ALL ROADS LEAD TO MARKOV KERNEL NETWORKS: A REPRESENTATION THEOREM FOR DISCRETIZATION-STABLE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A central goal of operator learning is to design networks that are stable to changes in the discretization of the input function, a property sometimes called "discretization invariance". We consider a quantitative refinement of this idea called *uniform discretization stability* that provides a Hölder-stability bound controlling changes in a network's output in terms of the Wasserstein distance between input measures. Our main result is a representation theorem showing all uniformly discretization stable operators can be written exactly as a *stable Markov kernel network*—an abstract architecture that is a joint generalization of transformers, Fourier neural operators (FNO), Flowers, and other successful learned operators. Our representation theorem characterizes a design space for discretization-stable operator learning; designing new architectures becomes a question of which parts of a Markov kernel network to parameterize, and how. We further show that the read kernels of these networks can be realized as solutions of regularized variational problems, with particular choices of costs and regularizers recovering the kernels of transformers, FNO, Flowers, their hybrids, and beyond. This variational formulation provides a systematic approach to architecture design, together with sufficient conditions for uniform discretization stability.

1 INTRODUCTION

Operator learning requires neural networks that approximate maps between functions. The input functions of the target operator are continuum objects that can not be represented exactly on a computer. Furthermore, in scientific applications using real-world data, we generally do not have complete knowledge of the input function f, but rather only a finite collection of (sometimes noisy) samples from the graph of a function $(x, f(x))$. To accurately model the continuum target operator, a neural network should not be sensitive to small changes in the discretization of the input. The output of a "discretization invariant" network converges as the discretization of the input function is refined. Quantitative guarantees about the trained network, however, require stability bounds, not merely convergence under discretization refinement. This leads to the following three questions:

1. What discretization stability properties are relevant for operator learning?
2. What network architectures satisfy these stability properties?
3. How can these architectures be characterized and systematically explored?

Our approach is to lift operators between functions to operators between measures. This provides a unified viewpoint of continuum inputs (functions) and their noisy discretizations. The necessary stability properties can then be stated as a Hölder stability with respect to a Wasserstein metric. *Uniform discretization stability* controls the behavior of learned operator networks under different discretizations of the input functions. For example, it provides quantitative guarantees for networks that are trained at one resolution or on a uniform grid, but evaluated on non-uniform point clouds sampled at different resolution. It also provides a unified framework for that analysis of errors arising from both discretization and noisy measurements.

We then turn to the second question, and introduce *Markov kernel networks*, a shared generalization of transformers (Vaswani et al., 2017), FNO (Li et al., 2021), Flowers (Muser et al., 2026), and other

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

operator networks. These networks alternate nonlinear integral operators built using Markov kernels to mix information spatially with channel-wise mixing operators to mix information pointwise in the embedding space. When the individual components of these networks satisfy relevant stability properties, the entire network becomes uniformly discretization stable. Conversely, our main theorem shows that every uniformly discretization stable operator can be written as a stable Markov kernel network. This is not an approximation theorem, the representation is exact.

With the design space of discretization-stable architectures fixed by the representation theorem, we propose a variational approach to explore the design space. We investigate Markov kernels that minimize regularized objectives; the induced read behavior is determined by the choice of regularizer and cost function. We provide sufficient conditions on the regularizers and costs so that the resulting networks are guaranteed to be uniformly discretization stable, and identify the regularizers corresponding to the popular network architectures discussed earlier. We aim to provide a principled framework for adaptive (Cortes et al., 2017) and hybrid (Balaji et al., 2026) architectures, and a useful belief space for agentic (Toscano et al., 2026) approaches to operator learning.

1.1 Prior work

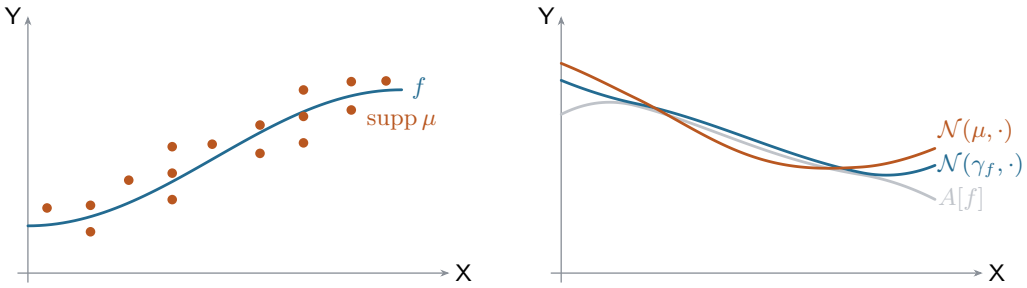
Discretization invariance was initially explored in the context of ill-posed inverse problems. The term was first defined in Lassas & Siltanen (2004), with important antecedents in Lasanen (2002); D'Ambrogi et al. (1999); Lehtinen et al. (1989), see Lassas et al. (2009) and Stuart (2010) for additional historical remarks. It was popularized in operator learning by (earlier preprints of) the seminal work Kovachki et al. (2023), which includes analysis of operators that converge under discretization refinement. Various notions of quantitative discretization invariance and stability bounds have been developed in Wang & Golland (2023); Bartolucci et al. (2023); Lanthaler et al. (2024); de Bie et al. (2019); Ruiz et al. (2020); Gao et al. (2025); Subedi & Tewari (2025), and Bendahi et al. (2026). Stability of transformers was investigated in Castin et al. (2024). The notion of uniform discretization stability we propose in Section 2 has roots in Transportation-L^p theory (García Trillos & Slepčev, 2016; Thorpe et al., 2017); we are unaware of prior work using these Hölder bounds in the context of operator learning, but Furuya et al. (2026) discusses a similar Lipschitz bound for transformers. This paper builds on both the stability bound and "graph measure" viewpoint of that work.

The mindset for our representation theorem and variational characterizations is inspired by Unser (2019), which proved that optimal activation functions (in a total-variation sense) for multi-layer perceptrons take the form of linear splines with adaptive knot locations. Our presentation of Markov kernel networks directly mirrors measure-theoretic transformers (Furuya et al., 2025b;a; 2026). Nicolae & Blondel (2017) characterized softmax attention as a minimizer of a KL-regularized objective, which was further explored in Martins et al. (2022); Terjék (2021).

1.2 Paper outline and contributions

- In Section 2, we propose *uniform discretization-stability*, a Hölder stability property, as a quantitative notion of discretization invariance.
- In Section 3, we define *Markov kernel networks*, an abstract architecture that generalizes transformers, neural operators, Flowers, and others. Our main results are the representation theorems 3.5 and 3.6 establishing an equivalence between uniformly discretization-stable operators and stable Markov kernel networks. This suggests Markov kernel networks provide a natural design space for uniformly discretization-stable operator learning.
- Section 4 explores the design space of Markov kernel networks using a variational characterization of discretization stable networks. We investigate Markov kernels that minimize a particular regularized objective function, and show how certain regularizers determine different read behavior, including those corresponding to existing architectures and hybrids. We establish stability bounds for variationally defined Markov kernels.

Section 5 discusses some limitations of our results. The appendices contain background and notation used throughout the document (Appendix A), proofs (Appendix B), a universal approximation result for Markov kernel networks (Appendix C), constructions for popular architectures as Markov kernel networks (Appendix D), and an additional discussion of hybrid architectures defined by variational Markov reads (Appendix E).



(a) Inputs (b) Outputs

Figure 1: Operator learning setup. The target operator is a continuum object $f \mapsto A[f]$. A neural network $\mathcal{N}$ takes as input a measure μ representing a noisy sample of points from the graph of f (denoted γ_f), and outputs a function that can be evaluated at any point in the domain X. The network $\mathcal{N}$ is *uniformly discretization stable* (Definition 2.1) if it satisfies a Hölder stability property that provides quantitative guarantees that $\mathcal{N}(\mu, \cdot)$ converges to $\mathcal{N}(\gamma_f, \cdot)$, the output had the network been provided with the ideal (continuum) input γ_f.

2 Uniform discretization stability

Let $(\mathsf{X}, \rho_{\mathsf{X}})$ be a compact metric space with at least two elements, and $\mathsf{Y} \subset \mathbb{R}^{d_{\text{out}}}$ be non-empty, compact, and convex. Define the *graph metric space* $\mathsf{G} := \mathsf{X} \times \mathsf{Y}$ equipped with the metric

$$\rho_{\mathsf{G}} : \mathsf{G} \times \mathsf{G} \to \mathbb{R}_{\geq 0}, \qquad \rho_{\mathsf{G}}((x_1, y_1), (x_2, y_2)) := \sqrt{\rho_{\mathsf{X}}(x_1, x_2)^2 + |y_1 - y_2|^2}. \tag{1}$$

Choose a Radon probability measure $\lambda \in \mathcal{P}(\mathsf{X})$ with $\operatorname{supp} \lambda = \mathsf{X}$. Let $L^2(\mathsf{X}, \mathsf{Y})$ be the space of Radon-measurable functions $f : \mathsf{X} \to \mathsf{Y}$ that are square integrable with respect to λ. Observe that since Y is compact, $L^2(\mathsf{X}, \mathsf{Y}) \subset L^\infty(\mathsf{X}, \mathsf{Y})$. In this paper, we consider operators $A : L^2(\mathsf{X}, \mathsf{Y}) \to L^2(\mathsf{X}, \mathsf{Y})$. All maps in the paper are Borel measureable.

We identify a function $f \in L^2(\mathsf{X}, \mathsf{Y})$ with its *graph measure* (Furuya et al., 2026),

$$\gamma_f \in \mathcal{P}(\mathsf{G}), \qquad \gamma_f := (\mathrm{id}, f)_{\#}\lambda, \quad (\mathrm{id}, f) : \mathsf{X} \to \mathsf{X} \times \mathsf{Y}. \tag{2}$$

(See Appendix A for definition of pushforward operator "$_{\#}$".) The map $f \mapsto \gamma_f$ is injective. For helpful intuition, when f is continuous (so it is defined pointwise and has a closed graph) we have $\operatorname{supp} \gamma_f = \{(x, f(x)) : x \in \mathsf{X}\}$.

Computer implementations of neural networks for function-to-function maps generally do not accept literal functions $f \in L^2(\mathsf{X}, \mathsf{Y})$, but instead f is approximated by a finite collection of points on γ_f, or only near γ_f if measurements include noise. We model these sampled points as a mixture of Dirac measures. Numerically, $f \in L^2(\mathsf{X}, \mathsf{Y})$ is approximated by sampling a finite collection of domain points $\{x_i\}_{i=1}^N$ points and forming the empirical approximation γ_f,

$$\mu \in \mathcal{P}(\mathsf{G}), \qquad \mu = \sum_{i=1}^N w_i \delta_{(x_i, y_i)}, \quad w_i \geq 0, \quad \sum_{i=1}^N w_i = 1. \tag{3}$$

When the samples (x_i, y_i) lie on the graph of a function f, we have $y_i = f(x_i)$. The mixture weights w_i allow the samples to approximate a graph measure even when the x_i locations are not distributed uniformly with respect to λ. This formulations captures both discretizations on uniform grids, as well as irregularly-sampled point clouds. Provided with samples μ to represent the input function f, an operator network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ can then be evaluated at an arbitrary $x \in \mathsf{X}$. Informally, a trained neural network $\mathcal{N}$ is a good approximation of A when

$$\mu \approx \gamma_f \implies \mathcal{N}(\mu, \cdot) \approx A[f] \qquad \text{for all } f \in L^2(\mathsf{X}, \mathsf{Y}). \tag{4}$$

To make the above precise, we require a metric on measures to compare the graph measure γ_f to the noisy samples μ. For this, we use an optimal transport distance called the Wasserstein-1 metric (see A.1 for the standard defintion) (Villani, 2009).

We now formulate *uniform discretization stability* as a Hölder-stability property for neural networks.

Definition 2.1 (Uniform discretization stability.)**.** A network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ is *uniformly discretization stable* on a set $\mathcal{S} \subseteq \mathcal{P}(\mathsf{G})$ if there exist constants $C > 0$, $\beta \in (0, 1]$ such that

$$\sup_{x \in \mathsf{X}} |\mathcal{N}(\mu, x) - \mathcal{N}(\nu, x)| \leq C\, W_1(\mu, \nu)^\beta \qquad \text{for all } \mu, \nu \in \mathcal{S}. \tag{DS}$$

Returning now to the approximation criterion (4), using the triangle inequality and the fact $\lambda(\mathsf{X}) = 1$,

$$\|\mathcal{N}(\mu, \cdot) - A[f]\|_{L^2(\mathsf{X}, \mathbb{R}^{d_{\text{out}}})} \leq \underbrace{\|\mathcal{N}(\gamma_f, \cdot) - A[f]\|_{L^2(\mathsf{X}, \mathbb{R}^{d_{\text{out}}})}}_{\text{ideal approximation error}} + \underbrace{\|\mathcal{N}(\mu, \cdot) - \mathcal{N}(\gamma_f, \cdot)\|_\infty}_{\text{discretization error}}. \tag{5}$$

The first term on the right is the approximation error of an ideal network given all information about the input function f; this quantity is controlled by training the network. The second term is the error introduced by the discretization μ. It is largely determined by the neural network architecture itself, and is not directly controlled by training. Discretization stable architectures satisfying (DS) control the discretization error, so they have the possibility to be trained to satisfy the desired property (4). In the next section, we will show that all such networks have a particular representation as stable Markov kernel networks. This allows us to view, for example, softmax attention, and SelfWarps (Muser et al., 2026) as two particular choices in the same space of discretization stable architectures.

Remark. Uniform discretization stability is a stronger condition than convergence under resolution refinement. In particular, (DS) allows a unified analysis of both resolution refinement and stability to noise in the measurements of f, both of which are important in applications. Furthermore, for sufficiently regular classes of target operators, it allows us to obtain quantitative guarantees for the performance of a network trained at one resolution but evaluated at another, which is important for scientific applications where large quantities of high-resolution training data may be prohibitively expensive. Indeed, if $\mathcal{N}$ is uniformly discretization stable, then

$$\|\mathcal{N}(\mu_{\text{test}}, \cdot) - A[f]\|_{L^2(\mathsf{X}, \mathbb{R}^{d_{\text{out}}})} \leq \|\mathcal{N}(\mu_{\text{train}}, \cdot) - A[f]\|_{L^2(\mathsf{X}, \mathbb{R}^{d_{\text{out}}})} + C\, W_1(\mu_{\text{test}}, \mu_{\text{train}})^\beta, \tag{6}$$

where μ_{train} and μ_{test} are discretizations of f at the training and test resolutions, respectively. Simple convergence under resolution refinement provides no such guarantees.

Remark. Definition 2.1 does not impose additional regularity requirements of the target operator A or the input functions f. See the discussion of universal approximation in Appendix C.

2.1 Non-collapsing measures $\mathcal{S}$

The set $\mathcal{S} \subset \mathcal{P}(\mathsf{G})$ in Definition 2.1 need not contain all probability measures to provide a useful notion of discretization stability. Our main result, Theorem 3.6, holds for an admissible classes that are relevant for operator learning, but not all possible inputs in $\mathcal{P}(\mathsf{G})$ (see Example 3.7).

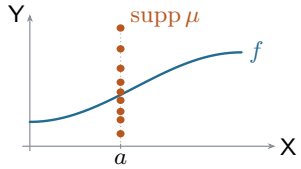


Figure 2: A measure that is not in any admissible set $\mathcal{S}$. Since $(\pi_{\mathsf{X}})_\# \mu = \delta_a$, every set containing μ, or a sequence converging to μ, fails the admissibility criterion (7).

A set of measurements at only a single value of x does not provide much information about an input function f, and we are not concerned with the behavior of networks on such pathological inputs. To avoid this pathology, we define a condition on admissible sets of measures to ensure there is enough variation in the X-marginal.

Definition 2.2 (Admissible sets of non-collapsing measures)**.** A non-empty set $\mathcal{S} \subset \mathcal{P}(\mathsf{G})$ is *admissible* if there exists $\epsilon > 0$ such that

$$\inf_{\mu \in \mathcal{S}} \inf_{a \in \mathsf{X}} W_1\big((\pi_{\mathsf{X}})_\# \mu, \delta_a\big) > \epsilon \tag{7}$$

where $\pi_{\mathsf{X}} : (x, y) \mapsto x$ is projection onto the X component.

Crucially, $\mathcal{S}$ can contain both graph measures γ_f and empirical measures μ built from finitely many samples, even in the presence of a large amount of noise, as long as the x coordinates do not collapse.

3 Markov kernel networks

In this section, we define an abstract architecture for operator learning called *Markov kernel networks* that generalizes transformers, Fourier neural operators, Flowers, and other successful network

architectures. We then come to the main representation theorem of this paper which shows that all uniformly discretization stable operators on admissible input classes can be exactly written in the form of a Markov kernel network.

3.1 DEFINITIONS

Let M be a metric space. A nonlinear *Markov kernel* is a map $K : \mathcal{P}_1(\mathsf{M}) \times \mathsf{M} \to \mathcal{P}_1(\mathsf{M})$. Markov kernels form the backbone of the "in-context maps" (Furuya et al., 2025b) that define the hidden layers of a Markov kernel network. An in-context map uses a state encoded as a probability measure, called a "context", to determine how to update a single point.

Definition 3.1 (Markov kernel in-context map)**.** Let E, E' be compact metric spaces. A *Markov kernel in-context map* $\Gamma : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \to \mathsf{E}'$ with $H > 0$ heads is of the form

$$r_h(\mu, z) = \int_{\mathsf{E}} v_h(w) K_h[\mu, z](\mathrm{d}w), \qquad h = 1, \dots, H \tag{8}$$

$$\Gamma(\mu, z) = m\big(z, r_1(\mu, z), \dots, r_H(\mu, z)\big). \tag{9}$$

For $h = 1, \dots, H$, the maps $K_h : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \to \mathcal{P}(\mathsf{E})$ are Markov kernels, bounded $v_h : \mathsf{E} \to \mathbb{R}^{d_h}$ are called *value maps*, and $m : \mathsf{E} \times \prod_{h=1}^{H} \mathbb{R}^{d_h} \to \mathsf{E}'$ are called *channel mixers*. Each $r_h : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \to \mathbb{R}^{d_h}$ is called a *read head*. Additionally, we require the output of the Markov kernels to be supported on the input context μ:

$$\operatorname{supp} K_h[\mu, z] \subset \operatorname{supp} \mu \qquad \text{for all } \mu \in \mathcal{P}(\mathsf{E}),\ z \in \mathsf{E},\ h = 1, \dots, H. \tag{10}$$

Remark. The support condition (10) says that we may only access information provided in the current context, which is derived from the sampled points on the graph of the input function f. We have no access to information about f outside the sampled points. The read heads mix information spatially across the embedding space E, conversely the channel-mixers are pointwise in E.

We use Markov kernel in-context maps to define neural network layers. The input to the layer is a current context that was output from the previous layer, and was ultimately derived from the sampled points of the input function. The layer updates the context by applying the in-context map to each point; that is, the context is pushed-forward by the in-context map.

Definition 3.2 (Markov kernel layer)**.** A Markov kernel layer $B : \mathcal{P}(\mathsf{E}) \to \mathcal{P}(\mathsf{E}')$ determined by the Markov kernel in-context map $\Gamma : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \to \mathsf{E}'$ is defined

$$B_\Gamma(\mu) = \Gamma(\mu, \cdot)_{\#}\mu. \tag{11}$$

Admittedly, this notation is quite compact. Appendix A.2 contains an explicit construction of a softmax attention layer from a transformer using our notation to provide a link to a farmiliar setting.

We can now define the full Markov kernel network abstract architecture. A Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ encodes the input context into a lifted embedding space, updates the context using Markov kernel layers, then decodes the final context at a query point to obtain the output.

Definition 3.3 (Markov kernel network)**.** Let $L > 0$, and E_ℓ be a compact metric space for $\ell = 0, \dots, L$. A *Markov kernel network* $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ is of the form

$$\begin{aligned} \mu_0 &:= q_{\mathsf{G}\#}\mu, \\ \mu_L &:= (B_{L-1} \circ \cdots \circ B_0)(\mu_0), \qquad && \mu \in \mathcal{P}(\mathsf{G}),\ x \in \mathsf{X} \\ \mathcal{N}(\mu, x) &:= \Gamma_p\big(\mu_L, q_{\mathsf{X}}(x)\big), \end{aligned} \tag{12}$$

where

- $q_{\mathsf{G}} : \mathsf{G} \to \mathsf{E}_0$ and $q_{\mathsf{X}} : \mathsf{X} \to \mathsf{E}_L$ are pointwise encoders that lift inputs to hidden embedding spaces. We note that $q_{\mathsf{X}} = q_{\mathsf{X}}(x)$ is independent of the context μ.
- $B_\ell : \mathcal{P}(\mathsf{E}_\ell) \to \mathcal{P}(\mathsf{E}_{\ell+1})$ are Markov kernel layers for $\ell = 0, \dots, L-1$.
- $\Gamma_p : \mathcal{P}(\mathsf{E}_L) \times \mathsf{E}_L \to \mathsf{Y}$ is a decoder in the form of a Markov kernel in-context map.

Remark. Markov kernel networks are a shared generalization of transformers (Vaswani et al., 2017), neural operators (Kovachki et al., 2023), and Flowers (Muser et al., 2026). Different concrete architectures are obtained by specifying the Markov kernels K, value maps v, and channel mixers m (all of these for both the hidden context layers B_ℓ and the decoder Γ_p), as well as the encoders q_{G} and q_{X}. The popular architectures listed above can largely be classified by how information is routed through the network by their Markov kernels. Transformers are distinguished by their specific form of scaled dot-product attention. Nonlinear neural operators can be realized by multiple heads and a signed-measure decomposition of a nonlinear integral kernel $k(\mu, z, w)\,\mu(\mathrm{d}w)$; Fourier neural operators (Li et al., 2021) can be realized by the simple choice $K[\mu, z](\mathrm{d}w) = \mu$ with spectral computations occuring in v and m. Flowers, on the other hand, uses Markov kernels that output a Dirac measure.

Furthermore, branch-and-trunk style networks, such as DeepONet (Chen & Chen, 1995; Lu et al., 2019) also fit into the Markov kernel network abstract architecture. The distinguishing feature of these networks are the forms of their decoders Γ_p.

Example constructions of various popular architectures in the notation of Markov kernel networks can be found in Appendix D.

3.2 A REPRESENTATION THEOREM FOR UNIFORMLY DISCRETIZATION-STABLE OPERATORS

We now define *stable Markov kernel networks* as the subset of Definition 3.3 whose individual components all satisfy certain stability properties, and show this guarantees the entire network is uniformly discretization stable. We then state our main result that every operator that is uniformly discretization stable on a class of input measures satisfying Definition 2.2 is exactly representable as a stable Markov kernel network.

Definition 3.4 (Stable Markov kernel network)**.** A Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ is *stable* if the following hold:

1. The encoders q_{G}, q_{X}, value maps v_h, and channel mixers m are Lipschitz.
2. In every hidden layer B_ℓ, every Markov kernel $K_h : \mathcal{P}(\mathsf{E}_\ell) \times \mathsf{E}_\ell \to \mathcal{P}(\mathsf{E}_\ell)$ is Hölder stable in both arguments: there exists $C > 0$ and $\beta \in (0, 1]$ such that
$$W_1\big(K_h[\mu, w], K_h[\nu, z]\big) \le C\big(W_1(\mu, \nu) + \rho_{\mathsf{E}}(w, z)\big)^\beta \text{ for all } \mu, \nu \in \mathcal{P}(\mathsf{E}_\ell),\ w, z \in \mathsf{E}_\ell, \tag{13}$$
3. In the decoder Γ_p, every Markov kernel $K_h : \mathcal{P}(\mathsf{E}_L) \times \mathsf{E}_L \to \mathcal{P}(\mathsf{E}_L)$ is Hölder stable in the context: there exists $C > 0$ and $\beta \in (0, 1]$ such that
$$W_1\big(K_h[\mu, z], K_h[\nu, z]\big) \le C W_1(\mu, \nu)^\beta \qquad \text{for all } \mu, \nu \in \mathcal{P}(\mathsf{E}_L),\ z \in \mathsf{E}_L. \tag{14}$$

Remark. In particular, the decoder Γ_p is not required to be continuous with respect to its second argument $z = q_{\mathsf{X}}(x)$.

Theorem 3.5. *Every stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ is uniformly discretization stable on $\mathcal{S} = \mathcal{P}(\mathsf{G})$.*

Our main result gives a constructive extension of operators that are uniformly discretization stable on admissible sets of non-collapsing measures to uniformly discretization stable operators on the entire graph space, with the extension being a stable Markov kernel network.

Theorem 3.6. *Let $\mathcal{S} \subset \mathcal{P}(\mathsf{G})$ be an admissible set of non-collapsing measures, and $\mathcal{N}^* : \mathcal{S} \times \mathsf{X} \to \mathsf{Y}$ be uniformly discretization stable on $\mathcal{S}$. Then there exists a stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ such that $\mathcal{N}|_{\mathcal{S}\times\mathsf{X}} = \mathcal{N}^*$.*

The main idea in the proof is to normalize spatial features and calibrate the decoder reads using techniques from de Carvalho et al. (2013). The proof is in Appendix B.

Example 3.7 (Counterexample when $\mathcal{S} = \mathcal{P}(\mathsf{G})$)**.** We can not use stable Markov kernel networks to represent all uniformly discretization stable operators without restricting the domain of input measures $\mathcal{S}$. Let $\mathsf{X} = \mathsf{Y} = [0, 1]$ equiped with the Lebesgue measure λ, and consider $\mathcal{N}^* : \mathcal{P}(\mathsf{G}) \times$

$\mathsf{X} \to \mathsf{Y}$ defined by

$$\mathcal{N}^*(\mu, x) = \sqrt{\int_{\mathsf{G}} s \; \mu(\mathrm{d}s, \mathrm{d}y)}. \tag{15}$$

This operator is uniformly discretization stable on the entire domain $\mathcal{S} = \mathcal{P}(\mathsf{G})$. It is constant on graph measures, $\mathcal{N}^*(\gamma_f, x) = 1/\sqrt{2}$ for all $f \in L^2[\mathsf{X}, \mathsf{Y}]$. However, for Dirac inputs $\delta_{(t,0)}, t \in [0, 1]$, the map $t \mapsto \mathcal{N}^*(\delta_{(t,0)}, x) = \sqrt{t}$ is not Lipschitz. On the other hand, every stable Markov kernel networks $\mathcal{N}$ forces the map $t \mapsto \mathcal{N}(\delta_{(t,0)}, x)$ to be Lipschitz (the support condition (10) forces every read head for Dirac inputs to be $r_h(\delta_z, z) = v_h(z)$, and the stable network becomes a composition of Lipshitz maps).

Example 3.7 illustrates the importance of the non-collapse condition (7) on the domain $\mathcal{S}$ in Theorem 3.6. We can match the behavior of $\mathcal{N}^*$ with a stable Markov kernel network on any reasonable discretized input measure, even very noisy approximations of graph measures, but not its behavior when the spatial marginal collapses. However, the theorem shows that when we restrict the domain of $\mathcal{N}^*$ to an admissible class $\mathcal{S}$, we can construct a stable Markov network that matches the behavior of $\mathcal{N}^*$ exactly on $\mathcal{S}$, and the resulting network is uniformly discretization stable on all of $\mathcal{P}(\mathsf{G})$.

We close this section by characterizing the operators represented by stable Markov kernel networks.

Corollary 3.8 (Operator characterization of stable Markov kernel networks)**.** *Let $A : L^2(\mathsf{X}, \mathsf{Y}) \to L^2(\mathsf{X}, \mathsf{Y})$ be a possibly non-linear operator mapping functions to functions. Then there exists $C > 0$, $\beta \in (0, 1]$ such that*

$$\|A[f] - A[g]\|_{L^\infty(\mathsf{X},\mathsf{Y})} \le C W_1(\gamma_f, \gamma_g)^\beta \qquad \textit{for all } f, g \in L^2(\mathsf{X}, \mathsf{Y}) \tag{16}$$

if and only if there exists a stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ such that

$$A[f](x) = \mathcal{N}(\gamma_f, x) \qquad \textit{for all } f \in L^2(\mathsf{X}, \mathsf{Y}) \textit{ and } \lambda\textit{-a.e. } x \in \mathsf{X}. \tag{17}$$

4 Variational principles for Markov kernel networks

Theorems 3.5 and 3.6 define the design space of uniformly discretization stable neural networks to be the stable Markov kernel networks, and Definition 3.4 gives stability properties of the individual components $K_h, v_h, m, q_{\mathsf{G}}, q_{\mathsf{X}}$ in such networks. Exploring new architectures in the design space becomes a question of specifying parameterizations or fixed implementations of these components. We propose one approach for designing the Markov kernels K_h in (8) based on variational principles. We fix q_{G}, q_{X}, and every v_h and m. The proof of Theorem 3.6 shows this can be done in a manner that does not diminish the expressive power of the network itself.

Our approach is inspired by the variational characterization of softmax from Niculae & Blondel (2017). For key, query matrices K, Q and temperature $\tau > 0$, the minimizer of

$$\operatorname*{arg\,min}_{\kappa \in \mathcal{P}(\mathsf{E})} \left\{ -\int_{\mathsf{E}} \langle Qz, \, Kw \rangle \, \kappa(dw) + \tau \, \mathrm{KL}(\kappa \| \mu) \right\}$$

is the classical softmax attention kernel

$$K[\mu, z](dw) = \frac{\exp\left(\frac{\langle Qz, \, Kw \rangle}{\tau}\right)}{\int_{\mathsf{E}} \exp\left(\frac{\langle Qz, \, Kw' \rangle}{\tau}\right) \, \mu(dw')} \, \mu(dw).$$

We consider a general cost $c : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \times \mathsf{E} \to \mathbb{R}$; and a regularizer $\mathcal{R} : \mathcal{P}(\mathsf{E}) \times \mathcal{P}(\mathsf{E}) \to [0, +\infty]$. Set

$$K[\mu, z] \in \operatorname*{arg\,min}_{\substack{\kappa \in \mathcal{P}(\mathsf{E}) \\ \operatorname{supp} \kappa \subseteq \operatorname{supp} \mu}} J_{\mu,z}(\kappa). \qquad J_{\mu,z}(\kappa) := \int_{\mathsf{E}} c(\mu, z, w) \kappa(dw) + \mathcal{R}(\kappa, \mu). \tag{18}$$

Here, $c(\mu, z, w)$ is the cost of reading from location w based on the input context μ and query z. The regularizer $\mathcal{R}$ penalizes large deviations from μ. The support condition restricts reads to the context μ, so when a minimizer K exists, it provides legal read behavior for a Markov kernel network consistent with Definition 3.1.

From this point of view, the design space is $(c, \mathcal{R})$-space. We ask: what are the conditions on costs and regularizers that guarantee the associated kernel is sufficiently stable? We can answer this question for a large class of regularizers constructed from f-divergences and Wasserstein distances.

4.1 Regularizers from f-divergences

An f-divergence $D_\phi : \mathcal{P}(\mathsf{E}) \times \mathcal{P}(\mathsf{E}) \to [0, +\infty]$ with generator $\phi : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ is of the form

$$D_\phi(\kappa \,\|\, \mu) = \begin{cases} \int_{\mathsf{E}} \phi(\frac{d\kappa}{d\mu}) d\mu, & \kappa \ll \mu \\ \infty, & \kappa \not\ll \mu. \end{cases} \tag{19}$$

where ϕ is convex and $\phi(1) = 0$. We further define $\phi(0) = \lim_{t\to 0} \phi(t)$, which could be infinite. This class covers many common divergences, including KL, Hellinger, Jensen-Shannon, and χ^2. Not all f-divergence regularizers lead to Markov kernels with our required stability properties; the next proposition gives sufficient conditions on the f-divergence regularizer and cost function for this to happen.

Proposition 4.1 (Kernel stability for f-divergence regularizers)**.** *Let $c : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \times \mathsf{E} \to \mathbb{R}$ satisfy*

$$\begin{aligned} |c(\mu, z, w)| &\le C, \\ |c(\mu, z, w) - c(\mu', z', w')| &\le L\rho_{\mathsf{E}}(w, w') + L(\rho_{\mathsf{E}}(z, z') + W_1(\mu, \mu'))^\beta, \end{aligned} \tag{20}$$

for some $C, L > 0$, $\beta \in (0, 1]$. Let D_ϕ have generator $\phi : [0, +\infty) \to [0, +\infty]$ that is lower semicontinuous, strictly convex, and finite and C^1 on $(0, +\infty)$, with $\phi(1) = \phi'(1) = 0$ and $\lim_{r\to+\infty} \phi'(r) = +\infty$. Further suppose the function

$$h(t) = \begin{cases} 0, & t \le \phi'(0+), \\ (\phi')^{-1}(t), & t > \phi'(0+) \end{cases} \tag{21}$$

is Hölder stable on $[-2C, 2C]$. Then (18) *with cost function c and regularizer $\mathcal{R} = D_\phi$ has a unique minimizer K satisfying the stability condition* (13)*. Moreover, the kernel takes the form*

$$K[\mu, z](dw) = h(b_{\mu,z} - c(\mu, z, w))\, \mu(dw), \qquad \int_{\mathsf{E}} h(b_{\mu,z} - c(\mu, z, w))\, \mu(dw) = 1, \tag{22}$$

where $b_{\mu,z} \in [-C, C]$ is the unique normalizing constant.

A proof is in Appendix B.2. Table 1 gives a non-exhaustive list of discretization stable Markov kernels generated in this way. We further note that the Pearson χ^2 regularizer corresponds to the construction used in the proof of Theorem 3.6.

4.2 Regularizers from projected Wasserstein distances

Regularizers constructed from f-divergences cover a range of Markov kernel read behavior, but they do not recover the SelfWarp behavior from Flowers. For this, we turn to regularizers built from projected Wasserstein distances. Let $\pi_{\mathsf{X}} : \mathsf{E} \to \mathsf{X}$ be fixed, continuous and surjective. Define

$$\mathcal{R}(\kappa, \mu) := \frac{1}{2} W_2(\pi_{\mathsf{X}\#}\kappa, \pi_{\mathsf{X}\#}\mu)^2. \tag{23}$$

The function π_{X} should be interpreted as a projection. The regularizer above penalizes reads in an optimal transport distance on the spatial domain X.

Proposition 4.2. *Let $\mathcal{R}$ be given by equation* (23)*. If $(\mu, z, w) \mapsto c(\mu, z, w)$ is continuous in w, then there exists a Markov kernel K such that $K[\mu, z]$ satisfies* (18) *for all μ and z.*

To recover SelfWarp on continuous graph contexts γ_f, let $\tau_h(x, a) = x + g_h(a) \in \mathsf{X}$, with periodic wrapping when appropriate. For $B > \operatorname{diam}(\mathsf{X})$, choose

$$c_h(\mu, z, w) = B\rho_{\mathsf{X}}\big(\pi_{\mathsf{X}}(w), \tau_h(z)\big).$$

The unique minimizing kernel is $K_h[\gamma_f, z] = \delta_{(\tau_h(z), f(\tau_h(z)))}$. With value map $v_h(y, b) = V_h b$, this gives the SelfWarp read $V_h f(x + g_h(f(x)))$ at $z = (x, f(x))$. The precise statement and proof are in Appendix B.3.

Divergence	Generator $\phi(r)$	Response $h(t)$	Attention mechanism
KL	$r \log r - r + 1$	e^t	Softmax (Vaswani et al., 2017)
Pearson χ^2	$\frac{1}{2}(r-1)^2$	$(1+t)_+$	Sparsemax (Martins & Astudillo, 2016)
Tsallis / power, $q > 1$	$\frac{r^q - qr + q - 1}{q(q-1)}$	$[1 + (q-1)t]_+^{1/(q-1)}$	Entmax (Martins et al., 2022)
Shifted entropy, $a > 0$	$(r+a)\log\frac{r+a}{1+a} - r + 1$	$[(1+a)e^t - a]_+$	—
KL plus Pearson, $a > 0$	$r \log r - r + 1 + \frac{a}{2}(r-1)^2$	$a^{-1} W_0(a e^{t+a})$	—
Centered power, $p > 1$	$\frac{1}{p}\lvert r-1\rvert^p$	$[1 + \mathrm{sgn}(t)\lvert t\rvert^{1/(p-1)}]_+$	Sparsemax for $p = 2$ (Martins & Astudillo, 2016)
Quadratic plus reverse KL	$\frac{r^2 - 1 - 2\log r}{4}$	$t + \sqrt{1+t^2}$	—
Jeffreys	$(r-1)\log r$	$[W_0(e^{1-t})]^{-1}$	f-softargmax (Roulet et al., 2025)

Table 1: Examples of f-divergences consistent with the hypotheses of Proposition 4.1. W_0 denotes the principal real Lambert W function. The associated stable read kernel K is given by Equation (22). A dash in the last column indicates we are unaware of previous investigations of these attention mechanisms.

Remark. Our final question is: can we construct a single parameterized family of regularizers such that the read behavior of FNO, transformers, and Flowers are recovered in different parameter regimes? We show this is possible in Appendix E, which constructs a convex space of regularizers as a linear combination of spatial Wasserstein and f-divergences such that the resulting Markov kernel reads are discretization stable. In future work, we aim to explore this space using agents with the goal of adaptively choosing architectures suited to a particular operator learning problem at hand.

5 Limitations

The representation theorem presented here does not specify how to choose subclasses of Markov kernel networks suitable for a particular task, nor does it determine the most efficient implementations. For example, although it is possible to implement the GroupNorm operation as part of a Markov kernel layer (see Equation (75) in Section D.5), it is another question to choose a efficient implementation in code. The practical importance of depth is not reflected in the proof of the representation theorem, which requires only two layers. Additionally, the kernels K_h in the representation might be complicated.

Acknowledgments

(Redacted for reviewing)

AI USE STATEMENT

The main portion of this document was written by humans; there is no LLM-generated text appearing on any page preceding this statement, except for light line-level edits. The tables and figures were generated with help from ChatGPT 6 Astra. The appendices interleaves text generated by ChatGPT 6 Astra and humans; all LLM-generated text was reviewed and edited by the authors for clarity and correctness. We also used ChatGPT and Claude for non-writing tasks, such as literature searches and exploring mathematical ideas.

REFERENCES

Advait Balaji, Trevor Teolis, S. David Mis, Jose Antonio Lara Benitez, Chao Wang, and Maarten V. de Hoop. Hybrid operator learning of wave scattering maps in high-contrast media, 2026. URL https://arxiv.org/abs/2602.11197v1. Preprint. Local PDF: arXiv:2602.11197v1.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonić, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.19913v2. Local copy: arXiv v2, 2 November 2023, 28 pages.

Abderrahim Bendahi, Adrien Fradin, Johan Peralez, Julie Digne, and Madiha Nadri. Stability and discretization error of state space model neural operators, 2026. URL https://arxiv.org/abs/2605.18905v1. Local copy: arXiv v1, 17 May 2026, 51 pages.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How Smooth Is Attention? In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2312.14820v2. Local PDF: arXiv:2312.14820v2.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995. doi: 10.1109/72.392253.

Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. AdaNet: Adaptive structural learning of artificial neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 874–883, 2017. URL https://proceedings.mlr.press/v70/cortes17a.html.

Barbara D'Ambrogi, Sari Mäenpää, and Markku Markkanen. Discretization independent retrieval of atmospheric ozone profile. *Geophysica*, 35(1-2):87–99, 1999. URL https://archive.geophysica.fi/pdf/geophysica_1999_35_1-2_087_dambrogi.pdf.

Gwendoline de Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1556–1565, 2019. URL https://proceedings.mlr.press/v97/de-bie19a.html. Local PDF: 10-page published ICML 2019 version; identity and Proposition 1 verified 21 September 2026.

Miguel de Carvalho, Boris Oumow, Johan Segers, and Michał Warchoł. A Euclidean likelihood estimator for bivariate tail dependence. *Communications in Statistics—Theory and Methods*, 42 (7):1176–1192, 2013. doi: 10.1080/03610926.2012.709905. URL https://arxiv.org/abs/1204.3524.

Takashi Furuya, Maarten V. de Hoop, and Matti Lassas. Transformers through the lens of support-preserving maps between measures, 2025a. URL https://arxiv.org/abs/2509.25611v1. Version 1, 30 September 2025; Local PDF: arXiv:2509.25611v1.

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://arxiv.org/abs/2408.01367v2. arXiv:2408.01367; Local PDF: arXiv:2408.01367v2.

Takashi Furuya, David Mis, Ivan Dokmanić, Maarten V. de Hoop, and Matti Lassas. Function graph transformers universally approximate operators between function spaces, 2026. URL https://arxiv.org/abs/2605.17968v1. Local PDF: arXiv:2605.17968v1.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? on the discretization mismatch errors in neural operators. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/313829757739365201b5adb3a1cbd9bd-Abstract-Conference.html. Local copy: ICLR 2025 publisher PDF, 25 pages including appendices.

Nicolás García Trillos and Dejan Slepčev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220:193–241, 2016. doi: 10.1007/s00205-015-0929-z. URL https://arxiv.org/abs/1403.6355v3. Local PDF is arXiv version 3, September 2014.

C. J. Himmelberg and F. S. Van Vleck. Some selection theorems for measurable functions. *Canadian Journal of Mathematics*, 21:394–399, 1969. doi: 10.4153/CJM-1969-041-7. URL https://www.cambridge.org/core/journals/canadian-journal-of-mathematics/article/some-selection-theorems-for-measurable-functions/30345D24D3DFFAAF9CE805C30EE959C6. Theorem 1, p. 395, restates the measurable selection theorem of Kuratowski and Ryll-Nardzewski (1965); local PDF is the published version.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://arxiv.org/abs/2108.08481v6. Local PDF: arXiv:2108.08481v6.

Samuel Lanthaler, Andrew M. Stuart, and Margaret Trautner. Discretization error of fourier neural operators, 2024. URL https://arxiv.org/abs/2405.02221. Version 2, 25 September 2025.

Sari Lasanen. *Discretizations of Generalized Random Variables with Applications to Inverse Problems*. PhD thesis, University of Oulu, Helsinki, 2002. URL https://www.finna.fi/Record/utu.999531305405971. 64 pages.

Matti Lassas and Samuli Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563, 2004. doi: 10.1088/0266-5611/20/5/013. URL https://siltanen-research.net/wp-content/uploads/publications/LassasSiltanen2004.pdf.

Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009. doi: 10.3934/ipi.2009.3.87. URL https://www.aimsciences.org/article/doi/10.3934/ipi.2009.3.87.

M. S. Lehtinen, L. Päivärinta, and E. Somersalo. Linear inverse problems for generalised random variables. *Inverse Problems*, 5(4):599–612, 1989. doi: 10.1088/0266-5611/5/4/011. URL https://doi.org/10.1088/0266-5611/5/4/011.

Zongyi Li, Nikola 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://arxiv.org/abs/2010.08895v3. Local PDF: arXiv:2010.08895v3.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators, 2019. URL https://arxiv.org/abs/1910.03193v3. Preprint. Local PDF: arXiv:1910.03193v3.

André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger

(eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623. PMLR, 2016. URL https://proceedings.mlr.press/v48/martins16.html. Local PDF: publisher conference version, 10 pages; title, authors, imprint and sparsemax formulas verified 20 September 2026.

André F. T. Martins, Marcos Treviso, António Farinhas, Pedro M. Q. Aguiar, Mário A. T. Figueiredo, Mathieu Blondel, and Vlad Niculae. Sparse continuous distributions and Fenchel–Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022. URL https://jmlr.org/papers/v23/21-0879.html.

Till Muser, Alexandra Spitzer, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić. Flowers: A warp drive for neural PDE solvers. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.04430v2. Version 2, 1 June 2026; Local PDF: arXiv:2603.04430v2.

Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3338–3348. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2d1b2a5ff364606ff041650887723470-Abstract.html.

Vincent Roulet, Tianlin Liu, Nino Vieillard, Michaël E. Sander, and Mathieu Blondel. Loss functions and operators generated by f-divergences. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52110–52138. PMLR, 2025. URL https://proceedings.mlr.press/v267/roulet25a.html.

Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/12bcd658ef0a540cabc36cdf2b1046fd-Abstract.html.

Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061. URL https://doi.org/10.1017/S0962492910000061.

Unique Subedi and Ambuj Tewari. Controlling statistical, discretization, and truncation errors in learning fourier linear operators, 2025. URL https://arxiv.org/abs/2408.09004. Version 2, 6 February 2025; first posted in 2024 as Error Bounds for Learning Fourier Linear Operators.

Dávid Terjék. Moreau-yosida f-divergences. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10214–10224. PMLR, 2021. URL https://proceedings.mlr.press/v139/terjek21a.html. Local PDF: published ICML proceedings version, 11 pages; title, author, and Section 4 verified 20 September 2026.

Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K. Rohde, and Dejan Slepčev. A transportation L^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2): 187–210, 2017. doi: 10.1007/s10851-017-0726-4. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6141213/.

Juan Diego Toscano, Daniel T. Chen, and George Em Karniadakis. Athena: Agentic team for hierarchical evolutionary numerical algorithms, 2026. URL https://arxiv.org/abs/2512.03476.

Michael Unser. A representer theorem for deep neural networks. *Journal of Machine Learning Research*, 20(110):1–30, 2019. URL https://www.jmlr.org/papers/v20/18-418.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.03762v7. Local PDF: arXiv:1706.03762v7.

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

Clinton J. Wang and Polina Golland. Discretization invariant networks for learning maps between neural fields. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=CpYBAqDgmz.

A Notation and background

We use sans-serif letters for metric spaces, lower-case Greek letters for measures, and ρ for a metric. The notation used most often is collected below.

Notation	Description
$(\mathsf{X}, \rho_{\mathsf{X}})$	Spatial domain
$\mathsf{Y} \subset \mathbb{R}^{d_{\text{out}}}$	Compact convex feature range
$\mathsf{G} = \mathsf{X} \times \mathsf{Y}$	Graph metric space
E_ℓ	Embedding space at layer ℓ
$\mathcal{P}(\mathsf{M})$	Radon probability measures on M
$\mathcal{P}_p(\mathsf{M})$	Such measures with finite pth moment
λ	Reference probability measure on X
$\gamma_f = (\mathrm{id}, f)_\# \lambda$	Graph measure of f
δ_z	Dirac probability measure at z
$\operatorname{supp} \mu$	Support of μ

A Borel probability measure μ on a metric space M is Radon if

$$\mu(A) = \sup\{\mu(K) : K \subseteq A \text{ is compact}\}$$

for every Borel set A. Every Borel probability measure on a compact metric space is Radon.

For a Borel map $F : \mathsf{M} \to \mathsf{M}'$, the pushforward of μ is

$$F_\# \mu(B) = \mu(F^{-1}(B)), \qquad B \subseteq \mathsf{M}' \text{ Borel.} \tag{24}$$

Equivalently, $\int \phi \, \mathrm{d}(F_\# \mu) = \int \phi \circ F \, \mathrm{d}\mu$ for every bounded Borel function ϕ.

Definition A.1 (Wasserstein distances)**.** For $1 \le p < \infty$, let $\mathcal{P}_p(\mathsf{M})$ denote the Radon probability measures satisfying $\int \rho_{\mathsf{M}}(z, z_0)^p \, \mu(\mathrm{d}z) < \infty$ for one, and hence every, $z_0 \in \mathsf{M}$. For $\mu, \nu \in \mathcal{P}_p(\mathsf{M})$, define

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathsf{M} \times \mathsf{M}} \rho_{\mathsf{M}}(z, w)^p \, \pi(\mathrm{d}z, \mathrm{d}w) \right)^{1/p}, \tag{25}$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathsf{M} \times \mathsf{M}$ with marginals μ and ν.

On a compact metric space, $\mathcal{P}_p(\mathsf{M}) = \mathcal{P}(\mathsf{M})$ and W_p metrizes weak convergence for every finite p (Villani, 2009, Chapter 6). This applies to G and the embedding spaces E_ℓ throughout the paper.

We regard $L^2(\mathsf{X}, \mathsf{Y})$ as equivalence classes modulo equality λ-almost everywhere. Borel representatives are used when pointwise values are needed. The graph measure is independent of this choice, and $\gamma_f = \gamma_g$ if and only if $f = g$ λ-almost everywhere. An L^∞ norm is an essential supremum; the supremum in (DS) is over every $x \in \mathsf{X}$.

A.1 Sequences and empirical measures

A sequence $z_1, \ldots, z_N$ determines the empirical measure $\mu = N^{-1} \sum_i \delta_{z_i}$. For fixed N, this identifies sequences up to permutation. For two equally weighted empirical measures on a metric space,

$$W_1\left(\frac{1}{N} \sum_i \delta_{z_i}, \frac{1}{N} \sum_i \delta_{w_i} \right) = \min_{\sigma \in S_N} \frac{1}{N} \sum_i \rho_{\mathsf{M}}(z_i, w_{\sigma(i)}).$$

The same measure can arise from different sequence lengths by repeating every token. A map defined on measures must give the same result under such repetitions. The attention formula below has this property.

A.2 SOFTMAX ATTENTION AS A CONTEXT PUSHFORWARD

Consider tokens $z_1, \dots, z_N$ in a compact subset E of a Euclidean space. For query, key, value, and output matrices Q, C, V, W, put

$$s(z,w) = \frac{\langle Qz, Cw\rangle}{\sqrt{k}}, \qquad K[\mu,z](\mathrm{d}w) = \frac{e^{s(z,w)}}{\int_{\mathsf{E}} e^{s(z,u)}\,\mu(\mathrm{d}u)}\,\mu(\mathrm{d}w),$$

where k is the query/key dimension. The value map is $v(w) = Vw$, and the mixer $m(z,r) = z + Wr$ includes a residual connection. For $\mu = N^{-1}\sum_j \delta_{z_j}$, this gives

$$\Gamma(\mu, z_i) = z_i + W\sum_{j=1}^{N} \frac{e^{s(z_i,z_j)}}{\sum_{q=1}^{N} e^{s(z_i,z_q)}} V z_j, \qquad B_\Gamma(\mu) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\Gamma(\mu,z_i)}. \tag{26}$$

This is the usual single-head softmax attention update, written as $B_\Gamma(\mu) = \Gamma(\mu,\cdot)_{\#}\mu$. For a weighted context $\sum_j \omega_j \delta_{z_j}$, the attention weights are proportional to $\omega_j e^{s(z_i,z_j)}$. Section D.2 gives the full block, including multiple heads, normalization, and the feed-forward map.

B PROOFS

B.1 SECTION 3

Theorem 3.5. *Every stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ is uniformly discretization stable on $\mathcal{S} = \mathcal{P}(\mathsf{G})$.*

Proof. We first derive stability of the in-context maps from the component assumptions in Definition 3.4. Consider a hidden layer or the decoder, with input space E, output space E', and H heads. Write $A_h > 0$ and $\beta_h \in (0,1]$ for its kernel stability constants, and let $V_h > 0$ and $M > 0$ be Lipschitz constants for its value maps and mixer, using the sum product metric for the mixer. Set

$$\beta = \min_{1\le h\le H} \beta_h, \qquad D = \max\{1, 2\,\mathrm{diam}(\mathsf{E})\}.$$

Since every distance in the kernel bounds is at most D, the inequality $t^{\beta_h} \le D^{\beta_h-\beta} t^\beta$ allows us to use the common exponent β after enlarging the constants A_h.

For probability measures $\alpha, \alpha' \in \mathcal{P}(\mathsf{E})$ and any coupling π of them, Lipschitz continuity of v_h gives

$$\left\| \int v_h \,\mathrm{d}\alpha - \int v_h \,\mathrm{d}\alpha' \right\| \le \int \|v_h(u) - v_h(v)\|\,\pi(\mathrm{d}u, \mathrm{d}v) \le V_h \int \rho_{\mathsf{E}}(u,v)\,\pi(\mathrm{d}u,\mathrm{d}v).$$

Taking the infimum over couplings, each read therefore satisfies

$$\|r_h(\mu,w) - r_h(\nu,z)\| \le V_h W_1\big(K_h[\mu,w], K_h[\nu,z]\big).$$

For a hidden layer, put $t = W_1(\mu,\nu) + \rho_{\mathsf{E}}(w,z)$. Joint kernel stability and the Lipschitz mixer imply

$$\begin{aligned}\rho_{\mathsf{E}'}\big(\Gamma(\mu,w), \Gamma(\nu,z)\big) &\le M\left(\rho_{\mathsf{E}}(w,z) + \sum_{h=1}^{H} V_h A_h t^\beta\right) \\ &\le M\left(D^{1-\beta} + \sum_{h=1}^{H} V_h A_h\right) t^\beta.\end{aligned}$$

Here $t \le D$ and $\beta \le 1$ imply $t \le D^{1-\beta} t^\beta$. For the decoder, the same argument at a fixed query gives

$$\rho_{\mathsf{E}'}\big(\Gamma(\mu,z), \Gamma(\nu,z)\big) \le M\left(\sum_{h=1}^{H} V_h A_h\right) W_1(\mu,\nu)^\beta,$$

uniformly in z. Applying these estimates to each layer yields joint stability constants $C_\ell > 0$, $\beta_\ell \in (0,1]$ for Γ_ℓ and context-only stability constants $C_p > 0$, $\beta_p \in (0,1]$ for Γ_p.

We next show that each context update B_ℓ is Hölder continuous in W_1. Fix $\eta, \xi \in \mathcal{P}(\mathsf{E}_\ell)$, write $d = W_1(\eta, \xi)$, and choose an optimal coupling π of η and ξ. Such a coupling exists because E_ℓ is compact. The pushforward of π under

$$(w, z) \longmapsto \big(\Gamma_\ell(\eta, w), \Gamma_\ell(\xi, z)\big)$$

is a coupling of $B_\ell(\eta)$ and $B_\ell(\xi)$. Consequently, joint stability gives

$$\begin{aligned} W_1\big(B_\ell(\eta), B_\ell(\xi)\big) &\le C_\ell \int \big(d + \rho_{\mathsf{E}_\ell}(w, z)\big)^{\beta_\ell} \, \pi(\mathrm{d}w, \mathrm{d}z) \\ &\le C_\ell \left(d + \int \rho_{\mathsf{E}_\ell}(w, z) \, \pi(\mathrm{d}w, \mathrm{d}z)\right)^{\beta_\ell} \\ &= 2^{\beta_\ell} C_\ell \, d^{\beta_\ell}. \end{aligned}$$

The second inequality is Jensen's inequality for the concave function $t \mapsto t^{\beta_\ell}$, and the last equality uses optimality of π.

Now fix arbitrary inputs $\mu, \nu \in \mathcal{P}(\mathsf{G})$, and denote their successive encoded contexts by μ_ℓ, ν_ℓ. If $Q > 0$ is a Lipschitz constant for q_{G}, then

$$W_1(\mu_0, \nu_0) = W_1\big(q_{\mathsf{G}\#}\mu, q_{\mathsf{G}\#}\nu\big) \le Q \, W_1(\mu, \nu).$$

Iterating the context-update estimate therefore yields a constant $A > 0$, depending only on the network, such that

$$W_1(\mu_L, \nu_L) \le A \, W_1(\mu, \nu)^\theta, \qquad \theta = \prod_{\ell=0}^{L-1} \beta_\ell.$$

At each output location x, both evaluations use the same query $q_{\mathsf{X}}(x)$. The decoder's stability bound is uniform over these queries, so

$$\begin{aligned} \sup_{x \in \mathsf{X}} |\mathcal{N}(\mu, x) - \mathcal{N}(\nu, x)| &= \sup_{x \in \mathsf{X}} \big|\Gamma_p(\mu_L, q_{\mathsf{X}}(x)) - \Gamma_p(\nu_L, q_{\mathsf{X}}(x))\big| \\ &\le C_p \, W_1(\mu_L, \nu_L)^{\beta_p} \\ &\le C_p A^{\beta_p} \, W_1(\mu, \nu)^{\beta_p \theta}. \end{aligned}$$

Since $\beta_p \theta \in (0, 1]$ and μ, ν were arbitrary, this is uniform discretization stability on $\mathcal{P}(\mathsf{G})$. □

Theorem 3.6. *Let $\mathcal{S} \subset \mathcal{P}(\mathsf{G})$ be an admissible set of non-collapsing measures, and $\mathcal{N}^* : \mathcal{S} \times \mathsf{X} \to \mathsf{Y}$ be uniformly discretization stable on $\mathcal{S}$. Then there exists a stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ such that $\mathcal{N}|_{\mathcal{S} \times \mathsf{X}} = \mathcal{N}^*$.*

Proof. We use one context layer to calibrate finitely many spatial values so that they have zero mean and unit total variance on $\mathcal{S}$. The decoder then recovers the target through weighted averages, followed by a fixed linear readout and projection onto Y.

Step 1: Extend the target in the context variable. Let $C > 0$ and $\beta \in (0, 1]$ be the discretization-stability constants of $\mathcal{N}^*$, write $c = d_{\text{out}}$, and choose $M \ge 1$ such that $\mathsf{Y} \subset [-M, M]^c$. Since $\mathcal{P}(\mathsf{G})$ is a compact metric space, $\mathcal{S}$ has a countable dense sequence $(\mu^{(i)})_{i \ge 1}$. For $j = 1, \dots, c$, define

$$U_j(\mu, x) = \mathrm{clip}_{[-M,M]}\left(\inf_{i \ge 1} \left\{\mathcal{N}_j^*(\mu^{(i)}, x) + C \, W_1(\mu, \mu^{(i)})^\beta\right\}\right), \qquad (\mu, x) \in \mathcal{P}(\mathsf{G}) \times \mathsf{X},$$

where $\mathrm{clip}_{[-M,M]}(t) = \max\{-M, \min\{t, M\}\}$. This is Borel measurable because the infimum is countable. The triangle inequality for the metric W_1^β, followed by the fact that clipping is 1-Lipschitz, gives

$$|U_j(\mu, x) - U_j(\nu, x)| \le C \, W_1(\mu, \nu)^\beta.$$

If $\mu \in \mathcal{S}$, discretization stability bounds every term in the infimum from below by $\mathcal{N}_j^*(\mu, x)$, while density provides a sequence of terms converging to that value. Thus $U = \mathcal{N}^*$ on $\mathcal{S} \times \mathsf{X}$. Globally, $|U| \le \sqrt{c}\, M$ and U is Hölder continuous in the context, uniformly in x, with constant $C_U = \sqrt{c}\, C$.

Step 2: Find finitely many spatial values with uniformly positive total variance. The set $\mathcal{T} = \overline{\mathcal{S}}$ is compact and remains non-collapsing: indeed,

$$\mu \longmapsto \inf_{a \in \mathsf{X}} W_1\big((\pi_{\mathsf{X}})_{\#}\mu, \delta_a\big)$$

is 1-Lipschitz. Consequently, every $\mu \in \mathcal{T}$ has a non-Dirac spatial marginal. For each such measure, some Lipschitz function $h : \mathsf{X} \to [0,1]$ has positive variance; one may take a rescaled distance from a point in the marginal's support. That variance remains positive in a neighborhood of μ. A finite subcover of $\mathcal{T}$ therefore provides Lipschitz functions $h_1, \dots, h_n : \mathsf{X} \to [0,1]$ whose variances have a strictly positive sum throughout $\mathcal{T}$.

Write $\phi_k = h_k \circ \pi_{\mathsf{X}}$ and, for every $\mu \in \mathcal{P}(\mathsf{G})$, set

$$a_k(\mu) = \int_{\mathsf{G}} \phi_k \,\mathrm{d}\mu, \qquad b_k(\mu) = \int_{\mathsf{G}} \phi_k^2 \,\mathrm{d}\mu - a_k(\mu)^2, \qquad V(\mu) = \sum_{k=1}^{n} b_k(\mu).$$

These functions are globally Lipschitz in W_1, since the ϕ_k are bounded and Lipschitz. By compactness there is $v_* > 0$ such that $V(\mu) \geq v_*$ on $\mathcal{T}$. Define

$$D(\mu) = \max\{V(\mu), v_*\}.$$

Step 3: Calibrate the values in a context layer. Take $L = 1$ and set

$$R = v_*^{-1/2}, \qquad \mathsf{E}_0 = \mathsf{G}, \qquad \mathsf{E}_1 = \mathsf{G} \times [-R, R]^n.$$

Equip E_1 with the sum of the graph metric and the Euclidean metric on the new coordinates. Choose $y_* \in \mathsf{Y}$ and use the 1-Lipschitz encoders

$$q_{\mathsf{G}} = \mathrm{id}_{\mathsf{G}}, \qquad q_{\mathsf{X}}(x) = \big((x, y_*), 0\big).$$

The context layer retains each original token and appends its calibrated spatial values:

$$\Gamma_0(\mu, g) = \left(g, \left(\frac{\phi_k(g) - a_k(\mu)}{\sqrt{D(\mu)}}\right)_{k=1}^{n}\right).$$

This is a Markov kernel in-context map with one head: use the uniform kernel $\widehat{K}[\mu, g] = \mu$ and the value map

$$\widehat{v}(w) = \big(\phi_k(w), \phi_k(w)^2\big)_{k=1}^{n}.$$

Its read supplies all first and second moments needed to compute a_k and D, so the mixer can apply the displayed formula using only the token and this read. To define the mixer on arbitrary read vectors, first clip every moment coordinate to $[0,1]$. Actual reads are unchanged, and the resulting feature coordinates always lie in $[-R, R]$.

The uniform kernel $\widehat{K}$ is jointly 1-Lipschitz, and $\widehat{v}$ is Lipschitz because the ϕ_k are bounded and Lipschitz. The mixer is globally Lipschitz as well: after clipping, its moment coordinates lie in $[0,1]$, squaring is Lipschitz on this interval, and $t \mapsto \max\{t, v_*\}^{-1/2}$ is bounded and Lipschitz. Its feature numerators are bounded and Lipschitz in the token and clipped moments. Thus the hidden layer satisfies the component assumptions of Definition 3.4 on its full domain.

Write $\pi : \mathsf{E}_1 \to \mathsf{G}$ for projection onto the retained token and $\psi_k : \mathsf{E}_1 \to [-R, R]$ for the kth new coordinate. For any $\eta \in \mathcal{P}(\mathsf{E}_1)$, define

$$\bar{a}_k(\eta) = \int_{\mathsf{E}_1} \psi_k \,\mathrm{d}\eta, \qquad T(\eta) = \sum_{k=1}^{n} \left(\int_{\mathsf{E}_1} \psi_k^2 \,\mathrm{d}\eta - \bar{a}_k(\eta)^2\right).$$

For the updated context $\eta = B_0(\mu) = \Gamma_0(\mu, \cdot)_{\#}\mu$, direct integration gives

$$\pi_{\#}\eta = \mu, \qquad \bar{a}_k(\eta) = 0, \qquad T(\eta) = \frac{V(\mu)}{D(\mu)}.$$

Thus the new coordinates have zero mean for every input, and their total variance is one whenever $\mu \in \mathcal{S}$. This completes the calibration.

Step 4: Decode with weighted averages and a fixed readout. For arbitrary $\eta \in \mathcal{P}(\mathsf{E}_1)$ and $z \in \mathsf{E}_1$, set

$$\theta_j(\eta, z) = \frac{U_j(\pi_{\#}\eta, \pi_{\mathsf{X}}(\pi z))}{4MR}, \qquad j = 1, \dots, c.$$

Use cn decoder heads, indexed by (j, k), with

$$K_{jk}[\eta, z](\mathrm{d}w) = \big[1 + \theta_j(\eta, z)(\psi_k(w) - \bar{a}_k(\eta))\big]\, \eta(\mathrm{d}w), \qquad v_{jk}(w) = \psi_k(w).$$

Since $|\theta_j| \le 1/(4R)$ and $|\psi_k - \bar{a}_k| \le 2R$, the densities lie in $[1/2, 3/2]$. Subtracting $\bar{a}_k(\eta)$ makes them integrate to one even for contexts outside the image of B_0. These are Borel probability kernels supported on $\operatorname{supp}\eta$, and the value maps are bounded and Lipschitz. Their reads satisfy

$$r_{jk}(\eta, z) = \bar{a}_k(\eta) + \theta_j(\eta, z) \left(\int \psi_k^2 \, \mathrm{d}\eta - \bar{a}_k(\eta)^2 \right).$$

Let $\Pi_{\mathsf{Y}} : \mathbb{R}^c \to \mathsf{Y}$ be nearest-point projection, which is 1-Lipschitz because Y is non-empty, compact, and convex. The decoder mixer is simply

$$m_p\big(z, (r_{jk})_{j,k}\big) = \Pi_{\mathsf{Y}} \left(4MR \left(\sum_{k=1}^{n} r_{jk} \right)_{j=1}^{c} \right).$$

This mixer is globally Lipschitz, being a fixed linear map followed by the 1-Lipschitz projection. There is no reference head or variance normalization in this mixer. Indeed, writing $A(\eta) = \sum_k \bar{a}_k(\eta)$ and $\mathbf{1}_c = (1, \dots, 1) \in \mathbb{R}^c$, the read identities give

$$\Gamma_p(\eta, z) = \Pi_{\mathsf{Y}}(4MR\, A(\eta)\mathbf{1}_c + T(\eta)\, U(\pi_{\#}\eta, \pi_{\mathsf{X}}(\pi z))).$$

It remains to verify context-only stability of the decoder kernels themselves. Fix j, k and z, and write

$$F_{\eta,z}(w) = 1 + \theta_j(\eta, z)(\psi_k(w) - \bar{a}_k(\eta)), \qquad d = W_1(\eta, \xi), \qquad \Delta = \operatorname{diam}(\mathsf{E}_1).$$

Since π and ψ_k are 1-Lipschitz, Step 1 and the averaging estimate in the proof of Theorem 3.5 give

$$|\theta_j(\eta, z) - \theta_j(\xi, z)| \le \frac{C}{4MR} d^{\beta}, \qquad |\bar{a}_k(\eta) - \bar{a}_k(\xi)| \le d.$$

Together with $|\theta_j| \le 1/(4R)$ and $|\psi_k - \bar{a}_k| \le 2R$, these imply

$$\|F_{\eta,z}\|_\infty \le \frac{3}{2}, \qquad \operatorname{Lip}_w(F_{\eta,z}) \le \frac{1}{4R}, \qquad \|F_{\eta,z} - F_{\xi,z}\|_\infty \le \frac{C}{2M} d^{\beta} + \frac{d}{4R}.$$

For a 1-Lipschitz test function f normalized to vanish at a fixed point of E_1, we have $\|f\|_\infty \le \Delta$ and $\operatorname{Lip}(fF_{\eta,z}) \le 3/2 + \Delta/(4R)$. Therefore

$$\begin{aligned} \left| \int f \, \mathrm{d}K_{jk}[\eta, z] - \int f \, \mathrm{d}K_{jk}[\xi, z] \right| &\le \left| \int f F_{\eta,z} \, \mathrm{d}(\eta - \xi) \right| + \left| \int f (F_{\eta,z} - F_{\xi,z}) \, \mathrm{d}\xi \right| \\ &\le \left(\frac{3}{2} + \frac{\Delta}{2R} \right) d + \frac{\Delta C}{2M} d^{\beta}. \end{aligned}$$

Kantorovich–Rubinstein duality, followed by $d \le \Delta$ and $\beta \le 1$, now gives

$$W_1\big(K_{jk}[\eta, z], K_{jk}[\xi, z]\big) \le C_K W_1(\eta, \xi)^{\beta}$$

for a constant C_K independent of z. This holds for all $\eta, \xi \in \mathcal{P}(\mathsf{E}_1)$, including contexts outside the image of B_0. All components therefore satisfy Definition 3.4, and Theorem 3.5 gives uniform discretization stability of the assembled network on its full domain.

Finally, on every context $\eta = B_0(\mu)$ the term $A(\eta)$ vanishes, so

$$\mathcal{N}(\mu, x) = \Gamma_p\big(B_0(\mu), q_{\mathsf{X}}(x)\big) = \Pi_{\mathsf{Y}} \left(\frac{V(\mu)}{D(\mu)}\, U(\mu, x) \right).$$

For $\mu \in \mathcal{S}$, the ratio V/D is one and $U(\mu, x) = \mathcal{N}^*(\mu, x) \in \mathsf{Y}$, proving the required equality. Only the decoder kernels use the query dependence of U; no spatial regularity of the target is needed. □

Corollary 3.8 (Operator characterization of stable Markov kernel networks)**.** *Let $A : L^2(\mathsf{X}, \mathsf{Y}) \to L^2(\mathsf{X}, \mathsf{Y})$ be a possibly non-linear operator mapping functions to functions. Then there exists $C > 0$, $\beta \in (0, 1]$ such that*

$$\|A[f] - A[g]\|_{L^\infty(\mathsf{X},\mathsf{Y})} \leq C W_1(\gamma_f, \gamma_g)^\beta \qquad \text{for all } f, g \in L^2(\mathsf{X}, \mathsf{Y}) \tag{16}$$

if and only if there exists a stable Markov kernel network $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$ such that

$$A[f](x) = \mathcal{N}(\gamma_f, x) \qquad \text{for all } f \in L^2(\mathsf{X}, \mathsf{Y}) \text{ and } \lambda\text{-a.e. } x \in \mathsf{X}. \tag{17}$$

Proof. Suppose first that the displayed Hölder estimate holds, and set

$$\mathcal{S} = \{\gamma_f : f \in L^2(\mathsf{X}, \mathsf{Y})\}.$$

This set is non-empty because Y is non-empty. Every $\mu \in \mathcal{S}$ has spatial marginal $(\pi_{\mathsf{X}})_{\#}\mu = \lambda$. Since λ has full support and X contains at least two points, λ is not a Dirac measure. The function

$$a \longmapsto W_1(\lambda, \delta_a) = \int_{\mathsf{X}} \rho_{\mathsf{X}}(x, a)\, \lambda(\mathrm{d}x)$$

is continuous and strictly positive on the compact space X. Its minimum c_λ is therefore positive, and

$$\inf_{\mu \in \mathcal{S}} \inf_{a \in \mathsf{X}} W_1\big((\pi_{\mathsf{X}})_{\#}\mu, \delta_a\big) = c_\lambda > 0.$$

Taking $\epsilon = c_\lambda/2$ shows that $\mathcal{S}$ is an admissible set of non-collapsing measures.

Since Y is compact, every output of A belongs to $L^\infty(\mathsf{X}, \mathbb{R}^{d_{\text{out}}})$. Define

$$Q : \mathcal{S} \to L^\infty(\mathsf{X}, \mathbb{R}^{d_{\text{out}}}), \qquad Q(\gamma_f) = A[f].$$

This is well-defined because equality of graph measures implies equality of the input L^2 classes. The assumed estimate states that

$$\|Q(\mu) - Q(\nu)\|_{L^\infty} \leq C\, W_1(\mu, \nu)^\beta \qquad (\mu, \nu \in \mathcal{S}).$$

We next choose representatives of these output classes so that this estimate holds pointwise and the resulting evaluation map is jointly Borel measurable.

As a subspace of the compact metric space $(\mathcal{P}(\mathsf{G}), W_1)$, the set $\mathcal{S}$ is separable. Choose a dense sequence $(\mu_i)_{i \geq 1}$ in $\mathcal{S}$ and Borel representatives $u_i : \mathsf{X} \to \mathsf{Y}$ of $Q(\mu_i)$. The set

$$E = \bigcup_{i,j \geq 1} \big\{x \in \mathsf{X} : |u_i(x) - u_j(x)| > C\, W_1(\mu_i, \mu_j)^\beta\big\}$$

is Borel and λ-null. Fix $y_* \in \mathsf{Y}$ and redefine every u_i to equal y_* on E. These remain Borel representatives of the same classes and now satisfy

$$|u_i(x) - u_j(x)| \leq C\, W_1(\mu_i, \mu_j)^\beta \qquad \text{for all } i, j \geq 1,\ x \in \mathsf{X}.$$

For $k \geq 1$ and $\mu \in \mathcal{S}$, let

$$i_k(\mu) = \min\{i \geq 1 : W_1(\mu, \mu_i) < 2^{-k}\}.$$

Density makes this index well-defined, and it is Borel measurable because each set on which it equals a given index is a relative open ball with finitely many earlier relative open balls removed. Moreover,

$$\sup_{x \in \mathsf{X}} |u_{i_k(\mu)}(x) - u_{i_\ell(\mu)}(x)| \leq C(2^{-k} + 2^{-\ell})^\beta.$$

Thus the limit

$$\mathcal{N}^*(\mu, x) = \lim_{k \to \infty} u_{i_k(\mu)}(x)$$

exists uniformly in x and takes values in the closed set Y. Each approximating map is jointly Borel measurable, so $\mathcal{N}^*$ is jointly Borel measurable as well. Passing to the limit in the pairwise pointwise bounds gives

$$|\mathcal{N}^*(\mu, x) - \mathcal{N}^*(\nu, x)| \leq C\, W_1(\mu, \nu)^\beta \qquad (\mu, \nu \in \mathcal{S},\ x \in \mathsf{X}).$$

Finally, $Q(\mu_{i_k(\mu)}) \to Q(\mu)$ in L^∞, while their chosen representatives converge uniformly to $\mathcal{N}^*(\mu, \cdot)$. Hence $\mathcal{N}^*(\mu, \cdot)$ represents $Q(\mu)$.

The map $\mathcal{N}^*$ is therefore uniformly discretization stable on $\mathcal{S}$. Theorem 3.6 provides a stable Markov kernel network $\mathcal{N}$ on $\mathcal{P}(\mathsf{G}) \times \mathsf{X}$ extending $\mathcal{N}^*$. For every $f \in L^2(\mathsf{X}, \mathsf{Y})$, we consequently have $\mathcal{N}(\gamma_f, x) = A[f](x)$ for λ-almost every x.

Conversely, suppose such a stable Markov kernel network $\mathcal{N}$ exists. By Theorem 3.5, there are constants $C > 0$ and $\beta \in (0, 1]$ such that

$$\begin{aligned} \|A[f] - A[g]\|_{L^\infty} &= \|\mathcal{N}(\gamma_f, \cdot) - \mathcal{N}(\gamma_g, \cdot)\|_{L^\infty} \\ &\le \sup_{x \in \mathsf{X}} |\mathcal{N}(\gamma_f, x) - \mathcal{N}(\gamma_g, x)| \\ &\le C\, W_1(\gamma_f, \gamma_g)^\beta \end{aligned}$$

for all $f, g \in L^2(\mathsf{X}, \mathsf{Y})$, as required. □

B.2 Section 4.1: stable variational reads

Regularized attention is a familiar instance of (18) (Niculae & Blondel, 2017; Martins et al., 2022). A broader family uses f-divergences. We denote the convex generator by φ to distinguish it from an input field. For a superlinear generator, its divergence and the corresponding regularizer are

$$D_\varphi(\kappa \| \mu) = \begin{cases} \displaystyle\int_{\mathsf{E}} \varphi(p)\, \mathrm{d}\mu, & \kappa = p\mu, \\ +\infty, & \kappa \not\ll \mu, \end{cases} \qquad \mathcal{R}_{\varphi,\tau} = \tau D_\varphi, \quad \tau > 0. \tag{27}$$

The integral has its extended meaning. The optimizer is obtained by applying the inverse derivative of φ to a shifted negative cost (Terjék, 2021). Quantitative regularity of this scalar response gives a stability certificate even when the reference context changes.

Proposition B.1 (Stable divergence-regularized kernels)**.** *Let E be compact and $\tau > 0$. Let $\varphi : [0, \infty) \to [0, \infty]$ be lower semicontinuous, strictly convex, finite and C^1 on $(0, \infty)$, with $\varphi(1) = \varphi'(1) = 0$ and $\lim_{r \to \infty} \varphi'(r) = +\infty$. Define its scalar response by*

$$h(u) = \begin{cases} 0, & u \le \varphi'(0+), \\ (\varphi')^{-1}(u), & u > \varphi'(0+), \end{cases} \tag{28}$$

where the first case is absent if $\varphi'(0+) = -\infty$. Suppose the Borel cost c satisfies, for some $M, L, H \ge 0$ and $\beta \in (0, 1]$,

$$\begin{aligned} |c(\mu, z, w)| &\le M, \\ |c(\mu, z, w) - c(\nu, z', w')| &\le L\rho_{\mathsf{E}}(w, w') + H\big(W_1(\mu, \nu) + \rho_{\mathsf{E}}(z, z')\big)^\beta. \end{aligned} \tag{29}$$

Assume that h is γ-Hölder on $I_M = [-2M/\tau, 2M/\tau]$, for some $\gamma \in (0, 1]$. Then (18) with $\mathcal{R} = \mathcal{R}_{\varphi,\tau}$ has a unique minimizer, defining a supported Borel Markov kernel

$$K[\mu, z](\mathrm{d}w) = h\big(b_{\mu,z} - c(\mu, z, w)/\tau\big)\, \mu(\mathrm{d}w), \tag{30}$$

where $b_{\mu,z} \in [-M/\tau, M/\tau]$ is the unique scalar making the density integrate to one. This kernel satisfies

$$W_1\big(K[\mu, z], K[\nu, z']\big) \le C\big(W_1(\mu, \nu) + \rho_{\mathsf{E}}(z, z')\big)^{\beta\gamma}. \tag{31}$$

Here C depends only on the displayed bounds, $\operatorname{diam} \mathsf{E}$, $h(2M/\tau)$, and the Hölder constant of h on I_M. If (29) is assumed only for $z' = z$, uniformly in z, then the conclusion holds for $z' = z$ with the same uniformity.

The normalization $\varphi'(1) = 0$ is harmless: subtracting a multiple of $p - 1$ does not change the divergence between probability measures. The response condition is a quantitative convexity requirement. For example, a lower bound $\varphi''(p) \ge m_B > 0$ for $0 < p \le B$, for every finite B, suffices for $\gamma = 1$, without requiring global strong convexity.

Proposition 4.1 is the case $\tau = 1$, $M = C$, and $H = L$ of this result. If the scalar response has exponent γ, the resulting kernel has exponent $\beta\gamma$.

Proof. Write $s = (\mu, z)$, $t = (\nu, z')$, $c_s(w) = c(\mu, z, w)$, and

$$D = \operatorname{diam} \mathsf{E}, \qquad a = W_1(\mu, \nu), \qquad \delta = a + \rho_{\mathsf{E}}(z, z'), \qquad \varepsilon = La + H\delta^\beta.$$

Let A be a γ-Hölder constant of h on I_M, and set $B = h(2M/\tau)$.

Normalization and optimality. The assumptions on φ imply that h is continuous and nondecreasing, strictly increasing wherever positive, with $h(0) = 1$, $\lim_{u\to-\infty} h(u) = 0$, and $\lim_{u\to\infty} h(u) = \infty$. Thus

$$F_s(b) := \int h(b - c_s/\tau)\,\mathrm{d}\mu$$

is continuous, goes from zero to infinity, and is strictly increasing wherever positive. It has a unique root $F_s(b_s) = 1$. Since $|c_s| \le M$ and $h(0) = 1$,

$$-M/\tau \le b_s \le M/\tau, \qquad 0 \le p_s(w) := h(b_s - c_s(w)/\tau) \le B.$$

The candidate $p_s\mu$ has finite objective. If $\varphi(0) < \infty$, the generator is bounded on $[0, B]$; if $\varphi(0) = \infty$, then $\varphi'(0+) = -\infty$ and $p_s \ge h(-2M/\tau) > 0$, so it is evaluated on a compact subinterval of $(0, \infty)$.

Extend φ by $+\infty$ to negative arguments. By the definition of h, the scalar $b_s - c_s(w)/\tau$ belongs to $\partial\varphi(p_s(w))$, including when $p_s(w) = 0$. For any probability density q of finite divergence, the supporting-line inequality gives

$$\varphi(q) - \varphi(p_s) \ge (b_s - c_s/\tau)(q - p_s).$$

Integrating, using $\int (q - p_s)\,\mathrm{d}\mu = 0$, and adding the linear cost proves $J_s(q\mu) \ge J_s(p_s\mu)$. Strict convexity forces $q = p_s$ almost everywhere in the equality case. All other measures have infinite objective, since c_s is bounded. This proves existence and uniqueness, and absolute continuity gives the support constraint. For fixed b, $s \mapsto F_s(b)$ is Borel, and

$$\{s : b_s < b\} = \{s : F_s(b) > 1\}.$$

Thus the normalizer is Borel, and parameter-dependent integration proves the Markov property.

Stability without estimating the normalizer. Let π be an optimal coupling of μ, ν. The cost assumptions imply

$$\int |c_s(w) - c_t(v)|\,\pi(\mathrm{d}w, \mathrm{d}v) \le \varepsilon.$$

On this coupling define $q_0(v) = h(b_s - c_t(v)/\tau)$. Since h is nondecreasing, $q_0 - p_t$ has the same sign everywhere. Both p_s and p_t integrate to one against π, so

$$\int |q_0 - p_t|\,\mathrm{d}\pi = \left| \int q_0\,\mathrm{d}\pi - 1 \right| \le \int |q_0 - p_s|\,\mathrm{d}\pi.$$

Every argument of h lies in I_M. The triangle inequality, the response bound, and concavity of $r \mapsto r^\gamma$ therefore give

$$\begin{aligned} \int |p_s(w) - p_t(v)|\,\mathrm{d}\pi &\le 2A\tau^{-\gamma} \int |c_s(w) - c_t(v)|^\gamma\,\mathrm{d}\pi \\ &\le 2A(\varepsilon/\tau)^\gamma. \end{aligned} \tag{32}$$

For any 1-Lipschitz test function v, subtract the midpoint of its range so that $|v| \le D/2$. This does not change its integral against a difference of probability measures. Since $p_s \le B$,

$$\left| \int v\,p_s\,\mathrm{d}\mu - \int v\,p_t\,\mathrm{d}\nu \right| \le Ba + \frac{D}{2} \int |p_s(w) - p_t(v')|\,\pi(\mathrm{d}w, \mathrm{d}v').$$

Kantorovich–Rubinstein duality yields the quantitative bound

$$W_1(K_s, K_t) \le Ba + DA\tau^{-\gamma}(La + H\delta^\beta)^\gamma. \tag{33}$$

If $D = 0$, all the measures coincide. Otherwise $a \le \delta \le 2D$, so boundedness converts the right-hand side into $C\delta^{\beta\gamma}$. Taking $z' = z$ throughout proves the fixed-query version under only its stated cost hypotheses. □

B.2.1 KL, PEARSON, AND TSALLIS EXAMPLES

Corollary B.2 (KL and Pearson reads)**.** *Under the cost assumptions of Proposition B.1, the regularizers*

$$\mathcal{R}_{\mathrm{KL},\tau}(\kappa,\mu)=\tau\,\mathrm{KL}(\kappa\|\mu),\qquad \mathcal{R}_{2,\tau}(\kappa,\mu)=\frac{\tau}{2}\chi^2(\kappa\|\mu) \tag{34}$$

yield unique kernels with densities

$$p^{\mathrm{KL}}_{\mu,z}(w)=\frac{e^{-c(\mu,z,w)/\tau}}{\int_{\mathsf{E}} e^{-c(\mu,z,v)/\tau}\,\mu(\mathrm{d}v)},\qquad p^{2}_{\mu,z}(w)=\big(b^2_{\mu,z}-c(\mu,z,w)/\tau\big)_+, \tag{35}$$

where $b^2_{\mu,z}$ normalizes the second density. Both satisfy (31) *with exponent β.*

Indeed, $\varphi_{\mathrm{KL}}(p)=p\log p-p+1$ gives $h(u)=e^u$, while $\varphi_2(p)=(p-1)^2/2$ gives $h(u)=(1+u)_+$. Both responses are Lipschitz on bounded intervals. These are the Gibbs and positive-part transformations underlying softmax and sparsemax (Martins & Astudillo, 2016; Martins et al., 2022). Here $\chi^2(\kappa\|\mu)=\int(p-1)^2\,\mathrm{d}\mu$ for $\kappa=p\mu$, and is infinite otherwise; the factor $1/2$ belongs to the regularizer.

The generators in Corollary B.2 have the claimed scalar responses, both locally Lipschitz. For KL, normalization gives $e^{b_s}=1/\int e^{-c_s/\tau}\,\mathrm{d}\mu$. For Pearson, set $b_s^2=1+b_s$ to obtain the stated positive-part density. In the notation of the proof, $B_{\mathrm{KL}}=e^{2M/\tau}$ and $A_{\mathrm{KL}}=B_{\mathrm{KL}}$, while $B_2=1+2M/\tau$ and $A_2=1$. Equation (33) therefore gives

$$W_1(K_s^{\mathrm{KL}},K_t^{\mathrm{KL}})\le B_{\mathrm{KL}}a+\frac{DB_{\mathrm{KL}}}{\tau}\varepsilon, \tag{36}$$

$$W_1(K_s^{2},K_t^{2})\le B_2 a+\frac{D}{\tau}\varepsilon. \tag{37}$$

No lower bound on the density or on positive empirical atom masses is used.

More generally, the power, or Tsallis, generators and their responses are

$$\varphi_\alpha(p)=\frac{p^\alpha-\alpha p+\alpha-1}{\alpha(\alpha-1)},\qquad h_\alpha(u)=\big[1+(\alpha-1)u\big]_+^{1/(\alpha-1)},\quad \alpha>1. \tag{38}$$

The density formula is the continuous entmax transformation (Martins et al., 2022, Proposition 8). For $1<\alpha\le 2$ the response is locally Lipschitz, so the kernel retains exponent β. For $\alpha>2$, it is $1/(\alpha-1)$-Hölder, giving exponent $\beta/(\alpha-1)$. These are sufficient exponents, not claims of optimality. The limit $\alpha\to 1$ gives KL, and $\alpha=2$ gives our Pearson normalization.

A direct sufficient curvature condition for $\gamma=1$ is $\varphi''(p)\ge m_B>0$ for $0<p\le B$. The mean value theorem then bounds the inverse derivative by $1/m_B$; the same bound holds across a finite zero threshold by taking a limit from positive densities. If $h(I_M)$ is bounded away from zero, the curvature bound is needed only on that smaller density interval. In particular, global strong convexity is unnecessary for KL.

The remaining responses in Table 1. Differentiating the shifted-entropy and KL-plus-Pearson generators gives, respectively, $\log((r+a)/(1+a))$ and $\log r+a(r-1)$. Inverting these derivatives, with truncation at zero when necessary, gives the displayed responses. The centered-power derivative is $\mathrm{sgn}(r-1)|r-1|^{p-1}$, so its response is locally Hölder with exponent $\min\{1,1/(p-1)\}$. The derivatives of the quadratic-plus-reverse-KL and Jeffreys generators are $(r-r^{-1})/2$ and $\log r+1-r^{-1}$, yielding the stated formulas. These last two generators have value $+\infty$ at zero, as allowed in the proposition. Apart from the two power families, all responses in the table are Lipschitz on bounded intervals: their generators have strictly positive continuous second derivative on the positive density range, and the shifted-entropy response is also Lipschitz across its zero threshold. All listed generators are normalized, strictly convex, and superlinear.

Why arbitrary divergences are not covered. Even smooth strict convexity does not prevent a support discontinuity when the generator has finite asymptotic slope. Consider reverse KL, with $\varphi(p)=-\log p+p-1$, on $\mathsf{E}=\{0,1\}$, and take

$$\mu_e=(1-e)\delta_0+e\delta_1,\qquad c(w)=-Cw,\qquad C>\tau.$$

For $0 < e < 1$, a row with mass x at 1 has objective $-Cx + \tau\,\mathrm{KL}(\mu_e\|\kappa)$. Its unique minimizer satisfies

$$Cx(1-x) = \tau(x-e), \qquad x = \frac{C - \tau + \sqrt{(C-\tau)^2 + 4C\tau e}}{2C} \longrightarrow 1 - \tau/C > 0.$$

At $e = 0$, the support constraint forces the row δ_0. Thus no context-continuous selection exists for this bounded Lipschitz cost. Here $\varphi'(p) \to 1$, violating the proposition's superlinearity assumption.

Nor does superlinearity and strict convexity alone ensure a Hölder rate. For example, choose a strictly increasing continuous derivative with

$$\varphi'(1+r) = \mathrm{sgn}(r)e^{-1/|r|} \quad (0 < |r| \le 1/2), \qquad \varphi'(1) = 0,$$

and extend it strictly increasingly to $(0,\infty)$ with limit $+\infty$ at infinity; integrating from 1, with the lower-semicontinuous extension at zero, gives a normalized, strictly convex, superlinear generator. On $\mathsf{E} = [0,1]$, fix $\mu = (\delta_0 + \delta_1)/2$ and take the bounded Lipschitz cost $c(z,w) = \tau\min\{z, e^{-3}\}(2w-1)$. For $0 < z \le e^{-3}$ the unique optimal densities at $0, 1$ are $1+r(z), 1-r(z)$, where $r(z) = 1/\log(1/z)$. Hence $W_1(K[\mu,z], K[\mu,0]) = r(z)/2$, which obeys no positive Hölder exponent at zero. This explains the quantitative response assumption.

Empirical normalization. For $\mu = \sum_i m_i\delta_{w_i}$ with $m_i > 0$, write $k_i = m_i p_i$ and $c_i = c(\mu, z, w_i)$. Pearson minimization is exactly

$$\min_{k\in\Delta_n}\left\{\sum_i c_i k_i + \frac{\tau}{2}\sum_i \frac{(k_i - m_i)^2}{m_i}\right\}.$$

For $m_i = 1/n$, the quadratic term differs from $(\tau n/2)\sum_i k_i^2$ by a constant on Δ_n. Completing the square gives the projection of $-c/(\tau n)$. Both quantitative estimates above are independent of the smallest positive m_i.

B.2.2 The fixed Pearson regularizer in the representation theorem

Use the notation of the proof of Theorem 3.6. The calibration head $\widehat{K}[\mu, g] = \mu$ is the unique optimizer for $c = 0$ and $\mathcal{R} = \frac{1}{2}\chi^2$. For the decoder, ψ_k is 1-Lipschitz, $|\psi_k| \le R$, and

$$\theta_j(\eta, z) = \frac{U_j(\pi_{\#}\eta, \pi_{\mathsf{X}}(\pi z))}{4MR}, \qquad |\theta_j| \le \frac{1}{4R}.$$

The decoder costs and corresponding densities are

$$c_{jk}(\eta, z, w) = -\theta_j(\eta, z)\psi_k(w), \tag{39}$$

$$p^*_{jk}(\eta, z, w) = 1 + \theta_j(\eta, z)\left(\psi_k(w) - \int \psi_k\,\mathrm{d}\eta\right) \in [1/2, 3/2]. \tag{40}$$

Here $|U_j| \le M$ and U_j is β-Hölder in its context, uniformly at every fixed spatial query. Since π is 1-Lipschitz, the costs in (39) obey

$$|c_{jk}| \le \tfrac{1}{4}, \qquad \mathrm{Lip}_w(c_{jk}) \le \frac{1}{4R}, \qquad \sup_w |c_{jk}(\eta, z, w) - c_{jk}(\xi, z, w)| \le \frac{C}{4M} W_1(\eta, \xi)^\beta.$$

They are Borel and satisfy the fixed-query hypotheses globally. The density p^*_{jk} in (40) integrates to one and lies in $[1/2, 3/2]$. Taking $b^2 = 1 - \theta_j\int\psi_k\,\mathrm{d}\eta$ in (35) gives precisely this density. Alternatively, for every probability density $q \in L^2(\eta)$,

$$J_{\eta,z}(q\eta) - J_{\eta,z}(p^*_{jk}\eta) = \tfrac{1}{2}\int (q - p^*_{jk})^2\,\mathrm{d}\eta,$$

because $c_{jk} + p^*_{jk} - 1 = -\theta_j\int\psi_k\,\mathrm{d}\eta$ is constant in the read token. All other measures have infinite objective. The variational heads therefore agree with the original construction on every context, so the exact representation and its global stability are unchanged. This concerns the representing network constructed in that theorem, not a variational realization of every prescribed supported kernel.

B.3 Projected Wasserstein regularization and SelfWarp reads

Proof of Proposition 4.2. Write $F(\mu) = \mathcal{P}(\operatorname{supp}\mu)$. Since E is compact, this is a nonempty compact subset of $\mathcal{P}(\mathsf{E})$. For each fixed (μ, z), the finite-valued cost is continuous and bounded in w, so its integral is continuous in κ. Continuous pushforward and continuity of W_2 on probabilities over compact X show that $J_{\mu,z}$ is continuous in κ. It therefore attains its minimum on $F(\mu)$.

To obtain a Borel choice of minimizers, first observe that F is lower hemicontinuous. Given $\kappa \in F(\mu)$ and an open neighborhood U of κ, choose a finitely supported probability $\sum_i a_i \delta_{w_i} \in U$ with $w_i \in \operatorname{supp}\mu$. There are open neighborhoods O_i of w_i such that $\sum_i a_i \delta_{w_i'} \in U$ whenever $w_i' \in O_i$. Each $\mu(O_i) > 0$. By the Portmanteau theorem, every sufficiently nearby μ' has $\mu'(O_i) > 0$, and hence a support point $w_i' \in O_i$. Thus $F(\mu') \cap U \neq \emptyset$. In particular, F is a measurable compact-valued correspondence.

By the standing measurability convention, c is jointly Borel. Parameter-dependent integration therefore makes $(\mu, z, \kappa) \mapsto J_{\mu,z}(\kappa)$ Borel; it is continuous in κ for fixed (μ, z). The measurable minimum theorem gives a measurable nonempty compact-valued argmin correspondence. The measurable selection theorem (Himmelberg & Van Vleck, 1969, Theorem 1) supplies a Borel selector $K[\mu, z]$. Since $K[\mu, z] \in F(\mu)$, this is a supported Markov kernel satisfying (18). No uniqueness or continuity of the selection is asserted. □

Proposition B.3 (Exact recovery of SelfWarp reads)**.** *Let $(\mathsf{X}, \rho_{\mathsf{X}})$ be a compact metric space, let $\mathsf{A} \subset \mathbb{R}^m$ be compact, and set $\mathsf{E} = \mathsf{X} \times \mathsf{A}$ with spatial projection π_{X}. Let $\lambda \in \mathcal{P}(\mathsf{X})$ have full support. Fix a Borel address map $\tau_h : \mathsf{E} \to \mathsf{X}$ and a constant $B > D := \operatorname{diam}(\mathsf{X})$. Use the regularizer* (23) *and the cost*

$$c_h(\mu, z, w) = B\rho_{\mathsf{X}}\big(\pi_{\mathsf{X}}(w), \tau_h(z)\big). \tag{41}$$

For every continuous $f : \mathsf{X} \to \mathsf{A}$, write $j_f(y) = (y, f(y))$ and $\gamma_f = (j_f)_{\#}\lambda$. Then, for every $z \in \mathsf{E}$, the unique minimizer in (18) *is*

$$K_h[\gamma_f, z] = \delta_{j_f(\tau_h(z))} = \delta_{(\tau_h(z), f(\tau_h(z)))}. \tag{42}$$

In particular, on a compact Euclidean domain or torus, take $\tau_h(x, a) = x + g_h(a)$, assuming the address lies in X or using periodic wrapping. With the value map $v_h(y, b) = V_h b$, the resulting read is

$$r_h(\gamma_f, (x, f(x))) = V_h f\big(x + g_h(f(x))\big), \tag{43}$$

which is the continuum SelfWarp head of Muser et al. (2026, Appendix C.2).

Proof. Fix f, z, and write $a = \tau_h(z)$. Continuity of f and full support of λ give $\operatorname{supp}\gamma_f = j_f(\mathsf{X})$, so $\kappa_a := \delta_{j_f(a)}$ is feasible. For any feasible κ, set $\nu = (\pi_{\mathsf{X}})_{\#}\kappa$, and let q be an optimal coupling of ν and λ for W_2. Since both distances below are bounded by D, the reverse triangle inequality gives

$$\frac{1}{2}\big(\rho_{\mathsf{X}}(y, t)^2 - \rho_{\mathsf{X}}(a, t)^2\big) \geq -D\rho_{\mathsf{X}}(y, a).$$

Integrating against $q(dy, dt)$ yields

$$\frac{1}{2}W_2(\nu, \lambda)^2 - \frac{1}{2}W_2(\delta_a, \lambda)^2 \geq -D \int_{\mathsf{X}} \rho_{\mathsf{X}}(y, a)\, \nu(dy).$$

Hence the objective corresponding to (41) satisfies

$$J_{\gamma_f,z}(\kappa) - J_{\gamma_f,z}(\kappa_a) \geq (B - D) \int_{\mathsf{X}} \rho_{\mathsf{X}}(y, a)\, \nu(dy). \tag{44}$$

As $B > D$, every minimizer has spatial marginal δ_a. The graph $j_f(\mathsf{X})$ contains exactly one token above a, so its only supported probability measure with that marginal is κ_a. This proves uniqueness and (42); integrating the value map proves (43). The cost is Borel and continuous in w, so every minimizing Markov kernel supplied by Proposition 4.2 has the stated restriction to continuous graph contexts. □

Scope of the recovery. For a general context μ, if $a = \tau_h(z) \in \pi_{\mathsf{X}}(\operatorname{supp}\mu)$, the same argument with $(\pi_{\mathsf{X}})_{\#}\mu$ in place of λ shows that the full set of minimizers is

$$\mathcal{P}\big(\operatorname{supp}\mu \cap \pi_{\mathsf{X}}^{-1}(\{a\})\big).$$

Thus the spatial cost alone does not select between multiple observed feature values at the same address. Proposition B.3 recovers the continuum graph-context reads; the finite-grid interpolation rows and boundary conventions are described separately in Section D.5. No stability assertion for arbitrary noisy contexts follows from this exact-recovery statement.

C Universal approximation by uniformly discretization stable networks

Uniform discretization stability does not prevent approximation of continuous operators on compact sets of input functions.

Proposition C.1. *Let $\mathcal{D} \subset L^2(\mathsf{X},\mathsf{Y})$ be compact. For every continuous $A : \mathcal{D} \to L^2(\mathsf{X},\mathsf{Y})$ and $\epsilon > 0$, there exists $\mathcal{N} : \mathcal{P}(\mathsf{G}) \times \mathsf{X} \to \mathsf{Y}$, uniformly discretization stable on $\mathcal{P}(\mathsf{G})$, such that*

$$\sup_{f\in\mathcal{D}} \|\mathcal{N}(\gamma_f,\cdot) - A[f]\|_{L^2(\mathsf{X},\mathbb{R}^{d_{out}})} < \epsilon. \tag{45}$$

Proof. We approximate A by weighted averages of finitely many of its outputs, with weights that depend Lipschitz continuously on the input graph measure. If $\mathcal{D}$ is empty, any constant map suffices, so assume $\mathcal{D} \neq \emptyset$.

First, the graph embedding is continuous because coupling the graph measures at the same spatial point gives

$$W_1(\gamma_f,\gamma_g) \le \int_{\mathsf{X}} |f(x) - g(x)|\,\lambda(\mathrm{d}x) \le \|f-g\|_{L^2}.$$

It is also injective. Since $\mathcal{D}$ is compact, this embedding is therefore a homeomorphism onto the compact set

$$\mathcal{K} = \{\gamma_f : f \in \mathcal{D}\} \subset \mathcal{P}(\mathsf{G}).$$

Consequently, the induced map $\widetilde{A} : \mathcal{K} \to L^2(\mathsf{X},\mathsf{Y})$, defined by $\widetilde{A}(\gamma_f) = A[f]$, is uniformly continuous. Choose $\delta > 0$ such that

$$W_1(\gamma_f,\gamma_g) < \delta \quad\Longrightarrow\quad \|A[f] - A[g]\|_{L^2} < \epsilon/2 \qquad \text{for } f,g \in \mathcal{D}.$$

Choose a finite $\delta/2$-net $\mu_i = \gamma_{f_i}$, $i = 1,\dots,m$, in $\mathcal{K}$. For each i, choose a Borel representative $h_i : \mathsf{X} \to \mathsf{Y}$ of $A[f_i]$. Also fix $y_* \in \mathsf{Y}$ and set $h_0(x) = y_*$.

We now define the weights on all of $\mathcal{P}(\mathsf{G})$. Put

$$b_i(\mu) = \max\{\delta - W_1(\mu,\mu_i), 0\}, \qquad B(\mu) = \max\left\{\sum_{i=1}^m b_i(\mu), \delta/2\right\},$$

and set

$$w_i(\mu) = \frac{b_i(\mu)}{B(\mu)} \quad (i = 1,\dots,m), \qquad w_0(\mu) = 1 - \sum_{i=1}^m w_i(\mu).$$

All these weights are nonnegative and sum to one. They are globally Lipschitz: the functions b_i are bounded and Lipschitz, and the Lipschitz denominator B is bounded below by $\delta/2$. If $\mu \in \mathcal{K}$, the net property gives $\sum_i b_i(\mu) \ge \delta/2$, so $w_0(\mu) = 0$. Moreover, $w_i(\mu) > 0$ implies $W_1(\mu,\mu_i) < \delta$.

Define

$$\mathcal{N}(\mu,x) = \sum_{i=0}^m w_i(\mu) h_i(x).$$

This map is Borel measurable and takes values in Y by convexity. Since the h_i are uniformly bounded and the weights are Lipschitz, there is a constant $C_\epsilon > 0$ such that, for all $\mu,\nu \in \mathcal{P}(\mathsf{G})$,

$$\begin{aligned}\sup_{x\in\mathsf{X}} |\mathcal{N}(\mu,x) - \mathcal{N}(\nu,x)| &\le \sum_{i=0}^m \left(\sup_{x\in\mathsf{X}} |h_i(x)|\right) |w_i(\mu) - w_i(\nu)| \\ &\le C_\epsilon\, W_1(\mu,\nu).\end{aligned}$$

Thus $\mathcal{N}$ is uniformly discretization stable on $\mathcal{P}(\mathsf{G})$ with $\beta = 1$; the constant C_ϵ may depend on the requested accuracy.

Finally, fix $f \in \mathcal{D}$. At $\mu = \gamma_f$, the weight w_0 vanishes and each active weight w_i corresponds to an output $A[f_i]$ within $\epsilon/2$ of $A[f]$ in L^2. Therefore

$$\|\mathcal{N}(\gamma_f, \cdot) - A[f]\|_{L^2} \le \sum_{i=1}^{m} w_i(\gamma_f)\, \|A[f_i] - A[f]\|_{L^2} \le \epsilon/2.$$

Taking the supremum over $f \in \mathcal{D}$ gives the required bound, since $\epsilon/2 < \epsilon$. □

The approximating map can also be chosen to be a stable Markov kernel network. For nonempty $\mathcal{D}$, all measures in $\mathcal{K}$ have spatial marginal λ, so $\mathcal{K}$ is admissible by the same argument used in the operator-characterization corollary. Applying Theorem 3.6 to the restriction of the constructed map to $\mathcal{K} \times \mathsf{X}$ gives a stable Markov kernel network with the same approximation guarantee. For empty $\mathcal{D}$, a constant Markov kernel network suffices. This conclusion concerns the abstract architecture of Definition 3.3.

D Concrete subclasses of Markov kernel networks

We give explicit realizations of transformers, neural operators, DeepONets, and Flowers in Definition 3.3. In each case, the Markov kernels specify where a head reads, the value maps specify what it reads, and the channel mixer combines the reads with the query token. The constructions reproduce the stated architectures exactly on their specified input contexts. They do not assert discretization stability: in particular, point evaluation may be unstable in the graph-Wasserstein metric.

D.1 Conventions, context updates, and output queries

We work with fixed parameters and finitely many layers on a compact metric spatial domain X with at least two points, as in the main text. Hidden tokens have the form $z = (x, a) \in \mathsf{E}_\ell = \mathsf{X} \times \mathsf{A}_\ell$, where $\mathsf{A}_\ell \subset \mathbb{R}^{d_\ell}$ is a compact box containing the feature values under consideration. Thus these are bounded-state realizations, rather than claims about an architecture on an entire unbounded function space. A pointwise input lifting is encoded as $q_{\mathsf{G}}(x, y) = (x, e^\theta(x, y))$. The coordinate x is retained as a spatial label; it need not itself be supplied as a feature to every learned map. Continuum realizations using point reads below assume continuous hidden fields.

For a continuous hidden field $u : \mathsf{X} \to \mathsf{A}$ and a fixed probability measure λ with $\operatorname{supp} \lambda = \mathsf{X}$, write

$$j_u(x) = (x, u(x)), \qquad \mu_u = (j_u)_\# \lambda. \tag{46}$$

If $\Gamma(\mu_u, (x, u(x))) = (x, u^+(x))$, its context update is exactly $B_\Gamma(\mu_u) = (\mathrm{id}, u^+)_\# \lambda$. Likewise, for an empirical context $\mu = \sum_i \omega_i \delta_{(x_i, a_i)}$, with positive weights summing to one, the pushforward updates all features simultaneously while retaining the labels and weights:

$$B_\Gamma(\mu) = \sum_i \omega_i \delta_{\Gamma(\mu, (x_i, a_i))}.$$

Empirical formulas below reproduce the corresponding discrete computation; they are not assertions that a quadrature sum equals its continuum integral.

Supported point reads. Continuity and full spatial support imply $\operatorname{supp} \mu_u = j_u(\mathsf{X})$. The graph law therefore determines u, and the addressed read

$$D_y[\mu_u] := \delta_{(y, u(y))}, \qquad y \in \mathsf{X}, \tag{47}$$

is supported on the context. For a finite context with distinct spatial labels, $D_{x_i}[\mu] = \delta_{(x_i, a_i)}$ is similarly valid at a supplied label. An arbitrary off-grid point requires a specified reconstruction rule. These reads are Borel maps on their stated domains. Indeed, on continuous graph contexts, $u \mapsto \mu_u$ is a continuous injection from the Polish space $C(\mathsf{X}, \mathsf{A})$ into $\mathcal{P}(\mathsf{E})$; its image and inverse are Borel, and point evaluation is jointly continuous in (u, y). For empirical contexts, normalizing

the restriction of μ to $\{y\} \times \mathsf{A}$ whenever this fiber has positive mass gives a jointly Borel supported read; it equals the stated Dirac measure when the feature at y is unique.

A final pointwise projection Q is consequently implemented by

$$\begin{aligned} q_{\mathsf{X}}(x) &= (x, a_*), \quad & K_p[\mu, z] &= D_{\pi_{\mathsf{X}} z}[\mu], \\ v_p(y, b) &= b, & m_p(z, r) &= Q(r), \end{aligned} \tag{48}$$

where a_* is any fixed feature vector in the final box. In particular, the encoder does not need access to the unknown output feature: the decoder kernel obtains it from the context. On a grid this realizes the projection at existing labels. Off-grid reconstruction requires its own supported reads and mixer, as illustrated for Flowers below.

Domains of the component maps. Each construction specifies the contexts on which equality with the original architecture is claimed. To obtain kernels on all of $\mathcal{P}(\mathsf{E}) \times \mathsf{E}$, use $K[\mu, z] = \mu$ whenever a stated domain or mass condition fails, including an absent graph, sensor, or positive denominator. The relevant graph and fixed-grid context classes are Borel, so this gives supported Borel extensions. Values are bounded on the compact feature spaces. Since probability reads lie in the closed convex hulls of their value ranges, the mixers can be extended to arbitrary read vectors by clipping to bounding boxes and then clipping their outputs to sufficiently large feature boxes. Such extensions leave the stated realizations unchanged. For a prescribed compact convex output set Y, projection onto Y likewise leaves represented outputs in Y unchanged. None of these extensions is claimed to preserve stability automatically.

D.2 Transformers

We first realize a deterministic Transformer block with scaled dot-product attention, a positionwise feed-forward network, residual connections, and layer normalization (Vaswani et al., 2017, Sections 3.1–3.3). We use additive positional features: the lifting is

$$q_{\mathsf{G}}(x, y) = \big(x, E^\theta(y) + p^\theta(x)\big), \tag{49}$$

where the embedding E^θ and positional map p^θ take values in $\mathbb{R}^d$. Thus a token is $z = (x, a)$, with $a \in \mathbb{R}^d$. Its first component retains the spatial label, while its second component is the usual Transformer feature vector, including its initial positional encoding. Subsequent attention and normalization act on the second component.

Suppress the layer index. For head h, let k_h be the query/key dimension and d_h the value dimension. Choose

$$Q_h^\theta, C_h^\theta \in \mathbb{R}^{k_h \times d}, \qquad V_h^\theta \in \mathbb{R}^{d_h \times d}, \qquad W_h^\theta \in \mathbb{R}^{d \times d_h}. \tag{50}$$

The symbol C denotes the key matrix, reserving K for a Markov kernel. For $z = (x, a)$ and $w = (y, b)$, define

$$s_h^\theta(z, w) := \frac{\langle Q_h^\theta a, C_h^\theta b \rangle}{\sqrt{k_h}}, \tag{51}$$

$$K_h[\mu, z](\mathrm{d}w) := \frac{\exp s_h^\theta(z, w)}{\int_{\mathsf{E}} \exp s_h^\theta(z, \widetilde{w})\, \mu(\mathrm{d}\widetilde{w})}\, \mu(\mathrm{d}w), \qquad v_h(w) := V_h^\theta b. \tag{52}$$

Each row is a probability measure supported on $\operatorname{supp} \mu$. The read r_h has dimension d_h, so $\sum_h W_h^\theta r_h \in \mathbb{R}^d$; this sum is the usual output projection of concatenated head values. The scaling involves the query/key dimension k_h, which need not equal the value dimension d_h.

The positionwise feed-forward map is

$$F^\theta(c) = A_2^\theta \sigma(A_1^\theta c + b_1^\theta) + b_2^\theta, \qquad A_1^\theta \in \mathbb{R}^{d_{\mathrm{ff}} \times d}, \quad A_2^\theta \in \mathbb{R}^{d \times d_{\mathrm{ff}}}, \tag{53}$$

where $b_1^\theta \in \mathbb{R}^{d_{\mathrm{ff}}}$, $b_2^\theta \in \mathbb{R}^d$, and σ is applied coordinatewise (ReLU in the original Transformer). Let $N_1^\theta, N_2^\theta : \mathbb{R}^d \to \mathbb{R}^d$ be tokenwise layer-normalization maps, each with its own scale and shift in $\mathbb{R}^d$ and a fixed positive ε in its denominator. The post-normalized block is precisely the channel mixer

$$t = N_1^\theta\left(a + \sum_{h=1}^{H} W_h^\theta r_h\right), \qquad m\big((x, a), r_1, \dots, r_H\big) = \big(x, N_2^\theta(t + F^\theta(t))\big). \tag{54}$$

These formulas describe evaluation with dropout disabled. Residual connections, the feed-forward map, and layer normalization are included in the mixer.

For the post-normalized block and $\mu = \sum_{j=1}^{n} \omega_j \delta_{z_j}$, with $\omega_j > 0$ and $\sum_j \omega_j = 1$, the heads give

$$r_h(\mu, z_i) = \sum_{j=1}^{n} \alpha_{ij}^h V_h^\theta a_j, \qquad \alpha_{ij}^h = \frac{\omega_j \exp s_h^\theta(z_i, z_j)}{\sum_{q=1}^{n} \omega_q \exp s_h^\theta(z_i, z_q)}. \tag{55}$$

Uniform weights recover ordinary rowwise softmax; general weights give its quadrature-weighted counterpart. If $m(z_i, r_1, \dots, r_H) = (x_i, a_i^+)$, the bulk update is

$$B_\Gamma(\mu) = \sum_{i=1}^{n} \omega_i \delta_{(x_i, a_i^+)}. \tag{56}$$

Thus every token is updated from the same previous context, and the pushforward retains its spatial label and sampling weight.

On a finite sequence, causal and padding masks are included by multiplying the exponential in both the numerator and denominator of (52) by the prescribed mask $M_h(z, w) \in \{0, 1\}$. Every represented query must have at least one permitted source token. This is again a supported probability row.

For a final pointwise output map O^θ, the supported graph read $D_x[\mu_L]$ from (47), followed by the feature projection and O^θ, realizes the continuum output $O^\theta(u_L(x))$; on an empirical context this is the native readout at supplied locations. Alternatively, an attention decoder uses a learned query feature $\psi^\theta(x)$ and the score

$$s_{p,h}^\theta((x, a), (y, b)) = \frac{\langle Q_{p,h}^\theta \psi^\theta(x), C_{p,h}^\theta b \rangle}{\sqrt{k_{p,h}}}. \tag{57}$$

Insert this score into (52), with value map $V_{p,h}^\theta b$. The decoder mixer combines these reads with the query feature $\psi^\theta(x)$ and applies the output projection. The query encoder may retain the fixed placeholder in (48), since the score and mixer use x directly. The query need not be an atom of the source context.

D.3 Neural operators

On a common spatial domain, with the identity local coordinate map and a normalized integration measure, the nonlinear integral layers of Kovachki et al. (2023, Section 3) have the form

$$u^+(x) = \sigma\left(W^\theta u(x) + b^\theta(x) + \int_{\mathsf{X}} \kappa^\theta[u](x, y) u(y)\, \lambda(\mathrm{d}y) \right). \tag{58}$$

Here $u(x) \in \mathbb{R}^d$, the updated feature lies in $\mathbb{R}^p$, $W^\theta \in \mathbb{R}^{p \times d}$, and the integral kernel is matrix-valued. Its dependence on the current features, or on retained input channels, is permitted. In graph-context notation, write the integral as

$$I_i^\theta(\mu, z) = \sum_{j=1}^{d} \int_{\mathsf{E}} k_{ij}^\theta(\mu, z, w) b_j\, \mu(\mathrm{d}w), \qquad z = (x, a), \quad w = (y, b). \tag{59}$$

For example, $k^\theta(\mu_u, (x, u(x)), (y, u(y))) = \kappa^\theta[u](x, y)$. We take $k^\theta : \mathcal{P}(\mathsf{E}) \times \mathsf{E} \times \mathsf{E} \to \mathbb{R}^{p \times d}$ Borel measurable, allowing context dependence subject to the mass bound below. The signed matrix entries need not themselves be probability densities. A fixed finite integration measure $\nu = c\lambda$ is included by absorbing $c = \nu(\mathsf{X})$ into the kernel.

Choose continuous $\phi_0, \phi_1 : \mathsf{X} \to [0, 1]$ and nonzero continuous $\rho_0, \rho_1 : \mathsf{X} \to [0, \infty)$ such that

$$\phi_0 + \phi_1 = 1, \qquad \phi_s \rho_s = 0, \qquad s \in \{0, 1\}. \tag{60}$$

Such functions exist because X has at least two points: choose a function with plateaus 0 and 1 on disjoint neighborhoods, its complement, and bumps inside their respective zero regions. The partition splits each value into two pieces. Each piece then has a region where additional probability mass contributes zero to its read.

Lemma D.1 (Markov realization of signed integral layers)**.** *Let the feature range be compact and fix the parameters θ. Suppose that, on the represented contexts and queries, the positive and negative parts of every entry satisfy*

$$a_{ij}^{\pm}(\mu, z) := \int_{\mathsf{E}} (k_{ij}^{\theta})^{\pm}(\mu, z, w)\, \mu(\mathrm{d}w) \le M \tag{61}$$

for some fixed $M > 0$. For the functions in (60)*, also assume $c_s(\mu) := \int_{\mathsf{E}} \rho_s(y)\, \mu(\mathrm{d}w) > 0$ for $s = 0, 1$. Then* (59) *is recovered exactly from $4pd$ scalar Markov read heads, each absolutely continuous with respect to μ, by a linear channel mixer.*

Proof. Define the completion probabilities

$$\nu_s[\mu](\mathrm{d}w) := \frac{\rho_s(y)}{c_s(\mu)}\, \mu(\mathrm{d}w). \tag{62}$$

For the full-support graph contexts of Section D.1, $c_s(\mu_u) = \int_{\mathsf{X}} \rho_s \, \mathrm{d}\lambda > 0$ is independent of u.

Index the heads by $h = (i, j, s, \epsilon)$, where $\epsilon \in \{+, -\}$. Set

$$K_{ij,s,\epsilon}[\mu, z](\mathrm{d}w) := \frac{(k_{ij}^{\theta})^{\epsilon}(\mu, z, w)}{M}\, \mu(\mathrm{d}w) + \left(1 - \frac{a_{ij}^{\epsilon}(\mu, z)}{M}\right) \nu_s[\mu](\mathrm{d}w), \tag{63}$$

$$v_{ij,s,\epsilon}(w) := \phi_s(y) b_j. \tag{64}$$

The two nonnegative summands in (63) have masses a_{ij}^{ϵ}/M and $1 - a_{ij}^{\epsilon}/M$, respectively. Hence $K_{ij,s,\epsilon}$ is a probability measure supported on $\operatorname{supp} \mu$. The bounded value map depends only on w. Since it vanishes $\nu_s[\mu]$-almost everywhere, its read is

$$r_{ij,s,\epsilon}(\mu, z) = \frac{1}{M} \int_{\mathsf{E}} (k_{ij}^{\theta})^{\epsilon}(\mu, z, w) \phi_s(y) b_j \, \mu(\mathrm{d}w). \tag{65}$$

Summing over the partition and subtracting the signed parts gives

$$I_i^{\theta}(\mu, z) = M \sum_{j=1}^{d} \sum_{s=0}^{1} \big(r_{ij,s,+}(\mu, z) - r_{ij,s,-}(\mu, z) \big). \tag{66}$$

□

Thus the channel mixer for (58) is

$$\begin{aligned} m^{\theta}\big((x, a), r\big) &:= \big(x, \sigma\big(W^{\theta} a + b^{\theta}(x) + J(r)\big)\big), \\ J_i(r) &:= M \sum_{j=1}^{d} \sum_{s=0}^{1} (r_{ij,s,+} - r_{ij,s,-}). \end{aligned} \tag{67}$$

In particular, the masses $a_{ij}^{\pm}$ remain inside the kernels; the mixer does not require them as additional inputs. Preserving x makes the context pushforward precisely the graph law of u^{+}. Applying this construction at every layer, together with the lifting of Section D.1 and the decoder (48), recovers the complete network whenever the represented hidden fields remain continuous. Usual continuous local kernels and pointwise maps have this property. Borel kernels alone need not preserve it. Any original-input channels used by the kernels are retained through the layers.

The pointwise bound $|k_{ij}^{\theta}| \le M$ is sufficient for (61), but a uniform row-mass bound is all the proof uses. For the usual continuous local kernel parameterizations, compact feature ranges give a pointwise bound. General dependence on the entire context requires the mass bound separately. The same identities hold for a weighted empirical context $\mu = \sum_n \omega_n \delta_{(x_n, u_n)}$ whenever (61) holds and $\sum_n \omega_n \rho_s(x_n) > 0$ for both s. They then reproduce the corresponding quadrature layer exactly. The completion condition holds eventually along spatial quadratures converging weakly to λ, but need not hold on arbitrary coarse samples. These are representation statements; no discretization-stability assertion follows from the signed decomposition alone.

D.3.1 FOURIER NEURAL OPERATORS

The finite-mode spectral layers of Fourier neural operators (Li et al., 2021) admit a simpler realization. Let $\mathsf{X} = \mathbb{T}^{d_x}$ with normalized volume λ, and let $\Lambda \subset \mathbb{Z}^{d_x}$ be the finite set of retained modes. With $R_k^\theta \in \mathbb{C}^{p\times d}$, their spectral contribution is

$$S^\theta[u](x) = \sum_{k\in\Lambda} e^{2\pi \mathrm{i} k\cdot x} R_k^\theta \widehat{u}(k), \qquad \widehat{u}(k) = \int_{\mathsf{X}} e^{-2\pi \mathrm{i} k\cdot y} u(y)\, \lambda(\mathrm{d}y). \tag{68}$$

Take the usual conjugate symmetry of the modes and multipliers so that $S^\theta[u]$ is real. A single Markov head with

$$K[\mu, z] = \mu, \qquad v(y, b) = \big(\cos(2\pi k\cdot y)b,\, \sin(2\pi k\cdot y)b\big)_{k\in\Lambda} \tag{69}$$

returns $2d|\Lambda|$ real Fourier moments $r = (C_k, S_k)_{k\in\Lambda}$, with $\widehat{u}(k) = C_k - \mathrm{i}S_k$. The channel mixer is

$$\begin{aligned} m^\theta((x, a), r) &:= \big(x, \sigma\big(W^\theta a + b^\theta(x) + F^\theta(x, r)\big)\big)\,, \\ F^\theta(x, r) &:= \mathrm{Re} \sum_{k\in\Lambda} e^{2\pi \mathrm{i} k\cdot x} R_k^\theta (C_k - \mathrm{i}S_k). \end{aligned} \tag{70}$$

Complex notation in this expression abbreviates a real linear map on the sine and cosine moments. Thus the Markov kernel performs an unweighted context read, while the value map collects the Fourier moments and the mixer applies the learned multipliers and synthesizes the result at x. The number of read channels depends on the retained modes, not on the spatial resolution.

On a uniform periodic grid, $\mu = N^{-1}\sum_{n=1}^N \delta_{(x_n,u_n)}$ gives normalized discrete Fourier coefficients, and (70) agrees with the corresponding FFT layer when the forward and inverse normalization conventions are matched. Retained modes must be represented with the same grid-frequency and conjugate-symmetry conventions. Other positive quadrature weights give the associated quadrature Fourier layer; equality to the continuum coefficients requires a separate quadrature argument. The formulas use the actual normalized-volume integration measure, or its declared discrete quadrature. Other fixed domain volumes or transform conventions require their known normalization factors. The local term, pointwise maps, and final projection are handled as in Section D.1.

D.4 DEEPONETS

A fixed-sensor DeepONet pairs coefficients from a branch network with features from a trunk network (Lu et al., 2019, Section 2.1). For scalar output, sensors $s_1, \dots, s_S \in \mathsf{X}$, and branch/trunk width P, its map is

$$\mathcal{N}(\mu_f, x) = \big\langle B^\theta\big(f(s_1), \dots, f(s_S)\big), T^\theta(x)\big\rangle + b_0^\theta, \tag{71}$$

where $B^\theta : \mathsf{A}^S \to \mathbb{R}^P$ and $T^\theta : \mathsf{X} \to \mathbb{R}^P$ are the branch and trunk maps. Vector outputs are obtained by applying this construction to each output component. We use the common input/output domain of Definition 3.3.

Take the identity lifting $q_{\mathsf{G}}(x, a) = (x, a)$. To satisfy $L > 0$ in that definition, use one identity context layer: $K_0[\mu, z] = \mu$, $v_0 \equiv 0$, and $m_0(z, r) = z$ give $B_0(\mu) = \mu$. All sensor extraction and branch–trunk synthesis can then occur in the decoder. Using the supported evaluation rows (47), choose

$$K_{p,j}[\mu, (x, a)] = D_{s_j}[\mu], \qquad v_{p,j}(y, b) = b, \qquad j = 1, \dots, S. \tag{72}$$

On a continuous graph context μ_f, the jth read is exactly $r_j = f(s_j)$. The decoder mixer

$$m_p\big((x, a), r_1, \dots, r_S\big) = \langle B^\theta(r_1, \dots, r_S), T^\theta(x)\rangle + b_0^\theta \tag{73}$$

therefore gives (71). It ignores the placeholder feature a and uses x in the trunk. Signed branch and trunk outputs are permitted because the probability constraint applies to the read measures.

For an empirical input, the same realization is exact when every sensor location occurs in the context with its observed value. The output query x may be any point of X, since its dependence enters through the trunk and does not require a sample at x. If a sensor is absent, its value requires a declared reconstruction rule. The fixed-sensor realization itself provides no such reconstruction and no uniform graph-W_1 stability assertion.

D.5 FLOWERS

The SelfWarp heads, residual blocks, and multiscale composition of Flowers are specified in Muser et al. (2026, Appendix C). We first represent a block using the graph-context conventions of Section D.1, with $u : \mathsf{X} \to \mathbb{R}^{C_{\mathrm{in}}}$ continuous and bounded. Here X is a compact Euclidean domain or a torus. Let $C_{\mathrm{out}} = HC_h$, and write the value projection in head blocks as $V_h \in \mathbb{R}^{C_h \times C_{\mathrm{in}}}$. The displacement maps $g_h : \mathbb{R}^{C_{\mathrm{in}}} \to \mathbb{R}^{d_x}$ use all input channels. For $z = (x, a)$, let $\tau_h(z) = x + g_h(a)$, interpreted with periodic wrapping, or assume that this address lies in X. Set

$$K_h[\mu_u, z] = D_{\tau_h(z)}[\mu_u] = \delta_{(\tau_h(z), u(\tau_h(z)))}, \qquad v_h(y, b) = V_h b. \tag{74}$$

Then $r_h(\mu_u, (x, u(x))) = V_h u(x + g_h(u(x)))$. Thus the mixer

$$m_{\mathrm{warp}}((x, a), r_1, \ldots, r_H) = \left(x,\ Pa + \bigoplus_{h=1}^{H} r_h \right)$$

realizes SelfWarp plus its pointwise residual projection $P \in \mathbb{R}^{C_{\mathrm{out}} \times C_{\mathrm{in}}}$. For affine projections, value biases are included in v_h and residual biases in the mixer. The output retains the target coordinate x. Since λ has full support, the point selected in (74) belongs to $\operatorname{supp} \mu_u$.

Normalization. Write b for the field produced by this first layer. GroupNorm uses moments of b across space and channel groups, so it is a second Markov kernel layer. For every group $G \subset \{1, \ldots, C_{\mathrm{out}}\}$, take $K_G[\mu_b, z] = \mu_b$ and

$$v_G(y, b) = \left(\frac{1}{|G|} \sum_{c \in G} b_c,\ \frac{1}{|G|} \sum_{c \in G} b_c^2 \right).$$

The resulting read gives the group mean M_G and second moment S_G. Its mixer preserves x and maps each feature channel $c \in G$ to

$$\mathrm{GELU}\left(\gamma_c \frac{b_c - M_G}{\sqrt{\max\{S_G - M_G^2, 0\} + \varepsilon}} + \beta_c \right), \qquad \varepsilon > 0. \tag{75}$$

For actual moment reads $S_G - M_G^2 \geq 0$; the maximum only extends the mixer to arbitrary read vectors. This reproduces the order residual sum, GroupNorm, then GELU. On a grid use $K_G = n^{-1} \sum_{j=1}^{n} \delta_{z_j}$ to obtain the required uniform spatial average regardless of context weights. When supplied as input, conditioning variables for FiLM can be retained in the token and used to determine γ_c, β_c. Stacking these two-layer realizations and using the decoder (48) gives the corresponding single-scale network, including its pointwise lifting and final projection.

Interpolation and boundaries. On a fixed grid with tokens $z_j = (x_j, a_j)$, let $b_j(y) \geq 0$ be the sampler's multilinear interpolation weights. For constant-preserving boundary conventions, the implemented warp has

$$K_h[\mu, z] = \sum_j b_j(\tau_h(z)) \delta_{z_j}, \qquad \sum_j b_j(\tau_h(z)) = 1.$$

These rows are supported on the sampled tokens. With zero padding, the sum $s_h(z) := \sum_j b_j(\tau_h(z))$ can be less than one. For $s_h(z) > 0$, divide this row by $s_h(z)$ and multiply the resulting read by $s_h(z)$ in m_{warp}; when $s_h(z) = 0$, choose $K_h[\mu, z] = \mu$ and multiply the read by zero. The factor $s_h(z)$ is determined by the fixed sampler geometry and the local source address, so the mixer can compute it from z. This implements zero padding without adding unsupported zero-valued tokens.

The full multiscale network. The U-Net in Muser et al. (2026, Appendix C.4) additionally uses strided and transposed convolutions and stored encoder skips. To represent it on its finite grid hierarchy $\Lambda_0, \ldots, \Lambda_J$, retain every finest-grid label $i \in \Lambda_0$ throughout the MKN. Choose surjections $c_\ell : \Lambda_0 \to \Lambda_\ell$, with $c_0 = \mathrm{id}$, and representatives $\iota_\ell(y) \in c_\ell^{-1}(y)$. At scale ℓ, token i stores the active feature at $c_\ell(i)$ and the previously saved skip features at their corresponding coordinates. Warp addresses use the physical coordinate $c_\ell(i)$ at that scale; the original label i is retained only to identify

the carrier token. A read of a coarse site y selects the existing token $z_{\iota_\ell(y)}$ and extracts its active channels. Consequently, each finite convolution stencil is represented by Dirac heads followed by the original, possibly signed, filter matrices in the mixer; the same construction realizes transposed convolutions and changes the active scale while retaining all finest-grid labels. Out-of-domain stencil terms receive zero coefficients. Coarse-grid interpolation uses the preceding formula with these representative tokens. GroupNorm uses the probability row

$$K_{\mathrm{norm},\ell}[\mu, z] = \frac{1}{|\Lambda_\ell|} \sum_{y \in \Lambda_\ell} \delta_{z_{\iota_\ell(y)}},$$

which gives exactly the uniform coarse-grid moments, irrespective of the sizes of the fibers of c_ℓ. Convolutional biases and the ReLU following each down/up convolution are applied in the mixer. Saved skip slots are copied unchanged through the active-channel operations; copying and concatenation are pointwise mixers, performed at the original architecture's stages. The encoder is $q_{\mathsf{G}}(x, a) = (x, \mathrm{Lift}(a \oplus x))$, with empty skip slots appended; the final supported point read at a finest-grid query, followed by the original pointwise projection, is Γ_p. Hence the complete finite-grid Flowers architecture is an MKN. Retaining the finest-grid labels avoids requiring a pushforward layer to create new tokens during upsampling. These exact realizations do not by themselves assert discretization stability of the learned point reads or of the multiscale network.

E Hybrid architectures through variational Markov reads

We construct a convex family of regularizers that recovers attention, uniform reads, and continuum SelfWarp at different parameter values. The read kernels satisfy a common Hölder bound on a fixed class of regular graph contexts. We also give a grid approximation bound and explain why the graph-context restriction is needed.

E.1 Graph contexts and costs

Let $\mathsf{X} = [0,1]^d$, with $d \geq 1$ and normalized Lebesgue probability measure $\lambda = \mathrm{d}x$. Let V be a nonempty compact convex subset of a Euclidean space, and put $\mathsf{E} = \mathsf{X} \times \mathsf{V}$ with its Euclidean product metric. For $L_u \geq 0$, define

$$\mathcal{U}_{L_u} := \{u : \mathsf{X} \to \mathsf{V} : \mathrm{Lip}(u) \leq L_u\}, \qquad j_u(y) := (y, u(y)), \qquad \gamma_u := (j_u)_{\#}\lambda.$$

Every probability measure supported on $\operatorname{supp}\gamma_u = j_u(\mathsf{X})$ has the unique form $\kappa = (j_u)_{\#}\nu$, with $\nu \in \mathcal{P}(\mathsf{X})$. Consequently,

$$\tfrac{1}{2}W_2^2\big((\pi_{\mathsf{X}})_{\#}\kappa, (\pi_{\mathsf{X}})_{\#}\gamma_u\big) = \tfrac{1}{2}W_2^2(\nu, \lambda), \qquad \mathrm{KL}(\kappa\|\gamma_u) = \mathrm{KL}(\nu\|\lambda).$$

The KL identity follows because j_u is a homeomorphism onto its graph; it also holds when the value is infinite.

To specify a common cost class, set

$$\mathcal{T}_{L_u} := \{(\gamma_u, z, j_u(y)) : u \in \mathcal{U}_{L_u},\ z \in \mathsf{E},\ y \in \mathsf{X}\}.$$

Fix $M, L_c, L_y \geq 0$ and $\beta \in (0,1]$. Let $\mathfrak{C}$ consist of the continuous real costs on $\mathcal{T}_{L_u}$ satisfying

$$\begin{aligned} |c(\gamma_u, z, j_u(y))| &\leq M, \\ |c(\gamma_u, z, j_u(y)) - c(\gamma_v, z', j_v(y'))| &\leq L_c\big(W_1(\gamma_u, \gamma_v) + \rho_{\mathsf{E}}(z, z')\big)^\beta + L_y|y - y'|. \end{aligned} \tag{76}$$

E.2 A stable Wasserstein–KL family

Fix $\tau > 0$. For $\alpha \in [0,1]$, define

$$J_{\alpha,c;u,z}(\nu) := \int_{\mathsf{X}} c(\gamma_u, z, j_u(y))\,\nu(dy) + \frac{1-\alpha}{2}W_2^2(\lambda, \nu) + \alpha\tau\,\mathrm{KL}(\nu\|\lambda), \qquad \nu \in \mathcal{P}(\mathsf{X}). \tag{77}$$

At $\alpha = 0$ the entropy term is omitted, rather than interpreted as $0 \cdot (+\infty)$.

Theorem E.1 (A convex class of stable variational reads)**.** *For every $c \in \mathfrak{C}$, $\alpha \in [0,1]$, $u \in \mathcal{U}_{L_u}$, and $z \in \mathsf{E}$, the objective* (77) *has a unique minimizer $\nu_{\alpha,c;u,z}$. Define the supported graph-token read*

$$K_{\alpha,c}[\gamma_u, z] := (j_u)_{\#}\nu_{\alpha,c;u,z}. \tag{78}$$

Set

$$\eta := \min\left\{\frac{\beta}{3}, \frac{1}{d+1}\right\}, \qquad \omega(0) := 0, \qquad \omega(h) := h\log(e/h) \quad (0 < h \le 1).$$

There is a constant A, depending only on $d, L_u, M, L_c, L_y, \beta, \tau$, and $\operatorname{diam}(\mathsf{V})$, such that

$$\begin{aligned} &W_1\big(K_{\alpha,c}[\gamma_u, z], K_{\widetilde{\alpha},\widetilde{c}}[\gamma_v, z']\big) \\ &\quad \le A\left[\big(W_1(\gamma_u,\gamma_v) + \rho_{\mathsf{E}}(z,z')\big)^{\eta} + \big(\|c-\widetilde{c}\|_{\infty,\mathcal{T}_{L_u}} + \omega(|\alpha-\widetilde{\alpha}|)\big)^{1/3}\right]. \end{aligned} \tag{79}$$

In particular, each fixed design defines a Hölder-stable Borel Markov kernel on this graph-context class. The design space $\mathfrak{C} \times [0,1]$ is convex and compact for uniform convergence of costs and ordinary convergence of α.

The exponents are not claimed optimal. The estimate controls read kernels on the stated graph class; applying it through multiple layers requires preservation of the spatial marginal and the uniform graph regularity. It does not supply the all-context kernel bound in Definition 3.4.

E.3 Proof of Theorem E.1

The main step is cubic transport growth under *linear interpolation of measures*; entropy adds quadratic growth. Together they control the spatial minimizer uniformly. We then lift to graph tokens and compare coefficients using finite-entropy histograms, since a point read itself has infinite entropy.

Throughout this appendix, $D = \sqrt{d}$ and spatial Wasserstein distances use Euclidean distance on $\mathsf{X} = [0,1]^d$. For the two penalties, write $F(\nu) = \frac{1}{2}W_2^2(\lambda,\nu)$ and $H(\nu) = \mathrm{KL}(\nu\|\lambda)$. Set

$$q_d = (12dD)^{-1}, \qquad m_{d,\tau} = \min\{q_d,\, 2\tau/D^3\} > 0. \tag{80}$$

For the spatial proof, $c \in C(\mathsf{X})$ denotes a single continuous function of position, and $\nu_{\alpha,c}$ denotes its minimizing spatial measure. The corresponding token read is its pushforward through j_u. Write

$$J_{\alpha,c}(\nu) = \int_{\mathsf{X}} c\,\mathrm{d}\nu + (1-\alpha)F(\nu) + \alpha\tau H(\nu), \qquad V(c) = \min_{\nu} J_{0,c}(\nu).$$

We use standard compact Kantorovich duality and standard Wasserstein facts; the estimates needed below are proved explicitly.

Transport provides quantitative growth

Growth at zero entropy. For every $c \in C(\mathsf{X})$, $J_{0,c}$ has a unique minimizer ν_c. For every $\nu \in \mathcal{P}(\mathsf{X})$,

$$J_{0,c}(\nu) - V(c) \ge q_d W_1(\nu,\nu_c)^3. \tag{81}$$

Consequently, if $\nu_b = (1-b)\nu_0 + b\nu_1$ and $r = W_1(\nu_0,\nu_1)$, then

$$(1-b)F(\nu_0) + bF(\nu_1) - F(\nu_b) \ge q_d b(1-b)\big(b^2 + (1-b)^2\big)r^3. \tag{82}$$

Proof. Identify the optimizer. Define on all of $\mathbb{R}^d$

$$\Phi_c(x) = \sup_{y\in\mathsf{X}}\{x\cdot y - \tfrac{1}{2}|y|^2 - c(y)\}.$$

This convex Lipschitz function is differentiable almost everywhere, and its gradients lie in X. At a differentiability point its maximizing destination is unique and equals $T_c(x) = \nabla\Phi_c(x)$. Define T_c arbitrarily on the null exceptional set. For every source–destination pair,

$$\tfrac{1}{2}|x-y|^2 + c(y) \ge \tfrac{1}{2}|x|^2 - \Phi_c(x).$$

The coupling $(\mathrm{id}, T_c)_{\#}\lambda$ attains equality. Integrating against an optimal coupling for any candidate row gives

$$\nu_c = (T_c)_{\#}\lambda, \qquad V(c) = \int_{\mathsf{X}} \big(\tfrac{1}{2}|x|^2 - \Phi_c(x)\big)\,\mathrm{d}\lambda. \tag{83}$$

Every minimizing coupling must have destination $T_c(x)$ for almost every source, proving uniqueness.

Control the response to a cost perturbation. If Φ, Ψ are smooth convex functions with gradients in X, put $u = \Phi - \Psi$ and $e = \|u\|_{L^\infty(\mathsf{X})}$. Integration by parts on the cube gives

$$\int_{\mathsf{X}} |\nabla u|^2\,\mathrm{d}x = \int_{\partial\mathsf{X}} u\,\partial_n u\,\mathrm{d}S - \int_{\mathsf{X}} u(\Delta\Phi - \Delta\Psi)\,\mathrm{d}x \le 4de.$$

Indeed, each gradient component belongs to $[0,1]$, so the boundary integral of $|\partial_n u|$ is at most $2d$. Convexity and the divergence theorem give $0 \le \int_{\mathsf{X}} \Delta\Phi \le d$ and the same bound for Ψ. For $\Phi_c, \Phi_{\widetilde{c}}$, global mollification preserves convexity, the gradient range, and the bound $\|\Phi_c - \Phi_{\widetilde{c}}\|_{L^\infty(\mathbb{R}^d)} \le \|c - \widetilde{c}\|_\infty$. Dominated convergence of the mollified gradients therefore yields

$$W_1(\nu_c, \nu_{\widetilde{c}}) \le W_2(\nu_c, \nu_{\widetilde{c}}) \le 2\sqrt{d}\,\|c - \widetilde{c}\|_\infty^{1/2}. \tag{84}$$

Turn response control into growth. Set $r = W_1(\nu, \nu_c)$ and choose a 1-Lipschitz Kantorovich–Rubinstein test h with $\int h\,\mathrm{d}(\nu - \nu_c) = r$ and $\|h\|_\infty \le D$. Testing the minimizers at neighboring costs in each other's objectives, then using (84), gives

$$\frac{\mathrm{d}}{\mathrm{d}t} V(c - th) = -\int h\,\mathrm{d}\nu_{c-th}.$$

The derivative is continuous by (84). Integrating it and using $\mathrm{Lip}(h) \le 1$ gives

$$V(c - th) \ge V(c) - t\int h\,\mathrm{d}\nu_c - \tfrac{4}{3}\sqrt{dD}\,t^{3/2}.$$

On the other hand, $V(c - th) \le J_{0,c}(\nu) - t\int h\,\mathrm{d}\nu$. Hence

$$J_{0,c}(\nu) - V(c) \ge tr - \tfrac{4}{3}\sqrt{dD}\,t^{3/2}.$$

Choosing $t = r^2/(4dD)$ proves (81); the case $r = 0$ follows from minimality.

Finally, compact Kantorovich duality supplies a continuous destination potential b_* supporting F at ν_b: $F(\nu) \ge F(\nu_b) + \int b_*\,\mathrm{d}(\nu - \nu_b)$. Thus ν_b minimizes $J_{0,-b_*}$. Apply (81) to ν_0, ν_1 and add with weights $1-b, b$. The linear terms cancel, while the dual norm formula for W_1 gives $W_1(\nu_0, \nu_b) = br$ and $W_1(\nu_1, \nu_b) = (1-b)r$. This proves (82). □

Uniform control of the spatial minimizers

Existence for every α follows from weak lower semicontinuity on compact $\mathcal{P}(\mathsf{X})$ and the finite competitor λ. When the endpoint entropies are finite, linear interpolation has the identity

$$(1-b)H(\nu_0) + bH(\nu_1) - H(\nu_b) = (1-b)\,\mathrm{KL}(\nu_0\|\nu_b) + b\,\mathrm{KL}(\nu_1\|\nu_b).$$

Pinsker's inequality and $W_1 \le (D/2)\|\cdot\|_{\mathrm{TV}}$, with $\|\cdot\|_{\mathrm{TV}}$ the total variation norm, bound this below by $2b(1-b)W_1(\nu_0,\nu_1)^2/D^2$. Combine this with (82), take ν_0 to be a minimizer, divide the interpolation deficit by b, and let $b \downarrow 0$. For every finite-objective competitor, with $r = W_1(\nu, \nu_{\alpha,c})$, this proves

$$J_{\alpha,c}(\nu) - J_{\alpha,c}(\nu_{\alpha,c}) \ge (1-\alpha)q_d r^3 + \frac{2\alpha\tau}{D^2} r^2 \ge m_{d,\tau} r^3. \tag{85}$$

At $\alpha = 0$ no entropy assertion is used. The last inequality uses $r \le D$; it also proves uniqueness. Apply (85) in both directions to the minimizers for c and $\widetilde{c}$. Adding cancels the regularizer, so that

$$W_1(\nu_{\alpha,c}, \nu_{\alpha,\widetilde{c}}) \le m_{d,\tau}^{-1/3}\|c - \widetilde{c}\|_\infty^{1/3}, \qquad 0 \le \alpha \le 1. \tag{86}$$

This proves the spatial estimate with no positive lower bound on the entropy coefficient.

Continuity in the entropy coefficient. If $\mathrm{Lip}(c) \le L_y$, then

$$W_1(\nu_{\alpha,c}, \nu_{\widetilde{\alpha},c}) \le C_{d,L_y,\tau}\, \omega(|\alpha - \widetilde{\alpha}|)^{1/3} \quad \text{for all } \alpha, \widetilde{\alpha} \in [0,1]. \tag{87}$$

Proof. Approximate the singular endpoint. Divide X into q^d cubes and let $S_q\nu$ spread each cell's mass uniformly within that cell. Then

$$W_1(S_q\nu, \nu) \le D/q, \qquad H(S_q\nu) \le d \log q, \qquad H(S_q\nu) \le H(\nu) \tag{88}$$

when the last right-hand side is finite. The entropy statements follow from the cell probabilities and Jensen's inequality. Also

$$|F(\nu) - F(\nu')| \le D W_1(\nu, \nu'). \tag{89}$$

To see this, glue an optimal coupling from λ to ν to a coupling between ν and ν', and use that $|x-y|^2/2$ is D-Lipschitz in y on X; then interchange the targets.

Set $C_0 = D(L_y + D)$ and $V(c) = J_{0,c}(\nu_c)$. Testing $S_q\nu_c$ in the α-problem, using $0 \le F \le D^2/2$ and $H \ge 0$, gives

$$J_{0,c}(\nu_{\alpha,c}) - V(c) + \alpha\tau H(\nu_{\alpha,c}) \le C_0/q + \alpha\tau d \log q + \alpha D^2/2. \tag{90}$$

For $\alpha > 0$, take $q = \lceil 1/\alpha \rceil$. The transport growth estimate and the nonnegativity of $J_{0,c} - V(c)$ imply, respectively,

$$W_1(\nu_{\alpha,c}, \nu_c)^3 \le \frac{\alpha}{q_d}\big(C_0 + D^2/2 + \tau d \log(2/\alpha)\big), \tag{91}$$

$$H(\nu_{\alpha,c}) \le B_0 + d \log(2/\alpha), \qquad B_0 = (C_0 + D^2/2)/\tau. \tag{92}$$

Compare two positive coefficients. By symmetry assume $0 \le \alpha \le \widetilde{\alpha} \le 1$, and put $h = \widetilde{\alpha} - \alpha$. For $\alpha \ge h > 0$, write $J_{\widetilde{\alpha},c} = J_{\alpha,c} + h(\tau H - F)$ and test each minimizer in the other's objective. Equation (85) bounds the sum of the two gaps below by $2m_{d,\tau} W_1(\nu_{\alpha,c}, \nu_{\widetilde{\alpha},c})^3$. The sum equals

$$h\left[(\tau H - F)(\nu_{\alpha,c}) - (\tau H - F)(\nu_{\widetilde{\alpha},c})\right] \le h\left[\tau B_0 + \tau d \log(2/h) + D^2/2\right],$$

by (92). This proves the required modulus in this regime. If $0 \le \alpha < h$ and $h < 1/4$, both coefficients are below $2h$; compare each row to ν_c using (91). If $h \ge 1/4$, use $W_1 \le D$. The case $h = 0$ is immediate. □

FROM POSITIONS TO TOKENS AND COMPLETION OF THE PROOF

First control the graph lift. Put $\delta = W_1(\gamma_u, \gamma_v)$ and $L_j = \sqrt{1 + L_u^2}$. A coupling of the graph measures projects to a coupling of λ with itself. At a paired location (x, y),

$$|u(x) - v(x)| \le |u(x) - v(y)| + L_u|x - y| \le L_j \rho_{\mathsf{E}}(j_u(x), j_v(y)).$$

Integration and minimization give $\int_{\mathsf{X}} |u - v|\, \mathrm{d}\lambda \le L_j\delta$. Put $M_{uv} = \|u - v\|_\infty$ and $B_* = 1 + 2L_u + \operatorname{diam}\mathsf{V}$. At a maximizer, $|u - v| \ge M_{uv}/2$ within distance $M_{uv}/(2B_*) \le 1/2$. The intersection of that ball with the cube has volume bounded below by a dimensional constant times its radius to the power d. Thus

$$\|u - v\|_\infty \le C_G \delta^{1/(d+1)}, \qquad C_G = C_G(d, L_u, \operatorname{diam}\mathsf{V}). \tag{93}$$

For fixed (α, c), apply (86) to the two functions $y \mapsto c(\gamma_u, z, j_u(y))$ and $y \mapsto c(\gamma_v, z', j_v(y))$. Couple their minimizing spatial measures, then change the graph lift at the same position. Equations (76) and (93) yield

$$\begin{aligned} &W_1(K_{\alpha,c}[\gamma_u, z], K_{\alpha,c}[\gamma_v, z']) \\ &\quad \le L_j (L_c/m_{d,\tau})^{1/3} \big(W_1(\gamma_u, \gamma_v) + \rho_{\mathsf{E}}(z, z')\big)^{\beta/3} + C_G \delta^{1/(d+1)}. \end{aligned}$$

All input distances are bounded, so this is at most a constant times $(W_1(\gamma_u, \gamma_v) + \rho_{\mathsf{E}}(z, z'))^\eta$. At fixed (γ_v, z'), combine (86), the coefficient estimate (87), and the Lipschitz pushforward bound to control the change from (α, c) to $(\widetilde{\alpha}, \widetilde{c})$. The triangle inequality with intermediate row $K_{\alpha,c}[\gamma_v, z']$ proves (79).

The class $\mathcal{U}_{L_u}$ is compact in the uniform norm by Arzelà–Ascoli; $u \mapsto \gamma_u$ is continuous and injective, since λ has full support and u is continuous. Hence the set of admissible triples $(\gamma_u, z, j_u(y))$ is compact. The bounds in (76) make $\mathfrak{C}$ convex, closed, bounded, and equicontinuous there, so it is compact. Existence and uniqueness were proved in (85); pushing the spatial optimizer through j_u gives the unique token row. Continuity implies measurability. Compactness is asserted for the cost–coefficient coordinates, not for uniform values of extended-valued regularizers at $\alpha = 0$. □

Read / architecture	Regularizer	Cost
SelfWarp	$F_\mu(\kappa)$	$B\lvert\pi_{\mathsf{X}}(w) - a(\mu,z)\rvert$, $B > \sqrt{d}$
Attention	$\tau\,\mathrm{KL}(\kappa\Vert\mu)$	$-\langle Qz, K_{\mathrm{key}}w\rangle$
FNO uniform read	$\mathcal{R}_\alpha(\kappa,\mu)$	0
Representation theorem	$\frac{1}{2}\chi^2(\kappa\Vert\mu)$	$c_{jk}(\mu,z,w) = -\theta_j(\mu,z)\psi_k(w)$

Table 2: Variational read constructions. The first three rows belong to the Wasserstein–KL family on the stated graph class, with an admissible Hölder address $a(\mu,z) = \pi_{\mathsf{X}}(z) + g(\mu,z) \in \mathsf{X}$ for SelfWarp. The last row is the separate Pearson realization; its calibration head has zero cost.

E.4 Architecture examples

For graph contexts, write

$$F_\mu(\kappa) := \tfrac{1}{2}W_2^2\big((\pi_{\mathsf{X}})_\#\kappa, (\pi_{\mathsf{X}})_\#\mu\big), \qquad \mathcal{R}_\alpha(\kappa,\mu) := (1-\alpha)F_\mu(\kappa) + \alpha\tau\,\mathrm{KL}(\kappa\Vert\mu).$$

Table 2 records the three architecture reads in this family and the separate Pearson construction from the representation theorem. The latter is proved in Section B.2.2; its decoder costs need not belong to the joint cost class $\mathfrak{C}$.

At $\alpha = 1$, put $Z = \int_{\mathsf{X}} e^{-c/\tau}\,\mathrm{d}\lambda$. Direct expansion gives

$$\int_{\mathsf{X}} c\,\mathrm{d}\nu + \tau H(\nu) = \tau\,\mathrm{KL}\Big(\nu \,\Big\|\, Z^{-1}e^{-c/\tau}\lambda\Big) - \tau\log Z.$$

This proves the normalized attention row in Table 2, including extended values outside the entropy domain. For the SelfWarp cost in the table, (89) gives

$$\begin{aligned}\int_{\mathsf{X}} B\lvert y - \pi_{\mathsf{X}}z - g(\gamma_u,z)\rvert\,\nu(\mathrm{d}y) + F(\nu) - F\big(\delta_{\pi_{\mathsf{X}}z+g(\gamma_u,z)}\big)\\ \geq (B-D)W_1\big(\nu, \delta_{\pi_{\mathsf{X}}z+g(\gamma_u,z)}\big)\,.\end{aligned}$$

Therefore $B > D$ suffices for the unique spatial optimizer $\delta_{\pi_{\mathsf{X}}z+g(\gamma_u,z)}$. It need not be necessary for a particular address. For zero cost, both F and H are nonnegative and vanish only at λ, so every α gives $K[\gamma_u, z] = \gamma_u$. At $\alpha = 0$, the spatial optimizer also has the transport-map description (83). The point-read identification agrees with Proposition B.3.

A bounded Lipschitz attention score in (z,w) is spatially Lipschitz after substitution $w = j_u(y)$. Equation (93) controls its dependence on u, giving (76), for example with $\beta = 1/(d+1)$. For the warp costs,

$$\begin{aligned}&\big\lvert B\lvert y - \pi_{\mathsf{X}}z - g(\gamma_u,z)\rvert - B\lvert y' - \pi_{\mathsf{X}}z' - g(\gamma_v,z')\rvert\big\rvert\\ &\qquad \leq B\big(\lvert y - y'\rvert + \rho_{\mathsf{E}}(z,z') + \lvert g(\gamma_u,z) - g(\gamma_v,z')\rvert\big).\end{aligned}$$

A Hölder displacement map therefore gives an admissible cost. Choose a common exponent no larger than $1/(d+1)$ and the displacement exponent, and common bounds covering both families. Then the endpoints and their straight interpolation belong to $\mathfrak{C} \times [0,1]$.

E.5 Grid consistency and the obstruction from changing supports

The graph theorem does not by itself certify optimization on empirical contexts. The following approximation statement supplies a separate numerical guarantee. Partition X into cubes Q_i of side $h = 1/q$, with centers x_i and weights $w_i = q^{-d}$. Let $\lambda_h = \sum_i w_i\delta_{x_i}$. For a probability vector $p = (p_i)$, set $\nu_h(p) = \sum_i p_i\delta_{x_i}$ and

$$J^h_{\alpha,\widehat{c}}(p) = \sum_i \widehat{c}_i p_i + \frac{1-\alpha}{2}W_2^2(\lambda_h, \nu_h(p)) + \alpha\tau\sum_i p_i\log(p_i/w_i). \tag{94}$$

This is a convex problem. Its entropy acts on the destination row, not on the transport coupling.

Uniform grid consistency. Suppose $\mathrm{Lip}(c) \le L_y$, $\max_i |\widehat{c}_i - c(x_i)| \le e_c$, and p_h has objective gap at most e_{solve} in (94). Then, for all $\alpha \in [0,1]$,

$$W_1(\nu_h(p_h), \nu_{\alpha,c}) \le Dh + \left[\frac{(3DL_y + 5D^2)h + 2e_c + e_{\mathrm{solve}}}{m_{d,\tau}}\right]^{1/3}. \tag{95}$$

For exact samples of $u \in \mathcal{U}_{L_u}$, pushing the row to $j_u(x_i)$ multiplies this bound by at most $\sqrt{1+L_u^2}$.

Proof. Lift p to the histogram $L_h p = \sum_i (p_i/w_i)\mathbf{1}_{Q_i}\lambda$. Its entropy equals the discrete entropy, and $W_1(L_h p, \nu_h(p)) \le Dh$. Coupling source and destination points to their cell centers shows that the transport terms differ by at most $2D^2h$; the cost terms differ by at most $DL_y h + e_c$. Thus the continuum and discrete objectives differ by at most $C_1 h + e_c$, where $C_1 = DL_y + 2D^2$.

Take as a discrete competitor the cell masses of $\nu_{\alpha,c}$. Its histogram is $S_q \nu_{\alpha,c}$, whose continuum objective is at most the minimum plus $C_0 h$, with $C_0 = DL_y + D^2$, by (88) and (89). When $\alpha = 0$ no entropy comparison is needed. Comparing the discrete optimizer to this competitor gives a continuum objective gap for $L_h p_h$ of at most

$$(C_0 + 2C_1)h + 2e_c + e_{\mathrm{solve}}.$$

Apply (85) and add the lifting distance Dh. This proves (95), for every approximate minimizer, including nonunique discrete minimizers at $\alpha = 0$. □

Why non-collapse is insufficient. Take the ambient interval $\mathsf{E} = \mathsf{X} = [0,2]$, $\pi_{\mathsf{X}} = \mathrm{id}$, the fixed cost $c(w) = -3w$, and

$$\mu_\varepsilon = \frac{1-\varepsilon}{2}(\delta_0 + \delta_1) + \varepsilon\delta_2, \qquad 0 \le \varepsilon \le \tfrac{1}{4}.$$

These measures remain a positive distance from every Dirac measure and $W_1(\mu_\varepsilon, \mu_0) = 3\varepsilon/2$. At $\alpha = 0$, the source–destination objective is $-3y + |x-y|^2/2$, strictly decreasing in the available destination $y \in [0,2]$. Its unique minimizing row is therefore δ_2 for $\varepsilon > 0$, but δ_1 for $\varepsilon = 0$. Thus the input distance tends to zero while the distance between the minimizing reads remains one. The theorem therefore cannot extend to arbitrary non-collapsing contexts. The grid bound concerns a specified approximation of the continuum problem.