Function graph transformers universally approximate operators between function spaces

Takashi Furuya
Doshisha University, RIKEN AIP

David Mis
Rice University

Ivan Dokmanić
University of Basel

Maarten V. de Hoop
Rice University

Matti Lassas
University of Helsinki

Abstract

We study the approximation of nonlinear operators between function spaces by transformers. Our approach is to lift functions to measures supported on their graphs and leverage a recently introduced measure-theoretic view of transformers. A function h is represented by its graph measure γ_h, while finite tokens $\{(x_j, h(x_j))\}_{j=1}^N$ provide empirical approximations of this measure. We show that this framework naturally describes discretization refinement through convergence of measures and provides a setting for operator learning. Within this framework, we introduce *function-graph preserving maps*, a structured class of measure maps that send graph measures to graph measures, ensuring that outputs remain single-valued functions. Crucially, this additional structure does not reduce generality: we prove that the resulting function-graph preserving maps can be approximated by finite compositions of standard softmax self-attention layers and pointwise MLPs, yielding universal approximation results for broad classes of nonlinear operators. Unlike existing theoretical approaches to operator learning with transformers, the measure-theoretic framework also accommodates bounded rough inputs in negative-order Sobolev topologies through regularization, for which discretization invariance is particularly challenging, as well as query points on different output domains. Overall, function-graph preserving maps provide a continuum viewpoint and a mathematical toolkit for transformer-based operator learning, clarifying the roles of positional encodings, graph structure, and regularization in ensuring consistency across discretizations.

1 Introduction

In this paper, we study operator learning with transformers from a foundational point of view and revisit universal approximation. Learning operators between function spaces has emerged as a central problem at the interface of analysis, scientific computing, and machine learning. A defining requirement is discretization invariance: the learned model should act consistently across different samplings of the underlying continuum. Transformer architectures, originally introduced for sequence modeling [Vaswani et al., 2017], provide a natural and increasingly powerful framework for operator learning. By representing a function h of x through a collection of tokens $\{(x_j, h(x_j))\}_{j=1}^N$, transformers define mappings capable of modeling long-range interactions. The self-attention mechanism computes pairwise interactions via learned kernels, which can be interpreted as nonlinear integral operators in the continuum limit. This observation places transformers squarely within the landscape of operator approximation, with attention playing the role of a data-adaptive kernel.

When $h : \overline{\Omega} \to [-L, L]^n$ is a continuous function, we choose a finite set of points $x_j \in \overline{\Omega} \subset \mathbb{R}^d$ and approximate the graph $\mathbf{G}(h) = \{(x, h(x)) : \ x \in \overline{\Omega}\}$ by a finite sequence of tokens $\xi_j =$

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

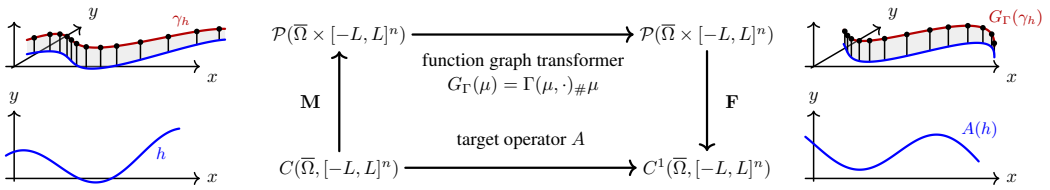


Figure 1: Operators between function spaces are universally approximated by function-graph preserving maps. Functions (blue) are identified with their graph measures (red), which are approximated by sums of Dirac measures (vertical black lines). We construct tokens from these empirical approximations, and a transformer outputs tokens corresponding to an approximation of the target graph measure. This diagram illustrates the simplest case of a target operator A between continuous functions; with appropriate mollification and regularization, universal approximation can be extended to operators acting on (generalized) functions in negative-order Sobolev spaces.

$(x_j, h(x_j)) \in \mathbb{R}^{d+n}$, $j = 1, \dots, N$. A mathematically fruitful perspective is to represent functions by empirical (graph) measures,

$$\gamma_h^{(N)} = \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_j} = \frac{1}{N} \sum_{j=1}^{N} \delta_{(x_j, h(x_j))}, \tag{1}$$

that approximate a continuous measure $\gamma_h \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ supported on $\mathbf{G}(h)$, see Appendix B. For a continuous, possibly nonlinear operator A, we show that the operator between graph measures, $\gamma_h \to \gamma_{A(h)}$ can be approximated by composition of MLPs and traditional softmax-based transformers, which in the single-head case are given by

$$\xi_i' = \xi_i + W \sum_{j=1}^{N} \frac{\exp\left(\frac{1}{\sqrt{k}} \langle Q\xi_i, K\xi_j \rangle\right)}{\sum_{\ell=1}^{N} \exp\left(\frac{1}{\sqrt{k}} \langle Q\xi_i, K\xi_\ell \rangle\right)} V\xi_j. \tag{2}$$

where Q, K, V, W are the query, key, value and head matrices, and k is the dimension of the key/query vectors $Q\xi_i$ and $K\xi_j$. By denoting the empirical graph measure by $\mu = \gamma_h^{(N)}$, we can write (2) using a map that takes in a measure μ and the token ξ_i,

$$\xi_i' = \Gamma_\theta(\mu, \xi_i) \tag{3}$$

with parameters $\theta = (Q, K, V, W)$, where

$$\Gamma_\theta(\mu, \xi) = \xi + W \int_{\mathbb{R}^{d+n}} \frac{\exp\left(\frac{1}{\sqrt{k}} \langle Q\xi, K\xi' \rangle\right)}{\int_{\mathbb{R}^{d+n}} \exp\left(\frac{1}{\sqrt{k}} \langle Q\xi, K\zeta \rangle\right) d\mu(\zeta)} V\xi' d\mu(\xi'). \tag{4}$$

Using this notation, a transformer layer defines a map from discrete measures to discrete measures,

$$G_\Gamma : \mu_N = \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_j} \mapsto \mu_N' = \frac{1}{N} \sum_{j=1}^{N} \delta_{(x_j', y_j')}, \quad (x_j', y_j') = \Gamma_\theta(\mu, (x_j, y_j)), \tag{5}$$

Using the notation of the push forward $F_{\#}$ of a map F used in measure theory, we can write (5) as

$$G_\Gamma(\mu) = \Gamma_\theta(\mu, \cdot)_{\#}\mu, \tag{6}$$

that is $G_\Gamma(\mu)$ is obtained by pushing forward the measure μ using a context function $\Gamma_\theta(\mu, \cdot)$ that depends on μ, see (37) in Appendix A. The formula (6) defines a map on probability measures, $G_\Gamma : \mathcal{P}(\mathbb{R}^D) \to \mathcal{P}(\mathbb{R}^D)$ that treats measures μ and their empirical approximations μ_N in a uniform way. This unified treatment makes it possible to use measure theory to study limits of long token sequences by taking $N \to \infty$. In this paper we will consider measure-theoretic transformers G_Γ with a context function $\Gamma = (\Gamma^{(x)}, \Gamma^{(y)})$ whose x-coordinate function $\Gamma^{(x)}$ depends only on the variable x and defines a one-to-one map $x \mapsto \Gamma^{(x)}(x)$. This property will ensure that graph measures are mapped to graph measures. We also generalize these considerations to the case when $A : C(\overline{\Omega}; [-L, L]^n) \to C(\overline{D}; [-L, L]^m)$ is a map between different domains $\Omega \subset \mathbb{R}^d$ and $D \subset \mathbb{R}^{d'}$. Summarizing, we view transformer layers as maps acting on measures and operators as maps from graph measures to graph measures. Under this formulation, self-attention induces a nonlinear pushforward of measures through kernels that depend on both spatial location and function values.

Prior work. Early work on operator learning with transformers showed that attention can facilitate discretization-robust PDE surrogate modeling. OFormer [Li et al., 2023] combines self-attention, cross-attention, and pointwise networks to learn mappings from sampled input functions to queried outputs, while Cao [2021] interprets attention in relation to operator approximation and Galerkin-type projections. GNOT [Hao et al., 2023] extends this line to irregular meshes, heterogeneous input fields, and multiscale structure through normalized attention and geometric gating. PiT [Chen and Wu, 2024] builds attention weights from spatial relations among sample locations, giving a position-induced kernel with a direct discretization interpretation. Other work demonstrated that the choice of attention kernel and positional encoding affects generalization across sequence lengths and sampling regimes [Choromanski et al., 2021, Katharopoulos et al., 2020, Li et al., 2024].

A second line of work has moved from task-specific neural operators toward foundation-model architectures for scientific machine learning. Vision transformers and patch-based representations [Dosovitskiy et al., 2021] motivate models such as CViT [Wang et al., 2025], which combines transformer encoders with query-wise decoding to evaluate outputs at arbitrary coordinates. Universal Physics Transformers [Alkin et al., 2024] use a latent transformer framework for Eulerian grids, meshes, and Lagrangian particles, while Poseidon [Herde et al., 2024] and Walrus [McCabe et al., 2025] are pretrained transformer-based models for PDEs and continuum dynamics, emphasizing transfer across equations, domains, resolutions, and physical regimes.

The mathematical theory of transformers has also developed rapidly. Attention has been studied through smoothness, continuum, and measure-theoretic viewpoints [Shih et al., 2025, Castin et al., 2024, Geshkovski et al., 2025]. Universal in-context approximation by transformers was established by Furuya et al. [2025b], while Furuya et al. [2025a] studied the closure of transformers as support-preserving maps between measures. Related perspectives include measure-to-measure interpolation and mean-field approximation results [Geshkovski et al., 2024, Biswal et al., 2024, Mei et al., 2019, Yang, 2020]. The closest prior work is Calvello et al. [2025], who formulate continuum attention directly on function spaces and prove universality for operators on C^s and nonnegative-order Sobolev spaces. Our approach is complementary: functions are represented by graph measures and transformers by graph-preserving measure maps, yielding universality results that extend to negative-order Sobolev spaces. A key difficulty for negative-order Sobolev inputs is that pointwise evaluation, and hence direct tokenization, is not generally well-defined. We address this by tokenizing a mollified input and proving convergence to the target operator as the mollification and regularization are removed. The measure formulation provides a common space for finite token sets and continuum graph measures, allowing us to analyze discretization refinement within the same framework. See Appendix J for further discussion.

Despite these advances, the theory of transformers for operator learning remains incomplete. New perspectives are required to illuminate this problem given the success of transformers on hard operator learning tasks. Fundamental open questions concern approximation under low regularity, stability under input perturbations, and the role of depth in capturing compositional structure.

Our contributions

We develop a mathematical framework for transformers as operator learners. We obtain a universal approximation result through the introduction of function-graph preserving maps imposing a natural structure on the context mappings.

1. Graph measure formulation of operator learning. We represent functions by graph measures and identify finite token sets with empirical approximations of these measures. We show that continuous nonlinear operators, $A : C(\overline{\Omega}; [-L, L]^n) \to C^1(\overline{D}; [-L, L]^m)$, can be represented by maps from graph measures to graph measures.

2. Function-graph preserving maps. We define function-graph preserving maps through structured context maps that send graph measures to graph measures. On empirical graph measures, these maps act on finite sequences of tokens while preserving their spatial coordinates. We prove continuity under discretization refinement: as the empirical graph measures converge in W_1, their images converge to the corresponding continuum output measure.

3. Universal approximation. We show that function-graph preserving maps can be approximated by compositions of self-attention layers and pointwise MLPs.

4. Generalized functions and regularization. We generalize the universal approximation to operators between generalized function spaces through a regularization procedure that can be realized via a composition with a linear neural operator.

2 Graph measures and maps between functions and measures

Let $\Omega \subset \mathbb{R}^d$ be an open bounded set with C^∞ boundary, or a cube $[0,1]^d$, $\mathcal{B}$ the family of Borel sets in Ω and $\lambda_\Omega : \mathcal{B} \to [0,\infty)$ be the Borel measure on Ω that coincides with the Lebesgue measure for all open subsets of Ω. (For basic notions related to this paper, see Appendix A.) Measures, particularly finite measures supported on finite sets, are closely related to sequences of tokens via sums of Dirac measures, see Appendix B. We argue that non-discrete measures can be considered as limits of sequences of N tokens as $N \to \infty$. We use point-sample tokens here for simplicity; Appendix E describes a patch-based tokenization compatible with the same graph-measure viewpoint.

Let $L^\infty_{\mathcal{B}}(\Omega; [-L,L]^n)$ denote the essentially bounded Borel measurable functions. For $h : \overline{\Omega} \to [-L,L]^n$, we define the "graph measure",

$$\gamma_h = (\mathrm{id}, h)_\# \overline{\lambda}_\Omega, \quad \text{i.e.,} \quad \gamma_h(A) = \overline{\lambda}_\Omega(\{x \in \Omega : \ (x, h(x)) \in A\}), \tag{7}$$

where $\overline{\lambda}_\Omega = \frac{1}{\mathrm{vol}(\Omega)} \lambda_\Omega$ and $A \subset \Omega \times [-L,L]^n$ is a Borel set. Thus γ_h is a measure supported on the graph of h that satisfies

$$\int_{\Omega \times [-L,L]^n} \phi(x,y)\, \mathrm{d}\gamma_h(x,y) = \frac{1}{\mathrm{vol}(\Omega)} \int_\Omega \phi(x, h(x))\, \mathrm{d}\lambda_\Omega(x) \quad \text{for all } \phi \in C(\overline{\Omega} \times [-L,L]^n). \tag{8}$$

Definition 1. *We define* $\mathbf{M} : L^\infty_{\mathcal{B}}(\Omega; [-L,L]^n) \to \mathcal{P}(\Omega \times [-L,L]^n)$ *to be the map from a function to its graph measure. That is,* $\mathbf{M}(h) := \gamma_h = (\mathrm{id}, h)_\# \overline{\lambda}_\Omega$.

The left inverse for the map $\mathbf{M}$ is given by

$$\mathbf{F} : \ \gamma_h \mapsto \mathbf{F}(\gamma_h)(x) := \lim_{r \to 0+} \frac{\mathrm{vol}(\Omega)}{\omega_d r^d} \int_{(B(x,r) \cap \Omega) \times [-L,L]^n} y' \, d\gamma_h(x', y'), \quad x \in \Omega, \tag{9}$$

where $B(x,r)$ is the ball centered at x with radius r, and ω_d is the volume of the unit ball in $\mathbb{R}^d$. For any $h \in L^\infty_{\mathcal{B}}(\Omega; [-L,L]^n)$ formula (9) is valid in a Borel set whose complement has measure zero [Bogachev, 2007]. We note that operators $\mathbf{M}$ and $\mathbf{F}$ are defined for graph measures while we also pay attention to sums of delta measures. Because of this, we define the regularized versions of $\mathbf{M}$ and $\mathbf{F}$.

Definition 2. *For* $\varepsilon > 0$*, we define the map from general measures to continuous functions,*

$$(\mathbf{F}_\varepsilon \gamma)(x) = \frac{\mathrm{vol}(\Omega)}{\varepsilon^d} \int_{\overline{\Omega} \times [-L,L]^n} \rho\left(\frac{x' - x}{\varepsilon}\right) y' \, d\gamma(x', y'). \tag{10}$$

We use a standard mollifier $\rho \in C^\infty_c(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \rho(x')dx' = 1$, $\mathrm{supp}\,(\rho) \subset B(0,1)$, $\rho(-x) = \rho(x)$ and set $\rho_\tau(x) = \tau^{-d}\rho(x/\tau)$ and $\rho_\tau * h(x) = \int_{\mathbb{R}^d} \rho_\tau(x - x')h(x')\, dx'$.

For functions $h \in L^\infty_{\mathcal{B}}(\Omega; [-L,L]^n)$ and their graph measures γ_h we have $(\mathbf{F}_\varepsilon \gamma_h)(x) = \rho_\varepsilon * \tilde{h}(x)$, where $\tilde{h}$ is the zero-extension of h to $\mathbb{R}^d$. For $h \in L^p(\Omega)$, for all $1 \le p < \infty$, it holds that

$$h = \lim_{\varepsilon \to 0+} (\mathbf{F}_\varepsilon \gamma_h) \quad \text{in } L^p(\Omega). \tag{11}$$

For $h \in H^{-s}(\Omega)$, $s \in \mathbb{R}$ we introduce a regularized version of h,

$$R_\tau h = \rho_\tau * (h \cdot (\rho_\tau * \mathbf{1}_{\Omega_\tau})), \quad \tau > 0, \tag{12}$$

with $\Omega_\tau = \{x \in \Omega \ : \ \mathrm{dist}(x, \partial\Omega) > 4\tau\}$. We note that, here, $\rho_\tau * \mathbf{1}_{\Omega_\tau}$ vanishes near the boundary and convolution with ρ_τ makes a function continuous (in fact, C^∞). Below, we will consider functions $h \in H^{-s}(\Omega)$ with $s > 0$, $s - \frac{1}{2} \notin \mathbb{Z}$. Such functions can be discontinuous, finite measures, or even Dirac delta distributions when $-s < -\frac{d}{2}$. We show in Appendix C that for Sobolev functions $h \in H^{-s}(\Omega)$ it holds that $R_\tau h \to h$ in $H^{-s}(\Omega)$ as $\tau \to 0$.

3 Function-graph preserving maps

Below, let $\pi_1 : \overline{\Omega} \times \mathbb{R}^n \to \overline{\Omega}$ be the projection $\pi_1(x, y) = x$ and $\pi_2 : \overline{\Omega} \times \mathbb{R}^n \to \mathbb{R}^n$ be the projection $\pi_2(x, y) = y$. We consider the set of probability measures $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ as a metric space with 1-Wasserstein distance W_1.

Below, we will consider measure-theoretic transformers $G_\Gamma : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$. These are continuous maps in the 1-Wasserstein topology, see Furuya et al. [2025a],

$$G_\Gamma(\mu) = \Gamma(\mu, \cdot)_{\#}\mu$$

where $_\#$ denotes push forward, see (37) in Appendix A, and

$$\Gamma : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n \tag{13}$$

is a continuous map called the *context map*. We will study G_Γ with context maps Γ having the property that for all graph measures γ_h of continuous functions $h : \overline{\Omega} \to [-L, L]^n$ the image $G_\Gamma(\gamma_h)$ is a graph measure γ_g of some continuous function $g = J_\Gamma(h) : \overline{\Omega} \to [-L, L]^n$. Next, we show that such maps can be represented using a structured form discussed in the following lemma that is proven in Appendix D.

Lemma 1. *Assume that the context function $\Gamma : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$ is continuous and has the property that it preserves graph measures:*

(P) For all continuous $h : \overline{\Omega} \to [-L, L]^n$ there is a continuous function $g = J_\Gamma(h) : \overline{\Omega} \to [-L, L]^n$ such that $\Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_g$,

and moreover, that the map $\gamma_h \mapsto J_\Gamma(h)$ is Lipschitz from

$$\mathcal{G} := \{\gamma_h : h \in C(\overline{\Omega}; [-L, L]^n)\},$$

equipped with the 1-Wasserstein distance, to $C(\overline{\Omega}; [-L, L]^n)$, equipped with the uniform norm, with $\mathrm{Lip}(J_\Gamma(h))$ uniformly bounded for $h \in C(\overline{\Omega}; [-L, L]^n)$.

Then, there exists a context map $\widehat{\Gamma} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$ that is equivalent to Γ in the sense that it satisfies

$$\Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h \tag{14}$$

for all continuous functions $h : \overline{\Omega} \to [-L, L]^n$; the context map $\widehat{\Gamma}$ has the coordinate functions

$$\widehat{\Gamma}(\mu, (x, y)) = (\widehat{\Gamma}^{(x)}(\mu, (x, y)), \widehat{\Gamma}^{(y)}(\mu, (x, y))) \in \overline{\Omega} \times [-L, L]^n, \text{ where} \tag{15}$$

$$\widehat{\Gamma}^{(x)}(\mu, (x, y)) = x \quad \text{and} \quad \widehat{\Gamma}^{(y)}(\mu, (x, y)) = \widehat{p}(\mu, x), \tag{16}$$

and $\widehat{p} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times \overline{\Omega} \to [-L, L]^n$; i.e., the x-coordinate $\widehat{\Gamma}^{(x)}$ is the identity map in the variable x and the y-coordinate $\widehat{\Gamma}^{(y)}(\mu, (x, y))$ is independent of the variable y.

Motivated by the above result, Lemma 12 in Appendix D and Corollary 1 concerning the continuity of maps, we define the following class of function-graph preserving maps.

Definition 3. *A map $G_{\widehat{\Gamma}} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ is a* function-graph preserving map *if*

$$G_{\widehat{\Gamma}}(\mu) = \widehat{\Gamma}(\mu, \cdot)_{\#}\mu,$$

where $\widehat{\Gamma}(\mu, (x, y))$ is defined by formula (15) with

$$\widehat{\Gamma}^{(x)}(\mu, (x, y)) = \Phi(x) \quad \text{and} \quad \widehat{\Gamma}^{(y)}(\mu, (x, y)) = \widehat{p}(\mu, x). \tag{17}$$

Here, $\widehat{p} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times \overline{\Omega} \to [-L, L]^n$ is uniformly continuous in μ and uniformly Lipschitz bounded in (x, y), and the map $\Phi : \overline{\Omega} \to \overline{\Omega}$ is a volume preserving C^1-diffeomorphism, i.e., the Jacobian determinant satisfies $|det\,(D\Phi)| = 1$. In the case when $\Phi(x) = x$, $\widehat{\Gamma}$ coincides with (16).

The map Φ in (17) preserves all information of the coordinates, x, throughout the network, and can be interpreted as a coordinate feature embedding in a similar spirit to the feature mappings used in [Cao, 2021, Sec. 4.1.1] and [Li et al., 2023, Fig. 1], as well as the absolute positional encoding in [Wang et al., 2025, Eq. (1)].

Below, for the sake of notational simplicity, we assume that $\Phi(x) = x$. However, results can be easily adapted for a non-identity map. In essence, a function-graph preserving map never generates implicitly multi-valued functions and, hence, preserves the function space aspect of operator learning. Function-graph preserving maps define maps between functions as follows:

$$\mathbf{F} \circ G_{\widehat{\Gamma}} \circ \mathbf{M} : \{h \in C(\overline{\Omega}) : \ \|h\|_\infty \leq L\} \to C^1(\overline{\Omega}).$$

Approximation of graph measures and tokenization. Next, our aim is to consider approximation of graph measures by finite sums of Dirac measures yielding a discretization or tokenization; see Appendix B for a more detailed discussion.

Definition 4. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. The points $\bar{x}_j \in \Omega \subset \mathbb{R}^d$, $j = 1, 2, \ldots$ are* almost surely *regularly distributed if for all continuous functions $h \in C(\overline{\Omega})$ it holds that* $\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} h(\bar{x}_j) = \frac{1}{\mathrm{vol}(\Omega)} \int_\Omega h(x)\, dx$.

Varadarajan's theorem (Varadarajan [1958] and Theorem 11.4.1 in Dudley [1989]), implies that if $X_1, X_2, \ldots$ are independent, uniformly distributed points in Ω, then they are almost surely regularly distributed.

We denote for $h \in C(\overline{\Omega}; \mathbb{R}^n)$ and $g \in H^{-s}(\Omega; \mathbb{R}^n)$

$$\mathbf{M}_N(h) = \sum_{j=1}^{N} \frac{1}{N} \delta_{(\bar{x}_j, h(\bar{x}_j))} = (\mathrm{id}, h)_\# \left(\sum_{j=1}^{N} \frac{1}{N} \delta_{\bar{x}_j} \right), \quad \mathbf{M}_{N,\tau}(g) = \mathbf{M}_N(R_\tau g). \tag{18}$$

Later, in Section 5, we study maps where $\mathbf{F}$ and $\mathbf{M}$ are replaced by the regularized decoder $\mathbf{F}_\varepsilon$ and the empirical encoder $\mathbf{M}_N$, respectively. For rough inputs, we use the regularized empirical encoder $\mathbf{M}_{N,\tau}$. In this setting, the transformer maps discrete measures to discrete measures.

Then, by the assumption that $\bar{x}_j \in \Omega \subset \mathbb{R}^d$, $j = 1, 2, \ldots$ are regularly distributed points, we have for $h \in C(\overline{\Omega})$ (see Appendix F.1),

$$\lim_{N\to\infty} \mathbf{M}_N(h) = \gamma_h \text{ in the } W_1\text{-metric}. \tag{19}$$

Moreover, for $h \in H^{-s}(\Omega)$ and fixed $\tau > 0$ and $h_\tau = R_\tau h$ we have (see Appendix F.1),

$$\lim_{N\to\infty} \mathbf{M}_{N,\tau}(h) = \gamma_{h_\tau} \text{ in the } W_1\text{-metric}. \tag{20}$$

In the next section, we design function-graph preserving maps to act on graph measures and their empirical approximations in a discretization invariant manner. We illustrate these results in Appendix K: a function-graph transformer trained at a fixed input resolution is evaluated on increasingly fine input point clouds, and the error decreases as N grows, consistent with convergence of empirical graph measures toward the continuum graph measure; see Table 3.

4 Universal approximation of function-graph preserving maps

In this section, we discuss approximating function-graph preserving maps by traditional transformers. General results on approximation of function-graph preserving maps are considered in Appendix F. We employ a measure-theoretic framework [Furuya et al., 2025b, Castin et al., 2024] and the following composition. Below, we denote coordinates in the space $\mathbb{R}^{d+d'}$ by $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. Let $\pi_d : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}^d$ be the projection $\pi_d(x, y) := x$ onto the first d coordinates. A context map as it appears in a single-layer "measure-theoretic" transformer, based on multi-head self-attention,[1] is of

[1] Attention can also be written as an expectation, as was done in Calvello et al. [2025, Definition 5 and Eq (14)]:

$$\mathrm{Att}(z) = \mathbb{E}_{z' \sim \rho_{\theta,\mu,z}} [Vz'], \quad d\rho_{\theta,\mu,z}(z') = \frac{\exp\left(\frac{1}{\sqrt{k}} \langle Qz,\, Kz' \rangle\right)}{\int_{\mathbb{R}^{d+d'}} \exp\left(\frac{1}{\sqrt{k}} \langle Qz,\, Kz'' \rangle\right) d\mu(z'')} d\mu(z').$$

the form $\Gamma_\theta : \mathcal{P}(\mathbb{R}^{d+d'}) \times \mathbb{R}^{d+d'} \to \mathbb{R}^{d+d''}$,

$$\Gamma_\theta(\mu, z) := \left(\pi_d(z)\ ,\ W^L z + \sum_{h=1}^{H} W^h \mathrm{Att}^h(z)\right), \text{ where} \tag{21}$$

$$\mathrm{Att}^h(z) := \int_{\mathbb{R}^{d+d'}} \frac{\exp\left(\frac{1}{\sqrt{k}}\langle Q^h z,\ K^h z'\rangle\right)}{\int_{\mathbb{R}^{d+d'}} \exp\left(\frac{1}{\sqrt{k}}\langle Q^h z,\ K^h z''\rangle\right) d\mu(z'')} V^h z'\, d\mu(z'). \tag{22}$$

Here, $W^L : \mathbb{R}^{d+d'} \to \mathbb{R}^{d''}$ is a weight matrix corresponding to the linear part of the context map (introduced to make skip connections possible), K^h and Q^h are the multi-head key and query matrices in $\mathbb{R}^{k\times(d+d')}$, V^h are the multi-head value matrices in $\mathbb{R}^{d_{\text{head}}\times(d+d')}$, and W^h are the multi-head weight matrices in $\mathbb{R}^{d''\times d_{\text{head}}}$, respectively. We have written $\theta = \{Q^h, K^h, V^h, W^L, W^h\}_{h=1}^H$. By abuse of notation, $z \to \Gamma_\theta(\mu)(z) = \Gamma_\theta(\mu, z)$ defines a map $\mathbb{R}^{d+d'} \to \mathbb{R}^{d+d''}$. For two context maps, $\Gamma_j : \mathcal{P}(\mathbb{R}^{d_j}) \times \mathbb{R}^{d_j} \to \mathbb{R}^{d_{j+1}}$, $j = 1, 2$, the composition $\Gamma_2 \diamond \Gamma_1 : \mathcal{P}(\mathbb{R}^{d_1}) \times \mathbb{R}^{d_1} \to \mathbb{R}^{d_3}$ is defined as

$$(\mu, z) \mapsto (\Gamma_2 \diamond \Gamma_1)(\mu, z) := \Gamma_2(\nu, \Gamma_1(\mu, z)), \quad \nu := \Gamma_1(\mu)_\sharp \mu. \tag{23}$$

Moreover, $F_\xi : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}^d \times \mathbb{R}^{d''}$ has the form

$$F_\xi(z) = (\pi_d(z), H_\xi(z)), \quad z \in \mathbb{R}^{d+d'}, \tag{24}$$

where $H_\xi : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}^{d''}$ is an MLP. We view F_ξ also as maps $(\mu, z) \to F_\xi(\mu, z) = F_\xi(z)$ that are, however, independent of μ. This makes $F_\xi \diamond \Gamma_\theta$ well defined. Finally, let $y_0 \in [-L, L]^n$ and $w_{y_0} : \Omega \times [-L, L]^n \to \Omega \times [-L, L]^n$ be the map $w_{y_0}(x, y) = (x, y_0)$. Moreover, let $w_{y_0}^* : F \to F \circ w_{y_0}$ be the pull-back, or in terminology of ML, a warping operator with w_{y_0}, see Muser et al. [2026]. In Appendix F.2 we prove the following approximation theorem.

Theorem 1. *Let $\widehat{\Gamma} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$ be a continuous map of the form (15) and (16) and let $G_{\widehat{\Gamma}}(\mu) = \widehat{\Gamma}(\mu, \cdot)_\# \mu$ be a function-graph preserving map. Then, for any $\varepsilon \in (0, 1)$, there exists a sufficiently deep measure-theoretic transformer, $G_{\Gamma^{\text{tran}}} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ of the form*

$$G_{\Gamma^{\text{tran}}}(\mu) := \Gamma^{\text{tran}}(\mu, \cdot)_\# \mu, \quad \Gamma^{\text{tran}}(\mu, \cdot) := w_{y_0}^*(F_{\xi_{D_{\text{tr}}}} \diamond \Gamma_{\theta_{D_{\text{tr}}}} \diamond \ldots \diamond F_{\xi_1} \diamond \Gamma_{\theta_1}(\mu, \cdot)), \tag{25}$$

where $y_0 \in [-L, L]^n$ and $\Gamma_{\theta_\ell} : \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d_\ell}) \times \mathbb{R}^d \times \mathbb{R}^{d_\ell} \to \mathbb{R}^d \times \mathbb{R}^{d'_\ell}$ and $F_{\xi_\ell} : \mathbb{R}^d \times \mathbb{R}^{d'_\ell} \to \mathbb{R}^d \times \mathbb{R}^{d_{\ell+1}}$ are defined by (21) and (24) with $d_1 = n$ and $d_\ell = 4d + 4n$ for $\ell = 2, \ldots, D_{\text{tr}}$, and $d_{D_{\text{tr}}} = n$ and $d'_\ell = 4d + 4n$ for $\ell = 1, 2, \ldots, D_{\text{tr}}$, that satisfies

$$\sup_{\mu \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)} W_1(G_{\Gamma^{\text{tran}}}(\mu), G_{\widehat{\Gamma}}(\mu)) \le \varepsilon. \tag{26}$$

In particular, there exists $G_{\Gamma_m^{\text{tran}}}$ that converges uniformly in W_1 to $G_{\widehat{\Gamma}}$ as $m \to \infty$. Moreover,

$$\Gamma^{\text{tran}}(\mu, (x, y)) = (x, \Gamma_{\text{tran}}^{(y)}(\mu, x)), \quad \text{where } \Gamma_{\text{tran}}^{(y)}(\mu, x) \text{ is independent of } y. \tag{27}$$

Remark 1. *The pullback $w_{y_0}^*$ is used to make the context map independent of y. However, this is not necessary to preserve graph measures. Indeed, if*

$$\Gamma(\mu, (x, y)) = (x, T(\mu, x, y)),$$

then, for any graph measure γ_h, we have

$$\Gamma(\gamma_h, \cdot)_\# \gamma_h = \gamma_g, \qquad g(x) = T(\gamma_h, x, h(x)).$$

Moreover, g is continuous if T and h are continuous. Thus, the graph-preserving property and the approximation result remain valid even without the pullback.

5 Expressibility results for nonlinear operators

5.1 Regularization of measures originating from functions in $H^{-s}(\Omega)$, $s \geq 0$

Here, we construct function-graph preserving maps that represent nonlinear operators $A : H^{-s}(\Omega;\mathbb{R}^n) \to H^{-s'}(\Omega;\mathbb{R}^{n'})$, where H^s denotes a Sobolev space, $s \geq 0$, and $s' \in \mathbb{R}$. This extends the framework beyond pointwise-defined continuous inputs to low-regularity objects such as distributions and finite measures. Such spaces are also natural in microlocal analysis, where one studies how operators act on singularities; our result suggests that graph-transformer representations can be formulated in a setting compatible with those distributional viewpoints.

In Definition 2, we mapped measures to continuous functions. Below, we map measures to rough Sobolev functions in a way that makes the errors caused by regularization effects weaker. To this end, we introduce a *regularized map from (non-graph) measures to (generalized) functions*.

We introduce two parameters, $\eta > 0$ and $K \in \mathbb{N}$. Let $k \in \mathbb{Z}_+$ and $r > \frac{d}{2} + k$ be an even integer.

Let $\psi_1, \psi_2, \dots$ be a complete orthonormal basis in $L^2(\Omega;\mathbb{R}^n)$ of eigenfunctions satisfying $(I - \Delta)\psi_j = \lambda_j \psi_j$ and $\psi_j|_{\partial\Omega} = 0$. In Appendix G.3 we define a regularization operator $\gamma \to S_{\eta,K}(\gamma)$ that maps a measure γ to a function that has r (weak) derivatives. When $\gamma = \frac{1}{N}\sum_{j=1}^N \delta_{(x_j,y_j)}$ is a discrete measure, $S_{\eta,K}(\gamma)$ is the solution of the minimization problem,

$$S_{\eta,K}(\gamma) = \underset{\tilde{h}\in H^r_D(\Omega)}{\arg\min}\ \eta\|(I-\Delta)^{r/2}\tilde{h}\|^2_{L^2(\Omega)} + \sum_{k=1}^K \left| \sum_{j=1}^N \frac{\psi_k(x_j)}{N}\cdot y_j - \int_{\overline{\Omega}} \psi_k(x)\,\tilde{h}(x)dx \right|^2, \quad (28)$$

where $H^r_D(\Omega)$ is the Sobolev space of functions with the Navier boundary conditions (that is, iterated Dirichlet conditions). Formula (28) is generalized to arbitrary probability measures γ on $\overline{\Omega}\times[-L,L]^n$ in Appendix G.3. When h is a continuous function and γ_h is its graph measure, the operator $h \mapsto S_{\eta,K}(\gamma_h)$ can be viewed as a spectral low-pass filter with multiplier $(1+\eta\lambda_j^r)^{-1}$ for $j \leq K$ and zero otherwise. Here, K determines a sharp spectral cutoff at $\lambda_K \sim CK^{2/d}$, while r and η control the decay of the retained spectral modes. The explicit spectral formula in Appendix G.3 shows that $S_{\eta,K}(\gamma)$ depends linearly on the measure moments. In particular, $h \mapsto S_{\eta,K}(\gamma_h)$ is a linear operator and can be viewed as a linear neural operator. We choose a suitable $K = K(\eta)$ and denote $S_\eta := S_{\eta,K(\eta)}$. For each fixed $\tau > 0$, $S_\eta(\gamma_{R_\tau h}) \to R_\tau h$ as $\eta \to 0$. Consequently,

$$\lim_{\tau\to 0}\lim_{\eta\to 0} S_\eta(\gamma_{R_\tau h}) = h$$

in $H^{-s}(\Omega)$. See Lemma 17 in Appendix G.3.

5.2 Representation of a nonlinear operator using a function-graph preserving map

Let $L_1 > L$, $k \geq 0$, $0 < s \leq r - \frac{d}{2}$, $s - \frac{1}{2} \notin \mathbb{Z}$, $s' \geq 0$. Assume that A is a possibly nonlinear operator with the properties that

$$A : \mathrm{cl}_{H^{-s}(\Omega;\mathbb{R}^n)}(C(\overline{\Omega};[-L_1,L_1]^n)) \to \mathrm{cl}_{H^{-s'}(\Omega;\mathbb{R}^n)}(C(\overline{\Omega};[-L,L]^n)), \quad (29)$$

$$A : C^k(\overline{\Omega};[-L_1,L_1]^n) \to C^1(\overline{\Omega};[-L,L]^n) \quad (30)$$

are continuous. The domain in (29) contains bounded discontinuous functions, but is restricted to the negative-order Sobolev closure of a class with a fixed amplitude bound, rather than the whole space $H^{-s}(\Omega;\mathbb{R}^n)$.

The two continuity assumptions serve different purposes. Condition (29) allows us to recover $A(h)$ from $A(R_\tau h)$ as $\tau \to 0$. Condition (30) ensures that smooth regularized inputs produce C^1 outputs, which are Lipschitz in the spatial variable. Together with the continuity of the spectral regularizer, this yields a continuous context map and its associated push forward map. These assumptions are sufficient for our construction; we do not claim that they are minimal.

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ be C^∞, that is the identity function in $[-L,L]^n$ and whose image is contained in $[-L_1,L_1]^n$. The function Ψ is defined precisely in Appendix A. We use Ψ to define a Nemytskii operator [Isaia, 2022] that clips the values of $S_\eta(\gamma)$ and define $\tilde{S}_\eta(\gamma) := \Psi \circ S_\eta(\gamma)$. The following lemmas are proven in Appendices G.1 and G.2.

Lemma 2. *The context map* $\Gamma_\eta : \mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n) \to \overline{\Omega}\times[-L,L]^n$ *and associated* $G_{\Gamma_\eta} : \mathcal{P}(\overline{\Omega}\times[-L,L]^n) \to \mathcal{P}(\overline{\Omega}\times[-L,L]^n)$*, define continuous maps*

$$\Gamma_\eta(\mu,(x,y)) = (x, A(\tilde{S}_\eta(\mu))(x)) \in \overline{\Omega}\times[-L,L]^n, \quad G_{\Gamma_\eta}(\mu) = \Gamma_\eta(\mu,\cdot)_{\#}\mu. \tag{31}$$

Lemma 3. *For* $h \in C(\overline{\Omega})$ *we have* $G_{\Gamma_\eta}(\gamma_h) = \gamma_g$*, where* $g(x) = A(\tilde{S}_\eta(\gamma_h))(x)$.

The discrete measure $\mathbf{M}_N(h)$ in (18) can be considered as a tokenized version of the function h. When N is large, the continuity of G_{Γ_η} implies that the measure $\nu = \frac{1}{N}\sum_{j=1}^N \delta_{(\bar{x}_j,g_j)} := G_{\Gamma_\eta}(\mathbf{M}_N(h))$ approximates γ_g, where g is given in Lemma 3. The measure ν can be considered as a tokenized version of g. Then, formula

$$\mathbf{F}_\varepsilon\nu = \sum_{j=1}^N g_j \frac{\mathrm{vol}(\Omega)}{N}\frac{1}{\varepsilon^d}\rho\left(\frac{x-\bar{x}_j}{\varepsilon}\right) \tag{32}$$

produces a function approximating g from the tokens $\xi_j = (\bar{x}_j, g_j)$, $j = 1,\dots,N$, see App. H.

Combining Theorem 1 and Lemmas 2 and 3, we can prove the following universal approximation theorem; see Appendix G.4 for the proof.

Theorem 2. *For* $h \in cl_{H^{-s}(\Omega;\mathbb{R}^n)}(C(\Omega;[-L,L]^n))$ *it holds that*

$$A(h) = \lim_{\tau\to 0}\lim_{\eta\to 0}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\lim_{m\to\infty} \mathbf{F}_\varepsilon \circ G_{\Gamma^{\mathrm{tran}}_{\eta,m}} \circ \mathbf{M}_{N,\tau}(h), \tag{33}$$

where the limit is valid in $H^{-s'}(\Omega)$ *and* $G_{\Gamma^{\mathrm{tran}}_{\eta,m}}$ *are softmax-based traditional transformers that approximate* G_{Γ_η}*, given by Theorem 1. Moreover, for* $h \in C^r_c(\Omega;[-L,L]^n)$

$$A(h) = \lim_{\eta\to 0}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\lim_{m\to\infty} \mathbf{F}_\varepsilon \circ G_{\Gamma^{\mathrm{tran}}_{\eta,m}} \circ \mathbf{M}_N(h), \tag{34}$$

where the limit is valid in $H^{-s'}(\Omega)$.

Theorem 2 approximates a fixed operator A on rough inputs through regularized tokenizations. For fixed $\tau > 0$, the tokens represent the smooth function $R_\tau h$. After passing to the continuum limit and removing the spectral regularization, continuity of A in the negative-order Sobolev topology allows us to recover $A(h)$ as $\tau \to 0$.

The approximation is pointwise with respect to the input h, with convergence in $H^{-s'}(\Omega)$. Uniform approximation over an input class is not asserted here.

Remark 2 (Order of limits)**.** *The order of limits in Theorem 2 is important. For example, if we first take* $\tau \to 0$*, then we cannot in general define* $\mathbf{M}_N(h)$*, since point values of negative-order Sobolev inputs are not well-defined. We therefore keep* $\tau > 0$ *fixed when tokenizing* $R_\tau h$ *and passing to the limit* $N \to \infty$*. The theorem establishes convergence in the stated order; exchanging the limits requires additional uniform estimates.*

Remark 3 (Quantitative estimates)**.** *Theorem 2 is qualitative. Quantitative estimates for individual stages of the construction are given in Appendix F.1 for tokenization and in Appendix G.3.1 for the spectral cutoff* $K(\eta)$*. Under additional Hölder regularity and polynomial growth assumptions on* A*, Theorem 7 in Appendix L provides an end-to-end error bound separating the contributions from input mollification, spectral regularization, tokenization, transformer approximation, and decoding.*

5.3 Representation of a nonlinear operator between different function spaces

Let $\Omega \subset \mathbb{R}^d$ and $D \subset \mathbb{R}^{d'}$ be open bounded sets with C^∞-smooth boundary, or cubes. Below, we use in $\overline{\Omega}\times[-L,L]^n$ the coordinates (x,y) and in $\overline{D}\times[-L,L]^n$ the coordinates (x',y'). When μ is a measure on $\overline{\Omega}\times[-L,L]^n$ and ν is a measure on $\overline{D}\times[-L,L]^n$, their tensor product, $\mu\otimes\nu$ is just the same as the product of their volume elements, that is, $d\mu\otimes\nu(x,y,x',y') = d\mu(x,y)d\nu(x',y')$. On $X = (\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n)$ we use the maps $\rho_1((x,y),(x',y')) = (x,y)$ and $\rho_2((x,y),(x',y')) = (x',y')$, and denote $\pi_2(x',y') = y'$ and $\tilde{\pi}_2 = \pi_2\circ\rho_2 : ((x,y),(x',y')) \to y'$

Proposition 1. *Let* $A : L^2(\Omega;[-L_1,L_1]^n) \to C^1(\overline{D};[-L,L]^n)$ *be a continuous, possibly nonlinear operator. Then there are function-graph preserving maps* $G_{\tilde{\Gamma}_\eta}(\tilde{\mu}) = \tilde{\Gamma}_\eta(\tilde{\mu},\cdot)_{\#}\tilde{\mu}$ *with context*

functions $\tilde{\Gamma}_\eta : \mathcal{P}(X) \times X \to X$, $\eta > 0$, *such that for all* $h \in C(\overline{\Omega}; [-L, L]^n)$ *and* $x' \in D$, *the function* $A(h)$ *at* x' *can be represented as*

$$A(h)(x') = \lim_{\eta \to 0} \langle (\rho_2)_\# \tilde{G}_{\tilde{\Gamma}_\eta}(\gamma_h \otimes \delta_{(x',0)}), \pi_2 \rangle = \lim_{\eta \to 0} \langle \tilde{G}_{\tilde{\Gamma}_\eta}(\gamma_h \otimes \delta_{(x',0)}), \tilde{\pi}_2 \rangle. \tag{35}$$

The proof is given in Appendix G.5. Thus an operator mapping functions on Ω to functions on D can be represented by a function-graph preserving map on $X = (\overline{\Omega} \times [-L, L]^n) \times (\overline{D} \times [-L, L]^n)$, with inputs sampled at $x_j \in \Omega$ and outputs queried at $x'_k \in D$. In fact, we can generalize formula (35) to represent $A(h)$ in such a way that the function h is approximated using its values $h(x_j)$ at the points $x_j \in \Omega$, $j = 1, \ldots, N$ but $A(h)$ can be evaluated at different points $x'_k \in D$, $k = 1, \ldots, K$. To accomplish this, we write (35) in the more general form that is closely related to cross attention. It is shown in (156) in Appendix G.5 that we can choose $\tilde{\Gamma}_\eta(\tilde{\mu}, ((x, y), (x', y'))) = \big((x, y), \Gamma_\eta((\rho_1)_\# \tilde{\mu}, (x', y'))\big)$. We consider the product measure $\mu \otimes \nu$, where $\mu = \gamma_h$ is a measure on $\overline{\Omega} \times [-L, L]^n$ and $\nu = \frac{1}{K} \sum_{k=1}^K \delta_{(x'_k,0)}$ is a measure on $\overline{D} \times [-L, L]^n$. Then, as seen in (164) in Appendix G.5,

$$(\rho_2)_\# G_{\tilde{\Gamma}_\eta}(\mu \otimes \nu) = \Gamma_\eta(\mu, \cdot)_\# \nu = \sum_{k=1}^K \frac{1}{K} \delta_{(x'_k, y_k)}, \quad y_k = A(\tilde{S}_\eta(\gamma_h))(x'_k). \tag{36}$$

Here, $\Gamma_\eta(\mu, \cdot)_\# \nu$ can be interpreted as a cross-attention: the measure $\mu = \gamma_h$ determines how the measure $\nu = \frac{1}{K} \sum \delta_{(x'_k,0)}$ associated to the grid point x'_k is pushed forward. This means that in (36), the input measure γ_h determines the regularized output function $A(\tilde{S}_\eta(\gamma_h))$, while the query measure specifies the evaluation points x'_k. The target values $A(h)(x'_k)$ are recovered as $\eta \to 0$. From the practical point of view, this suggests a simplified architecture that one can use to learn the operator A. We discuss the architecture and loss functions suitable for function-graph preserving maps in Appendix I, and illustrate this query-point formulation in Appendix K.

6 Conclusions

We have introduced function-graph preserving maps as a measure-theoretic framework for operator learning with transformers. By representing functions through graph measures, the proposed formulation connects finite tokenized inputs with continuum objects and makes discretization refinement amenable to analysis. Within this framework, structured measure maps preserve graph measures, allowing operators between function spaces to be represented through maps between measures.

The main results show that graph-preserving measure maps can be approximated by standard softmax self-attention layers and pointwise multilayer perceptrons. Consequently, continuous nonlinear operators between function spaces, including operators acting on low-regularity or distributional inputs after suitable regularization through a neural operator, admit approximation by transformer architectures in a discretization-invariant sense. The framework also accommodates operators between different domains, clarifying the role of spatial coordinates, positional encodings, and query points in transformer-based neural operators.

There are several limitations. While the main results are qualitative, we derive quantitative approximation estimates under additional regularity and stability assumptions, including an end-to-end convergence rate. However, these estimates do not establish sharp rates or optimal parameter complexity for transformers. The theory explains how standard transformer layers approximate the proposed measure maps, but does not identify optimal tokenization strategies. The results suggest that the geometry of function spaces, function-graph structure, and stability under discretization refinement should be central design principles for scientific transformer models.

Acknowledgments

TF is supported by JSPS KAKENHI Grant Number JP24K16949, 25H01453, JST CREST JPMJCR24Q5, JST ASPIRE JPMJAP2329. MVdH gratefully acknowledges the support of the Department of Energy BES, under grant DE-SC0020345, Oxy and the Simons Foundation under the MATH + X Program. The research of M.L. was partially supported by European Research Council

Advanced Grant 101097198 (PDE-Inverse) and the Research Council of Finland, grant 359182 (FAME Flagship). I.D. was partially supported by the European Research Council Consolidator Grant 101232533 (PhaseShift). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the other funding organizations.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, 2 edition, 2003.

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: a framework for efficiently scaling neural operators. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Shiba Biswal, Karthik Elamvazhuthi, and Rishi Sonthalia. Universal approximation of mean-field models via transformers. *arXiv preprint arXiv:2410.16295*, 2024.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private measures, random walks, and synthetic data. *Probability theory and related fields*, 189(1):569–611, 2024.

V. I. Bogachev. *Measure theory. Vol. I.* Springer-Verlag, Berlin, 2007. ISBN 978-3-540-34513-8; 3-540-34513-2. doi: 10.1007/978-3-540-34514-5. URL https://doi.org/10.1007/978-3-540-34514-5.

Haïm Brezis and Moshe Marcus. Hardy's inequalities revisited. *Annali della Scuola Normale Superiore di Pisa – Classe di Scienze (4)*, 25:217–237, 1997.

Edoardo Calvello, Nikola B. Kovachki, Matthew E. Levine, and Andrew M. Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1–52, 2025. URL http://jmlr.org/papers/v26/24-0879.html.

Shuhao Cao. Choose a transformer: fourier or galerkin. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5817–5840. PMLR, 21–27 Jul 2024.

Junfeng Chen and Kailiang Wu. Positional knowledge is all you need: position-induced transformer (pit) for operator learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Richard M. Dudley. *Real Analysis and Probability*. The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Takashi Furuya, Maarten de Hoop, and Matti Lassas. Transformers through the lens of support-preserving maps between measures, 2025a.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6S4WQD1LZR.

Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

David Gilbarg, Neil S Trudinger, David Gilbarg, and NS Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 1977.

Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman, Boston, 1985.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. Gnot: a general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes, 2024. URL https://arxiv.org/abs/2405.19101.

Florin Isaia. On the autonomous nemytskii operator between sobolev spaces in the critical and supercritical cases: Well-definedness and higher-order chain rule. *Nonlinear Analysis*, 214: 112576, 2022. ISSN 0362-546X. doi: https://doi.org/10.1016/j.na.2021.112576. URL https://www.sciencedirect.com/science/article/pii/S0362546X21002066.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rR03qFesqk.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I*. Die Grundlehren der mathematischen Wissenschaften, Band 181. Springer-Verlag, New York-Heidelberg, 1972a. Translated from the French by P. Kenneth.

Jacques-Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications, Vol. I*, volume 181 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1972b.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Regaldo-Saint Blancard, Francois Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, Francois

Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics, 2025. URL https://arxiv.org/abs/2511.15684.

William McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge, 2000.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/mei19a.html.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. URL https://arxiv.org/abs/2003.08934.

Till Muser, Alexandra Spitzer, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić. Flowers: A warp drive for neural pde solvers, 2026. URL https://arxiv.org/abs/2603.04430.

Jindřich Nečas. *Direct Methods in the Theory of Elliptic Equations*. Springer Monographs in Mathematics. Springer, Berlin, 2012.

Yu Safarov and Dmitri Vassiliev. *The asymptotic distribution of eigenvalues of partial differential operators*. Number 155. American Mathematical Soc., 1996.

Benjamin Shih, Ahmad Peyvan, Zhongqiang Zhang, and George Em Karniadakis. Transformers as neural operators for solutions of differential equations with finite regularity. *Computer Methods in Applied Mechanics and Engineering*, 434:117560, 2025. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2024.117560. URL https://www.sciencedirect.com/science/article/pii/S0045782524008144.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Vladimir M Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. In *Selected works of AN Kolmogorov*, pages 86–170. Springer, 1993.

V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958. ISSN 00364452. URL http://www.jstor.org/stable/25048365.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009a. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009b. doi: 10.1007/978-3-540-71050-9.

Sifan Wang, Jacob H Seidman, Shyam Sankaran, Hanwen Wang, George J. Pappas, and Paris Perdikaris. CVit: Continuous vision transformer for operator learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cRnCcuLvyr.

Greg Yang. Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation, April 2020.

A Basic notions

Below, $\Omega \subset \mathbb{R}^d$ is assumed to be bounded and open set.

For a Borel measure μ on $\overline{\Omega} \times [-L, L]^n$ and $\phi \in C_b(\overline{\Omega} \times \mathbb{R}^n)$ we denote

$$\langle \mu, \phi \rangle = \int_{\overline{\Omega} \times [-L,L]^n} \phi(x, y)\, d\mu(x, y).$$

When $F : \Omega \to D$ is a continuous map (or sometimes, even less regular) and μ is a measure on Ω, the push forward of the measure μ in the map F, denoted $F_{\#}\mu$, is a measure on D that gives for measurable sets $B \subset D$ the measure

$$F_{\#}\mu(B) = \mu(F^{-1}(B)), \quad F^{-1}(B) = \{x \in \Omega :\ F(x) \in B\}. \tag{37}$$

For probability measures μ_n and μ supported on the compact set $\overline{\Omega} \times [-L, L]^n$, we recall that μ_n converges to μ in p-Wasserstein metric if and only if probability measures μ_n converge to μ weakly, that is, for all continuous functions $\phi \in C(\overline{\Omega} \times \mathbb{R}^n)$ it holds that $\langle \mu_n, \phi \rangle \to \langle \mu, \phi \rangle$ as $n \to \infty$, see [Villani, 2009a, Theorem 7.12].

We denote $C_0(\overline{\Omega}) = \{u \in C(\overline{\Omega}) :\ u|_{\partial\Omega} = 0\}$. Also, for $s \in \mathbb{N}$,

$$H^s(\Omega) = \{u \in L^2(\Omega) :\ \partial^\alpha u \in L^2(\Omega) \text{ for all multi-indices } \alpha \text{ with } |\alpha| \le s\}$$

are the Sobolev spaces,

$$H^s_0(\Omega) = \mathrm{cl}_{H^s(\Omega)}(C^\infty_c(\Omega)).$$

Finally, $H^{-s}(\Omega)$ are the dual spaces of $H^s_0(\Omega)$. Sobolev spaces are defined also with $s \in \mathbb{R}$. By Sobolev embedding results in an open bounded set with C^∞-smooth boundary, or a cube $[0, 1]^d$, $H^s(\Omega) \subset C^k(\overline{\Omega})$ when $s > k + \frac{d}{2}$. On the cube this follows by the Sobolev extension theorem.

Recall that $I - \Delta_D$ is a positive selfadjoint operator on $L^2(\Omega)$, where Δ_D is the Dirichlet Laplacian with domain $\mathcal{D}(\Delta_D) = H^2(\Omega) \cap H^1_0(\Omega) = \{u \in H^2(\Omega) :\ u|_{\partial\Omega} = 0\}$. Let

$$A_D := I - \Delta_D, \qquad \mathcal{L} := A^r_D = (I - \Delta_D)^r,$$

where $r \in \mathbb{Z}_+$. Let $A_D\psi_j = \lambda_j\psi_j$ where $\{\psi_j\}_{j=1}^\infty$ is an $L^2(\Omega)$-orthonormal basis,

$$0 < \lambda_1 \le \lambda_2 \le \cdots, \qquad \lambda_j \to \infty.$$

Note the $\psi_j \in H^1_0(\Omega)$ satisfy $\psi_j|_{\partial\Omega} = 0$. The operator $\mathcal{L}$ is positive selfadjoint and has compact resolvent. Then,

$$\mathcal{L}\psi_j = \lambda^r_j\psi_j.$$

By Weyl's law for the Dirichlet Laplacian on bounded Lipschitz domains,

$$\lambda_j = \left(\frac{(2\pi)^d}{\omega_d|\Omega|}\right)^{2/d} j^{2/d}(1 + o(1)), \qquad j \to \infty.$$

In particular, after increasing the constant if necessary, there is a constant $C_{\mathcal{L}} > 0$ such that

$$\lambda_j \le C_{\mathcal{L}} j^{2/d}, \qquad j \ge 1. \tag{38}$$

Define the spectral Sobolev space associated with the Dirichlet operator by

$$H^r_D(\Omega) := \mathcal{D}(\mathcal{L}^{1/2}) = \mathcal{D}\big((I - \Delta_D)^{r/2}\big),$$

with norm

$$\|h\|^2_{H^r_D(\Omega)} := \|\mathcal{L}^{1/2}h\|^2_{L^2(\Omega)} = \sum_{j=1}^\infty \lambda^r_j |\langle h, \psi_j\rangle_{L^2(\Omega)}|^2.$$

For vector-valued functions $h = (h_1, \ldots, h_n)$, the norm is defined componentwise:

$$\|h\|^2_{H^r_D(\Omega;\mathbb{R}^n)} := \sum_{\ell=1}^n \|h_\ell\|^2_{H^r_D(\Omega)}.$$

As we have assumed that the boundary of $\partial\Omega$ is C^∞ smooth or that Ω is a cube, the operator $\mathcal{L}^{1/2}$ is selfadjoint and its domain $H^r_D(\Omega;\mathbb{R}^n)$ consists of Sobolev functions in $h \in H^r(\Omega;\mathbb{R}^n)$ that satisfy the Navier boundary conditions specified by its spectral definition above. This is the main reason for the assumptions on the boundary.

Let us next define a capping function Ψ that has the preferred properties: Let $\Psi : \mathbb{R}^n \to \mathbb{R}^n$, $\Psi(y) = (\psi(y_i))_{i=1}^n$ be a component-vice function where $\psi : \mathbb{R} \to \mathbb{R}$ is an odd C^∞-smooth function such that $\psi(y) = y$ for $|y| \leq (L + L_1)/2$, $\psi(\mathbb{R}) \subset [-L_1, L_1]$, and $\psi(y) = L_1$ for $y > L_1$. As described in the main text, we use the bounded function $\Psi \in C^\infty(\mathbb{R}^n;\mathbb{R}^n)$ to define a Nemytskii operator Isaia [2022] that caps values of function to the box $[-L_1, L_1]^n$ and denote $\tilde{S}_\eta(\gamma) = \Psi \circ S_\eta(\gamma)$.

B Mapping a sequence of tokens to a sequence of tokens and corresponding mapping from measures to measures

B.1 Part I: Simplest case when the input and output grids coincide

Functions between finite sequences of tokens are closely related to functions between discrete measures. To describe this relation, let $(t_1, t_2, \dots, t_N) \in (\mathbb{R}^D)^N$ be sequences of N tokens in $\mathbb{R}^D$. All sequences that are same up to a permutation of the order in the sequence, are considered to be equivalent, and we denote such equivalence classes of sequences of N tokens by X_D^N, and let the union of all these be $X_D = \bigcup_{N=1}^\infty X_D^N$. A sequence $(t_1, t_2, \dots, t_N) \in X_D^N$ can be identified with the probability measure $\mu = \sum_{i=1}^N \frac{1}{N}\delta_{t_i}$ that is a discrete measure supported in the set $\{t_1, \dots, t_N\}$. We denote the corresponding identification map by

$$\iota = \iota_N : (t_1, t_2, \dots, t_N) \to \sum_{i=1}^N \frac{1}{N}\delta_{t_i}. \tag{39}$$

Then a map $g : X_D^N \to X_D^N$ that for any N maps a sequence $(t_1, t_2, \dots, t_N)$ of D-dimensional tokens to a sequence $(y_1, y_2, \dots, y_N)$ of D-dimensional tokens, defines a map $G_N = \iota_N \circ g \circ \iota_N^{-1}$ from discrete probability measures

$$\iota_N(X_D^N) = \{\sum_{i=1}^N \frac{1}{N}\delta_{t_i} : \ t_i \in \mathbb{R}^D\}$$

to itself, $G_N : \iota_N(X_D^N) \to \iota_N(X_D^N)$, defined by

$$G_N(\sum_{i=1}^N \frac{1}{N}\delta_{t_i}) = \sum_{i=1}^N \frac{1}{N}\delta_{\hat{t}_i}, \quad (\hat{t}_1, \hat{t}_2, \dots, \hat{t}_N) = g(t_1, t_2, \dots, t_N),$$

Under suitable conditions, the maps G_N are compatible and define a map on the empirical probability measures, $G : \bigcup_{N\geq 1} \iota_N(X_D^N) \to \bigcup_{N\geq 1} \iota_N(X_D^N)$. When these maps satisfy a common uniform continuity estimate with respect to the Wasserstein metric, see Furuya et al. [2025b] for details, these maps extends to a map from probability measures with finite first moment to probability measures with finite first moment, $\overline{G} : \mathcal{P}_1(\mathbb{R}^D) \to \mathcal{P}_1(\mathbb{R}^D)$ that is continuous with respect to the Wasserstein metric. One can consider the Wasserstein distance as a generalization of the permutation invariant distance of sequences of tokens: We recall that the 1-Wasserstein distance of the measures $\mu = \sum_{i=1}^N \frac{1}{N}\delta_{t_i}$ and $\mu' = \sum_{i=1}^N \frac{1}{N}\delta_{t'_i}$ is given by

$$W_1(\mu, \mu') = \min_\sigma \frac{1}{N}\sum_{i=1}^N |t_i - t'_{\sigma(i)}|,$$

where the minimum is taken over the permutations, $\sigma : \{1, 2, \dots, N\} \to \{1, 2, \dots, N\}$. When the number, N, of tokens grows, the discrete measures $\sum_{i=1}^N \frac{1}{N}\delta_{t_i}$ can converge to continuous measures (or a sum of continuous measures, point measures, and measures that are singular with respect to the standard measure of $\mathbb{R}^D$.) Thus to understand properties of transformers on arbitrarily long sequences, it is useful to consider mappings between "general" measures.

Let us next consider an approximation of nonlinear operator $A : C(\overline{\Omega}; [-L, L]^n) \to C(\overline{\Omega}; [-L, L]^n)$, using transformers, where $\Omega \subset \mathbb{R}^d$. Let $x_j \in \overline{\Omega}$, $j = 1, 2, \ldots, N$ be computation grid points in the domain Ω where both the input and output functions are given (This is simple case where the construction of the context functions and a flow diagram are easier to describe). When $h \in C(\overline{\Omega}; [-L, L]^n)$ is the input function, we its point values determine a token sequences of N tokens

$$\Big((x_j, h(x_j)) \Big)_{j \in [N]} \tag{40}$$

With the above preparations, let consider the function-graph preserving map $\overline{G} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$, $\Omega \subset \mathbb{R}^d$, see Definition 3, from a point of view of token sequences. When $\overline{G}$ maps the graph measure γ_h of a function $h \in C(\overline{\Omega}; [-L, L]^n)$ to $\overline{G}(\gamma_h) = \gamma_{J(h)}$ that is the graph measure of a function $J(h) \in C(\overline{\Omega}; [-L, L]^n)$, we define a sequence-to-sequence map between the token sequences,

$$\overline{g} = \mathbf{j}_{x,\varepsilon} \circ \overline{G} \circ \iota : X^N_{d+n} \to X^N_{d+n} \tag{41}$$

where $\mathbf{j}_{x,\varepsilon}$ is approximation of the map ι_N^{-1}, associated to grid points $x = (x_1, \ldots, x_N)$, that is given by

$$\mathbf{j}_{x,\varepsilon} : \gamma \to \Big((x_1, \mathbf{F}_\varepsilon \gamma(x_1)) \,,\, (x_2, \mathbf{F}_\varepsilon \gamma(x_2)) \,, \ldots ,\, (x_N, \mathbf{F}_\varepsilon \gamma(x_N)) \Big), \tag{42}$$

where $\mathbf{F}_\varepsilon$ is the convolution operator with a mollifier function, see formula (10). This mapping has the form (when the domain deformation map $\Phi : \overline{\Omega} \to \overline{\Omega}$ in Def. 3 is the identity map)

$$\overline{g} : \Big((x_1, h(x_1)), (x_2, h(x_2)), \ldots, (x_N, h(x_N)) \Big) \to \Big((x_1, u_1), (x_2, u_2), \ldots, (x_N, u_N) \Big), \tag{43}$$

where $x_j \in \overline{\Omega}$ are computation grid points in the domain Ω and $u_j \approx J(h)(x_j)$. Our considerations show that u_j converges to $J(h)(x_j)$ when the grid points x_j are suitably chosen and $N \to \infty$. Moreover, we show that the function-graph preserving maps can be approximations by traditional softmax-based transformers $\overline{G}^{\mathrm{tran}} = G^{\mathrm{tran}}_{\Gamma_{\eta,m}}$, given by Theorem 1, that satisfies

$$\sup_{\mu \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)} W_1(\overline{G}^{\mathrm{tran}}(\mu), \overline{G}(\mu)) \leq \varepsilon, \tag{44}$$

where $\varepsilon > 0$ depends on the depth of the transformer $\overline{G}^{\mathrm{tran}}$ and W_1 is the 1-Wasserstein distance. This defines a sequence of tokens to sequence of tokens map

$$\overline{g}^{\mathrm{tran}} = \mathbf{j}_{x,\varepsilon} \circ \overline{G}^{\mathrm{tran}} \circ \iota : X^N_{d+n} \to X^N_{d+n}. \tag{45}$$

B.2 Part II: The case when the input and output grids are different

Let us next consider a nonlinear operator $A : C(\overline{\Omega}; [-L, L]^n) \to C(\overline{\Omega}; [-L, L]^n)$, where $\Omega \subset \mathbb{R}^d$. Let $x_j \in \overline{\Omega}$, $j = 1, 2, \ldots, J$ are computation grid points in the domain Ω where the input function is given and $z_k \in \overline{\Omega}$, $k = 1, 2, \ldots, K$ are another computation grid points in the domain Ω where the output function as given.

When $h \in \mathrm{cl}_{H^{-s}(\Omega;\mathbb{R}^n)}(C(\overline{\Omega}; [-L, L]^n))$, or alternatively, $h \in C(\overline{\Omega}; [-L, L]^n)$, is the input function, we consider the regularized input function $h_\tau = R_\tau(h)$,

$$R_\tau h = \rho_\tau * (h \cdot (\rho_\tau * \mathbf{1}_{\Omega_\tau})),$$

where $\rho_\tau \in C^\infty_c(B(0, \tau))$ is a smooth mollifier function, see (12). These point values determine the token sequences of $N = J \cdot K$ tokens (that are tensorized)

$$\Big((x_j, h_\tau(x_j)); (z_k, 0)) \Big)_{(j,k) \in [J] \times [K]} \tag{46}$$

Observe that this sequence of has a large number, $N = J \cdot K$, of tokens. However, even for linear maps we observe that if one wants to parametrize a general linear map (i.e., a matrix) $A_{lin} = (h_\tau(x_j))_{j \in [J]} \to (g(x_k)))_{k \in [K]}$, one needs $N = J \cdot K$ matrix elements.

Let us also consider the function-graph preserving map

$$\overline{G} : \mathcal{P}((\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n)) \to \mathcal{P}((\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n)),$$

see Definition 3, from a point of view of token sequences (when the domain deformation map $\Phi : \overline{\Omega} \to \overline{\Omega}$ in Def. 3 is the identity map). When $\overline{G}$ maps $\gamma_h \otimes \nu$, where $h \in C(\overline{\Omega}; [-L, L]^n)$ and $\nu = \frac{1}{K} \sum_{k=1}^{K} \delta_{(z_k, 0)}$, to

$$\overline{G}(\gamma_h \otimes \nu) = \gamma_h \otimes \tfrac{1}{K} \sum_{k=1}^{K} \delta_{(z_k, J(h)(z_k))},$$

where $J(h) \in C(\overline{\Omega}; [-L, L]^n)$ is approximately the function $A(h) \in C(\overline{\Omega}; [-L, L]^n)$. Also, we use the map $(\rho_2)_\# : \mu_1 \otimes \nu_1 \to \nu$. Then, we define the sequence-to-sequence map between the token sequences,

$$\overline{g} = \mathbf{j}_{\vec{z}, \varepsilon} \circ (\rho_2)_\# \circ \overline{G} \circ \iota : X_{d+n}^N \to X_{d+n}^K \tag{47}$$

where $\mathbf{j}_{\vec{z}, \varepsilon}$ is a map similar to (42), but is associated to $\vec{z} = (z_1, \ldots, z_K)$, that is given by

$$\mathbf{j}_{\vec{z}, \varepsilon} : \gamma \to \Big((z_1, \mathbf{F}_\varepsilon \gamma(z_1)), (z_2, \mathbf{F}_\varepsilon \gamma(z_2)), \ldots, (z_K, \mathbf{F}_\varepsilon \gamma(z_K)) \Big), \tag{48}$$

where $\mathbf{F}_\varepsilon$ is the convolution operator with a mollifier function, see formula (10). This gives a map the form

$$\overline{g} : \Big((x_j, h(x_j)); (z_k, 0)) \Big)_{(j,k) \in [J] \times [K]} \to \Big((z_1, t_1), (z_2, t_2), \ldots, (z_K, t_K) \Big) \tag{49}$$

where and t_k is approximately equal to $J(h)(z_k)$, and our considerations show that t_k converges to $J(h)(z_k)$ when the grid points z_k are suitably chosen, $J, K \to \infty$ and $\varepsilon \to 0$.

C Regularized functions $R_\tau h$ converge to h.

In this section we show the following result:

Proposition 2. *Let $s > 0$ with $s - \frac{1}{2} \notin \mathbb{Z}$. Then for every $h \in H^{-s}(\Omega)$, $\lim_{\tau \to 0^+} \big\| R_\tau h - h \big\|_{H^{-s}(\Omega)} = 0$.*

C.1 Setting and notation

Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with C^∞-smooth boundary, or a cube $[0, 1]^d$, , and let $d(x) := \operatorname{dist}(x, \partial\Omega)$. For $s > 0$ we use the convention

$$H_0^s(\Omega) := \overline{C_c^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}}, \qquad H^{-s}(\Omega) := \big(H_0^s(\Omega)\big)^*,$$

where $\overline{C_c^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}}$ is the closure of the set $C_c^\infty(\Omega)$ in $H^s(\Omega)$ and $H^{-s}(\Omega)$ is the topological dual of $H_0^s(\Omega)$, with duality pairing $\langle \cdot, \cdot \rangle_{H^{-s}, H_0^s}$ extending the $L^2(\Omega)$ inner product.

Fix $\rho \in C_c^\infty(\mathbb{R}^d)$ even, nonnegative, with $\int_{\mathbb{R}^d} \rho dx = 1$ and supp $\rho \subset \overline{B(0, 1)}$. Set

$$\rho_\tau(x) := \tau^{-d} \rho(x/\tau), \qquad \Omega_\tau := \{x \in \Omega : d(x) > 4\tau\}, \qquad \chi_\tau := \rho_\tau * \mathbf{1}_{\Omega_\tau},$$
$$T_r := \{x \in \mathbb{R}^d : d(x) \le r\}.$$

Let $\tau_0 > 0$ be small enough that $\Omega_{\tau_0} \neq \emptyset$. The regularization studied is

$$R_\tau h := \rho_\tau * \big(h \cdot \chi_\tau\big), \qquad h \in H^{-s}(\Omega),\ \tau \in (0, \tau_0].$$

Observe that since ρ_τ is non-negative and its integral is 1, we have for $h \in C(\Omega)$

$$\|R_\tau h\|_{L^\infty(\Omega)} \le \|h\|_{L^\infty(\Omega)}. \tag{50}$$

Since $\chi_\tau \in C_c^\infty(\Omega)$ (Lemma 4 below), the product $h\chi_\tau$ is defined as the distribution (i.e., a generalized function)

$$\langle h\chi_\tau, \psi \rangle_{\mathcal{D}'(\mathbb{R}^d), \mathcal{D}(\mathbb{R}^d)} := \langle h, \chi_\tau \psi \rangle_{H^{-s}, H_0^s} \qquad (\psi \in \mathcal{D}(\mathbb{R}^d)),$$

which lies in the space of compactly supported distributions $\mathcal{E}'(\mathbb{R}^d)$ with support in supp $\chi_\tau \Subset \Omega$. Convolution with ρ_τ thus yields $R_\tau h \in C_c^\infty(\mathbb{R}^d)$ with support in $\overline{\Omega}$.

C.2 Approximation result via convolutions

Proposition 3. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set with C^∞-smooth boundary, or a cube $[0,1]^d$. Let $s > 0$ with $s - \frac{1}{2} \notin \mathbb{Z}$. Then for every $h \in H^{-s}(\Omega)$,*

$$\lim_{\tau \to 0^+} \left\| R_\tau h - h \right\|_{H^{-s}(\Omega)} = 0.$$

The proof proceeds via Banach–Steinhaus (uniform boundedness) plus a density argument. We first collect the necessary lemmas.

Auxiliary lemmas

Lemma 4 (Geometry of χ_τ)**.** *For all $\tau \in (0, \tau_0]$:*

(i) $\chi_\tau \in C^\infty(\mathbb{R}^d)$, $0 \le \chi_\tau \le 1$;

(ii) *supp* $\chi_\tau \subset \{x \in \Omega : d(x) \ge 3\tau\} \Subset \Omega$;

(iii) $\chi_\tau(x) = 1$ *whenever* $x \in \Omega$ *and* $d(x) \ge 5\tau$;

(iv) $\|\partial^\alpha \chi_\tau\|_{L^\infty(\mathbb{R}^d)} \le C_\alpha \, \tau^{-|\alpha|}$ *for every multi-index* α, *with* $C_\alpha = \|\partial^\alpha \rho\|_{L^1(\mathbb{R}^d)}$;

(v) *for every multi-index* α *with* $|\alpha| \ge 1$, *supp* $(\partial^\alpha \chi_\tau) \subset T_{5\tau}$.

Proof. (i) χ_τ is the convolution of an L^∞ indicator with a C_c^∞ kernel, hence C^∞; the bound $0 \le \chi_\tau \le 1$ follows from $0 \le \mathbf{1}_{\Omega_\tau} \le 1$ and $\int \rho_\tau dx = 1$.

(ii) If $\chi_\tau(x) \ne 0$, there exists $y \in \Omega_\tau$ with $|x - y| < \tau$; then $d(x) \ge d(y) - |x - y| > 4\tau - \tau = 3\tau$.

(iii) If $x \in \Omega$ and $d(x) \ge 5\tau$, then for every $|x - y| \le \tau$ one has $d(y) \ge 5\tau - \tau = 4\tau$, so $B(x, \tau) \subset \Omega_\tau$ and $\chi_\tau(x) = \int_{B(x,\tau)} \rho_\tau(x - y) dy = 1$.

(iv) $\partial^\alpha \chi_\tau = (\partial^\alpha \rho_\tau) * \mathbf{1}_{\Omega_\tau}$, so by Young's inequality, $\|\partial^\alpha \chi_\tau\|_\infty \le \|\partial^\alpha \rho_\tau\|_{L^1(\mathbb{R}^d)} = \tau^{-|\alpha|} \|\partial^\alpha \rho\|_{L^1(\mathbb{R}^d)}$.

(v) On the open set $\{x \in \Omega : d(x) > 5\tau\}$ one has $\chi_\tau \equiv 1$ by (iii); outside $\{x \in \Omega : d(x) \ge 3\tau\}$ one has $\chi_\tau \equiv 0$ by (ii). Hence every derivative of χ_τ of order α, $|\alpha| \ge 1$ vanishes there, so its support lies in $\{x \in \Omega : 3\tau \le d(x) \le 5\tau\} \subset T_{5\tau}$. □

Lemma 5 (Density of $C_c^\infty(\Omega)$ in $H^{-s}(\Omega)$)**.** *For every $s > 0$, $C_c^\infty(\Omega)$ is dense in $H^{-s}(\Omega)$.*

Proof. The inclusion $H_0^s(\Omega) \hookrightarrow L^2(\Omega)$ is continuous and dense ($C_c^\infty(\Omega)$ is dense in both). Taking duals (of a continuous, dense embedding of Hilbert spaces) gives a continuous embedding $L^2(\Omega) \hookrightarrow H^{-s}(\Omega)$ with dense image. Since $C_c^\infty(\Omega)$ is dense in $L^2(\Omega)$, it is dense in $H^{-s}(\Omega)$. □

Lemma 6 (Adjoint identity)**.** *Define, for $\varphi \in H_0^s(\Omega)$ with zero extension $\widetilde{\varphi} \in H^s(\mathbb{R}^d)$,*

$$A_\tau \varphi := \chi_\tau \cdot (\rho_\tau * \widetilde{\varphi}).$$

Then $A_\tau \varphi \in C_c^\infty(\Omega)$, and for every $h \in H^{-s}(\Omega)$,

$$\langle R_\tau h, \varphi \rangle_{H^{-s}, H_0^s} = \langle h, A_\tau \varphi \rangle_{H^{-s}, H_0^s}.$$

In particular,

$$\|R_\tau\|_{H^{-s}(\Omega) \to H^{-s}(\Omega)} = \|A_\tau\|_{H_0^s(\Omega) \to H_0^s(\Omega)}. \tag{51}$$

Proof. The function $\rho_\tau * \widetilde{\varphi}$ is C^∞ on $\mathbb{R}^d$, and $\chi_\tau \in C_c^\infty(\Omega)$, so $A_\tau \varphi \in C_c^\infty(\Omega) \subset H_0^s(\Omega)$.

Next we use the duality: It holds that $R_\tau h \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^n) \subset L^2(\Omega; \mathbb{R}^n)$, so

$$\langle R_\tau h, \varphi \rangle = \int_\Omega R_\tau h \, \varphi dx = \int_{\mathbb{R}^d} (\rho_\tau * u)(x) \, \widetilde{\varphi}(x) dx, \qquad u := h \chi_\tau \in \mathcal{E}'(\mathbb{R}^d).$$

Because ρ_τ is even, the standard adjoint identity for convolution gives

$$\int_{\mathbb{R}^d} (\rho_\tau * u) \, \widetilde{\varphi} dx = \langle u, \rho_\tau * \widetilde{\varphi} \rangle_{\mathcal{E}', C^\infty},$$

which can be checked for test functions and extended by approximation since u has compact support. Then by the definition of $h\chi_\tau$ as a distribution,

$$\langle u, \rho_\tau * \widetilde{\varphi}\rangle = \langle h, \chi_\tau \cdot (\rho_\tau * \widetilde{\varphi})\rangle_{H^{-s},H_0^s} = \langle h, A_\tau\varphi\rangle.$$

Identifying $H^{-s}(\Omega)$ with $(H_0^s(\Omega))^*$ shows that R_τ is the Banach-space adjoint of A_τ, hence we obtain the identity (51). □

Lemma 7 (Iterated Hardy inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with smooth boundary or a cube, and $j \in \mathbb{Z}_+$. There exists $C = C(\Omega, j)$ such that*

$$\int_\Omega \frac{|g(x)|^2}{d(x)^{2j}}dx \le C\,\|g\|^2_{H^j(\Omega)} \qquad \text{for all } g \in H_0^j(\Omega).$$

Proof. For $g \in C_c^\infty(\Omega)$, use the one-dimensional iterated Hardy inequality $\int_0^a |v(t)|^2 t^{-2j}\,dt \le C_j \int_0^a |v^{(j)}(t)|^2\,dt$ for $v \in C_c^\infty(0,a)$; see [Grisvard, 1985, Thm. 1.4.4.4], [Nečas, 2012, Thm. 21.3], and [Adams and Fournier, 2003, Thm. 4.36]. On a cube, d is the minimum of the face distances t_ℓ, so $d^{-2j} \le \sum_{\ell=1}^{2d} t_\ell^{-2j}$. Apply the one-dimensional inequality along each face normal and integrate over the other coordinates. On a smooth domain, use a finite partition of unity in smooth normal-coordinate boundary charts, where the normal variable is the distance to the boundary. The same inequality applies along normals; coordinate changes and cutoff derivatives are controlled by the H^j norm. The interior contribution is bounded by the L^2 norm because d is bounded below there. See also Brezis and Marcus [1997]. Approximation in H^j by C_c^∞ functions and Fatou's lemma give the result for $H_0^j(\Omega)$. □

Lemma 8 (Boundary-layer mollification estimate)**.** *Let $j \in \mathbb{Z}_+ \cup \{0\}$. There exists $C = C(\Omega, \rho, j)$ such that for every $g \in H_0^j(\Omega)$ (with the convention $H_0^0(\Omega) := L^2(\Omega)$) and every $\tau \in (0, \tau_0]$,*

$$\big\|\rho_\tau * \widetilde{g}\big\|_{L^2(T_{5\tau})} \le C\,\tau^j\,\|g\|_{H^j(\Omega)},$$

where $\widetilde{g}$ denotes the zero extension of g to $\mathbb{R}^d$.

Proof. By the Cauchy–Schwarz inequality applied with the probability measure $\rho_\tau(x-\cdot)dy$,

$$\big|(\rho_\tau * \widetilde{g})(x)\big|^2 \le 1 \cdot \int_{\mathbb{R}^d} \rho_\tau(x-y)\,|\widetilde{g}(y)|^2 dy = \big(\rho_\tau * |\widetilde{g}|^2\big)(x).$$

Integrating over $T_{5\tau}$ and applying Fubini,

$$\big\|\rho_\tau * \widetilde{g}\big\|^2_{L^2(T_{5\tau})} \le \int_{\mathbb{R}^d} |\widetilde{g}(y)|^2 \Big(\int_{T_{5\tau}} \rho_\tau(x-y)dx\Big)dy.$$

The inner integral is bounded by $\int_{\mathbb{R}^d} \rho_\tau(x-y)dx = 1$, and is nonzero only when $\operatorname{dist}(y, T_{5\tau}) \le \tau$, i.e. $y \in T_{6\tau}$. Combined with $\widetilde{g} = 0$ outside $\overline{\Omega}$,

$$\big\|\rho_\tau * \widetilde{g}\big\|^2_{L^2(T_{5\tau})} \le \int_{T_{6\tau}\cap\Omega} |g(y)|^2 dy.$$

For $j = 0$ this gives $\|\rho_\tau * \widetilde{g}\|_{L^2(T_{5\tau})} \le \|g\|_{L^2(\Omega)}$, i.e. the claim with $\tau^0 = 1$. For $j \ge 1$ we apply Lemma 7:

$$\int_{T_{6\tau}\cap\Omega} |g|^2 dy = \int_{T_{6\tau}\cap\Omega} \frac{|g|^2}{d^{2j}}\,d^{2j}dy \le (6\tau)^{2j} \int_\Omega \frac{|g|^2}{d^{2j}}dy \le C(6\tau)^{2j}\,\|g\|^2_{H^j(\Omega)}.$$

Taking square roots of both sides of the above inequality completes the proof. □

Lemma 9 (Uniform boundedness of A_τ at integer order)**.** *For each $k \in \mathbb{Z}_+ \cup \{0\}$ there exists $C = C(\Omega, \rho, k)$ such that*

$$\|A_\tau\varphi\|_{H^k(\Omega)} \le C\,\|\varphi\|_{H^k(\Omega)} \qquad \text{for all } \varphi \in H_0^k(\Omega),\ \tau \in (0, \tau_0].$$

Proof. The case $k = 0$: $\|A_\tau\varphi\|_{L^2} = \|\chi_\tau(\rho_\tau * \widetilde{\varphi})\|_{L^2} \le \|\rho_\tau * \widetilde{\varphi}\|_{L^2} \le \|\widetilde{\varphi}\|_{L^2} = \|\varphi\|_{L^2(\Omega)}$, using $\|\chi_\tau\|_\infty \le 1$ and Young's inequality, as mollification is an L^2-contraction.

Assume $k \geq 1$. Recall $A_\tau \varphi \in C_c^\infty(\Omega)$, so its zero extension to $\mathbb{R}^d$ is itself a $C_c^\infty(\mathbb{R}^d)$ function and

$$\|A_\tau \varphi\|_{H^k(\Omega)} = \|A_\tau \varphi\|_{H^k(\mathbb{R}^d)} = \Big(\sum_{|\alpha| \leq k} \|\partial^\alpha A_\tau \varphi\|_{L^2(\mathbb{R}^d)}^2 \Big)^{1/2}.$$

By Leibniz rule, for $|\alpha| \leq k$

$$\partial^\alpha A_\tau \varphi = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial^\beta \chi_\tau) \cdot \partial^{\alpha-\beta}(\rho_\tau * \widetilde{\varphi}).$$

We use that for $\beta \leq \alpha$ and $\varphi \in H_0^k(\Omega)$, $\partial^{\alpha-\beta}\varphi \in H_0^{k-|\alpha-\beta|}(\Omega)$ and $\partial^{\alpha-\beta}\widetilde{\varphi} = \widetilde{\partial^{\alpha-\beta}\varphi}$ in $\mathcal{D}'(\mathbb{R}^d)$, where $\widetilde{\partial^{\alpha-\beta}\varphi}$ is the zero-extension of $\partial^{\alpha-\beta}\varphi$ to $\mathbb{R}^d$. Consequently, $\partial^{\alpha-\beta}(\rho_\tau * \widetilde{\varphi}) = \rho_\tau * \widetilde{\partial^{\alpha-\beta}\varphi}$.

Term $\beta = 0$. Using $\|\chi_\tau\|_\infty \leq 1$ and Young's inequality,

$$\big\|\chi_\tau \cdot \rho_\tau * \widetilde{\partial^\alpha \varphi}\big\|_{L^2} \leq \|\partial^\alpha \varphi\|_{L^2(\Omega)} \leq \|\varphi\|_{H^k(\Omega)}.$$

Term $|\beta| \geq 1$. By Lemma 4(iv,v), $\partial^\beta \chi_\tau$ is supported in $T_{5\tau}$ with $\|\partial^\beta \chi_\tau\|_\infty \leq C_\beta \tau^{-|\beta|}$, so

$$\big\|(\partial^\beta \chi_\tau) \cdot \rho_\tau * \widetilde{\partial^{\alpha-\beta}\varphi}\big\|_{L^2} \leq C_\beta\, \tau^{-|\beta|}\, \big\|\rho_\tau * \widetilde{\partial^{\alpha-\beta}\varphi}\big\|_{L^2(T_{5\tau})}. \tag{52}$$

Set $g := \partial^{\alpha-\beta}\varphi \in H_0^{k-|\alpha-\beta|}(\Omega)$. Since $|\alpha| \leq k$ implies $k - |\alpha-\beta| = k - |\alpha| + |\beta| \geq |\beta|$, we have $g \in H_0^{|\beta|}(\Omega)$ with $\|g\|_{H^{|\beta|}} \leq \|\varphi\|_{H^k}$. Lemma 8 with $j = |\beta|$ gives

$$\big\|\rho_\tau * \widetilde{g}\big\|_{L^2(T_{5\tau})} \leq C\, \tau^{|\beta|}\, \|g\|_{H^{|\beta|}(\Omega)} \leq C\, \tau^{|\beta|}\, \|\varphi\|_{H^k(\Omega)}. \tag{53}$$

Combining the formulas (52) and (53), we obtain

$$\big\|(\partial^\beta \chi_\tau) \cdot \rho_\tau * \widetilde{\partial^{\alpha-\beta}\varphi}\big\|_{L^2} \leq C\, \|\varphi\|_{H^k(\Omega)},$$

with C independent of τ (the powers $\tau^{-|\beta|}$ and $\tau^{|\beta|}$ cancel exactly).

Summing over $\beta \leq \alpha$ and over $|\alpha| \leq k$ yields $\|A_\tau \varphi\|_{H^k(\mathbb{R}^d)} \leq C(k)\, \|\varphi\|_{H^k(\Omega)}$, which is the claim because $\|A_\tau \varphi\|_{H^k(\Omega)} = \|A_\tau \varphi\|_{H^k(\mathbb{R}^d)}$. □

Lemma 10 (Uniform boundedness of A_τ at fractional order)**.** *Let $s > 0$ with $s - \frac{1}{2} \notin \mathbb{Z}$. There exists $C = C(\Omega, \rho, s)$ such that*

$$\|A_\tau \varphi\|_{H_0^s(\Omega)} \leq C\, \|\varphi\|_{H_0^s(\Omega)} \qquad \text{for all } \varphi \in H_0^s(\Omega),\ \tau \in (0, \tau_0].$$

Equivalently, $\sup_{\tau \in (0,\tau_0]} \|R_\tau\|_{H^{-s}(\Omega) \to H^{-s}(\Omega)} < \infty$.

Proof. If $s \in \mathbb{Z}_+$, the claim is Lemma 9. Assume $s \notin \mathbb{Z}_+$ and choose $N := \lceil s \rceil \in \mathbb{Z}_+$, so that $\theta := s/N \in (0, 1)$. By Lemma 9, the linear maps $A_\tau : L^2(\Omega) \to L^2(\Omega)$ and $A_\tau : H_0^N(\Omega) \to H_0^N(\Omega)$ are bounded with norms uniformly bounded by some M_0, M_N independent of $\tau \in (0, \tau_0]$.

The standard interpolation identity for the spaces H_0^s on a Lipschitz domain reads

$$\big[L^2(\Omega),\, H_0^N(\Omega)\big]_\theta = H_0^{\theta N}(\Omega) = H_0^s(\Omega),$$

provided that $s = \theta N \notin \frac{1}{2} + \mathbb{Z}$, i.e. $s - \frac{1}{2} \notin \mathbb{Z}$, see [McLean, 2000, Thm. B.8 and Thm. 3.33] or [Lions and Magenes, 1972b, Ch. 1, §9 and Thm. 11.6].

The complex interpolation theorem for bounded linear operators [McLean, 2000, Thm. B.2(iii)] then yields

$$\|A_\tau\|_{H_0^s(\Omega) \to H_0^s(\Omega)} \leq M_0^{1-\theta} M_N^\theta,$$

which is uniform in $\tau \in (0, \tau_0]$. The equivalent statement on R_τ follows from Lemma 6. □

Remark 4 (On the integer case)**.** *For integer $s \in \mathbb{Z}_+$ no interpolation is needed: Lemma 9 gives uniform boundedness directly, so the conclusion of Proposition 3 holds for every $s \in \mathbb{Z}_+$ (which all satisfy $s - \frac{1}{2} \notin \mathbb{Z}$).*

Lemma 11. *For every $h \in C_c^\infty(\Omega)$,*

$$\lim_{\tau \to 0^+} \|R_\tau h - h\|_{H^{-s}(\Omega)} = 0.$$

Proof. Set $\delta := \operatorname{dist}(\operatorname{supp} h, \partial\Omega) > 0$. For $\tau < \delta/5$, Lemma 4(iii) gives $\chi_\tau \equiv 1$ on supp (h), hence $h\chi_\tau = h$ and $R_\tau h = \rho_\tau * h$. Standard properties of mollifiers yield $\rho_\tau * h \to h$ in $C_c^\infty(\mathbb{R}^d)$, and in particular in $L^2(\Omega)$. The continuous embedding $L^2(\Omega) \hookrightarrow H^{-s}(\Omega)$ (which is the dual of the continuous embedding $H_0^s(\Omega) \hookrightarrow L^2(\Omega)$) yields convergence in $H^{-s}(\Omega)$. □

Proof. Let $\varepsilon > 0$ and $h \in H^{-s}(\Omega)$. By Lemma 5 there exists $h_\varepsilon \in C_c^\infty(\Omega)$ with $\|h - h_\varepsilon\|_{H^{-s}(\Omega)} < \varepsilon$. Let us decompose functions as

$$R_\tau h - h = R_\tau(h - h_\varepsilon) + (R_\tau h_\varepsilon - h_\varepsilon) - (h - h_\varepsilon).$$

By Lemma 10 there is C^* independent of $\tau \in (0, \tau_0]$ such that

$$\|R_\tau(h - h_\varepsilon)\|_{H^{-s}(\Omega)} \le C^* \|h - h_\varepsilon\|_{H^{-s}(\Omega)} < C^*\varepsilon.$$

By above assumption on h_ε, we have $\|h - h_\varepsilon\|_{H^{-s}(\Omega)} < \varepsilon$. By Lemma 11 there exists $\tau_\varepsilon \in (0, \tau_0]$ such that

$$\|R_\tau h_\varepsilon - h_\varepsilon\|_{H^{-s}(\Omega)} < \varepsilon \qquad \text{for all } \tau \in (0, \tau_\varepsilon).$$

Combining,

$$\|R_\tau h - h\|_{H^{-s}(\Omega)} \le (C^* + 2)\,\varepsilon \qquad \text{for all } \tau \in (0, \tau_\varepsilon).$$

Since $\varepsilon > 0$ was arbitrary, $\|R_\tau h - h\|_{H^{-s}(\Omega)} \to 0$ as $\tau \to 0^+$. □

D Proof of Lemma 1

Let us recall the setting. Let $\Omega \subset \mathbb{R}^d$ be a non-empty bounded open set, and set

$$\overline{\Omega} = \operatorname{cl}(\Omega).$$

Let $L > 0$, $n \in \mathbb{Z}_+$, and define

$$X := \overline{\Omega} \times [-L, L]^n \subset \mathbb{R}^{d+n}.$$

The space X is equipped with the Euclidean metric inherited from $\mathbb{R}^{d+n}$. Let $\overline{\lambda}_{\overline{\Omega}} = \frac{1}{\operatorname{vol}(\Omega)} \lambda_{\overline{\Omega}}$ denote the normalized Lebesgue measure on $\overline{\Omega}$. Thus $\operatorname{spt} \lambda_{\overline{\Omega}} = \overline{\Omega}$.

For $h \in C(\overline{\Omega}; [-L, L]^n)$ define its graph measure by

$$\gamma_h := (\mathrm{id}_{\overline{\Omega}}, h)_\# \overline{\lambda}_\Omega \in \mathcal{P}(X).$$

Set

$$\mathcal{G} := \{\gamma_h : h \in C(\overline{\Omega}; [-L, L]^n)\}.$$

Since X is compact, $\mathcal{P}(X) = \mathcal{P}_2(X)$, and we equip $\mathcal{P}(X)$ with the 1-Wasserstein distance W_1 induced by the Euclidean metric on X.

Let

$$\Gamma : \mathcal{P}(X) \times X \to X$$

be continuous, and assume:

$$\forall h \in C(\overline{\Omega}; [-L, L]^n)\ \exists g \in C(\overline{\Omega}; [-L, L]^n) \qquad \text{such that } \Gamma(\gamma_h, \cdot)_\# \gamma_h = \gamma_g. \tag{P}$$

$$\|J(h_1) - J(h_2)\|_{C(\overline{\Omega}; \mathbb{R}^n)} \le L_0\, W_1(\gamma_{h_1}, \gamma_{h_2}) \tag{L1}$$

for all $h_1, h_2 \in C(\overline{\Omega}; [-L, L]^n)$.

$$|J(h)(x_1) - J(h)(x_2)|_{\mathbb{R}^n} \le L_1 |x_1 - x_2| \tag{L2$_{\text{lin}}$}$$

for all $h \in C(\overline{\Omega}; [-L, L]^n)$ and all $x_1, x_2 \in \overline{\Omega}$.

Here J is the graph-update map determined by (P); its well-definedness is proved below.

Proposition 4. *Assume* (P)*,* (L1)*, and* ($\mathrm{L2}_{\mathrm{lin}}$)*. Then there exists a map*

$$\widehat{p} : \mathcal{P}(X) \times \overline{\Omega} \to [-L, L]^n$$

such that

$$\widehat{p}(\gamma_h, x) = J(h)(x) \qquad \forall h \in C(\overline{\Omega}; [-L, L]^n), \quad \forall x \in \overline{\Omega},$$

and

$$|\widehat{p}(\mu_1, x_1) - \widehat{p}(\mu_2, x_2)|_{\mathbb{R}^n} \le \sqrt{n}\, L_0\, W_1(\mu_1, \mu_2) + \sqrt{n}\, L_1\, |x_1 - x_2| \tag{54}$$

for all $\mu_1, \mu_2 \in \mathcal{P}(X)$ and all $x_1, x_2 \in \overline{\Omega}$.

Define

$$\widehat{\Gamma} : \mathcal{P}(X) \times X \to X, \qquad \widehat{\Gamma}(\mu, (x, y)) \coloneqq (x, \widehat{p}(\mu, x)).$$

Then $\widehat{\Gamma}$ is Lipschitz with respect to the product metric

$$d\big((\mu, z), (\nu, z')\big) \coloneqq W_1(\mu, \nu) + |z - z'|_{\mathbb{R}^{d+n}}, \qquad z, z' \in X,$$

and

$$\Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_{J(h)}$$

for every $h \in C(\overline{\Omega}; [-L, L]^n)$.

Proof. We split the proof into three steps.

Step 1: Well-definedness of the graph-update map.

We first show that the function g in (P) is unique.

Suppose

$$\gamma_{g_1} = \gamma_{g_2}$$

for some $g_1, g_2 \in C(\overline{\Omega}; [-L, L]^n)$. Consider the continuous function

$$\Phi : X \to \mathbb{R}, \qquad \Phi(x, y) \coloneqq |y - g_1(x)|^2_{\mathbb{R}^n}.$$

Since $\gamma_{g_1} = \gamma_{g_2}$, we have

$$\int_X \Phi \, d\gamma_{g_1} = \int_X \Phi \, d\gamma_{g_2}.$$

By the definition of graph measures,

$$\int_X \Phi \, d\gamma_{g_1} = \int_{\overline{\Omega}} |g_1(x) - g_1(x)|^2_{\mathbb{R}^n} \, d\overline{\lambda}_\Omega(x) = 0,$$

whereas

$$\int_X \Phi \, d\gamma_{g_2} = \int_{\overline{\Omega}} |g_2(x) - g_1(x)|^2_{\mathbb{R}^n} \, d\overline{\lambda}_\Omega(x).$$

Therefore

$$\int_{\overline{\Omega}} |g_2(x) - g_1(x)|^2_{\mathbb{R}^n} \, d\overline{\lambda}_\Omega(x) = 0.$$

Hence $g_1 = g_2$ $\lambda_{\overline{\Omega}}$-almost everywhere.

Since $g_1 - g_2$ is continuous and $\mathrm{supp}(\lambda_{\overline{\Omega}}) = \overline{\Omega}$, it follows that

$$g_1 = g_2 \qquad \text{on } \overline{\Omega}.$$

Thus the graph update in (P) is uniquely determined by h, and we may define

$$J(h) \coloneqq g, \qquad \Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_{J(h)}.$$

Step 2: Scalar McShane extensions.

Fix $x \in \overline{\Omega}$ and $i \in \{1, \dots, n\}$. Define

$$\psi_{x,i} : \mathcal{G} \to [-L, L], \qquad \psi_{x,i}(\gamma_h) \coloneqq J(h)(x)_i.$$

This is well-defined because $h \mapsto \gamma_h$ is injective by Step 1 applied to $g_1 = h_1$ and $g_2 = h_2$.

By (L1), for all $h_1, h_2 \in C(\overline{\Omega}; [-L, L]^n)$,

$$\begin{aligned} |\psi_{x,i}(\gamma_{h_1}) - \psi_{x,i}(\gamma_{h_2})| &= |J(h_1)(x)_i - J(h_2)(x)_i| \\ &\leq |J(h_1)(x) - J(h_2)(x)|_{\mathbb{R}^n} \\ &\leq \|J(h_1) - J(h_2)\|_{C(\overline{\Omega};\mathbb{R}^n)} \\ &\leq L_0 W_1(\gamma_{h_1}, \gamma_{h_2}). \end{aligned}$$

Thus $\psi_{x,i}$ is L_0-Lipschitz on $(\mathcal{G}, W_1)$.

Define its McShane extension by

$$\widetilde{\psi}_{x,i}(\mu) \coloneqq \inf_{h \in C(\overline{\Omega};[-L,L]^n)} \big\{ J(h)(x)_i + L_0 W_1(\mu, \gamma_h) \big\}, \qquad \mu \in \mathcal{P}(X). \tag{55}$$

We claim that

$$\widetilde{\psi}_{x,i}(\gamma_h) = J(h)(x)_i \tag{56}$$

for every $h \in C(\overline{\Omega}; [-L, L]^n)$.

Indeed, from the choice h in the infimum in (55),

$$\widetilde{\psi}_{x,i}(\gamma_h) \leq J(h)(x)_i.$$

Conversely, for arbitrary $k \in C(\overline{\Omega}; [-L, L]^n)$, the L_0-Lipschitz property of $\psi_{x,i}$ gives

$$J(k)(x)_i \geq J(h)(x)_i - L_0 W_1(\gamma_h, \gamma_k).$$

Therefore

$$J(k)(x)_i + L_0 W_1(\gamma_h, \gamma_k) \geq J(h)(x)_i.$$

Taking the infimum over k gives

$$\widetilde{\psi}_{x,i}(\gamma_h) \geq J(h)(x)_i.$$

Hence (56) holds.

Next, for arbitrary $\mu_1, \mu_2 \in \mathcal{P}(X)$ and arbitrary $h \in C(\overline{\Omega}; [-L, L]^n)$, the triangle inequality for W_1 implies

$$W_1(\mu_1, \gamma_h) \leq W_1(\mu_1, \mu_2) + W_1(\mu_2, \gamma_h).$$

Thus

$$J(h)(x)_i + L_0 W_1(\mu_1, \gamma_h) \leq L_0 W_1(\mu_1, \mu_2) + J(h)(x)_i + L_0 W_1(\mu_2, \gamma_h).$$

Taking the infimum over h gives

$$\widetilde{\psi}_{x,i}(\mu_1) \leq \widetilde{\psi}_{x,i}(\mu_2) + L_0 W_1(\mu_1, \mu_2).$$

Exchanging μ_1 and μ_2, we obtain

$$|\widetilde{\psi}_{x,i}(\mu_1) - \widetilde{\psi}_{x,i}(\mu_2)| \leq L_0 W_1(\mu_1, \mu_2). \tag{57}$$

We now prove Lipschitz regularity in the x-variable. Fix $\mu \in \mathcal{P}(X)$, $x_1, x_2 \in \overline{\Omega}$, and $\varepsilon > 0$. By the definition of the infimum, choose $h_\varepsilon \in C(\overline{\Omega}; [-L, L]^n)$ such that

$$J(h_\varepsilon)(x_2)_i + L_0 W_1(\mu, \gamma_{h_\varepsilon}) \leq \widetilde{\psi}_{x_2,i}(\mu) + \varepsilon.$$

Using the same function h_ε as a competitor in the infimum defining $\widetilde{\psi}_{x_1,i}(\mu)$, we get

$$\widetilde{\psi}_{x_1,i}(\mu) \leq J(h_\varepsilon)(x_1)_i + L_0 W_1(\mu, \gamma_{h_\varepsilon}).$$

Hence

$$\begin{aligned} \widetilde{\psi}_{x_1,i}(\mu) - \widetilde{\psi}_{x_2,i}(\mu) &\leq J(h_\varepsilon)(x_1)_i - J(h_\varepsilon)(x_2)_i + \varepsilon \\ &\leq |J(h_\varepsilon)(x_1) - J(h_\varepsilon)(x_2)|_{\mathbb{R}^n} + \varepsilon \\ &\leq L_1 |x_1 - x_2| + \varepsilon, \end{aligned}$$

where we used ($\mathrm{L2}_{\mathrm{lin}}$). Exchanging x_1 and x_2, and then letting $\varepsilon \downarrow 0$, gives

$$|\widetilde{\psi}_{x_1,i}(\mu) - \widetilde{\psi}_{x_2,i}(\mu)| \le L_1|x_1 - x_2|. \tag{58}$$

Step 3: Construction of $\widehat{p}$ and $\widehat{\Gamma}$.

Let

$$T_L : \mathbb{R} \to [-L, L], \qquad T_L(r) := \max\{-L, \min\{L, r\}\}.$$

The map T_L is 1-Lipschitz. Define

$$\widehat{p}(\mu, x)_i := T_L\big(\widetilde{\psi}_{x,i}(\mu)\big), \qquad i = 1, \dots, n.$$

Then

$$\widehat{p} : \mathcal{P}(X) \times \overline{\Omega} \to [-L, L]^n.$$

By (56), since $J(h)(x)_i \in [-L, L]$, the clipping is inactive on graph measures:

$$\widehat{p}(\gamma_h, x)_i = T_L\big(J(h)(x)_i\big) = J(h)(x)_i.$$

Therefore

$$\widehat{p}(\gamma_h, x) = J(h)(x) \qquad \forall h,\ x. \tag{59}$$

Since T_L is 1-Lipschitz, estimates (57) and (58) imply

$$|\widehat{p}(\mu_1, x)_i - \widehat{p}(\mu_2, x)_i| \le L_0 W_1(\mu_1, \mu_2) \tag{60}$$

and

$$|\widehat{p}(\mu, x_1)_i - \widehat{p}(\mu, x_2)_i| \le L_1|x_1 - x_2| \tag{61}$$

for every coordinate i.

Summing the coordinate estimates in the Euclidean norm gives

$$|\widehat{p}(\mu_1, x) - \widehat{p}(\mu_2, x)|_{\mathbb{R}^n} \le \sqrt{n}\, L_0 W_1(\mu_1, \mu_2) \tag{62}$$

and

$$|\widehat{p}(\mu, x_1) - \widehat{p}(\mu, x_2)|_{\mathbb{R}^n} \le \sqrt{n}\, L_1|x_1 - x_2|. \tag{63}$$

Using the triangle inequality,

$$\begin{aligned} |\widehat{p}(\mu_1, x_1) - \widehat{p}(\mu_2, x_2)|_{\mathbb{R}^n} &\le |\widehat{p}(\mu_1, x_1) - \widehat{p}(\mu_2, x_1)|_{\mathbb{R}^n} + |\widehat{p}(\mu_2, x_1) - \widehat{p}(\mu_2, x_2)|_{\mathbb{R}^n} \\ &\le \sqrt{n}\, L_0 W_1(\mu_1, \mu_2) + \sqrt{n}\, L_1|x_1 - x_2|. \end{aligned}$$

This proves (54).

Now define

$$\widehat{\Gamma}(\mu, (x, y)) := (x, \widehat{p}(\mu, x)).$$

Because $\widehat{p}(\mu, x) \in [-L, L]^n$, we have $\widehat{\Gamma}(\mu, z) \in X$ for all $(\mu, z) \in \mathcal{P}(X) \times X$.

Let $z = (x, y)$ and $z' = (x', y')$. Then

$$\begin{aligned} |\widehat{\Gamma}(\mu, z) - \widehat{\Gamma}(\nu, z')|_{\mathbb{R}^{d+n}} &= \big(|x - x'|^2_{\mathbb{R}^d} + |\widehat{p}(\mu, x) - \widehat{p}(\nu, x')|^2_{\mathbb{R}^n}\big)^{1/2} \\ &\le |x - x'|_{\mathbb{R}^d} + |\widehat{p}(\mu, x) - \widehat{p}(\nu, x')|_{\mathbb{R}^n} \\ &\le \sqrt{n}\, L_0 W_1(\mu, \nu) + (1 + \sqrt{n}\, L_1)|x - x'|_{\mathbb{R}^d} \\ &\le \sqrt{n}\, L_0 W_1(\mu, \nu) + (1 + \sqrt{n}\, L_1)|z - z'|_{\mathbb{R}^{d+n}}. \end{aligned}$$

Thus $\widehat{\Gamma}$ is Lipschitz on $\mathcal{P}(X) \times X$ with respect to the product metric

$$d\big((\mu, z), (\nu, z')\big) = W_1(\mu, \nu) + |z - z'|_{\mathbb{R}^{d+n}}.$$

Finally, fix $h \in C(\overline{\Omega}; [-L, L]^n)$. By (59),

$$\widehat{\Gamma}(\gamma_h, (x, y)) = (x, J(h)(x))$$

for all $(x, y) \in X$. Equivalently,

$$\widehat{\Gamma}(\gamma_h, \cdot) = (\mathrm{id}_{\overline{\Omega}}, J(h)) \circ \pi_1,$$

where

$$\pi_1 : X \to \overline{\Omega}, \qquad \pi_1(x, y) = x.$$

Since

$$(\pi_1)_{\#}\gamma_h = \overline{\lambda}_{\Omega},$$

we get

$$\begin{aligned}
\widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h &= \big((\mathrm{id}_{\overline{\Omega}}, J(h)) \circ \pi_1\big)_{\#}\gamma_h \\
&= (\mathrm{id}_{\overline{\Omega}}, J(h))_{\#}(\pi_1)_{\#}\gamma_h \\
&= (\mathrm{id}_{\overline{\Omega}}, J(h))_{\#}\overline{\lambda}_{\Omega} \\
&= \gamma_{J(h)}.
\end{aligned}$$

By the definition of J from (P),

$$\Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_{J(h)}.$$

Therefore

$$\Gamma(\gamma_h, \cdot)_{\#}\gamma_h = \widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_{J(h)}.$$

The proof is complete. □

The above proves Lemma 1.

We include also the following result that essentially gives the converse to Lemma 1.

Lemma 12. *If $\widehat{p} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to [-L, L]^n$ is a continuous function and $\widehat{\Gamma}(\mu, (x, y))$ is defined by formulas (15) and (16), then $\widehat{\Gamma} : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$ is a continuous map and for all continuous functions $h : \overline{\Omega} \to [-L, L]^n$ the graph measure $\gamma_h \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ satisfies*

$$\widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h = \gamma_g, \quad \textit{where } g = \widehat{p}(\gamma_h, \cdot) \in C(\overline{\Omega}; [-L, L]^n) \tag{64}$$

Proof. Let us define

$$\widehat{\Gamma}(\gamma_h, (x, y)) = (x, \widehat{p}(\gamma_h, x)) \tag{65}$$

Then, as $\widehat{\Gamma}(\gamma_h, (x, y))$ is independent of the variable y we have $\widehat{p}(\gamma_h, \cdot)(x, y) = (\widehat{p}(\gamma_h, \cdot) \circ \pi_1)(x, y)$ and hence

$$\begin{aligned}
\widehat{\Gamma}(\gamma_h, \cdot)_{\#}\gamma_h &= (Id \times \widehat{p}(\gamma_h, \cdot))_{\#}\gamma_h && (66)\\
&= ((Id \times \widehat{p}(\gamma_h, \cdot) \circ \pi_1)_{\#}\gamma_h && (67)\\
&= ((Id \times \widehat{p}(\gamma_h, \cdot))_{\#}\Big((\pi_1)_{\#}\gamma_h\Big) && (68)\\
&= (Id \times \widehat{p}(\gamma_h, \cdot))_{\#}\overline{\lambda}_{\Omega} && (69)\\
&= \gamma_{\widehat{p}(\gamma_h, \cdot)}. && (70)
\end{aligned}$$

This proves the claim. □

Justification of the original Definition 3

Below, we need the following well-known result. This result makes it possible to define function-graph preserving map for graph measures γ_h and the approximate the measure γ_h by a finite sum of delta-distributions.

Lemma 13. *Let (X, d) be a metric space and let $p \geq 1$. Assume $F_j : X \to X$ converge to F in the L^∞ topology and that $Lip(F_j)$ are uniformly bounded, that is,*

$$\|F_j - F\|_\infty \to 0 \qquad \textit{and there is } C_0 > 0 \textit{ such that} \qquad \mathrm{Lip}(F_j) \leq C_0. \tag{71}$$

Assume also that $\mu_j \to \mu$ in W_p. Then the pushforward measures satisfy

$$W_p((F_j)_{\#}\mu_j,\ F_{\#}\mu) \longrightarrow 0.$$

Proof. The condition (71) implies the Lipschitz constants $\mathrm{Lip}(F_j)$ are uniformly bounded. Using the triangle inequality for W_p,

$$W_p((F_j)_{\#}\mu_j,\, F_{\#}\mu) \le W_p((F_j)_{\#}\mu_j,\, (F_j)_{\#}\mu) + W_p((F_j)_{\#}\mu,\, F_{\#}\mu)\,.$$

For the first term, since F_j is $\mathrm{Lip}(F_j)$-Lipschitz,

$$W_p((F_j)_{\#}\mu_j,\, (F_j)_{\#}\mu) \le \mathrm{Lip}(F_j)\, W_p(\mu_j, \mu). \tag{72}$$

Because $\mathrm{Lip}(F_j)$ is uniformly bounded and $W_p(\mu_j, \mu) \to 0$, the term (72) tends to 0.

For the second term, since $F_j \to F$ uniformly,

$$W_p((F_j)_{\#}\mu,\, F_{\#}\mu) \le \left(\int_X d(F_j(x), F(x))^p \, d\mu(x)\right)^{1/p} \le \|F_j - F\|_\infty.$$

Thus this term also converges to 0.

Combining the two estimates yields

$$W_p((F_j)_{\#}\mu_j,\, F_{\#}\mu) \to 0,$$

as desired. □

Corollary 1. *If the map $\Gamma : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$, $\Gamma : (\mu, (x, y)) \to \Gamma(\mu, (x, y))$ is uniformly continuous in μ and uniformly Lipschitz bounded in (x, y), that is,*

(i) if $\mu_j \to \mu$ in $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ then

$$\lim_{j\to\infty} \sup_{(x,y)\in\overline{\Omega}\times[-L,L]^n} |\Gamma(\mu_j, (x, y)) - \Gamma(\mu, (x, y))| = 0 \tag{73}$$

(ii) There in $L > 0$ such that for all $(x_1, y_1), (x_2, y_2) \in \overline{\Omega} \times [-L, L]^n$

$$\sup_{\mu\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)} |\Gamma(\mu, (x_1, y_1)) - \Gamma(\mu, (x_2, y_2))| \le L(|x_1 - x_2| + |y_1 - y_2|), \tag{74}$$

then the map $G : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ given by

$$G(\mu) = \Gamma(\mu, \cdot)_{\#}\mu, \tag{75}$$

is continuous.

Proof. Let $\mu_j \to \mu$ in $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$. Define $F_j := \Gamma(\mu_j, \cdot)$ and $F := \Gamma(\mu, \cdot)$, so that $G(\mu_j) = (F_j)_{\#}\mu_j$ and $G(\mu) = F_{\#}\mu$.

We verify the hypotheses of Lemma 13 (with $X = \overline{\Omega} \times [-L, L]^n$ and $p = 1$):

1. By assumption (i),
$$\|F_j - F\|_\infty = \sup_{(x,y)\in\overline{\Omega}\times[-L,L]^n} |\Gamma(\mu_j, (x, y)) - \Gamma(\mu, (x, y))| \to 0.$$
2. By assumption (ii), for every j and every $(x_1, y_1), (x_2, y_2) \in \overline{\Omega} \times [-L, L]^n$,
$|F_j(x_1, y_1) - F_j(x_2, y_2)| = |\Gamma(\mu_j, (x_1, y_1)) - \Gamma(\mu_j, (x_2, y_2))| \le L(|x_1 - x_2| + |y_1 - y_2|)$,
so $\mathrm{Lip}(F_j) \le L$ uniformly in j.
3. By hypothesis, $\mu_j \to \mu$ in W_1.

Therefore, by Lemma 13

$$W_1(G(\mu_j),\, G(\mu)) = W_1((F_j)_{\#}\mu_j,\, F_{\#}\mu) \to 0,$$

which proves that $G : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ is continuous. □

An auxiliary lemma on Wasserstein convergence

We need the following, quite well-known result. We recall its proof as it is essential below.

Lemma 14 (Uniform map approximation implies uniform W_1 approximation)**.** *Let (X, d_X) be a compact metric space and let (Y, d_Y) be a metric space. Let*

$$\Gamma_m, \Gamma : \mathcal{P}(X) \times X \to Y$$

be Borel maps such that, for every $\mu \in \mathcal{P}(X)$, the pushforward measures

$$G_m(\mu) := (\Gamma_m(\mu, \cdot))_{\#}\mu, \qquad G(\mu) := (\Gamma(\mu, \cdot))_{\#}\mu$$

belong to $\mathcal{P}_1(Y)$. Suppose that

$$\Delta_m := \sup_{\mu \in \mathcal{P}(X)} \sup_{z \in X} d_Y\big(\Gamma_m(\mu, z), \Gamma(\mu, z)\big) < \infty.$$

Then

$$\sup_{\mu \in \mathcal{P}(X)} W_1\big(G_m(\mu), G(\mu)\big) \le \Delta_m.$$

Proof. Fix $\mu \in \mathcal{P}(X)$. By the Kantorovich–Rubinstein duality,

$$W_1\big(G_m(\mu), G(\mu)\big) = \sup_{\mathrm{Lip}(\varphi) \le 1} \left| \int_Y \varphi(y)\, dG_m(\mu)(y) - \int_Y \varphi(y)\, dG(\mu)(y) \right|,$$

where the supremum is taken over all real-valued 1-Lipschitz functions $\varphi : Y \to \mathbb{R}$.

Using the definition of pushforward measure, for every such φ,

$$\begin{aligned}
&\left| \int_Y \varphi(y)\, dG_m(\mu)(y) - \int_Y \varphi(y)\, dG(\mu)(y) \right| \\
&\quad = \left| \int_X \Big[\varphi\big(\Gamma_m(\mu, z)\big) - \varphi\big(\Gamma(\mu, z)\big) \Big]\, d\mu(z) \right| \\
&\quad \le \int_X \left| \varphi\big(\Gamma_m(\mu, z)\big) - \varphi\big(\Gamma(\mu, z)\big) \right|\, d\mu(z) \\
&\quad \le \int_X d_Y\big(\Gamma_m(\mu, z), \Gamma(\mu, z)\big)\, d\mu(z) \\
&\quad \le \Delta_m.
\end{aligned}$$

Taking the supremum over all 1-Lipschitz φ gives

$$W_1\big(G_m(\mu), G(\mu)\big) \le \Delta_m.$$

Since the bound is independent of μ, taking the supremum over $\mu \in \mathcal{P}(X)$ yields

$$\sup_{\mu \in \mathcal{P}(X)} W_1\big(G_m(\mu), G(\mu)\big) \le \Delta_m.$$

This proves the claim. □

E Generalized patching and tokenization

Let $D \subset \mathbb{R}^d$ be bounded and open, with $\overline{D} = \bigcup_{\rho=1}^{P} \overline{D}_\rho$, where $D_\rho \subset D$, $D_\rho = D^0 + q_\rho$ are open sets, such that $D_{\rho_1} \cap D_{\rho_2} = \emptyset$ for $\rho_1 \neq \rho_2$. Fix distinct points $x_1, \ldots, x_J \in D^0$.

E.1 Patching for continuous functions

For $f :\ \overline{D} \to \mathbb{R}^n$ be continuous, $f(x) = (f_i(x))_{i=1}^n$, let $\hat{\mathbf{E}}(f) :\ D \to \mathbb{R}^C$ be

$$\hat{\mathbf{E}}(f)(x) = (\hat{\mathbf{E}}_c(f)(x))_{c=1}^C = \left(\sum_{\rho=1}^{P} \sum_{i=1}^{n} \mathbf{F}_c^i \left(\int_{D_\rho} f_i(z)\, dW_{c,\rho}(z) \right) \mathbb{I}_{D_\rho}(x) \right)_{c=1}^{C}, \tag{76}$$

where $\mathbf{F} = (\mathbf{F}_c^i)_{c\in[C],i\in[n]} \in \mathbb{R}^{C\times n}$ and

$$W_{c,\rho} = \sum_{j=1}^{J} W_{c,j}\delta_{x_j+q_\rho}.$$

For the componentwise recovery below, retain the nC separate components

$$\hat{\mathbf{E}}_c^{(i)}(f)(x) = \sum_{\rho=1}^{P}\sum_{j=1}^{J} \mathbf{F}_c^i W_{c,j} f_i(x_j+q_\rho)\mathbb{I}_{D_\rho}(x), \qquad \hat{\mathbf{E}}^{\mathrm{sep}}(f) = (\hat{\mathbf{E}}_c^{(i)}(f))_{c,i}.$$

The mixed encoder above sums these components over i. Recovering all samples from that sum requires injectivity of the full $C \times nJ$ matrix $(\mathbf{H}^{(1)} \ \cdots \ \mathbf{H}^{(n)})$, not just its individual blocks.

E.2 Patching for generalized functions

Let us next consider $f = (f_i)_{i=1}^n \in H^{-s}(D;\mathbb{R}^n)$ and replace the delta-distributions by smooth test functions. Choose $\tau > 0$ small enough that each $\rho_\tau(\cdot - (x_j+q_\rho))$ is supported inside D_ρ, and define $\hat{\mathbf{E}}_\tau(f):\ D \to \mathbb{R}^C$ to be

$$\hat{\mathbf{E}}_\tau(f)(x) = (\hat{\mathbf{E}}_{\tau,c}(f)(x))_{c=1}^C = \left(\sum_{\rho=1}^{P}\sum_{i=1}^{n} \mathbf{F}_c^i \langle f_i, W_{c,\rho,\tau}\rangle\, \mathbb{I}_{D_\rho}(x)\right)_{c=1}^{C}, \tag{77}$$

where

$$W_{c,\rho,\tau}(z) = \sum_{j=1}^{J} W_{c,j}\rho_\tau(z-(x_j+q_\rho))$$

and $\rho_\tau \in C_0^\infty(B(0,\tau))$ is a mollifier. The brackets denote distributional pairing. Retaining the component contributions separately gives

$$\begin{aligned}
\hat{\mathbf{E}}_\tau^{\mathrm{sep}}(f)(x) &= (\hat{\mathbf{E}}_{\tau,c}^{(i)}(f)(x))_{c=1,\dots,C,i=1,\dots,n} && (78)\\
&= \sum_{\rho=1}^{P}\left(\sum_{j=1}^{J} \mathbf{F}_c^i W_{c,j}\langle f_i, \rho_\tau(\cdot-(x_j+q_\rho))\rangle\right)_{c,i} \mathbb{I}_{D_\rho}(x) && (79)\\
&= \sum_{\rho=1}^{P}\left(\sum_{j=1}^{J} \mathbf{H}_{c,j}^{(i)}\langle f_i, \rho_\tau(\cdot-(x_j+q_\rho))\rangle\right)_{c,i} \mathbb{I}_{D_\rho}(x), && (80)
\end{aligned}$$

where $\mathbf{H}_{c,j}^{(i)} = \mathbf{F}_c^i W_{c,j}$.

E.3 From a patched function to a measure

Let us next consider $f \in C(\overline{D};\mathbb{R}^n)$ and the separate components $\hat{\mathbf{E}}^{\mathrm{sep}}(f)$. Let $\mathbf{H}^{(i)} = [\mathbf{H}_{c,j}^{(i)}]_{c\in[C],j\in[J]} \in \mathbb{R}^{C\times J}$ and assume that for all $i = 1,...,n$ the linear map $\mathbf{H}^{(i)}:\ \mathbb{R}^J \to \mathbb{R}^C$ is one-to-one, that is, the matrix $\mathbf{H}^{(i)}$ has a left inverse matrix $\mathbf{K}^{(i)} = [\mathbf{K}_{j,c}^{(i)}]_{j\in[J],c\in[C]} \in \mathbb{R}^{J\times C}$, that is, for all i we have

$$\sum_{c=1}^{C} \mathbf{K}_{j,c}^{(i)}\mathbf{H}_{c,j'}^{(i)} = \delta_{j,j'}. \tag{81}$$

Then, for $x \in D_\rho$,

$$\sum_{c=1}^{C} \mathbf{K}_{j,c}^{(i)}\hat{\mathbf{E}}_c^{(i)}(f)(x) = f^{(i)}(x_j+q_\rho). \tag{82}$$

Choose $z_\rho = x_1 + q_\rho \in D_\rho$. Let $\mathbf{N}_\rho : \ (C(D_\rho; \mathbb{R}^n))^C \to \mathcal{P}(D_\rho \times \mathbb{R}^n)$ be the operator

$$\mathbf{N}_\rho(g) = \frac{1}{J} \sum_{j=1}^{J} \delta_{(x_j + q_\rho, (\sum_{c=1}^{C} \mathbf{K}_{j,c}^{(i)} g_c^{(i)}(z_\rho))_{i=1}^n)} \tag{83}$$

and $\hat{\mathbf{N}} : \ C_{pc}(D)^{nC} \to \mathcal{P}(D \times \mathbb{R}^n)$ be the operator

$$\hat{\mathbf{N}}(g) \ = \ \frac{1}{P} \sum_{\rho=1}^{P} \mathbf{N}_\rho(g) \tag{84}$$

$$= \ \frac{1}{PJ} \sum_{\rho=1}^{P} \sum_{j=1}^{J} \delta_{(x_j + q_\rho, (\sum_{c=1}^{C} \mathbf{K}_{j,c}^{(i)} g_c^{(i)}(z_\rho))_{i=1}^n)} \tag{85}$$

where $C_{pc}(D)^{nC} = \bigoplus_{\rho=1}^{P} (C(D_\rho; \mathbb{R}^n))^C$ is the space of piecewise continuous functions that are continuous in the sets $D_\rho \subset D$. Then,

$$\hat{\mathbf{N}}(\hat{\mathbf{E}}^{\text{sep}}(f)) = \frac{1}{PJ} \sum_{\rho=1}^{P} \sum_{j=1}^{J} \delta_{(x_j + q_\rho, f(x_j + q_\rho))}. \tag{86}$$

For $\hat{\mathbf{E}}_\tau^{\text{sep}}(f)$ the same calculation recovers the measurements $\langle f_i, \rho_\tau(\cdot - (x_j + q_\rho)) \rangle$ in place of point values.

The inverse of patching

Next, we consider the map from a measure approximating a function to a measure approximating a patched function and the inverse operation. We consider the computational grids

$$G_{large} = \{x_j + q_\rho : \ j = 1, 2, \ldots, J, \ \rho = 1, 2, \ldots, P\}, \tag{87}$$

$$G_{small} = \{q_\rho : \ \rho = 1, 2, \ldots, P\}. \tag{88}$$

We also consider the operator $\hat{J}$ from $\mathcal{P}(G_{large} \times \mathbb{R}^n)$ to tuples $(\nu_j)_{j=1}^J$ of positive measures on $G_{small} \times \mathbb{R}^n$ whose masses sum to one,

$$\hat{J} : \sum_{\rho=1}^{P} \sum_{j=1}^{J} a_{j\rho} \delta_{(x_j + q_\rho, y_{j,\rho})} \to \left(\sum_{\rho=1}^{P} a_{j\rho} \delta_{(q_\rho, y_{j,\rho})} \right)_{j=1}^{J} \tag{89}$$

and its inverse operator

$$\hat{J}^{-1} : \left(\sum_{\rho=1}^{P} a_{j\rho} \delta_{(q_\rho, y_{j,\rho})} \right)_{j=1}^{J} \to \sum_{\rho=1}^{P} \sum_{j=1}^{J} a_{j\rho} \delta_{(x_j + q_\rho, y_{j,\rho})} \tag{90}$$

The index j must be retained when expressing this inverse as a pushforward. Define $T : \{1, \ldots, J\} \times G_{small} \times \mathbb{R}^n \to G_{large} \times \mathbb{R}^n$ by

$$T(j, q_\rho, y) = (x_j + q_\rho, y). \tag{91}$$

The distinct sample points lie in disjoint patches, so T is bijective and $\hat{J}^{-1}((\nu_j)_j) = T_\#(\sum_j \delta_j \otimes \nu_j)$. For a general probability measure μ on $G_{large} \times \mathbb{R}^n$, define $\hat{J}\mu = (\nu_j)_j$ by $\sum_j \delta_j \otimes \nu_j = (T^{-1})_\# \mu$.

F Auxiliary results on approximation of function-graph preserving maps

To prove universal approximation results in the main text, we consider approximation of measure-theoretic transformers in general.

Proposition 5. *Let $\varepsilon > 0$, $N \in \mathbb{N}$, and $h \in C(\overline{\Omega}; [-L, L]^n)$. Suppose that the maps $G_n, G : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ satisfy*

$$\lim_{n \to \infty} W_1(G_n(\mu), \, G(\mu)) = 0 \quad \textit{for all } \mu \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n). \tag{92}$$

Then for each $h \in C(\overline{\Omega}; [-L, L]^n)$

$$\mathbf{F}_\varepsilon \circ G \circ \mathbf{M}_N(h) = \lim_{n\to\infty} \mathbf{F}_\varepsilon \circ G_n \circ \mathbf{M}_N(h) \in C(\overline{\Omega}; \mathbb{R}^n), \tag{93}$$

where the convergence holds in $C(\overline{\Omega}; \mathbb{R}^n)$ *(and hence also in* $H^{-s}(\Omega; \mathbb{R}^n)$ *for any* $s \geq 0$*). Similar statement is valid if* G_n *is replaced by* G_η*,* $\eta > 0$ *and the limit* $n \to \infty$ *is replaced by* $\eta \to 0$*.*

Proof. The proof proceeds in two steps: we first show that $\mathbf{F}_\varepsilon$ is continuous, and then compose with the pointwise convergence of G_n.

Step 1. We show that for fixed $\varepsilon > 0$, the map $\mathbf{F}_\varepsilon : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to C(\overline{\Omega}; \mathbb{R}^n)$ is continuous, where $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ is equipped with the W_1-topology.

Recall from (10) that, that for $\gamma \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$,

$$(\mathbf{F}_\varepsilon \gamma)_i(x) = \frac{\mathrm{Vol}(\Omega)}{\varepsilon^d} \int_{\overline{\Omega}\times[-L,L]^n} \rho\left(\frac{x' - x}{\varepsilon}\right) y_i \, d\gamma(x', y), \quad i = 1, \ldots, n.$$

For each $x \in \overline{\Omega}$ and $i \in \{1, \ldots, n\}$, define $\varphi_{x,i} : \overline{\Omega} \times [-L, L]^n \to \mathbb{R}$ by

$$\varphi_{x,i}(x', y) = \frac{\mathrm{Vol}(\Omega)}{\varepsilon^d} \rho\left(\frac{x' - x}{\varepsilon}\right) y_i.$$

Since $\rho \in C_0^\infty(\mathbb{R}^d)$ and $|y_i| \leq L$, the function $\varphi_{x,i}$ is continuous and bounded on $\overline{\Omega} \times [-L, L]^n$, uniformly in x:

$$\|\varphi_{x,i}\|_\infty \leq \frac{L\mathrm{Vol}(\Omega)}{\varepsilon^d} \|\rho\|_\infty =: C_\varepsilon \quad \text{for all } x \in \overline{\Omega}. \tag{94}$$

Now let $\gamma_m \to \gamma$ in W_1 on $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$. Since $\overline{\Omega} \times [-L, L]^n$ is compact, convergence in W_1 implies weak convergence of measures together with convergence of total mass (see, e.g., Villani, Villani [2009a], Theorem 6.9). In particular, for each fixed x and i,

$$(\mathbf{F}_\varepsilon \gamma_m)_i(x) = \int_{\overline{\Omega}\times[-L,L]^n} \varphi_{x,i} \, d\gamma_m \to \int_{\overline{\Omega}\times[-L,L]^n} \varphi_{x,i} \, d\gamma = (\mathbf{F}_\varepsilon \gamma)_i(x). \tag{95}$$

It remains to upgrade pointwise convergence (95) to uniform convergence in $x \in \overline{\Omega}$. Since ρ is smooth, the family of functions $\{x \mapsto \varphi_{x,i}(x', y)\}$ satisfies, for all $(x', y) \in \overline{\Omega} \times [-L, L]^n$,

$$|\varphi_{x_1,i}(x', y) - \varphi_{x_2,i}(x', y)| \leq \frac{L\mathrm{Vol}(\Omega)}{\varepsilon^{d+1}} \|\nabla\rho\|_\infty \, |x_1 - x_2|.$$

Hence, for any positive measure ν on $\overline{\Omega} \times [-L, L]^n$,

$$\begin{aligned} |(\mathbf{F}_\varepsilon \nu)_i(x_1) - (\mathbf{F}_\varepsilon \nu)_i(x_2)| &\leq \int |\varphi_{x_1,i} - \varphi_{x_2,i}| \, d\nu \\ &\leq \frac{L\mathrm{Vol}(\Omega)}{\varepsilon^{d+1}} \|\nabla\rho\|_\infty \, \nu(\overline{\Omega} \times [-L, L]^n) \, |x_1 - x_2|. \end{aligned} \tag{96}$$

Since $\gamma_m \to \gamma$ in W_1 on a compact space, the total masses $\gamma_m(\overline{\Omega} \times [-L, L]^n)$ converge to $\gamma(\overline{\Omega} \times [-L, L]^n)$ and are thus uniformly bounded by some $C_1 > 0$. By (96), the functions $\{x \mapsto (\mathbf{F}_\varepsilon \gamma_m)_i(x)\}_{m\geq 1}$ are equicontinuous and uniformly bounded (by $C_\varepsilon C_1$, using (94)). Combined with the pointwise convergence (95), the Arzelà–Ascoli theorem yields

$$\|\mathbf{F}_\varepsilon \gamma_m - \mathbf{F}_\varepsilon \gamma\|_{C(\overline{\Omega};\mathbb{R}^n)} \to 0.$$

This proves that $\mathbf{F}_\varepsilon$ is continuous.

Step 2. Let $\mu_0 := \mathbf{M}_N(h) \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$, which is fixed. By the pointwise convergence assumption (92),

$$G_n(\mu_0) \to G(\mu_0) \quad \text{in } W_1.$$

Applying Step 1,

$$\mathbf{F}_\varepsilon(G_n(\mu_0)) \to \mathbf{F}_\varepsilon(G(\mu_0)) \quad \text{in } C(\overline{\Omega}; \mathbb{R}^n),$$

which is precisely (93). □

Remark 5. *The proof of Proposition 5 when a limit* $n \to \infty$ *of the integer valued index* n *of* G_n *is changed to the limit* $\eta \to 0$ *of a continuous valued parameter* η *(when one consider* G_η *instead of* G_n*) is analogous.*

F.1 Proof of (19) and (20)

F.1.1 Recalling the setup

Let $\Omega \subset \mathbb{R}^d$ be a nonempty bounded open set, $D := \operatorname{diam}(\Omega) < \infty$, and $\bar{\lambda}_\Omega$ its normalized Lebesgue measure. Since Ω is nonempty and open, $\operatorname{Vol}(\Omega) > 0$. Define the normalised uniform probability measure

$$\bar{\lambda}_\Omega := \frac{1}{\operatorname{Vol}(\Omega)} \lambda_\Omega \in \mathcal{P}(\overline{\Omega}). \tag{97}$$

Definition 5 (Graph measure)**.** *For a Borel map $h : \Omega \to \mathbb{R}^n$ we define*

$$\gamma_h \in \mathcal{P}(\overline{\Omega} \times \mathbb{R}^n), \qquad \int_{\overline{\Omega}\times\mathbb{R}^n} \phi \, d\gamma_h = \frac{1}{Vol(\Omega)} \int_\Omega \phi(x, h(x)) \, dx \quad \forall \phi \in C_b(\overline{\Omega} \times \mathbb{R}^n). \tag{98}$$

Equivalently, if h is defined on $\overline{\Omega}$, then $\gamma_h = (\mathrm{id}, h)_\# \bar{\lambda}_\Omega$. If $h \in L^p(\Omega; \mathbb{R}^n)$, then γ_h has finite p-th moment on $\overline{\Omega} \times \mathbb{R}^n$. The definition is independent of the representative of h up to $\bar{\lambda}_\Omega$-a.e. equality.

The points $\{\bar{x}_j\}_{j\geq 1} \subset \Omega$ are *regularly distributed with density* $\kappa \equiv 1$ if the empirical measure $\sigma_N := \frac{1}{N} \sum_{j=1}^N \delta_{\bar{x}_j}$ satisfies

$$\frac{1}{N} \sum_{j=1}^N \varphi(\bar{x}_j) \xrightarrow[N\to\infty]{} \int_{\overline{\Omega}} \varphi \, d\bar{\lambda}_\Omega \qquad \forall \varphi \in C(\overline{\Omega}), \tag{99}$$

i.e. $\sigma_N \rightharpoonup \bar{\lambda}_\Omega$ narrowly. For $h \in C(\overline{\Omega}; \mathbb{R}^n)$ the empirical graph measure is

$$\mathbf{M}_N(h) = (\mathrm{id} \times h)_\# \sigma_N = \frac{1}{N} \sum_{j=1}^N \delta_{(\bar{x}_j, h(\bar{x}_j))} \in \mathcal{P}(\overline{\Omega} \times \mathbb{R}^n). \tag{100}$$

Let $\rho_\tau \in C_0^\infty(\mathbb{R}^d)$ be a standard mollifier with $\operatorname{supp} \rho_\tau \subset B(0, \tau)$ and $\int_{\mathbb{R}^d} \rho_\tau dx = 1$. Define

$$\Omega_\tau := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > 4\tau\}, \qquad \psi_\tau := \rho_\tau * \mathbf{1}_{\Omega_\tau}, \qquad R_\tau h := \rho_\tau * \mathcal{E}_0(\psi_\tau h). \tag{101}$$

Here $\mathcal{E}_0$ denotes the canonical extension by zero to $\mathbb{R}^d$ of the compactly supported function or distribution $\psi_\tau h$ on Ω. For vector-valued h the definition is componentwise. The empirical regularised graph measure is

$$\mathbf{M}_{N,\tau}(h) := \mathbf{M}_N(R_\tau h).$$

The enlarged margin 4τ in the definition of Ω_τ is used only to ensure $\operatorname{supp} R_\tau h \Subset \Omega$.

The Wasserstein distance W_p, $p \in [1, \infty)$, on $\mathcal{P}(\overline{\Omega} \times \mathbb{R}^n)$ uses the Euclidean cost; see [Villani, 2009b, Def. 6.1].

F.1.2 Proof of (19)

Proposition 6 ((19))**.** *Under (99), for every $h \in C(\overline{\Omega}; \mathbb{R}^n)$ and every $p \in [1, \infty)$,*

$$\lim_{N\to\infty} W_p(\mathbf{M}_N(h), \gamma_h) = 0.$$

Proof. Set $L := \|h\|_\infty < \infty$ and $K := \overline{\Omega} \times [-L, L]^n$, which is compact. Both $\mathbf{M}_N(h)$ and γ_h are supported in K. For $\phi \in C(K)$, the composition $\Phi(x) := \phi(x, h(x))$ is in $C(\overline{\Omega})$, so (99) gives

$$\int_K \phi \, d\mathbf{M}_N(h) = \tfrac{1}{N} \sum_{j=1}^N \Phi(\bar{x}_j) \to \tfrac{1}{\operatorname{Vol}(\Omega)} \int_\Omega \Phi = \int_K \phi \, d\gamma_h,$$

i.e. $\mathbf{M}_N(h) \rightharpoonup \gamma_h$ narrowly. On a compact Polish space, narrow convergence is equivalent to W_p-convergence for every $p \in [1, \infty)$ ([Villani, 2009b, Thm. 6.9]), so $W_p(\mathbf{M}_N(h), \gamma_h) \to 0$. □

F.1.3 Stability of graph measures and proof of (19)B

Lemma 15 (Graph measure stability)**.** *Let $p \in [1,\infty)$ and let $f, g : \Omega \to \mathbb{R}^n$ be Borel maps with $f, g \in L^p(\Omega;\mathbb{R}^n)$. Then*

$$W_p(\gamma_f,\gamma_g) \le \|f-g\|_{L^p(\bar{\lambda}_\Omega)} = \mathit{Vol}(\Omega)^{-1/p}\,\|f-g\|_{L^p(\Omega)}\,. \tag{102}$$

In particular, if $f, g \in C(\overline{\Omega};\mathbb{R}^n)$, then

$$W_p(\gamma_f,\gamma_g) \le \|f-g\|_{C(\overline{\Omega})}\,.$$

Proof. The Borel map

$$\Psi : \Omega \to (\overline{\Omega}\times\mathbb{R}^n)^2, \qquad \Psi(x) = ((x,f(x)),(x,g(x))),$$

pushes $\bar{\lambda}_\Omega$ to a coupling of γ_f and γ_g. Hence

$$W_p(\gamma_f,\gamma_g)^p \le \int_\Omega \|(x,f(x))-(x,g(x))\|^p\; d\bar{\lambda}_\Omega(x) = \int_\Omega |f(x)-g(x)|^p\; d\bar{\lambda}_\Omega(x).$$

Taking p-th roots proves the claim. □

Lemma 16 (Smoothness and L^p approximation by R_τ)**.** *Fix $\tau > 0$ and $s \in \mathbb{Z}_+$. For every $h \in H^{-s}(\Omega;\mathbb{R}^n)$,*

$$R_\tau h \in C_c^\infty(\Omega;\mathbb{R}^n) \subset C(\overline{\Omega};\mathbb{R}^n).$$

Moreover, for every $p \in [1,\infty)$ and every $h \in L^p(\Omega;\mathbb{R}^n)$,

$$\|R_\tau h - h\|_{L^p(\Omega)} \xrightarrow[\tau\downarrow 0]{} 0. \tag{103}$$

Proof. It is enough to argue componentwise. By the definition of Ω_τ and the support condition $\operatorname{supp}(\rho_\tau) \subset B(0,\tau)$,

$$\operatorname{supp}(\psi_\tau) = \operatorname{supp}(\rho_\tau * \mathbf{1}_{\Omega_\tau}) \subset \{x\in\Omega : \operatorname{dist}(x,\partial\Omega) \ge 2\tau\} \Subset \Omega.$$

Hence $\psi_\tau h$ is a compactly supported distribution in Ω, and its extension $\mathcal{E}_0(\psi_\tau h)$ is a compactly supported distribution on $\mathbb{R}^d$. Convolution with $\rho_\tau \in C_0^\infty(\mathbb{R}^d)$ gives $R_\tau h \in C^\infty(\mathbb{R}^d)$. Its support satisfies

$$\operatorname{supp}(R_\tau h) \subset \operatorname{supp}(\psi_\tau h) + \operatorname{supp}\rho_\tau \subset \{x\in\Omega : \operatorname{dist}(x,\partial\Omega)\ge\tau\} \Subset \Omega,$$

so $R_\tau h \in C_c^\infty(\Omega)$.

Now let $h \in L^p(\Omega;\mathbb{R}^n)$ and let $\widetilde{h}$ denote its extension by zero to $\mathbb{R}^d$. Since $0 \le \psi_\tau \le 1$ and $\psi_\tau(x) \to 1$ for a.e. $x \in \Omega$, dominated convergence gives

$$\left\|\mathcal{E}_0(\psi_\tau h) - \widetilde{h}\right\|_{L^p(\mathbb{R}^d)} = \|(\psi_\tau - 1)h\|_{L^p(\Omega)} \to 0.$$

Therefore,

$$\begin{aligned}\left\|R_\tau h - \widetilde{h}\right\|_{L^p(\mathbb{R}^d)} &\le \left\|\rho_\tau * \left(\mathcal{E}_0(\psi_\tau h) - \widetilde{h}\right)\right\|_{L^p(\mathbb{R}^d)} + \left\|\rho_\tau * \widetilde{h} - \widetilde{h}\right\|_{L^p(\mathbb{R}^d)} \\ &\le \left\|\mathcal{E}_0(\psi_\tau h) - \widetilde{h}\right\|_{L^p(\mathbb{R}^d)} + \left\|\rho_\tau * \widetilde{h} - \widetilde{h}\right\|_{L^p(\mathbb{R}^d)} \to 0,\end{aligned}$$

where the first term uses Young's inequality and the second term is the standard approximation-of-the-identity theorem. Restricting to Ω proves (103). □

Remark 6. *The stronger assertion*

$$\|R_\tau h - h\|_{C(\overline{\Omega})} \to 0$$

is false for the operator (101) in general, because $R_\tau h$ is cut off near $\partial\Omega$. For example, if $\Omega = (0,1)$ and $h \equiv 1$, then $R_\tau h$ vanishes near the boundary while h does not. The correct convergence statement needed below is the L^p convergence (103).

Proposition 7 ((19)B)**.** *Under (99), for every $h \in C(\overline{\Omega};\mathbb{R}^n)$ and every $p \in [1,\infty)$,*

$$\lim_{\tau\downarrow 0}\lim_{N\to\infty} \mathbf{M}_{N,\tau}(h) = \gamma_h \quad \text{in } W_p.$$

Proof. By Lemma 16, $R_\tau h \in C_c^\infty(\Omega;\mathbb{R}^n) \subset C(\overline{\Omega};\mathbb{R}^n)$ for every $\tau > 0$. Applying Theorem 6 to $R_\tau h$ gives

$$\lim_{N\to\infty} \mathbf{M}_{N,\tau}(h) = \lim_{N\to\infty} \mathbf{M}_N(R_\tau h) = \gamma_{R_\tau h} \quad \text{in } W_p. \tag{104}$$

By Lemmas 15 and 16,

$$W_p(\gamma_{R_\tau h},\gamma_h) \le \|R_\tau h - h\|_{L^p(\bar{\lambda}_\Omega)} = \mathrm{Vol}(\Omega)^{-1/p}\,\|R_\tau h - h\|_{L^p(\Omega)} \underset{\tau\downarrow 0}{\longrightarrow} 0. \tag{105}$$

Combining (104) and (105) proves the iterated limit. □

F.1.4 Proof of (20)

Proposition 8 ((20))**.** *Under (99), for every $\tau > 0$, $s \in \mathbb{N}$, $h \in H^{-s}(\Omega;\mathbb{R}^n)$, and $p \in [1,\infty)$,*

$$\lim_{N\to\infty} W_p(\mathbf{M}_{N,\tau}(h),\gamma_{R_\tau h}) = 0,$$

where $\gamma_{R_\tau h} = (\mathrm{id}, R_\tau h)_\# \bar{\lambda}_\Omega$.

Proof. Lemma 16 gives $R_\tau h \in C(\overline{\Omega};\mathbb{R}^n)$. Apply Theorem 6 to $R_\tau h$: $\mathbf{M}_{N,\tau}(h) = \mathbf{M}_N(R_\tau h) \to \gamma_{R_\tau h}$ in W_p. □

F.2 Proof of Theorem 1

Let $\varepsilon > 0$. Later, we choose a sufficiently small parameter $\varepsilon_1 > 0$ that depends on ε.

Target map f has the form $\widehat{\Gamma}(\mu,(x,y)) = (x, p(\mu,(x,y)))$, where $p(\mu,(x,y)) = \widehat{p}(\mu,x)$ is independent of the y-varialbe (We note that μ is a measure in the $\Omega \times [-L,L]^n$, so $p(\mu,(x,y))$ depends on a measure defined using the (x,y) variables. To clarify the analysis we consider below $p(\mu,(x,y))$ as function of all variables $(\mu,(x,y))$.)

We first apply [Furuya et al., 2025b, Theorem 1] to the map $\widehat{p}: \mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n) \to [-L,L]^n$, then there exist a deep transformer $\hat{F}_{\xi_{D_{\mathrm{tr}}}} \diamond \hat{\Gamma}_{\theta_{D_{\mathrm{tr}}}} \diamond \ldots \diamond \hat{F}_{\xi_1} \diamond \hat{\Gamma}_{\theta_1}$ such that

$$\sup_{(\mu,(x,y))\in K\times(\overline{\Omega}\times[-L,L]^n)} |p(\mu,(x,y)) - \hat{F}_{\xi_{D_{\mathrm{tr}}}} \diamond \hat{\Gamma}_{\theta_{D_{\mathrm{tr}}}} \diamond \ldots \diamond \hat{F}_{\xi_1} \diamond \hat{\Gamma}_{\theta_1}(\mu,(x,y))| \le \varepsilon_1, \tag{106}$$

where $K = P(\overline{\Omega}\times[-L,L]^n)$ and $\hat{\Gamma}_{\theta_\ell}: \mathcal{P}(\mathbb{R}^{\hat{d}_\ell}) \times \mathbb{R}^{\hat{d}_\ell} \to \mathbb{R}^{\hat{d}_\ell}$ and $\hat{F}_{\xi_\ell}: \mathbb{R}^{\hat{d}_\ell} \to \mathbb{R}^{\hat{d}_{\ell+1}}$, where $\hat{d}_1 = d+n$, $\hat{d}_{D_{\mathrm{tr}}+1} = n$, and $\hat{d}_\ell = 4(d+n)$, for $\ell = 2,3,\ldots,D_{\mathrm{tr}}$ and $\hat{d}'_\ell = 4(d+n)$ for $\ell = 1,2,\ldots,D_{\mathrm{tr}}+1$. Here, the bound $4(d+n)$ for $\hat{d}_\ell$ and $\hat{d}'_\ell$ is the sum of the dimension $(d+n)$ of the domain $\overline{\Omega}\times[-L,L]^n$ of the map $\hat{\Gamma}_{\theta_\ell}$ and 3 times its co-domain $(d+n)$. We point out that the bound for the width, $4(d+n)$, is larger than in the corresponding results in [Furuya et al., 2025b, Theorem 1]. Let us now define the maps $\Gamma_{\theta_\ell}: \mathcal{P}(\mathbb{R}^d\times\mathbb{R}^{d_\ell}) \times \mathbb{R}^d \times \mathbb{R}^{d_\ell} \to \mathbb{R}^d \times \mathbb{R}^{d_\ell}$ and $F_{\xi_\ell}: \mathbb{R}^d \times \mathbb{R}^{d_\ell} \to \mathbb{R}^d \times \mathbb{R}^{d_{\ell+1}}$, where $d_1 = d+n$, $d_{D_{\mathrm{tr}}+1} = d+n$, and $d_\ell = 4(d+n)$, for $\ell = 2,3,\ldots,D_{\mathrm{tr}}$ and $d'_\ell = 4(d+n)$ for $\ell = 1,2,\ldots,D_{\mathrm{tr}}+1$, by setting

$$\Gamma_{\theta_1}\Big(\mu,(x,y)\Big) = \Big(x, \hat{\Gamma}_{\theta_1}(\mu,(x,y))\Big), \tag{107}$$

and

$$\Gamma_{\theta_\ell}\Big(\mu,(x,x',y)\Big) = \Big(x, \hat{\Gamma}_{\theta_\ell}(\mu,(x',y))\Big), \tag{108}$$

for $\ell = 2,3,\ldots,D_{\mathrm{tr}}$ and

$$\hat{F}_{\xi_\ell}\Big(x,x',y\Big) = \Big(x, \hat{F}_{\xi_\ell}(x',y)\Big) \tag{109}$$

for $\ell = 1, 2, \ldots, D_{\mathrm{tr}} - 1$ and finally, we define $F_{\xi_{D_{\mathrm{tr}}}}$ by scaling $\hat{F}_{\xi_{D_{\mathrm{tr}}}}$ by a factor $\frac{L}{L+\varepsilon_1}$ and capping it with the smooth function Ψ. This defines the function

$$F_{\xi_{D_{\mathrm{tr}}}}\Big(\mu, (x, x', y)\Big) = \Big(x, \Psi(\frac{L}{L+\varepsilon_1} \cdot \hat{F}_{\xi_{D_{\mathrm{tr}}}}(\mu, (x', y)))\Big). \tag{110}$$

Recall that the function $\widehat{p}(\mu, (x, y))$ obtains in values on the interval $[-L, L]^n$. Thus we see that when the approximation (106) is valid, we have $\hat{F}_{\xi_\ell}((\mu, (x', y))) \in [-(L+\varepsilon_1), (L+\varepsilon_1)]^n$. Then,

$$\frac{L}{L+\varepsilon_1}\hat{F}_{\xi_\ell}((\mu, (x', y))) \in [-L, L]^n,$$

and thus

$$\Psi(\frac{L}{L+\varepsilon_1}\hat{F}_{\xi_\ell}((\mu, (x', y))) = \frac{L}{L+\varepsilon_1}\hat{F}_{\xi_\ell}((\mu, (x', y))) \in [-L, L]^n$$

When ε_1 is small enough, the above construction implies that the putative context function

$$\begin{aligned}\Gamma^{\mathrm{put}}(\mu, (x, y)) &= \Big(x\,,\, \Psi(\frac{L}{L+\varepsilon_1} \cdot \hat{F}_{\xi_{D_{\mathrm{tr}}}} \diamond \hat{\Gamma}_{\theta_{D_{\mathrm{tr}}}} \diamond \ldots \diamond \hat{F}_{\xi_1} \diamond \hat{\Gamma}_{\theta_1}(\mu, (x, y)))\Big) \\ &= F_{\xi_{D_{\mathrm{tr}}}} \diamond \Gamma_{\theta_{D_{\mathrm{tr}}}} \diamond \ldots \diamond F_{\xi_1} \diamond \Gamma_{\theta_1}(\mu, (x, y))\end{aligned}$$

satisfies

$$\sup_{(\mu,(x,y))\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n)} \Big|\Gamma^{\mathrm{put}}(\mu, (x, y)) - \widehat{\Gamma}(\mu, (x, y))\Big| \leq \varepsilon. \tag{111}$$

Recall that for $y_0 \in [-L, L]^n$, $w_{y_0} : \Omega \times \mathbb{R}^n \to \Omega \times \mathbb{R}^n$ is the map $w_{y_0}(x, y) = (x, y_0)$ and $w^*_{y_0} : F \to F \circ w_{y_0}$ a warping operator. We define now

$$\Gamma^{\mathrm{tran}}(\mu, (\cdot, \cdot)) = w^*_{y_0}\Gamma^{\mathrm{put}}(\mu, (\cdot, \cdot)) = \Gamma^{\mathrm{put}}(\mu, w_{y_0}(\cdot, \cdot)), \tag{112}$$

that is,

$$\Gamma^{\mathrm{tran}}(\mu, (x, y)) = \Gamma^{\mathrm{put}}(\mu, w_{y_0}(x, y_0)) = \Gamma^{\mathrm{put}}(\mu, (x, y_0)) = (x\,,\, \Gamma^{\mathrm{put},(y)}(\mu, (x, y_0))), \tag{113}$$

We recall that the target function $\widehat{\Gamma}(\mu, (x, y)) = (x, \widehat{p}(\mu, x))$ is independent of the y-variable. Moreover, the x-coordinates of $\Gamma^{\mathrm{put}}(\mu, (x, y))$ and $\widehat{\Gamma}(\mu, (x, y))$ are both equal to x. Thus, formula (111) implies that

$$\sup_{(\mu,(x,y))\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n)} \Big|\Gamma^{\mathrm{put},(y)}(\mu, (x, y)) - \widehat{p}(\mu, x)\Big| \tag{114}$$

$$= \sup_{(\mu,(x,y))\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n)} \Big|\Gamma^{\mathrm{put}}(\mu, (x, y)) - \widehat{\Gamma}(\mu, (x, y_0))\Big| \leq \varepsilon \tag{115}$$

and as the supremum in (114) is taken on over all $y \in [-L, L]^n$ we trivially have for the individual value y_0 that

$$\sup_{(\mu,x)\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times\overline{\Omega})} \Big|\Gamma^{\mathrm{put},(y)}(\mu, (x, y_0)) - \widehat{p}(\mu, x)\Big| \leq \varepsilon. \tag{116}$$

By definition of $\Gamma^{\mathrm{tran}}(\mu, (x, y))$, this implies

$$\sup_{(\mu,(x,y))\in\mathcal{P}(\overline{\Omega}\times[-L,L]^n)\times(\overline{\Omega}\times[-L,L]^n)} \Big|\Gamma^{\mathrm{tran}}(\mu, (x, y)) - \widehat{\Gamma}(\mu, (x, y))\Big| \leq \varepsilon. \tag{117}$$

This and Lemma 14 imply (26) and (27).

□

Theorem 1 means that we can approximate operators mapping graph measures to graph measures by traditional transformers in the product space $\overline{\Omega} \times [-L, L]^n$. Later, we see that by composing the traditional transformers by $\mathbf{M}_N$ and $\mathbf{F}_\varepsilon$, we can use those to approximate operators A from functions to functions.

G Proofs in Section 5

G.1 Proof of Lemma 2

(i) Let us consider the map

$$G_\eta(\mu) = F_\eta(\mu, \cdot)_\# \mu, \tag{118}$$

where

$$F_\eta(\mu, (x, y)) = (x, A(\tilde{S}_\eta(\mu))(x)), \quad F_\eta(\mu, (\cdot, \cdot)) : \overline{\Omega} \times \mathbb{R} \to \overline{\Omega} \times \mathbb{R}. \tag{119}$$

We observe that $F_\eta(\mu, (x, y))$ is independent of the value of y.

By Lemma 19, the The map $\tilde{S}_\eta : \mathcal{P}(\Omega \times [-L, L]^n) \to L^2(\overline{\Omega}; [-L_1, L_1]^n)$ is continuous. This and the assumption (30) on the map A imply that $A \circ \tilde{S}_\eta : \mathcal{P}(\Omega \times [-L, L]^n) \to C^1(\overline{\Omega}; [-L, L]^n)$ is continuous. This implies that the map $\Gamma_\eta : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \times (\overline{\Omega} \times [-L, L]^n) \to \overline{\Omega} \times [-L, L]^n$ be the map, given in (31), is continuous. This proves the claim (i).

(ii) Next, we consider the function $\mu \to A(\tilde{S}_\eta(\mu))(\cdot)$ as a map $\mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to C^1(\overline{\Omega}) \subset \text{Lip}(\overline{\Omega})$.

When $\mu_j \to \mu$ as $j \to \infty$ in $\mathcal{P}(\overline{\Omega} \times [-L, L]^n)$, we see that $S_\eta(\mu_j) \to S_\eta(\mu)$ in $H^{2r}(\Omega)$. Hence, $A(\tilde{S}_\eta(\mu_j)) \to A(\tilde{S}_\eta(\mu))$ in $C^1(\overline{\Omega})$, and hence

$$\lim_{j \to \infty} F_\eta(\mu_j, (\cdot, \cdot)) = F_\eta(\mu, (\cdot, \cdot)) \quad \text{in } L^\infty(\overline{\Omega} \times [-L, L]^n) \cap \text{Lip}(\overline{\Omega} \times [-L, L]^n). \tag{120}$$

This and Lemma 13 with $F_j = F_\eta(\mu_j, \cdot)$ imply that

$$\lim_{j \to \infty} G_\eta(\mu_j) = \lim_{j \to \infty} F_\eta(\mu_j, \cdot)_\# \mu_j = G_\eta(\mu) \quad \text{in } \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \tag{121}$$

Hence, the map

$$G_\eta : \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \to \mathcal{P}(\overline{\Omega} \times [-L, L]^n) \tag{122}$$

is continuous. This proves the claim (ii). □

G.2 Proof of Lemma 3

We show below that the definition of the map Γ_η in (31) and the map G_η in (31) yield the claim.

To prove the claim, let us assume that $h \in C(\overline{\Omega})$. Recall that $A : C^k(\overline{\Omega}) \to C^1(\overline{\Omega})$ is a continuous map.

Assume that $h \in X$ and γ_h be the measure

$$\int_{\overline{\Omega} \times \mathbb{R}} \phi(x, y) \, \mathrm{d}\gamma_h(x, y) = \tfrac{1}{\text{Vol}(\Omega)} \int_{\overline{\Omega}} \phi(x, h(x)) \, \mathrm{d}\lambda_\Omega(x). \tag{123}$$

Observe that for $\phi(x, y) \in C_b(\overline{\Omega} \times \mathbb{R})$, we have by (123)

$$\begin{aligned}
\langle G_\eta(\gamma_h), \phi \rangle &= \langle \Gamma_\eta(\gamma_h, (\cdot, \cdot))_\# \gamma_h), \phi \rangle && (124)\\
&= \langle \gamma_h, \phi(\Gamma_\eta(\gamma_h, (\cdot, \cdot)))) \rangle && (125)\\
&= \int_{\overline{\Omega} \times \mathbb{R}} \phi(\Gamma_\eta(\gamma_h, (x, y))) \, d\gamma_h(x, y) && (126)\\
&= \int_{\overline{\Omega} \times \mathbb{R}} \phi(\Gamma_\eta(\gamma_h, (x, y))) \, \mathbf{1}(y) \, d\gamma_h(x, y) && (127)\\
&= \int_{\overline{\Omega} \times \mathbb{R}} \phi(x, A(\tilde{S}_\eta(\gamma_h))(x)) \, \mathbf{1}(y) \, d\gamma_h(x, y) && (128)\\
&= \tfrac{1}{\text{Vol}(\Omega)} \int_{\overline{\Omega}} \phi(x, A(\tilde{S}_\eta(\gamma_h))(x)) \, \mathrm{d}\lambda_\Omega(x) && (129)\\
&= \langle \gamma_{A(\tilde{S}_\eta(\gamma_h))}, \phi \rangle, && (130)
\end{aligned}$$

where $\mathbf{1}(y) = 1$ is the constant function having the value 1. Hence, $G_\eta(\gamma_h) = \gamma_{A(\tilde{S}_\eta(\gamma_h))}$.

□

G.3 Properties of $S_{\eta,K(\eta)}$

We recall that we use two parameters $\eta > 0$ and $K \in \mathbb{N}$. Also $k \in \mathbb{Z}_+$ and $r > \frac{d}{2} + k$ be an even integer, so that $H^r(\Omega) \subset C^k(\overline{\Omega})$ by Sobolev's embedding theorem [Adams and Fournier, 2003]. Let $A_D = (1 - \Delta)$, where Δ is the Laplacian with Dirichlet boundary condition, considered as an unbounded selfadjoint operator in $L^2(\Omega)^n$ with domain $\mathcal{D}(A_D) = H^2(\Omega;\mathbb{R}^n) \cap H_0^1(\Omega;\mathbb{R}^n)$. Let $\mathcal{L} = (I - \Delta)^r$ be an unbounded self-adjoint operator in $L^2(\Omega;\mathbb{R}^n)$.Let us denote $H_D^r(\Omega) := \mathcal{D}(A_D^{r/2})$. Let $\psi_1, \psi_2, \cdots \in H_D^r(\Omega)$ be a complete orthonormal basis in $L^2(\Omega)$ of scalar eigenfunctions satisfying $(I - \Delta)\psi_j = \lambda_j \psi_j$ and $\psi_j|_{\partial\Omega} = 0$. We use the same basis in each component of a vector-valued function, so $\int_\Omega \psi_j(x) h(x)\, dx \in \mathbb{R}^n$. We solve the minimization problem,

$$S_{\eta,K}(\gamma) = \operatorname*{arg\,min}_{\tilde{h} \in H_D^r(\Omega)} \left(\eta \mathcal{R}_r(\tilde{h}) + \sum_{k=1}^{K} \left| V_0 \int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x) y \, d\gamma(x,y) - \int_{\overline{\Omega}} \psi_k(x)\, \tilde{h}(x) dx \right|^2 \right), \quad (131)$$

where $V_0 = \mathrm{Vol}(\Omega)$ and we use the regularizer $\mathcal{R}_r(\tilde{h}) = \|A_D^{r/2}\tilde{h}\|^2_{L^2(\Omega;\mathbb{R}^n)} = \|\mathcal{L}^{1/2}\tilde{h}\|^2_{L^2(\Omega;\mathbb{R}^n)}$. Note that $\mathcal{R}_r(\tilde{h})$ is equivalent to $\|\tilde{h}\|^2_{H^r(\Omega;\mathbb{R}^n)}$ for $\tilde{h} \in H_D^r(\Omega)$ and as $H_0^s(\Omega) \subset H_D^s(\Omega)$, the dual spaces satisfy $H_D^s(\Omega)^* \subset H^{-s}(\Omega)$.

For the convergence statements below, we also use $S_{\eta,K}$ on probability measures on $\overline{\Omega} \times \mathbb{R}^n$ with $\int |y|\, d\gamma(x,y) < \infty$, defined by

$$S_{\eta,K}(\gamma) = V_0 \sum_{j=1}^{K} \frac{\psi_j}{1 + \eta\lambda_j^r} \int_{\overline{\Omega}\times\mathbb{R}^n} \psi_j(x) y \, d\gamma(x,y).$$

The boundedness of each ψ_j ensures that these moments exist, and the formula agrees with (131) on the bounded token space. For fixed $\tau > 0$, $R_\tau h$ is smooth and bounded, so its graph measure satisfies this moment condition for arbitrary $h \in H^{-s}$.

Lemma 17. *If $K(\eta) \to \infty$ as $\eta \to 0$, the regularizers $S_\eta = S_{\eta,K(\eta)}$ satisfy for all $h \in H^{-s}(\Omega;\mathbb{R}^n)$*

$$\lim_{\tau\to 0} \lim_{\eta\to 0} \|S_{\eta,K(\eta)}(\gamma_{R_\tau h}) - h\|_{H^{-s}(\Omega)} = 0. \quad (132)$$

We see below in (137) that in the case when $\gamma = \gamma_h$ we have

$$S_{\eta,K}(\gamma_h) = \operatorname*{arg\,min}_{\tilde{h} \in H_D^r(\Omega;\mathbb{R}^n)} \left(\eta \|(I - \Delta)^{r/2}\tilde{h}\|^2_{L^2(\Omega;\mathbb{R}^n)} + \sum_{k=1}^{K} \left| \int_{\overline{\Omega}} \psi_k(x)(h(x) - \tilde{h}(x))\, dx \right|^2 \right). \quad (133)$$

The spectral formula (136) shows that $S_{\eta,K}(\mu)$ depends linearly on the measure moments. For graph measures, these moments depend linearly on h, so $h \mapsto S_{\eta,K}(\gamma_h)$ is a linear operator and can be viewed as a linear neural operator.

Next, we show that the regularization method defines a continuous operator $S_\eta :\ \gamma \to S_\eta(\gamma)$, e.g. $\mathcal{P}(\overline{\Omega} \times [-L,L]^n) \to H_D^r(\Omega;\mathbb{R}^n)$. In particular, this is useful if we assume that $A : H_D^r(\Omega;\mathbb{R}^n) \subset C^k(\overline{\Omega};[-L,L]^n) \to C^1(\overline{\Omega};[-L,L]^n)$ is continuous. We need to study the limit when first $K \to \infty$ and then $\eta \to 0$.

Let

$$P_K h(x) = \sum_{k=1}^{K} \int_{\overline{\Omega}} \psi_k(x') h(x')\, dx'\, \psi_k(x)$$

be an orthogonal projection to K first eigenvalues. Then, denoting $V_0 = \mathrm{Vol}(\Omega)$,

$$\begin{aligned} S_{\eta,K}(\gamma) &= V_0(\eta(I - \Delta)^r + P_K)^{-1} \left(\sum_{k=1}^{K} \left[\int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x') y \, d\gamma(x',y) \right] \psi_k \right) \\ &= \sum_{k=1}^{K} V_0(\eta\lambda_k^r + 1)^{-1} \left[\int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x') y \, d\gamma(x',y) \right] \psi_k \\ &= \sum_{k=1}^{K} \Big\langle \gamma\ ,\ \psi_k(x) y \Big\rangle V_0(\eta\lambda_k^r + 1)^{-1} \psi_k. \end{aligned} \quad (134)$$

We denote

$$Q_K[\gamma](x) = V_0 \sum_{k=1}^{K} \left[\int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x') y \, d\gamma(x', y) \right] \psi_k(x) \in C^k(\overline{\Omega}; \mathbb{R}^n). \tag{135}$$

Then, we see that

$$S_{\eta,K}(\gamma) \quad = \quad (\eta(I-\Delta)^r + I)^{-1} Q_K[\gamma]. \tag{136}$$

When $\gamma = \gamma_h$, we have

$$\int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x') y \, d\gamma_h(x', y) = \frac{1}{\mathrm{vol}(\Omega)} \int_{\overline{\Omega}} \psi_k(x') h(x') \, d\lambda(x') \tag{137}$$

and thus we obtain

$$\begin{aligned} S_{\eta,K}(\gamma_h) \quad &= \quad \sum_{k=1}^{K} (\eta\lambda_k^r + 1)^{-1} \left[\int_{\overline{\Omega}} \psi_k(x') h(x') \, d\lambda(x') \right] \psi_k \\ &= \quad (\eta(I-\Delta)^r + I)^{-1} P_K h. \end{aligned} \tag{138}$$

Below, we also denote

$$S_{\eta,K} h = (\eta(I-\Delta)^r + I)^{-1} P_K h$$

G.3.1 Quantitative growth of $K(\eta)$

We define the spectral norm $\|\cdot\|_r$ on $\mathcal{D}(A_D^{r/2})$ by

$$\|h\|_r^2 := \sum_{j=1}^{\infty} \lambda_j^r \, |\langle h, \psi_j \rangle_{L^2}|^2.$$

We now choose K explicitly. Set

$$K(\eta) := \left\lceil \eta^{-d/(4r)} \right\rceil, \qquad 0 < \eta \le 1. \tag{139}$$

Then $K(\eta) \to \infty$ as $\eta \downarrow 0$, and

$$\eta \lambda_{K(\eta)}^r \longrightarrow 0 \qquad \text{as } \eta \downarrow 0. \tag{140}$$

Indeed, using (38) and $K(\eta) \le 2\eta^{-d/(4r)}$ for $0 < \eta \le 1$,

$$\eta \lambda_{K(\eta)}^r \le C_{\mathcal{L}}^r \eta K(\eta)^{2r/d} \le C_{\mathcal{L}}^r 2^{2r/d} \eta \left(\eta^{-d/(4r)} \right)^{2r/d} = C_{\mathcal{L}}^r 2^{2r/d} \eta^{1/2} \to 0.$$

This choice quantifies the spectral cutoff used in the regularization step; it does not by itself give a rate for the full operator approximation in Theorem 2.

Lemma 18. *Let $K(\eta)$ be given by (139). Then:*

1. *For every $h \in L^2(\Omega; \mathbb{R}^n)$,*
$$\lim_{\eta \downarrow 0} \|S_{\eta,K(\eta)} h - h\|_{L^2(\Omega;\mathbb{R}^n)} = 0.$$
2. *For every $h \in H_D^r(\Omega; \mathbb{R}^n)$,*
$$\lim_{\eta \downarrow 0} \|S_{\eta,K(\eta)} h - h\|_{H_D^r(\Omega;\mathbb{R}^n)} = 0.$$
3. *If $f \in L^2(\Omega; [-L, L]^n)$, then*
$$\lim_{\eta \downarrow 0} \|S_{\eta,K(\eta)}(\gamma_f) - f\|_{L^2(\Omega;\mathbb{R}^n)} = 0.$$
If in addition $f \in H_D^r(\Omega; \mathbb{R}^n)$, then
$$\lim_{\eta \downarrow 0} \|S_{\eta,K(\eta)}(\gamma_f) - f\|_{H_D^r(\Omega;\mathbb{R}^n)} = 0.$$

Proof. It is enough to prove the scalar-valued case; the vector-valued case follows by summing over the components.

Let

$$h = \sum_{j=1}^{\infty} a_j \psi_j, \qquad a_j := \langle h, \psi_j \rangle_{L^2(\Omega)}.$$

Then

$$S_{\eta,K} h - h = -\sum_{j=1}^{K} \frac{\eta \lambda_j^r}{1 + \eta \lambda_j^r} a_j \psi_j - \sum_{j>K} a_j \psi_j.$$

First assume $h \in L^2(\Omega)$. Since the two sums are orthogonal in $L^2(\Omega)$,

$$\|S_{\eta,K} h - h\|_{L^2(\Omega)}^2 = \sum_{j=1}^{K} \left(\frac{\eta \lambda_j^r}{1 + \eta \lambda_j^r} \right)^2 |a_j|^2 + \sum_{j>K} |a_j|^2.$$

Taking $K = K(\eta)$ and using monotonicity of λ_j^r,

$$\begin{aligned} \sum_{j=1}^{K(\eta)} \left(\frac{\eta \lambda_j^r}{1 + \eta \lambda_j^r} \right)^2 |a_j|^2 &\le \left(\eta \lambda_{K(\eta)}^r \right)^2 \sum_{j=1}^{K(\eta)} |a_j|^2 \\ &\le \left(\eta \lambda_{K(\eta)}^r \right)^2 \|h\|_{L^2(\Omega)}^2. \end{aligned}$$

By (140), this term tends to zero. Also, since $K(\eta) \to \infty$ and $\sum_{j=1}^{\infty} |a_j|^2 < \infty$,

$$\sum_{j>K(\eta)} |a_j|^2 \to 0.$$

Therefore

$$\|S_{\eta,K(\eta)} h - h\|_{L^2(\Omega)} \to 0.$$

Now assume $h \in H_D^r(\Omega) = \mathcal{D}(\mathcal{L}^{1/2})$. Then

$$\sum_{j=1}^{\infty} \lambda_j^r |a_j|^2 < \infty.$$

Again by orthogonality, now after applying $\mathcal{L}^{1/2}$,

$$\|S_{\eta,K} h - h\|_{H_D^r(\Omega)}^2 = \sum_{j=1}^{K} \left(\frac{\eta \lambda_j^r}{1 + \eta \lambda_j^r} \right)^2 \lambda_j^r |a_j|^2 + \sum_{j>K} \lambda_j^r |a_j|^2.$$

With $K = K(\eta)$,

$$\begin{aligned} \sum_{j=1}^{K(\eta)} \left(\frac{\eta \lambda_j^r}{1 + \eta \lambda_j^r} \right)^2 \lambda_j^r |a_j|^2 &\le \left(\eta \lambda_{K(\eta)}^r \right)^2 \sum_{j=1}^{K(\eta)} \lambda_j^r |a_j|^2 \\ &\le \left(\eta \lambda_{K(\eta)}^r \right)^2 \|h\|_{H_D^r(\Omega)}^2. \end{aligned}$$

This tends to zero by (140). The tail term also tends to zero because

$$\sum_{j=1}^{\infty} \lambda_j^r |a_j|^2 < \infty \qquad \text{and} \qquad K(\eta) \to \infty.$$

Hence

$$\|S_{\eta,K(\eta)} h - h\|_{H_D^r(\Omega)} \to 0.$$

Finally, if $f \in L^2(\Omega; [-L, L]^n)$, then

$$S_{\eta,K(\eta)}(\gamma_f) = S_{\eta,K(\eta)} f.$$

The preceding L^2 and H_D^r convergence results therefore apply directly to $S_{\eta,K(\eta)}(\gamma_f)$.

□

Corollary 2. *Let $s \geq 0$ with $s - \frac{1}{2} \notin \mathbb{Z}$. Suppose that $h \in H^{-s}(\Omega;\mathbb{R}^n)$. Then*

$$\lim_{\tau\downarrow 0}\lim_{\eta\downarrow 0} \|S_{\eta,K(\eta)}(\gamma_{R_\tau h}) - h\|_{H^{-s}(\Omega;\mathbb{R}^n)} = 0.$$

Proof. We have seen above that $R_\tau h \in L^2(\Omega;\mathbb{R}^n)$ for all sufficiently small $\tau > 0$ and that $R_\tau h \to h$ in $H^{-s}(\Omega;\mathbb{R}^n)$ as $\tau \downarrow 0$, by Proposition 3 when $s > 0$. For $s = 0$, the cutoff converges to one by dominated convergence in L^2, and convolution with ρ_τ is an L^2 contraction converging to the identity.

For fixed $\tau > 0$, Proposition 18 gives

$$S_{\eta,K(\eta)}(\gamma_{R_\tau h}) \to R_\tau h \qquad \text{in } L^2(\Omega;\mathbb{R}^n).$$

Since $s \geq 0$, the embedding

$$L^2(\Omega;\mathbb{R}^n) \hookrightarrow H^{-s}(\Omega;\mathbb{R}^n)$$

is continuous. Hence

$$S_{\eta,K(\eta)}(\gamma_{R_\tau h}) \to R_\tau h \qquad \text{in } H^{-s}(\Omega;\mathbb{R}^n).$$

Therefore

$$\lim_{\eta\downarrow 0} \|S_{\eta,K(\eta)}(\gamma_{R_\tau h}) - h\|_{H^{-s}} = \|R_\tau h - h\|_{H^{-s}}.$$

Letting $\tau \downarrow 0$ gives the claim. □

Below, we choose $K(\eta)$ to be as in (139) and we denote

$$S_\eta = S_{\eta,K(\eta)}. \tag{141}$$

G.3.2 Properties of the map S_η

We point out that when $h \in C(\overline{\Omega})$, we have

$$\begin{aligned}\int_{\overline{\Omega}\times[-L,L]^n} |\tilde{h}(x) - y|^2 \,\mathrm{d}\gamma_h(x,y) &= \tfrac{1}{\mathrm{Vol}(\Omega)} \int_{\overline{\Omega}} |\tilde{h}(x) - h(x)|^2 \,\mathrm{d}\lambda_\Omega(x) \\ &= \tfrac{1}{\mathrm{Vol}(\Omega)} \int_{\overline{\Omega}} (|h(x)|^2 - 2\tilde{h}(x)\cdot h(x) + |\tilde{h}(x)|^2) \,\mathrm{d}\lambda_\Omega(x). \end{aligned} \tag{142}$$

Lemma 19. *For a fixed $\eta > 0$, the map $S_\eta : \mathcal{P}(\Omega \times [-L,L]^n) \to H^{2r}(\Omega;\mathbb{R}^n)$ is continuous.*

Proof. By (136), we have

$$\begin{aligned} S_\eta(\gamma) = S_{\eta,K(\eta)}(\gamma) &= \mathrm{Vol}(\Omega) \sum_{k=1}^{K(\eta)} (\eta\lambda_k^r + I)^{-1} \left[\int_{\overline{\Omega}\times[-L,L]^n} \psi_k(x') y \, d\gamma(x',y) \right] \psi_k \\ &= (\eta(I-\Delta)^r + I)^{-1} Q_{K(\eta)}[\gamma]. \end{aligned} \tag{143}$$

Here, the map Q_K given in (135) is a continuous map $Q_K : \mathcal{P}(\overline{\Omega}\times[-L,L]^n) \to C^k(\overline{\Omega};\mathbb{R}^n) \subset L^2(\overline{\Omega};\mathbb{R}^n)$. As $(\eta(I-\Delta)^r + I)^{-1} : L^2(\overline{\Omega};\mathbb{R}^n) \to H^{2r}(\Omega;\mathbb{R}^n)$ is continuous, we obtain the claim. □

Lemma 20. *Let $h \in C(\overline{\Omega};[-L,L]^n)$. Then $\lim_{\eta\to 0} S_\eta(\gamma_h) = h$ in $L^2(\overline{\Omega};\mathbb{R}^n)$. Moreover, when $h \in H^r_D(\Omega) \cap C(\overline{\Omega};[-L,L]^n)$, we have $\lim_{\eta\to 0} S_\eta(\gamma_h) = h$ in $H^r_D(\Omega)$ and therefore in $C^k(\overline{\Omega};\mathbb{R}^n)$.*

Proof. By (137),

$$\begin{aligned} S_{\eta,K(\eta)}(\gamma_h) &= \sum_{k=1}^{K(\eta)} (\eta\lambda_k^r + I)^{-1} \left[\int_{\overline{\Omega}} \psi_k(x') h(x') \, d\lambda(x') \right] \psi_k \\ &= \sum_{k=1}^{K(\eta)} (\eta\lambda_k^r + 1)^{-1} \langle h, \psi_k\rangle_{L^2(\Omega)} \psi_k \end{aligned} \tag{144}$$

and by using Lebesgue dominated convergence theorem we see that $S_{\eta,K(\eta)}(\gamma_h) \to h$ in $L^2(\Omega)$ as $\eta \to 0$.

When $h \in H^r_D(\Omega)$, we use that the domain of the square root $\sqrt{\mathcal{L}}$ of the selfadjoint operator $\mathcal{L}$ is the space $H^r_D(\Omega)$, see Lions and Magenes [1972a]. We see by applying spectral theory, Lions and Magenes [1972a], that $\lim_{\eta\to 0} S_\eta(\gamma_h) = h$ in $H^r_D(\Omega)$ that is, in $H^r_D(\Omega)$. □

G.4 Proof of Theorem 2

Let $\tau > 0$, $\eta > 0$, and $h \in \mathrm{cl}_{H^{-s}(\Omega;\mathbb{R}^n)}(C(\overline{\Omega};[-L,L]^n))$

We see that $h_\tau = R_\tau(h) \in H^r_D(\Omega) \cap C(\overline{\Omega};[-L,L]^n)$. Thus, when $\tau > 0$ is small enough, we see that

$$S_\eta(\gamma_{h_\tau}) = (\eta(I-\Delta)^r + I)^{-1} P_{K(\eta)}(h_\tau) \to h_\tau \tag{145}$$

in $H^r_D(\Omega) \subset C^k(\overline{\Omega};\mathbb{R}^n)$ as $\eta \to 0$ and thus

$$\lim_{\eta\to 0} \|S_\eta(\gamma_{h_\tau})\|_{L^\infty(\Omega;\mathbb{R}^n)} = \|h_\tau\|_{L^\infty(\Omega;\mathbb{R}^n)} \tag{146}$$

Hence, when $\eta > 0$ is small enough, we have

$$S_\eta(\gamma_{h_\tau}) \in C^k(\overline{\Omega};[-(L+L_1)/2,(L+L_1)/2]^n). \tag{147}$$

Thus, as $\psi(y_i) = y_i$ for $y \in [-(L+L_1)/2,(L+L_1)/2]$, it holds for sufficiently small $\eta > 0$ that

$$\tilde{S}_\eta(\gamma_{h_\tau}) = \Psi \circ (\eta(I-\Delta)^r + I)^{-1} P_{K(\eta)}(h_\tau) = S_\eta(\gamma_{h_\tau}). \tag{148}$$

We can approximate G_{Γ_η} by softmax-based traditional transformers $G_{\Gamma^{\mathrm{tran}}_{\eta,m}}$ that are given by Theorem 1. The limit $m \to \infty$ can be included in the theorem by the formula (26). This gives us

$$\mathbf{F}_\varepsilon \circ G_\eta \circ \mathbf{M}_{N,\tau}(h) = \lim_{m\to\infty} \mathbf{F}_\varepsilon \circ G_{\Gamma^{\mathrm{tran}}_{\eta,m}} \circ \mathbf{M}_{N,\tau}(h) \tag{149}$$

in $C(\overline{\Omega};\mathbb{R}^n)$.

For fixed $\eta > 0$

$$A_{\eta,\tau}(h) := \lim_{\varepsilon\to 0} \lim_{N\to\infty} \mathbf{F}_\varepsilon \circ G_{\Gamma_\eta} \circ \mathbf{M}_{N,\tau}(h) \tag{150}$$

in $H^{-s'}(\Omega;\mathbb{R}^n)$ where the limit $\lim_{N\to\infty} \mathbf{M}_{N,\tau}(h) = \gamma_{h_\tau}$ converges in $\mathcal{P}(\overline{\Omega} \times [-L,L]^n)$.

Then we see that the limit $\lim_{N\to\infty} S_\eta(\mathbf{M}_{N,\tau}(h))$ converges in $H^r_D(\Omega)$. The definition of the map G_{Γ_η} in (31) and Lemma 13 imply that $\lim_{N\to\infty} G_{\Gamma_\eta} \circ \mathbf{M}_{N,\tau}(h)$ converges to γ_b in $\mathcal{P}(\overline{\Omega} \times [-L,L]^n)$ when $b = A(\tilde{S}_\eta(\gamma_{h_\tau}))$. Then, by (11) to

$$\lim_{\varepsilon\to 0} \left(\lim_{N\to\infty} \mathbf{F}_\varepsilon \circ G_{\Gamma_\eta} \circ \mathbf{M}_{N,\tau}(h) \right) = A(\tilde{S}_\eta(\gamma_{h_\tau})), \tag{151}$$

where the limit takes place in $H^{-s'}(\Omega)$.

Then,

$$\begin{aligned} A(\tilde{S}_\eta(\gamma_{h_\tau})) &= A(S_\eta(\gamma_{h_\tau})) && (152)\\ &= A(\Psi \circ (\eta(I-\Delta)^r + I)^{-1} P_{K(\eta)}(h_\tau)). && (153) \end{aligned}$$

Moreover,

$$A_\eta(h_\tau) = A((\eta(I-\Delta)^r + I)^{-1} P_{K(\eta)} h_\tau).$$

As

$$\lim_{\eta\to 0} (\eta(I-\Delta)^r + I)^{-1} P_{K(\eta)} h_\tau = h_\tau$$

converges in $H^r_D(\Omega) \subset C^k(\overline{\Omega})$, this implies that

$$\lim_{\eta\to 0} A(\Psi \circ (\eta(I-\Delta)^r + 1)^{-1} P_{K(\eta)} h_\tau)) = A(h_\tau)$$

converges in $H^{-s'}(\Omega)$. This implies that

$$\lim_{\tau\to 0} A(h_\tau) = A(h)$$

converges in $H^{-s'}(\Omega)$.

The proof in the case when $h \in C^r_c(\Omega;\mathbb{R}^n) \cap C(\overline{\Omega};[-L,L]^n)$ follows similarly as the above arguments for h_τ. □

G.5 Proof of Proposition 1

Recall that we suppose

$$A : L^2(\Omega; [-L, L]^n) \to C^1(\overline{D}; [-L, L]^n). \tag{154}$$

is continuous, possibly nonlinear, function. Here, $\Omega \subset \mathbb{R}^d$ and $D \subset \mathbb{R}^{d'}$ are open bounded sets with C^∞-smooth boundary, or a cube. Below, we use in $\overline{\Omega} \times [-L, L]^n$ the coordinates (x, y) and in $\overline{D} \times [-L, L]^n$ the coordinates (x', y'). On $X = (\overline{\Omega} \times [-L, L]^n) \times (\overline{D} \times [-L, L]^n)$ we consider the maps $\rho_1((x, y), (x', y')) = (x, y)$ and $\rho_2((x, y), (x', y')) = (x', y')$, and denote $\pi_2(x', y') = y'$.

Our next goal is to express the map $\widehat{G}_\eta(\mu, x') = A(\tilde{S}_\eta(\mu))(x')$ using the same self-attention construction as in the measure-to-measure setting.

Let $\mu \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ and $x' \in D$, and define the product measure

$$\tilde{\mu} := \mu \otimes \delta_{(x',0)} \in \mathcal{P}\big((\overline{\Omega} \times [-L, L]^n) \times (D \times [-L, L]^n)\big). \tag{155}$$

Define the lifted map $\tilde{\Gamma}_\eta : \mathcal{P}(X) \times X \to X$,

$$\tilde{\Gamma}_\eta(\tilde{\mu}, ((x, y), (x', y'))) := \big((x, y),\, \Gamma_\eta((\rho_1)_\# \tilde{\mu}, (x', y'))\big), \tag{156}$$

where $\Gamma_\eta : \mathcal{P}\big(\overline{\Omega} \times [-L, L]^n\big) \times (D \times [-L, L]^n) \to (D \times [-L, L]^n)$ is the mapping

$$\Gamma_\eta(\mu, (x', y')) = \big(x', A(\tilde{S}_\eta(\mu))(x')\big). \tag{157}$$

We emphasize that here the context function $\Gamma_\eta(\mu, (x', y'))$ is independent of the y' variable and the output-value component of $\tilde{\Gamma}_\eta(\tilde{\mu}, ((x, y), (x', y')))$ is independent of the y and y' variables. However, $\Gamma_\eta(\mu, (x', y'))$ is not a function-graph preserving map, but an auxiliary map just used in this proof.

Let $\tilde{G}_\eta : \mathcal{P}(X) \to \mathcal{P}(X)$ be the measure map with the context function $\tilde{\Gamma}_\eta$, that is,[2]

$$\tilde{G}_\eta(\tilde{\mu}) := \tilde{\Gamma}_\eta(\tilde{\mu}, \cdot)_\# \tilde{\mu} \in \mathcal{P}\big(X\big). \tag{158}$$

Then

$$\begin{aligned}
(\rho_2)_\# \tilde{G}_\eta(\tilde{\mu}) &= (\rho_2)_\# \Big(\tilde{\Gamma}_\eta(\tilde{\mu}, ((\cdot, \cdot), (\cdot, \cdot)))_\# \tilde{\mu}\Big) \\
&= \Big(\rho_2 \circ \tilde{\Gamma}_\eta(\tilde{\mu}, ((\cdot, \cdot), (\cdot, \cdot)))\Big)_\# \tilde{\mu} \\
&= \Big(\Gamma_\eta((\rho_1)_\# \tilde{\mu}, \rho_2((\cdot, \cdot), (\cdot, \cdot)))\Big)_\# \tilde{\mu} \\
&= \Big(\Gamma_\eta((\rho_1)_\# \tilde{\mu}, (\cdot, \cdot))\Big)_\# ((\rho_2)_\# \tilde{\mu}) \\
&= \Gamma_\eta(\mu, \cdot)_\# \delta_{(x',0)} \in \mathcal{P}\big(\overline{D} \times [-L, L]^n\big).
\end{aligned} \tag{159}$$

Since $\Gamma_\eta(\mu, (x', y'))$ is independent of y', we obtain

$$\Gamma_\eta(\mu, \cdot)_\# \delta_{(x',0)} = \delta_{(x', A(\tilde{S}_\eta(\mu))(x'))}, \tag{160}$$

and hence

$$\widehat{G}_\eta(\mu, x') = \langle \Gamma_\eta(\mu, \cdot)_\# \delta_{(x',0)}, \pi_2 \rangle = A(\tilde{S}_\eta(\mu))(x'). \tag{161}$$

□

[2]We emphasize that here we use only self-attention with a context map $\tilde{\Gamma}_\eta$ in a product space, that is, we have raised the dimension of sets where the measures are considered.

Further observations. The map $(\mu, x') \mapsto \widehat{G}_\eta(\mu, x') = A(\tilde{S}_\eta(\mu))(x')$ admits the factorization

$$(\mu, x') \;\mapsto\; \mu \otimes \delta_{(x',0)} \;\mapsto^{\tilde{G}_\eta}\; \tilde{G}_\eta(\tilde{\mu}) \;\mapsto^{(\rho_2)_\#}\; (\rho_2)_\# \tilde{G}_\eta(\tilde{\mu}) \;\mapsto^{\langle \cdot, \pi_2 \rangle}\; A(\tilde{S}_\eta(\mu))(x'), \tag{162}$$

i.e., as a composition of push forward of projections and a self-attention map. If we use $\mu = \gamma_h$, where $h \in C(\overline{\Omega})$, and $x' \in \overline{D}$, and use for $\eta = 0$ the notation $S_0(\gamma_h) = \lim_{\eta \to 0} S_\eta(\gamma_h)) = h$, we have

$$\begin{aligned}(h, x') \mapsto^{\mathbf{M} \times Id} (\gamma_h, x') \;\mapsto\; \gamma_h \otimes \delta_{(x',0)} \;\mapsto^{\tilde{g}_0}\; \tilde{g}_0(\gamma_h \otimes \delta_{(x',0)}) \;\mapsto^{(\rho_2)_\#} \\ (\rho_2)_\# \tilde{g}_0(\gamma_h \otimes \delta_{(x',0)}) \;\mapsto^{\langle \cdot, \pi_2 \rangle}\; A(h)(x').\end{aligned} \tag{163}$$

That is,

$$A(h)(x') = \langle (\rho_2)_\# \tilde{g}_0(\gamma_h \otimes \delta_{(x',0)}), \pi_2 \rangle$$

or equivalently,

$$A(h)(x') = \langle \tilde{g}_0(\gamma_h \otimes \delta_{(x',0)}) \,,\, (\pi_2 \circ \rho_2) \rangle.$$

or,

$$A(h)(x') = \langle \tilde{g}_0(\gamma_h \otimes \delta_{(x',0)}) \,,\, y' \rangle.$$

Let us also do a computation analogous to (159) for the product measure $\mu \otimes \nu$, where $\mu \in \mathcal{P}(\overline{\Omega} \times [-L, L]^n)$ and $\nu \in \mathcal{P}(\overline{D} \times [-L, L]^n)$. We see that

$$\begin{aligned}(\rho_2)_\# \tilde{G}_\eta(\mu \otimes \nu) &= (\rho_2)_\# \Big(\tilde{\Gamma}_\eta(\mu \otimes \nu, ((\cdot, \cdot), (\cdot, \cdot)))_\# (\mu \otimes \nu) \Big) \\ &= \Big(\rho_2 \circ \tilde{\Gamma}_\eta((\mu \otimes \nu), ((\cdot, \cdot), (\cdot, \cdot))) \Big)_\# (\mu \otimes \nu) \\ &= \Big(\Gamma_\eta((\rho_1)_\#(\mu \otimes \nu), \rho_2((\cdot, \cdot), (\cdot, \cdot))) \Big)_\# (\mu \otimes \nu) \\ &= \Big(\Gamma_\eta((\rho_1)_\#(\mu \otimes \nu), (\cdot, \cdot)) \Big)_\# ((\rho_2)_\#(\mu \otimes \nu)) \\ &= \Gamma_\eta(\mu, \cdot)_\# \nu \in \mathcal{P}\big(\overline{D} \times [-L, L]^n\big).\end{aligned} \tag{164}$$

Next, we prove that

$$(\rho_2)_\# \tilde{\Gamma}_\eta\big(\mathbf{M}_N(h) \otimes \nu, \cdot\big)_\# \big(\mathbf{M}_N(h) \otimes \nu\big) = \sum_{k=1}^K \frac{1}{K} \delta_{(x'_k, y'_k)}, \quad y'_k = A(h_{\eta,N})(x'_k). \tag{165}$$

We recall that $\nu = \sum_{k=1}^K \frac{1}{K} \delta_{(x'_k, 0)}$ and $\mathbf{M}_N(h) = \sum_{j=1}^N \frac{1}{N} \delta_{(x_j, h(x_j))}$. Then, for any $\phi \in C(\overline{D} \times [-L, L]^n)$

$$\begin{aligned}&\langle (\rho_2)_\# \tilde{\Gamma}_\eta\big(\mathbf{M}_N(h) \otimes \nu, \cdot\big)_\# \big(\mathbf{M}_N(h) \otimes \nu\big), \phi \rangle \\ &= \langle \tilde{\Gamma}_\eta\big(\mathbf{M}_N(h) \otimes \nu, \cdot\big)_\# \big(\mathbf{M}_N(h) \otimes \nu\big), \phi \circ \rho_2 \rangle \\ &= \langle \big(\mathbf{M}_N(h) \otimes \nu\big), \phi \circ \rho_2 \circ \tilde{\Gamma}_\eta\big(\mathbf{M}_N(h) \otimes \nu, \cdot\big) \rangle \\ &= \langle \big(\mathbf{M}_N(h) \otimes \nu\big), \phi\big(x', A(\tilde{S}_\eta(\mathbf{M}_N(h)))(x')\big) \rangle \\ &= \sum_{k=1}^K \frac{1}{K} \phi(x'_k, A(\tilde{S}_\eta(\mathbf{M}_N(h)))(x'_k)).\end{aligned} \tag{166}$$

This proves (165). By taking limit $\eta \to 0$, we obtain

$$\lim_{\eta \to 0} (\rho_2)_\# \tilde{\Gamma}_\eta\big(\gamma_h \otimes \nu, \cdot\big)_\# \big(\gamma_h \otimes \nu\big) = \sum_{k=1}^K \frac{1}{K} \delta_{(x'_k, y_k)}, \quad y_k = A(h)(x'_k),\; \nu = \sum_{k=1}^K \frac{1}{K} \delta_{(x'_k, 0)}. \tag{167}$$

G.6 Cross-attention as block-constrained product-measure self-attention

The product-measure construction above gives a self-attention representation on the product space

$$X = (\overline{\Omega} \times [-L, L]^n) \times (\overline{D} \times [-L, L]^n).$$

It is useful to compare this product-space self-attention mechanism with the cross-attention modules that are often used in practical query-decoder architectures. The point of the comparison is that standard cross-attention is not a different kind of measure map: it can be written as a product-measure self-attention layer with a special block structure in the query, key, value, and output matrices. However, the converse is not true: a general product-measure self-attention layer is a larger class.

We write

$$s = (x, y) \in \overline{\Omega} \times [-L, L]^n, \qquad r = (x', y') \in \overline{D} \times [-L, L]^n,$$

and denote a product token by

$$\xi = (s, r) = ((x, y), (x', y')) \in X.$$

Let $\tilde{\mu}$ be a measure on X. The attention part of a single product-space self-attention head, suppressing the graph-preserving projection used in (21), can be written as

$$\mathrm{Att}^X_{\tilde{Q},\tilde{K},\tilde{V}}(\tilde{\mu}, \xi) = \frac{\displaystyle\int_X \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle \tilde{Q}\xi, \tilde{K}\xi' \rangle\right) \tilde{V}\xi' \, d\tilde{\mu}(\xi')}{\displaystyle\int_X \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle \tilde{Q}\xi, \tilde{K}\xi'' \rangle\right) d\tilde{\mu}(\xi'')}. \tag{168}$$

With respect to the splitting $\xi = (s, r)$, a general product-space attention head has matrices of the form

$$\tilde{Q} = [Q_s \quad Q_r], \qquad \tilde{K} = [K_s \quad K_r], \qquad \tilde{V} = [V_s \quad V_r]. \tag{169}$$

Thus the attention score between $\xi = (s, r)$ and $\xi' = (s', r')$ may depend on all four variables s, r, s', r'. In the discrete product measure

$$\tilde{\mu}_{N,K} = \mathbf{M}_N(h) \otimes \nu_K, \qquad \nu_K = \frac{1}{K}\sum_{k=1}^{K} \delta_{(z_k,0)},$$

this corresponds to self-attention over the NK tokens

$$\xi_{ik} = \big((x_i, h(x_i)), (z_k, 0)\big), \qquad i = 1, \dots, N, \quad k = 1, \dots, K.$$

A fully general product-space layer may therefore attend from (i, k) to (j, ℓ) for all i, j, k, ℓ.

We now recover standard query-to-input cross-attention from (168). We choose the block-constrained matrices

$$\tilde{Q}_{\mathrm{ca}} = [0 \quad Q_r], \qquad \tilde{K}_{\mathrm{ca}} = [K_s \quad 0], \qquad \tilde{V}_{\mathrm{ca}} = [V_s \quad 0]. \tag{170}$$

Then

$$\tilde{Q}_{\mathrm{ca}}\xi = Q_r r, \qquad \tilde{K}_{\mathrm{ca}}\xi' = K_s s', \qquad \tilde{V}_{\mathrm{ca}}\xi' = V_s s'.$$

For a product measure $\tilde{\mu} = \mu \otimes \nu$, where μ is a measure on $\overline{\Omega} \times [-L, L]^n$ and ν is a probability measure on $\overline{D} \times [-L, L]^n$, substitution into (168) gives

$$\mathrm{Att}^X_{\mathrm{ca}}(\mu \otimes \nu, (s, r)) = \frac{\displaystyle\iint \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle Q_r r, K_s s' \rangle\right) V_s s' \, d\mu(s') d\nu(r')}{\displaystyle\iint \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle Q_r r, K_s s'' \rangle\right) d\mu(s'') d\nu(r'')} \tag{171}$$

$$= \frac{\displaystyle\int \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle Q_r r, K_s s' \rangle\right) V_s s' \, d\mu(s')}{\displaystyle\int \exp\left(\frac{1}{\sqrt{d_{\mathrm{att}}}}\langle Q_r r, K_s s'' \rangle\right) d\mu(s'')}. \tag{172}$$

The integration over the attended query coordinate r' cancels. The right hand side of (171) is exactly the usual cross-attention map from the query token r to the input or memory measure μ. This identity is a layerwise statement for factorized product measures. The block-constrained input, query, and

cross-attention modules considered below preserve this factorization when their output projections are placed in the corresponding input or query blocks. In contrast, an unrestricted product-space self-attention layer need not preserve a product form.

The same block calculation also identifies other familiar attention modules as special product-space self-attention layers. With the output projection placed in the corresponding block, one obtains

$$\begin{aligned}
\text{input self-attention:} \quad & (\tilde{Q},\tilde{K},\tilde{V}) = ([Q_s \quad 0]\,,[K_s \quad 0]\,,[V_s \quad 0]),\\
\text{query self-attention:} \quad & (\tilde{Q},\tilde{K},\tilde{V}) = ([0 \quad Q_r]\,,[0 \quad K_r]\,,[0 \quad V_r]),\\
\text{query-to-input cross-attention:} \quad & (\tilde{Q},\tilde{K},\tilde{V}) = ([0 \quad Q_r]\,,[K_s \quad 0]\,,[V_s \quad 0]),\\
\text{input-to-query cross-attention:} \quad & (\tilde{Q},\tilde{K},\tilde{V}) = ([Q_s \quad 0]\,,[0 \quad K_r]\,,[0 \quad V_r]).
\end{aligned}$$

The first two choices update the input block and query block, respectively, and the last two choices give the two possible directions of cross-attention.

This yields the following practical hierarchy in terms of computation cost. Full product-measure self-attention on $\mathbf{M}_N(h)\otimes\nu_K$ uses NK product tokens and allows all couplings $(i,k)\leftrightarrow(j,\ell)$; its direct attention cost is $O(N^2K^2)$. The fiberwise product self-attention described in Appendix I instead evaluates the K measures

$$\mathbf{M}_N(h)\otimes\delta_{(z_k,0)}, \qquad k=1,\dots,K,$$

with the same product-space transformer and can treat k as a batch index; this is still literal product-space self-attention for each query, and processing all K queries has total attention cost $O(KN^2)$. The block-constrained cross-attention form (170)–(172) collapses the product measure by the calculation above. After any chosen input encoding of μ, the query-to-input cross-attention sublayer has cost $O(NK)$. This count does not include the input encoder cost, for instance $O(N^2)$ per full input self-attention layer, nor any query self-attention cost, for instance $O(K^2)$ per full query self-attention layer.

Thus standard cross-attention is a special case of product-measure self-attention, but only after imposing block constraints on the weight matrices. The unrestricted product-measure transformer used in the representation theorem is a larger class. Therefore, the approximation results for unrestricted product-measure self-attention do not by themselves prove an approximation theorem for every architecture built only from the restricted cross-attention blocks.

H Example of a computation of $G_\eta(\mu)$ for a discrete measure, μ

We assume, here, that $h\in C(\overline{\Omega})$. The analogous considerations can be made also for $h\in H^{-s}(\Omega)$, but then we start by replacing h by $h_\tau = R_\tau(h)$. Let $x_j\in\Omega$, $j=1,2,\dots,N$. Let $y_j = h(x_j)$ and

$$\mu = \mathbf{M}_N(h) = \sum_{j=1}^{N}\frac{1}{N}\delta_{(x_j,y_j)} \in \mathcal{M}(\overline{\Omega}\times\mathbb{R}), \tag{173}$$

$$S_\eta(\mu)(\cdot) = \sum_{k=1}^{K(\eta)} V_0(\eta\lambda_k^r+1)^{-1}\left[\int_{\overline{\Omega}\times\mathbb{R}}\psi_k(x')y'\,d\mu(x',y')\right]\psi_k(\cdot)\in C^k(\overline{\Omega}), \tag{174}$$

see (136). Denote $h_\eta := \tilde{S}_\eta(\mu) = \Psi\circ S_\eta(\mu)$. Then, $\Gamma_\eta(\mu,(x,y)) = \big(x, A(h_\eta)(x)\big)\in\overline{\Omega}\times\mathbb{R}$ and

$$G_\eta(\mu) = \Gamma_\eta(\mu,\cdot)_\#\mu = \sum_{j=1}^{N}\frac{1}{N}\delta_{(x_j,z_j)}, \quad \text{where } z_j = A(h_\eta)(x_j).$$

I Training of transformers, architecture, and the loss function

In this appendix we consider the representations (167) and (165) of nonlinear operator A. We recall that by (156) and (157), the map $\tilde{\Gamma}_\eta$ in (167) satisfies

$$\tilde{\Gamma}_\eta(\gamma_h\otimes\nu,((x,y),(x',y'))) := \big((x,y),\,\Gamma_\eta(\gamma_h,(x',y'))\big), \quad \gamma_h = (\rho_1)_\#(\gamma_h\otimes\nu) \tag{175}$$

where

$$\Gamma_\eta(\mu,(x',y')) = \big(x', A(\tilde{S}_\eta(\mu))(x')\big) \tag{176}$$

is a mapping that is independent of the measure ν, and therefore of the points x'_j. This suggests that one can learn an approximation for the operator A by training a neural network of the form

$$\mu = \mathbf{M}_N(h) \to \langle \tilde{\Gamma}_\eta(\mu \otimes \delta_{(x',0)},\cdot))_\#(\mu \otimes \delta_{(x',0)}), \tilde{\pi}_2\rangle \tag{177}$$

Observe that the target function $\Gamma_\eta(\mu,(x',y'))$ has the property that its y'-component,

$$\Gamma_\theta^{\mathrm{tran},(y')}(\mu\otimes\nu,((x,y),(x',y'))) = \Gamma_\theta^{\mathrm{tran},(y')}(\mu\otimes\nu;x')$$

depends only on the variable x'. Hence, similarly to the formulas (175) and (176), we may assume that we design the architecture of the neural networks that we use in training,

$$(\mu\otimes\nu,((x,y),(x',y'))) \to \tilde{\Gamma}_\theta^{\mathrm{tran}}(\mu\otimes\nu,((x,y),(x',y')))$$

so that their y'-components,

$$\Gamma_\theta^{\mathrm{tran},(y')}(\mu\otimes\nu,((x,y),(x',y'))) = \Gamma_\theta^{\mathrm{tran},(y')}(\mu\otimes\nu;x')$$

depend only on the variable x'. It defines a map

$$\begin{aligned}\Gamma_\theta^{\mathrm{tran}} \quad : \quad & \mathcal{P}((\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n))\times((\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n)) \\ & \to (\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n)\end{aligned}$$

is of the form (25) and depend on parameters $\theta\in\Theta$ and minimize the cost function

$$L(\theta) = \sum_{p=1}^{P}\sum_{k=1}^{K}\left|\left\langle\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)_\#(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}),\tilde{\pi}_2\right\rangle - A(h_p)(x'_k)\right|^2, \tag{178}$$

when training data consists of function $h_p\in C(\overline{\Omega};[-L,L]^n)$, $p=1,2,\ldots,P$ and the values of the functions $A(h_p)$ at the points $x'_k\in\overline{D}$, $k=1,\ldots,K$. In more explicit way, in (178) we have

$$\left\langle\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)_\#(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}),\tilde{\pi}_2\right\rangle \tag{179}$$

$$= \left\langle \mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}\,,\,\tilde{\pi}_2\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)\right\rangle \tag{180}$$

$$= \left\langle \mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}\,,\,\pi_2\circ\rho_2\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)\right\rangle \tag{181}$$

As we have assumed that $(\mu\otimes\nu,((x,y),(x',y'))) \to \tilde{\Gamma}_\theta^{\mathrm{tran}}(\mu\otimes\nu,((x,y),(x',y')))$ is a context function whose y'-component depends only on the variable x', the above implies that

$$\left\langle\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)_\#(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}),\tilde{\pi}_2\right\rangle \tag{182}$$

$$= \left\langle \mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}\,,\,\tilde{\pi}_2\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)\right\rangle \tag{183}$$

$$= \left\langle \mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}\,,\,\pi_2\circ\rho_2\left(\tilde{\Gamma}_\theta^{\mathrm{tran}}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right)\right\rangle \tag{184}$$

$$= \left\langle \mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)}\,,\,\tilde{\Gamma}_\theta^{\mathrm{tran},(y')}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)},\cdot)\right\rangle \tag{185}$$

$$= \left\langle \delta_{x'_k}\,,\,\Gamma_\theta^{\mathrm{tran},(y')}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)};x')\right\rangle \tag{186}$$

$$= \Gamma_\theta^{\mathrm{tran},(y')}(\mathbf{M}_N(h_p)\otimes\delta_{(x'_k,0)};x'_k) \tag{187}$$

where $\Gamma_\theta^{\mathrm{tran}} : \mathcal{P}((\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n))\times((\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n)) \to (\overline{\Omega}\times[-L,L]^n)\times(\overline{D}\times[-L,L]^n)$ is a composition of MLPs and transformers of the form (25) and $\Gamma_\theta^{\mathrm{tran},(y')}(\mu\otimes\nu;x'_k)$ is its y'-component evaluated at the point x'_k (and at the arbitrary values of (x,y) and y').

J Rough inputs and the graph-measure formulation

The basic operation in measure-theoretic attention is integration of a continuous function against a token measure. Thus, finite token sets and continuum graph measures can be treated within the same measure space. On a compact token space Y, the space $\mathcal{P}(Y)$ is compact in the 1-Wasserstein topology, and empirical graph measures converge in this topology under regularly distributed sampling.

For a general negative-order Sobolev input h, however, point values and the graph measure γ_h need not be defined. Our construction therefore first mollifies h and uses the empirical graph measure $\mathbf{M}_{N,\tau}(h) = \mathbf{M}_N(R_\tau h)$. The spectral regularizer then produces a smooth input for the target operator, so that the corresponding context map is continuous and can be approximated by softmax attention and pointwise MLPs. For inputs satisfying its assumptions, Theorem 2 shows that these approximations recover $A(h)$ when the regularization parameters are removed in the stated order.

This distinguishes the construction from applying pointwise nonlinearities directly to distributions, which is not generally well-defined. It does not imply that other operator-learning architectures cannot be extended through suitable regularization. Our approximation proof specifically uses the universality of softmax-based transformers on measures.

K Numerical illustrations

We now illustrate that a function-graph preserving map can recover a target operator from samples of its input-output behavior. The purpose of this appendix is to provide a numerical illustration of the theory, not to claim state-of-the-art performance on any particular operator learning task.

K.1 Setup and learning task

Our domain is the two-dimensional torus

$$\mathbb{T}^2 = (\mathbb{R}/\mathbb{Z})^2,$$

identified with the unit square with periodic boundary conditions. Inputs are scalar functions sampled at point clouds. For an input function h and support points $x_j \in \mathbb{T}^2$, the model receives the unordered graph tokens

$$\{(x_j, h(x_j))\}_{j=1}^N, \qquad \text{corresponding to} \qquad \mathbf{M}_N(h) = \frac{1}{N}\sum_{j=1}^N \delta_{(x_j,h(x_j))}.$$

The target operator A is a small, randomly initialized one-layer Fourier neural operator (FNO) [Li et al., 2021]. The target FNO used Fourier modes $(6, 6)$, hidden width 16, projection width 128, with a total of 8753 fixed random parameters.

K.2 Transformer configurations

Sinusoidal positional encoding. We encode coordinates with fixed sinusoidal features rather than raw coordinates [Vaswani et al., 2017, Tancik et al., 2020, Mildenhall et al., 2020]. Specifically, each sampled input is represented by tokens $(x_j, r_j^{(0)})$, where $x_j \in \mathbb{T}^2$ is the support coordinate and $r_j^{(0)}$ is an initial latent embedding of the observed value $h(x_j)$, conditioned on fixed sinusoidal features $\phi(x_j)$. The coordinate feature map is $\phi : \mathbb{R}^2 \to \mathbb{R}^{24}$,

$$\phi(x_1, x_2) := \Big(\sin(2\pi\omega_m x_\ell), \cos(2\pi\omega_m x_\ell)\Big)_{\ell=1,2;\, m=1,\dots,6}$$

with frequencies

$$\omega_m = 16^{(m-1)/5}, \qquad m = 1, \dots, 6.$$

Since some ω_m are nonintegers, these features are not 1-periodic and do not enforce the torus boundary condition. At every transformer layer, the coordinates x_j remain fixed. The attention and MLP layers update only the latent variables $r_j^{(\ell)}$, while receiving $\phi(x_j)$ as deterministic coordinate conditioning. Thus the layer map has the form

$$(x_j, r_j^{(\ell)})_{j=1}^N \longmapsto (x_j, r_j^{(\ell+1)})_{j=1}^N,$$

which preserves the support points.

Same-domain and query mode. Our transformer implementation supports two prediction modes. In *same-domain* mode, the model predicts output values on the same coordinates as the input samples,

$$\{(x_j, h(x_j))\}_{j=1}^N \longmapsto \{(x_j, A(h)(x_j))\}_{j=1}^N.$$

For same-domain prediction, the model applies a seven-layer transformer directly to the sampled input graph tokens $(x_j, h(x_j))$. The support coordinates are preserved throughout the network, and a pointwise readout maps the final latent at each input location x_j to the predicted value $A(h)(x_j)$.

In *query* mode, the model predicts at arbitrary query coordinates $y_k \in \mathbb{T}^2$,

$$\big(\{x_j, h(x_j)\}_{j=1}^N, x'_k\big) \longmapsto A(h)(x'_k).$$

For query prediction, the model first applies a six-layer input branch to the sampled graph tokens $\{x_j, h(x_j)\}_{j=1}^N$ and a two-layer pointwise query branch to the query coordinates. For each query point y_k, it then forms the fiber tokens $((x_j, h(x_j)), (x'_k, 0))$, applies a fiberwise product-space self-attention block followed by two residual token-wise MLP blocks, and reads out from the resulting fiber.

Both the same-domain and query mode transformers lift their input tokens to an embedding space with dimension of 144. All attention blocks use 4 heads.

K.3 Training protocol

Training data was generated online: at each optimization step, we sampled a fresh batch of random input functions h and fresh point clouds, rather than drawing from a fixed stored dataset. Input functions were sampled online from a truncated periodic Gaussian Fourier field: sine and cosine coefficients were drawn independently with frequency-decaying variance, rescaled to have approximate sup-norm 0.9. The target values $A(h)$ were computed on the fly by discretizing h onto a 64×64 torus grid, applying the fixed one-layer FNO teacher, and interpolating the output to the requested locations x_j or x'_k.

The transformers were trained over $20{,}000$ training steps with batch size 32. After a linear warmup period lasting 1000 steps, we used a cosine learning rate from 10^{-3} decaying to 10^{-4}. Each step of query-mode training used $K_{\text{train}} = 16$ query points.

In both modes, input support points were sampled uniformly at random during training. In query mode, the query points were also sampled uniformly at random; in same-domain mode there are no separate query points, since the model is trained to predict $A(h)(x_j)$ at the same sampled support locations x_j.

For predictions $\widehat{u}$ and targets $u = A(h)$, we report

$$\text{rel } L^2 = \frac{\|\widehat{u} - u\|_{L^2}}{\|u\|_{L^2}}, \qquad \text{rel } L^\infty = \frac{\|\widehat{u} - u\|_{L^\infty}}{\|u\|_{L^\infty}}.$$

The L^2 norms are computed by empirical quadrature over the evaluation points for each held-out function, the L^∞ norms by the corresponding pointwise maximum, and the reported values are averages over a number of held-out functions h specified in each table caption. Same-domain mode use $N = 625$ evaluation points, while query-mode norms use $K = 4096$ evaluation points.

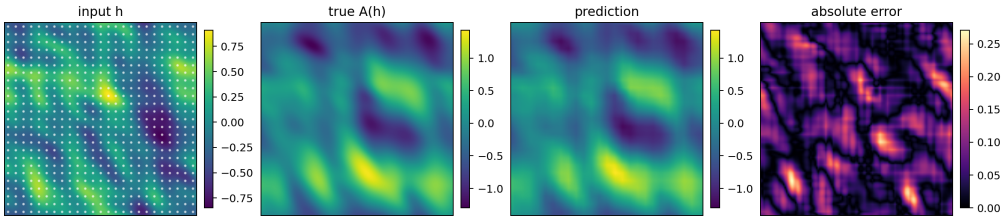
Where appropriate, the tables below report 95% confidence interval half-widths computed over per-function error samples.

K.4 Recovery results

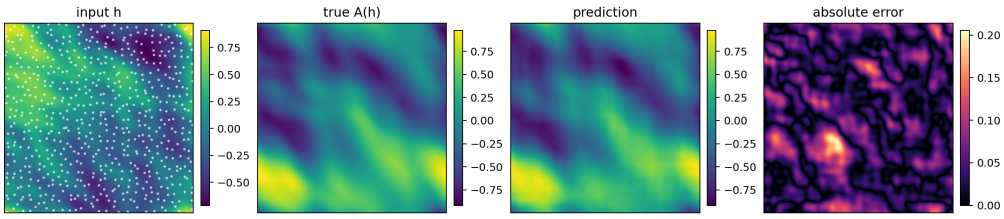
Table 1 gives the main recovery results over 128 held-out functions. Figure 2 shows representative samples on input functions and sampling regimes not seen during training.

K.5 Query-mode comparison

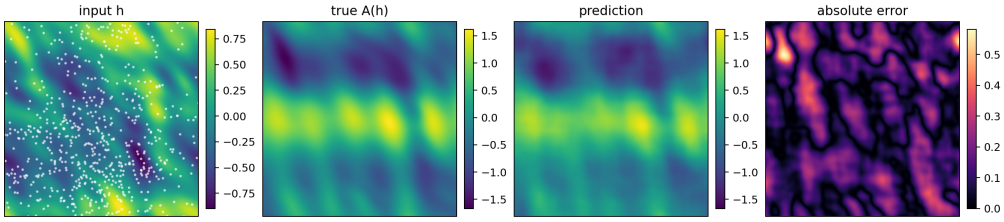
Table 2 compares the product-space query model against the same-domain transformer followed by interpolation to the query coordinates. Both methods are evaluated at $K = 4096$ query locations over 128 held-out functions. Product-space self-attention substantially improves arbitrary-coordinate prediction over the interpolation baseline.



(a) Same-domain mode, grid input sampling.



(b) Query mode, jittered-grid input sampling.



(c) Query mode, Gaussian input sampling.

Figure 2: Representative two-dimensional FNO-teacher recovery examples for the trained same-domain and query-mode transformers. The white dots in the input panel mark the sampling locations $\{x_j\}_{j=1}^{N}$. For the same-domain example (a), the network predicts at the input grid locations, and values at the displayed pixels are obtained by periodic bilinear interpolation from these predicted grid values. For the query-mode examples (b) and (c), the displayed pixel locations are used directly as query points, so all pixel values are returned by the trained network.

Table 1: Main recovery results for a fixed random one-layer FNO teacher. Errors are per-function means over 128 held-out functions; intervals are 95% confidence interval half-widths. Timings were recorded on a single NVIDIA RTX PRO 6000 Blackwell GPU with 96GB RAM.

mode	params	rel. L^2	rel. L^∞	train time
same-domain	2,315,666	0.0609 ± 0.0031	0.0689 ± 0.0034	406.7 s
query	3,381,410	0.1106 ± 0.0048	0.1566 ± 0.0065	1830.2 s

K.6 Irregular sampling and empirical measure convergence

We next probe robustness away from the training input resolution and sampling distribution. Table 3 fixes the evaluation set to $K = 4096$ query locations and varies the number of observed input samples N, using 16 held-out functions for each resolution. Although the query-mode model was trained only with $N = 625$ input samples, the relative L^2 error decreases steadily as more input samples are provided, from 0.3582 ± 0.0436 at $N = 128$ to 0.0800 ± 0.0058 at $N = 2048$. This resolution-transfer behavior is consistent with learning an approximation to the underlying continuous operator, rather than a predictor tied to a fixed input discretization.

Table 4 evaluates the trained query-mode model at $K = 4096$ query locations while changing the input sampling distribution. Each row uses 16 held-out functions. Although training used uniformly

Table 2: Query-mode vs interpolation. The interpolation baseline first applies the same-domain transformer and then interpolates its prediction to query points. Product-space self-attention provides substantially better accuracy at arbitrary query points than the interpolation baseline.

	mode	rel. L^2	rel. L^∞
product-space self-attention	query	0.1117 ± 0.0043	0.1588 ± 0.0080
interpolation	same-domain	0.2260 ± 0.0067	0.4764 ± 0.0162

Table 3: Resolution transfer for the trained query-mode product-space transformer. Each row uses 16 held-out functions and $K = 4096$ query locations. The transformer was trained with $N = 625$ input samples, and was never provided inputs with the values of N in this table.

N	rel. L^2	rel. L^∞
128	0.3582 ± 0.0436	0.4404 ± 0.0517
256	0.2004 ± 0.0260	0.2869 ± 0.0321
512	0.1184 ± 0.0115	0.1680 ± 0.0151
1024	0.0929 ± 0.0127	0.1306 ± 0.0117
2048	0.0800 ± 0.0058	0.1205 ± 0.0156

random input samples, the model attains its lowest errors on structured grid and jittered-grid inputs, and degrades gradually under uniform random and Gaussian sampling. We note that Gaussian sampling test is intended as a stress-test, and points $x_1, x_2, \ldots$ sampled in this way are not regularly distributed in the sense of Definition 4.

Table 4: Irregular-sampling evaluation for the trained query-mode model. Each row uses 16 held-out functions and $K = 4096$ query locations. The model was trained using uniformly random input sampling, and was not trained on the other sampling methods in this table.

input sampling	rel. L^2	rel. L^∞
grid	0.0714 ± 0.0063	0.1079 ± 0.0155
jittered grid	0.0832 ± 0.0095	0.1276 ± 0.0200
uniform random	0.1202 ± 0.0165	0.1821 ± 0.0253
Gaussian	0.1819 ± 0.0247	0.2436 ± 0.0322

K.7 Limitations

These numerical experiments are small-scale, and are only intended to illustrate certain aspects of the theory. Our numerical illustration uses an FNO as a target operator $A : C(\mathbb{T}^2, [-L, L]) \to C(\mathbb{T}^2, [-L, L])$, which fits the setting of Theorem 1. We do not illustrate a learned operator between negative-order Sobolev spaces, the setting of Theorem 2.

L Quantitative approximation rates for function-graph transformers

This appendix develops quantitative counterparts of the universal approximation result in Theorem 2 under additional Hölder assumptions. We first construct and realize finite-statistic approximations of general measure-dependent context maps, and then exploit the spectral factorization of the particular context map Γ_η. Estimates for tokenization, input mollification, spectral regularization, and decoding are combined into an end-to-end error bound and a coupled parameter schedule. Throughout, we use the cutoff $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$ from (139).

L.1 Setting and notation

Throughout, $\Omega \subset \mathbb{R}^d$ is bounded with C^∞ boundary, $V_0 := \mathrm{Vol}(\Omega)$, and $\bar{\lambda}_\Omega$ is normalized Lebesgue measure on Ω. The token space is

$$\Xi := \overline{\Omega} \times [-L, L]^n \subset \mathbb{R}^D, \qquad D := d + n, \qquad \Delta := \mathrm{diam}(\Xi),$$

and $\mathcal{P}(\Xi)$ denotes Borel probability measures on Ξ with the 1-Wasserstein metric and the Kantorovich–Rubinstein duality

$$W_1(\mu, \nu) = \sup\Big\{ \int_\Xi \varphi \, d(\mu - \nu) : \mathrm{Lip}(\varphi) \le 1 \Big\}. \tag{188}$$

Since Ξ is compact, $(\mathcal{P}(\Xi), W_1)$ is a compact metric space metrizing the weak* topology. For $h \in C(\overline{\Omega}; [-L, L]^n)$ the graph measure is $\gamma_h := (g_h)_\# \bar{\lambda}_\Omega$, $g_h(x) = (x, h(x))$; empirical tokenizations are $\mathbf{M}_N(h) = \frac{1}{N} \sum_{j=1}^N \delta_{(\bar{x}_j, h(\bar{x}_j))}$ for grid points $\bar{x}_1, \dots, \bar{x}_N \in \Omega$, and $\mathbf{M}_{N,\tau}(h) := \mathbf{M}_N(R_\tau h)$ with the mollifier R_τ of (214).

Let $(\phi_j, \lambda_j)_{j \ge 1}$ be the scalar Dirichlet eigenpairs of $I - \Delta$ on Ω, L^2-normalized, acting diagonally on $\mathbb{R}^n$-valued functions. By the Weyl law (Safarov and Vassiliev [1996]; also (38)) there is $c_W \ge 1$ with

$$c_W^{-1} j^{2/d} \;\le\; \lambda_j \;\le\; c_W j^{2/d}, \qquad j \ge 1. \tag{189}$$

For $m \in \mathbb{Z}_{\ge 0}$ let $s(m)$ be the smallest even integer $> \frac{d}{2} + m$; iterated elliptic regularity and the Sobolev embedding give (Gilbarg et al. [1977])

$$\|\phi_j\|_{C^m(\overline{\Omega})} \;\le\; C_S \, \lambda_j^{s(m)/2}, \qquad j \ge 1. \tag{190}$$

The Dirichlet spectral scale defines $\|u\|_{H^\sigma}^2 := \sum_j \lambda_j^\sigma |\hat{u}_j|^2$, $\hat{u}_j := \langle \phi_j, u \rangle \in \mathbb{R}^n$; for $|\sigma| < \frac{1}{2}$ this is equivalent to the standard $H^\sigma(\Omega)$ norm, and for larger $|\sigma|$ with $\sigma - \frac{1}{2} \notin \mathbb{Z}$ the identification holds on the Dirichlet domain scale, the setting of this paper; smooth compactly supported functions belong to the domain of every positive integer power of the Dirichlet realization, which is used in Subsection L.7.

Fix an even integer $r > \frac{d}{2} + k$ (with k as in Assumption 2 below). The spectral regularizer, in the V_0 normalization, is

$$S_{\eta,K}(\mu) \;=\; \sum_{j=1}^K \frac{b_j(\mu)}{1 + \eta \lambda_j^r} \, \phi_j, \qquad b_j(\mu) := V_0 \int_\Xi \phi_j(x) \, y \, d\mu(x, y) \in \mathbb{R}^n, \tag{191}$$

so that on graph measures $b_j(\gamma_h) = \langle \phi_j, h \rangle$ and

$$S_{\eta,K}(\gamma_h) = \big(\eta(I - \Delta)^r + I\big)^{-1} P_K h. \tag{192}$$

We abbreviate $S_\eta := S_{\eta, K(\eta)}$ with $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$. Let $\Psi \in C^\infty(\mathbb{R}^n; \mathbb{R}^n)$ be the clipping of this paper: the identity on $[-L, L]^n$, with image in $[-L_1, L_1]^n$; set $\tilde{S}_\eta := \Psi \circ S_\eta$. This paper's context function (31) is

$$\Gamma_\eta(\mu, (x, y)) \;:=\; \big(x,\; A(\tilde{S}_\eta(\mu))(x)\big), \qquad G_\Gamma(\mu) := \Gamma(\mu, \cdot)_\# \mu, \tag{193}$$

and the decoder is $\mathbf{F}_\varepsilon(\gamma)(x) := V_0 \varepsilon^{-d} \int_\Xi \rho\big((x' - x)/\varepsilon\big) \, y' \, d\gamma(x', y')$ for a fixed even mollifier $\rho \in C_c^\infty(B(0,1))$, $\int \rho = 1$; on graph measures $\mathbf{F}_\varepsilon(\gamma_g) = \rho_\varepsilon * \tilde{g}$ with $\tilde{g}$ the zero extension of g.

L.2 The Hölder classes and the entropy of the input space

Definition 6. *For $\alpha \in (0,1]$, $\Lambda > 0$, $M_\xi \geq 0$, set*

$$\mathcal{G}_\alpha(\Lambda) := \Big\{\Gamma : \mathcal{P}(\Xi) \times \Xi \to \Xi : \sup_{\xi\in\Xi} |\Gamma(\mu,\xi) - \Gamma(\nu,\xi)| \leq \Lambda W_1(\mu,\nu)^\alpha \ \ \forall \mu,\nu\Big\}, \tag{194}$$

$$\mathcal{G}_\alpha^{\text{joint}}(\Lambda, M_\xi) := \Big\{\Gamma \in \mathcal{G}_\alpha(\Lambda) : \sup_{\mu\in\mathcal{P}(\Xi)} \big[\Gamma(\mu,\cdot)\big]_{C^{0,\alpha}(\Xi)} \leq M_\xi\Big\}. \tag{195}$$

The two moduli enter different bounds: the measure modulus Λ alone determines the number of cylindrical statistics, hence the attention-head count (Subsections L.3–L.4); the token modulus M_ξ is what makes the approximation of the functions $\Gamma(\mu_j,\cdot)$, with μ_j fixed, by finite ReLU networks possible, and it enters only the size of the outer network. The lower bound of Theorem 4 holds for maps $\Gamma(\mu,(x,y)) = (x, F(\mu))$, whose value component is independent of the token. These maps belong to $\mathcal{G}_\alpha^{\text{joint}}(1, \Delta^{1-\alpha})$. By (188) and the compactness of $\mathcal{P}(\Xi)$, membership in $\mathcal{G}_\alpha(\Lambda)$ is the natural metric strengthening of the weak*-continuity hypothesis of [Furuya et al., 2025b, Thm. 1].

We write $\mathrm{Lip}_1^0(\Xi) := \{\varphi : \Xi \to \mathbb{R} : \mathrm{Lip}(\varphi) \leq 1,\ \varphi(x^*) = 0\}$ for the base-point-normalized unit Lipschitz ball at a fixed $x^* \in \Xi$; the normalization loses nothing in (188) (constants integrate to zero against $\mu - \nu$) and gives $\|\varphi\|_\infty \leq \Delta$ for every $\varphi \in \mathrm{Lip}_1^0(\Xi)$, whence, by Arzelà–Ascoli and the classical entropy of Lipschitz classes [Tikhomirov, 1993],

$$\log N\big(\mathrm{Lip}_1^0(\Xi), \|\cdot\|_\infty, \delta\big) \ \leq\ \Big(\tfrac{C\Delta}{\delta}\Big)^D \log \tfrac{C\Delta}{\delta}, \qquad 0 < \delta \leq \Delta. \tag{196}$$

Proposition 9 (Entropy of the input space)**.** *Let $\varrho_0 > 0$ be such that Ξ contains a ball of radius ϱ_0. Then for $0 < \varepsilon \leq \varrho_0/8$,*

$$\exp\big(c\,\varepsilon^{-D}\big) \ \leq\ N\big(\mathcal{P}(\Xi), W_1, \varepsilon\big) \ \leq\ \exp\Big(C\varepsilon^{-D}\log\tfrac{C}{\varepsilon}\Big), \tag{197}$$

with c, C depending only on D, Δ, ϱ_0.

Proof. The upper bound follows from an elementary discretization of the support and quantization of the probability weights. The lower bound follows from the general Wasserstein packing estimate of [Boedihardjo et al., 2024, Proposition 8.2]. For completeness, we give direct proofs below, using a constant-weight Gilbert–Varshamov construction for the lower bound.

Upper bound. Choose an $(\varepsilon/2)$-net $z_1, \ldots, z_M$ of Ξ, with

$$M \leq C_D \left(1 + \frac{\Delta}{\varepsilon}\right)^D.$$

A measurable nearest-point map transports every $\mu \in \mathcal{P}(\Xi)$ to a measure

$$\mu^z = \sum_{i=1}^M w_i \delta_{z_i}, \qquad w \in \Sigma_M := \left\{ w \in [0,1]^M : \sum_{i=1}^M w_i = 1 \right\},$$

at cost $W_1(\mu,\mu^z) \leq \varepsilon/2$.

Set

$$q := \left\lceil \frac{2\Delta M}{\varepsilon} \right\rceil.$$

For every $w \in \Sigma_M$, there are integers $a_i \geq 0$, $\sum_i a_i = q$, such that

$$\left| w_i - \frac{a_i}{q} \right| \leq \frac{1}{q} \qquad (i = 1, \ldots, M).$$

Indeed, first take $a_i = \lfloor q w_i \rfloor$, and then add one to the required number of entries with the largest fractional parts. Thus, writing $v_i = a_i/q$,

$$\|w - v\|_{\ell^1} \leq \frac{M}{q}.$$

By matching the common mass at each z_i and transporting the remaining mass over distances at most Δ,

$$W_1\left(\mu^z, \sum_{i=1}^{M} v_i \delta_{z_i}\right) \le \frac{\Delta}{2}\|w - v\|_{\ell^1} \le \frac{\Delta M}{2q} \le \frac{\varepsilon}{4}.$$

Consequently, the measures with weights in $q^{-1}\mathbb{Z}^M \cap \Sigma_M$ form an ε-net of $\mathcal{P}(\Xi)$.

The number of these weight vectors is

$$\binom{q+M-1}{M-1} \le \binom{q+M}{M} \le \left(\frac{e(q+M)}{M}\right)^M \le \left(C\left(1+\frac{\Delta}{\varepsilon}\right)\right)^M.$$

Since $\Delta \ge 2\varrho_0$ and $\varepsilon \le \varrho_0/8$, taking logarithms yields

$$\log N\big(\mathcal{P}(\Xi), W_1, \varepsilon\big) \le C\varepsilon^{-D} \log \frac{C}{\varepsilon}.$$

Lower bound. We first consider sufficiently small ε. A volumetric packing argument in $B(x_0, \varrho_0) \subset \Xi$ gives a 16ε-separated set

$$S = \{z_1, \dots, z_M\}, \qquad M \ge c_D \left(\frac{\varrho_0}{\varepsilon}\right)^D,$$

where, by discarding at most one point and adjusting c_D, we may assume that M is even. For sufficiently large M, the constant-weight Gilbert–Varshamov bound gives a family $\mathcal{C}$ of subsets of $\{1, \dots, M\}$ such that

$$|\mathcal{C}| \ge \exp(c_0 M), \qquad |\sigma| = \frac{M}{2}, \qquad |\sigma \triangle \sigma'| \ge \frac{M}{4} \quad (\sigma \ne \sigma',\ \sigma, \sigma' \in \mathcal{C}),$$

with an absolute constant $c_0 > 0$.

For $\sigma \in \mathcal{C}$, define

$$\mu_\sigma := \frac{2}{M} \sum_{i \in \sigma} \delta_{z_i}.$$

If $\sigma \ne \sigma'$, then the cardinality $|\sigma \setminus \sigma'|$ of the set $\sigma \setminus \sigma'$ is bounded by

$$|\sigma \setminus \sigma'| = \frac{1}{2}|\sigma \triangle \sigma'| \ge \frac{M}{8}.$$

The mass of μ_σ on $\{z_i : i \in \sigma \setminus \sigma'\}$ is therefore at least $1/4$. Since $\mu_{\sigma'}$ assigns no mass to these points, every coupling must transport this mass to distinct points of S, over distances at least 16ε. Hence

$$W_1(\mu_\sigma, \mu_{\sigma'}) \ge \frac{1}{4}\, 16\varepsilon = 4\varepsilon.$$

A ball of radius ε in the W_1 metric can contain at most one of these measures. It follows that

$$N\big(\mathcal{P}(\Xi), W_1, \varepsilon\big) \ge |\mathcal{C}| \ge \exp(c_0 M) \ge \exp(c_1 \varepsilon^{-D})$$

for $0 < \varepsilon \le \varepsilon_0$, where $\varepsilon_0 > 0$ and $c_1 > 0$ depend only on D and ϱ_0.

To cover the remaining range, choose two points $a, b \in B(x_0, \varrho_0)$ with $|a - b| = \varrho_0$. Then

$$W_1(\delta_a, \delta_b) = \varrho_0 > 2\varepsilon \qquad \text{for } 0 < \varepsilon \le \varrho_0/8,$$

so the covering number is at least 2 throughout the stated range. Replacing c_1 by

$$c := \min\{c_1, (\log 2)\varepsilon_0^D\},$$

we obtain $\exp(c\varepsilon^{-D}) \le 2$ for $\varepsilon \ge \varepsilon_0$. This proves the lower bound for the entire range $0 < \varepsilon \le \varrho_0/8$. □

L.3 Constructive Stone–Weierstrass approximation and conditional complexity lower bounds

The proof of [Furuya et al., 2025b, Thm. 1] has three steps: (i) a Stone–Weierstrass approximation of the target by cylindrical maps $\tilde{\Gamma}(\mu,\xi) = g(\xi, \int \varphi_1 d\mu, \dots, \int \varphi_m d\mu)$; (ii) approximation of the multiplication map by context-free MLPs; (iii) realization of cylindrical maps by attention layers. Step (i) does not provide explicit complexity bounds. This subsection makes the approximation constructive and establishes a conditional lower bound on parameter complexity.

Theorem 3 (Constructive cylindrical approximation)**.** *Let $\Gamma \in \mathcal{G}_\alpha(\Lambda)$ be spatial-coordinate preserving, that is,*

$$\Gamma(\mu,(x,y)) = (x, \Gamma_2(\mu,(x,y))).$$

Let $0 < \varepsilon \le \Lambda\Delta^\alpha$, and set $\varepsilon' := \frac{1}{3}(\varepsilon/\Lambda)^{1/\alpha}$. There exist an integer m with

$$m \;\le\; \exp\Big(\big(\tfrac{C\Delta}{\varepsilon'}\big)^D \log \tfrac{C\Delta}{\varepsilon'}\Big) \;\le\; \exp\Big(C'\big(\tfrac{\Lambda^{1/\alpha}}{\varepsilon^{1/\alpha}}\big)^D \log \tfrac{C'}{\varepsilon}\Big),$$

test functions $\varphi_1,\dots,\varphi_m \in \mathrm{Lip}_1^0(\Xi)$ (hence $\|\varphi_i\|_\infty \le \Delta$), and an explicit function $g : \Xi \times [-\Delta,\Delta]^m \to \Xi$, Lipschitz in the statistics for the ℓ^∞ norm, such that $\tilde{\Gamma}(\mu,\xi) := g\big(\xi, (\int \varphi_i \, d\mu)_{i\le m}\big)$ satisfies

$$\sup_{\mu\in\mathcal{P}(\Xi)} \sup_{\xi\in\Xi} \big|\tilde{\Gamma}(\mu,\xi) - \Gamma(\mu,\xi)\big| \;\le\; \varepsilon.$$

Moreover, $\tilde{\Gamma}$ preserves the spatial coordinate. The test functions are normalized at a fixed base point, so their sup norms are uniformly bounded by Δ.

Proof. Choose an ε'-net $\mu_1,\dots,\mu_{N'}$ of $(\mathcal{P}(\Xi), W_1)$. Choose also an $(\varepsilon'/4)$-net $\varphi_1,\dots,\varphi_m$ of $\mathrm{Lip}_1^0(\Xi)$ in the sup norm. The entropy estimate (196) gives

$$m \le \exp\left(\left(\frac{C\Delta}{\varepsilon'}\right)^D \log \frac{C\Delta}{\varepsilon'}\right),$$

after adjusting C. Since each φ_i is 1-Lipschitz and vanishes at the fixed base point, $\|\varphi_i\|_\infty \le \Delta$.

Write

$$t_i(\mu) := \int_\Xi \varphi_i \, d\mu, \qquad t(\mu) := (t_i(\mu))_{i=1}^m, \qquad c_{ij} := t_i(\mu_j).$$

For $t \in [-\Delta,\Delta]^m$, define

$$\widehat{W}_j(t) := \max_{1\le i\le m} (t_i - c_{ij}). \tag{198}$$

For every $\mu \in \mathcal{P}(\Xi)$, Kantorovich–Rubinstein duality and the choice of the test-function net imply

$$W_1(\mu,\mu_j) - \frac{\varepsilon'}{2} \le \widehat{W}_j(t(\mu)) \le W_1(\mu,\mu_j). \tag{199}$$

Indeed, the upper bound follows because every φ_i belongs to $\mathrm{Lip}_1^0(\Xi)$. For the lower bound, replacing any $\varphi \in \mathrm{Lip}_1^0(\Xi)$ by a net function changes its integral against $\mu - \mu_j$ by at most $2(\varepsilon'/4)$.

Define, on the entire statistics cube,

$$h_j(t) := \left(1 - \frac{\widehat{W}_j(t)}{2\varepsilon'}\right)_+, \qquad S(t) := \sum_{j=1}^{N'} h_j(t), \qquad h_0(t) := \left(\frac{1}{4} - S(t)\right)_+. \tag{200}$$

Set

$$H(t) := S(t) + h_0(t) = \max\left\{S(t), \frac{1}{4}\right\}, \qquad \psi_j(t) := \frac{h_j(t)}{H(t)} \quad (j = 0,\dots,N').$$

These weights are well defined and nonnegative throughout $[-\Delta,\Delta]^m$, and

$$\sum_{j=0}^{N'} \psi_j(t) = 1.$$

Each $\widehat{W}_j$ is 1-Lipschitz for the ℓ^∞ norm, so each h_j, $j \geq 1$, is $(2\varepsilon')^{-1}$-Lipschitz. Moreover,

$$|h_0(t) - h_0(t')| \leq |S(t) - S(t')| \leq \sum_{j=1}^{N'} |h_j(t) - h_j(t')|.$$

Using $H(t) \geq 1/4$ and the normalization of the weights, we obtain

$$\begin{aligned}\sum_{j=0}^{N'} |\psi_j(t) - \psi_j(t')| &\leq \frac{2}{H(t)} \sum_{j=0}^{N'} |h_j(t) - h_j(t')| \\ &\leq \frac{CN'}{\varepsilon'} \|t - t'\|_{\ell^\infty}.\end{aligned}$$

Take $\mu_0 := \mu_1$ and define

$$g(\xi, t) := \sum_{j=0}^{N'} \psi_j(t) \Gamma(\mu_j, \xi). \tag{201}$$

By the spatial-coordinate-preserving assumption, for $\xi = (x, y)$ this reads

$$g((x, y), t) = \left(x, \sum_{j=0}^{N'} \psi_j(t) \Gamma_2(\mu_j, (x, y)) \right).$$

The second component is a convex combination of points in $[-L, L]^n$. Hence g takes values in Ξ and preserves the spatial coordinate.

For fixed ξ, subtracting $\Gamma(\mu_0, \xi)$ inside the difference and using $\operatorname{diam}(\Xi) = \Delta$ gives

$$\begin{aligned}|g(\xi, t) - g(\xi, t')| &\leq \Delta \sum_{j=0}^{N'} |\psi_j(t) - \psi_j(t')| \\ &\leq \frac{C\Delta N'}{\varepsilon'} \|t - t'\|_{\ell^\infty}.\end{aligned}$$

Thus g is uniformly Lipschitz in the statistics.

It remains to verify the approximation error. For every μ, some μ_j satisfies $W_1(\mu, \mu_j) \leq \varepsilon'$. By (199), $\widehat{W}_j(t(\mu)) \leq \varepsilon'$, and therefore $h_j(t(\mu)) \geq 1/2$. Consequently,

$$S(t(\mu)) \geq \frac{1}{2}, \qquad h_0(t(\mu)) = \psi_0(t(\mu)) = 0.$$

Furthermore, if $\psi_j(t(\mu)) > 0$ for $j \geq 1$, then $\widehat{W}_j(t(\mu)) < 2\varepsilon'$. Equation (199) yields

$$W_1(\mu, \mu_j) < \frac{5}{2}\varepsilon'.$$

Using the measure-variable Hölder assumption, we conclude that, uniformly in μ and ξ,

$$\begin{aligned}|g(\xi, t(\mu)) - \Gamma(\mu, \xi)| &\leq \sum_{j=1}^{N'} \psi_j(t(\mu)) |\Gamma(\mu_j, \xi) - \Gamma(\mu, \xi)| \\ &\leq \Lambda \left(\frac{5}{2}\varepsilon' \right)^\alpha \\ &\leq \Lambda (3\varepsilon')^\alpha = \varepsilon.\end{aligned}$$

This proves the claim. □

Theorem 4 (Conditional lower bound on parameter complexity)**.** *Let*

$$\mathcal{H}_\alpha := \{\Gamma_F(\mu, (x, y)) = (x, F(\mu)) : F : \mathcal{P}(\Xi) \to [-L, L]^n,\ |F(\mu) - F(\nu)| \leq W_1(\mu, \nu)^\alpha\}.$$

There exist constants $c > 0$ and $\varepsilon_0 > 0$, depending only on the fixed geometric data, L, and α, with the following property.

Let $B, L_{\mathrm{par}} > 0$, and let $\{G_\theta : \theta \in [-B, B]^P\}$ be a family of maps $\mathcal{P}(\Xi) \times \Xi \to \Xi$ satisfying

$$\|G_\theta - G_{\theta'}\|_\infty \le L_{\mathrm{par}} \|\theta - \theta'\|_{\ell^\infty}.$$

Suppose that, for some $0 < \varepsilon \le \varepsilon_0$, every $\Gamma_F \in \mathcal{H}_\alpha$ admits a parameter $\theta \in [-B, B]^P$ such that

$$\|G_\theta - \Gamma_F\|_\infty \le \varepsilon.$$

Then

$$P \log\left(1 + \frac{2BL_{\mathrm{par}}}{\varepsilon}\right) \ge \exp\left(c\varepsilon^{-D/\alpha}\right).$$

In particular, if

$$\log\left(1 + \frac{2BL_{\mathrm{par}}}{\varepsilon}\right) \le \exp\left(\frac{c}{2}\varepsilon^{-D/\alpha}\right),$$

then

$$P \ge \exp\left(\frac{c}{2}\varepsilon^{-D/\alpha}\right).$$

Proof. Set

$$r_\varepsilon := (16\varepsilon)^{1/\alpha}.$$

For sufficiently small ε, the entropy lower bound in Proposition 9, together with the covering property of a maximal separated set, yields measures $\mu_1, \dots, \mu_M$ such that

$$W_1(\mu_s, \mu_{s'}) > r_\varepsilon \qquad (s \ne s'), \qquad M \ge \exp\left(c_0 \varepsilon^{-D/\alpha}\right),$$

where $c_0 > 0$ depends only on the fixed data. Indeed, a maximal r_ε-separated set is an r_ε-net, so its cardinality is at least $N(\mathcal{P}(\Xi), W_1, r_\varepsilon)$.

For every $\sigma \subset \{1, \dots, M\}$, define

$$u_\sigma(\mu) := \max_{s \in \sigma} \left(8\varepsilon - W_1(\mu, \mu_s)^\alpha\right)_+,$$

with $u_\varnothing := 0$, and set

$$F_\sigma(\mu) := u_\sigma(\mu) e_1, \qquad \Gamma_\sigma(\mu, (x, y)) := (x, F_\sigma(\mu)),$$

where e_1 is the first coordinate vector of $\mathbb{R}^n$.

Each function $\mu \mapsto W_1(\mu, \mu_s)$ is 1-Lipschitz. Since $0 < \alpha \le 1$,

$$|a^\alpha - b^\alpha| \le |a - b|^\alpha \qquad (a, b \ge 0).$$

Taking the positive part and a finite maximum preserves this Hölder bound. Consequently,

$$|u_\sigma(\mu) - u_\sigma(\nu)| \le W_1(\mu, \nu)^\alpha.$$

Moreover, $0 \le u_\sigma \le 8\varepsilon$. By choosing $\varepsilon_0 \le L/8$, we ensure that F_σ takes values in $[-L, L]^n$. Thus $\Gamma_\sigma \in \mathcal{H}_\alpha$.

If $\sigma \ne \sigma'$, interchange them if necessary and choose $s \in \sigma \setminus \sigma'$. At μ_s we have

$$u_\sigma(\mu_s) = 8\varepsilon, \qquad u_{\sigma'}(\mu_s) = 0,$$

because $W_1(\mu_s, \mu_{s'})^\alpha > 16\varepsilon$ for every $s' \ne s$. Hence

$$\|\Gamma_\sigma - \Gamma_{\sigma'}\|_\infty = 8\varepsilon.$$

There are therefore 2^M targets with pairwise sup distance 8ε.

Choose an $\varepsilon / L_{\mathrm{par}}$-net Θ_ε of the parameter cube for the ℓ^∞ norm, with

$$|\Theta_\varepsilon| \le \left(1 + \frac{2BL_{\mathrm{par}}}{\varepsilon}\right)^P.$$

For each σ, choose an approximating parameter θ_σ satisfying

$$\|G_{\theta_\sigma} - \Gamma_\sigma\|_\infty \le \varepsilon,$$

and then choose $\widehat{\theta}_\sigma \in \Theta_\varepsilon$ within $\varepsilon/L_{\mathrm{par}}$ of θ_σ. The parameter Lipschitz bound gives

$$\|G_{\widehat{\theta}_\sigma} - \Gamma_\sigma\|_\infty \le 2\varepsilon.$$

Distinct targets must have distinct $\widehat{\theta}_\sigma$: otherwise their mutual distance would be at most 4ε, contradicting the 8ε separation. Therefore,

$$2^M \le \left(1 + \frac{2BL_{\mathrm{par}}}{\varepsilon}\right)^P.$$

Taking logarithms yields

$$P \log\left(1 + \frac{2BL_{\mathrm{par}}}{\varepsilon}\right) \ge M \log 2 \ge (\log 2) \exp\left(c_0 \varepsilon^{-D/\alpha}\right).$$

After decreasing c_0 to a positive constant c and, if necessary, decreasing ε_0, this implies the first assertion.

Under the additional hypothesis on BL_{par}, division gives

$$P \ge \frac{\exp(c\varepsilon^{-D/\alpha})}{\exp((c/2)\varepsilon^{-D/\alpha})} = \exp\left(\frac{c}{2}\varepsilon^{-D/\alpha}\right).$$

This proves the claim. □

L.4 Explicit realization by a transformer: formulas and architecture bounds

This subsection realizes the cylindrical map of Theorem 3 by an explicit transformer, with all weight matrices displayed, and counts heads, dimensions, depths, widths, and parameters. A feedforward network evaluates the test functions at each token. These values are then averaged using an attention layer with all scores set to zero, which gives uniform attention weights $1/N$ on a sequence of N tokens. A second feedforward network evaluates g from the resulting averages and the current token. The token modulus is needed here, and only here.

L.4.1 The ReLU approximation $\tilde{g}$

Write $\gamma_j := \Gamma(\mu_j, \cdot)_2 : \Xi \to [-L, L]^n$ for the value component with the measure fixed at μ_j, so that $g(\xi, t) = (x, \sum_j \psi_j(t)\gamma_j(\xi))$ in (201). The constituents of g are replaced as follows. The maxima in (198) and the positive-part functions in (200) can be represented exactly by ReLU networks: $\max(a, b) = a + \mathrm{ReLU}(b - a)$ iterated in a balanced binary tree of depth $\lceil \log_2 m \rceil$ computes each $\widehat{W}_j$ exactly with $m - 1$ gates, and $h_j = \mathrm{ReLU}(1 - \widehat{W}_j/(2\varepsilon'))$ is one gate; the sum S is linear. For the remaining operations, recall $H(t) = \max\{S(t), 1/4\}$ and set $U_* := 2 + N'(1 + \Delta/\varepsilon')$, so that $1/4 \le H(t) \le U_*$ on $[-\Delta, \Delta]^m$. Let $\widetilde{R}$ be a ReLU network approximating $1/s$ on $[1/4, U_*]$ to accuracy δ_g, with its values clipped to $[0, 4]$. Use ReLU multiplication networks with error at most δ_g on $[0, U_*] \times [0, 4]$ for the inner multiplication and on $[-1, 4U_* + 1] \times [-L, L]$ for the outer multiplication below, componentwise. Using Assumption 1 below, choose ReLU networks $\tilde{\gamma}_j$, clipped to $[-L, L]^n$, with $\|\tilde{\gamma}_j - \gamma_j\|_{L^\infty} \le \varepsilon/8$. The parameter bounds for these networks are proved in Proposition 10. With the exact clipping $\mathrm{clip}_L(u) = \min(L, \max(-L, u))$ componentwise, define

$$\tilde{g}(\xi, t) := \Big(x,\ \mathrm{clip}_L\Big(\sum_{j=0}^{N'} \widetilde{\times}\big(\widetilde{\times}\big(h_j(t), \widetilde{R}(H(t))\big),\ \tilde{\gamma}_j(\xi)\big)\Big)\Big). \tag{202}$$

Choosing $\delta_g = \min\{1, \varepsilon/(C_0 N' U_*(1 + L))\}$ with C_0 sufficiently large bounds the combined reciprocal and multiplication errors by $\varepsilon/8$: the reciprocal error is multiplied by at most U_* and a component bounded by L, and the errors are summed over $N' + 1$ terms and n components. Replacing the functions γ_j contributes at most $\varepsilon/8$ since the weights sum to one; clipping only decreases the error. Hence

$$\sup_{\xi, t} |\tilde{g}(\xi, t) - g(\xi, t)| \le \tfrac{\varepsilon}{4}. \tag{203}$$

Assumption 1 (Token regularity of the slices)**.** $\sup_\mu [\Gamma(\mu, \cdot)]_{C^{0,\alpha}(\Xi)} \le M_\xi$, *i.e.* $\Gamma \in \mathcal{G}_\alpha^{\mathrm{joint}}(\Lambda, M_\xi)$.

This is an additional assumption to (194), which controls only the measure variable. The token Hölder bound provides quantitative network approximations of $\xi \mapsto \Gamma(\mu_j, \xi)$ for each fixed measure μ_j. The conditional lower bound in Theorem 4 already holds for maps whose second component is independent of the token.

L.4.2 The transformer layers, with explicit weights

The embedding space is $\mathbb{R}^{D+m+n}$ with coordinates $e = (\xi; u_1, \dots, u_m; w)$; the lift is $\iota(\xi) = (\xi; 0_m; 0_n)$. Fix ReLU networks $\tilde{\varphi}_i$ with $\|\tilde{\varphi}_i - \varphi_i\|_\infty \le \varepsilon'/4$, clipped to $[-\Delta, \Delta]$ (size at most $C(1 + \Delta/\varepsilon')^{2D}$ each, as proved below). The realized map is $T = F_{\text{out}} \diamond \Gamma_\theta \diamond F_{\text{enc}}$ in the block notation of Furuya et al. [2025b]:

$$F_{\text{enc}}(e) := \big(\xi;\ \tilde{\varphi}_1(\xi), \dots, \tilde{\varphi}_m(\xi);\ 0_n\big) \qquad \text{(token-wise MLP)}, \tag{204}$$

$$Q^i := 0, \quad K^i := 0, \quad V^i := e_{D+i}^\top, \quad W^i := e_{D+i} \qquad (i \le m) \qquad (m \text{ heads}, d_{\text{head}} = 1), \tag{205}$$

$$\Gamma_\theta(\hat{\mu})(e) = e + \sum_{i \le m} W^i \int V^i e'\, d\hat{\mu}(e') = \Big(\xi;\ u_i + \int u_i'\, d\hat{\mu}\ (i \le m);\ w\Big). \tag{206}$$

Setting the query and key matrices to zero makes every attention score zero. For N equally weighted tokens, the softmax therefore assigns weight $1/N$ to each token; for a general measure, the layer averages with respect to that measure. Applied after F_{enc} to $\hat{\mu} = (F_{\text{enc}} \circ \iota)_\# \mu$, it adds the mean of $\tilde{\varphi}_i$ to the coordinate that already contains $\tilde{\varphi}_i(\xi)$, giving exactly

$$u_i = \tilde{\varphi}_i(\xi) + \tilde{t}_i(\mu), \qquad \tilde{t}_i(\mu) := \int_\Xi \tilde{\varphi}_i\, d\mu. \tag{207}$$

To recover the mean $\tilde{t}_i(\mu)$, the following feedforward network subtracts $\tilde{\varphi}_i(\xi)$, which it can evaluate from the retained token ξ. Define this network and the final linear projection by

$$\hat{g}(\xi, u) := \tilde{g}\big(\xi,\ u - \tilde{\varphi}(\xi)\big), \qquad F_{\text{out}}(e) := \big(\xi; u; \hat{g}_2(\xi, u)\big), \qquad \pi(e) := (x, w), \tag{208}$$

where the subtraction is realized by a second copy of the encoder subnetworks prepended to $\tilde{g}$ (a factor ≤ 2 on P_{enc}, absorbed below). The realized context map is then, in closed form,

$$\Gamma_T(\mu, \xi) := \pi\Big(F_{\text{out}}\big(\Gamma_\theta\big((F_{\text{enc}} \circ \iota)_\# \mu\big)(F_{\text{enc}}(\iota(\xi)))\big)\Big) = \big(x,\ \tilde{g}_2(\xi, \tilde{t}(\mu))\big). \tag{209}$$

Proposition 10 (Architecture bound for a cylindrical approximation)**.** *Let $\Gamma \in \mathcal{G}_\alpha^{\text{joint}}(\Lambda, M_\xi)$ preserve the spatial coordinate, and let*

$$0 < \varepsilon \le \min\{1, 2\Lambda\Delta^\alpha\}.$$

Set

$$a := \frac{1}{3}\left(\frac{\varepsilon}{2\Lambda}\right)^{1/\alpha}.$$

There exist integers $m, N' \ge 1$, satisfying

$$\log m + \log N' \le C\left(1 + \frac{\Delta}{a}\right)^D \log\left(C\left(1 + \frac{\Delta}{a}\right)\right),$$

and a transformer Γ_T consisting of a token-wise ReLU encoder, averaging by attention with uniform weights on token sequences (implemented by setting all scores to zero), and a token-wise ReLU network evaluating the output such that

$$\sup_{\mu \in \mathcal{P}(\Xi)} \sup_{\xi \in \Xi} |\Gamma_T(\mu, \xi) - \Gamma(\mu, \xi)| \le \varepsilon.$$

The transformer preserves the spatial coordinate and satisfies

$$\sup_{\mu \in \mathcal{P}(\Xi)} W_1\big(G_{\Gamma_T}(\mu), G_\Gamma(\mu)\big) \le \varepsilon.$$

Network size below means the number of nonzero weights and biases. The construction has the following resource bounds, where C depends only on the fixed geometric data, D, n, α, and the choice of norms.

1. Attention. *The total number of scalar heads is m, each computing one average. They may be placed in one layer or applied successively in m layers with one head each. The embedding dimension is at most $D + m + n$. The attention matrices have at most Cm nonzero entries; counting zero entries as well, they contain at most $4m(D + m + n)$ entries.*

2. Encoder. *The total number of nonzero weights and biases in the encoder and its second copy used to subtract $\widetilde{\varphi}_i(\xi)$ after attention satisfies*

$$P_{\text{enc}} \le Cm\left(1+\frac{\Delta}{a}\right)^{2D}.$$

3. Network approximating g. *Set*

$$U := 2 + N'\left(1+\frac{\Delta}{a}\right).$$

The network approximating g may be chosen with size

$$P_{\text{out}} \le C\left[N'm + N'n\left(1+\frac{M_\xi+\Delta+L+1}{\varepsilon}\right)^{2D/\alpha} + \left(\frac{N'U(1+L)}{\varepsilon}\right)^4\right].$$

4. Total size. *For fixed $\Lambda > 0$ and M_ξ, there is a constant C independent of ε such that*

$$\log\bigl(1+P_{\text{tot}}\bigr) \le C\left(\frac{\Lambda}{\varepsilon}\right)^{D/\alpha}\log\frac{C}{\varepsilon}.$$

Proof. Apply Theorem 3 with target accuracy $\varepsilon/2$. Let $\varphi_1,\dots,\varphi_m$ and $\mu_1,\dots,\mu_{N'}$ be the test functions and representative measures used in its construction.

Write

$$\Gamma(\mu,(x,y)) = (x,\Gamma_2(\mu,(x,y))), \qquad t_i(\mu) := \int_\Xi \varphi_i\,d\mu, \qquad c_{ij} := t_i(\mu_j).$$

Approximate statistics and attention realization. Choose scalar ReLU networks $\widetilde{\varphi}_i$ such that

$$\sup_{\xi\in\Xi}|\widetilde{\varphi}_i(\xi)-\varphi_i(\xi)| \le a/4, \qquad \widetilde{\varphi}_i(\Xi)\subset[-\Delta,\Delta].$$

The range condition is enforced by exact ReLU clipping, which does not increase the approximation error. Define

$$\widetilde{t}_i(\mu) := \int_\Xi \widetilde{\varphi}_i\,d\mu.$$

Since μ is a probability measure,

$$\|\widetilde{t}(\mu)-t(\mu)\|_{\ell^\infty} \le a/4.$$

The encoder places ξ, the values $\widetilde{\varphi}_i(\xi)$, and zero output coordinates in an embedding of dimension $D+m+n$. Head i has zero query and key matrices and computes the mean $\widetilde{t}_i(\mu)$ of $\widetilde{\varphi}_i$. The attention block adds this mean to the coordinate that contains $\widetilde{\varphi}_i(\xi)$, so that coordinate becomes

$$\widetilde{\varphi}_i(\xi)+\widetilde{t}_i(\mu).$$

A second copy of the encoder functions in the output network subtracts $\widetilde{\varphi}_i(\xi)$, leaving $\widetilde{t}_i(\mu)$ exactly.

Each head computes the mean of a different function $\widetilde{\varphi}_i$ and changes only the coordinate holding that function's value. Applying the heads successively therefore gives the same result as applying them simultaneously.

Stable interpolation weights. For $t\in[-\Delta,\Delta]^m$, define

$$\widehat{W}_j(t) := \max_{i\le m}(t_i-c_{ij}), \qquad h_j(t) := \left(1-\frac{\widehat{W}_j(t)}{2a}\right)_+,$$

and set

$$S(t) := \sum_{j=1}^{N'} h_j(t), \qquad h_0(t) := \left(\frac{1}{4}-S(t)\right)_+, \qquad H(t) := S(t)+h_0(t) = \max\{S(t),1/4\}.$$

Let

$$\psi_j(t) := \frac{h_j(t)}{H(t)} \quad (j = 0, \dots, N'), \qquad \mu_0 := \mu_1, \qquad \gamma_j(\xi) := \Gamma_2(\mu_j, \xi).$$

Then

$$g((x, y), t) = \left(x, \sum_{j=0}^{N'} \psi_j(t)\gamma_j(x, y) \right)$$

is defined on the entire statistics cube. The weights are nonnegative and sum to one.

By the test-function net construction,

$$W_1(\mu, \mu_j) - a/2 \le \widehat{W}_j(t(\mu)) \le W_1(\mu, \mu_j).$$

Since $\widehat{W}_j$ is 1-Lipschitz for ℓ^∞, the approximate statistics satisfy

$$W_1(\mu, \mu_j) - 3a/4 \le \widehat{W}_j(\widetilde{t}(\mu)) \le W_1(\mu, \mu_j) + a/4.$$

Some representative measure satisfies $W_1(\mu, \mu_j) \le a$. For this index, $h_j(\widetilde{t}(\mu)) \ge 3/8$. Thus

$$S(\widetilde{t}(\mu)) \ge 3/8, \qquad \psi_0(\widetilde{t}(\mu)) = 0.$$

Moreover, if $\psi_j(\widetilde{t}(\mu)) > 0$ for $j \ge 1$, then

$$W_1(\mu, \mu_j) < 11a/4.$$

The measure-variable Hölder bound therefore gives

$$\begin{aligned} |g(\xi, \widetilde{t}(\mu)) - \Gamma(\mu, \xi)| &\le \sum_{j=1}^{N'} \psi_j(\widetilde{t}(\mu)) |\Gamma(\mu_j, \xi) - \Gamma(\mu, \xi)| \\ &\le \Lambda(11a/4)^\alpha \\ &\le \Lambda(3a)^\alpha = \varepsilon/2. \end{aligned}$$

ReLU approximation of the interpolation formula. The maximum, positive-part, and clipping operations can be represented exactly by ReLU networks. Choose vector-valued ReLU networks $\widetilde{\gamma}_j$, clipped componentwise to $[-L, L]$, such that

$$\sup_{\xi \in \Xi} |\widetilde{\gamma}_j(\xi) - \gamma_j(\xi)| \le \varepsilon/8.$$

Replacing all slices by these networks changes their weighted sum by at most $\varepsilon/8$.

On the statistics cube,

$$0 \le h_j(t) \le 1 + \Delta/a, \qquad 1/4 \le H(t) \le U.$$

Approximate $s \mapsto 1/s$ on $[1/4, U]$ and all multiplications in

$$\sum_{j=0}^{N'} \frac{h_j(t)}{H(t)} \widetilde{\gamma}_j(\xi)$$

by ReLU networks. Taking their individual errors at most

$$\delta := \frac{\varepsilon}{CN'U(1 + L)}$$

with a sufficiently large constant C ensures that their combined Euclidean output error is at most $\varepsilon/8$. Indeed, an error δ in the reciprocal is multiplied by at most U and then by a bounded slice value; the errors are summed over $N' + 1$ terms and n components.

Finally, clip the output componentwise to $[-L, L]^n$, which does not increase its distance from the exact convex combination. The resulting network $\widetilde{g}$ preserves x and satisfies

$$\sup_{\xi, t} |\widetilde{g}(\xi, t) - g(\xi, t)| \le \varepsilon/4.$$

The transformer computes $\Gamma_T(\mu,\xi) = \widetilde{g}(\xi, \widetilde{t}(\mu))$. Consequently,

$$|\Gamma_T(\mu,\xi) - \Gamma(\mu,\xi)| \le \varepsilon/4 + \varepsilon/2 \le \varepsilon.$$

For the push forward estimate, use the coupling obtained by applying both maps to the same $\xi \sim \mu$:

$$W_1(G_{\Gamma_T}(\mu), G_\Gamma(\mu)) \le \int_\Xi |\Gamma_T(\mu,\xi) - \Gamma(\mu,\xi)|\, d\mu(\xi) \le \varepsilon.$$

Resource bounds. For completeness, conservative ReLU size bounds suffice. A scalar Lipschitz function on a fixed-dimensional compact set can first be extended to an enclosing box and then approximated by continuous piecewise-affine interpolation on a simplicial grid. At accuracy $a/4$, each test function needs at most $C(1+\Delta/a)^D$ affine pieces. A max–min representation of a continuous piecewise-affine function with J pieces yields a ReLU realization with at most CJ^2 nonzero parameters in fixed dimension. This gives the encoder bound, including its duplicated copy.

The same construction for each Hölder slice uses at most

$$C\left(1 + \frac{M_\xi + \Delta + L + 1}{\varepsilon}\right)^{D/\alpha}$$

affine pieces. Applying the piecewise-affine realization bound to its n components and to the $N'+1$ slices gives the second term in the output-network estimate. Computing the maxima $\widehat{W}_j$ and positive-part functions h_j, h_0 requires at most $CN'm$ nonzero parameters.

The reciprocal can be approximated by univariate piecewise-linear interpolation with at most $C(1 + U/\sqrt{\delta})$ parameters, because its second derivative is bounded on $[1/4, U]$. Multiplication is reduced to squaring through

$$uv = \frac{1}{4}\big((u+v)^2 - (u-v)^2\big).$$

Piecewise-linear approximation of the square on the required bounded intervals gives product networks with at most $C(1+U(1+L)/\sqrt{\delta})$ parameters each. There are at most $CN'n$ product networks. Substituting the chosen δ and using $\varepsilon \le 1$ yields the conservative bound

$$C\left(\frac{N'U(1+L)}{\varepsilon}\right)^4$$

for these arithmetic subnetworks. Exact clipping and linear operations are absorbed into the displayed bounds.

Each scalar attention head has four vectors of length at most $D+m+n$, counting zero entries as well. Its query and key vectors vanish, while its value and output vectors have only a constant number of nonzero entries. This proves the attention bounds.

Finally, the entropy estimates give, after adjusting constants,

$$\log m + \log N' \le C\left(1 + \frac{\Delta}{a}\right)^D \log\left(C\left(1 + \frac{\Delta}{a}\right)\right).$$

All displayed resource bounds are polynomial in $m, N', 1/a, 1/\varepsilon$, for fixed M_ξ and the other data. Taking logarithms and substituting $a = \frac{1}{3}(\varepsilon/(2\Lambda))^{1/\alpha}$ therefore proves the total-size estimate. □

Remark 7 (Averaging)**.** *All attention scores in this construction are zero. Thus attention is used only to compute averages of the functions evaluated at each token, and the approximation is realized by pointwise ReLU networks combined with averaging.*

Remark 8 (Embedding dimension)**.** *The construction stores all m statistics simultaneously, so its embedding dimension grows with m. The qualitative universality result of [Furuya et al., 2025b, Thm. 1] uses an embedding dimension independent of the approximation accuracy. The present argument does not establish the same quantitative size bound under that fixed-dimension constraint.*

L.5 The operator assumption and the context function Γ_η

L.5.1 The Hölder assumption on A

To obtain quantitative rates, we impose the following additional regularity and growth assumptions on A.

Assumption 2 (Quantitative hypotheses on A)**.** *Let $\alpha, \beta \in (0,1]$, $k \in \mathbb{Z}_{\geq 0}$, $s > 0$, $0 < \delta \leq s$, $s' \geq 0$, and $B_0 > 0$. Define*

$$\mathcal{B} := \left\{ h \in H^{-s+\delta}(\Omega;\mathbb{R}^n) \cap L^\infty(\Omega;[-L,L]^n) : \|h\|_{H^{-s+\delta}} \leq B_0 \right\},$$

and

$$\mathcal{K}_{\mathcal{B}} := \overline{\mathcal{B} \cup \{R_\tau h : \ h \in \mathcal{B},\ 0 < \tau \leq 1\}}^{H^{-s}(\Omega;\mathbb{R}^n)}.$$

For $\rho > 0$, set

$$B_k(\rho) := \left\{ u \in C^k(\overline{\Omega};[-L_1,L_1]^n) : \|u\|_{C^k} \leq \rho \right\}.$$

Assume that there exist constants $M_A(\rho), [A]_{\alpha,\rho} \geq 0$ such that:

1. *For every $u \in B_k(\rho)$,*
$$A(u) \in C^{0,\alpha}(\overline{\Omega};[-L,L]^n), \qquad \|A(u)\|_{C^{0,\alpha}} \leq M_A(\rho).$$
2. *For every $u, w \in B_k(\rho)$,*
$$\|A(u) - A(w)\|_{L^\infty} \leq [A]_{\alpha,\rho} \|u - w\|_{C^k}^\alpha.$$

These constants satisfy, for some $C_A > 0$ and $p \geq 0$,

$$M_A(\rho) + [A]_{\alpha,\rho} \leq C_A(1+\rho)^p \qquad (\rho > 0).$$

For the end-to-end approximation result, assume additionally that A is defined on $\mathcal{K}_{\mathcal{B}}$ and that, for some $C_\omega > 0$,

$$\|A(h_1) - A(h_2)\|_{H^{-s'}(\Omega)} \leq C_\omega \|h_1 - h_2\|_{H^{-s}(\Omega)}^\beta \qquad (h_1, h_2 \in \mathcal{K}_{\mathcal{B}}). \tag{210}$$

L.5.2 Membership in the joint class, with explicit constants

Define the η-dependent constants

$$\begin{aligned} \rho_\eta^S :=& V_0 L\sqrt{n} \sum_{j \leq K(\eta)} \tfrac{\|\phi_j\|_{C^0}\|\phi_j\|_{C^k}}{1+\eta\lambda_j^r}, \quad C_\Sigma(\eta) := \sum_{j \leq K(\eta)} \tfrac{\|\phi_j\|_{C^k}}{1+\eta\lambda_j^r}, \\ C_k^W(\eta) :=& \sqrt{n} V_0 C_\Omega (1+L) \sum_{j \leq K(\eta)} \tfrac{\|\phi_j\|_{C^k}\|\phi_j\|_{C^1}}{1+\eta\lambda_j^r}, \end{aligned} \tag{211}$$

where $C_\Omega \geq 1$ satisfies

$$\mathrm{Lip}_{\overline{\Omega}}(f) \leq C_\Omega \|f\|_{C^1(\overline{\Omega})} \qquad (f \in C^1(\overline{\Omega})).$$

and

$$\rho_\eta := c_k \|\Psi\|_{C_b^k} (1 + \rho_\eta^S)^k.$$

For $0 \leq m \leq k$, define $C_m^W(\eta)$ by the formula for $C_k^W(\eta)$ with $\|\phi_j\|_{C^k}$ replaced by $\|\phi_j\|_{C^m}$.

Lemma 21 (Nemytskii bounds)**.** *Let $k \in \mathbb{Z}_{\geq 0}$ and $\Psi \in C_b^{k+1}(\mathbb{R}^n;\mathbb{R}^n)$. There exists a constant $c_k \geq 1$, depending only on k, d, n and the conventions for the C^k norms, such that, for every $\rho \geq 0$ and $v, w \in C^k(\overline{\Omega};\mathbb{R}^n)$ satisfying*

$$\|v\|_{C^k}, \|w\|_{C^k} \leq \rho,$$

we have

$$\|\Psi \circ v\|_{C^k} \leq c_k \|\Psi\|_{C_b^k} (1+\rho)^k$$

and

$$\|\Psi \circ v - \Psi \circ w\|_{C^k} \leq \Lambda_{\Psi,k}(\rho) \|v - w\|_{C^k},$$

where

$$\Lambda_{\Psi,k}(\rho) := (k+1) c_k \|\Psi\|_{C_b^{k+1}} (1+\rho)^k.$$

Here the C_b^j norm includes the supremum norms of the function and its derivatives through order j.

Proof. For $k = 0$, the first estimate follows from the boundedness of Ψ. The second follows from

$$\Psi(v(x)) - \Psi(w(x)) = \int_0^1 D\Psi\big(w(x) + t(v(x) - w(x))\big)(v(x) - w(x))\, dt.$$

The constants associated with the chosen norms are absorbed into c_0.

Now let $k \geq 1$. For a multi-index β with $1 \leq |\beta| \leq k$, the multivariate chain rule expresses $\partial^\beta(\Psi \circ v)$ as a finite sum of terms of the form

$$C_{\beta_1,\dots,\beta_\ell} D^\ell \Psi(v)\big[\partial^{\beta_1} v, \dots, \partial^{\beta_\ell} v\big],$$

where

$$1 \leq \ell \leq |\beta|, \qquad |\beta_i| \geq 1, \qquad \beta_1 + \cdots + \beta_\ell = \beta.$$

Each term is bounded in supremum norm by

$$C\|\Psi\|_{C_b^k} \rho^\ell \leq C\|\Psi\|_{C_b^k}(1 + \rho)^k.$$

Including the zeroth-order term and summing over the finitely many derivatives proves the first estimate.

For the difference estimate, subtract the corresponding expressions for v and w. First, changing the coefficient tensor gives

$$\|D^\ell \Psi(v) - D^\ell \Psi(w)\|_\infty \leq C\|\Psi\|_{C_b^{k+1}} \|v - w\|_{C^0}.$$

Its contribution, after multiplication by the ℓ derivative factors, is bounded by

$$C\|\Psi\|_{C_b^{k+1}}(1 + \rho)^k \|v - w\|_{C^k}.$$

Next, telescope the derivative factors one at a time. Each replacement contributes at most

$$C\|\Psi\|_{C_b^k} \rho^{\ell-1} \|v - w\|_{C^k} \leq C\|\Psi\|_{C_b^{k+1}}(1 + \rho)^k \|v - w\|_{C^k}.$$

There are at most $\ell + 1 \leq k + 1$ contributions per chain-rule term. The zeroth-order difference is controlled by the mean value formula above. After enlarging c_k to absorb the chain-rule coefficients and norm constants, the second estimate follows. □

Lemma 22 (W_1-Lipschitz bound for the filter)**.** *For every $\mu, \nu \in \mathcal{P}(\Xi)$ and every integer $0 \leq m \leq k$,*

$$\|S_\eta(\mu) - S_\eta(\nu)\|_{C^m} \leq C_m^W(\eta) W_1(\mu, \nu).$$

Moreover,

$$\sup_{\mu \in \mathcal{P}(\Xi)} \|S_\eta(\mu)\|_{C^k} \leq \rho_\eta^S.$$

Proof. Fix $j \leq K(\eta)$ and $i \in \{1, \dots, n\}$. For $\xi = (x, y)$ and $\xi' = (x', y')$ in Ξ,

$$\begin{aligned} |\phi_j(x)y_i - \phi_j(x')y_i'| &\leq L|\phi_j(x) - \phi_j(x')| + \|\phi_j\|_{C^0}|y_i - y_i'| \\ &\leq C_\Omega \|\phi_j\|_{C^1}\big(L|x - x'| + |y - y'|\big) \\ &\leq C_\Omega(1 + L)\|\phi_j\|_{C^1}|\xi - \xi'|. \end{aligned}$$

Thus the scalar test function $\xi \mapsto \phi_j(x)y_i$ has Lipschitz constant at most $C_\Omega(1 + L)\|\phi_j\|_{C^1}$. Kantorovich–Rubinstein duality gives

$$|b_{j,i}(\mu) - b_{j,i}(\nu)| \leq V_0 C_\Omega(1 + L)\|\phi_j\|_{C^1} W_1(\mu, \nu).$$

Combining the n components yields

$$|b_j(\mu) - b_j(\nu)| \leq \sqrt{n}\, V_0 C_\Omega(1 + L)\|\phi_j\|_{C^1} W_1(\mu, \nu).$$

Using the spectral formula and the triangle inequality, we obtain

$$\begin{aligned} \|S_\eta(\mu) - S_\eta(\nu)\|_{C^m} &\leq \sum_{j \leq K(\eta)} \frac{|b_j(\mu) - b_j(\nu)|\|\phi_j\|_{C^m}}{1 + \eta\lambda_j^r} \\ &\leq C_m^W(\eta) W_1(\mu, \nu). \end{aligned}$$

Finally, since $|y| \le L\sqrt{n}$ on Ξ and μ is a probability measure,

$$|b_j(\mu)| \le V_0 \int_\Xi |\phi_j(x)|\,|y|\,d\mu(x,y) \le V_0 L\sqrt{n}\,\|\phi_j\|_{C^0}.$$

Substituting this estimate into the spectral formula gives

$$\|S_\eta(\mu)\|_{C^k} \le V_0 L\sqrt{n} \sum_{j \le K(\eta)} \frac{\|\phi_j\|_{C^0}\|\phi_j\|_{C^k}}{1+\eta\lambda_j^r} = \rho_\eta^S.$$

Theorem 5 (Quantitative regularity of the context map)**.** *Assume the smooth-input bounds of Assumption 2. Define*

$$\Lambda_\eta := [A]_{\alpha,\rho_\eta}\Lambda_{\Psi,k}(\rho_\eta^S)^\alpha C_k^W(\eta)^\alpha, \qquad M_{\xi,\eta} := M_A(\rho_\eta) + \operatorname{diam}(\overline{\Omega})^{1-\alpha}.$$

Then

$$\Gamma_\eta \in \mathcal{G}_\alpha^{\mathrm{joint}}(\Lambda_\eta, M_{\xi,\eta}).$$

Moreover, for every $\mu,\nu \in \mathcal{P}(\Xi)$,

$$W_1\big(G_{\Gamma_\eta}(\mu), G_{\Gamma_\eta}(\nu)\big) \le (\Lambda_\eta + M_{\xi,\eta}) W_1(\mu,\nu)^\alpha.$$

With the polynomial growth bounds of Assumption 2, the Weyl estimate (189), and the eigenfunction estimate (190), we have

$$\Lambda_\eta \le CK(\eta)^{q_K}, \qquad M_{\xi,\eta} \le CK(\eta)^{q_K'}, \tag{212}$$

where

$$q_K = \alpha\left(1 + \frac{s(k)+s(1)}{d}\right) + (\alpha k + kp)\left(1 + \frac{s(k)+s(0)}{d}\right),$$
$$q_K' = kp\left(1 + \frac{s(k)+s(0)}{d}\right).$$

The constant C is independent of η.

Proof. *Measure-variable regularity.* Write

$$u_\mu := \widetilde{S}_\eta(\mu) = \Psi \circ S_\eta(\mu).$$

The filter bound and the Nemytskii estimate give

$$\|u_\mu\|_{C^k} \le \rho_\eta,$$

so $u_\mu \in B_k(\rho_\eta)$ for every μ. They also give

$$\begin{aligned}\|u_\mu - u_\nu\|_{C^k} &\le \Lambda_{\Psi,k}(\rho_\eta^S)\|S_\eta(\mu) - S_\eta(\nu)\|_{C^k} \\ &\le \Lambda_{\Psi,k}(\rho_\eta^S) C_k^W(\eta) W_1(\mu,\nu).\end{aligned}$$

Since

$$\Gamma_\eta(\mu,(x,y)) = (x, A(u_\mu)(x)),$$

the spatial components coincide when the token is fixed. The smooth-input Hölder bound on A therefore yields

$$\begin{aligned}\sup_{\xi\in\Xi}|\Gamma_\eta(\mu,\xi) - \Gamma_\eta(\nu,\xi)| &\le \|A(u_\mu) - A(u_\nu)\|_{L^\infty} \\ &\le [A]_{\alpha,\rho_\eta}\|u_\mu - u_\nu\|_{C^k}^\alpha \\ &\le \Lambda_\eta W_1(\mu,\nu)^\alpha.\end{aligned}$$

Token-variable regularity. Fix μ and take $\xi = (x,y)$ and $\xi' = (x',y')$. The map $\Gamma_\eta(\mu,\cdot)$ is independent of y. Since $\|A(u_\mu)\|_{C^{0,\alpha}} \le M_A(\rho_\eta)$,

$$\begin{aligned}|\Gamma_\eta(\mu,\xi) - \Gamma_\eta(\mu,\xi')| &\le |x-x'| + |A(u_\mu)(x) - A(u_\mu)(x')| \\ &\le \big(\operatorname{diam}(\overline{\Omega})^{1-\alpha} + M_A(\rho_\eta)\big)\,|x-x'|^\alpha \\ &\le M_{\xi,\eta}|\xi-\xi'|^\alpha.\end{aligned}$$

This proves the asserted membership.

Push forward estimate. Let π be an optimal coupling of μ and ν. Applying the two context maps to the respective components of this coupling gives

$$\begin{aligned} &W_1\big(G_{\Gamma_\eta}(\mu), G_{\Gamma_\eta}(\nu)\big) \\ &\quad\le \int_{\Xi\times\Xi} |\Gamma_\eta(\mu,\xi) - \Gamma_\eta(\nu,\xi')|\, d\pi(\xi,\xi') \\ &\quad\le \Lambda_\eta W_1(\mu,\nu)^\alpha + M_{\xi,\eta} \int_{\Xi\times\Xi} |\xi-\xi'|^\alpha\, d\pi(\xi,\xi'). \end{aligned}$$

By Jensen's inequality and the optimality of π,

$$\int |\xi-\xi'|^\alpha\, d\pi \le \left(\int |\xi-\xi'|\, d\pi\right)^\alpha = W_1(\mu,\nu)^\alpha.$$

This proves the push forward bound.

Growth of the constants. Write $K := K(\eta)$ and set

$$\sigma_0 := 1 + \frac{s(k)+s(0)}{d}, \qquad \sigma_W := 1 + \frac{s(k)+s(1)}{d}.$$

The Weyl and eigenfunction estimates imply

$$\|\phi_j\|_{C^m} \le C j^{s(m)/d}.$$

Since $(1+\eta\lambda_j^r)^{-1} \le 1$, summing these bounds gives

$$\rho_\eta^S \le CK^{\sigma_0}, \qquad C_k^W(\eta) \le CK^{\sigma_W}.$$

Consequently,

$$\rho_\eta = c_k \|\Psi\|_{C_b^k}(1+\rho_\eta^S)^k \le CK^{k\sigma_0}$$

and

$$\Lambda_{\Psi,k}(\rho_\eta^S) \le CK^{k\sigma_0}.$$

These estimates also hold for $k=0$, with the corresponding powers equal to zero. The polynomial bounds on A yield

$$[A]_{\alpha,\rho_\eta} + M_A(\rho_\eta) \le C(1+\rho_\eta)^p \le CK^{kp\sigma_0}.$$

It follows that

$$\begin{aligned} \Lambda_\eta &\le CK^{kp\sigma_0} K^{\alpha k\sigma_0} K^{\alpha\sigma_W} \\ &= CK^{\alpha\sigma_W + (\alpha k + kp)\sigma_0}, \end{aligned}$$

and, since $K \ge 1$,

$$M_{\xi,\eta} \le CK^{kp\sigma_0}.$$

These are precisely the exponents in (212). □

L.5.3 Realization of Γ_η: polynomial resources at fixed η

The context map Γ_η depends on μ only through the nK scalar components of

$$b_j(\mu) = V_0 \int_\Xi \phi_j(x) y\, d\mu(x,y), \qquad j = 1,\dots,K,$$

where $K = K(\eta)$. Set

$$r_j := V_0 L \|\phi_j\|_{C^0}, \qquad \mathcal{B}_\eta := \prod_{j=1}^K [-r_j, r_j]^n, \qquad r_{\max} := \max_{j\le K} r_j.$$

Then $b(\mu) \in \mathcal{B}_\eta$ and $r_{\max} \le CK^{s(0)/d}$. Define

$$G_\eta(b,x) := A\left(\Psi\left(\sum_{j=1}^K \frac{b_j}{1+\eta\lambda_j^r}\phi_j\right)\right)(x), \qquad (b,x) \in \mathcal{B}_\eta \times \overline{\Omega}. \tag{213}$$

Thus

$$\Gamma_\eta(\mu,(x,y)) = \big(x, G_\eta(b(\mu),x)\big).$$

This finite-statistic representation is exact. Consequently, no finite coverings of the measure space or the class of test functions, and no partition of unity on the measure space, are needed.

Proposition 11 (Quantitative realization of Γ_η)**.** *Assume the smooth-input bounds of Assumption 2. Let $\eta \in (0,1]$, $K = K(\eta)$, and $0 < \varepsilon_{\mathrm{T}} \le 1$. Equip the coefficient space $\mathbb{R}^{nK}$ with the componentwise ℓ^∞ norm, and define*

$$\widehat{Q}_\eta := [A]_{\alpha,\rho_\eta} \Lambda_{\Psi,k} (\rho_\eta^S)^\alpha \big(\sqrt{n}\, C_\Sigma(\eta)\big)^\alpha,$$

$$\delta_{\mathrm{stat}} := \min\left\{1, \left(\frac{\varepsilon_{\mathrm{T}}}{4(1+\widehat{Q}_\eta)}\right)^{1/\alpha}\right\}.$$

There exists a spatial-coordinate-preserving transformer $T_{\eta,\varepsilon_{\mathrm{T}}}$ such that

$$\sup_{\mu\in\mathcal{P}(\Xi)} \sup_{\xi\in\Xi} |T_{\eta,\varepsilon_{\mathrm{T}}}(\mu,\xi) - \Gamma_\eta(\mu,\xi)| \le \varepsilon_{\mathrm{T}},$$

and

$$\sup_{\mu\in\mathcal{P}(\Xi)} W_1\Big(G_{T_{\eta,\varepsilon_{\mathrm{T}}}}(\mu), G_{\Gamma_\eta}(\mu)\Big) \le \varepsilon_{\mathrm{T}}.$$

The construction has the following resource bounds. Size means the number of nonzero weights and biases; we also state the number of attention-matrix entries when zeros are included.

1. Attention. *There are nK scalar attention heads, each with zero query and key matrices so that it computes an average. They may be applied simultaneously or successively in nK layers, each containing one head. The embedding dimension is at most $D + nK + n$. The attention matrices have at most CnK nonzero entries; counting zero entries as well, they contain at most*

$$4nK(D + nK + n).$$

2. Encoder. *Including the second evaluation used to subtract the original feature values after attention, the encoder size satisfies*

$$P_{\mathrm{enc}} \le CnK\left(1 + \frac{CK^{s(1)/d}}{\delta_{\mathrm{stat}}}\right)^{2D}.$$

In particular, this bound is polynomial in K and $\varepsilon_{\mathrm{T}}^{-1}$.

3. Output network approximating G_η. *Set $q := nK + d$ and*

$$\Theta_\eta := C'_\Omega\left(1 + \widehat{Q}_\eta r_{\max}^\alpha + M_A(\rho_\eta)\right),$$

where $C'_\Omega \ge 1$ depends only on the fixed domain, α, and the coordinate normalization. There is a ReLU output network with parameter count P_G satisfying

$$\log(1 + P_G) \le Cq \log\left(\frac{Cq\Theta_\eta}{\varepsilon_{\mathrm{T}}}\right).$$

4. Total size. *With*

$$\sigma_0 := 1 + \frac{s(k) + s(0)}{d}, \qquad \widehat{q} := (\alpha + \alpha k + kp)\sigma_0,$$

we have

$$\Theta_\eta \le CK^{\widehat{q}}, \qquad \log(1 + P_{\mathrm{tot}}) \le CK \log\left(\frac{CK}{\varepsilon_{\mathrm{T}}}\right).$$

In particular, for $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$,

$$\log(1 + P_{\mathrm{tot}}) \le C\eta^{-d/(4r)} \log\left(\frac{C}{\eta\varepsilon_{\mathrm{T}}}\right).$$

The constants are independent of η and ε_{T}.

Proof. *Regularity of the outer function.* For every $b \in \mathcal{B}_\eta$, $|b_j| \le \sqrt{n}\, r_j$, and hence

$$\left\| \sum_{j=1}^{K} \frac{b_j}{1+\eta\lambda_j^r} \phi_j \right\|_{C^k} \le \rho_\eta^S.$$

The clipped reconstruction therefore belongs to $B_k(\rho_\eta)$.

For $b, b' \in \mathcal{B}_\eta$, the componentwise ℓ^∞ norm gives

$$\left\| \sum_{j=1}^{K} \frac{b_j - b'_j}{1+\eta\lambda_j^r} \phi_j \right\|_{C^k} \le \sqrt{n}\, C_\Sigma(\eta) \|b - b'\|_{\ell^\infty}.$$

The Nemytskii bound and the smooth-input Hölder bound on A yield

$$|G_\eta(b,x) - G_\eta(b',x)| \le \widehat{Q}_\eta \|b - b'\|_{\ell^\infty}^{\alpha}.$$

The spatial output regularity gives

$$|G_\eta(b,x) - G_\eta(b,x')| \le M_A(\rho_\eta)|x - x'|^\alpha.$$

After affine normalization of the coefficient box and a fixed enclosing spatial box, these estimates give a joint Hölder constant bounded by Θ_η. If necessary, extend each output component from $\mathcal{B}_\eta \times \overline{\Omega}$ to the enclosing box using the Hölder extension formula, and clip it to $[-L, L]$.

Statistics and attention. For $j \le K$ and $i \le n$, approximate

$$f_{j,i}(x,y) := V_0 \phi_j(x) y_i$$

by a ReLU network to accuracy δ_{stat}, and clip its output to $[-r_j, r_j]$. This clipping does not increase the approximation error. With zero query and key matrices, a scalar attention head computes the mean of each approximated feature with respect to μ. The residual connection also contributes the feature value at the current token; evaluating the same feature a second time and subtracting it leaves its mean. The resulting coefficient vector $\widetilde{b}(\mu)$ satisfies

$$\widetilde{b}(\mu) \in \mathcal{B}_\eta, \qquad \|\widetilde{b}(\mu) - b(\mu)\|_{\ell^\infty} \le \delta_{\mathrm{stat}}.$$

Thus

$$|G_\eta(\widetilde{b}(\mu), x) - G_\eta(b(\mu), x)| \le \widehat{Q}_\eta \delta_{\mathrm{stat}}^\alpha \le \varepsilon_{\mathrm{T}}/4.$$

The integrands have Lipschitz constants bounded by $CK^{s(1)/d}$. Each integrand has $D = d+n$ inputs, independently of K. Piecewise-affine interpolation in this dimension, followed by representation of the interpolants by ReLU networks, gives the stated encoder bound. Evaluating the same functions a second time for the subtraction above changes only its constant factor. The attention counts follow from assigning one scalar head to each coefficient component.

Approximation of the outer function. A tensor-product piecewise-linear partition of unity on the normalized q-dimensional box gives an interpolant with error at most $\varepsilon_{\mathrm{T}}/4$ when the mesh size is chosen sufficiently small. The number of grid vertices is bounded by

$$J \le \left(\frac{Cq\Theta_\eta}{\varepsilon_{\mathrm{T}}} \right)^{q/\alpha}.$$

Each partition function is a product of q one-dimensional piecewise-linear basis functions, one for each coordinate. These basis functions are one at their grid point and zero at the other grid points, with linear interpolation between adjacent grid points; they are often called hat functions. They can be represented exactly by ReLU networks, and their products can be approximated using ReLU multiplication networks. Choosing the individual multiplication errors sufficiently small gives an additional uniform Euclidean output error at most $\varepsilon_{\mathrm{T}}/4$. Componentwise clipping to $[-L, L]^n$ does not increase the error.

For example, piecewise-linear approximation of squaring, together with

$$uv = \frac{1}{4}\big((u+v)^2 - (u-v)^2\big),$$

provides multiplication networks whose size is polynomial in the inverse accuracy on bounded intervals. Approximating the J partition functions in this way therefore requires a number of parameters polynomial in J, q, and $\varepsilon_{\mathrm{T}}^{-1}$, with fixed polynomial degree. Consequently,

$$\log(1 + P_G) \leq Cq \log\left(\frac{Cq\Theta_\eta}{\varepsilon_{\mathrm{T}}}\right).$$

Combining the coefficient error and the output-network error gives

$$\sup_{\mu,x} \left|\widetilde{G}_\eta(\widetilde{b}(\mu), x) - G_\eta(b(\mu), x)\right| \leq 3\varepsilon_{\mathrm{T}}/4.$$

The transformer outputs $(x, \widetilde{G}_\eta(\widetilde{b}(\mu), x))$, retaining the spatial coordinate x. For the push forward bound, pair the transformer output and the exact output $\Gamma_\eta(\mu, \xi)$ for each token ξ drawn from μ; their mean distance is bounded by the uniform approximation error.

Growth estimates. The eigenfunction bounds imply

$$r_{\max} \leq CK^{s(0)/d}, \qquad C_\Sigma(\eta) \leq CK^{1+s(k)/d}, \qquad \rho_\eta^S \leq CK^{\sigma_0}.$$

Therefore,

$$\rho_\eta \leq CK^{k\sigma_0}, \qquad \Lambda_{\Psi,k}(\rho_\eta^S) \leq CK^{k\sigma_0}, \qquad [A]_{\alpha,\rho_\eta} + M_A(\rho_\eta) \leq CK^{kp\sigma_0}.$$

It follows that

$$\begin{aligned}\widehat{Q}_\eta r_{\max}^\alpha &\leq CK^{kp\sigma_0} K^{\alpha k\sigma_0} K^{\alpha(1+s(k)/d)} K^{\alpha s(0)/d} \\ &= CK^{(kp+\alpha k+\alpha)\sigma_0}.\end{aligned}$$

This proves the bound on Θ_η. The definition of δ_{stat} also shows that its inverse is polynomial in K and $\varepsilon_{\mathrm{T}}^{-1}$. Combining the attention, encoder, and output-network counts yields

$$\log(1 + P_{\mathrm{tot}}) \leq CK \log\left(\frac{CK}{\varepsilon_{\mathrm{T}}}\right).$$

Substitution of $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$ proves the final estimate. □

Corollary 3 (A one-parameter approximation schedule)**.** *Assume the hypotheses of Proposition 11. Fix $\theta > 0$ and, for $0 < t \leq 1$, set*

$$T_t := T_{t,t^\theta}.$$

Then

$$\sup_{\mu \in \mathcal{P}(\Xi)} W_1\big(G_{T_t}(\mu), G_{\Gamma_t}(\mu)\big) \leq t^\theta.$$

Consequently, if for some $\mu \in \mathcal{P}(\Xi)$ there exists $\nu_\mu \in \mathcal{P}(\Xi)$ such that

$$G_{\Gamma_\eta}(\mu) \longrightarrow \nu_\mu \quad \textit{in } W_1 \quad \textit{as } \eta \downarrow 0,$$

then

$$G_{T_t}(\mu) \longrightarrow \nu_\mu \quad \textit{in } W_1 \quad \textit{as } t \downarrow 0.$$

Moreover, the parameter count satisfies

$$\log\big(1 + P_{\mathrm{tot}}(T_t)\big) \leq Ct^{-d/(4r)} \left(1 + (1+\theta)\log\frac{1}{t}\right),$$

where C is independent of t and θ.

Proof. The approximation and parameter-count bounds follow from Proposition 11 with $\eta = t$ and $\varepsilon_{\mathrm{T}} = t^\theta$. For each μ satisfying the stated convergence assumption, the triangle inequality gives

$$W_1(G_{T_t}(\mu), \nu_\mu) \leq t^\theta + W_1(G_{\Gamma_t}(\mu), \nu_\mu) \longrightarrow 0.$$

□

L.6 Rates for the remaining approximation steps

L.6.1 Tokenization

Lemma 23 (Push forward under Hölder maps)**.** *Let $\alpha_h \in (0,1]$, and suppose that $g : \overline{\Omega} \to \Xi$ satisfies*

$$|g(x) - g(x')| \le C_g|x - x'| + H_g|x - x'|^{\alpha_h} \qquad (x, x' \in \overline{\Omega}),$$

for constants $C_g, H_g \ge 0$. Then, for every $\mu, \nu \in \mathcal{P}(\overline{\Omega})$,

$$W_1(g_\#\mu, g_\#\nu) \le C_g W_1(\mu,\nu) + H_g W_1(\mu,\nu)^{\alpha_h}.$$

Proof. Let π be an optimal coupling of μ and ν. Then $(g \times g)_\#\pi$ is a coupling of $g_\#\mu$ and $g_\#\nu$, so

$$\begin{aligned} W_1(g_\#\mu, g_\#\nu) &\le \int_{\overline{\Omega}\times\overline{\Omega}} |g(x) - g(x')|\, d\pi(x,x') \\ &\le C_g \int |x - x'|\, d\pi + H_g \int |x - x'|^{\alpha_h}\, d\pi. \end{aligned}$$

Since π is a probability measure, Jensen's inequality gives

$$\int |x - x'|^{\alpha_h}\, d\pi \le \left(\int |x - x'|\, d\pi\right)^{\alpha_h}.$$

The optimality of π proves the assertion.

Proposition 12 (Tokenization rate)**.** *Let $\alpha_h \in (0,1]$ and $h \in C^{0,\alpha_h}(\overline{\Omega}; [-L, L]^n)$. For sampling points $\bar{x}_1, \dots, \bar{x}_N$, define*

$$\mu_N := \frac{1}{N}\sum_{j=1}^N \delta_{\bar{x}_j}, \qquad \mathbf{M}_N(h) := \frac{1}{N}\sum_{j=1}^N \delta_{(\bar{x}_j, h(\bar{x}_j))}.$$

Then

$$W_1(\mathbf{M}_N(h), \gamma_h) \le W_1(\mu_N, \bar{\lambda}_\Omega) + [h]_{C^{0,\alpha_h}} W_1(\mu_N, \bar{\lambda}_\Omega)^{\alpha_h}.$$

If $d \ge 3$ and the sampling points are i.i.d. with law $\bar{\lambda}_\Omega$, then

$$\mathbb{E}\, W_1(\mathbf{M}_N(h), \gamma_h) \le C\big(1 + [h]_{C^{0,\alpha_h}}\big) N^{-\alpha_h/d},$$

where C depends only on Ω, d, α_h.

For deterministic sampling points satisfying

$$W_1(\mu_N, \bar{\lambda}_\Omega) \le C_0 N^{-1/d},$$

the same rate holds without expectation, with a constant depending additionally on C_0.

Proof. Define the graph map

$$g_h(x) := (x, h(x)).$$

Then

$$\mathbf{M}_N(h) = (g_h)_\#\mu_N, \qquad \gamma_h = (g_h)_\#\bar{\lambda}_\Omega,$$

and

$$|g_h(x) - g_h(x')| \le |x - x'| + [h]_{C^{0,\alpha_h}}|x - x'|^{\alpha_h}.$$

The first estimate follows from Lemma 23 with $C_g = 1$ and $H_g = [h]_{C^{0,\alpha_h}}$.

For i.i.d. sampling and $d \ge 3$, the empirical-measure estimate of Fournier and Guillin [2015], applied to the boundedly supported measure $\bar{\lambda}_\Omega$, gives

$$\mathbb{E}\, W_1(\mu_N, \bar{\lambda}_\Omega) \le C N^{-1/d}.$$

Taking expectations in the first estimate and using Jensen's inequality yields

$$\begin{aligned} \mathbb{E}\, W_1(\mathbf{M}_N(h), \gamma_h) &\le \mathbb{E}\, W_1(\mu_N, \bar{\lambda}_\Omega) \\ &\quad + [h]_{C^{0,\alpha_h}} \left(\mathbb{E}\, W_1(\mu_N, \bar{\lambda}_\Omega)\right)^{\alpha_h} \\ &\le C N^{-1/d} + C[h]_{C^{0,\alpha_h}} N^{-\alpha_h/d}. \end{aligned}$$

Since $N^{-1/d} \le N^{-\alpha_h/d}$, this proves the probabilistic rate. The deterministic assertion follows by substituting the assumed Wasserstein bound into the first estimate.

L.6.2 Mollification

Let $\rho \in C_c^\infty(B(0,1))$ be nonnegative and even, with $\int_{\mathbb{R}^d} \rho = 1$. For $0 < \tau \le 1$, define

$$\rho_\tau(x) := \tau^{-d}\rho(x/\tau), \qquad \Omega_\tau := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > 4\tau\},$$

and

$$\chi_\tau := \rho_\tau * \mathbf{1}_{\Omega_\tau}, \qquad R_\tau h := \rho_\tau * (\chi_\tau h). \tag{214}$$

Here $\mathbf{1}_{\Omega_\tau}$ is extended by zero to $\mathbb{R}^d$, and $\chi_\tau h$, which has compact support in Ω, is likewise extended by zero before convolution.

Proposition 13 (Mollification estimates)**.** *Let $s > 1/2$, $0 < \delta \le s$, and assume that*

$$0 \le s - \delta < \frac{1}{2}, \qquad s \notin \frac{1}{2} + \mathbb{Z}.$$

For every $h \in H^{-s+\delta}(\Omega;\mathbb{R}^n)$ and $0 < \tau \le 1$,

$$\begin{aligned}\|R_\tau h - h\|_{H^{-s}(\Omega)} &\le C\left(\tau^{\min\{\delta,2\}} + \tau^{\delta/(2s)}\right)\|h\|_{H^{-s+\delta}(\Omega)} \\ &\le C\tau^{\delta/(2s)}\|h\|_{H^{-s+\delta}(\Omega)}.\end{aligned}$$

For every fixed $\sigma \ge -s + \delta$, there exists $C_\sigma > 0$ such that

$$\|R_\tau h\|_{H^\sigma(\Omega)} \le C_\sigma \tau^{-(\sigma+s-\delta)}\|h\|_{H^{-s+\delta}(\Omega)}.$$

Consequently, for every integer $m \ge 0$,

$$\|R_\tau h\|_{C^m(\overline{\Omega})} \le C_m \tau^{-\kappa_m}\|h\|_{H^{-s+\delta}(\Omega)}, \qquad \kappa_m := s(m) + s - \delta.$$

All constants are independent of h and τ.

Moreover, $R_\tau h \in C_c^\infty(\Omega;\mathbb{R}^n)$. In particular, $R_\tau h$ belongs to the domain of every positive integer power of the Dirichlet realization of $I - \Delta$.

If additionally $h(x) \in [-L, L]^n$ almost everywhere, then

$$R_\tau h(x) \in [-L, L]^n \qquad (x \in \Omega).$$

Proof. Set $a := s - \delta$ and assume $0 \le a < 1/2$. We use the convention $H^{-b}(\Omega) = (H_0^b(\Omega))^*$ for $b \ge 0$, see Lions and Magenes [1972b]. Constants below are independent of h and τ.

We first recall the cutoff estimates needed in the proof. For $0 \le a < 1/2$, the zero extension

$$u_\tau := \operatorname{Ext}_0(\chi_\tau h)$$

satisfies

$$\|u_\tau\|_{H^{-a}(\mathbb{R}^d)} \le C\|h\|_{H^{-a}(\Omega)}. \tag{215}$$

Indeed, multiplication by χ_τ is uniformly bounded on $H^a(\Omega)$ in this range, and $H_0^a(\Omega) = H^a(\Omega)$ with equivalent norms, see Lions and Magenes [1972b]. Duality therefore gives (215).

We also use the boundary-cutoff estimate

$$\|(1 - \chi_\tau)\varphi\|_{H^s(\Omega)} \le C\|\varphi\|_{H^s(\Omega)} \qquad (\varphi \in H_0^s(\Omega)). \tag{216}$$

This estimate follows from the boundary Hardy estimates and the derivative bounds $|\partial^\beta \chi_\tau| \le C_\beta \tau^{-|\beta|}$. The exclusion $s \notin 1/2 + \mathbb{Z}$ avoids the critical trace indices.

Decompose

$$R_\tau h - h = \big(\rho_\tau * u_\tau - u_\tau\big)|_\Omega + (\chi_\tau - 1)h =: T_1 + T_2.$$

Estimate of T_1. Since ρ is even and has integral one,

$$|\widehat{\rho}(\zeta) - 1| \le C\min\{1, |\zeta|^2\}.$$

Consequently, for $0 < \tau \le 1$,

$$\sup_{\xi \in \mathbb{R}^d} |\widehat{\rho}(\tau\xi) - 1|(1 + |\xi|^2)^{-\delta/2} \le C\tau^{\min\{\delta,2\}}.$$

The Fourier multiplier estimate, restriction to Ω, and (215) imply

$$\|T_1\|_{H^{-s}(\Omega)} \le C\tau^{\min\{\delta,2\}}\|h\|_{H^{-s+\delta}(\Omega)}.$$

Estimate of T_2. Let $\varphi \in H_0^s(\Omega)$. By duality,

$$|\langle T_2, \varphi\rangle| \le \|h\|_{H^{-a}(\Omega)}\|(1-\chi_\tau)\varphi\|_{H_0^a(\Omega)}.$$

Because $a < 1/2$, the H_0^a norm is equivalent to the H^a norm, see Lions and Magenes [1972b].

Inside Ω, the function $1-\chi_\tau$ is supported in the boundary collar

$$\Sigma_\tau := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \le 5\tau\}.$$

For $s > 1/2$, the boundary-collar estimate gives

$$\|(1-\chi_\tau)\varphi\|_{L^2(\Omega)} \le C\tau^{1/2}\|\varphi\|_{H^s(\Omega)}.$$

Combining this with (216) and the Sobolev interpolation inequality yields

$$\begin{aligned}\|(1-\chi_\tau)\varphi\|_{H^{s-\delta}(\Omega)} &\le C\|(1-\chi_\tau)\varphi\|_{L^2(\Omega)}^{\delta/s}\|(1-\chi_\tau)\varphi\|_{H^s(\Omega)}^{1-\delta/s}\\ &\le C\tau^{\delta/(2s)}\|\varphi\|_{H^s(\Omega)}.\end{aligned}$$

Taking the supremum over the unit ball of $H_0^s(\Omega)$ therefore gives

$$\|T_2\|_{H^{-s}(\Omega)} \le C\tau^{\delta/(2s)}\|h\|_{H^{-s+\delta}(\Omega)}.$$

Together with the estimate of T_1, this proves the approximation bound. The simplified rate follows from $\delta/(2s) \le \min\{\delta, 2\}$.

Smoothing estimates. For every fixed $\sigma \ge -a$, the rapid decay of $\widehat{\rho}$ gives

$$\sup_{\xi\in\mathbb{R}^d} |\widehat{\rho}(\tau\xi)|(1+|\xi|^2)^{(\sigma+a)/2} \le C_\sigma\tau^{-(\sigma+a)}.$$

Using (215) and restricting to Ω, we obtain

$$\|R_\tau h\|_{H^\sigma(\Omega)} \le C_\sigma\tau^{-(\sigma+s-\delta)}\|h\|_{H^{-s+\delta}(\Omega)}.$$

The C^m estimate follows by taking $\sigma = s(m)$ and applying the Sobolev embedding.

Support and range. The support of χ_τ lies at distance at least 3τ from $\partial\Omega$. Convolution with ρ_τ therefore gives a smooth function supported at distance at least 2τ from the boundary. Hence $R_\tau h \in C_c^\infty(\Omega;\mathbb{R}^n)$.

Finally, $0 \le \chi_\tau \le 1$ and $\rho_\tau \ge 0$. If $h(x) \in [-L, L]^n$ almost everywhere, then each component of the zero extension of $\chi_\tau h$ lies in $[-L, L]$. Convolution with the probability kernel ρ_τ preserves this interval, proving the range assertion. □

L.6.3 Spectral regularization, with exact and empirical data

In the following two results, all Sobolev norms are taken in the Dirichlet spectral scale:

$$\|h\|_{H_D^t}^2 := \sum_{j=1}^{\infty} \lambda_j^t |\widehat{h}_j|^2, \qquad \widehat{h}_j := \langle h, \phi_j\rangle.$$

Define the spectral filter

$$\mathscr{R}_{\eta,K} h := \sum_{j=1}^{K} \frac{\widehat{h}_j}{1+\eta\lambda_j^r}\phi_j.$$

For bounded measurable $h : \Omega \to [-L, L]^n$, the normalization of the graph measure gives

$$\mathscr{R}_{\eta,K} h = S_{\eta,K}(\gamma_h).$$

Proposition 14 (Source-condition rate)**.** *Let $t < \sigma$, $0 < \sigma - t \le 2r$, and $h \in H_D^\sigma(\Omega;\mathbb{R}^n)$. For every $\eta > 0$ and integer $K \ge 1$,*

$$\|\mathscr{R}_{\eta,K}h - h\|_{H_D^t} \le \left(\eta^{\frac{\sigma-t}{2r}} + c_W^{\frac{\sigma-t}{2}}(K+1)^{-\frac{\sigma-t}{d}}\right)\|h\|_{H_D^\sigma}.$$

In particular, for $t = -s$, $s \ge 0$, and $0 < s + \sigma \le 2r$,

$$\|\mathscr{R}_{\eta,K}h - h\|_{H_D^{-s}} \le \left(\eta^{\frac{s+\sigma}{2r}} + c_W^{\frac{s+\sigma}{2}}(K+1)^{-\frac{s+\sigma}{d}}\right)\|h\|_{H_D^\sigma}.$$

For the cutoff $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$ and $0 < \eta \le 1$,

$$\|\mathscr{R}_{\eta,K(\eta)}h - h\|_{H_D^t} \le C\eta^{\frac{\sigma-t}{4r}}\|h\|_{H_D^\sigma}.$$

At $\sigma = t + 2r$, this becomes

$$\|\mathscr{R}_{\eta,K(\eta)}h - h\|_{H_D^t} \le C\eta^{1/2}\|h\|_{H_D^{t+2r}}.$$

Proof. Set $b := \sigma - t \in (0, 2r]$. The error has the spectral expansion

$$\mathscr{R}_{\eta,K}h - h = -\sum_{j=1}^{K}\frac{\eta\lambda_j^r}{1+\eta\lambda_j^r}\widehat{h}_j\phi_j - \sum_{j>K}\widehat{h}_j\phi_j.$$

For the filtered part, write

$$\frac{\eta\lambda^r}{1+\eta\lambda^r}\lambda^{-b/2} = \eta^{b/(2r)}\frac{z^{1-b/(2r)}}{1+z}, \qquad z := \eta\lambda^r.$$

Since $0 < b/(2r) \le 1$, the final fraction is at most one. Consequently,

$$\left\|\sum_{j=1}^{K}\frac{\eta\lambda_j^r}{1+\eta\lambda_j^r}\widehat{h}_j\phi_j\right\|_{H_D^t} \le \eta^{b/(2r)}\|h\|_{H_D^\sigma}.$$

For the tail,

$$\begin{aligned}\left\|\sum_{j>K}\widehat{h}_j\phi_j\right\|_{H_D^t}^2 &= \sum_{j>K}\lambda_j^\sigma|\widehat{h}_j|^2\lambda_j^{-b}\\ &\le \lambda_{K+1}^{-b}\|h\|_{H_D^\sigma}^2.\end{aligned}$$

The lower Weyl bound implies

$$\lambda_{K+1}^{-b/2} \le c_W^{b/2}(K+1)^{-b/d}.$$

The triangle inequality proves the first estimate.

For $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$,

$$(K(\eta)+1)^{-b/d} \le \eta^{b/(4r)}.$$

Since $\eta^{b/(2r)} \le \eta^{b/(4r)}$ for $0 < \eta \le 1$, the remaining assertions follow. □

Theorem 6 (Mean-square bound for empirical data)**.** *Let $s \ge 0$, $\sigma > -s$, and $0 < s + \sigma \le 2r$. Suppose*

$$h \in H_D^\sigma(\Omega;\mathbb{R}^n) \cap L^\infty(\Omega;[-L,L]^n).$$

Let $\bar{x}_1, \dots, \bar{x}_N$ be i.i.d. with law $\bar{\lambda}_\Omega$, and define $\mathbf{M}_N(h)$ using a measurable representative of h. Then

$$\begin{aligned}&\left(\mathbb{E}\|S_{\eta,K}(\mathbf{M}_N(h)) - h\|_{H_D^{-s}}^2\right)^{1/2}\\ &\quad\le \left(\eta^{\frac{s+\sigma}{2r}} + c_W^{\frac{s+\sigma}{2}}(K+1)^{-\frac{s+\sigma}{d}}\right)\|h\|_{H_D^\sigma} + \left(\frac{V_0 nL^2}{N}\right)^{1/2}\mathcal{N}_s(\eta,K),\end{aligned}$$

where

$$\mathcal{N}_s(\eta, K) := \left(\sum_{j=1}^{K} \frac{\lambda_j^{-s}}{(1+\eta\lambda_j^r)^2} \right)^{1/2}.$$

For every $\eta > 0$ and $K \geq 1$,

$$\mathcal{N}_s(\eta, K) \leq \begin{cases} C, & 2s > d, \\ C(1+\log K)^{1/2}, & 2s = d, \\ CK^{(d-2s)/(2d)}, & 2s < d. \end{cases}$$

Fix $B_0 > 0$ and restrict to inputs satisfying $\|h\|_{H_D^\sigma} \leq B_0$. With $K = K(\eta)$ from (139), the following root-mean-square bounds hold uniformly over these inputs:

$$\left(\mathbb{E} \|S_{\eta_N, K(\eta_N)}(\mathbf{M}_N(h)) - h\|_{H_D^{-s}}^2 \right)^{1/2} \leq \begin{cases} CN^{-1/2}, & 2s > d, \\ CN^{-1/2}(1+\log N)^{1/2}, & 2s = d, \\ CN^{-(s+\sigma)/(d+2\sigma)}, & 2s < d, \end{cases}$$

where one may choose

$$\eta_N = \begin{cases} N^{-2r/(s+\sigma)}, & 2s \geq d, \\ N^{-4r/(d+2\sigma)}, & 2s < d. \end{cases}$$

The constants are independent of N and h.

Proof. Define the empirical coefficient error

$$e_j := \frac{V_0}{N} \sum_{\ell=1}^{N} \phi_j(\bar{x}_\ell) h(\bar{x}_\ell) - \widehat{h}_j.$$

Uniform sampling gives $\mathbb{E}e_j = 0$. Independence and L^2 normalization of ϕ_j imply

$$\begin{aligned} \mathbb{E}|e_j|^2 &= \frac{V_0^2}{N} \mathbb{E} \left| \phi_j(\bar{x}) h(\bar{x}) - \mathbb{E}[\phi_j(\bar{x}) h(\bar{x})] \right|^2 \\ &\leq \frac{V_0^2}{N} \mathbb{E}\left[|\phi_j(\bar{x})|^2 |h(\bar{x})|^2 \right] \\ &\leq \frac{V_0 n L^2}{N}. \end{aligned}$$

Here we used $\mathbb{E}|\phi_j(\bar{x})|^2 = V_0^{-1}$.

The estimator decomposes as

$$S_{\eta,K}(\mathbf{M}_N(h)) - h = (\mathscr{R}_{\eta,K} h - h) + \sum_{j=1}^{K} \frac{e_j}{1+\eta\lambda_j^r} \phi_j.$$

By spectral orthogonality,

$$\begin{aligned} \mathbb{E} \left\| \sum_{j=1}^{K} \frac{e_j}{1+\eta\lambda_j^r} \phi_j \right\|_{H_D^{-s}}^2 &= \sum_{j=1}^{K} \frac{\lambda_j^{-s} \mathbb{E}|e_j|^2}{(1+\eta\lambda_j^r)^2} \\ &\leq \frac{V_0 n L^2}{N} \mathcal{N}_s(\eta, K)^2. \end{aligned}$$

No independence between different coefficient errors is required for this identity. Minkowski's inequality in $L^2(\mathbb{P}; H_D^{-s})$, together with Proposition 14, proves the stated error bound.

Since $(1+\eta\lambda_j^r)^{-2} \leq 1$, the Weyl estimate gives

$$\mathcal{N}_s(\eta, K)^2 \leq \sum_{j=1}^{K} \lambda_j^{-s} \leq C \sum_{j=1}^{K} j^{-2s/d}.$$

The three bounds follow from the corresponding power-sum estimates.

Finally, put $b := s + \sigma > 0$. Under $K(\eta) = \lceil \eta^{-d/(4r)} \rceil$, the deterministic error is bounded by $CB_0\eta^{b/(4r)}$. If $2s \geq d$, choosing $\eta_N = N^{-2r/b}$ makes this term $O(N^{-1/2})$. The sampling term is $O(N^{-1/2})$ when $2s > d$, and $O(N^{-1/2}(1 + \log N)^{1/2})$ when $2s = d$.

If $2s < d$, then

$$\mathcal{N}_s(\eta, K(\eta)) \leq C\eta^{-(d-2s)/(8r)}.$$

Notice that $d + 2\sigma > 0$, since $\sigma > -s$ and $2s < d$. The choice $\eta_N = N^{-4r/(d+2\sigma)}$ balances

$$\eta_N^{b/(4r)} \quad \text{and} \quad N^{-1/2}\eta_N^{-(d-2s)/(8r)},$$

both of which equal $N^{-(s+\sigma)/(d+2\sigma)}$. This completes the proof. □

L.6.4 Decoding

Lemma 24 (Decoder bounds)**.** *Let $s' \geq 0$ and $0 < \varepsilon \leq 1$. Assume that $\rho \in C_c^\infty(B(0,1))$ is nonnegative and even, with $\int \rho = 1$. For a function g on Ω, denote its zero extension to $\mathbb{R}^d$ by $\widetilde{g}$. Then the following estimates hold.*

(i) Stability between graph measures. *For $g_1, g_2 \in L^\infty(\Omega; [-L, L]^n)$,*

$$\begin{aligned}\|\mathbf{F}_\varepsilon\gamma_{g_1} - \mathbf{F}_\varepsilon\gamma_{g_2}\|_{H^{-s'}(\Omega)} &\leq C\|\rho_\varepsilon * (\widetilde{g}_1 - \widetilde{g}_2)\|_{L^2(\mathbb{R}^d)} \\ &\leq C\|g_1 - g_2\|_{L^2(\Omega)} \leq CV_0^{1/2}\|g_1 - g_2\|_{L^\infty(\Omega)}.\end{aligned}$$

(ii) Stability between general measures. *For every $\gamma, \gamma' \in \mathcal{P}(\Xi)$,*

$$\|\mathbf{F}_\varepsilon\gamma - \mathbf{F}_\varepsilon\gamma'\|_{H^{-s'}(\Omega)} \leq C\varepsilon^{-(d+1)}W_1(\gamma, \gamma').$$

(iii) Approximation of an exact graph measure. *For $\alpha \in (0, 1]$ and $g \in C^{0,\alpha}(\overline{\Omega}; [-L, L]^n)$,*

$$\|\mathbf{F}_\varepsilon\gamma_g - g\|_{H^{-s'}(\Omega)} \leq C\left([g]_{C^{0,\alpha}}\varepsilon^\alpha + L\varepsilon^{1/2}\right).$$

The constants depend on the fixed domain, s', ρ, and n; in part (ii) they may additionally depend on L. They are independent of ε and the functions or measures being compared.

Proof. We use the continuous embedding $L^2(\Omega) \hookrightarrow H^{-s'}(\Omega)$, see Lions and Magenes [1972b].

(i). By the definition of the graph measure and the evenness of ρ,

$$\mathbf{F}_\varepsilon\gamma_g = (\rho_\varepsilon * \widetilde{g})|_\Omega.$$

Young's inequality and $\|\rho_\varepsilon\|_{L^1(\mathbb{R}^d)} = 1$ give

$$\|\rho_\varepsilon * (\widetilde{g}_1 - \widetilde{g}_2)\|_{L^2(\mathbb{R}^d)} \leq \|\widetilde{g}_1 - \widetilde{g}_2\|_{L^2(\mathbb{R}^d)} = \|g_1 - g_2\|_{L^2(\Omega)}.$$

The remaining estimates follow from the embedding and the finite volume of Ω.

(ii). For fixed $x \in \Omega$ and $i \in \{1, \ldots, n\}$, define

$$f_{x,i}(x', y') := V_0\varepsilon^{-d}\rho\left(\frac{x' - x}{\varepsilon}\right)y_i'.$$

For $\xi' = (x', y')$ and $\xi'' = (x'', y'')$ in Ξ,

$$\begin{aligned}|f_{x,i}(\xi') - f_{x,i}(\xi'')| &\leq V_0L\varepsilon^{-(d+1)}\|\nabla\rho\|_\infty|x' - x''| \\ &\quad + V_0\varepsilon^{-d}\|\rho\|_\infty|y_i' - y_i''| \\ &\leq C\varepsilon^{-(d+1)}|\xi' - \xi''|,\end{aligned}$$

where we used $\varepsilon \leq 1$. Kantorovich–Rubinstein duality, applied componentwise, therefore yields

$$\sup_{x\in\Omega}|\mathbf{F}_\varepsilon\gamma(x) - \mathbf{F}_\varepsilon\gamma'(x)| \leq C\varepsilon^{-(d+1)}W_1(\gamma, \gamma').$$

Passing first to $L^2(\Omega)$ and then to $H^{-s'}(\Omega)$ proves the assertion.

(iii). Set

$$I_\varepsilon := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > \varepsilon\}, \qquad \Sigma_\varepsilon := \Omega \setminus I_\varepsilon.$$

For $x \in I_\varepsilon$, the convolution samples only points inside Ω. Hence

$$\begin{aligned} |\rho_\varepsilon * \widetilde{g}(x) - g(x)| &\le [g]_{C^{0,\alpha}} \int_{\mathbb{R}^d} \rho_\varepsilon(z)|z|^\alpha \, dz \\ &\le [g]_{C^{0,\alpha}} \varepsilon^\alpha. \end{aligned}$$

On Σ_ε, nonnegativity and normalization of the kernel imply

$$|\rho_\varepsilon * \widetilde{g}(x) - g(x)| \le 2L\sqrt{n}.$$

The smooth bounded domain satisfies $|\Sigma_\varepsilon| \le C_\Omega \varepsilon$ for $0 < \varepsilon \le 1$. Consequently,

$$\|\rho_\varepsilon * \widetilde{g} - g\|_{L^2(\Omega)} \le C \left([g]_{C^{0,\alpha}} \varepsilon^\alpha + L\varepsilon^{1/2} \right).$$

The embedding into $H^{-s'}(\Omega)$ completes the proof. □

L.7 End-to-end rates

Assumption 3 (Standing hypotheses)**.** *The following conditions hold.*

1. *$\Omega \subset \mathbb{R}^d$ is a bounded domain with C^∞ boundary and $d \ge 3$. The regularity parameters satisfy*

$$s > \frac{1}{2}, \qquad 0 < \delta \le s, \qquad 0 \le s - \delta < \frac{1}{2}, \qquad s \notin \frac{1}{2} + \mathbb{Z}, \qquad s' \ge 0.$$

Furthermore, $k \in \mathbb{Z}_{\ge 0}$, $r > d/2 + k$ is an even integer, and $\alpha, \beta \in (0, 1]$. We use

$$K(\eta) = \lceil \eta^{-d/(4r)} \rceil, \qquad 0 < \eta \le 1.$$

The mollifier ρ is nonnegative, even, smooth, supported in $B(0, 1)$, and has integral one.

2. *The input h ranges over the set $\mathcal{B}$ defined in Assumption 2, with $\|h\|_{H^{-s+\delta}} \le B_0$. Write*

$$\kappa_m := s(m) + s - \delta \qquad (m \in \mathbb{Z}_{\ge 0}).$$

3. *The operator A satisfies the smooth-input bounds of Assumption 2, including their polynomial growth bound with exponent p.*

4. *The rough-input Hölder estimate (210) holds on $\mathcal{K}_\mathcal{B}$.*

5. *The sampling points $\bar{x}_1, \dots, \bar{x}_N$ are i.i.d. with law $\bar{\lambda}_\Omega$. Expectations below are taken over these points.*

Theorem 7 (End-to-end error bound)**.** *Suppose Assumption 3 holds. Let $\tau, \eta, \varepsilon \in (0, 1]$, $N \in \mathbb{N}$, and $0 < \varepsilon_{\mathrm{T}} \le 1$. Let Γ^{tran} be a transformer satisfying*

$$\sup_{\mu \in \mathcal{P}(\Xi)} W_1\big(G_{\Gamma^{\mathrm{tran}}}(\mu), G_{\Gamma_\eta}(\mu)\big) \le \varepsilon_{\mathrm{T}}.$$

Such a transformer is supplied by Proposition 11.

Define

$$B_{\mathrm{sp}} := \alpha(2r + \kappa_k) + k(p + \alpha)\kappa_k.$$

Then, uniformly for $h \in \mathcal{B}$,

$$\begin{aligned} &\mathbb{E} \, \|\mathbf{F}_\varepsilon G_{\Gamma^{\mathrm{tran}}} \mathbf{M}_{N,\tau}(h) - A(h)\|_{H^{-s'}(\Omega)} \\ &\le C\tau^{\beta\delta/(2s)} \\ &\quad + C\tau^{-p\kappa_k}\varepsilon^\alpha + C\varepsilon^{1/2} \\ &\quad + C\eta^{\alpha/2}\tau^{-B_{\mathrm{sp}}} \\ &\quad + C\varepsilon^{-(d+1)}(\Lambda_\eta + M_{\xi,\eta})\tau^{-\alpha\kappa_1} N^{-\alpha/d} \\ &\quad + C\varepsilon^{-(d+1)}\varepsilon_{\mathrm{T}}. \end{aligned} \tag{217}$$

The constant depends only on the fixed data in Assumption 3, the clipping map Ψ, and the chosen norm conventions. It is independent of $h, \tau, \eta, \varepsilon, N$, and ε_{T}.

For deterministic sampling points satisfying

$$W_1\left(\frac{1}{N}\sum_{j=1}^{N}\delta_{\bar{x}_j}, \bar{\lambda}_\Omega\right) \le C_0 N^{-1/d},$$

the same estimate holds without expectation, with a constant depending additionally on C_0.

Proof. Write

$$g := R_\tau h, \qquad v_\eta := S_\eta(\gamma_g), \qquad g_\eta := \Psi \circ v_\eta, \qquad \mu_N := \mathbf{M}_N(g).$$

The range assertion of Proposition 13 gives $g(x) \in [-L, L]^n$. In particular, $\Psi \circ g = g$. Moreover,

$$G_{\Gamma_\eta}(\gamma_g) = \gamma_{A(g_\eta)}.$$

Using the linear extension of the decoder to differences of measures, decompose

$$\begin{aligned}
&\mathbf{F}_\varepsilon G_{\Gamma^{\mathrm{tran}}}(\mu_N) - A(h) \\
&\quad = \underbrace{A(g) - A(h)}_{T_1} + \underbrace{\mathbf{F}_\varepsilon \gamma_{A(g)} - A(g)}_{T_2} \\
&\qquad + \underbrace{\mathbf{F}_\varepsilon\big(\gamma_{A(g_\eta)} - \gamma_{A(g)}\big)}_{T_3} \\
&\qquad + \underbrace{\mathbf{F}_\varepsilon\big(G_{\Gamma_\eta}(\mu_N) - G_{\Gamma_\eta}(\gamma_g)\big)}_{T_4} \\
&\qquad + \underbrace{\mathbf{F}_\varepsilon\big(G_{\Gamma^{\mathrm{tran}}}(\mu_N) - G_{\Gamma_\eta}(\mu_N)\big)}_{T_5}.
\end{aligned}$$

All output norms below are taken in $H^{-s'}(\Omega)$.

Input mollification. Since $h, g \in \mathcal{K}_{\mathcal{B}}$, the rough-input Hölder estimate and Proposition 13 yield

$$\|T_1\|_{H^{-s'}} \le C_\omega \|R_\tau h - h\|^{\beta}_{H^{-s}} \le C\tau^{\beta\delta/(2s)}.$$

Decoding the exact output. The smoothing bound gives

$$\|g\|_{C^k} \le C\tau^{-\kappa_k}.$$

Consequently, the smooth-input assumptions imply

$$[A(g)]_{C^{0,\alpha}} \le C(1 + \|g\|_{C^k})^p \le C\tau^{-p\kappa_k}.$$

Lemma 24(iii) therefore gives

$$\|T_2\|_{H^{-s'}} \le C\tau^{-p\kappa_k}\varepsilon^\alpha + C\varepsilon^{1/2}.$$

Spectral regularization. We first obtain bounds independent of η for the regularized smooth input. The spectral multiplier of $\mathscr{R}_{\eta,K}$ has absolute value at most one, so

$$\|v_\eta\|_{H_D^{s(k)}} \le \|g\|_{H_D^{s(k)}}.$$

Both $s(k)$ and $2r + s(k)$ are even integers. For compactly supported smooth functions, the corresponding Dirichlet spectral norms are bounded by the standard Sobolev norms of the same orders, see Lions and Magenes [1972b]. Thus Proposition 13 and the spectral Sobolev embedding imply

$$\|v_\eta\|_{C^k} + \|g\|_{C^k} \le C\tau^{-\kappa_k}.$$

Similarly,

$$\|g\|_{H_D^{2r+s(k)}} \le C\tau^{-(2r+\kappa_k)}.$$

Proposition 14, applied with target order $s(k)$ and source order $2r + s(k)$, therefore yields

$$\|v_\eta - g\|_{C^k} \le C\eta^{1/2}\tau^{-(2r+\kappa_k)}.$$

Applying the Nemytskii difference estimate on a ball of radius $C\tau^{-\kappa_k}$ gives

$$\begin{aligned}\|g_\eta - g\|_{C^k} &= \|\Psi \circ v_\eta - \Psi \circ g\|_{C^k} \\ &\le C\tau^{-k\kappa_k}\|v_\eta - g\|_{C^k} \\ &\le C\eta^{1/2}\tau^{-(2r+\kappa_k)-k\kappa_k}.\end{aligned}$$

The composition bound also gives

$$\|g_\eta\|_{C^k} \le C\tau^{-k\kappa_k}.$$

For $k \ge 1$, the same upper bound controls $\|g\|_{C^k}$. For $k = 0$, both functions instead have uniformly bounded C^0 norms by their range constraints. Thus, in all cases, g and g_η belong to $B_k(C\tau^{-k\kappa_k})$, after increasing C.

The polynomial growth bound on A now implies

$$\begin{aligned}\|A(g_\eta) - A(g)\|_{L^\infty} &\le C\tau^{-kp\kappa_k}\|g_\eta - g\|_{C^k}^{\alpha} \\ &\le C\eta^{\alpha/2}\tau^{-\alpha(2r+\kappa_k)-k(p+\alpha)\kappa_k}.\end{aligned}$$

Applying Lemma 24(i) gives

$$\|T_3\|_{H^{-s'}} \le C\eta^{\alpha/2}\tau^{-B_{\mathrm{sp}}}.$$

Tokenization. By smoothing and the fixed-domain extension estimate,

$$[g]_{C^{0,1}} \le C\|g\|_{C^1} \le C\tau^{-\kappa_1}.$$

Proposition 12 gives

$$\mathbb{E}W_1(\mu_N, \gamma_g) \le C\tau^{-\kappa_1}N^{-1/d}.$$

The push forward estimate of Theorem 5 and Lemma 24(ii) imply

$$\|T_4\|_{H^{-s'}} \le C\varepsilon^{-(d+1)}(\Lambda_\eta + M_{\xi,\eta})W_1(\mu_N, \gamma_g)^\alpha.$$

Taking expectations and applying Jensen's inequality, we obtain

$$\mathbb{E}\|T_4\|_{H^{-s'}} \le C\varepsilon^{-(d+1)}(\Lambda_\eta + M_{\xi,\eta})\tau^{-\alpha\kappa_1}N^{-\alpha/d}.$$

Transformer approximation. The assumed measure-level approximation and Lemma 24(ii) give

$$\|T_5\|_{H^{-s'}} \le C\varepsilon^{-(d+1)}\varepsilon_{\mathrm{T}}.$$

Combining the five estimates proves (217).

For deterministic sampling, use the deterministic version of Proposition 12. The rest of the proof is unchanged. □

Corollary 4 (An algebraic rate in the token number)**.** *Suppose Assumption 3 holds. Let q_K, q'_K be the exponents in (212), and define*

$$u := p\kappa_k, \qquad v := \frac{d}{4r}\max\{q_K, q'_K\}, \qquad w := \alpha\kappa_1.$$

Set

$$b := \max\left\{2, \frac{u+1}{\alpha}\right\}, \qquad c := \frac{2(B_{\mathrm{sp}}+1)}{\alpha},$$

$$Q := (d+1)b + vc + w, \qquad \lambda := \frac{\alpha}{2d(1+Q)}.$$

Choose

$$a_\tau := \lambda, \qquad a_\varepsilon := b\lambda, \qquad a_\eta := c\lambda, \qquad a_{\mathrm{T}} := ((d+1)b+1)\lambda,$$

and

$$\varkappa := \lambda \min\left\{1, \frac{\beta\delta}{2s}\right\} > 0.$$

For each $N \in \mathbb{N}$, set

$$\tau_N := N^{-a_\tau}, \qquad \varepsilon_N := N^{-a_\varepsilon}, \qquad \eta_N := N^{-a_\eta}, \qquad \varepsilon_{\mathrm{T},N} := N^{-a_{\mathrm{T}}},$$

and take

$$\Gamma_N^{\mathrm{tran}} := T_{\eta_N, \varepsilon_{\mathrm{T},N}}$$

from Proposition 11. Then

$$\sup_{h\in\mathcal{B}} \mathbb{E}\left\|\mathbf{F}_{\varepsilon_N} G_{\Gamma_N^{\mathrm{tran}}} \mathbf{M}_{N,\tau_N}(h) - A(h)\right\|_{H^{-s'}(\Omega)} \le CN^{-\varkappa}.$$

The constant C is independent of N and h.

Proof. By (212) and the prescribed cutoff,

$$\Lambda_\eta + M_{\xi,\eta} \le C\eta^{-v}.$$

The right-hand side R of (217) is bounded by

$$R \le C(N^{-\lambda\beta\delta/(2s)} + N^{-\lambda(\alpha b - u)} + N^{-\lambda b/2} + N^{-\lambda(\alpha c/2 - B_{\mathrm{sp}})} + N^{-\alpha/d + \lambda Q} + N^{-\lambda}).$$

The definitions of b and c give

$$\alpha b - u \ge 1, \qquad b/2 \ge 1, \qquad \alpha c/2 - B_{\mathrm{sp}} = 1.$$

Furthermore,

$$\frac{\alpha}{d} - \lambda Q = \frac{\alpha(2+Q)}{2d(1+Q)} \ge \lambda.$$

Every term is therefore bounded by $CN^{-\varkappa}$. The conclusion follows from Theorem 7. □

Corollary 5 (A diagonal approximation sequence)**.** *Under the hypotheses and parameter choices of Corollary 4, let $N_j := 2^j$. Then*

$$\sup_{h\in\mathcal{B}} \mathbb{E}\left\|\mathbf{F}_{\varepsilon_{N_j}} G_{\Gamma_{N_j}^{\mathrm{tran}}} \mathbf{M}_{N_j,\tau_{N_j}}(h) - A(h)\right\|_{H^{-s'}(\Omega)} \le C2^{-\varkappa j} \longrightarrow 0.$$

Thus the coupled sequence gives uniform convergence in expected $H^{-s'}$ error using finite transformers. Moreover,

$$\log\bigl(1 + P_{\mathrm{tot}}(\Gamma_N^{\mathrm{tran}})\bigr) \le CN^{a_\eta d/(4r)}(1 + \log N).$$

Proof. The convergence estimate follows by substituting $N = 2^j$ into Corollary 4. The parameter-count bound follows from Proposition 11 with $\eta = \eta_N$ and $\varepsilon_{\mathrm{T}} = \varepsilon_{\mathrm{T},N}$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See the abstract and Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: See Sections 2, 3, 4, 5.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See Appendix K.

   Guidelines:

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

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

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

Answer: [No]

Justification: We have not provided code at the time of submission. We emphasize that the main contributions of this paper are the theorems and proofs, and numerical examples are provided only to illustrate the theory.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: See Appendix K.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Numerical results in Appendix K report 95% confidence intervals where appropriate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The GPU model, memory, and compute time for numerical illustrations is supplied in Appendix K.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This research contributes to the theory of machine learning. We are not aware of any ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This work is primarily theoretical, and we are not aware of any direct or immediate societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This work is theoretical and does not involve the release of data or models with potential misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: This work does not use external datasets or third-party assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: This work does not introduce or release new datasets, code, or models.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work is theoretical and does not involve crowdsourcing or human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: This work does not involve the use of large language models in the core methodology.

    Guidelines:

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