UNIVERSAL APPROXIMATION OF MEASURE-TO-MEASURE OPERATORS BY PUSHFORWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many learning tasks map an input distribution to an output distribution. A natural way to model such an operator is to transform each input sample using a continuous function that may depend on the entire input distribution, and then take the distribution of the transformed samples. This defines a measure-dependent pushforward model and includes measure-theoretic formulations of transformers. We ask when such models can approximate arbitrary continuous operators between spaces of probability measures. We first show that universal approximation fails when atomic inputs are allowed: some continuous measure-to-measure operators that split or redistribute atomic mass cannot be approximated arbitrarily well by deterministic pushforward models. We then introduce the uniform level set condition, which requires a continuous measure-dependent scalarization whose shrinking level set neighborhoods carry uniformly vanishing mass over the input family. This condition is satisfied, in particular, by compact families of absolutely continuous measures. On every compact family satisfying this condition, we prove that any continuous measure-to-measure operator with outputs of finite p-th moment can be uniformly approximated, in the p-Wasserstein distance, by continuous measure-dependent pushforwards. Combining our theorem with existing approximation results for measure-dependent in-context maps yields universal approximation by measure-theoretic transformers. We also extend the framework to continuously-varying source measures, yielding a corresponding universality result for a class of pushforward models that are closely aligned with cross-attention architectures.

1 INTRODUCTION

Many scientific and learning problems involving distributions can naturally be formulated as maps between spaces of probability measures. Given an input distribution, one would like to produce another distribution that depends continuously on the input. Examples include the prior-to-posterior map in Bayesian inference (Sprungk, 2020; Stuart, 2010), proximal maps on Wasserstein spaces (Ambrosio et al., 2008), and solution maps of McKean–Vlasov dynamics (Kolokoltsov, 2010).

A particularly important class of measure-to-measure maps is given by *pushforward models*. In such models, each sample from the input distribution is transformed by a map that may itself depend on the entire input distribution, and the collection of transformed samples determines the output distribution. This structure appears naturally in measure-theoretic formulations of transformers and attention mechanisms (Castin et al., 2025; Furuya et al., 2025b; Geshkovski et al., 2025; Rigollet, 2026; Sander et al., 2022; Vuckovic et al., 2020). Closely related pushforward constructions also arise in generative modeling, including flow matching (Lipman et al., 2023) and conditional normalizing flows (Winkler et al., 2019). This leads to the following fundamental approximation question:

To what extent can an arbitrary continuous operator between spaces of probability measures be approximated by continuous pushforward models whose transformation is allowed to depend on the input measure?

In this paper, we first show that universal approximation by continuous pushforward models fails in general when atomic input measures are allowed. We provide counterexamples induced by continuous

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.

measure-to-measure operators that split or redistribute atomic mass, which cannot in general be approximated by deterministic pushforward models.

We then identify a sufficient geometric-measure-theoretic condition on a compact family of input measures, which we call the *uniform level set condition*. Roughly speaking, this condition requires the existence of a continuous measure-dependent scalarization whose shrinking level set neighborhoods carry uniformly vanishing mass across the entire family of input measures. Under this condition, we prove that every continuous measure-to-measure operator with finite p-th moment outputs can be uniformly approximated in the p-Wasserstein distance by continuous pushforward models.

The uniform level set condition is substantially more flexible than absolute continuity with respect to the Lebesgue measure. In particular, it is automatically satisfied by compact families of non-atomic measures in one dimension, as well as by compact families dominated by a common non-atomic reference measure in higher dimensions. It also applies to certain families of singular measures supported on continuously-varying lower-dimensional structures, such as normalized arclength measures on a compact family of injective curves. At the same time, it necessarily excludes atomic measures, which is consistent with the obstruction described previously.

Our results have direct implications for the expressive power of transformers. Measure-theoretic transformers act on an input distribution through a measure-dependent transformation of its tokens and therefore induce a pushforward map at the level of probability measures. Combining our main approximation theorem with existing universality results for in-context maps, we establish universal approximation of continuous measure-to-measure operators by measure-theoretic transformers on compact families satisfying the uniform level set condition. Last, we extend the pushforward approximation framework to continuously-varying source measures, which relates to cross-attention.

Related work. The universal approximation properties of transformers have been extensively studied in the classical sequence-to-sequence setting (Vaswani et al., 2017). Early results established universality for transformers and their sparse variants (Yun et al., 2020a;b), followed by extensions concerning constrained architectures, prompting, approximation rates, and more refined descriptions of the expressive power of attention (Calvello et al., 2025; Chen et al., 2025; Cheng et al., 2026; Havrilla & Liao, 2024; Jiang & Li, 2024; Jiao et al., 2026; Kratsios et al., 2022; Liu et al., 2026; Petrov et al., 2024; Shen et al., 2026; Takakura & Suzuki, 2023; Wang & E, 2024).

A complementary line of work studies transformers in a measure-theoretic formulation, where a collection of tokens is represented by its empirical measure and attention is extended to probability measures. This viewpoint has led to mathematical descriptions of attention and transformer dynamics (Bach et al., 2025; 2026; Burger et al., 2025; Castin et al., 2025; Geshkovski et al., 2025; Rigollet, 2026; Sander et al., 2022; Vuckovic et al., 2020), as well as approximation results for functions whose inputs involve probability measures (Biswal et al., 2025; Cole et al., 2026; Fraiman, 2026; Furuya et al., 2025b; 2026; Geshkovski et al., 2026) and characterizations of the smoothness of attention layers acting on probability measures (Castin et al., 2024; Kim et al., 2021; Vuckovic et al., 2021).

More closely related to the present work are approximation and exact representation results in which both the input and output are probability measures. In work of Furuya et al. (2025a), transformer-induced measure-to-measure maps are studied through the lens of support-preserving transformations. Related exact transport map representations of continuous measure-to-measure transformations are investigated by Lavenant & Savaré (2026). These works naturally lead to structural conditions under which the output does not split the mass carried by individual input particles. Such conditions are particularly well suited to empirical measures, since an empirical input is then transformed into another empirical measure with compatible particle structure.

The question studied in the present paper is different. We ask under what conditions on a family of input measures can an arbitrary continuous measure-to-measure operator be approximated by deterministic pushforward models *without imposing a support-preserving or non-splitting condition on the target map*. This distinction is *essential*: natural maps such as Bayesian updating or joint measure conditioning may continuously change the weights of atoms and hence fall outside exact deterministic pushforward representations. Our results characterize a regime in which this obstruction disappears at the level of uniform approximation and thereby connect general continuous measure-to-measure operators with transformer-induced pushforward models.

Contributions. The main contributions of this work are summarized as follows.

(C1) **An obstruction caused by atomic measures.** We show that continuous measure-dependent pushforward models are not universal on general compact subsets of probability measures (Proposition 2.2). In particular, deterministic pushforwards cannot split atomic mass or freely modify atomic weights. This yields explicit continuous measure-to-measure operators that remain separated from every deterministic pushforward model by a strictly positive 1-Wasserstein distance (Examples 2.3 and 2.4).

(C2) **Universal approximation under a uniform level set condition.** We introduce the uniform level set condition for compact families of input measures (Assumption 3.1). We prove that, under this condition, every continuous measure-to-measure operator with outputs of finite p-th moment can be uniformly approximated in the p-Wasserstein distance by continuous measure-dependent pushforward maps (Theorem 3.6). The condition necessarily excludes atomic measures, but is satisfied by compact families of non-atomic measures in one dimension, compact families dominated by a common non-atomic reference measure, and certain families of singular measures supported on continuously-varying lower-dimensional structures (Examples 3.3, 3.4, and 3.5).

(C3) **Consequences for transformer architectures.** By combining our main approximation theorem with existing universality results for measure-theoretic in-context maps, we establish universal approximation of continuous measure-to-measure operators by measure-theoretic transformers on compact families satisfying the uniform level set condition (Theorem 4.4). We further extend the pushforward approximation framework to continuously-varying source measures, which has applications to cross-attention (Corollary 4.6).

Notation. Let $|\cdot|$ denote the Euclidean norm. We denote by $\mathcal{P}(\mathbb{R}^d)$ the set of all Borel probability measures on $\mathbb{R}^d$ and, for $1 \le p < \infty$, by $\mathcal{P}_p(\mathbb{R}^d) \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^d) \colon \int |x|^p \, \mu(dx) < \infty\}$ the set of probability measures with finite p-th moment. The subset of *non-atomic* or *atomless* probability measures is

$$\mathcal{P}_{\mathrm{na}}(\mathbb{R}^d) \coloneqq \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \colon \mu(\{x\}) = 0 \ \text{ for every } x \in \mathbb{R}^d\right\}.$$

We equip $\mathcal{P}(\mathbb{R}^d)$ with the topology of weak convergence of probability measures. Whenever $\mathcal{P}_p(\mathbb{R}^d)$ is considered, it is equipped with the p-Wasserstein metric W_p given by

$$\mathsf{W}_p(\mu,\nu) \coloneqq \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, \pi(dx,dy)\right)^{1/p},$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν. For a measurable map $T \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, the pushforward measure $T_{\#}\mu \in \mathcal{P}(\mathbb{R}^{d'})$ is defined by $T_{\#}\mu(A) \coloneqq \mu(T^{-1}(A))$ for all $A \subset \mathbb{R}^{d'}$ Borel. For topological spaces X and Y, we denote by $\mathcal{C}(X,Y)$ the space of continuous maps from X to Y; the topology on spaces of probability measures is always understood as above.

2 The pushforward approximation problem

This section formulates the approximation problem studied in this work, illustrates it with two examples of continuous measure-to-measure operators, and identifies atomic obstructions.

2.1 The measure-to-measure approximation problem

In what follows, let $d \in \mathbb{N}$, $d' \in \mathbb{N}$. We investigate whether for every compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$ with respect to W_p, every $F \in \mathcal{C}(K, \mathcal{P}_p(\mathbb{R}^{d'}))$, and every $\varepsilon > 0$, there exists a continuous map $G \in \mathcal{C}(K \times \mathbb{R}^d, \mathbb{R}^{d'})$ such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(G(\mu, \cdot)_{\#}\mu, F(\mu)\big) < \varepsilon.$$

In other words, we ask whether continuous measure-to-measure operators can be uniformly approximated by continuous measure-dependent pushforward models. This approximation problem arises naturally in several settings. We illustrate it with an example.

***Example* 2.1** (Prior-to-posterior map in Bayesian inference)**.** In Bayesian inverse problems, updating a prior probability measure using observed data modeled by a likelihood naturally defines a map from prior to posterior measures (Nelsen & Yang, 2026; Stuart, 2010). Let $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ denote the negative log-likelihood associated with the observed data. For a prior measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, the corresponding posterior measure is defined by

$$F_\Phi(\mu)(du) \coloneqq \frac{\exp\bigl(-\Phi(v)\bigr)}{Z_\mu}\,\mu(dv), \quad \text{where} \quad Z_\mu \coloneqq \int_{\mathbb{R}^d} \exp\bigl(-\Phi(u)\bigr)\,\mu(du).$$

Thus, Bayesian inference defines the prior-to-posterior map $F_\Phi\colon \mathcal{P}_p(\mathbb{R}^d) \to \mathcal{P}_p(\mathbb{R}^d)$. If Φ is continuous and bounded from below, then F_Φ is well-defined and W_p-continuous (Sprungk, 2020, Lemma 16). Notice that F_Φ generally changes the weights of the input measure while preserving its null sets. In particular, for an atomic prior $\mu = \sum_{i=1}^N a_i \delta_{x_i}$, the posterior is

$$F_\Phi(\mu) = \sum_{i=1}^N \frac{a_i e^{-\Phi(x_i)}}{\sum_{j=1}^N a_j e^{-\Phi(x_j)}}\delta_{x_i}.$$

Hence, although the support is unchanged, the weights are in general different. Consequently, $F_\Phi(\mu)$ cannot in general be represented as a deterministic pushforward of the form $G(\mu, \cdot\,)_\#\mu$. This provides a natural example of a continuous measure-to-measure operator that lies beyond exact deterministic pushforward representations on atomic and nonatomic measures, while still falling within the approximation problem considered in this work.

See Appendix A.1 for a further example based on proximal maps between Wasserstein spaces.

2.2 Obstructions caused by atomic measures

In general, the answer to the approximation question posed in Section 2.1 is negative when atomic input measures are allowed. More precisely, the following holds.

Proposition 2.2. *There exist a compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$ in W_p, a continuous map $F \in \mathcal{C}(K, \mathcal{P}_p(\mathbb{R}^{d'}))$, and $\varepsilon > 0$ such that there is no continuous map $G\colon K \times \mathbb{R}^d \to \mathbb{R}^{d'}$ satisfying*

$$\sup_{\mu\in K} \mathsf{W}_p\bigl(G(\mu, \cdot\,)_\#\mu, F(\mu)\bigr) < \varepsilon.$$

The obstruction already appears in the simplest possible examples.

***Example* 2.3** (Different support cardinalities)**.** Let $K \coloneqq \{\delta_0\} \subset \mathcal{P}(\mathbb{R})$ and define

$$F(\delta_0) \coloneqq \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1.$$

Since K is a singleton, $F\colon K \to \mathcal{P}_1(\mathbb{R})$ is continuous. For any continuous $G\colon K \times \mathbb{R} \to \mathbb{R}$, we have $G(\delta_0, \cdot\,)_\#\delta_0 = \delta_{G(\delta_0,0)}$. Therefore,

$$\mathsf{W}_1\bigl(G(\delta_0, \cdot\,)_\#\delta_0, F(\delta_0)\bigr) = \frac{1}{2}\,|G(\delta_0, 0) + 1| + \frac{1}{2}\,|G(\delta_0, 0) - 1| \geq 1$$

by the triangle inequality. Hence

$$\inf_{G\in\mathcal{C}(K\times\mathbb{R},\mathbb{R})} \sup_{\mu\in K} \mathsf{W}_1\bigl(G(\mu, \cdot\,)_\#\mu, F(\mu)\bigr) \geq 1.$$

Thus, a deterministic pushforward cannot approximate a map that splits a single input atom into two.

The preceding example changes the number of atoms. However, this is not the essential issue: an obstruction remains even when the input and output measures have the same support cardinality.

***Example* 2.4** (Same support cardinality, different weights)**.** Let $K \coloneqq \bigl\{\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1\bigr\}$ and define

$$\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1 \quad \text{and} \quad F(\nu) \coloneqq \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_1.$$

Again, $F\colon K \to \mathcal{P}_1(\mathbb{R})$ is continuous. For any continuous map $G\colon K \times \mathbb{R} \to \mathbb{R}$, it holds that

$$G(\nu, \cdot\,)_\#\nu = \frac{1}{2}\delta_{G(\nu,-1)} + \frac{1}{2}\delta_{G(\nu,1)}.$$

Proposition A.2 delivers the lower bound

$$\mathsf{W}_1\big(G(\nu,\,\cdot\,)_{\#}\nu, F(\nu)\big) \geq \frac{1}{3}.$$

Consequently,

$$\inf_{G\in\mathcal{C}(K\times\mathbb{R},\mathbb{R})} \sup_{\mu\in K} \mathsf{W}_1\big(G(\mu,\,\cdot\,)_{\#}\mu, F(\mu)\big) \geq \frac{1}{3}.$$

Thus, even when the input and output distributions have the same number of atoms, a deterministic pushforward cannot in general modify their weights.

3 Universal approximation by continuous pushforward models

This section establishes our main universal approximation result for continuous measure-dependent pushforward models.

3.1 Uniform level set condition

We introduce a sufficient condition on the family K of input measures under which universal approximation holds.

Assumption 3.1. *Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be W_p-compact. There exists $\varphi \in \mathcal{C}(K \times \mathbb{R}^d, \mathbb{R})$ such that*

$$\lim_{r\downarrow 0} \sup_{\mu\in K} \sup_{t\in\mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(\mu, x) - t| \leq r\}\big) = 0. \tag{3.1}$$

We call (3.1) the *uniform level set condition*. Intuitively, Assumption 3.1 requires the existence of a continuous scalarization φ whose level sets carry uniformly vanishing mass over the family K. In particular, the condition is satisfied when all measures in K are absolutely continuous with respect to Lebesgue measure: one may take, for example, any linear projection $\varphi(\mu, x) \coloneqq x \cdot v$ in nonzero direction v independently of μ. More generally, as shown below, absolute continuity with respect to Lebesgue measure is far from necessary. The importance of Assumption 3.1 is that it enables us to approximate the one-dimensional uniform distribution $\mathrm{Unif}[0, 1]$ by a continuous pushforward model. This is a key ingredient in the proof of Theorem 3.6, our main approximation result.

***Remark* 3.2** (Uniform level set condition implies non-atomic family)**.** If K is such that Assumption 3.1 holds, then $K \subset \mathcal{P}_{\mathrm{na}}(\mathbb{R}^d)$. Indeed, fix $\mu \in K$ and $x_0 \in \mathbb{R}^d$, and set $t = \varphi(\mu, x_0)$. Then

$$\{x_0\} \subset \{x \in \mathbb{R}^d \colon |\varphi(\mu, x) - \varphi(\mu, x_0)| \leq r\}$$

for any $r \geq 0$. Hence

$$\mu\big(\{x_0\}\big) \leq \sup_{\nu\in K} \sup_{t\in\mathbb{R}} \nu\big(\{x \in \mathbb{R}^d \colon |\varphi(\nu, x) - t| \leq r\}\big) \to 0$$

as $r \downarrow 0$ by (3.1). Since x_0 was arbitrary, μ is non-atomic as asserted.

To better understand Assumption 3.1, we present several examples of compact families that satisfy it.

***Example* 3.3** (One-dimensional case)**.** Let $K \subset \mathcal{P}_{\mathrm{na}}(\mathbb{R})$ be compact with respect to W_p. Then the identity function $\varphi \colon \mathbb{R} \to \mathbb{R}$ given by $\varphi(x) \coloneqq x$ satisfies

$$\lim_{r\downarrow 0} \sup_{\mu\in K} \sup_{t\in\mathbb{R}} \mu\big(\{x \in \mathbb{R} \colon |\varphi(x) - t| \leq r\}\big) = 0.$$

In particular, every compact family of non-atomic probability measures on $\mathbb{R}$ satisfies the uniform level set condition (3.1). See Appendix B.1 for the proof.

Remark 3.2 shows that if a compact subset $K \subset \mathcal{P}_p(\mathbb{R}^d)$ satisfies Assumption 3.1, then every element of K is atomless. Example 3.3 establishes the converse when $d = 1$; this gives a complete characterization of subsets satisfying Assumption 3.1 in the one-dimensional setting.

***Example* 3.4** (Families dominated by a common non-atomic measure)**.** Let $\rho \in \mathcal{P}_{\mathrm{na}}(\mathbb{R}^d)$ and let

$$K \subset \mathcal{P}_{\mathrm{ac},p}(\rho) \coloneqq \{\mu \in \mathcal{P}_p(\mathbb{R}^d) \colon \mu \ll \rho\}$$

be compact with respect to W_p. Then there exists a continuous function $\varphi \colon \mathbb{R}^d \to \mathbb{R}$ such that

$$\lim_{r\downarrow 0} \sup_{\mu\in K} \sup_{t\in\mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(x) - t| \leq r\}\big) = 0.$$

See Appendix B.2 for the proof. By taking ρ to be the Lebesgue measure, Example 3.4 implies that Assumption 3.1 is satisfied by any W_p-compact family of densities. Importantly, the only requirement in Example 3.4 is that ρ be non-atomic. In particular, Example 3.4 applies to compact families of probability measures that are supported on a low-dimensional manifold $\mathcal{M}$ and admit a density with respect to some fixed atomless reference measure on $\mathcal{M}$, e.g., a uniform measure on $\mathcal{M}$. Such families of measures arise naturally under the manifold hypothesis in machine learning.

***Example* 3.5** (Measures supported on a compact family of curves)**.** Let $Q \subset \mathbb{R}^d$ be compact and let $K' \subset \mathcal{C}^1([0,1], Q)$ be compact. Assume that there exists $c > 0$ such that

$$|\dot{\gamma}(s)| \geq c \quad \text{for} \quad \gamma \in K' \quad \text{and} \quad s \in [0,1],$$

and that every $\gamma \in K'$ is injective. For each $\gamma \in K'$, define the normalized arclength measure

$$\sigma_\gamma \coloneqq \left(\int_0^1 |\dot{\gamma}(s)|\, ds \right)^{-1} \gamma_{\#}\big(|\dot{\gamma}(s)|\, ds\big).$$

Then there exists a continuous function $\widetilde{\varphi}\colon K' \times Q \to [0,1]$ such that $\widetilde{\varphi}(\gamma, \gamma(s)) = s$ for $\gamma \in K'$ and $s \in [0,1]$, and

$$\lim_{r \downarrow 0} \sup_{\gamma \in K'} \sup_{t \in \mathbb{R}} \sigma_\gamma\big(\{x \in Q\colon |\widetilde{\varphi}(\gamma, x) - t| \leq r\}\big) = 0.$$

Assume, in addition, that the curves in K' are uniquely determined by their images, that is,

$$\gamma_1([0,1]) = \gamma_2([0,1]) \quad \text{implies} \quad \gamma_1 = \gamma_2.$$

Then the family $K \coloneqq \{\sigma_\gamma\colon \gamma \in K'\} \subset \mathcal{P}_p(Q)$ satisfies the uniform level set condition (3.1).

See Appendix B.3 for the proof. Example 3.5 illustrates that the uniform level set condition can accommodate singular measures whose supports themselves vary continuously. Such families naturally arise when data distributions are concentrated near evolving or parameter-dependent low-dimensional structures, rather than on a single fixed reference manifold.

3.2 Main result

We are now ready to state our main universal approximation theorem. The key idea is to use the uniform level set condition to approximately push forward each input measure to a common uniform distribution on $[0,1]$.

Theorem 3.6. *Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ satisfy Assumption 3.1. Then for any $F \in \mathcal{C}(K, \mathcal{P}_p(\mathbb{R}^{d'}))$ and any $\varepsilon > 0$, there exists a continuous map $G\colon K \times \mathbb{R}^d \to \mathbb{R}^{d'}$ such that*

$$\sup_{\mu \in K} \mathsf{W}_p\big(G(\mu, \cdot)_{\#}\mu, F(\mu)\big) < \varepsilon.$$

Sketch of proof. The full proof is given in Appendix B.4. Here we describe the construction together with the estimates used below. For $R > 0$, let

$$F_R(\mu) \coloneqq (\Pi_R)_{\#} F(\mu),$$

where Π_R is the metric projection onto the closed ball $\overline{B_R(0)}$ of radius R. Define

$$E_{\mathrm{tail},p}(R) \coloneqq \sup_{\nu \in F(K)} \left(\int_{\{|y| > R\}} |y|^p\, \nu(dy) \right)^{1/p}.$$

By Lemma B.2, $E_{\mathrm{tail},p}(R) \to 0$ as $R \to \infty$ and

$$\sup_{\mu \in K} \mathsf{W}_p\big(F(\mu), F_R(\mu)\big) \leq E_{\mathrm{tail},p}(R).$$

Next, for $h > 0$, choose an h-net $\{y_1, \ldots, y_N\} \subset \overline{B_R(0)}$ and a continuous partition of unity subordinate to the corresponding cover. The covering number may be chosen so that

$$N = N(R,h) \leq C_{d'} \left(1 + \frac{R}{h}\right)^{d'},$$

where $C_{d'} > 0$ depends only on d'. This yields continuous weights $\alpha_i \colon K \to [0,1]$ such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(F_R(\mu), F_{R,h}(\mu)\big) \le h, \quad \text{where} \quad F_{R,h}(\mu) \coloneqq \sum_{i=1}^{N} \alpha_i(\mu)\delta_{y_i}.$$

Next, for $r > 0$, set

$$\omega_K(r) \coloneqq \sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(\mu, x) - t| \le r\}\big).$$

By Assumption 3.1, it holds that $\omega_K(r) \to 0$ as $r \downarrow 0$. The construction in Lemma B.3 delivers the existence of $H_r \in \mathcal{C}(K \times \mathbb{R}^d, [0,1])$ such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(H_r(\mu, \cdot)_{\#}\mu, \mathrm{Unif}[0,1]\big) \le \omega_K(r).$$

Writing $s_i(\mu) \coloneqq \sum_{j=1}^{i} \alpha_j(\mu)$ and $s_0(\mu) \coloneqq 0$, the discontinuous interval map

$$u \mapsto Q_0(\mu, u) \coloneqq \sum_{i=1}^{N} \mathbf{1}_{(s_{i-1}(\mu), s_i(\mu)]}(u)\, y_i$$

pushes $\mathrm{Unif}[0,1]$ exactly onto $F_{R,h}(\mu)$. We smooth its transition points on intervals of width $\delta > 0$ to obtain a bounded continuous map Q_δ satisfying

$$\mathrm{Lip}\big(Q_\delta(\mu, \cdot)\big) \le C_R \delta^{-1},$$

while the W_p contribution of the transition regions is bounded by $2R\big(2N(R,h)\delta\big)^{1/p}$. Defining

$$G(\mu, x) \coloneqq Q_\delta\big(H_r(\mu, x), \mu\big),$$

the triangle inequality gives the asserted approximation after successively choosing R, h, δ, and r appropriately as functions of ε. □

For a fixed $\mu \in \mathcal{P}_{\mathrm{na}}(\mathbb{R}^d)$, it is known that the set of deterministic Monge couplings $(\mathrm{id}, g)_{\#}\mu$ induced by continuous transport maps $g \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ with source measure μ is dense in the set of all joint distributions in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'})$ with first marginal μ (Beiglböck & Lacker, 2018, Prop. 2.2). Theorem 3.6 can then be interpreted as the operator analog of this "diagonal approximation" result in which μ is no longer fixed, but is instead allowed to vary uniformly over the compact set K.

***Remark* 3.7** (Quantitative error bound)**.** The preceding proof actually yields the upper bound

$$\sup_{\mu \in K} \mathsf{W}_p\big(G(\mu, \cdot)_{\#}\mu, F(\mu)\big) \le E_{\mathrm{tail},p}(R) + h + \frac{C_R}{\delta}\omega_K(r) + 2R\left[2C_{d'}\left(1 + \frac{R}{h}\right)^{d'} \delta\right]^{1/p}.$$

Moreover, the construction gives the uniform bound $\|G\|_{\mathcal{C}(K \times \mathbb{R}^d)} \le R$.

4 Applications to transformer architectures

We now apply our main approximation theorem to measure-theoretic transformer architectures.

4.1 Measure-theoretic transformers

We first recall the measure-theoretic formulation of transformers.

***Definition* 4.1** (Measure-theoretic in-context attention maps)**.** For $H \in \mathbb{N}$, a *measure-theoretic in-context multi-head attention map* $\Gamma_\theta \colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ is defined by

$$\Gamma_\theta(\mu, x) \coloneqq x + \sum_{h=1}^{H} W^h \int \frac{\exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x, K^h z\rangle\big)\, \mu(dz)}\, V^h y\, \mu(dy), \tag{4.1}$$

where $\theta \coloneqq (W^h, Q^h, K^h, V^h)_{h=1}^{H}$, $(\mu, x) \in \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d$, and W^h, Q^h, K^h, and V^h are the head, query, key, and value matrices, respectively. See Remark 4.5 for a note about finiteness of (4.1).

In (4.1), Γ_θ acts on a distinguished token x while depending on the input probability measure μ, and can therefore be viewed as an in-context map. The standard discrete attention mechanism is recovered when μ is an empirical measure $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ associated with tokens $x_1, \dots, x_n$. In this case, the integral in the definition of Γ_θ reduces to the usual finite sum over the tokens. The measure-theoretic formulation removes the need to fix the number of tokens in advance and provides a unified description of attention for arbitrary probability measures, including non-atomic measures. For further details on measure-theoretic formulations of attention and in-context maps, see, e.g., the work of Castin et al. (2024; 2025); Furuya et al. (2025a;b); Geshkovski et al. (2023; 2025; 2026).

***Definition* 4.2** (Composition of measure-theoretic in-context maps)**.** Let

$$\Gamma_1 \colon \mathcal{P}(\mathbb{R}^{d_0}) \times \mathbb{R}^{d_0} \to \mathbb{R}^{d_1} \quad \text{and} \quad \Gamma_2 \colon \mathcal{P}(\mathbb{R}^{d_1}) \times \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}.$$

Their *composition* is defined by

$$\big(\Gamma_2 \diamond \Gamma_1\big)(\mu, x) \coloneqq \Gamma_2\big(\Gamma_1(\mu, \cdot)_\# \mu, \Gamma_1(\mu, x)\big).$$

***Definition* 4.3** (Measure-theoretic transformer)**.** Let $\Gamma_{\theta_1}, \dots, \Gamma_{\theta_L}$ be measure-theoretic in-context attention maps as in (4.1) and $\phi_1, \dots, \phi_L$ be context-free multilayer perceptron (MLP) maps. The associated deep transformer in-context map $\mathsf{T} \colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^{d'}$ is defined by

$$\mathsf{T} \coloneqq \phi_L \diamond \Gamma_{\theta_L} \diamond \cdots \diamond \phi_1 \diamond \Gamma_{\theta_1}. \tag{4.2}$$

The corresponding *measure-theoretic transformer* is the measure-to-measure operator defined by

$$\mathcal{P}(\mathbb{R}^d) \ni \mu \mapsto \mathsf{T}(\mu, \cdot)_\# \mu \in \mathcal{P}(\mathbb{R}^{d'}).$$

Furuya et al. (2025b) show that transformer in-context maps of the form (4.2) are universal approximators for general continuous in-context mappings defined on compact token domains; the crucial difficulty is that the present paper works on the whole of $\mathbb{R}^d$, not a compact domain. Nevertheless, combining modifications of this universality result (Proposition A.6) with Theorem 3.6 yields the following universal approximation result for measure-to-measure operators.

Theorem 4.4. *Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be a W_p-compact set satisfying Assumption 3.1. Then for any $F \in \mathcal{C}(K, \mathcal{P}_p(\mathbb{R}^{d'}))$ and any $\varepsilon > 0$, there exists a deep transformer in-context map $\mathsf{T} \colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^{d'}$ of the form* (4.2) *such that*

$$\sup_{\mu \in K} \mathsf{W}_p\big(\mathsf{T}(\mu, \cdot)_\# \mu, F(\mu)\big) < \varepsilon.$$

See Appendix C.1 for the proof. Theorem 4.4 proves that the measure-to-measure operators defined by transformers are universal approximators of continuous measure-to-measure operators defined on a compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$, provided that K satisfies the uniform level set condition (3.1). In many practical scenarios, the transformer T is not applied to $\mu \in K$ directly—which is atomless by Remark 3.2—but instead to an atomic measure μ_N supported on N samples from μ. The error analysis in this case can still be handled by first controlling the W_p error between $\mathsf{T}(\mu_N, \cdot)_\# \mu_N$ and $\mathsf{T}(\mu, \cdot)_\# \mu$ uniformly over K and then applying Theorem 4.4 to bound the W_p error between $\mathsf{T}(\mu, \cdot)_\# \mu$ and $F(\mu)$. This approach requires Assumption 3.1 and discretizes $\mu \in K$ with samples. A different approach removes Assumption 3.1 on the W_p-compact set K altogether by convolving the (now possibly atomic) input measure $\mu \in K$ with a standard isotropic Gaussian distribution that has a small variance σ^2 to obtain a distribution μ_σ; see Corollary A.5 in Appendix A.2 for details. Given samples $X_1, \dots, X_N$ from μ, one can combine the two approaches by approximating μ_σ by an empirical measure of the form $\widehat{\mu}_\sigma = \frac{1}{N}\sum_{i=1}^N \delta_{X_i + \sigma Z_i}$, where $Z_1, \dots, Z_N$ are independent Gaussian samples with mean zero and identity covariance matrix.

***Remark* 4.5** (Domain of measure-theoretic attention)**.** To be precise, the measure-theoretic attention map Γ_θ in (4.1) is only well-defined for inputs (μ, x) such that

$$\int \exp\big(\langle Q^h x, K^h y\rangle\big)\, \mu(dy) < \infty \quad \text{and} \quad \int \exp\big(\langle Q^h x, K^h y\rangle\big) V^h y\, \mu(dy) < \infty$$

for all $h \in \{1, \dots, H\}$. These conditions always hold when μ has compact support. Although Theorem 4.4 applies to input measures $\mu \in K$ with possibly full support on $\mathbb{R}^d$, the specific construction only requires evaluating self-attention layers at compactly supported measures. This is because the transformer constructed in the proof of Theorem 4.4 utilizes an explicit initial MLP layer which truncates the input measure to a common compact set.

4.2 CONTINUOUSLY-VARYING SOURCE MEASURES AND CONNECTION TO CROSS-ATTENTION

The preceding result is an approximation theorem for measure-theoretic transformers based on self-attention, in which the source measure of the transport is the input measure itself. The next corollary generalizes the main pushforward approximation result, Theorem 3.6 from Section 3.2, to the setting in which the source measure of the pushforward is allowed to depend continuously on the input measure, but need not equal the input measure. This is particularly relevant for analyzing cross-attention architectures, where the measure carrying the query tokens may differ from the measure representing the context and may even live on a different ambient space.

Corollary 4.6. *Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be W_p-compact. Let $\eta\colon K \to \mathcal{P}_p(\mathbb{R}^m)$ be W_p-continuous. Assume that there exists $\varphi \in \mathcal{C}(K \times \mathbb{R}^m, \mathbb{R})$ such that*

$$\lim_{r\downarrow 0} \sup_{\mu\in K} \sup_{t\in\mathbb{R}} \eta(\mu)\big(\{x \in \mathbb{R}^m\colon |\varphi(\mu, x) - t| \le r\}\big) = 0. \tag{4.3}$$

Then for every $F \in \mathcal{C}(K, \mathcal{P}_p(\mathbb{R}^{d'}))$ and every $\varepsilon > 0$, there exists $G \in \mathcal{C}(K \times \mathbb{R}^m, \mathbb{R}^{d'})$ such that

$$\sup_{\mu\in K} \mathsf{W}_p\big(G(\mu, \cdot)_{\#}\eta(\mu), F(\mu)\big) < \varepsilon.$$

Typical source measures covered by Corollary 4.6 include a fixed non-atomic measure, $\eta(\mu) \equiv \rho$ (e.g., an isotropic Gaussian), and a smoothed input measure, $\eta(\mu) = \rho * \mu$, where ρ is absolutely continuous. The proof of Corollary 4.6 follows the same construction as that of Theorem 3.6 and may be found in Appendix C.2.

***Remark* 4.7** (Relation to cross-attention)**.** Cross-attention allows one collection of query tokens to attend to a distinct collection of context tokens and is a central component of encoder–decoder transformers (Vaswani et al., 2017) and related architectures such as the Perceiver (Jaegle et al., 2021). Corollary 4.6 is directly connected to architectures in which the source tokens attend to the context tokens through cross-attention and are otherwise processed token-wise, without self-attention among the source tokens. Combined with universality for measure-dependent in-context maps—see, e.g., Theorem 1 due to Furuya et al. (2025b) or Proposition A.6—it yields an analog of Theorem 4.4 for this particular cross-attention configuration. However, this argument does not by itself establish universality for more general architectures that combine both self-attention among source tokens with cross-attention among context tokens.

5 CONCLUSION

In this paper, we studied the approximation of continuous measure-to-measure operators by continuous *pushforward models*, a subset of mappings which are naturally motivated by transformers. We showed that atomic measures present an obstruction to universal approximation due to the inability of pushforward models to split or adjust the weights of atoms. To remove this obstruction, we identified a sufficient condition for universal approximation on a compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$, which we termed the uniform level set condition. Notably, the uniform level set condition is weaker than absolute continuity and allows for measures whose supports admit low-dimensional structure. Our main result also implies a uniform approximation theorem for measure-theoretic transformers.

This work is the first to study universal approximation properties of pushforward models under general assumptions on the target operator; however, several important questions still remain. First, while the uniform level set condition is sufficient for universal approximation, it may not be necessary. An interesting mathematical question is to identify the minimal assumptions required on a compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$ to enable universal approximation of continuous operators on K by pushforward models. Second, certain measure-to-measure operators of interest, such as the joint-to-conditional operator in Bayesian inference, may not be continuous (Tsimpos et al., 2026, Appendix E); approximation of such maps by pushforward models is thus beyond the scope of this paper. Last, an important problem is obtaining quantitative approximation results for measure-theoretic transformers, where the number of parameters is controlled by the smoothness of the target operator and the covering number of the domain of approximation. We leave these important directions to future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, formulating mathematical claims and refining them, providing critical ingredients for proving mathematical claims, and assist with translation. We also used generative AI tools for checking proof arguments, improving the organization, presentation, readability, and structure of the paper, language editing, brainstorming, and literature search assistance. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Moreover, we did not use generative AI tools to generate synthetic data sets, implement methods, clean and reformat datasets, or support qualitative and thematic data analysis; these tasks are not applicable to this theoretical work.

All mathematical statements and proofs suggested or modified with the assistance of generative AI were independently examined and verified by the authors. Relevant references and attribution were checked against the original sources, and the final mathematical arguments and wording were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and other artifacts produced with the aid of generative AI.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: In metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Eviatar Bach, Ricardo Baptista, Edoardo Calvello, Bohan Chen, and Andrew M Stuart. Learning enhanced ensemble filters. *Journal of Computational Physics*, art. 114550, 2025.

Eviatar Bach, Ricardo Baptista, Jochen Bröcker, Bohan Chen, and Andrew M Stuart. Learning probabilistic filters with strictly proper scoring rules. *preprint arXiv:2606.26497*, 2026.

Mathias Beiglböck and Daniel Lacker. Denseness of adapted processes among causal couplings. *preprint arXiv:1805.03185*, 2018.

Shiba Biswal, Karthik Elamvazhuthi, and Rishi Sonthalia. Universal approximation of mean-field models via transformers. In *International Conference on Machine Learning*, pp. 4455–4470, 2025.

Martin Burger, Samira Kabri, Yury Korolev, Tim Roith, and Lukas Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2298):20240233, 2025.

Edoardo Calvello, Nikola B Kovachki, Matthew E Levine, and Andrew M Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1–52, 2025.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *International Conference on Machine Learning*, pp. 5817–5840. PMLR, 2024.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *Foundations of Computational Mathematics*, 2025.

Yifang Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Fundamental limits of visual autoregressive transformers: Universal approximation abilities. In *International Conference on Machine Learning*, 2025.

Jingpu Cheng, Ting Lin, Zuowei Shen, and Qianxiao Li. A unified framework for establishing the universal approximation of transformer-type architectures. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129131–129163, 2026.

Frank Cole, Dixi Wang, Yineng Chen, Yulong Lu, and Rongjie Lai. In-context operator learning on the space of probability measures. *preprint arXiv:2601.09979*, 2026.

James Dugundji. An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1(3):353–367, 1951.

Demián Fraiman. On the expressive power of contextual relations in transformers. *preprint arXiv:2603.25860*, 2026.

Takashi Furuya, Maarten V de Hoop, and Matti Lassas. Transformers through the lens of support-preserving maps between measures. *preprint arXiv:2509.25611*, 2025a.

Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, 2025b.

Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. Approximation theory for Lipschitz continuous transformers. In *International Conference on Machine Learning*, 2026.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57026–57037, 2023.

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

Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *Foundations of Computational Mathematics*, pp. 1–50, 2026.

Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42162–42210, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68926–68955, 2024.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for transformer networks with application to regression. In *International Conference on Machine Learning*, 2026.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Vassili N Kolokoltsov. *Nonlinear Markov processes and kinetic equations*, volume 182. Cambridge University Press, 2010.

Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, 2022.

Hugo Lavenant and Giuseppe Savaré. Continuous transformations of probability measures and their transport representations. *preprint arXiv:2604.16653*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Hude Liu, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. Attention mechanism, max-affine partition, and universal approximation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 120443–120511, 2026.

Nicholas H Nelsen and Yunan Yang. Operator learning meets inverse problems: A probabilistic perspective. In Andreas Hauptmann, Bangti Jin, and Carola-Bibiane Schönlieb (eds.), *Handbook of Numerical Analysis*, volume 27. Elsevier, 2026.

Aleksandar Petrov, Philip Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. In *International Conference on Machine Learning*, pp. 40523–40550. PMLR, 2024.

Philippe Rigollet. The mean-field dynamics of transformers. In *International Congress of Mathematicians 2026*, pp. 389–404. SIAM, 2026.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530, 2022.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. In *International Conference on Learning Representations*, volume 2026, pp. 42067–42103, 2026.

Björn Sprungk. On the local Lipschitz stability of Bayesian inverse problems. *Inverse Problems*, 36 (5):055015, 2020.

Andrew M Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pp. 33416–33447. PMLR, 2023.

Panos Tsimpos, Edoardo Calvello, Ayoub Belhadji, and Nicholas H Nelsen. One operator for many densities: Amortized approximation of conditioning by neural operators. *preprint arXiv:2605.06873*, 2026.

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, 2017.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *preprint arXiv:2007.02876*, 2020.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. On the regularity of attention. *preprint arXiv:2102.05628*, 2021.

Mingze Wang and Weinan E. Understanding the expressive power and mechanisms of transformer for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25781–25856, 2024.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *preprint arXiv:1912.00042*, 2019.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020a.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. $O(n)$ connections are expressive enough: Universal approximability of sparse transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13783–13794, 2020b.

Supplementary Material for:
Universal Approximation of Measure-to-Measure Operators by Pushforwards

A AUXILIARY RESULTS

This appendix contains auxiliary results and additional supporting material.

A.1 ADDITIONAL MATERIAL FOR SECTION 2

We now provide another example of a continuous measure-to-measure operator; this one is based on proximal maps. Proximal maps and minimizing-movement schemes can be defined on general metric spaces; see, e.g., the work of Ambrosio et al. (2008). Here we consider such a construction on the 1-Wasserstein space.

***Example* A.1** (W_1-proximal map)**.** Let $\Omega \subset \mathbb{R}^d$ be compact with positive Lebesgue measure. Define the entropy functional

$$\mathrm{Ent}(\nu) \coloneqq \begin{cases} \displaystyle\int_\Omega \rho(x)\log\rho(x)\,dx, & \text{if density } \rho \coloneqq \tfrac{d\nu}{dx} \text{ exists},\\ +\infty, & \text{otherwise.}\end{cases}$$

For $\tau > 0$ and $\mu \in \mathcal{P}_1(\Omega)$, consider the variational problem

$$F_\tau(\mu) \coloneqq \operatorname*{argmin}_{\nu\in\mathcal{P}_1(\Omega)} \left\{ \mathrm{Ent}(\nu) + \frac{1}{2\tau}\mathsf{W}_1^2(\mu,\nu) \right\}.$$

Since Ω is compact, $\mathcal{P}_1(\Omega) = \mathcal{P}(\Omega)$ and this space is compact with respect to W_1. Moreover, Ent is lower semicontinuous and strictly convex, while

$$\nu \mapsto \mathsf{W}_1^2(\mu,\nu)$$

is convex. Hence, the above minimization problem admits a unique minimizer. Therefore, F_τ defines a measure-to-measure operator

$$\begin{aligned} F_\tau\colon \mathcal{P}_1(\Omega) &\to \mathcal{P}_1(\Omega),\\ \mu &\mapsto F_\tau(\mu). \end{aligned}$$

We next observe that F_τ is continuous with respect to W_1. Indeed, let $\mu_n \to \mu$ in W_1 and set

$$\nu_n \coloneqq F_\tau(\mu_n).$$

By compactness of $\mathcal{P}_1(\Omega)$, every subsequence of $(\nu_n)_{n\in\mathbb{N}}$ has a further subsequence, still denoted by $(\nu_n)_{n\in\mathbb{N}}$, such that

$$\nu_n \to \nu_* \quad \text{in} \quad \mathsf{W}_1$$

for some $\nu_* \in \mathcal{P}_1(\Omega)$. For every $\eta \in \mathcal{P}_1(\Omega)$, the minimizing property gives

$$\mathrm{Ent}(\nu_n) + \frac{1}{2\tau}\mathsf{W}_1^2(\mu_n,\nu_n) \le \mathrm{Ent}(\eta) + \frac{1}{2\tau}\mathsf{W}_1^2(\mu_n,\eta).$$

Using the lower semicontinuity of Ent and the continuity of W_1, we obtain

$$\mathrm{Ent}(\nu_*) + \frac{1}{2\tau}\mathsf{W}_1^2(\mu,\nu_*) \le \mathrm{Ent}(\eta) + \frac{1}{2\tau}\mathsf{W}_1^2(\mu,\eta).$$

Since η is arbitrary, ν_* is a minimizer of the variational problem associated with μ. By uniqueness,

$$\nu_* = F_\tau(\mu).$$

Hence every convergent subsequence of $(\nu_n)_{n\in\mathbb{N}}$ has the same limit $F_\tau(\mu)$, and therefore

$$F_\tau(\mu_n) \to F_\tau(\mu) \quad \text{in} \quad \mathsf{W}_1.$$

Thus,

$$F_\tau \in \mathcal{C}\big(\mathcal{P}_1(\Omega), \mathcal{P}_1(\Omega)\big).$$

This map also illustrates a limitation of exact deterministic pushforward representations on atomic measures. Suppose that $0 \in \Omega$ and consider the input

$$\mu := \delta_0.$$

The minimum value in the definition of $F_\tau(\delta_0)$ is finite, since there exist absolutely continuous probability measures on Ω with finite entropy. Consequently,

$$\mathrm{Ent}\big(F_\tau(\delta_0)\big) < \infty,$$

and hence

$$F_\tau(\delta_0) \ll dx.$$

In particular, $F_\tau(\delta_0)$ is non-atomic. On the other hand, for every deterministic map

$$G\colon \mathcal{P}_1(\Omega) \times \Omega \to \Omega,$$

we have

$$G(\delta_0, \cdot\,)_\# \delta_0 = \delta_{G(\delta_0,0)},$$

which is a Dirac measure. Therefore,

$$F_\tau(\delta_0) \neq G(\delta_0, \cdot\,)_\# \delta_0$$

for every such G. Thus, W_1-proximal maps provide natural examples of continuous measure-to-measure operators which, on atomic inputs, cannot in general be represented exactly by deterministic pushforward models.

To conclude this appendix, we provide the proof of the lower bound asserted in Example 2.4.

Proposition A.2. *Let*

$$\nu_{u,v} := \frac{1}{2}(\delta_u + \delta_v) \quad \textit{and} \quad \nu^* := \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_1,$$

where $u \le v$. Then

$$\inf_{u \le v} \mathsf{W}_1(\nu_{u,v}, \nu^*) = \frac{1}{3}.$$

The minimum is attained at $(u, v) = (-1, 1)$.

Proof. In one dimension, the 1-Wasserstein distance admits the quantile representation

$$\mathsf{W}_1(\mu, \nu) = \int_0^1 |Q_\mu(t) - Q_\nu(t)|\, dt,$$

where Q_μ and Q_ν denote the corresponding quantile functions. For

$$\nu_{u,v} = \frac{1}{2}(\delta_u + \delta_v) \quad \text{and} \quad u \le v,$$

we have

$$Q_{\nu_{u,v}}(t) = \begin{cases} u, & 0 < t < \frac{1}{2}, \\ v, & \frac{1}{2} \le t < 1, \end{cases}$$

whereas

$$Q_{\nu^*}(t) = \begin{cases} -1, & 0 < t < \frac{1}{3}, \\ 1, & \frac{1}{3} \le t < 1. \end{cases}$$

Hence

$$\begin{aligned} \mathsf{W}_1(\nu_{u,v}, \nu^*) &= \int_0^{1/3} |u+1|\, dt + \int_{1/3}^{1/2} |u-1|\, dt + \int_{1/2}^1 |v-1|\, dt \\ &= \frac{1}{3}|u+1| + \frac{1}{6}|u-1| + \frac{1}{2}|v-1|. \end{aligned}$$

We now minimize this expression under the constraint $u \le v$. If $u \le 1$, the choice $v = 1$ is admissible and minimizes the last term. Thus it remains to minimize

$$f(u) \coloneqq \frac{1}{3}|u+1| + \frac{1}{6}|u-1| \quad \text{for} \quad u \le 1.$$

For $u \le -1$,

$$f(u) = -\frac{1}{2}u - \frac{1}{6},$$

which is decreasing as u increases and therefore attains its minimum at $u = -1$. For $-1 \le u \le 1$,

$$f(u) = \frac{1}{2} + \frac{1}{6}u,$$

which is increasing and hence again attains its minimum at $u = -1$. Therefore,

$$\inf_{u \le 1} f(u) = f(-1) = \frac{1}{3}.$$

If $u > 1$, then the constraint $v \ge u$ implies

$$|v-1| = v - 1 \ge u - 1.$$

Consequently,

$$\begin{aligned} \mathsf{W}_1(\nu_{u,v}, \nu^*) &\ge \frac{1}{3}(u+1) + \frac{1}{6}(u-1) + \frac{1}{2}(u-1) \\ &= u - \frac{1}{3} > \frac{1}{3}. \end{aligned}$$

Thus the global minimum is $\frac{1}{3}$, attained at

$$(u, v) = (-1, 1)$$

as claimed. □

A.2 Additional material for Section 4

By considering Gaussian convolutions, Theorem 4.4 can be generalized to compact sets K which *do not necessarily* satisfy the uniform level set condition; in particular, this generalization allows discrete measures to belong to K. To this end, given $\sigma > 0$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, let μ_σ denote the distribution of the random variable $X + \sigma Z$, where $X \sim \mu$ and $Z \sim \mathcal{N}(0, I_d)$ are independent. That is,

$$\mu_\sigma \coloneqq \mathrm{Law}(X + \sigma Z) = \int_{\mathbb{R}^d} \mathcal{N}(x, \sigma^2 I_d)\, \mu(dx)\,.$$

For a compact set $K \subset \mathcal{P}_p(\mathbb{R}^d)$, define

$$K_\sigma = \{\mu_\sigma \colon \mu \in K\}.$$

We first prove that the mapping $K \mapsto K_\sigma$ preserves W_p-compactness.

Lemma A.3. *If $K \subset \mathcal{P}_p(\mathbb{R}^d)$ is W_p-compact and $\sigma > 0$, then K_σ is W_p-compact.*

Proof. It suffices to show that the mapping $\mu \mapsto \mu_\sigma$ is W_p-continuous. In fact, we will show that it is 1-Lipschitz continuous uniformly in σ. Fix $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_p(\mathbb{R}^d)$. Let (X, Y) be random variables such that $X \sim \mu$, $Y \sim \nu$, and X and Y are W_p-optimally coupled. If $Z \sim \mathcal{N}(0, I_d)$, then $(X + \sigma Z, Y + \sigma Z)$ defines a coupling of μ_σ and ν_σ. Therefore,

$$\begin{aligned} \mathsf{W}_p(\mu_\sigma, \nu_\sigma) &\le \left(\mathbb{E}\|(X + \sigma Z) - (Y + \sigma Z)\|^p\right)^{1/p} \\ &= \left(\mathbb{E}\|X - Y\|^p\right)^{1/p} \\ &= \mathsf{W}_p(\mu, \nu). \end{aligned}$$

This proves the desired claim. □

Moreover, μ_σ uniformly approximates μ over K with an explicit rate.

Lemma A.4. *If $K \subset \mathcal{P}_p(\mathbb{R}^d)$ is W_p-compact and $\sigma > 0$, then*

$$\sup_{\mu \in K} \mathsf{W}_p(\mu_\sigma, \mu) \leq \big(\mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\|Z\|^p\big)^{1/p}\sigma.$$

Proof. If $\mu \in K$, $X \sim \mu$, and $Z \sim \mathcal{N}(0, I_d)$, then $(X, X + \sigma Z)$ defines a coupling of μ and μ_σ. Therefore,

$$\begin{aligned}\sup_{\mu\in K} \mathsf{W}_p(\mu, \mu_\sigma) &\leq \sup_{\mu\in K}\big(\mathbb{E}_{X\sim\mu, Z\sim\mathcal{N}(0,I_d)}\|X - (X + \sigma Z)\|^p\big)^{1/p} \\ &= \big(\mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\|Z\|^p\big)^{1/p}\sigma.\end{aligned}$$

This completes the proof. □

Notice that $K_\sigma \subset \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ for any $\sigma > 0$, even if K contains singular or atomic measures. Application of Theorem 4.4 to K_σ, together with the preceding lemmas, yields the following generalization to Theorem 4.4 that is valid for *any* W_p-compact set K.

Corollary A.5. *Fix $p \in [1, \infty)$. Let $F\colon \mathcal{P}_p(\mathbb{R}^d) \to \mathcal{P}_p(\mathbb{R}^{d'})$ be W_p-continuous. Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be W_p-compact. For any $\varepsilon > 0$, there exists $\sigma > 0$ and a transformer T of the form* (4.2) *such that*

$$\sup_{\mu\in K} \mathsf{W}_p\big(\mathsf{T}(\mu_\sigma, \cdot)_\#\mu_\sigma, F(\mu)\big) < \varepsilon.$$

Proof. Let $\varepsilon > 0$. Fix $\sigma > 0$ to be determined. By Lemma A.3, K_σ is a W_p-compact subset of $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Thus, by Example 3.4, K_σ satisfies the uniform level set condition (3.1) from Assumption 3.1. Consequently, by Theorem 4.4, there exists a deep transformer in-context map $\mathsf{T} = \mathsf{T}_{\sigma,\varepsilon}\colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^{d'}$ of the form (4.2) such that

$$\sup_{\nu\in K_\sigma} \mathsf{W}_p\big(\mathsf{T}(\nu, \cdot)_\#\nu, F(\nu)\big) < \frac{\varepsilon}{2}.$$

By Lemma A.4, the continuity of F, and a result on uniform continuity near compacta (Tsimpos et al., 2026, Lemma C.2), we may choose $\sigma = \sigma(\varepsilon)$ such that

$$\sup_{\mu\in K} \mathsf{W}_p\big(F(\mu_\sigma), F(\mu)\big) < \frac{\varepsilon}{2}.$$

Application of the triangle inequality to obtain

$$\sup_{\mu\in K} \mathsf{W}_p\big(\mathsf{T}(\mu_\sigma, \cdot)_\#\mu_\sigma, F(\mu)\big) \leq \sup_{\nu\in K_\sigma} \mathsf{W}_p\big(\mathsf{T}(\nu, \cdot)_\#\nu, F(\nu)\big) + \sup_{\mu\in K} \mathsf{W}_p\big(F(\mu_\sigma), F(\mu)\big) < \varepsilon$$

completes the proof. □

We conclude with the following observation. The proof of Theorem 4.4—in particular the analysis leading to (C.15) and (C.18)—implies an independently interesting result on the universality of deep transformer in-context maps over W_p-compact sets $K \subset \mathcal{P}_p(\mathbb{R}^d)$ supported on the whole of $\mathbb{R}^d$ instead of on a compact token domain. We state this result as a proposition.

Proposition A.6. *Fix $p \in [1, \infty)$. Let $\Omega \subset \mathbb{R}^d$ be compact and $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be W_p-compact. Then for any $G \in \mathcal{C}(K \times \Omega, \mathbb{R}^{d'})$ and any $\varepsilon > 0$, there exists a deep transformer in-context map $\mathsf{T}\colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^{d'}$ of the form* (4.2) *such that*

$$\sup_{(\mu,x)\in K\times\Omega} |\mathsf{T}(\mu, x) - G(\mu, x)| < \varepsilon.$$

B Proofs of results in Section 3

This appendix provides the remaining proofs of results from Section 3 in the main text.

B.1 Proof of Example 3.3

Proof. Since $\varphi(x) = x$, we have

$$\mu\big(\{x \in \mathbb{R} \colon |\varphi(x) - t| \le r\}\big) = \mu\big([t-r, t+r]\big).$$

Thus, it is enough to prove that

$$\lim_{r \downarrow 0} \sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big([t-r, t+r]\big) = 0.$$

Suppose by contradiction that this fails. Then there exist

$$\varepsilon > 0, \qquad r_n \downarrow 0, \qquad \mu_n \in K, \qquad t_n \in \mathbb{R}$$

such that

$$\mu_n\big([t_n - r_n, t_n + r_n]\big) \ge \varepsilon \quad \text{for all} \quad n \ge 1.$$

Since K is compact, it is tight. Hence there exists $R > 0$ such that

$$\mu\big([-R, R]\big) > 1 - \frac{\varepsilon}{2} \quad \text{for all} \quad \mu \in K.$$

We claim that $(t_n)_{n \in \mathbb{N}}$ is bounded. Indeed, if $(t_n)_{n \in \mathbb{N}}$ were unbounded, then, after passing to a subsequence,

$$|t_n| \to \infty.$$

Since $r_n \to 0$, for all sufficiently large n, it holds that

$$[t_n - r_n, t_n + r_n] \subset \mathbb{R} \setminus [-R, R].$$

Therefore

$$\mu_n\big([t_n - r_n, t_n + r_n]\big) \le \mu_n\big(\mathbb{R} \setminus [-R, R]\big) < \frac{\varepsilon}{2},$$

which contradicts the choice of μ_n and t_n. Thus $(t_n)_{n \in \mathbb{N}}$ is bounded. Passing to a subsequence, we may assume that

$$t_n \to t_*$$

for some $t_* \in \mathbb{R}$. Since K is compact, after passing to a further subsequence,

$$\mu_n \to \mu_*$$

for some $\mu_* \in K$. Since μ_* is non-atomic,

$$\mu_*\big(\{t_*\}\big) = 0.$$

Therefore, by continuity from above,

$$\mu_*\big([t_* - \delta, t_* + \delta]\big) \to \mu_*\big(\{t_*\}\big) = 0 \quad \text{as} \quad \delta \downarrow 0.$$

Hence we may choose $\delta > 0$ such that

$$\mu_*\big([t_* - \delta, t_* + \delta]\big) < \frac{\varepsilon}{2}.$$

For all sufficiently large n, it holds that

$$[t_n - r_n, t_n + r_n] \subset [t_* - \delta, t_* + \delta].$$

Consequently,

$$\varepsilon \le \mu_n\big([t_* - \delta, t_* + \delta]\big).$$

Since $[t_* - \delta, t_* + \delta]$ is closed, the Portmanteau theorem gives

$$\limsup_{n \to \infty} \mu_n\big([t_* - \delta, t_* + \delta]\big) \le \mu_*\big([t_* - \delta, t_* + \delta]\big).$$

It follows that

$$\varepsilon \le \mu_*\big([t_* - \delta, t_* + \delta]\big) < \frac{\varepsilon}{2},$$

a contradiction. Therefore,

$$\lim_{r \downarrow 0} \sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big([t-r, t+r]\big) = 0,$$

which proves the assertion. □

B.2 Proof of Example 3.4

Proof. Since ρ is non-atomic, by Lemma B.1 there exists $v \in \mathbb{S}^{d-1}$ such that

$$x \mapsto \varphi(x) \coloneqq x \cdot v$$

satisfies

$$\varphi_{\#}\rho \in \mathcal{P}_{\mathrm{na}}(\mathbb{R}).$$

Fix $\mu \in K$. Since $\mu \ll \rho$, for every Borel set $B \subset \mathbb{R}$, it holds that

$$(\varphi_{\#}\rho)(B) = 0 \quad \text{implies} \quad \rho\big(\varphi^{-1}(B)\big) = 0.$$

Hence

$$\mu\big(\varphi^{-1}(B)\big) = 0,$$

and therefore

$$(\varphi_{\#}\mu)(B) = 0.$$

Thus

$$\varphi_{\#}\mu \ll \varphi_{\#}\rho.$$

Since $\varphi_{\#}\rho$ is non-atomic, $\varphi_{\#}\mu$ is also non-atomic. Define

$$L \coloneqq \{\varphi_{\#}\mu \colon \mu \in K\} \subset \mathcal{P}_p(\mathbb{R}).$$

Since φ is Lipschitz, the map

$$\mu \mapsto \varphi_{\#}\mu$$

is continuous. Therefore L is compact in $\mathcal{P}_p(\mathbb{R})$. Moreover,

$$L \subset \mathcal{P}_{\mathrm{na}}(\mathbb{R}).$$

We claim that

$$\lim_{r \downarrow 0} \sup_{\nu \in L} \sup_{t \in \mathbb{R}} \nu\big([t-r, t+r]\big) = 0.$$

Suppose by contradiction that this fails. Then there exist $\varepsilon > 0$, $r_n \downarrow 0$, $\nu_n \in L$, and $t_n \in \mathbb{R}$ such that

$$\nu_n\big([t_n - r_n, t_n + r_n]\big) \geq \varepsilon \quad \text{for all} \quad n \geq 1.$$

Since L is compact, it is tight. Hence there exists $R > 0$ such that

$$\nu\big([-R, R]\big) > 1 - \frac{\varepsilon}{2} \quad \text{for all} \quad \nu \in L.$$

Consequently, $(t_n)_{n \in \mathbb{N}}$ must be bounded; otherwise, for infinitely many n, it would hold that

$$[t_n - r_n, t_n + r_n] \subset \mathbb{R} \setminus [-R, R],$$

which would imply

$$\nu_n\big([t_n - r_n, t_n + r_n]\big) \leq \frac{\varepsilon}{2}.$$

This is a contradiction. Now passing to a subsequence,

$$t_n \to t_*$$

for some $t_* \in \mathbb{R}$. Since L is compact, after passing to a further subsequence,

$$\nu_n \to \nu_*$$

for some $\nu_* \in L$. Because ν_* is non-atomic,

$$\nu_*\big(\{t_*\}\big) = 0.$$

Hence there exists $\delta > 0$ such that

$$\nu_*\big([t_* - \delta, t_* + \delta]\big) < \frac{\varepsilon}{2}.$$

For sufficiently large n,

$$[t_n - r_n, t_n + r_n] \subset [t_* - \delta, t_* + \delta].$$

Therefore

$$\varepsilon \leq \nu_n\big([t_* - \delta, t_* + \delta]\big).$$

Since $[t_* - \delta, t_* + \delta]$ is closed, the Portmanteau theorem yields

$$\limsup_{n\to\infty} \nu_n\big([t_* - \delta, t_* + \delta]\big) \le \nu_*\big([t_* - \delta, t_* + \delta]\big).$$

Hence

$$\varepsilon \le \nu_*\big([t_* - \delta, t_* + \delta]\big),$$

contradicting the choice of δ. Thus

$$\lim_{r\downarrow 0} \sup_{\mu\in K} \sup_{t\in\mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(x) - t| \le r\}\big) = \lim_{r\downarrow 0} \sup_{\nu\in L} \sup_{t\in\mathbb{R}} \nu\big([t - r, t + r]\big) = 0$$

as asserted. □

In the preceding proof, we required the following lemma concerning the non-atomicity of almost every one-dimensional projection.

Lemma B.1. *Let $\rho \in \mathcal{P}_{\mathrm{na}}(\mathbb{R}^d)$. Then for surface-almost every $v \in \mathbb{S}^{d-1}$, it holds that*

$$(\pi_v)_{\#}\rho \in \mathcal{P}_{\mathrm{na}}(\mathbb{R}), \quad \textit{where} \quad \pi_v(x) \coloneqq x \cdot v.$$

Proof. The case $d = 1$ is immediate because $\mathbb{S}^0 = \{-1, 1\}$ and $\pi_v(x) = \pm x$. We therefore assume that $d \ge 2$. Let σ denote the surface measure on $\mathbb{S}^{d-1}$. For $v \in \mathbb{S}^{d-1}$, set

$$\nu_v \coloneqq (\pi_v)_{\#}\rho.$$

Let

$$\Delta \coloneqq \{(s, s) \colon s \in \mathbb{R}\} \subset \mathbb{R}^2.$$

By the definition of pushforward measures,

$$(\nu_v \otimes \nu_v)(\Delta) = (\rho \otimes \rho)\big(\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \colon v \cdot (x - y) = 0\}\big).$$

Integrating with respect to v and applying Tonelli's theorem gives

$$\int_{\mathbb{S}^{d-1}} (\nu_v \otimes \nu_v)(\Delta)\, \sigma(dv) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \sigma\big(\{v \in \mathbb{S}^{d-1} \colon v \cdot (x - y) = 0\}\big)\, (\rho \otimes \rho)(dx, dy).$$

If $x \ne y$, then

$$\{v \in \mathbb{S}^{d-1} \colon v \cdot (x - y) = 0\}$$

is a $(d - 2)$-dimensional great subsphere of $\mathbb{S}^{d-1}$ and therefore has zero surface measure. If $x = y$, the above set is the whole sphere. Hence

$$\int_{\mathbb{S}^{d-1}} (\nu_v \otimes \nu_v)(\Delta)\, \sigma(dv) = \sigma(\mathbb{S}^{d-1})\, (\rho \otimes \rho)\big(\{(x, x) \colon x \in \mathbb{R}^d\}\big).$$

Since ρ is non-atomic,

$$(\rho \otimes \rho)\big(\{(x, x) \colon x \in \mathbb{R}^d\}\big) = \int_{\mathbb{R}^d} \rho(\{x\})\, \rho(dx) = 0.$$

Therefore,

$$\int_{\mathbb{S}^{d-1}} (\nu_v \otimes \nu_v)(\Delta)\, \sigma(dv) = 0.$$

Since the integrand is nonnegative,

$$(\nu_v \otimes \nu_v)(\Delta) = 0$$

for σ almost every $v \in \mathbb{S}^{d-1}$. Last, for any Borel probability measure ν on $\mathbb{R}$, it holds that

$$(\nu \otimes \nu)(\Delta) = \int_{\mathbb{R}} \nu(\{t\})\, \nu(dt).$$

It follows that $(\nu \otimes \nu)(\Delta) = 0$ if and only if ν is non-atomic because if ν has an atom t_0 of mass $w > 0$, then

$$(\nu \otimes \nu)(\Delta) \ge (\nu \otimes \nu)\big(\{(t_0, t_0)\}\big) = (\nu \otimes \nu)\big(\{t_0\} \times \{t_0\}\big) = \nu(\{t_0\})^2 = w^2 > 0.$$

Consequently,

$$\nu_v = (\pi_v)_{\#}\rho \in \mathcal{P}_{\mathrm{na}}(\mathbb{R})$$

for σ almost every $v \in \mathbb{S}^{d-1}$. □

B.3 Proof of Example 3.5

Proof. Define

$$E \coloneqq \Big\{ \big(\gamma, \gamma(s)\big) : \gamma \in K',\ s \in [0,1] \Big\} \subset K' \times Q.$$

Consider the map

$$\Psi \colon K' \times [0,1] \to E, \qquad \Psi(\gamma, s) \coloneqq \big(\gamma, \gamma(s)\big).$$

The map Ψ is continuous. Indeed, convergence in $\mathcal{C}^1([0,1], Q)$ implies uniform convergence, so the evaluation map $(\gamma, s) \mapsto \gamma(s)$ is continuous. Moreover, Ψ is injective. In fact, if

$$\big(\gamma, \gamma(s)\big) = \big(\widetilde{\gamma}, \widetilde{\gamma}(t)\big),$$

then $\gamma = \widetilde{\gamma}$, and hence $\gamma(s) = \gamma(t)$. Since γ is injective, we obtain $s = t$. Since $K' \times [0,1]$ is compact and $K' \times Q$ is Hausdorff, Ψ is a homeomorphism from $K' \times [0,1]$ onto E. In particular, E is compact. Define

$$f \colon E \to [0,1], \quad \text{where} \quad f\big(\gamma, \gamma(s)\big) \coloneqq s.$$

Equivalently,

$$f = \pi_2 \circ \Psi^{-1},$$

where $\pi_2(\gamma, s) = s$. Hence f is continuous.

Since $K' \times Q$ is a compact metric space, it is normal. By the Tietze extension theorem, there exists

$$\widetilde{\varphi} \in \mathcal{C}(K' \times Q, [0,1])$$

such that

$$\widetilde{\varphi}|_E = f.$$

Therefore,

$$\widetilde{\varphi}\big(\gamma, \gamma(s)\big) = s \quad \text{for all} \quad \gamma \in K' \quad \text{and} \quad s \in [0,1].$$

Since K' is compact in $\mathcal{C}^1([0,1], Q)$, there exists $C < \infty$ such that

$$|\dot{\gamma}(s)| \le C \qquad \text{for all} \quad \gamma \in K' \quad \text{and} \quad s \in [0,1].$$

On the other hand, by assumption,

$$\int_0^1 |\dot{\gamma}(s)|\, ds \ge c \quad \text{for all} \quad \gamma \in K'.$$

Fix $\gamma \in K'$ and $t \in \mathbb{R}$. By the definition of σ_γ and the identity $\widetilde{\varphi}(\gamma, \gamma(s)) = s$, we have

$$\sigma_\gamma\big\{x \in Q : |\widetilde{\varphi}(\gamma, x) - t| \le r\big\}\big) = \frac{\displaystyle\int_0^1 \mathbf{1}_{\{|s-t| \le r\}} |\dot{\gamma}(s)|\, ds}{\displaystyle\int_0^1 |\dot{\gamma}(s)|\, ds}.$$

Hence

$$\sigma_\gamma\big(\{x \in Q : |\widetilde{\varphi}(\gamma, x) - t| \le r\}\big) \le \frac{C}{c} \mathcal{L}^1\big([0,1] \cap [t-r, t+r]\big) \le \frac{2C}{c} r.$$

The estimate is uniform in $\gamma \in K'$ and $t \in \mathbb{R}$. Therefore,

$$\lim_{r \downarrow 0} \sup_{\gamma \in K'} \sup_{t \in \mathbb{R}} \sigma_\gamma\big(\{x \in Q : |\widetilde{\varphi}(\gamma, x) - t| \le r\}\big) = 0.$$

It remains to relate this parametrized statement to the uniform level set condition on a family of measures. First, the map

$$K' \ni \gamma \mapsto \sigma_\gamma \in \mathcal{P}_p(Q)$$

is continuous. Indeed, if $\gamma_n \to \gamma$ in $\mathcal{C}^1$, then

$$\gamma_n \to \gamma \quad \text{and} \quad |\dot{\gamma}_n| \to |\dot{\gamma}|$$

uniformly on $[0,1]$, and hence, for every $h \in \mathcal{C}(Q)$, it holds that

$$\int_Q h \, d\sigma_{\gamma_n} = \frac{\int_0^1 h(\gamma_n(s))|\dot{\gamma}_n(s)| \, ds}{\int_0^1 |\dot{\gamma}_n(s)| \, ds} \to \frac{\int_0^1 h(\gamma(s))|\dot{\gamma}(s)| \, ds}{\int_0^1 |\dot{\gamma}(s)| \, ds} = \int_Q h \, d\sigma_\gamma.$$

Thus, $\sigma_{\gamma_n} \to \sigma_\gamma$ weakly. Since Q is compact, this is equivalent to convergence in W_p. Consequently,

$$K \coloneqq \{\sigma_\gamma \colon \gamma \in K'\}$$

is compact in $\mathcal{P}_p(Q)$.

Assume now that distinct curves in K' have distinct images, i.e.,

$$\gamma_1([0,1]) = \gamma_2([0,1]) \quad \text{implies} \quad \gamma_1 = \gamma_2.$$

Then, $\gamma \mapsto \sigma_\gamma$ is injective. Since it is a continuous bijection from the compact space K' onto the Hausdorff space K, its inverse

$$K \ni \sigma_\gamma \mapsto \gamma \in K'$$

is continuous. We may therefore define

$$\varphi \colon K \times Q \to [0,1], \quad \text{where} \quad \varphi(\sigma_\gamma, x) \coloneqq \widetilde{\varphi}(\gamma, x).$$

This map is continuous, and the preceding estimate gives

$$\sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big(\{x \in Q \colon |\varphi(\mu, x) - t| \le r\}\big) \le \frac{2C}{c} r.$$

Hence, K satisfies the uniform level set condition. □

B.4 Proof of Theorem 3.6

Proof. The proof proceeds in six steps.

Step 1 (Truncate the output space). For $R > 0$, define the truncation map

$$\Pi_R \colon \mathbb{R}^{d'} \to \overline{B_R(0)} \subset \mathbb{R}^{d'} \quad \text{by} \quad \Pi_R(y) \coloneqq \begin{cases} y, & |y| \le R, \\ R\dfrac{y}{|y|}, & |y| > R, \end{cases}$$

and set

$$F_R(\mu) \coloneqq (\Pi_R)_{\#} F(\mu) \quad \text{for} \quad \mu \in K.$$

Then a direct calculation shows that

$$\mathsf{W}_p^p\big(F(\mu), F_R(\mu)\big) \le \int_{\{|y|>R\}} |y|^p \, F(\mu)(dy).$$

Hence

$$\sup_{\mu \in K} \mathsf{W}_p\big(F(\mu), F_R(\mu)\big) \le E_{\mathrm{tail},p}(R),$$

where

$$E_{\mathrm{tail},p}(R) \coloneqq \left(\sup_{\nu \in F(K)} \int_{\{|y|>R\}} |y|^p \, \nu(dy) \right)^{1/p}.$$

Since $F(K)$ is compact in $(\mathcal{P}_p(\mathbb{R}^{d'}), \mathsf{W}_p)$ by the continuity of F, Lemma B.2 shows that

$$E_{\mathrm{tail},p}(R) \to 0 \quad \text{as} \quad R \to \infty.$$

Step 2 (Cover the truncated output space). Fix $h > 0$. Since $\overline{B_R(0)} \subset \mathbb{R}^{d'}$ is compact, we can choose an h-net

$$\{y_1, \dots, y_N\} \subset \overline{B_R(0)}$$

and a continuous partition of unity $\{\psi_i\}_{i=1}^N \subset \mathcal{C}(\mathbb{R}^{d'})$ such that

$$0 \le \psi_i \le 1, \qquad \sum_{i=1}^N \psi_i \equiv 1 \quad \text{on} \quad \overline{B_R(0)},$$

and

$$\operatorname{supp}(\psi_i) \cap \overline{B_R(0)} \subset B_h(y_i)$$

for each i. Moreover, the covering may be chosen so that

$$N = N(R,h) \le C_{d'} \left(1 + \frac{R}{h}\right)^{d'}$$

by a standard volumetric bound. Next, define

$$\alpha_i(\mu) \coloneqq \int_{\mathbb{R}^{d'}} \psi_i(y)\, F_R(\mu)(dy) \quad \text{and} \quad F_{R,h}(\mu) \coloneqq \sum_{i=1}^N \alpha_i(\mu)\delta_{y_i}.$$

Since F_R is continuous by the 1-Lipschitz property of Π_R and each ψ_i is bounded and continuous, $\alpha_i \colon K \to [0,1]$ is continuous. For each $\mu \in K$, define a probability measure π_μ on $\mathbb{R}^{d'} \times \mathbb{R}^{d'}$ by

$$\pi_\mu \coloneqq \sum_{i=1}^N \big(\mathrm{id}, y_i\big)_{\#}\big(\psi_i F_R(\mu)\big) = \sum_{i=1}^N \big(\psi_i F_R(\mu)\big) \otimes \delta_{y_i},$$

where $\psi_i F_R(\mu)$ denotes the (subprobability) measure with density ψ_i with respect to $F_R(\mu)$. Because $F_R(\mu)$ is supported on $\overline{B_R(0)}$ and $\sum_i \psi_i = 1$ on $\overline{B_R(0)}$, it holds that π_μ is a coupling of $F_R(\mu)$ and $F_{R,h}(\mu)$. Moreover, for each i, the support condition implies that if $\psi_i(y) > 0$, then $|y - y_i| < h$ for $F_R(\mu)$-almost every y. Hence

$$\begin{aligned}
\mathsf{W}_p^p\big(F_R(\mu), F_{R,h}(\mu)\big) &\le \int_{\mathbb{R}^{d'} \times \mathbb{R}^{d'}} |y - z|^p \, \pi_\mu(dy, dz) \\
&= \sum_{i=1}^N \int_{\mathbb{R}^{d'}} |y - y_i|^p \psi_i(y)\, F_R(\mu)(dy) \\
&\le h^p \sum_{i=1}^N \int_{\mathbb{R}^{d'}} \psi_i(y)\, F_R(\mu)(dy) = h^p.
\end{aligned}$$

Therefore,

$$\sup_{\mu \in K} \mathsf{W}_p\big(F_R(\mu), F_{R,h}(\mu)\big) \le h \quad \text{for all} \quad R > 0.$$

Step 3 (Transport to and from the uniform distribution). Define

$$\omega_K(r) \coloneqq \sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(\mu, x) - t| \le r\}\big),$$

where φ is as in Assumption 3.1. Since this assumption holds true by hypothesis,

$$\omega_K(r) \to 0 \quad \text{as} \quad r \downarrow 0.$$

By the construction in Lemma B.3, for every $r > 0$, there exists

$$H_r \in \mathcal{C}(K \times \mathbb{R}^d, [0,1])$$

such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(H_r(\mu, \cdot)_{\#}\mu, \mathrm{Unif}[0,1]\big) \le \omega_K(r). \tag{B.1}$$

Define the cumulative weights

$$s_0(\mu) \coloneqq 0, \quad \text{and} \quad s_i(\mu) \coloneqq \sum_{\ell=1}^i \alpha_\ell(\mu) \quad \text{for all} \quad i = 1, \dots, N.$$

Then $s_N(\mu) = 1$. Also define the discontinuous map

$$(\mu, u) \mapsto Q_0(\mu, u) \coloneqq \sum_{i=1}^{N} \mathbf{1}_{(s_{i-1}(\mu), s_i(\mu)]}(u) y_i.$$

If $U \sim \mathrm{Unif}[0,1]$, then a direct calculation shows that

$$\mathrm{Law}\big(Q_0(\mu, U)\big) = Q_0(\mu, \cdot)_{\#}\mathrm{Unif}[0,1] = F_{R,h}(\mu).$$

Step 4 (Regularize the pushforward map). Let $\delta \in (0,1)$. Let $\kappa \in \mathcal{C}_c^\infty((-1,1))$ be nonnegative with $\int_{\mathbb{R}} \kappa(r)\, dr = 1$. Set

$$\kappa_\delta(r) \coloneqq \delta^{-1}\kappa(r/\delta).$$

Extend $Q_0(\mu, \cdot)$ to $\mathbb{R}$ by setting

$$\overline{Q}_0(\mu, u) \coloneqq \begin{cases} y_1, & u \le 0, \\ Q_0(\mu, u), & 0 < u \le 1, \\ y_N, & u > 1. \end{cases}$$

Define

$$Q_\delta(\mu, u) \coloneqq \int_{\mathbb{R}} \kappa_\delta(u - v)\overline{Q}_0(\mu, v)\, dv.$$

Equivalently,

$$Q_\delta(\mu, u) = \sum_{i=1}^{N} w_i^\delta(\mu, u) y_i,$$

where the nonnegative continuous weights w_i^δ are obtained by integrating $\kappa_\delta(u - \cdot)$ over the intervals associated with y_i. In particular,

$$\sum_{i=1}^{N} w_i^\delta(\mu, u) = 1.$$

Since $|\overline{Q}_0(\mu, v)| \le R$ uniformly in v and μ, it holds that

$$\mathrm{Lip}\big(Q_\delta(\mu, \cdot)\big) \le R\|\kappa_\delta'\|_{L^1(\mathbb{R})} = \frac{R\|\kappa'\|_{L^1(\mathbb{R})}}{\delta} = \frac{C_R}{\delta}, \tag{B.2}$$

where $C_R \coloneqq R\|\kappa'\|_{L^1(\mathbb{R})}$ is independent of μ, r, δ, h, and N. Last, define

$$G_{r,\delta,h}(\mu, x) \coloneqq Q_\delta\big(\mu, H_r(\mu, x)\big),$$

which also depends on R but we suppress this notation. Since H_r and Q_δ are continuous, it holds that

$$G_{r,\delta,h} \in \mathcal{C}(K \times \mathbb{R}^d, \mathbb{R}^{d'}).$$

Step 5 (Bound the regularized pushforward error). We first compare $G_{r,\delta,h}(\mu, \cdot)_{\#}\mu$ with $Q_\delta(\mu, \cdot)_{\#}\mathrm{Unif}[0,1]$. By (B.2) and the Lipschitz stability of W_p, it holds that

$$\mathsf{W}_p\big(G_{r,\delta,h}(\mu, \cdot)_{\#}\mu, Q_\delta(\mu, \cdot)_{\#}\mathrm{Unif}[0,1]\big) \le \frac{C_R}{\delta}\, \mathsf{W}_p\big(H_r(\mu, \cdot)_{\#}\mu, \mathrm{Unif}[0,1]\big).$$

Consequently, by (B.1), it holds that

$$\sup_{\mu \in K} \mathsf{W}_p\big(G_{r,\delta,h}(\mu, \cdot)_{\#}\mu, Q_\delta(\mu, \cdot)_{\#}\mathrm{Unif}[0,1]\big) \le \frac{C_R}{\delta}\omega_K(r).$$

We next compare Q_δ with Q_0. For each $\mu \in K$, let

$$T_\delta(\mu) \coloneqq \bigcup_{i=1}^{N-1} \{u \in [0,1] \colon |u - s_i(\mu)| < \delta\}.$$

Then

$$|T_\delta(\mu)| \le 2N(R,h)\delta,$$

and $Q_\delta(\mu,u) = Q_0(\mu,u)$ outside $T_\delta(\mu)$, apart from a set of Lebesgue measure zero. Since both maps take values in $\overline{B_R(0)}$, it holds that

$$|Q_\delta(\mu,u) - Q_0(\mu,u)| \le 2R$$

for each μ and u. Coupling them by the same random variable $U \sim \mathrm{Unif}[0,1]$ gives

$$\begin{aligned}\mathsf{W}_p^p\big(Q_\delta(\mu,\,\cdot\,)_\#\mathrm{Unif}[0,1], F_{R,h}(\mu)\big) &= \mathsf{W}_p^p\big(\mathrm{Law}(Q_\delta(\mu,U)), \mathrm{Law}(Q_0(\mu,U))\big)\\ &\le \mathbb{E}|Q_\delta(\mu,U) - Q_0(\mu,U)|^p\\ &\le (2R)^p |T_\delta(\mu)|\\ &\le (2R)^p\, 2N(R,h)\delta.\end{aligned}$$

Therefore,

$$\sup_{\mu\in K} \mathsf{W}_p\big(Q_\delta(\mu,\,\cdot\,)_\#\mathrm{Unif}[0,1], F_{R,h}(\mu)\big) \le 2R\big(2N(R,h)\delta\big)^{1/p}.$$

Step 6 (Combine the estimates). Combining Steps 1–5 and using the triangle inequality, we obtain

$$\sup_{\mu\in K} \mathsf{W}_p\big(G_{r,\delta,h}(\mu,\,\cdot\,)_\#\mu, F(\mu)\big) \le E_{\mathrm{tail},p}(R) + h + \frac{C_R}{\delta}\omega_K(r) + 2R\big(2N(R,h)\delta\big)^{1/p}. \quad \text{(B.3)}$$

Choose $R > 0$ sufficiently large such that

$$E_{\mathrm{tail},p}(R) < \frac{\varepsilon}{4}.$$

Next, choose $h > 0$ sufficiently small such that

$$h < \frac{\varepsilon}{4}.$$

With R and h fixed as in the preceding displays, choose $\delta \in (0,1)$ sufficiently small that

$$2R\big(2N(R,h)\delta\big)^{1/p} < \frac{\varepsilon}{4}.$$

Last, since $\omega_K(r) \to 0$ as $r \downarrow 0$, choose $r > 0$ sufficiently small such that

$$\frac{C_R}{\delta}\omega_K(r) < \frac{\varepsilon}{4}.$$

Then (B.3) yields

$$\sup_{\mu\in K} \mathsf{W}_p\big(G_{r,\delta,h}(\mu,\,\cdot\,)_\#\mu, F(\mu)\big) < \varepsilon.$$

This completes the proof. □

The following result is a well-known property of compact sets in Wasserstein space that is used in the preceding argument; we provide a proof for the sake of completeness.

Lemma B.2 (Uniform integrability of p-th moments)**.** *Let $d \in \mathbb{N}$ and $1 \le p < \infty$. Let $K \subset \mathcal{P}_p(\mathbb{R}^d)$ be W_p-compact. Then*

$$\lim_{R\to\infty} \sup_{\nu\in K} \int_{\{|y|>R\}} |y|^p\, \nu(dy) = 0. \quad \text{(B.4)}$$

Proof. Suppose by contradiction that (B.4) fails. Then there exist $\delta > 0$, a sequence $(\nu_n)_{n\in\mathbb{N}} \subset K$, and a sequence $R_n \to \infty$ such that

$$\int_{\{|y|>R_n\}} |y|^p\, \nu_n(dy) \ge \delta \quad \text{for all} \quad n \in \mathbb{N}.$$

Since K is compact with respect to W_p, after passing to a subsequence, we may assume that

$$\nu_n \to \nu \quad \text{in} \quad \mathsf{W}_p$$

for some $\nu \in K$.

Fix $\eta > 0$. Since $\nu \in \mathcal{P}_p(\mathbb{R}^d)$, there exists $M > 0$ such that

$$\int_{\{|y|>M\}} |y|^p \, \nu(dy) < \eta.$$

Choose a continuous function $\chi_M \colon [0,\infty) \to [0,1]$ such that

$$\chi_M(r) = 0 \quad \text{for all} \quad r \le M \quad \text{and} \quad \chi_M(r) = 1 \quad \text{for all} \quad r \ge M+1,$$

and define

$$\varphi_M(y) \coloneqq |y|^p \chi_M(|y|) \quad \text{for all} \quad y \in \mathbb{R}^d.$$

Then φ_M is continuous and satisfies

$$0 \le \varphi_M(y) \le |y|^p,$$

as well as

$$\varphi_M(y) = 0 \quad \text{if} \quad |y| \le M \quad \text{and} \quad \varphi_M(y) = |y|^p \quad \text{if} \quad |y| \ge M+1.$$

Since $\nu_n \to \nu$ in W_p and φ_M is continuous with at most p-th-order growth,

$$\int_{\mathbb{R}^d} \varphi_M(y) \, \nu_n(dy) \to \int_{\mathbb{R}^d} \varphi_M(y) \, \nu(dy) \quad \text{as} \quad n \to \infty.$$

Moreover,

$$\int_{\mathbb{R}^d} \varphi_M(y) \, \nu(dy) \le \int_{\{|y|>M\}} |y|^p \, \nu(dy) < \eta.$$

Hence, for all sufficiently large n, it holds that

$$\int_{\mathbb{R}^d} \varphi_M(y) \, \nu_n(dy) < 2\eta.$$

Since $R_n \to \infty$, we also have $R_n \ge M+1$ for all sufficiently large n. Therefore,

$$\begin{aligned} \int_{\{|y|>R_n\}} |y|^p \, \nu_n(dy) &\le \int_{\{|y|>M+1\}} |y|^p \, \nu_n(dy) \\ &\le \int_{\mathbb{R}^d} \varphi_M(y) \, \nu_n(dy) \\ &< 2\eta. \end{aligned}$$

Choosing $\eta < \delta/2$ gives a contradiction. □

The core of the proof of Theorem 3.6 relies on transporting to the one-dimensional uniform distribution with a jointly continuous in-context map.

Lemma B.3 (Uniform approximate uniformizer)**.** *Let $p \in [1,\infty)$ and $K \subset \mathcal{P}_p(\mathbb{R}^d)$ satisfy Assumption 3.1 with continuous scalarization φ. Then for every $r > 0$, there exists*

$$H_r \in \mathcal{C}(K \times \mathbb{R}^d, [0,1])$$

such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(H_r(\mu,\cdot)_{\#}\mu, \mathrm{Unif}[0,1]\big) \le \omega_K(r),$$

where

$$\omega_K(r) \coloneqq \sup_{\mu \in K} \sup_{t \in \mathbb{R}} \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(\mu,x) - t| \le r\}\big).$$

Proof. Fix $r > 0$. Let $\Theta \colon \mathbb{R} \to [0,1]$ be continuous and nondecreasing such that

$$\Theta(s) = 0 \quad \text{for all} \quad s \le -1 \quad \text{and} \quad \Theta(s) = 1 \quad \text{for all} \quad s \ge 1.$$

Define

$$H_r(\mu, x) \coloneqq \int_{\mathbb{R}^d} \Theta\left(\frac{\varphi(\mu, x) - \varphi(\mu, z)}{r}\right) \mu(dz).$$

Clearly H_r takes values in $[0, 1]$.

We first show that

$$H_r \in \mathcal{C}(K \times \mathbb{R}^d, [0, 1]).$$

To this end, let $(\mu_n, x_n) \to (\mu, x)$ in $K \times \mathbb{R}^d$ as $n \to \infty$ and set

$$g_n(z) \coloneqq \Theta\left(\frac{\varphi(\mu_n, x_n) - \varphi(\mu_n, z)}{r}\right) \quad \text{and} \quad g(z) \coloneqq \Theta\left(\frac{\varphi(\mu, x) - \varphi(\mu, z)}{r}\right).$$

Since φ is continuous, $g_n \to g$ uniformly on every compact subset of $\mathbb{R}^d$. Moreover,

$$0 \leq g_n \leq 1 \quad \text{and} \quad 0 \leq g \leq 1$$

and each are continuous. Since $\mu_n \to \mu$ in W_p, in particular μ_n converges to μ in distribution, and the family $\{\mu_n \colon n \in \mathbb{N}\} \cup \{\mu\}$ is uniformly tight. It follows that

$$\int_{\mathbb{R}^d} g_n(z)\, \mu_n(dz) \to \int_{\mathbb{R}^d} g(z)\, \mu(dz).$$

Hence

$$H_r(\mu_n, x_n) \to H_r(\mu, x) \quad \text{as} \quad n \to \infty,$$

so H_r is continuous.

Now fix $\mu \in K$ and define

$$\nu_\mu \coloneqq \varphi(\mu, \cdot)_{\#}\mu \in \mathcal{P}(\mathbb{R}).$$

Assumption 3.1 implies ν_μ is non-atomic. To see this, note that for every $t \in \mathbb{R}$, it holds that

$$\nu_\mu(\{t\}) = \mu\big(\{x \in \mathbb{R}^d \colon \varphi(\mu, x) = t\}\big)$$

and

$$\nu_\mu(\{t\}) \leq \mu\big(\{x \in \mathbb{R}^d \colon |\varphi(\mu, x) - t| \leq \eta\}\big)$$

for every $\eta > 0$. Letting $\eta \downarrow 0$ and using the uniform level set condition gives

$$\nu_\mu(\{t\}) = 0$$

as claimed.

Continuing, we let

$$F_\mu(t) \coloneqq \nu_\mu\big((-\infty, t]\big) \quad \text{for} \quad t \in \mathbb{R}$$

be the cumulative distribution function of ν_μ. Define its smoothed version by

$$F_{\mu,r}(t) \coloneqq \int_{\mathbb{R}} \Theta\left(\frac{t - s}{r}\right) \nu_\mu(ds).$$

Then

$$H_r(\mu, x) = F_{\mu,r}\big(\varphi(\mu, x)\big),$$

and consequently

$$H_r(\mu, \cdot)_{\#}\mu = (F_{\mu,r})_{\#}\nu_\mu.$$

For every $t \in \mathbb{R}$, it holds that

$$|F_{\mu,r}(t) - F_\mu(t)| \leq \nu_\mu([t - r, t + r]).$$

Indeed, the functions

$$s \mapsto \Theta\left(\frac{t - s}{r}\right) \quad \text{and} \quad s \mapsto \mathbf{1}_{(-\infty, t]}(s)$$

coincide outside the set $[t - r, t + r]$. Hence,

$$\sup_{t \in \mathbb{R}} |F_{\mu,r}(t) - F_\mu(t)| \leq \omega_K(r) \quad \text{for all} \quad \mu \in K.$$

Since ν_μ is non-atomic, F_μ is continuous, and the probability integral transform gives

$$(F_\mu)_\# \nu_\mu = \mathrm{Unif}[0,1].$$

Using the coupling $\mathrm{Law}(F_{\mu,r}(t), F_\mu(t))$ with $t \sim \nu_\mu$, we obtain the upper bound

$$\begin{aligned}
\mathsf{W}_p^p\big(H_r(\mu,\,\cdot\,)_\#\mu, \mathrm{Unif}[0,1]\big) &= \mathsf{W}_p^p\big((F_{\mu,r})_\#\nu_\mu, (F_\mu)_\#\nu_\mu\big) \\
&\le \int_{\mathbb{R}} |F_{\mu,r}(t) - F_\mu(t)|^p \, \nu_\mu(dt) \\
&\le \sup_{t\in\mathbb{R}} |F_{\mu,r}(t) - F_\mu(t)|^p \\
&\le \omega_K(r)^p.
\end{aligned}$$

Since this estimate is uniform in $\mu \in K$, we conclude that

$$\sup_{\mu\in K} \mathsf{W}_p\big(H_r(\mu,\,\cdot\,)_\#\mu, \mathrm{Unif}[0,1]\big) \le \omega_K(r).$$

The proof is complete. □

C Proofs of results in Section 4

This appendix provides the remaining proofs of results from Section 4 in the main text.

C.1 Proof of Theorem 4.4

Proof. Fix $\varepsilon > 0$. By Theorem 3.6 and Remark 3.7, there exists a *bounded* continuous map

$$G\colon K \times \mathbb{R}^d \to \mathbb{R}^{d'}$$

such that

$$\sup_{\mu\in K} \mathsf{W}_p\big(G(\mu,\,\cdot\,)_\#\mu, F(\mu)\big) < \frac{\varepsilon}{2}. \tag{C.1}$$

Set

$$M_G \coloneqq \sup_{(\mu,x)\in K\times\mathbb{R}^d} |G(\mu,x)| < \infty.$$

We first compactify the input space. Let

$$\Omega \coloneqq [-1,1]^d$$

and define

$$\begin{aligned}
\iota\colon \mathbb{R}^d &\to (-1,1)^d, \\
x &\mapsto \iota(x) \coloneqq \frac{x}{1+|x|}.
\end{aligned}$$

The map ι is bounded, continuous, and injective. Since it is 1-Lipschitz, the induced pushforward

$$\begin{aligned}
\iota_\#\colon \mathcal{P}_p(\mathbb{R}^d) &\to \mathcal{P}(\Omega), \\
\mu &\mapsto \iota_\#\mu
\end{aligned}$$

is continuous with respect to W_p. Define

$$K^\iota \coloneqq \{\iota_\#\mu \in \mathcal{P}(\Omega)\colon \mu \in K\} \subset \mathcal{P}_p(\mathbb{R}^d).$$

Since K is compact, K^ι is also compact by continuity of $\iota_\#$. Moreover,

$$\iota_\#|_K\colon K \to K^\iota$$

is a homeomorphism.

Since K is compact in $(\mathcal{P}_p(\mathbb{R}^d), \mathsf{W}_p)$, it is uniformly tight. Hence

$$\tau_R \coloneqq \sup_{\mu\in K} \mu\big(\mathbb{R}^d \setminus B_R(0)\big) \to 0 \quad \text{as} \quad R \to \infty.$$

Choose $R > 0$ sufficiently large such that

$$(3 + 4M_G)\tau_R^{1/p} < \frac{\varepsilon}{4}. \tag{C.2}$$

Set

$$C_R \coloneqq \iota\big(\overline{B_R(0)}\big) \subset \Omega.$$

Then C_R is compact. Define

$$\widetilde{G}_R \colon K^\iota \times C_R \to \overline{B_{M_G}(0)}$$

by

$$\widetilde{G}_R\big(\iota_\# \mu, \iota(x)\big) \coloneqq G(\mu, x) \quad \text{for all} \quad \mu \in K \quad \text{and} \quad x \in \overline{B_R(0)}.$$

This map is well-defined and continuous because ι and $\iota_\#|_K$ are homeomorphisms onto their images.

Since $K^\iota \times C_R$ is a closed subset of the compact metric space $\mathcal{P}(\Omega) \times \Omega$, the Dugundji extension theorem (Dugundji, 1951) delivers the existence of a continuous map

$$\widehat{G}_R \colon \mathcal{P}(\Omega) \times \Omega \to \overline{B_{M_G}(0)}$$

such that

$$\widehat{G}_R\big(\iota_\# \mu, \iota(x)\big) = G(\mu, x) \quad \text{for all} \quad \mu \in K \quad \text{and} \quad x \in \overline{B_R(0)}, \tag{C.3}$$

and

$$\sup_{(\nu,z)\in\mathcal{P}(\Omega)\times\Omega} |\widehat{G}_R(\nu, z)| \le M_G. \tag{C.4}$$

Choose $\delta \in (0, 1)$ sufficiently small such that

$$\delta < \frac{\varepsilon}{8}. \tag{C.5}$$

By the universal approximation theorem for measure-theoretic transformer in-context maps on the compact domain $\mathcal{P}(\Omega) \times \Omega$ (Furuya et al., 2025b, Theorem 1), there exists a measure-theoretic transformer

$$\mathsf{T}_\delta \colon \mathcal{P}(\Omega) \times \Omega \to \mathbb{R}^{d'}$$

of the form (4.2)—depending on R—such that

$$\sup_{(\nu,z)\in\mathcal{P}(\Omega)\times\Omega} \big|\mathsf{T}_\delta(\nu, z) - \widehat{G}_R(\nu, z)\big| < \delta. \tag{C.6}$$

In particular, by (C.4) and the preceding display, it holds that

$$\sup_{(\nu,z)\in\mathcal{P}(\Omega)\times\Omega} |\mathsf{T}_\delta(\nu, z)| \le \delta + M_G < 1 + M_G. \tag{C.7}$$

Since Ω is compact, $\mathcal{P}(\Omega) \times \Omega$ is compact. Since T_δ is continuous on $\mathcal{P}(\Omega) \times \Omega$ (Furuya et al., 2026), it is uniformly continuous. Therefore, there exists a nondecreasing function

$$\omega_\delta \colon [0, \infty) \to [0, \infty)$$

such that $\lim_{r\downarrow 0} \omega_\delta(r) = 0$ and

$$\big|\mathsf{T}_\delta(\nu, z) - \mathsf{T}_\delta(\nu', z')\big| \le \omega_\delta\big(\mathsf{W}_p(\nu, \nu') + |z - z'|\big) \tag{C.8}$$

for every ν and ν' in $\mathcal{P}(\Omega)$ and z and z' in Ω. Now choose $\rho > 0$ such that

$$\omega_\delta(\rho) < \frac{\varepsilon}{8}. \tag{C.9}$$

After δ and ρ have been fixed, choose $S \ge R$ sufficiently large such that

$$\tau_S \coloneqq \sup_{\mu\in K} \mu\big(\mathbb{R}^d \setminus B_S(0)\big) < \left(\frac{\rho}{4\sqrt{d}}\right)^p. \tag{C.10}$$

We next approximate ι on $\overline{B_S(0)}$ by a token-wise MLP. Define

$$\mathrm{clip}(t) \coloneqq -1 + \mathrm{ReLU}(t + 1) - \mathrm{ReLU}(t - 1)$$

and

$$\mathrm{Clip}(z) \coloneqq \big(\mathrm{clip}(z_1), \dots, \mathrm{clip}(z_d)\big).$$

Then Clip is exactly representable by a shallow ReLU network and satisfies

$$\mathrm{Clip}(\mathbb{R}^d) \subset \Omega.$$

For any $\eta > 0$, let $\widetilde{N}_\eta \colon \mathbb{R}^d \to \mathbb{R}^d$ be a ReLU MLP satisfying

$$\sup_{x \in \overline{B_S(0)}} |\widetilde{N}_\eta(x) - \iota(x)| < \eta.$$

This is always possible by the universal approximation theorem for ReLU MLPs. Set

$$N_\eta \coloneqq \mathrm{Clip} \circ \widetilde{N}_\eta,$$

which is also a MLP. Since Clip is the metric projection onto $\Omega = [-1, 1]^d$, it is 1-Lipschitz. Thus,

$$\sup_{x \in \overline{B_S(0)}} |N_\eta(x) - \iota(x)| = \sup_{x \in \overline{B_S(0)}} \big|\mathrm{Clip}\big(\widetilde{N}_\eta(x)\big) - \mathrm{Clip}\big(\iota(x)\big)\big| < \eta. \tag{C.11}$$

Moreover,

$$N_\eta(\mathbb{R}^d) \subset \Omega.$$

Choose $\eta > 0$ sufficiently small such that

$$\eta < \frac{\rho}{4}. \tag{C.12}$$

Let Γ_{id} be an attention layer of the form (4.1) with $Q^h = K^h = V^h = 0$ for every head h. Then

$$\Gamma_{\mathrm{id}}(\mu, x) = x$$

for every $(\mu, x) \in \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d$. Define $\mathsf{T} \colon \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^{d'}$ by

$$\mathsf{T} \coloneqq \mathsf{T}_\delta \diamond N_\eta \diamond \Gamma_{\mathrm{id}}. \tag{C.13}$$

Equivalently,

$$\mathsf{T}(\mu, x) = \mathsf{T}_\delta\big((N_\eta)_{\#}\mu, N_\eta(x)\big).$$

The initial identity attention layer and the token-wise MLP N_η put T in the form (4.2). The initial layer Γ_{id} is defined for every probability measure. Then $\mu \mapsto (N_\eta)_{\#}\mu$ maps $\mathcal{P}(\mathbb{R}^d)$ into $\mathcal{P}([-1, 1]^d)$. Consequently, all subsequent attention layers act on measures supported on compact subsets of Euclidean space because measure-theoretic attention operators map compactly supported probability measures to compactly supported probability measures. Thus, T is well-defined on $\mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d$.

We next estimate the error incurred by replacing ι with N_η. For $\mu \in K$, the measure $(N_\eta, \iota)_{\#}\mu$ is a coupling of $(N_\eta)_{\#}\mu$ and $\iota_{\#}\mu$. Hence

$$\begin{aligned}\mathsf{W}_p^p\big((N_\eta)_{\#}\mu, \iota_{\#}\mu\big) &\le \int_{\mathbb{R}^d} |N_\eta(x) - \iota(x)|^p \, \mu(dx) \\ &\le \eta^p + (2\sqrt{d})^p \mu\big(\mathbb{R}^d \setminus B_S(0)\big)\end{aligned}$$

by splitting the integral and invoking (C.11). Using $(|a|^p + |b|^p)^{1/p} \le |a| + |b|$ yields

$$\sup_{\mu \in K} \mathsf{W}_p\big((N_\eta)_{\#}\mu, \iota_{\#}\mu\big) \le \eta + 2\sqrt{d}\, \tau_S^{1/p}, \tag{C.14}$$

with τ_S as in (C.10). So, for $\mu \in K$ and $x \in B_R(0) \subset B_S(0)$, it holds that

$$\mathsf{W}_p\big((N_\eta)_{\#}\mu, \iota_{\#}\mu\big) + |N_\eta(x) - \iota(x)| \le 2\eta + 2\sqrt{d}\, \tau_S^{1/p} < \frac{\rho}{2} + \frac{\rho}{2} = \rho$$

by (C.12) and (C.10). Then by (C.8) and (C.9), it holds that

$$\big|\mathsf{T}_\delta\big((N_\eta)_{\#}\mu, N_\eta(x)\big) - \mathsf{T}_\delta\big(\iota_{\#}\mu, \iota(x)\big)\big| \le \omega_\delta(\rho) < \frac{\varepsilon}{8} \tag{C.15}$$

for every $\mu \in K$ and $x \in B_R(0)$. For $\mu \in K$ and $x \in \mathbb{R}^d \setminus B_R(0)$, (C.7) shows that

$$\left|\mathsf{T}_\delta\big((N_\eta)_\#\mu, N_\eta(x)\big) - \mathsf{T}_\delta\big(\iota_\#\mu, \iota(x)\big)\right| \le 2(1 + M_G). \tag{C.16}$$

Consequently, with $D_\mu(x)$ denoting the difference on the left-hand side of (C.16), we have

$$\left(\int_{\mathbb{R}^d} |D_\mu(x)|^p \, \mu(dx)\right)^{1/p} \le \frac{\varepsilon}{8} + 2(1 + M_G)\tau_R^{1/p}. \tag{C.17}$$

Similarly, for $x \in B_R(0)$, (C.3) and (C.6) yield

$$\sup_{\mu \in K} \left|\mathsf{T}_\delta\big(\iota_\#\mu, \iota(x)\big) - G(\mu, x)\right| < \delta. \tag{C.18}$$

For $x \in \mathbb{R}^d \setminus B_R(0)$, (C.7) gives

$$\sup_{\mu \in K} \left|\mathsf{T}_\delta\big(\iota_\#\mu, \iota(x)\big) - G(\mu, x)\right| \le (1 + M_G) + M_G = 1 + 2M_G.$$

Therefore,

$$\left(\int_{\mathbb{R}^d} \left|\mathsf{T}_\delta\big(\iota_\#\mu, \iota(x)\big) - G(\mu, x)\right|^p \mu(dx)\right)^{1/p} \le \delta + (1 + 2M_G)\tau_R^{1/p}. \tag{C.19}$$

Using the common random variable $X \sim \mu$ to couple $\mathsf{T}(\mu, X)$ and $G(\mu, X)$, followed by Minkowski's inequality and (C.17)–(C.19), we obtain the estimate

$$\sup_{\mu \in K} \mathsf{W}_p\big(\mathsf{T}(\mu, \cdot)_\#\mu, G(\mu, \cdot)_\#\mu\big) \le \frac{\varepsilon}{8} + \delta + (3 + 4M_G)\tau_R^{1/p} < \frac{\varepsilon}{2} \tag{C.20}$$

by (C.5) and (C.2). Last, the triangle inequality, (C.20), and (C.1) deliver the bound

$$\begin{aligned}\sup_{\mu \in K} \mathsf{W}_p\big(\mathsf{T}(\mu, \cdot)_\#\mu, F(\mu)\big) &\le \sup_{\mu \in K} \mathsf{W}_p\big(\mathsf{T}(\mu, \cdot)_\#\mu, G(\mu, \cdot)_\#\mu\big) \\ &\qquad + \sup_{\mu \in K} \mathsf{W}_p\big(G(\mu, \cdot)_\#\mu, F(\mu)\big) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\end{aligned}$$

as asserted. □

C.2 Proof of Corollary 4.6

Proof. The proof follows the same construction as that of Theorem 3.6, but now with the input measure μ replaced by the continuously-varying source measure $\eta(\mu)$. Indeed, the assumption (4.3), together with the continuity of η, yields, by the same argument as in Lemma B.3, a continuous map

$$H_r \in \mathcal{C}\big(K \times \mathbb{R}^m, [0, 1]\big)$$

such that

$$\sup_{\mu \in K} \mathsf{W}_p\big(H_r(\mu, \cdot)_\#\eta(\mu), \mathrm{Unif}[0, 1]\big) \to 0 \quad \text{as} \quad r \downarrow 0.$$

The remainder of the proof is identical to that of Theorem 3.6, except now with $H_r(\mu, \cdot)_\#\mu$ replaced by $H_r(\mu, \cdot)_\#\eta(\mu)$. □