$$

where $C_n \overset{\text{def.}}{=} \frac{2nM}{e\ln n}(1+o(1))$ and $c_{d_{\mathrm{in}},d_K} \overset{\text{def.}}{=} 2 d_{\mathrm{in}} d_K$. ■

F.4.3. THE ACTIVATION FUNCTIONS

Lemma 29 (Derivatives of splus**)** *For*

$$\mathrm{splus} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \ln(1+\exp(\cdot))$$

it holds

$$\mathrm{splus}^{(1)}(x) = \mathrm{sig}(x) \overset{\text{def.}}{=} 1/(1+\exp(-x))$$

and for $n \in \mathbb{N}$

$$\mathrm{splus}^{(n+1)}(x) = \mathrm{sig}^{(n)}(x) = \sum_{k=0}^{n} (-1)^{n+k} k! S_{n,k} \,\mathrm{sig}(x)(1-\mathrm{sig}(x))^k,$$

where $S_{n,k}$ *are the Stirling numbers of the second kind,* $S_{n,k} \overset{\text{def.}}{=} \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j} j^n$.

Proof We start with Faà di Bruno's formula,

$$\frac{d^n}{dt^n}\mathrm{sig}(x) = \frac{d^n}{dx^n}\frac{1}{f(x)} = \sum_{k=0}^{n}(-1)^k k! f^{-(k+1)}(x) B_{n,k}(f(x)),$$

where $f(x) \overset{\text{def.}}{=} 1+\exp(-x)$ and $B_{n,k}(f(x))$ denotes the Bell polynomials evaluated on $f(x)$. Next, we know the k-th derivative of $f(x)$ is given by

$$\frac{d^k}{dt^k} f(x) = (1-k)_k + k e^{-x}.$$

Now, using the definition of the Bell polynomials $B_{n,k}(f(t))$, we have

$$B_{n,k}(f(x)) = (-1)^n S_{n,k} e^{-kx},$$

where $S_{n,k}$ represents the Stirling numbers of the second kind. Substituting the expression for $B_{n,k}(f(x))$ into the derivative of $\mathrm{sig}(x)$, we obtain

$$\frac{d^n}{dx^n}\mathrm{sig}(x) = \sum_{k=0}^{n}(-1)^{n+k}k!S_{n,k}\,\mathrm{sig}(x)(1-\mathrm{sig}(x))^k.$$

■

Corollary 30 *In the setting of Theorem 29, for $n \in \mathbb{N}$,*

$$C^{\mathrm{splus}}(n) \leqslant \sum_{k=0}^{n}\frac{k^k k! S_{n,k}}{(k+1)^{k+1}}.$$

Proof For $k \in \mathbb{N}$ and $x \in [0,1]$, we have

$$f^k(x) \overset{\text{def.}}{=} x(1-x)^k, \quad (f^k)'(x) = (1-(k+1)x)(1-x)^{k-1};$$

which amounts to $(f^k)'(x) = 0$ at $1/(k+1)$, i.e.

$$\max_{x\in[0,1]} f(x) = \frac{1}{k+1}\left(\frac{k}{k+1}\right)^k = \frac{k^k}{(k+1)^{k+1}}.$$

■

Lemma 31 (Derivatives of GELU) *For*

$$\mathrm{GELU} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x\Phi(x),$$

it holds

$$\begin{aligned}\mathrm{GELU}'(x) &= \Phi(x) + x\varphi(x),\\ \mathrm{GELU}^{(n)}(x) &= n\varphi^{(n-2)}(x) + x\varphi^{(n-1)}(x), \quad n \geqslant 2,\end{aligned} \tag{33}$$

with $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $\Phi(x) = \int_{-\infty}^{x}\varphi(u)du$. The n-th derivative of $\varphi(x)$ is given by

$$\frac{d^n}{dx^n}\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\Big[\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{2k}2^k\frac{\Gamma(\frac{2k+1}{2})}{\Gamma(\frac{1}{2})}x^{n-2k}\Big].$$

Proof By induction we can show that (33) holds. And the representation of the n-th derivative of $\varphi(x)$ follows from de Oliveira and Ikeda (2012). ■

Corollary 32 *For $n \geqslant 2$, it holds in the setting of Theorem 31*

$$C^{\mathrm{GELU}}(n) \leqslant \frac{1}{\sqrt{2\pi}\Gamma(\frac{1}{2})}\left(na_{n-2}b_{n-2} + a_{n-1}c_{n-1}\right),$$

where

$$a_n \overset{\text{def.}}{=} \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k} 2^k \Gamma\left(\frac{2k+1}{2}\right), \quad b_n \overset{\text{def.}}{=} \max_{x\in\mathbb{R}} e^{-\frac{x^2}{2}} \sum_{k=0}^{\lfloor n/2 \rfloor} x^{n-k}, \quad c_n \overset{\text{def.}}{=} \max_{x\in\mathbb{R}} e^{-\frac{x^2}{2}} x \sum_{k=0}^{\lfloor n/2 \rfloor} x^{n-k}.$$

Lemma 33 (Derivatives of $\tanh$**)**
For $\tanh : \mathbb{R} \mapsto \mathbb{R}$*, the n-th derivatives have the representation*

$$\frac{d^n}{dx^n}\tanh x = C_n(\tanh x),$$
$$C_n(z) = (-2)^n(z+1)\sum_{k=0}^{n}\frac{k!}{2^k}\binom{n}{k}(z-1)^k, \quad n \geqslant 1.$$

Proof See Boyadzhiev (2007). ■

Corollary 34 *In the setting of Theorem 33, for $n \in \mathbb{N}$, $C^{\tanh}(n) = \max_{z\in[-1,1]} C_n(z)$.*

Lemma 35 (Derivatives of SWISH**)** *For*

$$\mathrm{SWISH} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{x}{1+e^{-x}}$$

it holds for $n \geqslant 1$

$$\frac{d^n}{dx^n}\mathrm{SWISH}(x) = n\sum_{k=1}^{n}(-1)^{k-1}(k-1)!S_{n,k}\,\mathrm{sig}^k(x) + x\sum_{k=1}^{n+1}(-1)^{k-1}(k-1)!S_{n+1,k}\,\mathrm{sig}^k(x),$$

where $S_{n,k}$ are Stirling numbers of the second kind, i.e.,

$$S_{n,k} = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(k-j)^n.$$

Proof By induction, we can show that

$$\frac{d^n}{dx^n}\mathrm{SWISH}(x) = n\,\mathrm{sig}^{(n-1)}(x) + x\,\mathrm{sig}^{(n)}(x), \quad n \geqslant 1.$$

By (Minai and Williams, 1993, Theorem 2), the derivatives of the sigmod function can be represented as

$$\mathrm{sig}^{(n)}(x) = \sum_{k=1}^{n+1}(-1)^{k-1}(k-1)!S_{n+1,k}\,\mathrm{sig}^k(x), \; n \geqslant 1.$$

Combining the above two equations, we obtain the general form of the n-th derivative of the SWISH function. ■

F.4.4. THE LAYER NORM

Lemma 36 (Bound of the Layer Norm for Derivative Type) *Fix $k \in \mathbb{N}$, $\beta \in \mathbb{R}^k$, $\gamma \in \mathbb{R}$, and $w \in [0,1]$. For the layer norm, given by*

$$\mathcal{LN} : \mathbb{R}^k \to \mathbb{R}^k, \quad x \mapsto \gamma\, f(x) g \circ \Sigma(x) + \beta;$$
$$f : \mathbb{R}^k \to \mathbb{R}^k, \quad x \mapsto x - M(x); \qquad g : \mathbb{R} \to \mathbb{R}, \quad u \mapsto \frac{1}{\sqrt{1+u}};$$
$$M : \mathbb{R}^k \to \mathbb{R}, \quad x \mapsto \frac{w}{k}\sum_{i=1}^{k} x_i; \qquad \Sigma : \mathbb{R}^k \to \mathbb{R}, \quad x \mapsto \frac{w}{k}\sum_{i=1}^{k}(x_i - M(x))^2;$$

holds for a compact symmetric set K (using Notation 3)

$$C_K^{\mathcal{LN}}(\alpha) \leqslant \alpha!\gamma \sum_{m=1}^{n=1} \frac{(2m+1)!!}{2^{2m}} \Bigg(\sum_{\substack{\alpha' \leqslant \alpha \\ |\alpha'|=n-1}} \sum_{\eta,\zeta \in \mathcal{P}'(\alpha',m)} \prod_{j=1}^{n} \frac{C_K^{\Sigma}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} + \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,m)} \prod_{j=1}^{n} \frac{C_K^{\Sigma}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \Bigg),$$

where $C_K^{\Sigma}(\alpha) = 2w\|K\|$ for $|\alpha| = 1$, $C_K^{\Sigma}(\alpha) = 2w$ for $|\alpha| = 2$, and $C_K^{\Sigma}(\alpha) = 0$ otherwise.

Proof Note that

$$g^{(n)}(x) = (-1)^n \frac{(2n+1)!}{n!2^{2n}}(1+x)^{-\frac{1}{2}-n},$$

implying $C_K^g(n) \leqslant (2n+1)!!2^{-2n}$, !! denoting the double factorial. We have further $C_K^f(\alpha) \leqslant \mathbb{1}_{|\alpha|=1}$ and a direct computation yields

$$C_K^{\Sigma}(\alpha) \leqslant \begin{cases} 2w\|K\| & \text{for } |\alpha| = 1 \\ 2w & \text{for } |\alpha| = 2 \\ 0 & \text{else.} \end{cases}$$

By Theorem 19,

$$C_K^{g\circ\Sigma}(\alpha) \leqslant \alpha! \sum_{m=1}^{n} C_{\Sigma[K]}^{g}(m) \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,m)} \prod_{j=1}^{n} \frac{C_K^{\Sigma}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}.$$

According to the general multivariate Leibnitz rule, it holds that

$$D^{\alpha}(f \cdot (g \circ \Sigma)) = \sum_{\beta \leqslant \alpha} \frac{\alpha!}{\beta!(\alpha-\beta)!} D^{\beta} f \cdot D^{\alpha-\beta}(g \circ \Sigma)$$

which implies

$$C_K^{\mathcal{LN}}(\alpha) \leqslant C_K^{g\circ\Sigma}(\alpha)\|K\| + \sum_{\beta \leqslant \alpha, |\beta|=1} \frac{\alpha!}{(\alpha-\beta)!} C_K^{g\circ\Sigma}(\alpha-\beta).$$

■

Corollary 37 (Bound of the Layer Norm for Derivative Level) *In the setting of Theorem 36, it holds that*

$$C_K^{\mathcal{LN}}(\leqslant n) \leqslant 2w\|K\|(2n+1)!!2^{-2n}(\|K\|+kn)\Big[\frac{2n}{e\ln n}(1+o(1))\Big]^n.$$

Furthermore, we have the asymptotic estimate

$$C_K^{g\circ\Sigma}(\leqslant n) \in \mathcal{O}\Big(w\|K\|\, n^{1/2}\big(\frac{n^{5/2}}{e^{3/4}\ln(n)}\,(1+o(1))\big)^n\Big).$$

Proof Analogue to the proof of Theorem 36,

$$C_K^{\mathcal{LN}}(\leqslant n) \leqslant \|K\|C_K^{g\circ\Sigma}(\leqslant n) + knC_K^{g\circ\Sigma}(\leqslant n-1) \leqslant (\|K\|+kn)C_K^{g\circ\Sigma}(\leqslant n),$$

where we can use Theorem 18 to bound

$$\begin{aligned} C_K^{g\circ\Sigma}(\leqslant n) &\leqslant C^g_{\Sigma[K]}(\leqslant n)C_K^{\Sigma}(\leqslant n)^n\Big[\frac{2n}{e\ln n}(1+o(1))\Big]^n \\ &\leqslant 2w\|K\|(2n+1)!!2^{-2n}\Big[\frac{2n}{e\ln n}(1+o(1))\Big]^n. \end{aligned} \tag{34}$$

Since $2n+1$ is odd, for each $n\in\mathbb{N}_+$, then sterling approximation *for double factorial* yields the asymptotic

$$(2n+1)!! \in \mathcal{O}\Big(\sqrt{2n}\,\Big(\frac{n}{e}\Big)^{n/2}\Big). \tag{35}$$

Merging (35) with the right-hand side of (34) yields

$$C_K^{g\circ\Sigma}(\leqslant n) \in \mathcal{O}\Big(w\|K\|\, n^{1/2}\big(\frac{n^{5/2}}{e^{3/4}\ln(n)}\,(1+o(1))\big)^n\Big).$$

■

F.4.5. THE MULTILAYER PERCEPTRON (FEEDFORWARD NEURAL NETWORK) WITH SKIP CONNECTION

Definition 38 (Single-Layer Feedforward Neural Network with Skip Connection) *Fix a non-affine activation function $\sigma\in C^\infty(\mathbb{R})$ and dimensions $d_{\text{in}}, d_{\text{ff}}, d_{\text{out}}\in\mathbb{N}$. A feedforward neural network is a map $\mathcal{PL}:\mathbb{R}^{d_{\text{in}}}\to\mathbb{R}^{d_{\text{out}}}$ represented for each $x\in\mathbb{R}^{d_{\text{in}}}$ by*

$$\mathcal{PL}(x) \overset{\text{def.}}{=} B^{(1)}x + B^{(2)}\big(\sigma\bullet(A\,x+a)\big) \tag{36}$$

for $A\in\mathbb{R}^{d_{\text{ff}}\times d_{\text{in}}}$, $a\in\mathbb{R}^{d_{\text{ff}}}$, $B^{(1)}\in\mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$, and $B^{(2)}\in\mathbb{R}^{d_{\text{out}}\times d_{\text{ff}}}$.

Lemma 39 (Bound of Neural Networks for Derivative Type) *In the notation of Notation 3, Theorem 25, and Theorem 38, it holds that*

$$C_K^{\mathcal{PL}}(\alpha) \leqslant C^{B^{(1)}}\mathbb{1}_{|\alpha|=1} + d_{\text{ff}}(\alpha!)^2C^{B^{(2)}}\sum_{m=1}^n C^{\sigma}_{h[K]}(m)\cdot(C^A)^m\sum_{\eta,\zeta\in\mathcal{P}'(\alpha,m)}\prod_{j=1}^n\frac{\mathbb{1}_{|\zeta^{(j)}|\leqslant 1}}{\eta^{(j)}!},$$

where $h[K]$ is defined as the image of $h(x)\overset{\text{def.}}{=}Ax+a$ on K.

Proof Write $\mathcal{PL}(x) = B^{(1)}x + B^{(2)}((g_i(x))_{i=1}^{d_{\text{ff}}})$, where for $i \in \{1, \dots, d_{\text{ff}}\}$

$$g_i(x) \overset{\text{def.}}{=} \sigma((Ax + a)_i).$$

If we define $h_i(x) \overset{\text{def.}}{=} h(x)_i$, we follow with Theorem 19

$$C_K^{g_i}(\alpha) \leqslant \alpha! \sum_{m=1}^{n} C_{h[K]}^{\sigma}(m) \cdot (C^A)^m \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,m)} \prod_{j=1}^{n} \frac{\mathbb{1}_{|\zeta^{(j)}| \leqslant 1}}{\eta^{(j)}!} \overset{\text{def.}}{=} C_K^g(\alpha),$$

and due to the component wise application of the activation function it holds that

$$\|D^\alpha \max_{i \in \{1,\dots,d_{\text{ff}}\}} g_i(x)\|_K = \max_{i \in \{1,\dots,d_{\text{ff}}\}} C_K^{g_i}(\alpha) \leqslant C_K^g(\alpha).$$

Using Theorem 20, we obtain

$$C_K^{\mathcal{PL}}(\alpha) \leqslant C^{B^{(1)}} \mathbb{1}_{|\alpha|=1} + d_{\text{ff}} \alpha! C^{B^{(2)}} C_K^g(\alpha).$$

■

Corollary 40 (Bound of Neural Networks for Derivative Level) *In the setting of Theorem 39,*

$$C_K^{\mathcal{PL}}(\leqslant n) \leqslant C^{B^{(1)}} + d_{\text{ff}} n! C^{B^{(2)}} C_{h[K]}^{\sigma}(\leqslant n)(C^A)^n \Big[\frac{2n}{e \ln n}(1 + o(1))\Big]^n.$$

If, moreover, $K = [-M_1, M_2]^{d_{\text{in}}}$ *then*

$$C_K^{\mathcal{PL}}(\leqslant n) \in \mathcal{O}\Big(\|B^{(1)}\|_\infty + \|B^{(2)}\|_\infty \|A\|_\infty^n \|\sigma\|_{n:\text{Ball}(a,\sqrt{d_{\text{in}}|M_1+M_2|})} \text{Width}(\mathcal{PL})\, n^{1/2} \big(\frac{n}{e}\big)^n C_n^n\Big)$$

Proof Arguing analogously to the proof of Theorem 39, barring the usage of Theorem 18, we obtain the estimate

$$C_K^{\mathcal{PL}}(\leqslant n) \leqslant C^{B^{(1)}} + d_{\text{ff}} n! C^{B^{(2)}} C_{h[K]}^{\sigma}(\leqslant n)(C^A)^n \Big[\frac{2n}{e \ln n}(1 + o(1))\Big]^n. \tag{37}$$

Let $C_n \overset{\text{def.}}{=} \frac{2n}{e \ln n}(1 + o(1))$ Using Stirling's approximation and the definition of the component-wise $\|\cdot\|_\infty$ norm of a matrix, (37) becomes

$$C_K^{\mathcal{PL}}(\leqslant n) \in \mathcal{O}\Big(\|B^{(1)}\|_\infty + \|B^{(2)}\|_\infty \|A\|_\infty^n C_{h[K]}^{\sigma}(\leqslant n) d_{\text{ff}}\, n^{1/2} \big(\frac{n}{e}\big)^n C_n^n\Big). \tag{38}$$

If, there is some $M_1, M_2 \leqslant 0$, such that $K = [0, \beta]^d$ then using the estimate between the $\|\cdot\|_2$ and $\|\cdot\|_\infty$ norms on $\mathbb{R}^{d_{\text{in}}}$ and the linearity of A we estimate

$$C_{h[K]}^{\sigma}(\leqslant n) \leqslant C_{\text{Ball}(a,\sqrt{d_{\text{in}}|M_1+M_2|})}^{\sigma}(\leqslant n) \leqslant \|\sigma\|_{n:\text{Ball}(a,\sqrt{d_{\text{in}}|M_1+M_2|})}.$$

Upon $\text{Width}(\mathcal{PL}) \overset{\text{def.}}{=} \max\{d_{\text{in}}, d_{\text{out}}, d_{\text{ff}}\}$, the estimate (37) implies that $C_K^{\mathcal{PL}}(\leqslant n)$ is of the order of

$$\mathcal{O}\Big(\|B^{(1)}\|_\infty + \|B^{(2)}\|_\infty \|A\|_\infty^n \|\sigma\|_{n:[-\|a\|_\infty - \sqrt{d_{\text{in}}|M_1+M_2|}, \|a\|_\infty + \sqrt{d_{\text{in}}|M_1+M_2|}]} \tag{39}$$
$$\times \text{Width}(\mathcal{PL})\, n^{1/2} \big(\frac{n}{e}\big)^n C_n^n\Big).$$

■

F.5. Step 2 (B) - Transformers

We may now merge the computations in Subsection F.4, with the Fa'a di Bruno-type from Section F.2 to uniformly bound the C^s-norms of the relevant class transformer networks. Our results are derived in two verions: the first is of "derivative type" (which is much smaller and more precise but consequentially more complicated) and the second is in "derivative level" form (cruder but simpler but also looser).

Theorem 41 (By Derivative Type) *Let K be a compact set, $\mathcal{TB}$ a transformer block as in Theorem 11, and $\alpha \in \mathfrak{O}_n^{Md_{\text{in}}}, n \in \mathbb{N}$. Then,*

$$C_K^{\mathcal{TB}}(\alpha) \leqslant \alpha! \sum_{\beta \in \mathfrak{O}_{\leqslant n}^{d_{\text{out}}}} N(\beta) C_{K^{(3)}}^{\mathcal{LN}}(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\alpha,\beta)} \prod_{j=1}^{n} \frac{C_K^{(3)}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}},$$

where for all $\gamma \in \mathfrak{O}_{\leqslant n}^{Md_{\text{in}}}$:

$$C_K^{(3)}(\gamma) \overset{\text{def.}}{=} \gamma! \sum_{\beta \in \mathfrak{O}_{\leqslant n}^{d_{\text{in}}}} N(\beta) C_{K^{(2)}}^{\mathcal{PL}}(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\gamma,\beta)} \prod_{j=1}^{n} \frac{C_K^{(2)}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}},$$

$$C_K^{(2)}(\gamma) \overset{\text{def.}}{=} \gamma! \sum_{\beta \in \mathfrak{O}_{\leqslant n}^{d_{\text{in}}}} N(\beta) C_{K^{(1)}}^{\mathcal{LN}}(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\gamma,\beta)} \prod_{j=1}^{n} \frac{C_K^{(1)}(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}},$$

$$C_K^{(1)}(\gamma) \overset{\text{def.}}{=} \mathbb{1}_{|\gamma|=1} + C_K^{\mathcal{MH}}(\gamma).$$

In the above, $K^{(1)} = \bigcup_{m=0}^{M} \mathcal{MH}_m[K]$, $K^{(2)} = \mathcal{LN}[K^{(1)}]$, and $K^{(3)} = \mathcal{PL}[K^{(2)}]$.

For respective multi-indices, a bound for $C_{K^{(3)}}^{\mathcal{LN}}, C_{K^{(1)}}^{\mathcal{LN}}$ is given by Theorem 36, $C_{K^{(2)}}^{\mathcal{PL}}$ is bounded in Theorem 39, and a bound for $C_K^{\mathcal{MH}}$ is given in Theorem 28.

Proof This is a direct consequence of Theorem 19. ■

Theorem 42 (By Derivative Level) *Let K be a compact set, $\mathcal{TB}$ a transformer block as in Theorem 11, and $n \in \mathbb{N}$. Then,*

$$C_K^{\mathcal{TB}}(\leqslant n) \leqslant C_{K^{(3)}}^{\mathcal{LN}}(\leqslant n) \left(d_{\text{out}} C_{K^{(2)}}^{\mathcal{PL}}(\leqslant n)\right)^n \left(d_{\text{in}}^2 C_{K^{(1)}}^{\mathcal{LN}}(\leqslant n)\right)^{n^2} \cdot \left(1 + C_K^{\mathcal{MH}}(\leqslant n)\right)^{n^3} \left[\frac{2n}{e \ln n}(1 + o(1))\right]^{n+n^2+n^3}$$

where, $K^{(1)} = \bigcup_{m=0}^{M} \mathcal{MH}_m[K]$, $K^{(2)} = \mathcal{LN}[K^{(1)}]$, and $K^{(3)} = \mathcal{PL}[K^{(2)}]$.

A bound for $C_{K^{(3)}}^{\mathcal{LN}}, C_{K^{(1)}}^{\mathcal{LN}}$ is given by Theorem 37, $C_{K^{(2)}}^{\mathcal{PL}}$ is bounded in Theorem 40, and a bound for $C_K^{\mathcal{MH}}$ is given in Theorem 28.

Proof Theorem 18 yields

$$C_K^{\mathcal{TB}}(\leqslant n) \leqslant C_{K^{(3)}}^{\mathcal{LN}}(\leqslant n) C_K^{(3)}(\leqslant n)^n \left[\frac{2d_{\text{out}} n}{e \ln n}(1 + o(1))\right]^n.$$

where

$$C_K^{(3)}(\leqslant n) \overset{\text{def.}}{=} C_{K^{(2)}}^{\mathcal{PL}}(\leqslant n) C_K^{(2)}(\leqslant n)^n \left[\frac{2d_{\text{in}} n}{e \ln n}(1+o(1))\right]^n,$$
$$C_K^{(2)}(\leqslant n) \overset{\text{def.}}{=} C_{K^{(1)}}^{\mathcal{LN}}(\leqslant n) C_K^{(1)}(\leqslant n)^n \left[\frac{2d_{\text{in}} n}{e \ln n}(1+o(1))\right]^n,$$
$$C_K^{(1)}(\leqslant n) \overset{\text{def.}}{=} 1 + C_K^{\mathcal{MH}}(\leqslant n),$$

which concludes the proof. ■

Theorem 43 (C^s-Norm Bound of Transformers) *Fix $n, L, H, C, D, d, M \in \mathbb{N}_+$ for a transformer class $\mathcal{TC}$. For any $\mathcal{T} \in \mathcal{TC}$, any compact $K_0 \subset \mathbb{R}^{M\times D}$, and any $\alpha \in \mathbb{N}^{M\times D}$, $|\alpha| \overset{\text{def.}}{=} n$ we have*

$$C_{K_0}^{\mathcal{T}}(\alpha) \leqslant d_{\text{out}}^L M\alpha! \cdot C^A \cdot C^L(\alpha), \tag{40}$$

where $C^1(\alpha) \overset{\text{def.}}{=} C_{K_0}^{\mathcal{TB}_1}(\alpha)$ and for $l \in \{2, \dots, L\}$,

$$C^l(\alpha) \overset{\text{def.}}{=} \leqslant \alpha! \sum_{\beta \in \mathfrak{D}_{\leqslant n}^{\tilde{d}_l}} N(\beta) C_{K_{l-1}}^{\mathcal{TB}_l}(\beta) \sum_{\eta,\zeta \in \mathcal{P}'(\mathfrak{o}(\alpha),\beta)} \prod_{j=1}^{n} \frac{C^l(\mathfrak{o}(\zeta^{(j)}))^{|\eta^{(j)}|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \tag{41}$$

where $K_l \overset{\text{def.}}{=} \mathcal{TB}_l[K_{l-1}]$, $\tilde{d}_l \overset{\text{def.}}{=} M_l d_{\text{in}}^l$, and a bound for $C_{K_{l-1}}^{\mathcal{TB}_l}(\beta)$ is given by Theorem 41, only depending on the transformer block class $\mathcal{TBC}_l$.

Proof [Proof of Theorem 43] The bounds (40) are a direct consequence of Theorem 19 and (41) follows directly from Theorem 20. ■

F.6. Step 2 (C) - Merging the C^s-Norm Bounds for Transformers with the Loss Function

In this section, we consider the following generalization of the class in Definition 1. As before, each result holds for input dimensions d just as much as any other input dimension, e.g. Md, with the only change being relabeling $d \leftarrow Md$. Therefore, for notational minimality, we chose to label the input dimension d and not dM.

Definition 44 (Smoothness Growth Rate) *Let $d, D \in \mathbb{R}$. A smooth function $g : \mathbb{R}^d \to \mathbb{R}^D$ is said to belong to the class $C_{poly:C,r}^{\infty}(\mathbb{R}^d, \mathbb{R}^D)$ (resp. $C_{exp:C,r}^{\infty}(\mathbb{R}^d, \mathbb{R}^D)$) if there exist $C, r \geqslant 0$ such that: for each $s \in \mathbb{N}_+$*

(i) ***Polynomial Growth -*** *$C_{poly:C,r}^{\infty}(\mathbb{R}^d, \mathbb{R}^D)$: $\|g\|_{C^s} \leqslant C\, s^r$,*

(ii) ***Exponential Growth -*** *$C_{exp:C,r}^{\infty}(\mathbb{R}^d, \mathbb{R}^D)$: $\|g\|_{C^s} \leqslant C\, e^{s\, r}$,*

The next lemma will help us relate the C^s-regularity of a model, a target function, and a loss function to their composition and product. We use it to relate the C^s-regularity of a transformed model $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^D$, the target function $f^\star : \mathbb{R}^d \to \mathbb{R}^D$, and the loss function $\ell : \mathbb{R}^{2D} \to \mathbb{R}$ to their composition

$$\begin{aligned} \ell_{\mathcal{T}} : \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto \ell\big(\mathcal{T}(x), f^\star(x)\big). \end{aligned} \tag{42}$$

One we computed have the C^s-regularity of $\ell_{\mathcal{T}}$, we can apply a concentration of measure-type argument based on an optimal transport-type duality, as in Amit et al. (2022); Hou et al. (2023b); Benitez et al. (2023); Kratsios et al. (2024), to obtain our generalization bounds. A key technical point where our analysis largely deviates from the mentioned derivations, is that we are not relying on any i.i.d. assumptions.

More generally, the next lemma allows us to bound the size of $\|\ell(\hat{f}, f^\star)\|_{C^s}$ using bounds on C^s norms of $\mathcal{TC}$ computed in Theorem 43, the target function $f^\star$, and on the loss function ℓ. Naturally, to use this result, we must assume a given level of regularity of the target function, as in Definition 1.

Lemma 45 (C^s-Norm of loss of between two functions) *Let $d, D, s \in \mathbb{N}_+$, $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}^D$ be of class C^s and $\ell : \mathbb{R}^{2D} \to \mathbb{R}$ be smooth. If there are constants $C_1, C_2, \widetilde{C}_1, \dots, \widetilde{C}_s \geqslant 0$ such that: $\|f_i\|_{C^s} \leqslant C_i$ for $i = 1, 2$ and for $j = 1, \dots, s$ we have $\|\ell\|_{C^j} \leqslant \widetilde{C}_j$ then for all $s > 0$ large it satisfies*

$$\big\|\ell(f_1, f_2)\big\|_{C^s} = \begin{cases} \mathcal{O}\big[\big(\frac{2Ds}{e \ln s}(1 + o(1))\big)^s\big], & \text{if } \max_{1\leqslant k} \widetilde{C}_k (C_1 C_2)^k \text{ is bounded,} \\ \mathcal{O}\big[\widetilde{C}_s \big(C_1 C_2 \frac{2Ds}{e \ln s}(1 + o(1))\big)^s\big], & \text{if } \max_{1\leqslant k} \widetilde{C}_k (C_1 C_2)^k \text{ is unbounded.} \end{cases} \tag{43}$$

Particularly, if $\ell \in C^\infty_{poly:C,r}(\mathbb{R}^{2D}, \mathbb{R})$, i.e., $\|\ell\|_{C^j} \leqslant C\, j^r$, then

$$\big\|\ell(f_1, f_2)\big\|_{C^s} = \mathcal{O}\big[C s^r \big(C_1 C_2 \frac{2Ds}{e \ln s}(1 + o(1))\big)^s\big]; \tag{44}$$

if $\ell \in C^\infty_{exp:C,r}(\mathbb{R}^{2D}, \mathbb{R})$, i.e., $\|\ell\|_{C^j} \leqslant C\, e^{j\, r}$, then

$$\big\|\ell(f_1, f_2)\big\|_{C^s} = \mathcal{O}\big[C e^{sr} \big(C_1 C_2 \frac{2Ds}{e \ln s}(1 + o(1))\big)^s\big]. \tag{45}$$

Lemma 45 allows us to obtain a bound on the term $\sup_{\hat{f} \in C^s_R(\mathbb{R}^d)} \|\ell(\hat{f}, f^\star)\|_{C^s}$ in (11), using Theorem 43 and our assumptions on ℓ and on $f^\star$.

Proof [Proof of Lemma 45] We first derive the general bound; which we then specialize to the case where the growth rate of ℓ is known. We first observe that

$$\begin{aligned} \big\|\ell(f_1, f_2)\big\|_{C^s} = & \underbrace{\max_{k=1,\cdots,s-1} \max_{\alpha \in \{1,\cdots,d\}^k} \|D^\alpha \ell(f_1(x), f_2(x))\|_\infty}_{\text{(VII)}} \\ & + \underbrace{\max_{\alpha \in \{1,\cdots,d\}^{s-1}} \operatorname{Lip}\big(D^\alpha \ell(f_1(x), f_2(x))\big)}_{\text{(VIII)}}. \end{aligned}$$

General Case - Term Term (VII): By Theorem 18, we have

$$\Big\|(D^\alpha \ell)(f_1(x), f_2(x))\Big\|_\infty \leqslant \big[\max_{1\leqslant k\leqslant s-1} \widetilde{C}_k (C_1 C_2)^k\big] \cdot \mathcal{O}\Big[\Big(\frac{2Dk}{e \ln k}(1 + o(1))\Big)^k\Big], \tag{46}$$

From (46) we have for all large $s > 0$ that

$$\begin{aligned} & \max_{k=1,\cdots,s-1} \max_{\alpha \in \{1,\cdots,d\}^k} \|D^\alpha \ell(f_1(x), f_2(x))\|_\infty \\ & = \begin{cases} \mathcal{O}\big[\big(\frac{2Ds}{e \ln s}(1 + o(1))\big)^s\big], & \text{if } \max_{1\leqslant k} \widetilde{C}_k (C_1 C_2)^k \text{ is bounded,} \\ \mathcal{O}\Big[\widetilde{C}_s \Big(C_1 C_2 \frac{2Ds}{e \ln s}(1 + o(1))\Big)^s\Big], & \text{if } \max_{1\leqslant k} \widetilde{C}_k (C_1 C_2)^k \text{ is unbounded.} \end{cases} \end{aligned}$$

General Case - Term Term (VIII): For each $\alpha \in \{1, \cdots, d\}^{s-1}$, by the multivariate Faà di Bruno formula, we have

$$D^{\alpha}\ell(f_1(x), f_2(x)) = \sum_{1 \leqslant |\beta| \leqslant s-1} (D^{\beta}\ell)(f_1(x), f_2(x)) \sum_{\eta,\zeta \in \mathcal{P}(\alpha,\beta)} \alpha! \prod_{j=1}^{s-1} \frac{[D^{\zeta^{(j)}}(f_1(x), f_2(x))]^{\eta^{(j)}}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}}.$$

The Lipschitz constants of the derivatives satisfy

$$\begin{aligned}
&\mathrm{Lip}\left(D^{\alpha}\ell(f_1(x), f_2(x))\right) \\
&= \sum_{1 \leqslant |\beta| \leqslant s-1} \mathrm{Lip}\left((D^{\beta}\ell)(f_1(x), f_2(x))\right) \sum_{\eta,\zeta \in \mathcal{P}(\alpha,\beta)} \alpha! \prod_{j=1}^{s-1} \frac{\mathrm{Lip}\left([D^{\zeta^{(j)}}(f_1(x), f_2(x))]^{\eta^{(j)}}\right)}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \\
&\leqslant \sum_{1 \leqslant |\beta| \leqslant s-1} \widetilde{C}_{|\beta|+1} \sum_{\eta,\zeta \in \mathcal{P}(\alpha,\beta)} \alpha! \prod_{j=1}^{s-1} \frac{(C_1C_2)^{|\eta(j)|}}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \\
&= \sum_{1 \leqslant |\beta| \leqslant s-1} \widetilde{C}_{|\beta|+1}(C_1C_2)^{|\beta|} \sum_{\eta,\zeta \in \mathcal{P}(\alpha,\beta)} \alpha! \prod_{j=1}^{s-1} \frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \\
&\leqslant \Big[\max_{1 \leqslant k \leqslant s-1} \widetilde{C}_{k+1}(C_1C_2)^{k}\Big] \sum_{1 \leqslant |\beta| \leqslant s-1} \sum_{\eta,\zeta \in \mathcal{P}(\alpha,\beta)} \alpha! \prod_{j=1}^{s-1} \frac{1}{\eta^{(j)}!(\zeta^{(j)}!)^{|\eta^{(j)}|}} \\
&= \Big[\max_{1 \leqslant k \leqslant s-1} \widetilde{C}_{k+1}(C_1C_2)^{k}\Big] \cdot \mathcal{O}\Big[\Big(\frac{2Ds}{e \ln s}(1 + o(1))\Big)^{s}\Big], \qquad (47)
\end{aligned}$$

where the last equality is due to Lemma 17.

From (47) we have for all $s > 0$ large that

$$\begin{aligned}
&\max_{\alpha \in \{1,\cdots,d\}^{s-1}} \mathrm{Lip}\left(D^{\alpha}\ell(f_1(x), f_2(x))\right) \\
&= \begin{cases} \mathcal{O}\big[\big(\frac{2Ds}{e \ln s}(1 + o(1))\big)^{s}\big], & \text{if } \max_{1 \leqslant k} \widetilde{C}_k(C_1C_2)^k \text{ is bounded,} \\ \mathcal{O}\big[\widetilde{C}_s\big(C_1C_2\frac{2Ds}{e \ln s}(1 + o(1))\big)^{s}\big], & \text{if } \max_{1 \leqslant k} \widetilde{C}_k(C_1C_2)^k \text{ is unbounded.} \end{cases}
\end{aligned}$$

Completing the General Case: Combining our estimates for terms Term (VII) and Term (VIII) respectively obtained in (46) and (47), we obtain an upper-bound for $\|\ell(f_1, f_2)\|_{C^s}$ via

$$\big\|\ell(f_1, f_2)\big\|_{C^s} = \begin{cases} \mathcal{O}\big[\big(\frac{2Ds}{e \ln s}(1 + o(1))\big)^{s}\big], & \text{if } \max_{1 \leqslant k} \widetilde{C}_k(C_1C_2)^k \text{ is bounded,} \\ \mathcal{O}\big[\widetilde{C}_s\big(C_1C_2\frac{2Ds}{e \ln s}(1 + o(1))\big)^{s}\big], & \text{if } \max_{1 \leqslant k} \widetilde{C}_k(C_1C_2)^k \text{ is unbounded.} \end{cases}$$

Special Cases of Interest: In particular, if ℓ belongs either to $C^{\infty}_{poly:C,r}(\mathbb{R}^d, \mathbb{R}^D)$ or to $C^{\infty}_{exp:C,r}(\mathbb{R}^d, \mathbb{R}^D)$, , as in Definition (1), then: there exists constants $C_\ell, r_\ell > 0$ s.t. for each $j = 1, \ldots, s$ we have

(i) **Polynomial Growth -** $C^{\infty}_{poly:C,r}(\mathbb{R}^{2D}, \mathbb{R})$ **Case:**

$$\|\ell\|_{C^j} \leqslant C\, j^r \overset{\text{def.}}{=} \tilde{C}_j,$$

(ii) **Exponential Growth -** $C^{\infty}_{exp:C,r}(\mathbb{R}^{2D},\mathbb{R})$ **Case:**

$$\|\ell\|_{C^j} \leqslant C\, e^{j\, r} \stackrel{\text{def.}}{=} \tilde{C}_j.$$

Consequentially, in cases (i) and (ii), the bound in (43) respectively becomes

(i) **Polynomial Growth -** $C^{\infty}_{poly:C,r}(\mathbb{R}^{2D},\mathbb{R})$ **Case:**

$$\big\|\ell(f_1,f_2)\big\|_{C^s} \leqslant \mathcal{O}\big[Cs^r\big(C_1C_2\frac{2Ds}{e\ln s}(1+o(1))\big)^s\big],$$

(ii) **Exponential Growth -** $C^{\infty}_{exp:C,r}(\mathbb{R}^{2D},\mathbb{R})$ **Case:**

$$\big\|\ell(f_1,f_2)\big\|_{C^s} \leqslant \mathcal{O}\big[Ce^{sr}\big(C_1C_2\frac{2Ds}{e\ln s}(1+o(1))\big)^s\big].$$

■

F.7. Step 3 - Combining Steps 1 and 2 and Completing The Proof of Theorem 3

We are now ready to complete the proof of our main result, namely Theorem 3. Before doing so, we state a more technical and general version, which we instead prove and which directly implies the simpler version found in the main body of our manuscript.

We operate under the following more general, but more technical set of assumptions than those considered in the main body of our text (in Setting 2.1).

Setting F.1 (Generalized Setting) *Let* $D,d,L,H,{*C'},C^A,C^b\in\mathbb{N}_+$, *set* $M\stackrel{\text{def.}}{=}0$, *and* $C\stackrel{\text{def.}}{=}({*C'},C^A,C^b)$, $r_f,r_\ell,C_f,C_\ell\geqslant 0$. *Suppose that Assumptions 3 and 4 hold.*

Fix a target function $f^\star:\mathbb{R}^d\to\mathbb{R}^D$ *and a loss function* $\ell:\mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}$. *Assume either that:*

(i) ***Polynomial Growth:*** $f^\star\in C^{\infty}_{poly:C_f,r_f}(\mathbb{R}^d,\mathbb{R}^D)$ *and* $\ell\in C^{\infty}_{poly:C_\ell,r_\ell}(\mathbb{R}^{2D},\mathbb{R})$,

(ii) ***Exponential Growth:*** $f^\star\in C^{\infty}_{exp:C_f,r_f}(\mathbb{R}^d,\mathbb{R}^D)$ *and* $\ell\in C^{\infty}_{exp:C_\ell,r_\ell}(\mathbb{R}^{2D},\mathbb{R})$,

(iii) ***No Growth:*** *There is a constant* $\bar{C}\geqslant 0$ *such that for all* $s>0$ *we have* $\|f^\star\|_{C^s},\|\ell\|_{C^s}\leqslant\bar{C}$.

Example 3 (Example of Generalized Setting (iii)) *For every* $d\in\mathbb{R}^d$, *the function* $f:\mathbb{R}^d\ni x\mapsto\cos\bullet x=\big(\cos(x_i)\big)_{i=1}^d$ *satisfies* $\|\frac{\partial^s}{\partial x_i^s}f\|_\infty\leqslant 1$ *for each* $s\in\mathbb{N}$ *and each* $i=1,\dots,d$. *Thus, it is an example of a function satisfying Assumption F.1.*

We are now ready to prove our main theorem, which is a combination of Theorems 2 and 3.

Theorem 46 (Pathwise Generalization Bounds for Transformers) *In Setting F.1, there is a* $\kappa\in(0,1)$, *depending only on* X., *and a* $t_0\in\mathbb{N}_0$ *such that: for each* $t_0\leqslant N\leqslant t\leqslant\infty$ *and* $\delta\in(0,1]$ *the following holds with probability at-least* $1-\delta$

$$\sup_{\mathcal{T}\in\mathcal{TC}}\big|\mathcal{R}_{\max\{t,N\}}(\mathcal{T})-\mathcal{R}^{(N)}(\mathcal{T})\big|\lesssim\sum_{s=1}^{\infty}I_{N\in[\tau_s,\tau_{s+1})}\,C_{\ell,\mathcal{TC},K,s-1}\bigg(I_{t<\infty}\,\kappa^t+\frac{\sqrt{2\ln(1/\delta)}}{N^{1/2}}+\mathrm{rate}_s(N)\bigg)$$

Table 6: Bounds on the terms in defining the constant $C_{\ell,\mathcal{TC},K,s}$, in Theorem 46, for a single attention block.

Term	Bound ($\mathcal{O}$)
$c_{\ell,f^\star}$	$C_f^s\, s^{r_\ell+2s^2} C_s^s$
$\mathcal{LN}$	$s^{(1+s)/2} C_s^s$
$\mathcal{PL}$	$\|B^{(1)}\| + \|B^{(2)}\| \|\|A\|^s \|\sigma\|_{s:\left[\pm\|a\|_\infty \pm\sqrt{d_{\text{in}}}\right]}$ $\text{Width}(\mathcal{PL})\, \tilde{C}_s^s$
$\mathcal{MH}$	$\|W\| \|\|V\| (\tilde{d} \|Q\| \|\|K\|)^s$ $\left(s^2 \left(\frac{s}{e}\right)^{2s} C_s^s\right)$

Here $C_s \overset{\text{def.}}{=} \frac{2s}{e \ln s}(1+o(1))$, $\tilde{C}_s \overset{\text{def.}}{=} s^{1/2}\left(\frac{n}{e}\right)^s C_s^s$, $c_d \overset{\text{def.}}{=} 2\max\{d_{\text{in}}, d_K, d_V, d_{\text{ff}}, d_{\text{out}}\}$, $\text{Width}(\mathcal{PL})$ is the width of the neural network $\mathcal{PL}$, where $\|\cdot\|$ denotes the componentwise max matrix/vector norm.

where $\text{rate}_s(N)$ *is defined in* (rate)*, the constant* $C_{\ell,\mathcal{TC},K,s} \overset{\text{def.}}{=} \sup_{\mathcal{T}\in\mathcal{TC}} \|\ell(\mathcal{T}, f^*)\|_{C^s}$*, is of order*

$$\mathcal{O}\Big(\underbrace{C^{\ell,f^\star}}_{\text{Loss \& Target}} \underbrace{C_{K^{(3)}}^{\mathcal{LN}}(\leqslant s)^s C_{K^{(1)}}^{\mathcal{LN}}(\leqslant s)^{s^3}}_{\text{Layernorms}} \underbrace{C_{K^{(2)}}^{\mathcal{PL}}(\leqslant s)^{s^2}}_{\text{Perceptron}} \underbrace{\left(1 + C_K^{\mathcal{MH}}(\leqslant s)\right)^{s^4}}_{\text{Multihead Attention}} \underbrace{D^{s^2}\, d^{2s^3}}_{\text{dimensions}} \underbrace{c_s^{s^s+s^3+s^4}}_{\text{Generic: } s\text{-th order Derivative}} \Big)$$

with terms according to Table 6 and the transition phases $(\tau_s)_{s=0}^\infty$ *are given iteratively by* $\tau_0 \overset{\text{def.}}{=} 0$ *and for each* $s \in \mathbb{N}_+$

$$\tau_s \overset{\text{def.}}{=} \inf\left\{ t \geqslant \tau_{s-1} : C_{\ell,\mathcal{TC},K,s}(\kappa^t + \text{rate}_s(N) + \tfrac{\sqrt{\log(1/\delta)}}{\sqrt{N}}) \leqslant C_{\ell,\mathcal{TC},K,s-1}(\kappa^t + \text{rate}_{s-1}(N) + \tfrac{\sqrt{\log(1/\delta)}}{\sqrt{N}}) \right\}.$$

Furthermore, $c \overset{\text{def.}}{=} 1-\kappa$*,* $c_2 \overset{\text{def.}}{=} c^{s/d}$*,* $\kappa^\infty \overset{\text{def.}}{=} \lim_{t\to\infty} \kappa^t = 0$*, and* $\lesssim$ *hides an absolute constant.*

Proof [Proof of Theorem 3] Since N is given, we may pick $s \in \mathbb{N}_+$ to ensure that $N \in [\tau_s, \tau_{s+1})$; where these are defined as in the statement of Theorem 46.

Since we are in Setting 2.1, then $\ell \in C^\infty_{poly:C_\ell,r_\ell}(\mathbb{R}^{2D}, \mathbb{R})$ (resp. $\ell \in C^\infty_{exp:C_\ell,r_\ell}(\mathbb{R}^{2D}, \mathbb{R})$) and $f^\star : \mathbb{R}^d \to \mathbb{R}^D$ is smooth. Therefore, Lemma 45 implies that there is an absolute constant $c_{\text{abs}} > 0$ such that for any transformer network $\mathcal{T} \in \mathcal{TC}$, the following bound holds

(i) **No Growth Case:** Using (43) we find that

$$\left\|\ell(\mathcal{T}, f^\star)\right\|_{C^s} \leqslant c_{\text{abs}} \left(\frac{2Ds}{e\ln s}(1+o(1))\right)^s \|\,\mathcal{T}\,\|_{C^s}^s \tag{48}$$

(ii) **Polynomial Growth Case -** $\ell \in C^\infty_{poly:C_\ell,r_\ell}(\mathbb{R}^{2D}, \mathbb{R})$ **Case:**

$$\left\|\ell(\mathcal{T}, f^\star)\right\|_{C^s} \leqslant c_{\text{abs}}\, s^{r_\ell} \left(\frac{2Ds}{e\ln s}(1+o(1))\right)^s \|f^\star\|_{C^s}^s \|\,\mathcal{T}\,\|_{C^s}^s \tag{49}$$

(iii) **Exponential Growth -** $C^\infty_{exp:C_\ell,r_\ell}(\mathbb{R}^{2D}, \mathbb{R})$ **Case:**

$$\left\|\ell(\mathcal{T}, f^\star)\right\|_{C^s} \leqslant c_{\text{abs}}\, e^{s\, r_\ell} \left(\frac{2Ds}{e\ln s}(1+o(1))\right)^s \|f^\star\|_{C^s}^s \|\,\mathcal{T}\,\|_{C^s}^s. \tag{50}$$

Since we have assumed that $f^\star \in C^\infty_{poly:C_f,r_f}(\mathbb{R}^d,\mathbb{R}^D)$ (resp. $C^\infty_{exp:C_f,r_f}(\mathbb{R}^d,\mathbb{R}^D)$ or the "no growth condition" in Setting F.1 (iii)) then the bounds in (48), (49), and (50), respectively, imply that

(i) **No Growth Case:**

$$\left\|\ell(\mathcal{T},f^\star)\right\|_{C^s} \leqslant c_{\text{abs}}\left(\frac{2Ds}{e\ln s}(1+o(1))\right)^s C_K^{\mathcal{TC}}(s)^s \tag{51}$$

(ii) **Polynomial Growth Case -** $\ell \in C^\infty_{poly:C_\ell,r_\ell}(\mathbb{R}^{2D},\mathbb{R})$ **Case:**

$$\left\|\ell(\mathcal{T},f^\star)\right\|_{C^s} \leqslant c_{\text{abs}}\, s^{r_\ell+2s^2}\left(\frac{C_f\,2D}{e\ln s}(1+o(1))\right)^s C_K^{\mathcal{TC}}(s)^s \tag{52}$$

(iii) **Exponential Growth -** $C^\infty_{exp:C_\ell,r_\ell}(\mathbb{R}^{2D},\mathbb{R})$ **Case:**

$$\left\|\ell(\mathcal{T},f^\star)\right\|_{C^s} \leqslant c_{\text{abs}}\, e^{s\,r_\ell+s^2 r_f}\left(\frac{2D\,s}{e\ln s}(1+o(1))\right)^s C_f^s\, C_{K_0}^{\mathcal{TC}}(s)^s, \tag{53}$$

where we have used the definition of the constant $C_K^{\mathcal{TC}}(s)$ as a uniform upper bound of $\sup_{\mathcal{T}\in\mathcal{TC}}$. Using Theorem 41 for the "derivative type estimate" (resp.42 for the "derivative level estimate") concludes the implies yields a uniform upper bound (of "derivative type" or "derivative level" respectively) on $C_{K_0}^{\mathcal{TC}}(s)$, i.e. independent of the particular transformer instance $\mathcal{T}\in\mathcal{TC}$. In either case, we respectively define $R>0$ to be the right-hand side of (52) or (53) depending on the respective assumptions made on ℓ and on $f^\star$.

The conclusion now follows upon applying Proposition 21 due to the inequality in (10). ■

Appendix G. Example of Additive Noise Using Stochastic Calculus

In this appendix, we briefly discuss why the seemingly *realizable* learning setting which we have placed ourselves in, i.e. $Y_n = f^\star(X_n)$, does not preclude additive noise. Our illustration considers the class of following Markov processes.

Assumption 5 (Structure on $X_\cdot$) *Let $g:\mathbb{R}^d\to[0,1]^d$ be a twice continuously differentiable function. Let $W_\cdot \overset{\text{def.}}{=} (W_t)_{t\geqslant 0}$ be d-dimensional Brownian motion and, for each $n\in\mathbb{N}$, define*

$$X_n \overset{\text{def.}}{=} g(W_n).$$

By construction, the boundedness of the change of variables-type function g in Assumption 5, implies that the process $X_\cdot = (X_n)_{n\in\mathbb{N}}$ is bounded (and can easily be seen to be Markovian since Brownian motion has the strong Markov property). However, we can say more, indeed under Assumption 5, the Itô Lemma (see e.g. (Cohen and Elliott, 2015, Theorem 14.2.4)) implies that X_n is given as the following stochastic differential equation (SDE) evaluated at integer times $n\in\mathbb{N}$

$$X_n = g(0) + \int_0^n \mu_s\,ds + \int_0^n \sigma_t^\top\,dW_s \tag{54}$$

where $\mu_\cdot = (\mu_t)_{t\geqslant 0}$ and $\sigma_\cdot = (\sigma_t)_{t\geqslant 0}$ are given by

$$\mu_t \overset{\text{def.}}{=} \frac{1}{2}\operatorname{tr}\left(H(g)(W_s)\right) \text{ and } \sigma_t \overset{\text{def.}}{=} \nabla g(W_t)$$

and $H(g)$ is the Hessian of g and tr is the trace of a matrix.

Example 4 *Set $d = 1$ and $g(x) = (\sin(x) + 1)/2$. Then, for each $n \in \mathbb{N}$ we have*

$$X_n = \int_0^n -\sin(W_s)/4 ds + \int_0^t \cos(W_s)/2 \, dW_s.$$

In particular, the expression (54) shows that the input process $X_\cdot$ is also defined for all intermediate times between non-negative integer times; i.e. for each $t \geqslant 0$ the process

$$X_t = g(0) + \int_0^t \mu_s \, ds + \int_0^t \sigma_t^\top \, dW_s \tag{55}$$

is well-defined and coincides with X_n whenever $t = n \in \mathbb{N}$. We may, therefore, also consider the "continuous-time extension" $Y_\cdot \overset{\text{def.}}{=} (Y_t)_{t \geqslant 0}$ of the target process defined for all intermediate times using (55) by

$$Y_t \overset{\text{def.}}{=} f^\star(X_t).$$

Note that Y_t coincides with the target process on non-negative integer times, as defined in our main text, by definition.

The convenience of these continuous-time extensions, of the discrete versions considered in our main text, is that now $Y_\cdot$ is the transformation of a continuous-time (Itô) process of satisfying the SDE (55) by a smooth function[4], namely $f^\star$. Therefore, we may again apply the Itô Lemma (again see e.g. (Cohen and Elliott, 2015, Theorem 14.2.4)) this time to the process $X_\cdot$ to obtain the desired signal and noise decomposition of the target process $Y_\cdot$ (both in discrete and continuous time). Doing so yields the following decomposition

$$\begin{aligned} Y_t = & \underbrace{f^\star(X_0) + \int_0^t \left((\nabla f^\star(X_s))^\top \mu_t + \frac{1}{2} \operatorname{tr}\left(\sigma_s^\top H(f^\star)(X_s) \sigma_s \right) \right) ds}_{\text{Signal (Target)}} \\ & + \underbrace{\int_0^t (\nabla f^\star)^\top \sigma_s \, dW_s}_{\text{Additive Noise}} . \end{aligned} \tag{56}$$

This shows that even if it a priori seemed that we are in the *realizable PAC setting* due to the structural assumption that $Y_n = f^\star(X_n)$ made when defining the target process, we are actually in the standard setting where the target data $(Y_n)_{n=0}^\infty$ can be written as a signal plus an additive noise term. Indeed, when $X_\cdot$ is simply a transformation of a Brownian motion by a bounded C^2-function, as in Assumption 5, then Assumption 1 held and Y_n admitted the signal-noise decomposition in (56).

4. Note that $f^\star$ was assumed to be smooth in our main result (Theorem 3).