PRESERVING REGULARITY IN FINITE-PRECISION TRANSFORMER APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers are the dominant architecture for large language models and are widely used in multimodal foundation models. Their substantial storage requirements have made low-bit quantization an important tool for practical deployment. Yet quantitative Transformer approximation theory predominantly treats weights as unconstrained real numbers, leaving a gap between theoretical guarantees and finite-precision representations. This question goes beyond approximation accuracy: quantization can increase a model's sensitivity to small input perturbations even when the approximation error remains small. It is therefore important to understand whether finite-precision Transformers can simultaneously achieve accurate approximation, controlled storage complexity, and preservation of a prescribed regularity bound. In this paper, we construct a finite-precision Transformer that approximates matrix-valued λ-Hölder functions on $[0,1]^{d\times n}$ while preserving a prescribed componentwise ℓ^1-Hölder bound L. Its hidden feature weights are discrete at $\{0,\pm 1,\pm\frac{1}{2}\}$ and target-independent, while target dependence is confined to the output coefficients. This separation allows us to directly control the effects of quantization on both approximation accuracy and regularity. For $D = dn$, the resulting Transformer achieves uniform error ε with $\mathcal{O}(\varepsilon^{-D/\lambda})$ nonzero parameters and a common $\mathcal{O}(\log(1/\varepsilon))$-bit fixed-point format for all parameters. These guarantees concern parameter storage under exact forward evaluation. As an application, empirical risk minimization over a structured finite-precision Transformer class achieves expected squared prediction risk $\mathcal{O}((\log(em)/m)^{2\lambda/(2\lambda+D)})$ under sub-Gaussian noise, matching the classical Hölder minimax exponent up to a logarithmic factor.

1 INTRODUCTION

Transformer architectures underpin modern large language models and many multimodal foundation models, making their approximation capabilities an important subject of theoretical study. Recent results establish explicit depth and parameter bounds for Hölder classes (He et al., 2026) and Sobolev targets (Jiao et al., 2026), together with nonparametric regression guarantees (Lai & Sun, 2026). However, these results treat stored parameters as unconstrained real numbers, so architectural complexity alone does not determine the precision or storage needed to achieve a given accuracy. Moreover, accurate approximation does not by itself ensure that the approximant preserves the target's regularity, which controls its sensitivity to input perturbations. This leaves an important question: can standard Transformers simultaneously achieve accurate approximation, controlled parameter storage, and preservation of a prescribed regularity bound under finite-precision parameterization?

Making stored precision explicit requires more than counting parameters. Feedforward approximation theory addresses this issue through quantized weights and encodable parameter classes, which provide finite representations of network parameters (Ding et al., 2019; Gühring & Raslan, 2021). Limited weight precision can itself become the principal constraint on approximation (Ou et al., 2024), while bit-complexity analyses relate approximation accuracy to the information required to encode a network (Mao & Xu, 2026). For standard Transformers approximating Hölder targets, however, a corresponding guarantee that also preserves the prescribed regularity bound remains, to our knowledge, unavailable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Achieving finite-precision approximation while preserving the prescribed regularity faces two fundamental obstacles. First, quantizing hidden weights can perturb intermediate representations and propagate errors through subsequent layers (Sakr et al., 2017; Ding et al., 2019; Frantar et al., 2023; Baek, 2026), complicating the control of both approximation error and output regularity (Lin et al., 2019; Furuya et al., 2026). Second, parameter discretization can violate the prescribed regularity bound even for structured interpolants. On a fine mesh, rounding coefficients can steepen local slopes, and the resulting perturbation to the Hölder seminorm is amplified by a factor that grows with the mesh resolution. At the resolution needed for ε-accuracy, an $\mathcal{O}(\varepsilon)$ bound on coefficient errors yields only an $\mathcal{O}(1)$ bound on the seminorm perturbation. Direct quantization therefore does not by itself guarantee the target bound L.

In this paper, we construct an explicit finite-precision Transformer for matrix-valued Hölder targets on $[0,1]^{d\times n}$, with flattened dimension $D = dn$ and componentwise bound $\|f_{pq}\|_{C^{0,\lambda}} \le L$, that simultaneously controls approximation accuracy, parameter storage, and output regularity. Our main contributions are summarized below.

- **Exact Transformer embedding with discrete hidden weights.** We construct a Transformer with one self-attention layer with one attention head whose target-independent hidden feature weights lie in $\{0, \pm 1, \pm\frac{1}{2}\}$, while target dependence is confined to the output-coefficient matrix (Theorem 4.1). Under exact real-arithmetic forward evaluation, all scalar parameters in the finite-precision construction share a common uniform fixed-point representation. For uniform approximation error ε, the nonzero parameter count M_{nz}, the common bit-width w_ε, and the total stored-value bit complexity satisfy $M_{\mathrm{nz}} = \mathcal{O}\big(\varepsilon^{-D/\lambda}\big)\,, w_\varepsilon = \mathcal{O}(\log(1/\varepsilon))\,, \mathrm{Bits}_{\mathrm{val}} = \mathcal{O}\big(\varepsilon^{-D/\lambda}\log(1/\varepsilon)\big)$ (Theorem 5.2).
- **Regularity-preserving finite-precision approximation.** Direct quantization of the output coefficients controls the uniform approximation error but guarantees only the relaxed seminorm bound $L + \tau_\varepsilon$, rather than the prescribed bound L. Contracting the coefficients before truncation creates a regularity margin that absorbs the coefficient perturbation. The resulting finite-precision approximant satisfies $[(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq}]_{\lambda,1} \le L$ for every component, without changing the proved upper-bound orders for the parameter count, common bit-width, or total stored-value bit complexity (Corollary 5.4).
- **Nonparametric regression with quantized Transformers.** The fixed-point parameterization defines a structured finite hypothesis class $\mathcal{T}_{N,s}$ whose statistical complexity is controlled by its log-cardinality. Under sub-Gaussian noise, empirical risk minimization over $\mathcal{T}_{N,s}$ achieves expected squared prediction risk $\mathcal{O}((\log(em)/m)^{2\lambda/(2\lambda+D)})$, matching the classical minimax exponent of Stone (1982) up to a logarithmic factor (Theorem 6.1).

2 Related Work

Quantitative Transformer approximation. Quantitative Transformer approximation has progressed from universality (Yun et al., 2020; Kajitsuka & Sato, 2024) to explicit rates for Hölder classes (He et al., 2026) and Sobolev spaces (Jiao et al., 2026), along with regression guarantees (Lai & Sun, 2026). Most results for standard Transformers quantify architectural resources such as depth, width, and parameter count while treating parameters as real-valued. For looped Transformer constructions, Xu & Sato (2025) additionally account for bounded weights and bit complexity (Theorem 3.6 therein). Our focus is on standard Transformer approximation of Hölder targets with simultaneous control of stored precision and output regularity.

ReLU conversion and functional regularity. Hu et al. (2026, Theorem 4.1) give a generic conversion from bounded ReLU networks to softmax-attention architectures. Their result concerns architectural simulation, whereas our analysis additionally tracks stored parameter precision and functional regularity. In another direction, Furuya et al. (2026, Theorem 8) analyze qualitative Lipschitz continuity for in-context Transformers on infinite-dimensional spaces $\mathcal{P}(\Omega)\times\Omega$, whereas our setting focuses on finite-dimensional Kuhn interpolants with explicit bit-width bounds. Lemma C.5 provides a constructive link by embedding the Kuhn-triangulation ReLU interpolant of Hong & Kratsios (2024) into a single-attention-layer Transformer via an exact algebraic identity, preserving the target Hölder bound L and every valid componentwise ℓ^1-modulus.

Table 1: Stored-precision and prescribed-regularity guarantees in closely related Transformer approximation results.

Work / setting	Stored precision	Prescribed regularity
Hu et al. (2026) *ReLU-to-attention conversion*	–	–
Xu & Sato (2025) *Looped Transformers*	$\mathcal{O}(\delta^{-(n+1)d})$ bits/weight[a]	–
Lai & Sun (2026) *Standard Transformers; $C^{s,\lambda}$ targets*	–	–
Jiao et al. (2026) *Standard Transformers; Hölder/Sobolev targets*	–	–
He et al. (2026) *Standard Transformers; Hölder targets*	–	–
This paper *Standard Transformers; Hölder targets*	$\boldsymbol{\mathcal{O}(\log(1/\varepsilon))}$ bits/parameter	✓

A dash indicates that the cited result does not establish the corresponding guarantee. Stored precision counts bits per parameter under exact evaluation.

[a]Xu & Sato (2025, Theorem 3.6) use mesh size δ and denote the sequence length by N, written as n here. Their bit complexity is the maximum number of bits required to represent a model weight.

Quantized and encodable approximation. While classical quantitative approximation theory for deep ReLU networks characterizes architectural complexity under unconstrained real parameters (Yarotsky, 2017), prior work on quantized and encodable networks shows that approximation guarantees depend on both the network architecture and the information stored in the weights (Ding et al., 2019; Gühring & Raslan, 2021; Ou et al., 2024; Mao & Xu, 2026). This literature does not address whether a standard Transformer can retain a prescribed regularity bound under finite-precision parameterization. In the standard Transformer setting considered here, we jointly bound the number of nonzero parameters and their fixed-point precision, while additionally controlling the regularity of the resulting approximant.

Finite-precision computation versus stored representation. A separate line of work studies Transformers whose computation is restricted to finite arithmetic precision. Merrill & Sabharwal (2023b;a) characterize computational expressivity under logarithmic precision, Chakrabarti et al. (2026) study expressivity thresholds across precision bit-widths, and Nowak et al. (2026) formalize memory semantics and compositionality under causal masking. Numerical inference stability and precision thresholds have also been investigated for arithmetic reasoning (Feng et al., 2025), mixed-precision stability (Budzinskiy et al., 2025), and layerwise risk (Baek, 2026). In our setting, parameter values are stored at explicitly bounded precision, whereas forward evaluation is idealized as exact real arithmetic.

3 Preliminaries

Notation. We write $[k] := \{1, \dots, k\}$ for $k \in \mathbb{N}_+$, with $\mathbb{N}_+ := \{1, 2, \dots\}$, $\mathbb{N} := \mathbb{N}_+ \cup \{0\}$, and $\mathbb{Z}$ denoting the set of integers. For any finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. Matrices and vectors are denoted by bold uppercase and lowercase letters, respectively. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_p$ denotes the ℓ^p-norm for $p \in [1, \infty]$. For a matrix $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{d\times n}$, its j-th column is $\boldsymbol{X}_{:,j}$, its i-th row is $\boldsymbol{X}_{i,:}$, and $\|\boldsymbol{X}\|_F$ denotes the Frobenius norm. $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix. Flattening $\boldsymbol{X}$ row-major into $\boldsymbol{X}^{(\mathrm{flt})} := (x_{11}, \dots, x_{1n}, x_{21}, \dots, x_{dn})^\top \in \mathbb{R}^D$ ($D := dn$) preserves the Euclidean norm $\|\boldsymbol{X}\|_F = \|\boldsymbol{X}^{(\mathrm{flt})}\|_2$. For a scalar function g, $\|g\|_{L^\infty(\Omega)} := \sup_{x\in\Omega} |g(x)|$ (abbreviated as $\|g\|_\infty$). For asymptotic complexity, we use standard notation: $a_n = \mathcal{O}(b_n)$ denotes an asymptotic upper bound ($|a_n| \le C|b_n|$ for some constant $C > 0$), while $a_n = \Theta(b_n)$ (or $a_n \asymp b_n$) denotes same-order scaling, meaning $c_1|b_n| \le |a_n| \le c_2|b_n|$ for universal or problem-dependent positive constants $c_1, c_2 > 0$. Appendix A indexes all notation.

Target functions and regularity. Without loss of generality, we consider inputs from $[0,1]^D$. A scalar target function belongs to the Hölder space $C^{0,\lambda}([0,1]^D)$, $0 < \lambda \leq 1$, with norm

$$|f|_{C^{0,\lambda}} := \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in[0,1]^D \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^{\lambda}}, \qquad \|f\|_{C^{0,\lambda}} := \max\{\|f\|_\infty, |f|_{C^{0,\lambda}}\}. \tag{1}$$

For a matrix-valued target $\boldsymbol{f} : [0,1]^{d\times n} \to \mathbb{R}^{d\times n}$ ($D = dn$), every scalar component f_{pq} satisfies $\|f_{pq}\|_{C^{0,\lambda}} \leq L$ under row-major flattening. we assume $L > 0$, since $L = 0$ is trivial. The regularity-preserving approximant we build on states its guarantees in terms of an ℓ^1-modulus of regularity. We therefore recall this notion.

Definition 3.1 (Modulus of regularity, Hong & Kratsios (2024, v4, Definition 3.1, p. 9))**.** *A function $\omega : [0,\tau] \to [0,\infty)$ is a* modulus of regularity *for $f : \Omega \to \mathbb{R}$ with respect to a norm $\|\cdot\|$ if it is nondecreasing, concave, $\omega(0) = 0$, and $|f(x) - f(y)| \leq \omega(\|x-y\|)$ for all $x, y \in \Omega$. Here τ is any number at least the diameter of Ω.*

Because $\|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq \|\boldsymbol{x}-\boldsymbol{y}\|_1$, every Hölder target with $\|f_{pq}\|_{C^{0,\lambda}} \leq L$ also satisfies the ℓ^1-Hölder seminorm bound $[g]_{\lambda,1} := \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in[0,1]^D \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|g(\boldsymbol{x})-g(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|_1^{\lambda}} \leq L$, which corresponds to the ℓ^1-modulus $\omega_H(t) = Lt^{\lambda}$. This ℓ^1-seminorm is the regularity measure controlled in our main results.

MLP and feedforward block conventions. Throughout, *MLP* (Multilayer Perceptron) and *FNN* (Feedforward Neural Network) are used synonymously for a vector-to-vector network, distinguished from a *Feedforward (FF) block* in a Transformer that applies a shared columnwise map to each token. An MLP $h : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ with $\Delta \geq 2$ affine maps is $h = A_\Delta \circ \sigma_R \circ A_{\Delta-1} \circ \cdots \circ \sigma_R \circ A_1$, where $A_\ell(\boldsymbol{z}) = \boldsymbol{W}^{(\ell)}\boldsymbol{z} + \boldsymbol{b}^{(\ell)}$ for $\ell \in [\Delta]$ and $\sigma_R(u) = \max\{0, u\}$ is applied elementwise. Its affine-map depth is Δ, its maximum hidden width is $\max_{1\leq\ell\leq\Delta-1} w_\ell$, and its stored parameter count is the number of nonzero entries in $(\boldsymbol{W}^{(\ell)}, \boldsymbol{b}^{(\ell)})_{\ell=1}^{\Delta}$. A Feedforward block $\boldsymbol{\mathcal{F}}_{FF} : \mathbb{R}^{d_{\text{in}}\times n} \to \mathbb{R}^{d_{\text{out}}\times n}$ with Δ affine maps applies such an MLP columnwise to $\boldsymbol{Z} \in \mathbb{R}^{d_{\text{in}}\times n}$, with shared weights and biases broadcast as $\boldsymbol{b}^{(\ell)}\boldsymbol{1}_{1\times n}$, so that $[\boldsymbol{\mathcal{F}}_{FF}(\boldsymbol{Z})]_{:,j} = h(\boldsymbol{Z}_{:,j})$ for $j \in [n]$.

Kuhn interpolation and underlying MLP. For an integer $N \geq 1$, consider the uniform grid $\{0, 1/N, \ldots, 1\}^D$ on $[0,1]^D$. The Kuhn triangulation (Kuhn, 1960) subdivides each grid cell into $D!$ simplices, each defined by a coordinate permutation whose vertex chain advances along coordinate directions. Enumerate the grid vertices as $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{G_N}$, where $G_N = (N+1)^D$. For each vertex $\boldsymbol{v}_k$, the associated Kuhn hat function $h_{k,N}$ is continuous, affine on each simplex of the Kuhn triangulation, equals one at $\boldsymbol{v}_k$, and vanishes at all other grid vertices. These hat functions are nonnegative and form a partition of unity on $[0,1]^D$. Ordering the target components row-major as $f_r := f_{pq}$ for $r = (p-1)n+q \in [D]$, following Hong & Kratsios (2024, Theorem 4.1) (Hong–Kratsios, hereafter HK), each component f_r is approximated by its *piecewise-affine Kuhn interpolant*

$$\Phi_r(\boldsymbol{x}) = \sum_{k=1}^{G_N} \boldsymbol{W}_{r,k}^{(\Delta)} h_{k,N}(\boldsymbol{x}), \qquad \boldsymbol{W}_{r,k}^{(\Delta)} = f_r(\boldsymbol{v}_k),$$

where the underlying hat functions $\{h_{k,N}\}$ are realized by a deep MLP.

Transformer architecture. We adopt the Transformer framework of Lai & Sun (2026) and define a Transformer $\boldsymbol{T} : \mathbb{R}^{d_{\text{in}}\times n} \to \mathbb{R}^{d_{\text{out}}\times n}$ of length K by $\boldsymbol{T} = \boldsymbol{\mathcal{F}}_{FF}^{(K)} \circ \boldsymbol{\mathcal{F}}_{SA}^{(K)} \circ \boldsymbol{\mathcal{F}}_{FF}^{(K-1)} \circ \cdots \circ \boldsymbol{\mathcal{F}}_{SA}^{(1)} \circ \boldsymbol{\mathcal{F}}_{FF}^{(0)} \circ \boldsymbol{\mathcal{F}}_{EB}$, initialized by an affine embedding $\boldsymbol{\mathcal{F}}_{EB}(\boldsymbol{X}) := \boldsymbol{W}_{EB}\boldsymbol{X} + \boldsymbol{B}_{EB} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n)\times n}$ that appends the fixed one-hot positional code $\boldsymbol{I}_n$. Set $d_0 := d_{\text{in}} + n$ and $d_{K+1} := d_{\text{out}}$. Each feedforward block $\boldsymbol{\mathcal{F}}_{FF}^{(k)} : \mathbb{R}^{d_k\times n} \to \mathbb{R}^{d_{k+1}\times n}$ has affine-map depth R_k and width W_k. Each self-attention layer $\boldsymbol{\mathcal{F}}_{SA}^{(k)} : \mathbb{R}^{d_k\times n} \to \mathbb{R}^{d_k\times n}$ with H_k heads and head dimension S_k is given by $\boldsymbol{\mathcal{F}}_{SA}^{(k)}(\boldsymbol{Z}) = \boldsymbol{Z} + \sum_{h=1}^{H_k} \boldsymbol{W}_O^{(k,h)} \boldsymbol{W}_V^{(k,h)} \boldsymbol{Z}\, \sigma_S\big(\boldsymbol{Z}^\top (\boldsymbol{W}_K^{(k,h)})^\top \boldsymbol{W}_Q^{(k,h)} \boldsymbol{Z}\big)$, where σ_S denotes columnwise softmax. Our construction specializes to length $K = 1$ with a single head ($H_1 = 1$), structured as $\boldsymbol{T} = \boldsymbol{\mathcal{F}}_{\text{post}} \circ \boldsymbol{\mathcal{F}}_{SA} \circ \boldsymbol{\mathcal{F}}_{\text{diag}} \circ \boldsymbol{\mathcal{F}}_{EB}$, where $\boldsymbol{\mathcal{F}}_{\text{diag}} := \boldsymbol{\mathcal{F}}_{FF}^{(0)}$ diagonalizes the input, $\boldsymbol{\mathcal{F}}_{SA} := \boldsymbol{\mathcal{F}}_{SA}^{(1)}$, and $\boldsymbol{\mathcal{F}}_{\text{post}} := \boldsymbol{\mathcal{F}}_{FF}^{(1)}$ evaluates the target network.

The network structure is characterized by its *dimension vector* $\dim(\boldsymbol{T}) := (d_{\text{in}}, d_0, \ldots, d_K, d_{\text{out}})$ and *size tuple* $\text{size}(\boldsymbol{T}) := \big((R_0, W_0), (H_1, S_1), \ldots, (R_K, W_K)\big)$. The entries of $\dim(\boldsymbol{T})$ specify the feature dimensions of the token representations, with d_0 after embedding and d_{out} at the output. Here R_k and W_k denote the affine-map depth and maximum hidden width of $\boldsymbol{\mathcal{F}}_{FF}^{(k)}$, and H_k and S_k the number of heads and head size of $\boldsymbol{\mathcal{F}}_{SA}^{(k)}$. We denote by $B_{EB}, B_{SA}, B_{FF} \geq 0$ the uniform magnitude bounds on the entries of the embedding, attention, and feedforward blocks, respectively, and by M_{EB}, M_{SA}, M_{FF} their stored nonzero parameter counts, with total stored nonzeros $M_{\text{nz}} := M_{EB} + M_{SA} + M_{FF}$.

4 Transformer Approximation with Discrete Hidden Weights

A standard obstacle in parameter quantization is that perturbing hidden weights alters intermediate representations, causing quantization errors to accumulate across layers (Sakr et al., 2017; Nagel et al., 2020; Frantar et al., 2023) and degrading regularity by amplifying perturbations across depth (Lin et al., 2019). The construction in this section bypasses this layerwise error accumulation entirely. Rather than quantizing an arbitrary continuous network, we design a Transformer whose feature-extraction layers contain only target-independent, discrete parameters. The remaining scaling and gating magnitudes are fixed structural constants. Target dependence enters strictly through the coefficient matrix $\boldsymbol{W}^{(\Delta)}$. Because the Kuhn interpolant is evaluated through exact algebraic identities, the network directly inherits the approximation error bound and regularity guarantees of the unquantized interpolant. Theorem 4.1 formalizes this result before we introduce coefficient quantization in Section 5.

Theorem 4.1 (Transformer Approximation with Discrete Hidden Weights)**.** *Let $0 < \lambda \leq 1$, $L > 0$, $D := dn$, and let $\boldsymbol{f} : [0,1]^{d\times n} \to \mathbb{R}^{d\times n}$ have components $f_{pq} \in C^{0,\lambda}([0,1]^D)$ satisfying $\|f_{pq}\|_{C^{0,\lambda}} \leq L$ and sharing a common ℓ^1-modulus of regularity $\omega : [0, D] \to [0, \infty)$. For every $\varepsilon \in (0,1)$, there exists a Transformer $\boldsymbol{T}_\varepsilon : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ such that the following hold.*

1. ***Architecture.*** *$\boldsymbol{T}_\varepsilon$ is a length $K = 1$ Transformer with a single self-attention layer and a single head ($H_1 = 1$), composed of the positional embedding $\boldsymbol{\mathcal{F}}_{EB} : \mathbb{R}^{d\times n} \to \mathbb{R}^{(d+n)\times n}$, the diagonalization block $\boldsymbol{\mathcal{F}}_{\text{diag}} = \boldsymbol{\mathcal{F}}_{FF}^{(0)}$, the gathering self-attention layer $\boldsymbol{\mathcal{F}}_{SA} = \boldsymbol{\mathcal{F}}_{SA}^{(1)}$, and the post-attention feedforward block $\boldsymbol{\mathcal{F}}_{\text{post}} = \boldsymbol{\mathcal{F}}_{FF}^{(1)}$:*

$$\boldsymbol{T}_\varepsilon = \boldsymbol{\mathcal{F}}_{\text{post}} \circ \boldsymbol{\mathcal{F}}_{SA} \circ \boldsymbol{\mathcal{F}}_{\text{diag}} \circ \boldsymbol{\mathcal{F}}_{EB}.$$

   It computes the piecewise-affine Kuhn interpolant $[\boldsymbol{T}_\varepsilon(\boldsymbol{X})]_{p,q} = \Phi_{(p-1)n+q}(\boldsymbol{X}^{(\text{flt})})$, $\boldsymbol{X} \in [0,1]^{d\times n}$. The underlying grid parameter is $N = \Theta\big(D(L/\varepsilon)^{1/\lambda}\big)$, with feedforward depth $\Delta + 2$ satisfying $\Delta \leq \lceil \log_2 D \rceil + 5$ and maximum width $\max\{W + n,\, D + n\}$ with $W \leq 8D(N+1)^D$. Its dimension vector and size tuple are

$$\dim(\boldsymbol{T}_\varepsilon) = (d,\, d+n,\, 2D+n,\, d),$$
$$\text{size}(\boldsymbol{T}_\varepsilon) = \big((2,\, 2D+n),\ (1,\, D),\ (\Delta+2,\, \max\{W+n,\, D+n\})\big).$$

2. ***Approximation and regularity preservation.*** *For every component $p \in [d]$ and $q \in [n]$,*

$$\|(\boldsymbol{T}_\varepsilon)_{pq} - f_{pq}\|_{L^\infty([0,1]^{d\times n})} \leq \varepsilon, \qquad [(\boldsymbol{T}_\varepsilon)_{pq}]_{\lambda,1} \leq L,$$

   and for all inputs $\boldsymbol{X}, \boldsymbol{Y} \in [0,1]^{d\times n}$, $\boldsymbol{T}_\varepsilon$ preserves the prescribed common ℓ^1-modulus:

$$|(\boldsymbol{T}_\varepsilon)_{pq}(\boldsymbol{X}) - (\boldsymbol{T}_\varepsilon)_{pq}(\boldsymbol{Y})| \leq \omega\big(\|\boldsymbol{X}^{(\text{flt})} - \boldsymbol{Y}^{(\text{flt})}\|_1\big).$$

3. ***Discrete hidden weights and weight bounds.*** *All routing and hidden parameters are target-independent and discrete: the embedding $\boldsymbol{\mathcal{F}}_{EB}$ and self-attention layer $\boldsymbol{\mathcal{F}}_{SA}$ use weights in $\{0, 1\}$; the diagonalization block $\boldsymbol{\mathcal{F}}_{\text{diag}}$ has weights and biases in $\{0, \pm 1\}$; the first hidden layer of $\boldsymbol{\mathcal{F}}_{\text{post}}$ has weights in $\{0, \pm 1\}$; and all internal hidden layers of $\boldsymbol{\mathcal{F}}_{\text{post}}$ have weights in $\{0, \pm\frac{1}{2}\}$. The structural constants in $\boldsymbol{\mathcal{F}}_{\text{post}}$ are the input scaling factor $c = nN$ and positional gating threshold $A = \lceil L \rceil$, with associated biases in $\{-A, 0, 1\}$. Target dependence is confined entirely to the output-coefficient matrix $\boldsymbol{W}^{(\Delta)} \in [-L, L]^{D\times G_N}$. The entrywise bounds satisfy $B_{EB} = 1$, $B_{SA} = 1$, and $B_{FF} = \max\{nN, 2A\}$.*

4. ***Parameter count.*** *The stored nonzero parameter counts of the constituent blocks satisfy*

$$M_{FF} = \mathcal{O}\big(D(N+1)^D\big), \qquad M_{SA} = \mathcal{O}(D), \qquad M_{EB} = d + n = \mathcal{O}(D).$$

Consequently, for fixed D, L, and λ, the total stored nonzeros scale as $M_{\mathrm{nz}} = \mathcal{O}(\varepsilon^{-D/\lambda})$, while both M_{SA} and M_{EB} remain entirely independent of the target accuracy ε.

Proof sketch. Choosing the grid resolution $N = \Theta\big(D(L/\varepsilon)^{1/\lambda}\big)$ ensures that the Kuhn interpolation error is bounded by ε. Because the Kuhn interpolant depends solely on the vertex values and the fixed hat functions, it preserves the prescribed modulus ω (Hong & Kratsios, 2024, Theorem 4.1). We realize this interpolant inside the Transformer in three stages. (i) *Embedding and diagonalization*. Append the fixed one-hot positional code $\boldsymbol{I}_n$ to $\boldsymbol{X}$ in n auxiliary channels, and let $\boldsymbol{\mathcal{F}}_{\mathrm{diag}}$ place x_{pq} in row $(p-1)n+q$ of column q. (ii) *Gathering and MLP evaluation*. Uniform attention $\boldsymbol{\mathcal{F}}_{SA}$ averages the columns to $\boldsymbol{X}^{(\mathrm{flt})}/n$. Multiplying by nN yields $N\boldsymbol{X}^{(\mathrm{flt})}$, and the feedforward layers evaluate the HK network identically across all tokens while the auxiliary channels carry the positional code. (iii) *Gating and output routing*. Since every token receives the full D-dimensional MLP output, the network uses the carried positional code to route the correct components to each token. Adding an offset $A = \lceil L \rceil$ allows a ReLU gate to retain the entries assigned to token q while zeroing out all others. The final affine layer subtracts A and redistributes these entries, recovering the exact $d \times n$ target matrix $[\boldsymbol{T}_\varepsilon(\boldsymbol{X})]_{p,q} = \Phi_{(p-1)n+q}(\boldsymbol{X}^{(\mathrm{flt})})$. Lemma C.5 establishes this exact output identity, transferring both the Hölder bound and modulus preservation to $\boldsymbol{T}_\varepsilon$. Appendices C and D provide the parameter matrices and complete proofs.

Remark 4.2 (Compatibility with residual feedforward layers)**.** *The feedforward blocks in our construction do not include skip connections around intermediate ReLU layers. If an architectural formulation strictly requires residual connections across every hidden layer, the technique of Jiao et al. (2026, Lemma A.1) can embed any token-wise network into strictly residual layers. Because this transformation introduces only auxiliary weights in $\{0, \pm 1\}$, it preserves our discrete parameter set without requiring higher precision, altering only the width and depth constants.*

5 Finite-Precision Approximation and Regularity

With all hidden and routing weights of $\boldsymbol{T}_\varepsilon$ restricted to fixed discrete sets and the scaling and gating magnitudes fixed by (n, N, L) in Theorem 4.1, the remaining task is to quantize the target-dependent coefficient matrix $\boldsymbol{W}^{(\Delta)}$ into finite bit-widths. Although fixed-point truncation directly controls uniform approximation error, preserving output regularity additionally requires controlling the amplification of coefficient perturbations in the ℓ^1-Hölder seminorm. To address this issue, Section 5.1 analyzes the stability of the Kuhn interpolant under coefficient perturbations and introduces coefficient contraction to retain the target Hölder bound. Section 5.2 then embeds these results into a uniform fixed-point representation, establishing our main approximation and storage complexity guarantees in Theorem 5.2 and Corollary 5.4.

5.1 Regularity Bounds for Coefficient Perturbations

For a target component f, let $a_k = f(\boldsymbol{v}_k)$ be the nodal value at grid node $\boldsymbol{v}_k$, and let q_k be a quantized coefficient with error $e_k = q_k - a_k$. The perturbation on the Kuhn interpolant is $\mathcal{E}_N := \sum_{k=1}^{G_N} e_k h_{k,N}$. While the partition of unity immediately bounds the uniform error by $\|\mathcal{E}_N\|_\infty \le \delta$ when $|e_k| \le \delta$, bounding the ℓ^1-Hölder seminorm requires controlling gradients across Kuhn simplices. Lemma 5.1 establishes these bounds and tracks the preserved regularity after multiplying the coefficients by $1-\rho$ before quantization.

Lemma 5.1 (Uniform and Hölder Bounds under Coefficient Perturbations)**.** *Fix $N \in \mathbb{N}_+$ and $0 < \lambda \le 1$. For any coefficient errors $|e_k| \le \delta$ with $\delta \ge 0$, the perturbation $\mathcal{E}_N := \sum_{k=1}^{G_N} e_k h_{k,N}$ on $[0,1]^D$ satisfies*

$$\|\mathcal{E}_N\|_\infty \le \delta, \qquad [\mathcal{E}_N]_{\lambda,1} \le 2\delta N^\lambda. \tag{2}$$

Let $L > 0$, $\|f\|_{C^{0,\lambda}([0,1]^D)} \le L$, and let $\Phi = \sum_{k=1}^{G_N} a_k h_{k,N}, a_k = f(\boldsymbol{v}_k), \|\Phi - f\|_\infty \le \eta$. If $0 \le \rho < 1$ and the perturbed coefficients q_k satisfy $|q_k| \le L$ and $|q_k - (1-\rho)a_k| \le \delta$, then $\Phi^q := \sum_{k=1}^{G_N} q_k h_{k,N}$ satisfies

$$\|\Phi^q\|_\infty \le L,$$

$$\|\Phi^q - f\|_\infty \le \underbrace{\eta}_{\text{interpolation error}} + \underbrace{\rho L}_{\text{contraction cost}} + \underbrace{\delta}_{\text{coefficient perturbation}}, \tag{3}$$

$$[\Phi^q]_{\lambda,1} \le \underbrace{(1-\rho)L}_{\text{contracted seminorm}} + \underbrace{2\delta N^\lambda}_{\text{seminorm perturbation}}. \tag{4}$$

The intuition behind this result is geometric and centers on how quantization affects local slopes. On any elementary Kuhn simplex, adjacent grid vertices are only $1/N$ apart. Rounding coefficients with tolerance δ can change an edge difference by up to 2δ, steepening local slopes by up to $2\delta N$, which underlies the $2\delta N^\lambda$ seminorm penalty in (2). Naive rounding can therefore push the Hölder seminorm beyond L. Contracting the coefficients by $1-\rho$ reduces the base seminorm bound to $(1-\rho)L$, leaving a margin of ρL that absorbs this penalty, so the quantized model still satisfies the target bound L. The cost is the uniform bias ρL in (3), which Section 5.2 keeps at the scale of ε by choosing a sufficient number of fixed-point bits.

5.2 Uniform Fixed-Point Quantization and Storage Bounds

Building on the perturbation bounds of Section 5.1, we quantize the output coefficients on a uniform fixed-point grid and derive bit-width bounds for stored parameters. Given a dynamic range bound $R > 0$ and a fractional-bit budget $s \in \mathbb{N}_+$, define

$$\mathcal{D}_s(R) := 2^{-s}\mathbb{Z} \cap [-R, R] = \left\{2^{-s}k : k \in \mathbb{Z},\ |k| \le \lfloor 2^s R \rfloor\right\}, \tag{5}$$

which contains at most $2^{s+1}(R+1)$ rational points. For any scalar value $z \in [-R, R]$, truncation toward zero is defined by

$$\mathsf{Q}_s(z) := 2^{-s}\operatorname{sgn}(z)\left\lfloor 2^s|z|\right\rfloor, \qquad \operatorname{sgn}(0) = 0. \tag{6}$$

By definition, the quantized value satisfies $\mathsf{Q}_s(z) \in \mathcal{D}_s(R)$, $|\mathsf{Q}_s(z)| \le |z|$, and $|\mathsf{Q}_s(z) - z| < 2^{-s}$. This guarantees that the round-off error on every parameter is bounded by 2^{-s}, directly matching the perturbation tolerance δ analyzed in Lemma 5.1.

Common fixed-point format. Let $\Theta_{\text{val}}(\boldsymbol{T})$ denote the collection of all scalar parameters in $\boldsymbol{T}$. When these parameters take values in $\mathcal{D}_{s_{\text{fp}}}(B)$ for a dynamic range bound $B \ge 1$, each entry requires at most $w = 1 + s_{\text{fp}} + \lceil \log_2(B+1) \rceil$ bits, covering one sign bit, s_{fp} fractional bits, and the necessary dynamic range. Storing the $M_{\text{nz}}(\boldsymbol{T})$ nonzero parameters thus requires $\text{Bits}_{\text{val}}(\boldsymbol{T}) = M_{\text{nz}}(\boldsymbol{T})w$ bits, representing storage footprint of the parameter values without structural indexing overhead. Applying this quantization scheme to Theorem 4.1 yields our main finite-precision approximation guarantee.

Theorem 5.2 (Finite-Precision Transformer Approximation with Controlled Regularity Bounds)**.** *Let $D = dn$, $L > 0$, and $0 < \lambda \le 1$. Let $f : [0,1]^{d\times n} \to \mathbb{R}^{d\times n}$ have components f_{pq} satisfying $\|f_{pq}\|_{C^{0,\lambda}} \le L$ under row-major flattening. For $0 < \varepsilon < 1$, set*

$$\eta = \frac{\varepsilon}{2}, \qquad N = \left\lceil \frac{D}{2}\left(\frac{2L}{\varepsilon}\right)^{1/\lambda} \right\rceil, \qquad s_{\text{fp}} = \left\lceil \log_2 \frac{2}{\varepsilon} \right\rceil. \tag{7}$$

Quantizing the output coefficients in Theorem 4.1 as $Q_{rk}^{\text{fp}} = \mathsf{Q}_{s_{\text{fp}}}(f_r(\boldsymbol{v}_k))$ yields a finite-precision Transformer $\boldsymbol{T}_\varepsilon^{\text{fp}}$ satisfying the following properties.

1. ***Approximation and regularity.*** *For every $p \in [d]$ and $q \in [n]$, $\|(\boldsymbol{T}_\varepsilon^{\text{fp}})_{pq} - f_{pq}\|_{L^\infty} \le \varepsilon$, $\|(\boldsymbol{T}_\varepsilon^{\text{fp}})_{pq}\|_{L^\infty} \le L$, and $[(\boldsymbol{T}_\varepsilon^{\text{fp}})_{pq}]_{\lambda,1} \le L + \tau_\varepsilon$ with $\tau_\varepsilon := 2^{1-s_{\text{fp}}}N^\lambda$.*
2. ***Common fixed-point representation.*** *All scalar parameters satisfy $\Theta_{\text{val}}(\boldsymbol{T}_\varepsilon^{\text{fp}}) \subseteq 2^{-s_{\text{fp}}}\mathbb{Z} \cap [-B_\varepsilon, B_\varepsilon]$ and admit representations of a common bit-width $w_\varepsilon = 1 + s_{\text{fp}} + \lceil \log_2(B_\varepsilon + 1) \rceil$, where $B_\varepsilon = \max\{nN, 2\lceil L \rceil\}$.*
3. ***Complexity orders.*** *$M_{\text{nz}}(\boldsymbol{T}_\varepsilon^{\text{fp}}) = \mathcal{O}(\varepsilon^{-D/\lambda})$, $w_\varepsilon = \mathcal{O}(\log(1/\varepsilon))$, and $\text{Bits}_{\text{val}}(\boldsymbol{T}_\varepsilon^{\text{fp}}) = \mathcal{O}(\varepsilon^{-D/\lambda}\log(1/\varepsilon))$.*

Remark 5.3 (Uniform format versus layer-wise precision)**.** *Theorem 5.2 packages every parameter into a single global fixed-point format to simplify theoretical accounting. The construction itself,*

however, is compatible with a layer-wise allocation of precision. The target-independent hidden weights and the attention weights require at most two code bits per entry. The structural scalars $c = nN$ and $A = \lceil L \rceil$ can be value-shared, while the first-layer grid biases and the output coefficient matrix carry the higher resolution. This mirrors common practice in low-bit network quantization, where internal projections are quantized aggressively while the embedding and output layers are kept at higher precision to avoid error propagation (Frantar et al., 2023; Lin et al., 2024; Ma et al., 2024).

While Theorem 5.2 controls uniform error with a fractional-bit count $s_{\mathrm{fp}} \asymp \log_2(1/\varepsilon)$, direct truncation leaves an $\mathcal{O}(1)$ perturbation $\tau_\varepsilon = 2^{1-s_{\mathrm{fp}}} N^\lambda$ in the Hölder seminorm, so the target bound L is no longer guaranteed. Retaining the exact bound L rather than the relaxed bound $L + \tau_\varepsilon$ requires absorbing this perturbation. By combining the coefficient contraction of Lemma 5.1 with a slightly larger fractional-bit count, Corollary 5.4 restores the bound L without changing the asymptotic parameter count or storage orders.

Corollary 5.4 (Quantized Approximation Preserving the Target Hölder Bound)**.** *Under the Hölder assumptions of Theorem 4.1, fix $0 < \varepsilon < 1$ and set*

$$\eta = \frac{\min\{\varepsilon, L\}}{3}, \qquad \rho = \frac{\eta}{L}, \qquad N = \left\lceil \frac{D}{2} \left(\frac{L}{\eta} \right)^{1/\lambda} \right\rceil, \qquad s_{\mathrm{fp}} = \left\lceil \log_2 \left(\frac{2N^\lambda}{\eta} \right) \right\rceil.$$

Quantizing the contracted coefficients as $Q_{rk}^{(\rho,\mathrm{fp})} = \mathsf{Q}_{s_{\mathrm{fp}}}((1-\rho) f_r(\boldsymbol{v}_k))$ produces a finite-precision Transformer $\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})}$ whose components satisfy

$$\|(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq} - f_{pq}\|_{L^\infty} \le \varepsilon, \qquad [(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq}]_{\lambda,1} \le L, \qquad \|(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq}\|_{L^\infty} \le (1-\rho)L$$

for all $p \in [d]$ and $q \in [n]$. This parameter schedule requires $s_{\mathrm{fp}} = 2\log_2(1/\varepsilon) + \mathcal{O}(1)$ fractional bits and preserves the same proved upper-bound orders $M_{\mathrm{nz}}(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})}) = \mathcal{O}(\varepsilon^{-D/\lambda}), w_\varepsilon = \mathcal{O}(\log(1/\varepsilon)), \mathrm{Bits}_{\mathrm{val}}(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})}) = \mathcal{O}(\varepsilon^{-D/\lambda}\log(1/\varepsilon))$.

6 Nonparametric Regression with Finite-Precision Transformers

We now study the statistical performance of the finite-precision Transformer class constructed above.

Regression model. Let $(\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^m$ be i.i.d. pairs in $[0,1]^{d\times n} \times \mathbb{R}^{d\times n}$ generated by

$$\boldsymbol{Y}_i = \boldsymbol{f}_0(\boldsymbol{X}_i) + \boldsymbol{\xi}_i, \qquad i = 1, \dots, m. \tag{8}$$

Fix $L > 0$ and $\lambda \in (0,1]$, and assume that each component of the unknown regression function $\boldsymbol{f}_0 : [0,1]^{d\times n} \to \mathbb{R}^{d\times n}$ satisfies $\|(f_0)_{pq}\|_{C^{0,\lambda}} \le L$. The noise matrices $\boldsymbol{\xi}_i \in \mathbb{R}^{d\times n}$ are conditionally mean-zero and sub-Gaussian with parameter σ^2 under the Frobenius inner product:

$$\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{X}_i] = \boldsymbol{0}_{d\times n}, \qquad \mathbb{E}\big[\exp\big(\langle \boldsymbol{u}, \boldsymbol{\xi}_i \rangle_F\big) \mid \boldsymbol{X}_i\big] \le \exp\left(\frac{\sigma^2}{2} \|\boldsymbol{u}\|_F^2 \right) \quad \text{a.s.,} \quad \forall \boldsymbol{u} \in \mathbb{R}^{d\times n}. \tag{9}$$

We evaluate predictors using the squared Frobenius loss $\ell(\boldsymbol{T}; \boldsymbol{X}, \boldsymbol{Y}) := \|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{Y}\|_F^2$. The prediction performance is measured by the population risk $\|\boldsymbol{T} - \boldsymbol{f}_0\|_{L^2(P_{\boldsymbol{X}})}^2 := \mathbb{E}_{\boldsymbol{X}} \|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{f}_0(\boldsymbol{X})\|_F^2$ evaluated on an independent test covariate $\boldsymbol{X} \sim P_{\boldsymbol{X}}$, with empirical risk $\frac{1}{m}\sum_{i=1}^m \|\boldsymbol{T}(\boldsymbol{X}_i) - \boldsymbol{Y}_i\|_F^2$ on the training sample.

Structured finite-precision hypothesis class. To formulate the statistical learning problem, we define a structured hypothesis class $\mathcal{T}_{N,s}$ of quantized Transformers indexed by grid resolution $N \in \mathbb{N}_+$ and fractional precision $s \in \mathbb{N}_+$. The deterministic prefix blocks $\boldsymbol{\mathcal{F}}_{SA} \circ \boldsymbol{\mathcal{F}}_{\mathrm{diag}} \circ \boldsymbol{\mathcal{F}}_{EB}$ are fixed to the constructive configuration of Section 4. Even if these entries were allowed to vary over their discrete value sets, they would contribute only an additive $\mathcal{O}(D)$ term to $\log|\mathcal{T}_{N,s}|$, which is negligible for the rates below. Within the post-attention block $\boldsymbol{\mathcal{F}}_{\mathrm{post}}$, the input-scaling factor nN and positional-gating constants $2A$ and $-A$ ($A = \lceil L \rceil$) remain fixed, while the variable parameters consist of an internal configuration $\vartheta \in \mathcal{P}_{N,s}$ and an output-coefficient matrix $Q \in \mathcal{D}_s(L)^{D\times G_N}$. Here ϑ specifies (i) the active hidden weights on the support $\mathcal{M}_N$ ($|\mathcal{M}_N| \le M_{FF} = \mathcal{O}(DG_N)$), taking values in $\{0, \pm 1\}$ in the first hidden layer and $\{0, \pm 1/2\}$ in internal hidden layers, and (ii) the

first-layer grid biases, taking integer values in $\{-N, \dots, N\}$. The finite set $\mathcal{P}_{N,s}$ contains the constructive choice $\vartheta^\star_N$, and $\mathcal{D}_s(L) = 2^{-s}\mathbb{Z} \cap [-L, L]$.

For $(\vartheta, Q) \in \mathcal{P}_{N,s} \times \mathcal{D}_s(L)^{D\times G_N}$, let $\boldsymbol{T}_{\vartheta,Q}$ denote the resulting Transformer. We append the elementwise clipping operator $\Pi_A(z) := \min\{A, \max\{-A, z\}\}$ to keep predictions in $[-A, A]^{d\times n}$. The resulting finite-precision hypothesis class is

$$\mathcal{T}_{N,s} := \left\{\Pi_A \circ \boldsymbol{T}_{\vartheta,Q} : \vartheta \in \mathcal{P}_{N,s},\ Q \in \mathcal{D}_s(L)^{D\times G_N}\right\}. \tag{10}$$

Theorem 6.1 (Nonparametric Regression with Quantized Transformers)**.** *In the regression setting above, let the sample size be $m \geq 2$. For the class $\mathcal{T}_{N_m,s_m}$ in* (10)*, choose grid resolution $N_m := \lceil \frac{D}{2}(\frac{m}{\log(em)})^{\frac{1}{2\lambda+D}} \rceil$ and fractional-bit count $s_m := \lceil \frac{\lambda}{2\lambda+D}\log_2 m \rceil$. Then any empirical risk minimizer $\hat{\boldsymbol{T}}_m \in \mathcal{T}_{N_m,s_m}$ for the squared Frobenius loss satisfies*

$$\mathbb{E}\left\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\right\|^2_{L^2(P_{\boldsymbol{X}})} = \mathcal{O}\left(\left(\frac{\log(em)}{m}\right)^{\frac{2\lambda}{2\lambda+D}}\right). \tag{11}$$

Remark 6.2 (Scope of the regularity guarantee)**.** *Corollary 5.4 guarantees the target Hölder bound only for the explicit construction obtained by contracting the coefficients before quantization. The ERM is then taken over the entire class without an explicit seminorm constraint, so its solution need not satisfy the same Hölder bound. Thus the guarantee of Theorem 6.1 is a prediction-risk bound rather than a regularity guarantee for the selected estimator. Imposing such a bound would require an additional constrained or regularized formulation.*

Remark 6.3 (Discussion of the convergence rate.)**.** *For real-valued Transformers, Lai & Sun (2026) showed that empirical risk minimization achieves the classical rate $m^{-2\lambda/(2\lambda+D)}$ for λ-Hölder regression (Stone, 1982) up to a logarithmic factor. Theorem 6.1 demonstrates that finite-precision parameterization attains the same rate, again up to a logarithmic factor. While continuous analyses control capacity via metric entropy or VC-dimension over continuous parameter manifolds (He et al., 2026; Jiao et al., 2026; Lai & Sun, 2026), our fixed-point parameterization naturally yields a finite hypothesis class whose metric entropy is bounded by its log-cardinality $\mathcal{O}(DG_{N_m}\log m)$. The fractional precision s_m is chosen to keep quantization bias negligible relative to spatial approximation error, while the grid resolution N_m balances spatial error against discrete variance. The logarithmic overhead in* (11) *arises from the capacity term $m^{-1}\log|\mathcal{T}_{N_m,s_m}|$ of the discrete, $\mathcal{O}(\log m)$-bit parameter class.*

7 Conclusion

This paper establishes explicit approximation and regularity-preservation guarantees for finite-precision Transformers approximating matrix-valued $C^{0,\lambda}$ targets. By embedding Kuhn interpolants into a single-attention-layer architecture with discrete hidden weights, our construction achieves uniform error ε using $\mathcal{O}(\varepsilon^{-D/\lambda})$ nonzero parameters and a common $\mathcal{O}(\log(1/\varepsilon))$-bit fixed-point format, while strictly preserving the prescribed ℓ^1-Hölder bound L. Crucially, pre-quantization coefficient contraction absorbs the N^λ-amplification of perturbations without increasing asymptotic parameter or bit complexity. For nonparametric regression, empirical risk minimization over this structured finite-precision class attains the classical Hölder minimax rate up to a logarithmic factor.

While our analysis focuses on stored parameter precision under exact evaluation and uniform attention, it opens several natural directions for future work, including (i) extending guarantees to full-network activation quantization, (ii) obtaining comparable approximation, bit-width, and regularity guarantees with content-adaptive attention, and (iii) developing computationally tractable estimators for the structured finite-precision class with risk bounds comparable to those of exact empirical risk minimization.

AI USE STATEMENT

In accordance with the ICLR author guidelines on AI assistance, the authors used large language models to assist with writing and proofreading, manuscript organization, literature search and cross-checking, and brainstorming of explanatory presentation. All cited sources were independently verified, and all mathematical statements, proofs, and final scientific claims were checked and approved by the authors, who take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

Appendix A collects the notation and norm conventions. The conversion, approximation and stored-precision proofs, and statistical class and risk analysis are given in Appendices C, D, E, and F.

REFERENCES

Jinwoo Baek. Numerical fragility in Transformers: A layer-wise theory for risk estimation and selective stabilization. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4987–4995. PMLR, 2026. URL https://proceedings.mlr.press/v300/baek26a.html.

Stanislav Budzinskiy, Wenyi Fang, Longbin Zeng, and Philipp Petersen. Numerical stability analysis of large language models. *arXiv preprint arXiv:2503.10251*, 2025. URL https://arxiv.org/abs/2503.10251.

Sayak Chakrabarti, Toniann Pitassi, and Josh Alman. Every bit counts: A theoretical study of precision-expressivity tradeoffs in quantized Transformers. *arXiv preprint arXiv:2602.02707*, 2026. URL https://arxiv.org/abs/2602.02707.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. On the universal approximability and complexity bounds of quantized ReLU neural networks. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJe9rh0cFX.

Guhao Feng, Kai Yang, Yuntian Gu, Xinyue Ai, Shengjie Luo, Jiacheng Sun, Di He, Zhenguo Li, and Liwei Wang. How numerical precision affects arithmetical reasoning capabilities of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 46–85. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.3. URL https://aclanthology.org/2025.findings-acl.3/.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. Approximation theory for Lipschitz continuous Transformers. *arXiv preprint arXiv:2602.15503*, 2026. URL https://arxiv.org/abs/2602.15503.

Ingo Gühring and Mones Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021. doi: 10.1016/j.neunet.2020.11.010. URL https://doi.org/10.1016/j.neunet.2020.11.010.

Xin He, Yuling Jiao, Xiliang Lu, and Jerry Zhijian Yang. Approximation error upper and lower bounds for Hölder class with Transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL https://openreview.net/forum?id=mWBxEAq9lv. arXiv:2605.07463.

Ruiyang Hong and Anastasis Kratsios. Bridging the gap between approximation and learning via optimal approximation by ReLU MLPs of maximal regularity. *arXiv preprint arXiv:2409.12335*, 2024. URL https://arxiv.org/abs/2409.12335.

Jerry Yao-Chieh Hu, Mingcheng Lu, Yi-Chen Lee, and Han Liu. Transformer approximations from ReLUs. *arXiv preprint arXiv:2604.24878*, 2026. URL https://arxiv.org/abs/2604.24878.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for Transformer networks with application to regression. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL https://openreview.net/forum?id=u7qkL3ryE4.

Tokio Kajitsuka and Issei Sato. Are Transformers with one layer self-attention using low-rank weight matrices universal approximators? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nJnky5K944. arXiv:2307.14023.

Harold W. Kuhn. Some combinatorial lemmas in topology. *IBM Journal of Research and Development*, 4(5):518–524, 1960. doi: 10.1147/rd.45.0518. URL https://doi.org/10.1147/rd.45.0518.

Yanming Lai and Defeng Sun. Standard Transformers achieve the minimax rate in nonparametric regression with $C^{s,\lambda}$ targets. *arXiv preprint arXiv:2602.20555*, 2026. URL https://arxiv.org/abs/2602.20555.

Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When fragile meets robust. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryetZ20ctX.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024. doi: 10.48550/arXiv.2402.17764. URL https://arxiv.org/abs/2402.17764.

Tong Mao and Jinchao Xu. Do neural networks really beat the curse of dimensionality? A bit-complexity view. *arXiv preprint arXiv:2608.01357*, 2026. URL https://arxiv.org/abs/2608.01357.

William Merrill and Ashish Sabharwal. A logic for expressing log-precision Transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35787–35812, 2023a. URL https://openreview.net/forum?id=uR8TtWCIsr.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision Transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023b. doi: 10.1162/tacl_a_00562. URL https://doi.org/10.1162/tacl_a_00562.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Franz Nowak, Ryan Cotterell, and Reda Boumasmoud. A compositional theory of causally masked Transformers. *arXiv preprint arXiv:2607.26988*, 2026. URL https://arxiv.org/abs/2607.26988.

Weigutian Ou, Philipp Schenkel, and Helmut Bölcskei. Three quantization regimes for ReLU networks. *arXiv preprint arXiv:2405.01952*, 2024. doi: 10.48550/arXiv.2405.01952. URL https://arxiv.org/abs/2405.01952.

Charbel Sakr, Jungwook Choi, Zhuo Wang, Kailash Gopalakrishnan, and Naresh Shanbhag. Analytical guarantees on numerical precision of deep neural networks. In *Proceedings of the 34th*

International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pp. 3007–3016. PMLR, 2017. URL https://proceedings.mlr.press/v70/sakr17a.html.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

Kevin Xu and Issei Sato. On expressive power of looped Transformers: Theoretical analysis and enhancement via timestep encoding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69613–69646. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25x.html.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002. URL https://doi.org/10.1016/j.neunet.2017.07.002.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

A Notation Reference and Architectural Conventions

Tables 2, 3, and 4 collect the main notation used throughout the paper. Each entry includes the symbol, its type or dimensions, a brief description, and the section or equation where it is first defined.

Table 2: Notation reference, Part I: dimensions, matrix operations, and function spaces.

Symbol	Type / Dimension	Description & First Definition
Matrix spaces and representations		
$d, n; D$	$\mathbb{N}_+$	Token dimension, sequence length, and flattened dimension $D = dn$ (§3)
$\boldsymbol{X}$	$\mathbb{R}^{d\times n}$	Input token matrix in $[0,1]^{d\times n}$ (§3)
$\boldsymbol{X}^{(\mathrm{flt})}$	$\mathbb{R}^{dn}$	Row-major flattening
$\widetilde{\boldsymbol{X}}$	$\mathbb{R}^{(d+n)\times n}$	Augmented input, (12)
$\boldsymbol{Z}$	$\mathbb{R}^{(2dn+n)\times n}$	Intermediate output of $\boldsymbol{\mathcal{F}}_{FF,1}$
$\|\boldsymbol{g}\|_{L^\infty,\ell^\infty(\Omega)}$	scalar	Supremum of the entrywise maximum norm on Ω (§C.4)
$\mathbf{u}_i$	$\mathbb{R}^D$	i-th standard unit basis vector in $\mathbb{R}^D$, (20)
$\boldsymbol{I}_n$	$\mathbb{R}^{n\times n}$	Fixed one-hot positional code block appended to the input (12)
$\mathrm{diag}(\boldsymbol{X})$	$\mathbb{R}^{dn\times n}$	Block-diagonal arrangement (Lemma C.1)
Function spaces and regularity		
$C^{0,\lambda}(\Omega)$	Banach space	Hölder space, (1)
$[g]_{\lambda,1}$	seminorm	Hölder seminorm on $[0,1]^D$ using the ℓ^1 metric (§3)
ω	$[0,\tau]\to[0,\infty)$	Modulus of regularity for $\tau \geq \mathrm{diam}_{\ell^1}(\Omega)$ (Def. 3.1)
Sets and asymptotic notation		
$\lvert\mathcal{S}\rvert$	$\mathbb{N}$	Cardinality of a finite set $\mathcal{S}$ (§3)
$\mathcal{O}(\cdot)$	asymptotic bound	Asymptotic upper bound: $\lvert a_n\rvert \leq C\lvert b_n\rvert$ for constant $C>0$ (§3)
$\Theta(\cdot), \asymp$	asymptotic order	Same-order scaling: $c_1\lvert b_n\rvert \leq \lvert a_n\rvert \leq c_2\lvert b_n\rvert$ for constants $c_1, c_2 > 0$ (§3)

Table 3: Notation reference, Part II: Transformer architecture and the Kuhn-triangulation MLP.

Symbol	Type / Dimension	Description & First Definition
Transformer architecture		
$\boldsymbol{\mathcal{F}}_{FF}$	$\mathbb{R}^{d_{in}\times n}\to\mathbb{R}^{d_{out}\times n}$	Feedforward block (column-wise)
$\boldsymbol{\mathcal{F}}_{SA}$	$\mathbb{R}^{d_{SA}\times n}\to\mathbb{R}^{d_{SA}\times n}$	Self-attention layer with skip connection
$\boldsymbol{\mathcal{F}}_{EB}$	$\mathbb{R}^{d_{\mathrm{in}}\times n}\to\mathbb{R}^{d_0\times n}$	Affine embedding layer
$\boldsymbol{T}$	$\mathbb{R}^{d_{in}\times n}\to\mathbb{R}^{d_{out}\times n}$	Full Transformer
K	$\mathbb{N}_+$	Transformer length; number of self-attention stages
$d_{\mathrm{in}}, d_k, d_{\mathrm{out}}$	$\mathbb{N}_+$	Channel dimensions in $\dim(\boldsymbol{T}) = (d_{\mathrm{in}}, d_0, \ldots, d_K, d_{\mathrm{out}})$, $k = 0,\ldots,K$
R_k, W_k	$\mathbb{N}_+$	Affine-map depth and maximum hidden width of $\boldsymbol{\mathcal{F}}_{FF}^{(k)}$, $k = 0,\ldots,K$
H_k, S_k	$\mathbb{N}_+$	Number of heads and head size of $\boldsymbol{\mathcal{F}}_{SA}^{(k)}$, $k = 1,\ldots,K$
B_{EB}, B_{FF}, B_{SA}	$[0,\infty)$	Weight bounds (embedding, FF, SA)
M_{EB}, M_{FF}, M_{SA}; M_{nz}	$\mathbb{N}$	Blockwise and total stored-nonzero counts (§3, §4, §5.2)
$\Theta_{\mathrm{val}}(\boldsymbol{T})$	set	Set of stored scalar parameter values in $\boldsymbol{T}$ (§5.2)

Symbol	Type / Dimension	Description & First Definition
$\mathrm{Bits}_{\mathrm{val}}(\boldsymbol{T})$	$\mathbb{N}$	Total bit count of stored scalar parameter values (§5.2)
w	$\mathbb{N}_+$	Uniform fixed-point bit-width per scalar (§5.2)
s_{fp}	$\mathbb{N}_+$	Fractional-bit count for dyadic quantization (§5.2, Theorem 5.2)
A	$\mathbb{N}_+$	Gating constant $\lceil L \rceil$, (17)
σ_R	$\mathbb{R} \to \mathbb{R}_{\geq 0}$	ReLU activation
σ_S	$\mathbb{R}^{n\times n} \to (0,1]^{n\times n}$	Column-wise softmax, $(\sigma_S(\boldsymbol{A}))_{ij} = \frac{e^{A_{ij}}}{\sum_{k=1}^{n} e^{A_{kj}}}$
Kuhn-triangulation MLP (Theorem B.1, Prop. B.2)		
Φ, $\widetilde{\Phi}$; $\boldsymbol{\Phi}$, $\widetilde{\boldsymbol{\Phi}}$	$\mathbb{R}^D \to \mathbb{R}$; $\mathbb{R}^D \to \mathbb{R}^D$	Global ReLU realizations of the scalar and vector-valued final-/integer-grid HK approximants. Their Kuhn-hat identities are used on $[0,1]^D$ and $[0,N]^D$, respectively
Φ_r, $\widetilde{\Phi}_r$	$\mathbb{R}^D \to \mathbb{R}$	Scalar components of Φ, $\widetilde{\Phi}$(vector-valued stacking convention in §B.1)
Δ	$\mathbb{N}_+$, $\Delta - 1 \leq \lceil \log_2 D \rceil + 4$	Affine-map count; Hong–Kratsios depth is $\Delta - 1$
W	$\mathbb{N}_+$, $\leq 8D(N+1)^D$	Width of the Kuhn-triangulation MLP
w_ℓ	$\mathbb{N}_+$, $w_\ell \leq W$	Width of hidden layer ℓ (§3)
N	$\mathbb{N}_+$	Grid parameter, $N = \Theta(D(L/\varepsilon)^{1/\lambda})$
G_N	$\mathbb{N}_+$, $(N+1)^D$	Number of Kuhn-hat functions at grid resolution N (vector-valued stacking convention in §B.1)
$\boldsymbol{v}_k$	$[0,1]^D$	Grid vertices $\boldsymbol{v}_k \in \{0, 1/N, \ldots, 1\}^D$, $k = 1, \ldots, G_N$ (§3)
$\boldsymbol{W}^{(1)}$	$\mathbb{R}^{d_1 \times D}$, $d_1 \leq W$	Stored first-layer weights, entries in $\{0, \pm 1\}$
$\boldsymbol{W}^{(l)}$ $(2 \leq l \leq \Delta - 1)$	$\mathbb{R}^{d_l \times d_{l-1}}$, $d_l \leq W$	HK internal hidden weights, entries in $\{0, \pm 1/2\}$
$\boldsymbol{W}^{(\Delta)}$	$\mathbb{R}^{D \times (N+1)^D}$	Coefficient matrix over Kuhn hat functions, $\max_{r,k} \lvert \boldsymbol{W}^{(\Delta)}_{r,k} \rvert \leq L$
$\boldsymbol{b}^{(1)}$	$\mathbb{R}^{d_1}$	First-layer biases, entries in $\{0, \pm 1, \ldots, \pm N\}$
$\boldsymbol{b}^{(l)}$ $(2 \leq l \leq \Delta - 1)$	$\mathbb{R}^{d_l}$, $d_l \leq W$	Comparison biases zero; HK penultimate hidden biases $+1$ (Prop. B.2)
$\boldsymbol{b}^{(\Delta)}$	$\mathbb{R}^D$	MLP output bias $(= \mathbf{0})$
$\hat{h}_k$	$[0,N]^D \to [0,1]$	Integer-grid Kuhn hat function
$\hat{\boldsymbol{h}}$	$[0,N]^D \to [0,1]^{G_N}$	Vector of integer-grid Kuhn hat functions (vector-valued stacking convention in §B.1)
$h_{k,N}$; $\boldsymbol{h}_N$	$[0,1]^D \to [0,1]$; $[0,1]^D \to [0,1]^{G_N}$	Kuhn hat function and stacked hat vector on the final domain $[0,1]^D$, $h_{k,N}(\boldsymbol{x}) = \hat{h}_k(N\boldsymbol{x})$

Table 4: Notation reference, Part III: FNN-to-Transformer conversion, self-attention, and generalization.

Symbol	Type / Dimension	Description & First Definition
FNN-to-Transformer conversion (§C)		
$\boldsymbol{\mathcal{F}}_{\mathrm{diag}} = \boldsymbol{\mathcal{F}}_{FF,1} = \boldsymbol{\mathcal{F}}^{(0)}_{FF}$	$\mathbb{R}^{(d+n)\times n} \to \mathbb{R}^{(2dn+n)\times n}$	Diagonalization block (Lemma C.1)
$\boldsymbol{\mathcal{F}}_{SA} = \boldsymbol{\mathcal{F}}^{(1)}_{SA}$	$\mathbb{R}^{(2dn+n)\times n} \to \mathbb{R}^{(2dn+n)\times n}$	Gathering SA (Lemma C.2)
$\boldsymbol{T}$	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	Converted Transformer, length $K = 1$ (Lemma C.5)
$\boldsymbol{T}^{\mathrm{fp}}_{\varepsilon}$	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	Finite-precision Transformer with direct truncation (Theorem 5.2)
$\boldsymbol{T}^{(\rho,\mathrm{fp})}_{\varepsilon}$	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	Finite-precision Transformer with coefficient contraction (Corollary 5.4)

Symbol	Type / Dimension	Description & First Definition
$\boldsymbol{\mathcal{F}}_{\text{post}} = \boldsymbol{\mathcal{F}}_{FF}^{(1)}$	$\mathbb{R}^{(2dn+n)\times n} \to \mathbb{R}^{d\times n}$	Post-attention feedforward block(Lemma C.4)
Self-attention weights (Lemma C.2)		
$\boldsymbol{W}_Q, \boldsymbol{W}_K$	$\mathbb{R}^{dn\times(2dn+n)}$	Query/Key matrices ($= \mathbf{0}$)
$\boldsymbol{W}_V$	$\mathbb{R}^{dn\times(2dn+n)}$	Value matrix, entries in $\{0,1\}$
$\boldsymbol{W}_O$	$\mathbb{R}^{(2dn+n)\times dn}$	Output matrix, entries in $\{0,1\}$
Generalization (§6)		
$\mathcal{F}$	finite set	Data-independent function class in Theorem E.1
$\boldsymbol{T}_{\vartheta,Q}$	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	Transformer parameterized by internal configuration ϑ and output matrix Q (§6, (10); §E.3)
$\mathcal{T}_{N,s}$	finite set	Clipped structured finite-precision Transformer class $\{\Pi_A \circ \boldsymbol{T}_{\vartheta,Q}\}$ (§6, (10))
ϑ; $\vartheta_N^\star$	$\in \mathcal{P}_{N,s}$	Internal parameter configuration and constructive choice (§6, §E.3)
$\mathcal{P}_{N,s}$	finite set	Finite set of allowable internal parameter configurations (§6, (10); §E.3)
Q; $Q^\star$	$\mathcal{D}_s(L)^{D\times G_N}$	Discrete output matrix and truncated constructive matrix (§6, Lemma E.3)
Q_s	$\mathbb{R} \to 2^{-s}\mathbb{Z}$	Deterministic dyadic quantizer truncating toward zero, (6)
$\mathcal{D}_s(R)$	$\subset [-R, R]$	Fixed-point codebook $2^{-s}\mathbb{Z} \cap [-R, R]$ (§5.2, (5))
N_m, s_m	$\mathbb{N}_+, \mathbb{N}_+$	Balanced grid resolution and fractional-bit count schedules (Theorem 6.1)
G_{N_m}	$\mathbb{N}_+$	Grid cardinality $(N_m+1)^D$
$\mathcal{M}_N$, $\lvert\mathcal{M}_N\rvert$	set, $\mathbb{N}$	Support of nonzero hidden MLP weights, and its cardinality (§6, §E.3)
$M_{\text{bias},N}$	$\mathbb{N}$	Number of first-layer grid biases(§6, §E.3)
Π_A	$\mathbb{R} \to [-A, A]$	Componentwise clipping map with $A = \lceil L \rceil$ (§6, §E.3)
$\boldsymbol{T}^*$; $\boldsymbol{T}^\star$	$\in \mathcal{F}$; $\in \mathcal{T}_{N_m,s_m}$	Reference predictor in $\mathcal{F}$; constructive quantized approximant $\Pi_A \circ \boldsymbol{T}_{\vartheta^\star_{N_m},Q^\star_m}$ (Theorem E.1, §E.3)
$\boldsymbol{T}_{N,Q^\star}$	$\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$	Constructive Transformer with truncated output coefficients (Lemma E.3)
$\hat{\boldsymbol{T}}_m$	$\in \mathcal{T}_{N_m,s_m}$	Empirical risk minimizer over the finite-grid Transformer class (Theorem 6.1)
$\lVert\cdot\rVert_{L^2(P_{\boldsymbol{X}})}$	seminorm	Population $L^2(P_{\boldsymbol{X}})$ seminorm
$\lVert\cdot\rVert_m$; $\langle\cdot,\cdot\rangle_m$	seminorm; bilinear form	Empirical L^2 seminorm and inner product (§E.1)
$\boldsymbol{f}_0$	$[0,1]^{d\times n} \to \mathbb{R}^{d\times n}$	True regression function
$\boldsymbol{\xi}_i$	$\mathbb{R}^{d\times n}$	Noise matrix, sub-Gaussian(σ^2)
σ^2	$[0,\infty)$	Conditional sub-Gaussian parameter

B Theoretical Background

B.1 The Hong–Kratsios (2025) approximation and its parameter realization

B.1.1 The interpolant and its regularity

For a scalar target $f : [0,1]^D \to \mathbb{R}$ and $N \in \mathbb{N}_+$, recall the Kuhn hat functions $\{h_{k,N}\}_{k=1}^{G_N}$ and the continuous piecewise-affine interpolant $\Phi(\boldsymbol{x}) = \sum_{k=1}^{G_N} f(\boldsymbol{v}_k) h_{k,N}(\boldsymbol{x})$ on $[0,1]^D$ from Section 3. Hong & Kratsios (2024, Theorem 4.1) establishes that this canonical interpolant preserves the prescribed ℓ^1-modulus of regularity of f and can be realized by a ReLU MLP with the explicit structural bounds restated below.

Theorem B.1 (Hong & Kratsios, 2024, Theorem 4.1)**.** *Let $f : [0,1]^D \to \mathbb{R}$ and let $\omega : [0, D] \to [0, \infty)$ be a modulus of regularity for f with respect to the ℓ^1 metric, in the sense of Definition 3.1. For any $N \in \mathbb{N}_+$, there exists a ReLU MLP $\Phi : \mathbb{R}^D \to \mathbb{R}$ represented by Δ affine maps, with $\Delta - 1 \leq \lceil \log_2 D \rceil + 4$, width at most $8D(N+1)^D$, and at most $16D(N+1)^D$ nonzero parameters, such that*

$$\|f - \Phi\|_{L^\infty([0,1]^D)} \leq \omega\left(\frac{D}{2N}\right),$$

and Φ interpolates f on the uniform grid $\{0, \frac{1}{N}, \ldots, 1\}^D$. The same ω is a modulus of regularity for Φ on $[0,1]^D$ with respect to the same ℓ^1 metric.

We call the ReLU realization specified below a *Kuhn-triangulation MLP*. We use Δ for its number of affine maps and W for its hidden width. The network is defined on $\mathbb{R}^D$, while the interpolation and regularity guarantees above concern its restriction to $[0,1]^D$.

B.1.2 Input scaling and stored parameters

To characterize the network parameters, we first construct the hat-function network on the integer grid $\{0, \ldots, N\}^D$, denoted by $\widetilde{\Phi} : \mathbb{R}^D \to \mathbb{R}$. Evaluating $\widetilde{\Phi}$ at $\boldsymbol{z} = N\boldsymbol{x}$ yields the desired interpolant on $[0,1]^D$. Decoupling this input scaling ensures that the first-layer weights belong to a fixed finite set.

Proposition B.2 (Parameter Realization of the HK Approximant)**.** *The interpolant supplied by Theorem B.1 admits the integer-grid realization*

$$\widetilde{\Phi}(\boldsymbol{z}) = \boldsymbol{W}^{(\Delta)}\sigma_R(\cdots\sigma_R(\boldsymbol{W}^{(1)}\boldsymbol{z} + \boldsymbol{b}^{(1)})\cdots) + \boldsymbol{b}^{(\Delta)},$$

with $\Phi(\boldsymbol{x}) = \widetilde{\Phi}(N\boldsymbol{x})$, the same depth and width upper bounds as in that theorem, and $\mathcal{O}(DG_N)$ nonzero parameters, where $G_N = (N+1)^D$. In this realization:

- *(i) $\boldsymbol{W}^{(\Delta)}$ consists of the target samples $f(\boldsymbol{v}_k)$ on the N-grid in $[0,1]^D$. All hidden weight matrices and all biases are target-independent.*
- *(ii) The HK penultimate hidden layer has $\boldsymbol{b}^{(\Delta-1)} = \mathbf{1}_{G_N \times 1}$, the intermediate comparison layers have $\boldsymbol{b}^{(l)} = \mathbf{0}$ for $2 \leq l \leq \Delta - 2$, and the final linear combination has $\boldsymbol{b}^{(\Delta)} = \mathbf{0}$.*
- *(iii) For $l = 2, \ldots, \Delta - 1$, $\boldsymbol{W}^{(l)}_{i,j} \in \{0, \pm 1/2\}$.*
- *(iv) $\boldsymbol{W}^{(1)}_{ij} \in \{0, \pm 1\}$, $\max_{i,j} |\boldsymbol{W}^{(\Delta)}_{ij}| \leq \max_k |f(\boldsymbol{v}_k)|$, and the first-layer bias vector $\boldsymbol{b}^{(1)}$ has dimension $2DG_N$ for $D \geq 2$ (and $4G_N$ for $D = 1$), with all entries being integer grid coordinates of magnitude at most N.*

Proof. Theorem 4.1 of Hong & Kratsios (2024) supplies the interpolation and regularity guarantees. We now describe a realization of the same interpolant with the stated parameter values, using the source's explicit formulas (Hong & Kratsios, 2024, v4, Eqs. (7.15)–(7.16), pp. 46–47).

Computing a single hat function. For an integer grid vertex $\boldsymbol{y}$, the integer-grid hat function is

$$\hat{h}_{\boldsymbol{y}}(\boldsymbol{z}) = \sigma_R\left(1 - \max_{i:\, y_i \text{ even}} |z_i - y_i| - \max_{i:\, y_i \text{ odd}} |z_i - y_i|\right), \qquad \boldsymbol{y} \in \{0, \ldots, N\}^D,$$

where the maximum of an empty group is zero. The two maxima group the coordinate distances according to whether y_i is even or odd. For $\boldsymbol{y} = N\boldsymbol{v}_k$, we denote this hat function by $\hat{h}_k$. The first layer computes $\sigma_R(z_i - y_i)$ and $\sigma_R(y_i - z_i)$, whose sum is $|z_i - y_i|$. It therefore uses only weights $\boldsymbol{W}^{(1)}_{ij} \in \{0, \pm 1\}$ and integer biases $b^{(1)}_r \in \{0, \pm 1, \ldots, \pm N\}$, with the active entries for vertex $\boldsymbol{y}$ taking values in $\{-y_i, y_i\}$.

Computing the maximum terms via a binary comparison tree. Each of the two maximum terms is computed via repeated pairwise comparisons. For two inputs a, b, the four ReLU features

$$\boldsymbol{u}(a,b) = \left(\sigma_R\left(\tfrac{a+b}{2}\right), \sigma_R\left(\tfrac{-a-b}{2}\right), \sigma_R\left(\tfrac{a-b}{2}\right), \sigma_R\left(\tfrac{-a+b}{2}\right)\right)$$

satisfy $\max\{a, b\} = u_1 - u_2 + u_3 + u_4$. Put $\kappa = \lceil \log_2 D \rceil$. For $D \geq 2$, arrange the comparisons in a balanced binary tree for each parity group, padding the leaves with zeros to 2^κ leaves; this

preserves the maximum because the leaf values are nonnegative. The first comparison takes the Layer-1 absolute-value sums as inputs, and each later comparison takes the four-feature sums. Since the two inputs occupy disjoint channels, the coefficients do not overlap or accumulate into integers, so the substitution produces only weights $\boldsymbol{W}_{ij}^{(l)} \in \{0, \pm 1/2\}$ and zero biases $\boldsymbol{b}^{(l)} = \mathbf{0}$.

Constructing the hat function with half-weights. The outer ReLU subtracts the two maxima from 1, so its bias must be $+1$. Directly feeding the final comparison features into this outer ReLU would require coefficients in $\{\pm 1\}$, violating the target set $\{0, \pm 1/2\}$. To avoid this without adding a layer, duplicate each feature in the last comparison layer by copying its affine row and bias. A feature u that would contribute to the hat function with coefficient $c \in \{0, \pm 1\}$ is then represented by two identical copies, each with coefficient $c/2$. The identity $(c/2)u + (c/2)u = cu$ preserves the value of the hat function and adds no layer. For $D = 1$, no comparison tree is needed. Duplicate the two first-layer features and use four weights $-1/2$ with bias $+1$ to obtain $\sigma_R(1 - |z_1 - y_1|)$. The unit bias is present in the source's Eq. (7.16) and its count of $2D + 1$ bias entries per hat function before Eq. (7.20), p. 48. We retain this bias from the explicit construction, despite the zero-intermediate-bias statement in Proposition 4.1(ii) of Hong & Kratsios (2024). This establishes statements (ii) and (iii) of Proposition B.2.

Layer-by-layer architecture. Table 5 summarizes the layer-by-layer structure of the scalar integer-grid network $\widetilde{\Phi} : \mathbb{R}^D \to \mathbb{R}$.

Table 5: Layer-by-layer architecture of the scalar Kuhn-triangulation MLP $\widetilde{\Phi}$.

Layer	Function / Operation	Width Bound	Weight Set	Bias Set
0	Input state $\boldsymbol{z} \in \mathbb{R}^D$	D	—	—
1	Differences $\sigma_R(\pm(z_i - y_i))$	$\leq 2DG_N$	$\{0, \pm 1\}$	$\{0, \pm 1, \ldots, \pm N\}$
Case $D \geq 2$ ($\kappa = \lceil \log_2 D \rceil \geq 1$):				
$2 \ldots \kappa$	Comparison trees	$\leq 8DG_N$	$\{0, \pm 1/2\}$	$\{0\}$
$\kappa + 1$	Terminal comparison & duplication	$\leq 16G_N \leq 8DG_N$	$\{0, \pm 1/2\}$	$\{0\}$
Case $D = 1$ ($\kappa = 0$):				
1	Differences & intra-layer duplication	$4G_N$	$\{0, \pm 1\}$	$\{0, \pm 1, \ldots, \pm N\}$
Common terminal stages:				
$\kappa + 2$	Hat basis evaluation: $\hat{h}_k(\boldsymbol{z})$	G_N	$\{0, \pm 1/2\}$	$\{+1\}$
$\Delta = \kappa + 3$	Output $\sum_k W_k^{(\Delta)} \hat{h}_k(\boldsymbol{z})$	1	$\{f(\boldsymbol{v}_k)\}_{k=1}^{G_N}$	$\{0\}$

Network size and output. For $D \geq 2$, the two comparison trees have width at most $4 \cdot 2^\kappa < 8D$ per hat subnetwork before duplication. Their duplicated final comparison layer has at most $16 \leq 8D$ features per hat subnetwork. For $D = 1$, the first layer has four features per hat function. The hidden depth is at most $\kappa + 2$, within the bound of Theorem B.1. Before duplication, each hat subnetwork uses $\mathcal{O}(D)$ tree nodes, each with at most four features and at most eight incoming entries per feature after substitution. Incomplete or empty parity groups (which have maximum zero) are padded with inactive zero leaves to 2^κ positions, depth-aligning all G_N subnetworks without corrupting active signals. Computing all G_N hat functions in parallel gives maximum hidden width $W \leq 8DG_N$ and $\mathcal{O}(DG_N)$ nonzeros, including terminal duplication and the unit biases.

Finally, the output layer forms a linear combination of the hat functions with scalar coefficients $W_k^{(\Delta)} = f(\boldsymbol{v}_k)$ and no additive constant. This gives (i) and the zero output bias in (ii). All preceding weights and biases depend solely on (D, N). The first layer contains at most $2D$ bias entries per hat subnetwork for $D \geq 2$, or four after duplication for $D = 1$, establishing the dimension counts in (iv). □

B.1.3 Vector-valued stacking convention

For a target $\boldsymbol{f}$ with flattened scalar components $(f_r)_{r=1}^D$, compute the hat functions once and use one output-coefficient row for each component. Write $\hat{\boldsymbol{h}}(\boldsymbol{z}) := (\hat{h}_1(\boldsymbol{z}), \ldots, \hat{h}_{G_N}(\boldsymbol{z}))^\top$ for the shared

integer-grid hat vector on $[0, N]^D$. Stacking the output rows of the scalar network gives the global ReLU MLP $\widetilde{\boldsymbol{\Phi}} = (\widetilde{\Phi}_r)_{r=1}^D : \mathbb{R}^D \to \mathbb{R}^D$. On $[0, N]^D$, it satisfies

$$\widetilde{\boldsymbol{\Phi}}(\boldsymbol{z}) = \boldsymbol{W}^{(\Delta)} \hat{\boldsymbol{h}}(\boldsymbol{z}), \qquad \boldsymbol{W}^{(\Delta)}_{r,k} = f_r(\boldsymbol{v}_k),$$

where $\boldsymbol{W}^{(\Delta)} \in \mathbb{R}^{D \times G_N}$ is the matrix of sampled target values. This is the *vector-valued integer-grid Kuhn-triangulation MLP* used in Appendix C.

Define the global final-domain map $\boldsymbol{\Phi} = (\Phi_r)_{r=1}^D : \mathbb{R}^D \to \mathbb{R}^D$ by $\boldsymbol{\Phi}(\boldsymbol{x}) := \widetilde{\boldsymbol{\Phi}}(N\boldsymbol{x})$. For $\boldsymbol{x} \in [0,1]^D$, set $h_{k,N}(\boldsymbol{x}) := \hat{h}_k(N\boldsymbol{x})$ and $\boldsymbol{h}_N(\boldsymbol{x}) := (h_{1,N}(\boldsymbol{x}), \ldots, h_{G_N,N}(\boldsymbol{x}))^\top$. Then $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{W}^{(\Delta)} \boldsymbol{h}_N(\boldsymbol{x})$ on $[0,1]^D$. Both hat vectors are nonnegative partitions of unity on their respective cubes. All components share the same hidden layers, so only the output-coefficient matrix depends on the target.

C FNN-TO-TRANSFORMER CONVERSION

We prove the conversion by refining Lemma 8 of Lai & Sun (2026). The proof has three steps: Lemma C.1 diagonalizes the token entries, Lemma C.2 gathers them, and Lemma C.4 combines scaling, MLP evaluation, and output routing. Their composition in Lemma C.5 preserves the hidden MLP weight sets, performs scaling and redistribution inside the architecture, and inherits the MLP modulus exactly (Remark C.6). Proposition D.1 collects the remaining parameter families.

C.1 AUGMENTED INPUT AND DIAGONALIZATION

Given an input matrix $\boldsymbol{X} \in [0,1]^{d \times n}$, define the *augmented input* $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{(d+n) \times n}$ as

$$\widetilde{\boldsymbol{X}} := \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{I}_n \end{pmatrix}. \tag{12}$$

The embedding $\boldsymbol{\mathcal{F}}_{EB} : \mathbb{R}^{d \times n} \to \mathbb{R}^{(d+n) \times n}$ given by $\boldsymbol{\mathcal{F}}_{EB}(\boldsymbol{X}) := \widetilde{\boldsymbol{X}}$ is realized by

$$\boldsymbol{W}_{EB} = \begin{pmatrix} \boldsymbol{I}_d \\ \boldsymbol{0}_{n \times d} \end{pmatrix} \in \mathbb{R}^{(d+n) \times d}, \qquad \boldsymbol{B}_{EB} = \begin{pmatrix} \boldsymbol{0}_{d \times n} \\ \boldsymbol{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n) \times n}. \tag{13}$$

Its entries lie in $\{0,1\}$, $B_{EB} = 1$, and its stored nonzero count is $M_{EB} = \mathrm{nnz}(\boldsymbol{W}_{EB}) + \mathrm{nnz}(\boldsymbol{B}_{EB}) = d + n = \mathcal{O}(D)$. The lower block $\boldsymbol{I}_n$ is a fixed one-hot positional code carried in n auxiliary channels: the j-th token carries the j-th column of $\boldsymbol{I}_n$, which is the j-th standard basis vector in $\mathbb{R}^n$. Thus this block plays the role of a positional embedding in a broad representational sense, but it is appended along the feature dimension rather than added to $\boldsymbol{X}$ as in the additive convention of Yun et al. (2020). In the affine embedding representation (13), the code is supplied by the fixed token-indexed coordinate matrix $\boldsymbol{B}_{EB}$, providing a deterministic position-augmented mapping rather than a token-invariant broadcast bias. The appended auxiliary-coordinate construction is adapted from Lemma 8 of Lai & Sun (2026), whose corresponding auxiliary block is $\boldsymbol{I}_n - \boldsymbol{1}_{n \times n}$. We use $\boldsymbol{I}_n$ so that the carried code is nonnegative and can pass through the ReLU layers unchanged. The added all-one offset is compensated by the -1 biases in the diagonalization block below.

Lemma C.1 (Diagonalization Block)**.** *There exists a feedforward block $\boldsymbol{\mathcal{F}}_{FF,1}$ of depth 2 and width $2dn + n$, mapping $\mathbb{R}^{(d+n) \times n} \to \mathbb{R}^{(2dn+n) \times n}$, with all weights in $\{0,1\}$ and all biases in $\{0,-1\}$, such that for any $\boldsymbol{X} \in [0,1]^{d \times n}$,*

$$\boldsymbol{\mathcal{F}}_{FF,1}(\widetilde{\boldsymbol{X}}) = \begin{pmatrix} \mathrm{diag}(x_{11}, x_{12}, \ldots, x_{1n}) \\ \mathrm{diag}(x_{21}, x_{22}, \ldots, x_{2n}) \\ \vdots \\ \mathrm{diag}(x_{d1}, x_{d2}, \ldots, x_{dn}) \\ \boldsymbol{I}_n \\ \boldsymbol{0}_{dn \times n} \end{pmatrix} \in \mathbb{R}^{(2dn+n) \times n}, \tag{14}$$

where $\mathrm{diag}(x_{i1}, \ldots, x_{in})$ *is the $n \times n$ diagonal matrix with x_{ij} in position (j, j).*

Proof. Define the first-layer weights $\boldsymbol{W}_1 \in \mathbb{R}^{(2dn+n)\times(d+n)}$ by the following block structure. For each feature dimension $i \in [d]$ and each position index $j \in [n]$, let the output row $r = (i-1)n + j$ (within the first dn rows) be:

$$(\boldsymbol{W}_1)_{r,:} := \big[\ \underbrace{0\ \cdots\ 0\ 1\ 0\ \cdots\ 0}_{d \text{ columns (position } i \text{ is 1)}}\ \big|\ \underbrace{0\ \cdots\ 0\ 1\ 0\ \cdots\ 0}_{n \text{ columns (position } j \text{ is 1)}}\ \big], \qquad b_r := -1.$$

The next n rows form $[0_{n\times d} \quad I_n]$ with bias 0. They copy the positional-code block I_n exactly. The bottom dn rows of $\boldsymbol{W}_1$ are all zeros. Set the remaining biases to 0.

For column q of $\widetilde{\boldsymbol{X}}$, the pre-activation at row $r = (i-1)n + j$ is:

$$(\boldsymbol{W}_1\widetilde{\boldsymbol{X}})_{r,q} + b_r = 1 \cdot x_{iq} + 1 \cdot (\boldsymbol{I}_n)_{jq} - 1 = x_{iq} + [\mathbf{1}_{j=q} - 1] = \begin{cases} x_{iq}, & j = q, \\ x_{iq} - 1, & j \neq q. \end{cases}$$

Since $x_{iq} \in [0, 1]$, applying ReLU to this case distinction gives

$$\sigma_R\big((\boldsymbol{W}_1\widetilde{\boldsymbol{X}})_{r,q} + b_r\big) = \begin{cases} x_{iq}, & j = q, \\ 0, & j \neq q. \end{cases}$$

Thus for each feature i, the n rows indexed by $(i-1)n+1$ through $i \cdot n$ form exactly $\mathrm{diag}(x_{i1}, \ldots, x_{in})$. The positional-code rows pass through the ReLU unchanged since $(\boldsymbol{I}_n)_{jq} \in \{0, 1\}$ is nonnegative. The second layer $\boldsymbol{W}_2 = \boldsymbol{I}_{2dn+n}$ (all entries in $\{0, 1\}$) with $\boldsymbol{b}_2 = \mathbf{0}$ preserves this output. The depth is 2, the width is $2dn + n$, all weights are in $\{0, 1\}$, and all biases are in $\{0, -1\}$. □

C.2 Token Gathering via Self-Attention

Attention reads the diagonalized input in the top dn rows of the $(2dn + n) \times n$ state and writes its uniform aggregate into the bottom dn rows. The skip connection preserves the middle n rows containing $\boldsymbol{I}_n$ for position-specific routing.

Lemma C.2 (Gathering Self-Attention)**.** *There exists a self-attention layer $\boldsymbol{\mathcal{F}}_{SA}$ mapping $\mathbb{R}^{(2dn+n)\times n} \to \mathbb{R}^{(2dn+n)\times n}$, with 1 head of size dn and all weight entries in $\{0, 1\}$, such that for $\boldsymbol{Z} = \boldsymbol{\mathcal{F}}_{FF,1}(\widetilde{\boldsymbol{X}})$ as in* (14)*,*

$$\boldsymbol{\mathcal{F}}_{SA}(\boldsymbol{Z}) = \begin{pmatrix} \mathrm{diag}(x_{11}, \ldots, x_{1n}) \\ \vdots \\ \mathrm{diag}(x_{d1}, \ldots, x_{dn}) \\ \boldsymbol{I}_n \\ \frac{1}{n}\,\boldsymbol{X}^{(flt)}\mathbf{1}_{1\times n} \end{pmatrix} \in \mathbb{R}^{(2dn+n)\times n}. \tag{15}$$

Proof. Choose the self-attention parameters as follows:

$$\boldsymbol{W}_Q := \mathbf{0}_{dn\times(2dn+n)}, \qquad \boldsymbol{W}_K := \mathbf{0}_{dn\times(2dn+n)},$$
$$\boldsymbol{W}_V := \begin{pmatrix} \boldsymbol{I}_{dn} & \mathbf{0}_{dn\times(dn+n)} \end{pmatrix} \in \mathbb{R}^{dn\times(2dn+n)},$$
$$\boldsymbol{W}_O := \begin{pmatrix} \mathbf{0}_{dn\times dn} \\ \mathbf{0}_{n\times dn} \\ \boldsymbol{I}_{dn} \end{pmatrix} \in \mathbb{R}^{(2dn+n)\times dn}.$$

With $\boldsymbol{W}_K = \boldsymbol{W}_Q = \mathbf{0}$, the attention score matrix is identically zero, so the softmax is uniform:

$$\sigma_S(\boldsymbol{Z}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{Z}) = \sigma_S(\mathbf{0}_{n\times n}) = \frac{1}{n}\mathbf{1}_{n\times n}.$$

Thus uniform averaging gathers the diagonalized entries into the flattened vector. The attention output before the skip connection is:

$$\boldsymbol{W}_O \boldsymbol{W}_V \boldsymbol{Z}\, \sigma_S(\cdots) = \begin{pmatrix} \mathbf{0}_{dn\times dn} \\ \mathbf{0}_{n\times dn} \\ \boldsymbol{I}_{dn} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{dn} & \mathbf{0}_{dn\times(dn+n)} \end{pmatrix} \begin{pmatrix} \mathrm{diag}(\boldsymbol{X}) \\ \boldsymbol{I}_n \\ \mathbf{0}_{dn\times n} \end{pmatrix} \cdot \frac{1}{n}\mathbf{1}_{n\times n}$$

$$= \begin{pmatrix} \mathbf{0}_{dn\times n} \\ \mathbf{0}_{n\times n} \\ \frac{1}{n}\operatorname{diag}(\boldsymbol{X})\mathbf{1}_{n\times n} \end{pmatrix}.$$

By definition, the block matrix $\operatorname{diag}(\boldsymbol{X})$ consists of diagonal blocks $\operatorname{diag}(\boldsymbol{X}_{i,:})$ for each row $i \in [d]$. Right-multiplying by the all-ones vector $\mathbf{1}_{n\times 1}$ sums each row block, which recovers the row-major flattening $\boldsymbol{X}^{(\text{flt})}$:

$$\operatorname{diag}(\boldsymbol{X})\mathbf{1}_{n\times 1} = \begin{pmatrix} \operatorname{diag}(\boldsymbol{X}_{1,:}) \\ \operatorname{diag}(\boldsymbol{X}_{2,:}) \\ \vdots \\ \operatorname{diag}(\boldsymbol{X}_{d,:}) \end{pmatrix} \mathbf{1}_{n\times 1} = \begin{pmatrix} \boldsymbol{X}_{1,:}^\top \\ \boldsymbol{X}_{2,:}^\top \\ \vdots \\ \boldsymbol{X}_{d,:}^\top \end{pmatrix} = \boldsymbol{X}^{(\text{flt})}.$$

Multiplying by the all-ones matrix $\mathbf{1}_{n\times n} = \mathbf{1}_{n\times 1}\mathbf{1}_{1\times n}$ then immediately yields $\operatorname{diag}(\boldsymbol{X})\mathbf{1}_{n\times n} = (\operatorname{diag}(\boldsymbol{X})\mathbf{1}_{n\times 1})\mathbf{1}_{1\times n} = \boldsymbol{X}^{(\text{flt})}\mathbf{1}_{1\times n}$. Adding the skip connection $\boldsymbol{Z}$, which has the diagonal blocks in the top dn rows and the positional code $\boldsymbol{I}_n$ in the middle n rows, gives (15). □

For the block-index notation of the general Transformer definition, set $\boldsymbol{\mathcal{F}}_{FF}^{(0)} := \boldsymbol{\mathcal{F}}_{FF,1}$, $\boldsymbol{\mathcal{F}}_{SA}^{(1)} := \boldsymbol{\mathcal{F}}_{SA}$. The subscripted or descriptive names are used in the local construction and the block-indexed names in the length-one architecture.

Remark C.3. *The factor $1/n$ in the bottom block of* (15) *arises from the uniform softmax averaging. Lai & Sun (2026) compensate for this factor by setting $(\boldsymbol{W}_O)_{dn+1:2dn,:} = n\boldsymbol{I}_{dn}$, which introduces a factor n in the self-attention parameters. We instead correct this factor by the scaling layer of the post-attention block below, keeping all self-attention weights in $\{0,1\}$ and all stored HK hidden MLP weight matrices unchanged.*

C.3 The Post-Attention Block: Scaling, MLP Evaluation, and Output Routing

The gathered state in (15) provides every token with the complete flattened input $\boldsymbol{X}^{(\text{flt})}$ alongside a positional code $\boldsymbol{I}_n$. The objective of the post-attention block $\boldsymbol{\mathcal{F}}_{\text{post}}$ is to process this data through the Kuhn-triangulation MLP and correctly assemble the 1D output back into a $d \times n$ matrix.

Lemma C.4 (Closed Post-Attention Realization)**.** *Let $\widetilde{\boldsymbol{\Phi}} = (\widetilde{\Phi}_r)_{r=1}^D : \mathbb{R}^D \to \mathbb{R}^D$ be the global vector-valued integer-grid Kuhn-triangulation MLP specified by the stacking convention in Section B.1, whose restriction to $[0,N]^D$ has penultimate Kuhn-hat representation $\hat{\boldsymbol{h}}$, represented by Δ affine maps and with width W, and suppose $\|\widetilde{\boldsymbol{\Phi}}(N\boldsymbol{x})\|_{\ell^\infty} \le L$ for $\boldsymbol{x} \in [0,1]^D$. The preceding stages form a feedforward block of affine-map depth $\Delta + 2$ and width $\max\{W + n,\ D + n\}$. For the input* (15)*, the composition $\boldsymbol{T} := \boldsymbol{\mathcal{F}}_{\text{post}} \circ \boldsymbol{\mathcal{F}}_{SA}^{(1)} \circ \boldsymbol{\mathcal{F}}_{FF}^{(0)} \circ \boldsymbol{\mathcal{F}}_{EB}$ satisfies, for all $\boldsymbol{X} \in [0,1]^{d\times n}$,*

$$[\boldsymbol{T}(\boldsymbol{X})]_{p,q} = \widetilde{\Phi}_{(p-1)n+q}\big(N\boldsymbol{X}^{(\text{flt})}\big), \qquad p \in [d],\ q \in [n], \tag{16}$$

exactly, computed inside the architecture. The resulting architecture is $\boldsymbol{\mathcal{F}}_{\text{post}} \circ \boldsymbol{\mathcal{F}}_{SA}^{(1)} \circ \boldsymbol{\mathcal{F}}_{FF}^{(0)} \circ \boldsymbol{\mathcal{F}}_{EB}$, a Transformer of length $K = 1$ with a single self-attention layer and a single head.

Proof. The proof constructs the feedforward block $\boldsymbol{\mathcal{F}}_{\text{post}}$ in three stages. The key invariant is that the upper block computes the hidden MLP state replicated across all tokens, while the lower block acts as an identity channel carrying the positional code $\boldsymbol{I}_n$.

Step 1: Scaling. With $c = nN$, the first affine layer extracts and scales the gathered attention output (15) via a $(D+n)\times(2D+n)$ block matrix, discarding the redundant top D diagonal channels. Since all inputs are nonnegative, the ReLU is inactive:

$$\boldsymbol{\Psi}^{(0)} = \begin{pmatrix} \mathbf{0}_{D\times D} & \mathbf{0}_{D\times n} & c\boldsymbol{I}_D \\ \mathbf{0}_{n\times D} & \boldsymbol{I}_n & \mathbf{0}_{n\times D} \end{pmatrix} \begin{pmatrix} \operatorname{diag}(\boldsymbol{X}) \\ \boldsymbol{I}_n \\ \frac{1}{n}\boldsymbol{X}^{(\text{flt})}\mathbf{1}_{1\times n} \end{pmatrix} = \begin{pmatrix} N\boldsymbol{X}^{(\text{flt})}\mathbf{1}_{1\times n} \\ \boldsymbol{I}_n \end{pmatrix}.$$

Step 2: Shared MLP evaluation. Write $h^{(0)}(\boldsymbol{z}) := \boldsymbol{z}$ and $h^{(\ell)}(\boldsymbol{z}) := \sigma_R\big(\boldsymbol{W}^{(\ell)}h^{(\ell-1)}(\boldsymbol{z}) + \boldsymbol{b}^{(\ell)}\big)$ for $1 \le \ell \le \Delta - 1$, so that $h^{(\Delta-1)} = \hat{\boldsymbol{h}}$ and $\widetilde{\boldsymbol{\Phi}}(\boldsymbol{z}) = \boldsymbol{W}^{(\Delta)}\hat{\boldsymbol{h}}(\boldsymbol{z})$. The next $\Delta - 1$ layers augment the

MLP weights with an identity block:

$$\mathbf{\Psi}^{(\ell)} = \sigma_R\left(\begin{pmatrix} \boldsymbol{W}^{(\ell)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_n \end{pmatrix} \mathbf{\Psi}^{(\ell-1)} + \begin{pmatrix} \boldsymbol{b}^{(\ell)} \\ \mathbf{0}_n \end{pmatrix} \mathbf{1}_{1\times n}\right) = \begin{pmatrix} h^{(\ell)}(N\boldsymbol{X}^{(\text{flt})})\mathbf{1}_{1\times n} \\ \boldsymbol{I}_n \end{pmatrix}.$$

This block matrix multiplication identically propagates the shared hidden state in the top block and the positional code in the bottom block.

Step 3: Position-gated routing. We now route the $D = dn$ dimensional shared MLP output $\widetilde{\mathbf{\Phi}}(N\boldsymbol{X}^{(\text{flt})})$ into the correct $d \times n$ matrix format. To make the token-routing explicit, we define the broadcasted positional code $\mathbf{P} \in \mathbb{R}^{D\times n}$ (stacking d copies of $\boldsymbol{I}_n$) and the block-summation matrix $\mathbf{S} \in \mathbb{R}^{d\times D}$:

$$\mathbf{P} := \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{I}_n \\ \vdots \\ \boldsymbol{I}_n \end{pmatrix}, \qquad \mathbf{S} := \begin{pmatrix} \mathbf{1}_{1\times n} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{1\times n} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{1\times n} \end{pmatrix}.$$

Setting $A := \lceil L \rceil$, the Transformer gating affine layer computes a gated matrix $\mathbf{\Psi}^{(\Delta)} \in \mathbb{R}^{D\times n}$. It applies the HK MLP output coefficients to the hidden state and uses $\mathbf{P}$ as a ReLU-based logical gate:

$$\mathbf{\Psi}^{(\Delta)} = \sigma_R\Big(\begin{pmatrix} \boldsymbol{W}^{(\Delta)} & 2A\,\mathbf{P} \end{pmatrix} \mathbf{\Psi}^{(\Delta-1)} - A\mathbf{1}_{D\times n}\Big).$$

To see how this exactly zeroes out invalid coordinates, consider column q of $\mathbf{\Psi}^{(\Delta)}$. For row $r = (i-1)n + j$ (where $i \in [d]$ and $j \in [n]$), the pre-activation is:

$$\begin{aligned} \Big(\boldsymbol{W}^{(\Delta)}\hat{\boldsymbol{h}}(N\boldsymbol{X}^{(\text{flt})})\Big)_r + 2A \cdot \mathbf{P}_{r,q} - A &= \widetilde{\Phi}_r(N\boldsymbol{X}^{(\text{flt})}) + 2A \cdot (\boldsymbol{I}_n)_{jq} - A \\ &= \widetilde{\Phi}_r(N\boldsymbol{X}^{(\text{flt})}) + A[2 \cdot \mathbf{1}_{j=q} - 1] \\ &= \begin{cases} \widetilde{\Phi}_r(N\boldsymbol{X}^{(\text{flt})}) + A, & j = q, \\ \widetilde{\Phi}_r(N\boldsymbol{X}^{(\text{flt})}) - A, & j \neq q. \end{cases} \end{aligned}$$

Because $\|\widetilde{\mathbf{\Phi}}(N\boldsymbol{X}^{(\text{flt})})\|_{\ell^\infty} \le L \le A$ for $\boldsymbol{X} \in [0,1]^{d\times n}$, applying the ReLU σ_R exactly preserves the values at the target token positions and outputs 0 otherwise:

$$\mathbf{\Psi}^{(\Delta)}_{r,q} = \begin{cases} \widetilde{\Phi}_r(N\boldsymbol{X}^{(\text{flt})}) + A, & j = q, \\ 0, & j \neq q. \end{cases} \tag{17}$$

Finally, the terminal affine map redistributes these entries into the $d \times n$ output and removes the offset A:

$$\boldsymbol{T}(\boldsymbol{X}) = \mathbf{S}\mathbf{\Psi}^{(\Delta)} - A\mathbf{1}_{d\times n} \in \mathbb{R}^{d\times n}.$$

The block-diagonal structure of $\mathbf{S}$ sums each n-dimensional block along the columns of $\mathbf{\Psi}^{(\Delta)}$. For column q and feature row $p \in [d]$, this sum extracts the single non-zero entry where $j = q$:

$$[\boldsymbol{T}(\boldsymbol{X})]_{p,q} = \sum_{j=1}^{n} \mathbf{\Psi}^{(\Delta)}_{(p-1)n+j,q} - A = \big(\widetilde{\Phi}_{(p-1)n+q}(N\boldsymbol{X}^{(\text{flt})}) + A\big) - A = \widetilde{\Phi}_{(p-1)n+q}(N\boldsymbol{X}^{(\text{flt})}).$$

This verifies (16). The post-attention block uses $1 + (\Delta - 1) + 1 + 1 = \Delta + 2$ affine maps and requires maximum width $\max\{W + n, D + n\}$. □

For the block-index notation of the general Transformer definition, set $\boldsymbol{\mathcal{F}}^{(1)}_{FF} := \boldsymbol{\mathcal{F}}_{\text{post}}$.

C.4 Exact Transformer Realization of the Kuhn-Triangulation MLP

For a vector- or matrix-valued function $\boldsymbol{g}$, $\|\boldsymbol{g}\|_{L^\infty,\ell^\infty(\Omega)} := \sup_{\boldsymbol{X}\in\Omega} \|\boldsymbol{g}(\boldsymbol{X})\|_{\ell^\infty}$ denotes the supremum entrywise maximum norm.

Lemma C.5 (Transformer Realization of a Kuhn-Triangulation MLP with Discrete Hidden MLP Weights)**.** *Let $\widetilde{\boldsymbol{\Phi}} = (\widetilde{\Phi}_r)_{r=1}^{D} : \mathbb{R}^D \to \mathbb{R}^D$ be the global vector-valued integer-grid Kuhn-triangulation MLP specified by the stacking convention in Section B.1 for a target $\boldsymbol{f}$ on $[0,1]^{d\times n}$ ($D = dn$ components), whose restriction to $[0,N]^D$ has the stated Kuhn-triangulation representation with Δ affine maps and width W, and set $\boldsymbol{\Phi} = (\Phi_r)_{r=1}^{D}$ by $\boldsymbol{\Phi}(\boldsymbol{x}) := \widetilde{\boldsymbol{\Phi}}(N\boldsymbol{x})$, with $\|\boldsymbol{\Phi}\|_{L^\infty,\ell^\infty} \le L$. Then there exists a Transformer $\boldsymbol{T} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of length $K = 1$ with*

$$\operatorname{size}(\boldsymbol{T}) = \big((2,\, 2D+n),\, (1,\, D),\, (\Delta+2,\, \max\{W+n,\, D+n\})\big)$$

and

$$\dim(\boldsymbol{T}) = (d,\, d+n,\, 2D+n,\, d)$$

such that for any $\boldsymbol{X} \in [0,1]^{d\times n}$,

$$[\boldsymbol{T}(\boldsymbol{X})]_{p,q} = \widetilde{\Phi}_{(p-1)n+q}\big(N\boldsymbol{X}^{(\mathit{flt})}\big) = \Phi_{(p-1)n+q}(\boldsymbol{X}^{(\mathit{flt})}), \qquad p \in [d],\, q \in [n], \tag{18}$$

computed inside the architecture (no external reshaping). The weight bounds are $B_{FF} = \max\{nN,\, 2A\}$ (with $A = \lceil L \rceil$), $B_{SA} = 1$, and $B_{EB} = 1$. The stored hidden MLP weights retain their $\{0, \pm 1\}/\{0, \pm 1/2\}$ structure. The complete localization of the remaining parameter families is collected later in Proposition D.1.

Proof. Compose Lemmas C.1, C.2, and C.4 with the embedding $\boldsymbol{\mathcal{F}}_{EB}$ of (13). This gives the stated $\operatorname{size}(\boldsymbol{T})$ and $\dim(\boldsymbol{T})$, together with identity (18), whose first equality is (16). The weight bounds follow from the three component lemmas. The hidden MLP weight sets follow from Proposition B.2(iii)–(iv). □

Remark C.6 (Regularity by exact equality)**.** *Because the identity of Lemma C.5 is an exact componentwise equality, each component $(\boldsymbol{T})_{p,q}$ inherits the regularity properties of $\boldsymbol{x} \mapsto \Phi_{(p-1)n+q}(\boldsymbol{x}) = \widetilde{\Phi}_{(p-1)n+q}(N\boldsymbol{x})$ without constant-factor loss. In particular, this exact realization transfers the target Hölder bound $[(\boldsymbol{T})_{pq}]_{\lambda,1} \le L$ and the prescribed common ℓ^1-modulus ω (Definition 3.1).*

D Approximation and Stored-Precision Proofs

D.1 Proof of Theorem 4.1

Proof. **Step 1: Vector-valued Kuhn-triangulation MLP construction and modulus preservation.** Identify $[0,1]^{d\times n}$ with $[0,1]^D$ via the flattening isomorphism $\boldsymbol{X} \mapsto \boldsymbol{X}^{(\mathrm{flt})}$. For $r = (p-1)n + q$, write $f_r := f_{pq}$. Define the vector-valued function $\boldsymbol{f}^{(\mathrm{flt})} : [0,1]^D \to \mathbb{R}^D$ whose r-th component is f_r. Each component belongs to $C^{0,\lambda}([0,1]^D)$ with $\|f_r\|_{C^{0,\lambda}} \le L$, admitting the ℓ^1-Hölder modulus $\omega_H(t) = Lt^\lambda$ and the prescribed common modulus ω (Definition 3.1).

We first fix the common grid parameter using the Hölder modulus $\omega_H(t) = Lt^\lambda$:

$$N := \left\lceil \frac{D}{2}\left(\frac{L}{\varepsilon}\right)^{1/\lambda} \right\rceil = \Theta\big(D(L/\varepsilon)^{1/\lambda}\big),$$

chosen so that $L(D/(2N))^\lambda \le \varepsilon$. On the grid vertices $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{G_N}$ of $\{0, 1/N, \dots, 1\}^D$ ($G_N = (N+1)^D$), the Kuhn interpolant of component f_r is

$$\Phi_{r,N}(\boldsymbol{x}) = \sum_{k=1}^{G_N} f_r(\boldsymbol{v}_k) h_{k,N}(\boldsymbol{x}), \tag{19}$$

where $\{h_{k,N}\}_{k=1}^{G_N}$ are the fixed Kuhn hat functions forming a continuous partition of unity affine on each Kuhn simplex. This interpolant does not depend on the modulus used to analyze it.

By Theorem B.1, using the Hölder modulus $\omega_H(t) = Lt^\lambda$ guarantees the uniform approximation error

$$\|\Phi_{r,N} - f_r\|_{L^\infty([0,1]^D)} \le \omega_H(D/(2N)) = L(D/(2N))^\lambda \le \varepsilon,$$

as well as the ℓ^1-Hölder seminorm bound $[\Phi_{r,N}]_{\lambda,1} \le L$.

Because $\Phi_{r,N}$ in (19) is precisely this explicit continuous interpolant affine on each Kuhn simplex, Hong & Kratsios (2024, Lemma 7.8) applied to this interpolant shows that the same ω is a modulus for $\Phi_{r,N}$:

$$|\Phi_{r,N}(\boldsymbol{x}) - \Phi_{r,N}(\boldsymbol{y})| \le \omega(\|\boldsymbol{x} - \boldsymbol{y}\|_1) \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in [0,1]^D.$$

We then realize this vector-valued interpolant $\boldsymbol{\Phi} = (\Phi_{1,N}, \dots, \Phi_{D,N})^\top$ as a ReLU MLP using Proposition B.2. By part (i) of that proposition, the hidden layers are target-independent and shared across all D outputs. The nodal values are stacked in the final coefficient matrix $\boldsymbol{W}_{r,k}^{(\Delta)} = f_r(\boldsymbol{v}_k)$, which is embedded in the Transformer gating affine layer. The cited Hong–Kratsios theorem and Proposition B.2 supply affine-map depth $\Delta - 1 \le \lceil \log_2 D \rceil + 4$, width $W \le 8D(N+1)^D$, and the stated hidden MLP weight sets $\{0, \pm 1\}$ and $\{0, \pm 1/2\}$.

Step 2: Closed Transformer realization. Let $\widetilde{\boldsymbol{\Phi}} : \mathbb{R}^D \to \mathbb{R}^D$ be the global integer-grid vector-valued HK network of Step 1, whose restriction to $[0, N]^D$ has the Kuhn-triangulation representation, so that $\boldsymbol{\Phi}(\boldsymbol{x}) = \widetilde{\boldsymbol{\Phi}}(N\boldsymbol{x})$. Apply Lemma C.5: this yields the Transformer

$$\boldsymbol{T}_\varepsilon = \boldsymbol{\mathcal{F}}_{\text{post}} \circ \boldsymbol{\mathcal{F}}_{SA}^{(1)} \circ \boldsymbol{\mathcal{F}}_{FF}^{(0)} \circ \boldsymbol{\mathcal{F}}_{EB} : \mathbb{R}^{d \times n} \to \mathbb{R}^{d \times n}$$

of length $K = 1$, with $[\boldsymbol{T}_\varepsilon(\boldsymbol{X})]_{p,q} = \Phi_{(p-1)n+q}(\boldsymbol{X}^{(\text{flt})})$ computed exactly inside the network for all $\boldsymbol{X} \in [0,1]^{d \times n}$. By Remark C.6, this exact entrywise algebraic equality transfers both regularity properties to $\boldsymbol{T}_\varepsilon$: for every $p \in [d], q \in [n]$ and all $\boldsymbol{X}, \boldsymbol{Y} \in [0,1]^{d \times n}$,

$$[(\boldsymbol{T}_\varepsilon)_{pq}]_{\lambda,1} \le L \qquad \text{and} \qquad |(\boldsymbol{T}_\varepsilon)_{pq}(\boldsymbol{X}) - (\boldsymbol{T}_\varepsilon)_{pq}(\boldsymbol{Y})| \le \omega\big(\|\boldsymbol{X}^{(\text{flt})} - \boldsymbol{Y}^{(\text{flt})}\|_1\big).$$

Step 3: Architecture and parameter counts. The augmented-input embedding (13) has stored nonzero count $M_{EB} = d + n = \mathcal{O}(D)$.

The complete dimension flow is:

$$\boldsymbol{X}(\mathbb{R}^{d \times n}) \xrightarrow{\boldsymbol{\mathcal{F}}_{EB}} \widetilde{\boldsymbol{X}}(\mathbb{R}^{(d+n) \times n}) \xrightarrow{\boldsymbol{\mathcal{F}}_{FF}^{(0)}} \mathbb{R}^{(2D+n) \times n} \xrightarrow{\boldsymbol{\mathcal{F}}_{SA}^{(1)}} \mathbb{R}^{(2D+n) \times n} \xrightarrow{\boldsymbol{\mathcal{F}}_{\text{post}}} \mathbb{R}^{d \times n},$$

consistent with $\dim(\boldsymbol{T}_\varepsilon) = (d,\ d+n,\ 2D+n,\ d)$ and the length $K = 1$. By Step 2, $[\boldsymbol{T}_\varepsilon(\boldsymbol{X})]_{p,q} = \Phi_{(p-1)n+q}(\boldsymbol{X}^{(\text{flt})})$ for $p \in [d]$, $q \in [n]$, computed inside the network (Lemma C.5).

Parameter count and size accounting. Lemma C.5 and the preceding steps give the architecture, size, and weight bounds. Proposition B.2(iii)–(iv) gives the hidden MLP weight sets. By Proposition B.2, the shared HK hidden part, including the G_N unit biases in its penultimate hidden layer and terminal-feature duplication, is bounded by $\mathcal{O}\big(D(N+1)^D\big)$ stored nonzeros, while stacking its D MLP output-coefficient rows contributes $D(N+1)^D$. The diagonalization, scaling, target-independent auxiliary position-gating entries, and terminal redistribution affine stages contribute $\mathcal{O}(D)$ nonzeros. In addition, each of the $\Delta - 1$ augmented hidden MLP layers stores a separate positional-code block $\boldsymbol{I}_n$, contributing $M_{\text{sig}} = n(\Delta - 1)$. As a result,

$$M_{FF} = \mathcal{O}\big(D(N+1)^D + n\Delta\big) = \mathcal{O}\big(D(N+1)^D\big),$$

because $n \le D$, $\Delta \le \lceil \log_2 D \rceil + 5$, and $(N+1)^D \ge 2^D$. Proposition D.1 summarizes the full parameter localization. □

Proposition D.1 (Parameter Families and Layerwise Representations). *For the architecture of $\boldsymbol{T}_\varepsilon$ constructed in Theorem 4.1 and its finite-precision variants, the parameter families admit the following descriptions:*

Parameter Family	*Value Set / Bound*	*Stored Nonzero Count*	*Layerwise Representation*
First hidden weights $\boldsymbol{W}^{(1)}$	$\{0, \pm 1\}$	$\leq 4DG_N$	*at most* 2 *code bits*
Internal hidden weights $\boldsymbol{W}^{(l)}$	$\{0, \pm 1/2\}$	$\mathcal{O}(DG_N)$	*at most* 2 *code bits*
Unquantized output coefficients $\boldsymbol{W}^{(\Delta)}$ *	$[-L, L]$	*at most* DG_N *nonzeros*	*no finite-bit guarantee*
Quantized output coefficients Q^{fp} *or* $Q^{(\rho,\mathrm{fp})}$ *	$\mathcal{D}_{s_{\mathrm{fp}}}(L)$	*at most* DG_N *nonzeros*	$1 + s_{\mathrm{fp}} + \lceil \log_2(L+1) \rceil$ *value bits*
First-layer grid biases $\boldsymbol{b}^{(1)}$	$\{0, \pm 1, \ldots, \pm N\}$	$\leq 4DG_N$	$\lceil \log_2(2N+1) \rceil$ *code bits*
HK penultimate unit biases $\boldsymbol{b}^{(\Delta-1)}$	$+1$	G_N	*fixed constant*
Scaling entries c	nN	D	$\lceil \log_2(nN+1) \rceil$ *value bits*
Routing / gating parameters	$\{0, 2A\}$, $\{0, 1\}$, *biases* $\{-A\}$	$3D + d$	$\mathcal{O}(1)$ *code bits;* $\mathcal{O}(\log(1+L))$ *value bits*

* *The unquantized and quantized rows are alternative realizations of the only target-dependent parameter family; all other parameters depend solely on* (D, n, N, L).

The constant routing and diagonalization weights lie in $\{0, 1\}$ *and their diagonalization biases in* $\{0, -1\}$. *Together with the embedding, gathering self-attention, and the identity blocks carrying the positional code, they are determined by* (D, n, N, L) *alone. The HK penultimate unit biases are all fixed at* $+1$*; only their number* G_N *varies with* N. *The first-layer grid biases and scaling entries depend on* ε *through* N, *while the gating family depends only on* L. *The layerwise code-bit counts in the table describe optional finite-codebook encodings and are not the common fixed-point stored-value count* $\mathrm{Bits}_{\mathrm{val}}$. *Under the common format of Section 5.2, every stored scalar uses the same bit-width* w. *The aggregate quantities* M_{EB}, M_{FF}, *and* M_{SA} *count nonzeros blockwise. The statistical class* $\mathcal{T}_{N,s}$ *of Section E.3 permits the active hidden weights and first-layer grid biases to vary over their respective finite codebooks, while input scaling and positional gating remain fixed. The displayed nonzero counts use a conservative expanded-matrix convention. An implementation with value sharing may store* $c = nN$ *and* $A = \lceil L \rceil$ *once and generate the fixed supports of* $c\boldsymbol{I}_D$, $2A\mathbf{P}$, *and the associated* $-A$ *offsets without storing repeated values. This implementation convention does not change* M_{nz} *or* $\mathrm{Bits}_{\mathrm{val}}$ *as defined here.*

D.2 Uniform and Hölder Bounds under Coefficient Perturbations

We use the ℓ^1-Hölder seminorm $[g]_{\lambda,1}$ defined in Section 3, distinct from the Euclidean convention in (1). Lemma 5.1 quantifies the seminorm perturbation under direct fixed-point truncation and the preserved regularity bound under coefficient contraction before quantization.

Proof of Lemma 5.1.

Proof. Step 1: Amplitude bound. Since the Kuhn hat functions $\{h_{k,N}\}_{k=1}^{G_N}$ are nonnegative and form a partition of unity on $[0,1]^D$ ($\sum_{k=1}^{G_N} h_{k,N}(\boldsymbol{x}) = 1$), for every $\boldsymbol{x} \in [0,1]^D$ we have

$$|\mathcal{E}_N(\boldsymbol{x})| = \left| \sum_{k=1}^{G_N} e_k h_{k,N}(\boldsymbol{x}) \right| \leq \sum_{k=1}^{G_N} |e_k| h_{k,N}(\boldsymbol{x}) \leq \delta \sum_{k=1}^{G_N} h_{k,N}(\boldsymbol{x}) = \delta,$$

which proves $\|\mathcal{E}_N\|_\infty \leq \delta$.

Step 2: Gradient bound on each Kuhn simplex. To control the regularity, we first bound the gradient of $\mathcal{E}_N$ on each simplex of the Kuhn triangulation. Consider an elementary grid cube $Q_{\boldsymbol{u}} = \prod_{i=1}^{D} [u_i/N, (u_i+1)/N]$ with multi-index $\boldsymbol{u} = (u_1, \ldots, u_D) \in \{0, \ldots, N-1\}^D$. Following Hong & Kratsios (2024, Lemma 7.5 and Lemma 7.8), to maintain facet-matching across adjacent cubes, the Kuhn triangulation incorporates coordinate reflections according to the parity of u_i. Define the reflection sign along coordinate i by $s_i := (-1)^{u_i} \in \{\pm 1\}$. The starting base corner

$\boldsymbol{w}_0 \in Q_{\boldsymbol{u}}$ of the simplex has coordinates

$$(\boldsymbol{w}_0)_i := \begin{cases} \frac{u_i}{N}, & \text{if } u_i \text{ is even } (s_i = +1), \\ \frac{u_i+1}{N}, & \text{if } u_i \text{ is odd } (s_i = -1). \end{cases}$$

When u_i is even, the simplex steps forward along the positive coordinate direction; when u_i is odd, the cube is reflected along coordinate i, so the simplex starts from the upper facet $(u_i + 1)/N$ and steps backward. For any permutation π of $[D]$, the corresponding Kuhn simplex $S \subset Q_{\boldsymbol{u}}$ is the convex hull of $D + 1$ grid vertices $\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_D$, generated along the coordinate staircase:

$$\boldsymbol{w}_j = \boldsymbol{w}_{j-1} + \frac{s_{\pi(j)}}{N}\mathbf{u}_{\pi(j)}, \qquad j = 1, \ldots, D, \tag{20}$$

where $\mathbf{u}_i \in \mathbb{R}^D$ denotes the i-th standard unit basis vector. Because each hat function $h_{k,N}$ is affine on S, the restriction $\mathcal{E}_N|_S$ is affine with constant gradient $\boldsymbol{g}_S := \nabla(\mathcal{E}_N|_S) \in \mathbb{R}^D$. Since $\mathcal{E}_N$ interpolates the perturbation values at all grid vertices, it satisfies $\mathcal{E}_N(\boldsymbol{v}_k) = e_k$. Letting $e(\boldsymbol{w}_j)$ denote the perturbation value at vertex $\boldsymbol{w}_j$, the difference along the j-th edge obeys

$$e(\boldsymbol{w}_j) - e(\boldsymbol{w}_{j-1}) = \mathcal{E}_N(\boldsymbol{w}_j) - \mathcal{E}_N(\boldsymbol{w}_{j-1}) = \boldsymbol{g}_S^\top(\boldsymbol{w}_j - \boldsymbol{w}_{j-1}) = \frac{s_{\pi(j)}}{N}(\boldsymbol{g}_S)_{\pi(j)}.$$

Since $s_{\pi(j)} \in \{\pm 1\}$, multiplying by $N s_{\pi(j)}$ gives the $\pi(j)$-th partial derivative:

$$(\boldsymbol{g}_S)_{\pi(j)} = N s_{\pi(j)}\big(e(\boldsymbol{w}_j) - e(\boldsymbol{w}_{j-1})\big).$$

Because π is a bijection of $[D]$, as j ranges over $\{1, \ldots, D\}$, $\pi(j)$ visits each coordinate $i \in [D]$ exactly once. Setting $j = \pi^{-1}(i)$ and using $|e(\boldsymbol{w})| \le \delta$, every coordinate satisfies

$$|(\boldsymbol{g}_S)_i| \le N\big(|e(\boldsymbol{w}_j)| + |e(\boldsymbol{w}_{j-1})|\big) \le 2N\delta.$$

Taking the maximum over $i \in [D]$ yields $\|\boldsymbol{g}_S\|_\infty \le 2N\delta$.

Step 3: Global Lipschitz bound and ℓ^1-Hölder seminorm. Fix arbitrary $\boldsymbol{x}, \boldsymbol{y} \in [0, 1]^D$ and consider the line segment $\gamma(t) := (1 - t)\boldsymbol{x} + t\boldsymbol{y}$ for $t \in [0, 1]$. Since $[0, 1]^D$ is partitioned into finitely many closed convex simplices, the pre-images $\gamma^{-1}(S)$ induce a finite partition $0 = t_0 < t_1 < \cdots < t_m = 1$ such that on each subinterval $[t_{r-1}, t_r]$, the segment $\gamma(t)$ lies in a single simplex S_r. On (t_{r-1}, t_r), the composite function $\mathcal{E}_N \circ \gamma$ is differentiable with

$$(\mathcal{E}_N \circ \gamma)'(t) = \boldsymbol{g}_{S_r}^\top(\boldsymbol{y} - \boldsymbol{x}).$$

By Hölder's inequality,

$$|(\mathcal{E}_N \circ \gamma)'(t)| \le \|\boldsymbol{g}_{S_r}\|_\infty \|\boldsymbol{y} - \boldsymbol{x}\|_1 \le 2N\delta\|\boldsymbol{y} - \boldsymbol{x}\|_1.$$

Integrating over $t \in [0, 1]$ yields the global ℓ^1-Lipschitz bound:

$$|\mathcal{E}_N(\boldsymbol{y}) - \mathcal{E}_N(\boldsymbol{x})| \le 2N\delta\|\boldsymbol{y} - \boldsymbol{x}\|_1.$$

Combining this Lipschitz bound with the amplitude bound $|\mathcal{E}_N(\boldsymbol{y}) - \mathcal{E}_N(\boldsymbol{x})| \le 2\|\mathcal{E}_N\|_\infty \le 2\delta$, we obtain

$$|\mathcal{E}_N(\boldsymbol{y}) - \mathcal{E}_N(\boldsymbol{x})| \le 2\delta \min\big\{1,\, N\|\boldsymbol{y} - \boldsymbol{x}\|_1\big\}.$$

Using the elementary inequality $\min\{1, t\} \le t^\lambda$ valid for all $t \ge 0$ and $\lambda \in (0, 1]$, setting $t = N\|\boldsymbol{y} - \boldsymbol{x}\|_1$ gives

$$|\mathcal{E}_N(\boldsymbol{y}) - \mathcal{E}_N(\boldsymbol{x})| \le 2\delta N^\lambda \|\boldsymbol{y} - \boldsymbol{x}\|_1^\lambda.$$

Dividing by $\|\boldsymbol{y} - \boldsymbol{x}\|_1^\lambda$ and taking the supremum over $\boldsymbol{x} \ne \boldsymbol{y}$ proves $[\mathcal{E}_N]_{\lambda,1} \le 2\delta N^\lambda$, establishing (2).

Step 4: Preserved regularity under coefficient contraction. For a target f with $\|f\|_{C^{0,\lambda}} \le L$, we have $\|f\|_\infty \le L$ and $[f]_{\lambda,1} \le L$ (since $\|\boldsymbol{x} - \boldsymbol{y}\|_2 \le \|\boldsymbol{x} - \boldsymbol{y}\|_1$). By Theorem B.1 (Hong & Kratsios, 2024, Theorem 4.1), the unquantized interpolant $\Phi := \sum_{k=1}^{G_N} a_k h_{k,N}$ with $a_k = f(\boldsymbol{v}_k)$ satisfies $\|\Phi\|_\infty \le L$, $[\Phi]_{\lambda,1} \le L$, and $\|\Phi - f\|_\infty \le \eta$. For perturbed coefficients q_k satisfying $|q_k| \le L$ and $|q_k - (1 - \rho)a_k| \le \delta$, define the quantized interpolant

$$\Phi^q := \sum_{k=1}^{G_N} q_k h_{k,N}.$$

Because $|q_k| \le L$ and the Kuhn hat functions form a partition of unity, we immediately have $\|\Phi^q\|_\infty \le \max_k |q_k| \le L$. Decomposing $q_k = (1-\rho)a_k + e_k$ with $|e_k| \le \delta$, we express Φ^q and its target error as

$$\Phi^q = (1-\rho)\Phi + \mathcal{E}_N, \qquad \Phi^q - f = (\Phi - f) - \rho\Phi + \mathcal{E}_N.$$

By the triangle inequality for $\|\cdot\|_\infty$,

$$\|\Phi^q - f\|_\infty \le \|\Phi - f\|_\infty + \rho\|\Phi\|_\infty + \|\mathcal{E}_N\|_\infty \le \underbrace{\eta}_{\text{interpolation error}} + \underbrace{\rho L}_{\text{contraction cost}} + \underbrace{\delta}_{\text{coefficient perturbation}},$$

which proves (3). Similarly, by the triangle inequality for the seminorm $[\cdot]_{\lambda,1}$ and the bound $[\mathcal{E}_N]_{\lambda,1} \le 2\delta N^\lambda$ from Step 3,

$$[\Phi^q]_{\lambda,1} \le (1-\rho)[\Phi]_{\lambda,1} + [\mathcal{E}_N]_{\lambda,1} \le \underbrace{(1-\rho)L}_{\text{contracted seminorm}} + \underbrace{2\delta N^\lambda}_{\text{seminorm perturbation}},$$

which proves (4). In particular, when $2\delta N^\lambda \le \rho L$, the perturbation $2\delta N^\lambda$ is absorbed by the contraction margin ρL, guaranteeing $[\Phi^q]_{\lambda,1} \le L$. □

D.3 Finite-Precision Approximation

Recall the deterministic quantizer $\mathsf{Q}_{s_{\mathrm{fp}}}$ defined in (6), which truncates its input toward zero using s_{fp} fractional bits. The floor identity $0 \le u - \lfloor u \rfloor < 1$ for $u \ge 0$ ensures that for every real scalar z,

$$\mathsf{Q}_{s_{\mathrm{fp}}}(z) \in 2^{-s_{\mathrm{fp}}}\mathbb{Z}, \qquad |\mathsf{Q}_{s_{\mathrm{fp}}}(z) - z| < 2^{-s_{\mathrm{fp}}}, \qquad |\mathsf{Q}_{s_{\mathrm{fp}}}(z)| \le |z|. \tag{21}$$

The common fixed-point format and parameter bit complexity follow Section 5.2.

Proof of Theorem 5.2.

Proof. The result combines the continuous approximation of Theorem 4.1 with the perturbation bound of Lemma 5.1 and the parameter characterization of Proposition D.1.

1. Approximation and regularity. For each component $r = (p-1)n + q$, Theorem 4.1 gives an unquantized interpolant Φ_r with $\|\Phi_r - f_r\|_\infty \le \eta$, $[\Phi_r]_{\lambda,1} \le L$, and $\|\Phi_r\|_\infty \le L$. Quantizing output coefficients gives errors $|Q^{\mathrm{fp}}_{rk} - f_r(\boldsymbol{v}_k)| < 2^{-s_{\mathrm{fp}}}$ with $|Q^{\mathrm{fp}}_{rk}| \le L$. Setting $\rho = 0$ and $\delta = 2^{-s_{\mathrm{fp}}}$ in Lemma 5.1 immediately yields $\|(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon)_{pq} - f_{pq}\|_{L^\infty} \le \eta + 2^{-s_{\mathrm{fp}}} \le \varepsilon$, $[(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon)_{pq}]_{\lambda,1} \le L + 2^{1-s_{\mathrm{fp}}}N^\lambda = L + \tau_\varepsilon$. Because $|Q^{\mathrm{fp}}_{rk}| \le L \le A$, the position gate (17) remains exact, ensuring $\|(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon)_{pq}\|_{L^\infty} \le L$ without an output-clipping head.

2. Common fixed-point representation. Proposition D.1 categorizes every parameter family in $\boldsymbol{T}_\varepsilon$: hidden weights belong to $\{0, \pm 1, \pm 1/2\}$, while grid biases, scaling entries, and routing matrices are integers bounded by B_ε. Because $s_{\mathrm{fp}} \ge 1$, half-integers and integers are all multiples of $2^{-s_{\mathrm{fp}}}$. The quantized output coefficients Q^{fp} also belong to $2^{-s_{\mathrm{fp}}}\mathbb{Z} \cap [-L, L]$ by (21). All stored network parameters therefore belong to $2^{-s_{\mathrm{fp}}}\mathbb{Z} \cap [-B_\varepsilon, B_\varepsilon]$.

3. Complexity orders. Quantization changes only coefficient values and leaves the parameter support unchanged. Therefore, the nonzero-count bound in Theorem 4.1 gives $M_{\mathrm{nz}}(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon) = \mathcal{O}(\varepsilon^{-D/\lambda})$. With $N = \Theta(\varepsilon^{-1/\lambda})$ and $B_\varepsilon = \mathcal{O}(N)$, the fractional-bit count is $s_{\mathrm{fp}} = \log_2(1/\varepsilon) + \mathcal{O}(1)$. The parameter bit-width and stored-value bit complexity from Section 5.2 are:

$$w_\varepsilon = 1 + s_{\mathrm{fp}} + \lceil \log_2(B_\varepsilon + 1) \rceil = \mathcal{O}(\log(1/\varepsilon)),$$
$$\mathrm{Bits}_{\mathrm{val}}(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon) = M_{\mathrm{nz}}(\boldsymbol{T}^{\mathrm{fp}}_\varepsilon) w_\varepsilon = \mathcal{O}(\varepsilon^{-D/\lambda}\log(1/\varepsilon)).$$

□

Proof of Corollary 5.4. The contraction schedule sets $\eta = \min\{\varepsilon, L\}/3$, $\rho = \eta/L \le 1/3$, and chooses s_{fp} such that

$$\delta = 2^{-s_{\mathrm{fp}}} \le \frac{\eta}{2N^\lambda}, \qquad \text{whence} \qquad 2\delta N^\lambda = 2^{1-s_{\mathrm{fp}}}N^\lambda \le \eta = \rho L.$$

Let $Q_{rk}^{(\rho,\mathrm{fp})} = \mathsf{Q}_{s_{\mathrm{fp}}}((1-\rho)f_r(\boldsymbol{v}_k))$. Then $|Q_{rk}^{(\rho,\mathrm{fp})} - (1-\rho)f_r(\boldsymbol{v}_k)| < \delta$ and $|Q_{rk}^{(\rho,\mathrm{fp})}| \le (1-\rho)L$. Applying Lemma 5.1 with contraction parameter ρ and error $\delta = 2^{-s_{\mathrm{fp}}}$ gives:

$$[(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq}]_{\lambda,1} \le (1-\rho)L + 2\delta N^\lambda \le (1-\rho)L + \rho L = L,$$
$$\|(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq} - f_{pq}\|_{L^\infty} \le \eta + \rho L + \delta \le \eta + \eta + \frac{\eta}{2} = \frac{5}{2}\eta \le \frac{5}{6}\varepsilon.$$

Since $|Q_{rk}^{(\rho,\mathrm{fp})}| \le (1-\rho)L$, the position gate (17) remains exact and $\|(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})_{pq}\|_{L^\infty} \le (1-\rho)L$.

For the complexity orders, $2N^\lambda/\eta = \Theta(\varepsilon^{-2})$ gives $s_{\mathrm{fp}} = 2\log_2(1/\varepsilon) + \mathcal{O}(1)$. Because $|Q_{rk}^{(\rho,\mathrm{fp})}| \le L$, the magnitude bound $B_\varepsilon = \mathcal{O}(N)$ and the nonzero count $M_{\mathrm{nz}}(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})})$ obey the same upper bound and asymptotic order $\mathcal{O}(\varepsilon^{-D/\lambda})$ as in Theorem 5.2. The storage orders $w_\varepsilon = \mathcal{O}(\log(1/\varepsilon))$ and $\mathrm{Bits}_{\mathrm{val}}(\boldsymbol{T}_\varepsilon^{(\rho,\mathrm{fp})}) = \mathcal{O}(\varepsilon^{-D/\lambda}\log(1/\varepsilon))$ are preserved. □

E The Statistical Class and Its Approximation Bound

E.1 Setup

We adopt the nonparametric regression model (8) and conditional sub-Gaussian noise assumption (9) from Section 6. For $r = (p-1)n+q$, let $f_{0,r} : [0,1]^D \to \mathbb{R}$ denote the flattened scalar component of $(f_0)_{pq}$. For any measurable matrix-valued function $g : [0,1]^{d\times n} \to \mathbb{R}^{d\times n}$, we denote its empirical norm and empirical inner product by

$$\|g\|_m^2 := \frac{1}{m}\sum_{i=1}^m \|g(\boldsymbol{X}_i)\|_F^2, \qquad \langle g,h\rangle_m := \frac{1}{m}\sum_{i=1}^m \langle g(\boldsymbol{X}_i), h(\boldsymbol{X}_i)\rangle_F.$$

For a sample-dependent predictor such as an empirical risk minimizer $\hat{\boldsymbol{T}}_m$, the population norm $\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\|_{L^2(P_{\boldsymbol{X}})}^2 = \mathbb{E}_{\boldsymbol{X}}[\|\hat{\boldsymbol{T}}_m(\boldsymbol{X}) - \boldsymbol{f}_0(\boldsymbol{X})\|_F^2]$ integrates over a fresh test point $\boldsymbol{X} \sim P_{\boldsymbol{X}}$ independent of the training data, while the outer expectation in risk bounds averages over the training pairs $(\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^m$. As in the main text, asymptotic constants may depend on $(D, L, \lambda, \sigma^2)$ but not on m.

Theorem E.1 (Finite-Class ERM Oracle Inequality)**.** *In the regression setup above, let $\mathcal{F}$ be any data-independent finite class of measurable matrix-valued predictors satisfying $\|\boldsymbol{T}(\boldsymbol{X})\|_F \le \sqrt{D}A$ for every $\boldsymbol{T} \in \mathcal{F}$ and every $\boldsymbol{X}$, with $A = \lceil L \rceil$. If some $\boldsymbol{T}^* \in \mathcal{F}$ satisfies $\|\boldsymbol{T}^* - \boldsymbol{f}_0\|_{L^\infty,\ell^\infty} \le a$, then every measurable exact ERM over $\mathcal{F}$ satisfies*

$$\mathbb{E}\left\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\right\|_{L^2(P_{\boldsymbol{X}})}^2 \le 50Da^2 + (128\sigma^2 + 240DA^2)\frac{\log(3|\mathcal{F}|)+1}{m}.$$

The numerical constants are explicit but not optimized. A fixed ordering of $\mathcal{F}$ provides measurable tie-breaking for an exact ERM. Appendix F proves the bound.

E.2 MLP Output-Coefficient Truncation and Stability

For any $N \in \mathbb{N}_+$, let $G_N = (N+1)^D$. For $s \in \mathbb{N}_+$, let $\boldsymbol{T}_{N,Q}$ denote the constructive Transformer obtained from Lemma C.5 at resolution N, with output-coefficient matrix replaced by $Q \in \mathcal{D}_s(L)^{D\times G_N}$ and all other parameters held fixed.

Lemma E.2 (Stability of the Kuhn-Hat Expansion)**.** *For the final-domain Kuhn hat functions $\{h_{k,N}\}_{k=1}^{G_N}$ (nonnegative, and summing to one pointwise) and any coefficient vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{G_N}$,*

$$\left\|\sum_{k=1}^{G_N}(\alpha_k - \beta_k)h_{k,N}\right\|_{L^\infty} \le \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_{\ell^\infty}, \qquad \left\|\sum_{k=1}^{G_N}\alpha_k h_{k,N}\right\|_{L^\infty} \le \|\boldsymbol{\alpha}\|_{\ell^\infty}.$$

Proof. By the partition of unity ($h_{k,N} \ge 0$, $\sum_k h_{k,N} = 1$), $\left|\sum_k(\alpha_k - \beta_k)h_{k,N}(\boldsymbol{x})\right| \le \sum_k |\alpha_k - \beta_k| h_{k,N}(\boldsymbol{x}) \le \max_k |\alpha_k - \beta_k|$, and likewise $\left|\sum_k \alpha_k h_{k,N}(\boldsymbol{x})\right| \le \max_k |\alpha_k|$. □

Lemma E.3 (Quantization Stability). *Under the Hölder assumptions of Theorem 4.1, fix any $N \in \mathbb{N}_+$ and let $\boldsymbol{T}$ be the constructive Transformer at that resolution. For $s \in \mathbb{N}_+$, let $Q^\star \in \mathcal{D}_s(L)^{D\times G_N}$ have entries $Q^\star_{rk} = \mathsf{Q}_s(W^{(\Delta)}_{rk})$ obtained by truncation toward zero of the MLP output-coefficient matrix $\boldsymbol{W}^{(\Delta)}$, with resulting Transformer $\boldsymbol{T}_{N,Q^\star}$.*

(i) For every $s \in \mathbb{N}_+$, the output truncation error satisfies

$$\|\boldsymbol{T} - \boldsymbol{T}_{N,Q^\star}\|_{L^\infty,\ell^\infty} < 2^{-s}. \tag{22}$$

(ii) If additionally $\varepsilon \in (0,1)$, $N \geq \lceil \frac{D}{2}(L/\varepsilon)^{1/\lambda}\rceil$, and $s \geq \lceil \log_2(1/\varepsilon)\rceil$, then

$$\|\boldsymbol{T} - \boldsymbol{T}_{N,Q^\star}\|_{L^\infty,\ell^\infty} \leq \varepsilon, \qquad \|\boldsymbol{T}_{N,Q^\star} - \boldsymbol{f}\|_{L^\infty,\ell^\infty} \leq 2\varepsilon.$$

Proof. By Proposition D.1, truncation changes only the MLP output-coefficient matrix, whose entries satisfy $|W^{(\Delta)}_{rk}| \leq L$ by Proposition B.2(iv). The quantizer Q_s satisfies $|\mathsf{Q}_s(z) - z| < 2^{-s}$ and $|\mathsf{Q}_s(z)| \leq |z| \leq L$ for all $|z| \leq L$, so $Q^\star \in \mathcal{D}_s(L)^{D\times G_N}$ with $|Q^\star_{rk}| \leq L \leq A$ ($A = \lceil L\rceil$). Applying Lemma E.2 rowwise, the partition of unity gives, for every $Q \in \mathcal{D}_s(L)^{D\times G_N}$ and every $r \in [D]$,

$$\big|(Q\boldsymbol{h}_N(\boldsymbol{x}))_r\big| \leq \sum_{k=1}^{G_N} |Q_{rk}| h_{k,N}(\boldsymbol{x}) \leq L \leq A.$$

Hence the position-gating identity (17) remains exact: the terminal redistribution affine layer returns the corresponding component of $Q\boldsymbol{h}_N(\boldsymbol{x})$. Applying Lemma E.2 to each row of $Q^\star - \boldsymbol{W}^{(\Delta)}$ gives

$$\|\boldsymbol{T} - \boldsymbol{T}_{N,Q^\star}\|_{L^\infty,\ell^\infty} = \|(\boldsymbol{W}^{(\Delta)} - Q^\star)\boldsymbol{h}_N\|_{L^\infty,\ell^\infty} < 2^{-s},$$

which proves (i) for every $N \in \mathbb{N}_+$. For (ii), $s \geq \lceil \log_2(1/\varepsilon)\rceil$ implies $2^{-s} \leq \varepsilon$, ensuring $\|\boldsymbol{T} - \boldsymbol{T}_{N,Q^\star}\|_{L^\infty,\ell^\infty} \leq \varepsilon$. By Theorem B.1, the grid condition gives $\|\boldsymbol{T} - \boldsymbol{f}\|_{L^\infty,\ell^\infty} \leq L(D/(2N))^\lambda \leq \varepsilon$. The triangle inequality yields $\|\boldsymbol{T}_{N,Q^\star} - \boldsymbol{f}\|_{L^\infty,\ell^\infty} \leq 2\varepsilon$. □

E.3 A Finite Structured Transformer Class for Nonparametric Regression

At resolution $N \in \mathbb{N}_+$ and fractional-bit count $s \in \mathbb{N}_+$, the architecture is that of Theorem 4.1 equipped with a fixed componentwise output-clipping head. To establish the regression rate, we define a structured finite-precision Transformer class whose internal parameters vary over discrete grids.

Layerwise parameter classes. We allow three parameter families to enter the data-independent discrete class: (i) *Active hidden weights*: on the active index set $\mathcal{M}_N$ of non-zero positions in the hidden matrices $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(\Delta-1)}$ ($|\mathcal{M}_N| = \mathcal{O}(DG_N)$, Proposition B.2), first-layer weights take values in $\{0, \pm 1\}$ and internal-layer weights take values in $\{0, \pm\frac{1}{2}\}$ by Proposition B.2(iii)–(iv); all entries outside $\mathcal{M}_N$ remain zero. (ii) *First-layer grid biases*: from the layer-by-layer architecture in Appendix B.1 (Table 5), the first hidden layer width is at most $2DG_N$ when $D \geq 2$ and equals $4G_N = 4DG_N$ when $D = 1$. Thus the fixed set of first-layer bias slots satisfies $M_{\mathrm{bias},N} \leq 4DG_N = \mathcal{O}(DG_N)$, with each slot taking values in $\{-N, -N+1, \ldots, N\}$. (iii) *Output coefficients*: the embedded coefficient matrix Q varies in $\mathcal{D}_s(L)^{D\times G_N}$, with $\mathcal{D}_s(L) = 2^{-s}\mathbb{Z} \cap [-L, L]$. The input-scaling scalar nN, positional-gating constants $2A$ and $-A$, and all other structural components, including attention matrices $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{0}$, $\boldsymbol{W}_V$ and $\boldsymbol{W}_O$, positional-code-carry blocks, routing and redistribution entries, penultimate unit biases ($+1$), and inactive coordinates, remain fixed to their constructive values. We collect the allowable internal configurations (families (i)–(ii)) into a finite set $\mathcal{P}_{N,s}$. The constructive internal configuration $\vartheta^\star_N$ from Proposition D.1 belongs to $\mathcal{P}_{N,s}$ for every $s \geq 1$.

Final clipping and class definition. Let $\boldsymbol{T}_{\vartheta,Q}$ denote the Transformer realized by internal configuration $\vartheta \in \mathcal{P}_{N,s}$ and output matrix $Q \in \mathcal{D}_s(L)^{D\times G_N}$. For a scalar z, define the componentwise clipping map $\Pi_A(z) := \min\{A, \max\{-A, z\}\}$ with $A = \lceil L\rceil$. For every coordinate $y = (f_0)_{pq}(\boldsymbol{X}) \in [-A, A]$, metric projection onto $[-A, A]$ gives $|\Pi_A(z) - y| \leq |z - y|$. Applied componentwise, this also yields $\|\Pi_A(\boldsymbol{T}(\boldsymbol{X}))\|_F \leq \sqrt{D}A$ for every network output. The clipping map admits the two-affine-stage ReLU realization

$$\Pi_A(z) = -A + \sigma_R(z + A) - \sigma_R(z - A),$$

which has hidden width $2D$ and at most $7D$ fixed entries when flattened (or width $2d$ and at most $7d$ entries columnwise). Merged with the terminal redistribution layer, it adds a single affine stage. Because its parameters depend only on L through A, clipping adds no class multiplicity. The class is

$$\mathcal{T}_{N,s} := \big\{\Pi_A \circ \boldsymbol{T}_{\vartheta,Q} : \vartheta \in \mathcal{P}_{N,s},\ Q \in \mathcal{D}_s(L)^{D\times G_N}\big\}.$$

Proof of Theorem 6.1.

Proof. First, the class $\mathcal{T}_{N,s}$ is finite and data-independent. Its cardinality is bounded by the product of the layerwise codebook sizes:

$$|\mathcal{T}_{N,s}| \le 3^{|\mathcal{M}_N|}(2N+1)^{M_{\mathrm{bias},N}}|\mathcal{D}_s(L)|^{DG_N}.$$

Using $|\mathcal{D}_s(R)| = 2\lfloor 2^s R\rfloor + 1 \le 2^{s+1}(R+1)$, the log-cardinality satisfies

$$\log|\mathcal{T}_{N,s}| \le |\mathcal{M}_N|\log 3 + M_{\mathrm{bias},N}\log(2N+1) + DG_N\big[(s+1)\log 2 + \log(L+1)\big].$$

Since $|\mathcal{M}_N| = \mathcal{O}(DG_N)$ and $M_{\mathrm{bias},N} = \mathcal{O}(DG_N)$, we have

$$\log|\mathcal{T}_{N,s}| = \mathcal{O}\big(DG_N(1+\log N + s + \log(1+L))\big).$$

At $N = N_m$ and $s = s_m$, because $\log N_m = \mathcal{O}(\log m)$ and $s_m = \mathcal{O}(\log m)$, this yields

$$\log|\mathcal{T}_{N_m,s_m}| = \mathcal{O}(DG_{N_m}\log m). \tag{23}$$

Define $\boldsymbol{T}^\star := \Pi_A \circ \boldsymbol{T}_{\vartheta^\star_{N_m},Q^\star_m} \in \mathcal{T}_{N_m,s_m}$ by selecting the exact constructive internal configuration $\vartheta^\star_{N_m} \in \mathcal{P}_{N_m,s_m}$ from Proposition D.1 and truncating only the nodal values via $Q^\star_{m,rk} = \mathsf{Q}_{s_m}(f_{0,r}(\boldsymbol{v}_k)) \in \mathcal{D}_{s_m}(L)$. By Theorem B.1, applied componentwise through the exact identity of Lemma C.5, the unquantized constructive Transformer $\boldsymbol{T}$ at resolution N_m satisfies $\|\boldsymbol{T} - \boldsymbol{f}_0\|_{L^\infty,\ell^\infty} \le L(D/(2N_m))^\lambda$. By Lemma E.3(i), replacing the unquantized output coefficients $\boldsymbol{W}^{(\Delta)}$ with $Q^\star_m = \mathsf{Q}_{s_m}(\boldsymbol{W}^{(\Delta)})$ incurs uniform error strictly bounded by 2^{-s_m}. Since $\boldsymbol{f}_0(\boldsymbol{X}) \in [-A,A]^{d\times n}$, componentwise projection onto this box cannot increase the Frobenius or componentwise ℓ^∞ distance, so

$$\|\Pi_A(\boldsymbol{T}_{\vartheta^\star_{N_m},Q^\star_m}) - \boldsymbol{f}_0\|_{L^\infty,\ell^\infty} \le \|\boldsymbol{T}_{\vartheta^\star_{N_m},Q^\star_m} - \boldsymbol{f}_0\|_{L^\infty,\ell^\infty} \le a_m,$$

where $a_m := L(D/(2N_m))^\lambda + 2^{-s_m}$. Clipping ensures $\|\Pi_A(\boldsymbol{T}_{\vartheta,Q})(\boldsymbol{X})\|_F \le \sqrt{D}A$ for every $\vartheta, Q, \boldsymbol{X}$. Applying Theorem E.1 to $\mathcal{F} = \mathcal{T}_{N_m,s_m}$ with $\boldsymbol{T}^\star$ as the comparison element yields

$$\mathbb{E}\left\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\right\|^2_{L^2(P_{\boldsymbol{X}})} \le 50Da_m^2 + (128\sigma^2 + 240DA^2)\frac{\log(3|\mathcal{T}_{N_m,s_m}|)+1}{m}.$$

We now evaluate the bias and variance terms. With $N_m = \lceil \frac{D}{2}(m/\log(em))^{\frac{1}{2\lambda+D}}\rceil$ and $s_m = \lceil \frac{\lambda}{2\lambda+D}\log_2 m\rceil$, the squared approximation bias satisfies

$$Da_m^2 = \mathcal{O}\big(N_m^{-2\lambda} + 2^{-2s_m}\big) = \mathcal{O}\left(\left(\frac{\log(em)}{m}\right)^{\frac{2\lambda}{2\lambda+D}}\right).$$

For the variance terms, since $G_{N_m} = (N_m+1)^D = \mathcal{O}((m/\log(em))^{D/(2\lambda+D)})$, inserting (23) gives

$$\frac{\log|\mathcal{T}_{N_m,s_m}|}{m} = \mathcal{O}\left(\frac{DG_{N_m}\log m}{m}\right) = \mathcal{O}\left(\left(\frac{\log(em)}{m}\right)^{\frac{2\lambda}{2\lambda+D}}\right).$$

Because $2\lambda/(2\lambda+D) < 1$ and $m \ge 2$, the remainder m^{-1} is of strictly smaller order $\mathcal{O}(m^{-1}) = o((\log(em)/m)^{\frac{2\lambda}{2\lambda+D}})$. Combining these bounds yields (11). □

F Proof of the Finite-Class ERM Oracle Inequality

Lemma F.1 (Bernstein's Inequality and Inversion)**.** *Let $\{Z_i\}_{i=1}^m$ be independent zero-mean real random variables satisfying $|Z_i| \le U$ almost surely and $\frac{1}{m}\sum_{i=1}^m \mathrm{Var}(Z_i) \le V$.*

(i) For any $t > 0$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m Z_i > t\right) \le \exp\left(-\frac{mt^2}{2V + \frac{2}{3}Ut}\right). \tag{24}$$

(ii) For any $x > 0$, with probability at least $1 - e^{-x}$,

$$\frac{1}{m}\sum_{i=1}^m Z_i \le \sqrt{\frac{2Vx}{m}} + \frac{2Ux}{3m}. \tag{25}$$

Proof. Part (i) is the standard Bernstein tail bound. For (ii), setting the right-hand side of (24) to e^{-x} yields the quadratic equation $mt^2 - \frac{2}{3}Uxt - 2Vx = 0$. Its unique positive root $t_0 = \frac{Ux}{3m} + \sqrt{\frac{2Vx}{m} + \frac{U^2x^2}{9m^2}}$ satisfies $t_0 \le \sqrt{\frac{2Vx}{m}} + \frac{2Ux}{3m}$ via $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Since the quadratic polynomial is monotonically increasing on $[t_0, \infty)$, setting $t = \sqrt{\frac{2Vx}{m}} + \frac{2Ux}{3m} \ge t_0$ satisfies the tail probability bound $\le e^{-x}$. □

Proof of Theorem E.1. Stage 1: ERM basic inequality. Write $\mathcal{F} = \{\boldsymbol{T}_1, \dots, \boldsymbol{T}_{N_\mathcal{F}}\}$, where $N_\mathcal{F} := |\mathcal{F}|$, and choose $\boldsymbol{T}^* \in \mathcal{F}$ from the theorem's approximation assumption. Then

$$r_* := \|\boldsymbol{T}^* - \boldsymbol{f}_0\|^2_{L^2(P_\boldsymbol{X})} \le Da^2.$$

By the theorem's uniform predictor bound and the componentwise bound $\|\boldsymbol{f}_0(\boldsymbol{X})\|_F \le \sqrt{D}L \le \sqrt{D}A$, for every $\boldsymbol{T}, \boldsymbol{T}' \in \mathcal{F}$,

$$\|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{T}'(\boldsymbol{X})\|^2_F \le M, \qquad \|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{f}_0(\boldsymbol{X})\|^2_F \le M, \qquad M := 4DA^2.$$

Set $\Gamma := \hat{\boldsymbol{T}}_m - \boldsymbol{T}^*$ and $e^* := \boldsymbol{T}^* - \boldsymbol{f}_0$. The ERM property and the squared-loss expansion give

$$\begin{aligned} 0 &\ge \frac{1}{m}\sum_{i=1}^m \left[\|\hat{\boldsymbol{T}}_m(\boldsymbol{X}_i) - \boldsymbol{Y}_i\|^2_F - \|\boldsymbol{T}^*(\boldsymbol{X}_i) - \boldsymbol{Y}_i\|^2_F\right] \\ &= \|\Gamma\|^2_m + 2\langle e^*, \Gamma\rangle_m - \frac{2}{m}\sum_{i=1}^m \langle \boldsymbol{\xi}_i, \Gamma(\boldsymbol{X}_i)\rangle_F. \end{aligned}$$

Since $-2\langle e^*, \Gamma\rangle_m \le \frac{1}{2}\|\Gamma\|^2_m + 2\|e^*\|^2_m$, this yields

$$\|\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*\|^2_m \le 4\|\boldsymbol{T}^* - \boldsymbol{f}_0\|^2_m + \frac{4}{m}\sum_{i=1}^m \langle \boldsymbol{\xi}_i, (\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*)(\boldsymbol{X}_i)\rangle_F. \tag{26}$$

Stage 2: empirical-to-population control. We first control the population distance from each class member to the oracle. For $j \in [N_\mathcal{F}]$, set

$$Z_i^{(j)} := \|\boldsymbol{T}_j(\boldsymbol{X}_i) - \boldsymbol{T}^*(\boldsymbol{X}_i)\|^2_F, \qquad \mu_j := \mathbb{E}Z_i^{(j)}, \qquad \widehat{\mu}_j := \frac{1}{m}\sum_{i=1}^m Z_i^{(j)}.$$

Then

$$0 \le Z_i^{(j)} \le M, \qquad |\mu_j - Z_i^{(j)}| \le M, \qquad \mathrm{Var}(Z_i^{(j)}) \le \mathbb{E}[(Z_i^{(j)})^2] \le M\mu_j.$$

Fix $0 < \delta < 1$ and let $x := \log(3N_\mathcal{F}/\delta)$. Applying the inverted Bernstein bound (25) of Lemma F.1 to the centered variables $\mu_j - Z_i^{(j)}$ gives, for each j with probability at least $1 - e^{-x}$,

$$\mu_j - \widehat{\mu}_j \le \sqrt{\frac{2M\mu_j x}{m}} + \frac{2Mx}{3m}.$$

Young's inequality bounds the square-root term $\sqrt{\frac{2M\mu_j x}{m}}$ by $\mu_j/2 + Mx/m$. Thus a union bound gives, with probability at least $1-\delta/3$ simultaneously for all j,

$$\|\boldsymbol{T}_j - \boldsymbol{T}^*\|_{L^2(P_{\boldsymbol{X}})}^2 \le 2\|\boldsymbol{T}_j - \boldsymbol{T}^*\|_m^2 + \frac{10M}{3m}\log\frac{3N_{\mathcal{F}}}{\delta}, \qquad j \in [N_{\mathcal{F}}]. \tag{27}$$

Stage 3: oracle empirical error. We next control the empirical error of the oracle itself. Set

$$Z_i^* := \|\boldsymbol{T}^*(\boldsymbol{X}_i) - \boldsymbol{f}_0(\boldsymbol{X}_i)\|_F^2, \qquad \mathbb{E}Z_i^* = r_*.$$

Since $0 \le Z_i^* \le M$ and $\mathrm{Var}(Z_i^*) \le Mr_*$, applying (25) with $x = \log(3/\delta)$ gives, with probability at least $1-\delta/3$,

$$\|\boldsymbol{T}^* - \boldsymbol{f}_0\|_m^2 \le r_* + \sqrt{\frac{2Mr_*}{m}\log\frac{3}{\delta}} + \frac{2M}{3m}\log\frac{3}{\delta} \le \frac{3}{2}r_* + \frac{5M}{3m}\log\frac{3}{\delta}. \tag{28}$$

Stage 4: noise multiplier. Finally, we control the noise multiplier uniformly over the finite class. For $\Gamma_j := \boldsymbol{T}_j - \boldsymbol{T}^*$, conditional independence and (9) give, for every $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\frac{t}{m}\sum_{i=1}^m \langle \boldsymbol{\xi}_i, \Gamma_j(\boldsymbol{X}_i)\rangle_F\right) \;\middle|\; \boldsymbol{X}_1, \dots, \boldsymbol{X}_m\right] \le \exp\left(\frac{\sigma^2 t^2}{2m}\|\Gamma_j\|_m^2\right).$$

Chernoff's bound therefore gives

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{\xi}_i, \Gamma_j(\boldsymbol{X}_i)\rangle_F > \sigma\sqrt{\frac{2x}{m}}\,\|\Gamma_j\|_m \;\middle|\; \boldsymbol{X}_1, \dots, \boldsymbol{X}_m\right) \le e^{-x}.$$

A union bound over j gives, with conditional probability at least $1-\delta/3$,

$$\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{\xi}_i, (\boldsymbol{T}_j - \boldsymbol{T}^*)(\boldsymbol{X}_i)\rangle_F \le \sigma\sqrt{\frac{2}{m}\log\frac{3N_{\mathcal{F}}}{\delta}}\,\|\boldsymbol{T}_j - \boldsymbol{T}^*\|_m, \qquad j \in [N_{\mathcal{F}}]. \tag{29}$$

This event holds simultaneously for every $\boldsymbol{T}_j \in \mathcal{F}$, so it may be evaluated at the data-dependent $\hat{\boldsymbol{T}}_m$. Its conditional probability bound is independent of the covariates, and the tower property gives the same unconditional bound.

Stage 5: combination and integration. On the intersection of these three events, insert (29) with $\boldsymbol{T}_j = \hat{\boldsymbol{T}}_m$ into (26) and use $2uv \le u^2/2 + 2v^2$. With $x = \log(3N_{\mathcal{F}}/\delta)$, this yields

$$\|\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*\|_m^2 \le 8\|\boldsymbol{T}^* - \boldsymbol{f}_0\|_m^2 + \frac{32\sigma^2 x}{m}. \tag{30}$$

Applying (27) to $\hat{\boldsymbol{T}}_m$ and combining (30) with (28) gives

$$\begin{aligned}
\|\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*\|_{L^2(P_{\boldsymbol{X}})}^2 &\le 2\|\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*\|_m^2 + \frac{10Mx}{3m} \\
&\le 16\|\boldsymbol{T}^* - \boldsymbol{f}_0\|_m^2 + \frac{64\sigma^2 x}{m} + \frac{10Mx}{3m} \\
&\le 24r_* + \frac{(64\sigma^2 + 30M)x}{m},
\end{aligned}$$

where the last line uses $\log(3/\delta) \le x$. As a result,

$$\begin{aligned}
\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\|_{L^2(P_{\boldsymbol{X}})}^2 &\le 2\|\hat{\boldsymbol{T}}_m - \boldsymbol{T}^*\|_{L^2(P_{\boldsymbol{X}})}^2 + 2r_* \\
&\le 50r_* + \frac{(128\sigma^2 + 60M)x}{m} \\
&\le 50Da^2 + (128\sigma^2 + 240DA^2)\frac{\log(3N_{\mathcal{F}}/\delta)}{m},
\end{aligned}$$

where $r_* \le Da^2$ and $M = 4DA^2$. Thus

$$\|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\|_{L^2(P_{\boldsymbol{X}})}^2 \le 50Da^2 + (128\sigma^2 + 240DA^2)\frac{\log(3N_{\mathcal{F}}/\delta)}{m} \tag{31}$$

with probability at least $1-\delta$. Let

$$\mathcal{R} := \|\hat{\boldsymbol{T}}_m - \boldsymbol{f}_0\|^2_{L^2(P_{\boldsymbol{X}})},$$
$$U_0 := 50Da^2 + (128\sigma^2 + 240DA^2)\frac{\log(3N_{\mathcal{F}})}{m},$$
$$V_0 := \frac{128\sigma^2 + 240DA^2}{m}.$$

Taking $\delta = e^{-t}$ in (31) shows that $\mathbb{P}(\mathcal{R} > U_0 + V_0 t) \leq e^{-t}$ for $t > 0$.

The tail bound can now be integrated to obtain the expectation bound:

$$\mathbb{E}\mathcal{R} \leq U_0 + \int_0^\infty \mathbb{P}(\mathcal{R} > U_0 + s)\, ds \leq U_0 + V_0$$

which is the stated expectation bound because $N_{\mathcal{F}} = |\mathcal{F}|$. □