ONE OPERATOR FOR MANY DENSITIES: AMORTIZED APPROXIMATION OF CONDITIONING BY NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic conditioning is concerned with the identification of a distribution of a random variable X given a random variable Y. It is a cornerstone of scientific and engineering applications where modeling uncertainty is key. This problem has traditionally been addressed in machine learning by directly learning the conditional distribution of a fixed joint distribution. This paper introduces a novel perspective: we propose to solve the conditioning problem by identifying a single operator that maps any joint density to its conditional, thus amortizing over joint-conditional pairs. We establish that the conditioning operator can be approximated to arbitrary accuracy by neural operators. Our proof relies on new results establishing continuity of the conditioning operator over suitable classes of densities, including densities defined on unbounded domains. Last, we learn the conditioning map for a class of Gaussian mixtures using neural operators, illustrating the promise of our framework. This work provides the theoretical underpinnings for general-purpose, amortized methods for probabilistic conditioning.

1 INTRODUCTION

Conditioning is a fundamental operation in probabilistic modeling. Given a joint distribution over random variables (X, Y), the conditional distribution $p(X \mid Y = y)$ for some observed y is the central object underlying prediction, inference, and decision-making. Modern machine learning systems traditionally solve the problem by learning the conditional distribution of a fixed joint, often from a statistical standpoint. In this work, we adopt an operator-theoretic perspective on probabilistic conditioning. Specifically, we study the map that takes a joint probability density ρ over (X, Y) and a conditioning variable y to the conditional density $\rho(\,\cdot \mid y)$. This defines a nonlinear operator

$$\Psi^\star \colon (\rho, y) \mapsto \rho(\,\cdot \mid y)$$

acting between function spaces. Under natural regularity assumptions, we show that $\Psi^\star$ is continuous between suitable metric spaces of probability densities. This seemingly simple observation has a powerful consequence: *conditioning can be universally approximated by neural operators*. Neural operators are learnable mappings between function spaces that serve as the natural generalization of deep neural networks to the infinite-dimensional setting considered in this paper. A major benefit of our operator learning perspective is that it enables amortization across joint distributions, which eliminates the need to retrain the neural operator for each new input density.

Contributions. Our main contributions are as follows.

- **Operator-theoretic formulation of conditioning.** We formalize probabilistic conditioning as a nonlinear operator acting on spaces of continuous probability densities and identify two complementary formulations: kernel conditioning and in-context conditioning.
- **Stability of conditioning.** We prove that the conditioning operator is continuous with respect to the supremum norm on suitable subsets of continuous densities. Under additional regularity assumptions, we establish quantitative Hölder stability bounds.
- **Universal approximation by neural operators.** Leveraging the stability results, we show that both formulations of the conditioning operator can be uniformly approximated on compact sets

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.

of probability densities by neural operators. In doing so, we propose trainable neural operator architectures for conditioning.

- **Unbounded domains and extensions.** We establish a general neural operator universal approximation theorem for continuous operators from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R}^{d'})$. We then use it to prove the approximation of the kernel conditioning operator by neural operators over unbounded L^1 densities. Our analysis also goes beyond uniformly lower bounded marginal densities and characterizes when conditioning admits continuous extensions.
- **Numerical proof-of-concept.** We go beyond theory by numerically demonstrating that *trained* neural operators can accurately implement conditioning in Gaussian mixture test problems.

Significance and related work. The approximation guarantees established in this paper pave the way for the development of a *data-driven model for amortized conditioning*: a machine learning model that, upon being trained on a sufficiently large and diverse dataset, can rapidly perform conditional inference on new, unseen joint probability densities. This has immediate implications for sequential Bayesian inference (Stuart, 2010) tasks such as nonlinear filtering in data assimilation (Calvello et al., 2025b; Sanz-Alonso et al., 2023) or optimal experimental design (Huan et al., 2024; Wu et al., 2023), where conditioning is a challenging but key computational primitive.

Our work builds on and connects several lines of research.

Operator learning. Operator learning was popularized in work of Bhattacharya et al. (2021); Kovachki et al. (2023); Li et al. (2021); Lu et al. (2021). The resulting architectures have achieved state-of-the-art performance for emulating partial differential equations (PDEs) (Li et al., 2021), with applications ranging from weather forecasting (Bonev et al., 2025; Kurth et al., 2023) to problems defined on complex geometries (Li et al., 2023; Zeng et al., 2025). One approach called DeepONet (Chen & Chen, 1995; Lu et al., 2021) approximates operators via a branch–trunk decomposition. Our in-context formulation can be viewed as a continuous analog of this architecture, where the conditioning variable plays the role of a query input. Another line of work introduced neural operators (Kovachki et al., 2023; 2024), which lift neural networks to function spaces, and are supported by extensive theoretical analysis (Brugiapaglia et al., 2026; Kovachki et al., 2024; Lanthaler et al., 2022; 2025). Recent work formulates continuum attention as an operator on function spaces, leading to the transformer neural operator (Calvello et al., 2025a). These methods have been successfully applied to PDEs (Boullé & Townsend, 2024; Li et al., 2021) and inverse problems (de Hoop et al., 2025; Nelsen & Yang, 2026). We extend this framework to probabilistic conditioning operators.

Conditional generative modeling and conditional kernel mean embeddings. Conditional GANs (Mirza & Osindero, 2014), conditional VAEs (Sohn et al., 2015), and flow-based models (Dinh et al., 2017; Lipman et al., 2023; Pierret et al., 2026) aim to model conditional distributions from data samples only. These approaches are often parametrized in finite-dimensional latent spaces and invoke measure transport (Marzouk et al., 2017). In contrast, we study conditioning at the level of function spaces and provide approximation theorems. A different direction studies conditional kernel mean embeddings (Li et al., 2022; Song et al., 2009) that represent conditional distributions in reproducing kernel Hilbert spaces (RKHS). This motivated work on rigorous operator-theoretic conditional density estimation methods (Schuster et al., 2020). Our operator learning formulation is related, but operates directly in spaces of probability densities rather than RKHS embeddings and amortizes over joint densities without retraining.

Learning on distribution-valued data and amortized inference. Machine learning in metric spaces such as the Wasserstein space of probability distributions (León & Niles-Weed, 2026; Lyu et al., 2026; Moosmüller & Cloninger, 2023; Panaretos & Zemel, 2020) is a contemporary research direction that aligns with our work. For distribution-to-distribution regression (Szabó et al., 2016; Vandergrift et al., 2026), measure-theoretic transformers are one prominent method (Bach et al., 2025; Biswal et al., 2025; Calvello et al., 2026; Castin et al., 2024; Fraiman, 2026; Furuya et al., 2025a;b; Geshkovski et al., 2025; Kawata & Suzuki, 2026; Lavenant & Savaré, 2026). They have been used for amortized unconditional density/score estimation (Ilin & Sushko, 2026), amortized Bayesian inference (Gloeckler et al., 2024; Whittle et al., 2026), and amortized optimal transport (Cole et al., 2026). These and related deep learning methods have been applied to yield fast non-Gaussian data assimilation algorithms (Al-Jarrah et al., 2025; Bach et al., 2025; Binder et al., 2026; Calvello et al., 2026; Cui et al., 2026; Revach et al., 2022; Tong et al., 2026).

2 FORMULATION

Let $D \subset \mathbb{R}^m$ and $E \subset \mathbb{R}^r$ be closures of bounded open sets with Lipschitz boundaries. Consider the space of continuous probability densities on $D \times E$ given by

$$C_{\mathrm{ac}}(D \times E) \coloneqq \left\{ \rho \in C(D \times E) \,\middle|\, \int_{D \times E} \rho(z)\, dz = 1 \quad \text{and} \quad \inf_{z \in D \times E} \rho(z) \geq 0 \right\}.$$

Write $z \coloneqq (x, y)$ for $x \in D$ the *state* or *target* variable and $y \in E$ the *query* or *conditioning* variable. For each query $y \in E$, we can define the *data marginal density* by

$$m_\rho(y) \coloneqq \int_D \rho(x, y)\, dx. \tag{1}$$

Finally, for a density $\rho \in C_{\mathrm{ac}}(D \times E)$ we can introduce the collection of queries with non-degenerate marginals by $E_\rho \coloneqq \{y \in E \,|\, m_\rho(y) > 0\}$, which allows us to define the *conditional density*

$$\rho(x \,|\, y) \coloneqq \frac{\rho(x, y)}{m_\rho(y)} \tag{2}$$

for all $x \in D$ and $y \in E_\rho$. For reasons that will soon become clear, we also need to introduce sets of queries with "well-behaved" data marginals. For a strictly positive number $\delta > 0$, let

$$\mathcal{X}_\delta \coloneqq \left\{ \rho \in C_{\mathrm{ac}}(D \times E) \,\middle|\, \inf_{y \in E} m_\rho(y) \geq \delta \right\}.$$

Writing $\mathsf{d}_{C(D \times E)}$ and $\mathsf{d}_{C(D)}$ for the metrics induced by the supremum norms on $C(D \times E)$ and $C(D)$, respectively, we can construct metric spaces

$$\mathcal{M}_\delta \coloneqq \big(\mathcal{X}_\delta, \mathsf{d}_{C(D \times E)}\big) \quad \text{and} \quad \mathcal{M}_\delta^E \coloneqq \big(\mathcal{X}_\delta \times E, \mathsf{d}_{C(D \times E) \times E}\big) \tag{3}$$

with $\mathsf{d}_{C(D \times E) \times E}$ induced by the *product norm* $\|(\rho, y)\|_{C(D \times E) \times E} \coloneqq \max\big(\|\rho\|_{C(D \times E)}, \|y\|\big)$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^r \supset E$. Note that this is the natural norm corresponding to the direct sum of Banach spaces $C(D \times E) \oplus \mathbb{R}^r$. Lemma B.1 establishes that (3) are indeed metric spaces.

Our goal in this section is to construct a well-behaved conditioning operator, that is, a mapping that takes any density $\rho \in C_{\mathrm{ac}}(D \times E)$ as input and returns its conditional. There exist at least two natural ways to do this: one can either input a ρ together with a query y and output the conditional function $\rho(\,\cdot \,|\, y)$, or one can input ρ alone and output the entire kernel function $\kappa_\rho(\,\cdot\,, \cdot\,) = \rho(\,\cdot \mid \cdot\,)$. As we will see, both definitions are possible and intimately linked.

2.1 THE IN-CONTEXT CONDITIONING OPERATOR

First, we define the *in-context conditioning operator* $\Psi^\star$ that takes a density ρ and a query y to produce the conditional density $\rho(\,\cdot \,|\, y)$. Noting that for $\rho \in \mathcal{X}_\delta$ we have $E_\rho = E$, we can define

$$\begin{aligned} \Psi^\star \colon \mathrm{Dom}(\Psi^\star) &\to C(D) \\ (\rho, y) &\mapsto \rho(\,\cdot \,|\, y) \end{aligned} \tag{4}$$

with domain $\mathrm{Dom}(\Psi^\star) \coloneqq \mathcal{X}_\delta \times E$. We equip $\mathrm{Dom}(\Psi^\star)$ with the metric $\mathsf{d}_{C(D \times E) \times E}$ to obtain the metric space $\mathcal{M}_\delta^E$, see (3).

2.2 THE KERNEL CONDITIONING OPERATOR

Now we define the *kernel conditioning operator* $\mathcal{G}^\star$ that takes a density ρ and produces the kernel function $(x, y) \mapsto \rho(x \,|\, y)$. To do so, we first define the *kernel function* for any $\rho \in C_{\mathrm{ac}}(D \times E)$ by

$$\begin{aligned} \kappa_\rho \colon D \times E_\rho &\to \mathbb{R}_+ \\ (x, y) &\mapsto \rho(x \,|\, y). \end{aligned} \tag{5}$$

We can now define the *kernel conditioning operator* by

$$\begin{aligned} \mathcal{G}^\star \colon \mathrm{Dom}(\mathcal{G}^\star) &\to C(D \times E) \\ \rho &\mapsto \kappa_\rho, \end{aligned} \tag{6}$$

with domain $\mathrm{Dom}(\mathcal{G}^\star) \coloneqq \mathcal{X}_\delta$. We equip $\mathrm{Dom}(\mathcal{G}^\star)$ with the metric induced by the supremum norm to obtain the metric space $\mathcal{M}_\delta$ (3). Last, we note that $\Psi^\star$ and $\mathcal{G}^\star$ are related by the *exponential map*; see Remark E.1.

3 MAIN RESULTS

This section presents the central results of the paper: uniform approximation of the conditioning operators $\mathcal{G}^\star$ and $\Psi^\star$ by *neural operators*. First, we define neural operators in Section 3.1 as learnable mappings between function spaces, implemented by neural networks. Then in Section 3.2, we discuss the first set of operator approximation results. These establish that $\mathcal{G}^\star$ and $\Psi^\star$ can be approximated to arbitrary accuracy by neural operators, uniformly over compact sets of continuous densities; examples of the latter include balls of α-Hölder densities. Last, in Section 3.3, we prove that neural operators can also approximate an extension of $\mathcal{G}^\star$ that acts on $\mathbb{R}^{m+r}$ densities that are bounded below, i.e., densities on an unbounded domain. In particular, this delivers approximation guarantees over compact sets of Gaussian mixtures. The technical backbone of the approximation theorems in this section is the continuity and stability results that will be presented later in Section 4.

3.1 NEURAL OPERATORS

First, we introduce neural operators following work of Lanthaler et al. (2025). Let $\Omega_{\text{in}} \subset \mathbb{R}^{d'_{\text{in}}}$ and $\Omega_{\text{out}} \subset \mathbb{R}^{d'_{\text{out}}}$ be Lipschitz domains and $\mathcal{F}_{\text{in}}$ and $\mathcal{F}_{\text{out}}$ be collections of functions $\Omega_{\text{in}} \to \mathbb{R}^{d_{\text{in}}}$ and $\Omega_{\text{out}} \to \mathbb{R}^{d_{\text{out}}}$, respectively. Moreover, for some $L \in \mathbb{N}$ and natural numbers $d_0, d_1, \dots, d_L$, we write $\mathcal{F}_i$ for a collection of functions $\Omega_{\text{in}} \to \mathbb{R}^{d_i}$. We define the following objects.

Definition 3.1 (Lifting layer)**.** Let $R\colon \mathbb{R}^{d'_{\text{in}}+d_{\text{in}}} \to \mathbb{R}^{d_0}$ be a neural network. A *lifting layer* is a map

$$\mathcal{R}\colon \mathcal{F}_{\text{in}} \to \mathcal{F}_0 \;\text{ with }\; \mathcal{R}(f)(x) = R\big(x, f(x)\big).$$

Definition 3.2 (Hidden layer)**.** The ith *hidden layer* is a map

$$\mathcal{L}_\ell\colon \mathcal{F}_{\ell-1} \to \mathcal{F}_\ell,$$
$$\mathcal{L}_\ell(f)(x) \coloneqq \sigma\left(W_\ell f(x) + b_\ell + \sum_{m=1}^{M_\ell} \langle T_{\ell,m} f, \psi_{\ell,m}\rangle \phi_{\ell,m}(x)\right),$$

where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a nonlinearity applied pointwise, $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{d_\ell}$ are weights and biases, and for each $m \in \{1, \dots, M_\ell\}$, we have that the $T_{\ell,m}\colon \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}$ are linear maps, the $\psi_{\ell,m} \in \mathcal{F}_\ell$ are *encoding functions*, and the $\phi_{\ell,m} \in \mathcal{F}_\ell$ are *decoding functions*. Finally, the brackets $\langle \cdot, \cdot \rangle$ denote the L^2 pairing for functions $\Omega_{\text{in}} \to \mathbb{R}^{d_\ell}$.

The collections of functions $\{T_{\ell,m} f\}_{m=1}^{M_\ell}$, $\{\phi_{\ell,m}\}_{m=1}^{M_\ell}$, and $\{\psi_{\ell,m}\}_{m=1}^{M_\ell}$ could be either learnable or pre-determined. In the latter case, one merely hard-codes them into the architecture, e.g., one could take $\psi_{\ell,m}(x) = \phi_{\ell,m}(x) = e^{\mathrm{i}m\cdot x}$ to be Fourier basis functions. In the former case, one may parametrize these functions by linear maps or neural networks and train their parameters.

Definition 3.3 (Decoding and projection Layers)**.** Fix $d_{L+1} \in \mathbb{N}$. A *decoding layer* is a map

$$\mathcal{D}\colon \mathcal{F}_L \to \mathbb{R}^{d_{L+1}} \quad \text{with} \quad \mathcal{D}(f) \coloneqq \sigma\left(b + \frac{1}{|\Omega_{\text{in}}|}\int_{\Omega_{\text{in}}} \widehat{R}(z, f(z))\, dz\right)$$

where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a component-wise non-linearity, $b \in \mathbb{R}^{d_{L+1}}$ is a vector, and $\widehat{R}\colon \Omega_{\text{in}} \times \mathbb{R}^{d_L} \to \mathbb{R}^{d_{L+1}}$ is a standard neural network. For a neural network $Q\colon \Omega_{\text{out}} \times \mathbb{R}^{d_{L+1}} \to \mathbb{R}^{d_{\text{out}}}$, a *$\mathcal{D}$-projection layer* is a map

$$\mathcal{Q}_{\mathcal{D}}\colon \mathcal{F}_{\text{L}} \to \mathcal{F}_{\text{out}} \;\text{ with }\; \mathcal{Q}_{\mathcal{D}}(f)(x) = Q\big(x, \mathcal{D}(f)\big).$$

Definition 3.4 (Neural operator)**.** A *neural operator (NO)* of depth $L > 0$ is a map of the form

$$\Psi\colon \mathcal{F}_{\text{in}} \to \mathcal{F}_{\text{out}} \quad \text{with} \quad f \mapsto \big(\mathcal{Q}_{\mathcal{D}} \circ \mathcal{L}_L \circ \cdots \circ \mathcal{L}_1 \circ \mathcal{R}\big)(f),$$

where $\mathcal{R}$ is a lifting layer as in Definition 3.1, each $\mathcal{L}_\ell$ is a hidden layer as in Definition 3.2, and $\mathcal{Q}_{\mathcal{D}}$ is a $\mathcal{D}$-projection layer as in Definition 3.3, depending on the decoding layer $\mathcal{D}$.

Definition 3.4 encompasses popular NOs such as Fourier neural operators (Li et al., 2021) and DeepONets (Lu et al., 2021). Key to our results is a generalization of NOs to mixed function-vector inputs (Bhattacharya et al., 2026; Huang et al., 2025), so we require an additional layer type.

Definition 3.5 (Vector-to-function layer)**.** Fix $d_{\text{vec}} \in \mathbb{N}$ and let $\Omega_{d_{\text{vec}}} \subset \mathbb{R}^{d_{\text{vec}}}$ be a subset. Let $h \in \mathcal{F}_{\text{in}}$ and $\psi\colon \mathbb{R}^{d_{\text{vec}}} \to \mathbb{R}$ be a neural network. A *vector-to-function layer* is a map of the form

$$V\colon \Omega_{d_{\text{vec}}} \to \mathcal{F}_{\text{in}} \;\text{ with }\; V(z)(x) \coloneqq h(x)\psi(z).$$

As above, the function h can be either learnable or pre-determined. We also accept the trivial linear neural network $\psi(z) = z$ as a valid choice for ψ.

Definition 3.6 (Augmented neural operator)**.** An *augmented neural operator* (AugNO) of depth $L > 0$ is a map of the form

$$\widetilde{\Psi} \colon \mathcal{F}_{\text{in}} \times \Omega_{d_{\text{vec}}} \to \mathcal{F}_{\text{out}} \quad \text{with} \quad (f, z) \mapsto \Psi\big(f \oplus V(z)\big),$$

where Ψ is a NO of depth L with input space $\mathcal{F}_{\text{in}}$ consisting of functions $\Omega_{\text{in}} \to \mathbb{R}^{d_{\text{in}}+d_{\text{in}}}$ and all remaining parameters as in Definition 3.4, V is a vector-to-function layer as in Definition 3.5, and $f \oplus V(z)$ denotes the concatenation of the maps f and $V(z)$, i.e., $(f \oplus V(z))(x) = (f(x), V(z)(x))$.

3.2 Operator approximation

Our first main result is an approximation theorem for the kernel conditioning operator $\mathcal{G}^\star$. To prove it, we combine the stability results from Section 4 with the Dugundji extension theorem (Dugundji, 1951) and a universal approximation theorem for neural operators (Lanthaler et al., 2025, Thm. 1, p. 267). All proofs can be found in Section C.

Theorem 3.7 (Approximation of $\mathcal{G}^\star$)**.** *Fix $\delta > 0$ and let $K \subseteq \mathcal{X}_\delta$ be compact. For any $\varepsilon > 0$, there exists a neural operator Ψ with input and output spaces $\mathcal{F}_{\text{in}} = \mathcal{F}_{\text{out}} = C(D \times E)$ such that*

$$\sup_{\rho \in K} \|\mathcal{G}^\star(\rho) - \Psi(\rho)\|_{C(D\times E)} \leq \varepsilon. \tag{7}$$

Notice that the above approximation result holds uniformly over the set of densities K. That is, there exists a single neural operator that approximates conditioning over the entire compact set K with arbitrarily high accuracy. However, this engenders the question: what subsets K of $\mathcal{X}_\delta$ are compact? A representative example is given in Corollary 3.9.

We now move on to our second main result. It is an approximation theorem for the in-context conditioning operator, paralleling Theorem 3.7.

Theorem 3.8 (Approximation of $\Psi^\star$)**.** *For any $\delta > 0$, any compact set $K \subseteq \mathcal{X}_\delta$, and any $\varepsilon > 0$, there exists an augmented neural operator $\widetilde{\Psi}$ with input space $F_{\text{in}} = C(D \times E)$ and $\Omega_{d_{\text{vec}}} = E$ as well as output space $F_{\text{out}} = C(D)$, as in Definition 3.6, such that*

$$\sup_{(\rho,y)\in K\times E} \|\Psi^\star(\rho, y) - \widetilde{\Psi}(\rho, y)\|_{C(D)} \leq \varepsilon. \tag{8}$$

The proof of this result extends that of Theorem 3.7. The central difficulty is the fact that $\Psi^\star$ takes as input both a function and a vector. To address this, we *lift* $\Psi^\star$ to the operator $\widetilde{\Psi}^\star$ by identifying points $y \in \mathbb{E}$ with constant functions $c_y(z) = y$ for all $z \in \mathbb{R}^d$. We then use Theorem 4.4, establishing the continuity of $\Psi^\star$, together with the Dugundji extension theorem (Dugundji, 1951) and a theorem of Lanthaler et al. (2025, Thm. 1) to obtain a neural operator that approximates $\widetilde{\Psi}^\star$.

We close this subsection by giving a concrete example of a set of densities K that is compact in $\mathcal{X}_\delta$.

Corollary 3.9 (Approximation on Hölder compact sets)**.** *Fix $\alpha \in (0, 1]$, $R > 0$, and $\delta > 0$. Define*

$$\mathcal{H}^\alpha_R := \big\{ f \in C^{0,\alpha}(D \times E) \,\big|\, \|f\|_{C^{0,\alpha}(D\times E)} \leq R \big\} \quad \text{and} \quad \mathcal{S} := \mathcal{X}_\delta \cap \mathcal{H}^\alpha_R. \tag{9}$$

For any $\varepsilon > 0$, there exist a neural operator $\Psi \colon C(D \times E) \to C(D \times E)$ and an augmented neural operator $\widetilde{\Psi} \colon C(D \times E) \times E \to C(D)$ such that

$$\sup_{\rho\in\mathcal{S}} \|\mathcal{G}^\star(\rho) - \Psi(\rho)\|_{C(D\times E)} \leq \varepsilon \quad \text{and} \quad \sup_{(\rho,y)\in\mathcal{S}\times E} \|\Psi^\star(\rho, y) - \widetilde{\Psi}(\rho, y)\|_{C(D)} \leq \varepsilon.$$

The proof combines a classical compact embedding theorem for Hölder spaces with the preceding two approximation theorems. We note that Hölder densities form a rather large class. Indeed, any Lipschitz density on a compact set is, in particular, α-Hölder for any $\alpha \in (0, 1]$. Moreover, rational power functions $x \mapsto x^\alpha$ are α-Hölder, for any $\alpha \in (0, 1)$, whilst not being Lipschitz near the origin (Gilbarg & Trudinger, 2001, Chapter 4.1). Finally, sample paths of a Brownian motion are locally α-Hölder for any $\alpha \in (0, \frac{1}{2})$ (Le Gall, 2016, Chapter 2.2).

3.3 UNBOUNDED DOMAINS

In this section, we extend the kernel conditioning operator $\mathcal{G}^\star$ to densities defined over the entire space $\mathbb{R}^{m+r}$. To do so, we introduce some additional machinery. For any $n \in \mathbb{N}$, write $L^1_{\mathrm{ac}}(\mathbb{R}^n)$ for the space of all probability density functions. Further recall the notation $\operatorname{ess\,inf}$ for the *essential infimum* of a function in $\mathbb{R}^n$, i.e., the greatest lower bound that holds almost everywhere; see Section A. Now fix a bounded open set $B \subset \mathbb{R}^r$ and some $\delta > 0$. Paralleling the construction of $\mathcal{X}_\delta$, define

$$\mathcal{U}_{\delta,B} \equiv \mathcal{U}_\delta \coloneqq \left\{ f \in L^1_{\mathrm{ac}}(\mathbb{R}^m \times \mathbb{R}^r) \,\middle|\, \operatorname*{ess\,inf}_{y \in B} m_f(y) \geq \delta \right\}, \tag{10}$$

where $m_f(y) \coloneqq \int_{\mathbb{R}^m} f(x,y)\,dx$. Note that B must be bounded, otherwise ρ would not be a probability density because $\int \rho(x,y)\,dx\,dy = \int m_\rho(y)\,dy \geq \delta \cdot \infty$. Now for $\rho \in L^1_{\mathrm{ac}}(\mathbb{R}^m \times \mathbb{R}^r)$, we extend κ_ρ and $\mathcal{G}^\star$ in the natural way. That is, we let $\kappa_\rho \colon \mathbb{R}^m \times B \to \mathbb{R}_+$ with $\kappa_\rho(x,y) \coloneqq \rho(x \,|\, y)$ and $\mathcal{G}^\star \colon \mathcal{U}_\delta \to L^1(\mathbb{R}^m \times B)$ with $\mathcal{G}^\star(\rho) \coloneqq \kappa_\rho$.[1]

Now, we define a class of neural operators that map between function spaces on unbounded domains.

Definition 3.10 (FullNO)**.** For any $n \in \mathbb{N}$ and any measurable set $\Omega \subset \mathbb{R}^n$, define the *zero extension operator* $\mathcal{Z}_\Omega \colon L^1(\Omega) \to L^1(\mathbb{R}^n)$ by $\mathcal{Z}_\Omega(f)(x) = f(x)$ for $x \in \Omega$ and $\mathcal{Z}_\Omega(f)(x) = 0$ for $x \notin \Omega$. We also define the *restriction operator* $\mathcal{R}_\Omega \colon L^1(\mathbb{R}^n) \to L^1(\Omega)$ by $f \mapsto f|_\Omega$. Fix d and d' in $\mathbb{N}$ and measurable sets $\Omega \subset \mathbb{R}^d$ and $\Omega' \subset \mathbb{R}^{d'}$. A *full space neural operator (FullNO)* from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R}^{d'})$ with truncation sets Ω and Ω' is a map $\Psi \colon L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^{d'})$ of the form

$$\Psi = \mathcal{Z}_{\Omega'} \circ \Psi_0 \circ S \circ \mathcal{R}_\Omega, \tag{11}$$

where $\Psi_0 \colon L^1(\Omega) \to L^1(\Omega')$ is a NO in the sense of Definition 3.4 and $S \colon \mathbb{R} \to \mathbb{R}$ is a shallow ReLU neural network acting on an input function f pointwise, i.e., $S(f)(x) = S(f(x))$ for each x.

To prove an approximation theorem for $\mathcal{G}^\star$ over function domain $\mathbb{R}^{m+r}$, we combine Lemma 4.3 with a novel unbounded, entire space universal approximation theorem. Specifically, we prove that FullNOs can universally approximate continuous operators $F^\star$ between $L^1(\mathbb{R}^n)$ spaces; the estimate is uniform over an L^1-compact set K of functions. The proof consists of five main steps. First, we cut off functions in K at some $M > 0$ to obtain a set K_M bounded in L^∞. In the second and third steps, we show that we can introduce truncations sets Ω and Ω' and a map $\Psi_0 \colon L^1(\Omega) \to L^1(\Omega')$ such that Ψ_0 is a good approximation of $F^\star$ uniformly over K_M. In the fourth step, we invoke an existing universal approximation theorem to obtain a NO Ψ_{00} that approximates Ψ_0 well over K_M. In the last step, we compose Ψ_{00} with appropriate truncation and zero extension operators to obtain a full space NO Ψ that approximates $F^\star$ well over K. The details are in Section C.

Theorem 3.11 (Universal approximation for functions with unbounded domain)**.** *Let $F^\star \colon L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^{d'})$ be continuous. Let $\varepsilon > 0$ be arbitrary. Let $K \subset L^1(\mathbb{R}^d)$ be any compact set. Then there exist $R > 0$, $R' > 0$, and a full space neural operator $\Psi \colon L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^{d'})$ with $\Omega \coloneqq (-R,R)^d$ and $\Omega' \coloneqq (-R',R')^{d'}$ truncation sets as in (3.10) such that*

$$\sup_{\rho \in K} \|F^\star(\rho) - \Psi(\rho)\|_{L^1(\mathbb{R}^{d'})} \leq \varepsilon. \tag{12}$$

Corollary 3.12 (Approximation of $\mathcal{G}^\star$ over functions with unbounded domain)**.** *Let $\varepsilon > 0$ and $\delta > 0$ be arbitrary. Let $B \subset \mathbb{R}^r$ be bounded, open, and measurable. Let $\mathcal{U}_\delta = \mathcal{U}_{\delta,B}$ be as in (10). Let $K \subset \mathcal{U}_\delta$ be compact in $L^1(\mathbb{R}^m \times \mathbb{R}^r)$. Then there exist $R > 0$ and a full space neural operator $\Psi \colon L^1(\mathbb{R}^m \times \mathbb{R}^r) \to L^1(\mathbb{R}^m \times \mathbb{R}^r)$ depending on $\Omega = \Omega' \coloneqq (-R,R)^{m+r}$ as in (3.10) such that*

$$\sup_{\rho \in K} \|\mathcal{G}^\star(\rho) - \Psi(\rho)\|_{L^1(\mathbb{R}^m \times B)} \leq \varepsilon. \tag{13}$$

The above theorem allows us to obtain an approximation over sets of densities K that embed compactly in $\mathcal{U}_\delta$. One such example is given by bounded sets in *weighted Sobolev spaces*.[2] Indeed, if K is a bounded subset of $W^{1,1}_\alpha(\mathbb{R}^{m+r}) \coloneqq W^{1,1}(\mathbb{R}^{m+r}) \cap L^1_\alpha(\mathbb{R}^{m+r})$ then by a result of Brezis (2011, Corollary 4.27) the set $K \cap \mathcal{U}_\delta \subset \mathcal{U}_\delta$ is compact. In particular, the density of any Gaussian mixture lies in $W^{1,1}_\alpha(\mathbb{R}^{m+r})$ for any $\alpha > 0$. Thus, we obtain approximation of the conditioning operator over sets of Gaussian mixtures with bounded means and covariance matrices, the latter also spectrally bounded away from zero.

[1] Note that $\kappa_\rho \in L^1(\mathbb{R}^m \times B)$ due to the uniform lower bound on the marginals.

[2] We use the notation $L^1_\alpha(\mathbb{R}^{m+r}) \coloneqq \left\{ f \in L^1(\mathbb{R}^{m+r}) \,\middle|\, \int |f(x)|\,(1 + \|x\|^\alpha)\,dx < \infty \right\}$ for $\alpha > 0$.

4 CONTINUITY OF CONDITIONING

The key to establishing the approximation theorems of Section 3 is to show that the conditioning operators $\mathcal{G}^\star$ and $\Psi^\star$ are continuous. Section 4.1 establishes stability results for the conditioning operators, in particular implying that they are continuous. In Section 4.2, the conditioning operators are continuously extended beyond $\mathcal{X}_\delta$ to the space of *positive densities* and these extensions are shown to be maximal. The reader can refer to Section 2 for a reminder on the domains and codomains of $\mathcal{G}^\star$ and $\Psi^\star$ as well as for the topologies used. All proofs may be found in Section D.

4.1 STABILITY RESULTS

First, we establish the local Lipschitz stability for the kernel conditioning operator $\mathcal{G}^\star$ on the metric space $\mathcal{M}_\delta$ defined in (3). That is, $\mathcal{G}^\star$ is Lipschitz continuous in a neighborhood of any point in $\mathcal{X}_\delta$ equipped with the topology of uniform convergence.

Lemma 4.1 ($\mathcal{G}^\star$ is locally Lipschitz)**.** *For any $p \in \mathcal{X}_\delta$ and $q \in \mathcal{X}_\delta$, it holds that*

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{C(D\times E)} \leq \frac{1}{\delta}\left(1 + \frac{|D| \min\big(\|p\|_{C(D\times E)}, \|q\|_{C(D\times E)}\big)}{\delta}\right)\|p - q\|_{C(D\times E)}. \quad (14)$$

The result is elementary, relying on the uniform lower bound $\delta > 0$ on densities in $\mathcal{X}_\delta$ to avoid a division by zero in the definition of the conditional density. The following corollary delivering global Lipschitz stability on a restricted domain is immediate; cf. Remark B.2.

Corollary 4.2 ($\mathcal{G}^\star$ is Lipschitz on balls)**.** *Fix $R > 0$ and define*

$$\mathcal{B}_R \coloneqq \big\{f \in C(D \times E) \,\big|\, \|f\|_{C(D\times E)} \leq R\big\}.$$

Then, for any p and q in $\mathcal{B}_R \cap \mathcal{X}_\delta$, it holds that

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{C(D\times E)} \leq \frac{1}{\delta}\left(1 + \frac{|D|R}{\delta}\right)\|p - q\|_{C(D\times E)}. \quad (15)$$

Finally, we can get an analogous result to Lemma 4.1 by changing the underlying function spaces.

Lemma 4.3 ($\mathcal{G}^\star$ is Lipschitz on functions with unbounded domain)**.** *Let $\mathcal{U}_\delta$ be as in (10). For any $p \in \mathcal{U}_\delta$ and $q \in \mathcal{U}_\delta$, it holds that*

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{L^1(\mathbb{R}^m\times B)} \leq \frac{2}{\delta}\|p - q\|_{L^1(\mathbb{R}^m\times\mathbb{R}^r)}. \quad (16)$$

We now turn our attention to the in-context conditioning operator $\Psi^\star$. The analysis of this operator is slightly more complicated than that of $\mathcal{G}^\star$. The next result shows that $\Psi^\star$ is continuous.

Theorem 4.4 (Continuity of $\Psi^\star$)**.** *The map $\Psi^\star$ in (4) is continuous.*

The main idea of the proof is to take a convergent sequence $(p_n, y_n) \to (p, y)$ in $\mathcal{X}_\delta \times E$ and use a triangle inequality to bound the error $\|\Psi^\star(p_n, y_n) - \Psi^\star(p, y)\|_{C(D)}$ by the sum of two terms: $\|\Psi^\star(p_n, y_n) - \Psi^\star(p, y_n)\|_{C(D)}$ and $\|\Psi^\star(p, y_n) - \Psi^\star(p, y)\|_{C(D)}$. The first term can be made small by taking a supremum over all $y \in E$ and using Lemma 4.1; the the second term can be made small by an elementary argument, using the uniform lower bound $\delta > 0$ as well as the fact that continuous densities p_n on the compact set $D \times E$ are uniformly continuous and hence $p_n \to p$ implies that marginals converge $m_{p_n} \to m_p$ as $n \to \infty$.

We can upgrade the qualitative continuity result of Theorem 4.4 to a quantitative Hölder stability estimate for $\Psi^\star$ by further restricting its domain. This is the content of the following corollary.

Corollary 4.5 (Hölder stability of $\Psi^\star$ on balls)**.** *Let $\alpha \in (0, 1]$. Fix $R > 0$ and define*

$$\mathcal{H}^\alpha_R \coloneqq \big\{f \in C^{0,\alpha}(D \times E) \,\big|\, \|f\|_{C^{0,\alpha}(D\times E)} \leq R\big\}.$$

For any (p, y) and (q, y') belonging to $(\mathcal{X}_\delta \cap \mathcal{H}^\alpha_R) \times E$, it holds that

$$\|\Psi^\star(p, y) - \Psi^\star(q, y')\|_{C(D)} \leq \frac{(1 + 2R)}{\delta}\left(1 + \frac{|D|R}{\delta}\right)\big\|(p, y) - (q, y')\big\|^\alpha_{C(D\times E)\times E}. \quad (17)$$

We note that the theorem remains true under slightly relaxed assumptions. Namely, one can ask for a continuous $f \in C(D \times E)$ and for $(x, y) \in D \times E$, ask for f to be α-Hölder continuous in y uniformly in x. See Remark E.2 for a more precise statement.

4.2 BEYOND UNIFORM LOWER BOUNDS

Let $\mathcal{X}_+ \coloneqq \bigcup_{\delta>0} \mathcal{X}_\delta$ be the union of all $\mathcal{X}_\delta$ for $\delta > 0$. It is not hard to see that this is exactly the set of densities with everywhere positive marginals, i.e., $\mathcal{X}_+ = \{\rho \in C_{\text{ac}}(D \times E) \,|\, m_\rho > 0\}$, see Lemma D.2. We can use this observation to show that $\mathcal{G}^\star$ and $\Psi^\star$ extend continuously to $\mathcal{X}_+$.

Theorem 4.6 (Continuous extension of conditioning operators)**.** *The operators $\mathcal{G}^\star$ and $\Psi^\star$ admit unique continuous extensions to the domains $\mathcal{X}_+$ and $\mathcal{X}_+ \times E$, denoted by $\mathcal{G}^\star_+$ and $\Psi^\star_+$, respectively.*

The key proof idea is to recognize that any convergent sequence $\rho_n \to \rho$ in $\mathcal{X}_+$ must eventually be contained in some $\mathcal{X}_\delta$ for $\delta > 0$. Then, we can use the continuity of $\mathcal{G}^\star$ and $\Psi^\star$ on each $\mathcal{X}_\delta$, shown in Section 4, to obtain the convergence $\Psi^\star(\rho_n, y_n) \to \Psi^\star(\rho, y)$ as $n \to \infty$ and similarly for $\mathcal{G}^\star$.

Finally, we can use the density of $\mathcal{X}_+$ in $C_{\text{ac}}(D \times E)$, established in Lemma D.3, to show that the extensions obtained above are, in fact, maximal.

Theorem 4.7 (Maximality of extension)**.** *Let $\mathcal{G}^\star_+$ and $\Psi^\star_+$ be the extension operators obtained in Theorem 4.6. There is no continuous extension of $\mathcal{G}^\star_+$ to any strict superset S of $\mathcal{X}_+$ in $C_{\text{ac}}(D \times E)$, that is, any set S satisfying $\mathcal{X}_+ \subsetneq S \subseteq C_{\text{ac}}(D \times E)$. Similarly, there is no continuous extension of $\Psi^\star_+$ to a set of the form $S' \times E$ where S' is some strict superset of $\mathcal{X}_+$ in $C_{\text{ac}}(D \times E)$.*

Note that the above result pertains to the $C(D \times E)$ norm, i.e., the topology of uniform convergence. The question of obtaining continuous conditioning operators on larger sets of measures under *different topologies* remains open. Nonetheless, we note that $\Psi^\star$ cannot be extended continuously to the full space of probability measures over $\mathbb{R}^{m+r}$ when equipped with the topology of *weak convergence*. One can show this by working with discrete measures and their conditionals, see Example E.3. This suggests that departing from the setting of probability measures with Lebesgue densities poses fundamental difficulties even for the simplest of examples.

5 NUMERICAL EXPERIMENTS

In this section, we demonstrate our results empirically by showcasing the approximation capabilities of neural operators for probabilistic conditioning. We train a Fourier neural operator (FNO) (Li et al., 2021) and a transformer neural operator (TNO) (Calvello et al., 2025a) to approximate the kernel conditioning operator $\mathcal{G}^\star$. We consider K-component Gaussian mixture joint distributions, for which a closed-form conditional kernel may be found. Section F.1 discusses our experimental setup, Section F.2 discusses the data generation process, and Section F.3 discusses architectural and training details. In Table 1, we report the results obtained by applying a trained TNO and FNO to a test set not used for training; Figure 1 illustrates the result of TNO on the median error sample for the experiment corresponding to $K = 3$.

The operator learning framework discussed in this paper leads to accuracy benefits compared to a traditional plug-in estimation approach, involving numerical integration for marginalization. This is demonstrated by deploying the trained models on kernel density estimate (KDE) joints in both experimental settings. Figure 2 showcases, in the context of $K = 1$, both the improvement in accuracy given by the TNO over the plug-in estimator in the test set for the KDE setting, as well as the predictions for the median error sample. Table 2 summarizes the accuracy results in both KDE setups. In the $K = 1$ setting, the learned models substantially outperform the plug-in baseline. In the more challenging $K = 3$ setting, learned models achieve lower median error but higher maximum error than the plug-in baseline. We expect more improvement with larger training sets and training data that includes KDEs as input functions. These results suggest that the conditioning operator may be learned with potential accuracy gains to traditional statistical approaches.

We note, however, that these experiments are intended as a proof of concept rather than a practical proposal. Their purpose is twofold: first, to numerically illustrate the theory developed, confirming that operator learning is a principled approach to kernel conditioning, and second, to motivate further methodological development. Indeed, grid discretization and the required density access pose limitations to high-dimensional scalability. This is discussed further in Section 6.

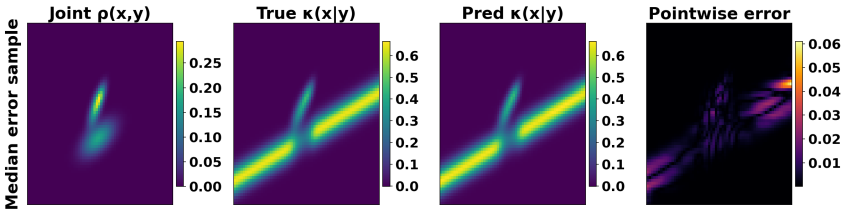

Figure 1: Trained TNO prediction on the median relative L^1 error sample from the Gaussian mixture test set.

Table 1: Accuracy in relative L^1 error achieved by the models on a test set not used for training.

Joint Class	FNO Med.	TNO Med.	FNO Max.	TNO Max.	FNO Params	TNO Params
Corr. Gaussian ($K = 1$)	0.0004	0.0006	0.0299	0.2683	9.6M	1.6M
Gaussian Mixture ($K = 3$)	0.0158	0.0251	0.7384	0.8938	9.6M	3.2M

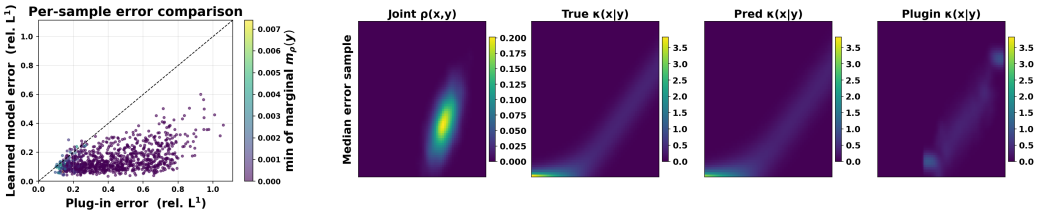

Figure 2: Per-sample relative L^1 test error comparison between plug-in estimator and learned model on the correlated Gaussian KDE test set (left). Predictions on the learned model median test error sample (right).

Table 2: Accuracy in relative L^1 error achieved by models and the plugin estimator on a KDE test set.

Joint Input	FNO Med.	TNO Med.	Plugin Med.	FNO Max.	TNO Max.	Plugin Max.
KDE Corr. Gaussian	0.1183	0.1244	0.4012	0.5822	0.6004	1.0597
KDE Gaussian Mixture	0.4829	0.4446	0.5079	1.1455	1.2511	0.9993

6 Discussion

In this paper, we introduce an operator learning framework for the conditioning of probability densities. Namely, we identify metric spaces of functions such that the maps $\rho \mapsto \kappa_\rho$, the *kernel conditioning operator*, and $(\rho, y) \mapsto \rho(\cdot \mid y)$, the *in-context conditioning operator*, are continuous. Continuity allows us to approximate these maps by neural operators up to arbitrary accuracy and uniformly over compact sets of densities. Our results lay the groundwork for a model for amortized probability density conditioning. That is, a model that has been trained on a large corpus of joint and conditional density pairs and can perform conditional inference on any unseen density within a class of admissible inputs, e.g., a bounded set of Hölder densities over a compact domain.

Open questions include:

1. *Development of scalable algorithms for operator learning in high dimensions*: Since neural operators receive and return density functions, they must be combined with methods that allow for efficient function representation in high dimensions, beyond simple grid evaluation.
2. *Learning from samples*: Commonly one has *sample access* as opposed to *density access* to the probability distribution of interest. Thus, classical density estimation or modern generative methods can be combined with neural operators to obtain pipelines for conditioning from point clouds of samples, or, equivalently, empirical measures.
3. *Conditioning of generic probability measures*: With an eye towards further approximation theorems, we wish to study the continuity—or lack thereof—of the conditioning operators over classes of more generic probability measures, e.g., discrete measures or singular measures.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, implement methods, clean and reformat datasets, provide critical ingredients for proving mathematical claims and check the final draft for typos and other issues. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Moreover, we did not use generative AI to assist with translation or support qualitative and thematic data analysis, as these tasks are not applicable to this work.

All portions of mathematical statements and proofs suggested 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.

REPRODUCIBILITY STATEMENT

Proofs are provided in the appendix, with Section B, Section C and Section D corresponding to the proofs of the results stated in Section 2, Section 3 and Section 4 of the main text, respectively. Moreover, all relevant notation is introduced in Section A of the appendix. Lastly, details regarding the numerical experiments are given in Section F of the appendix.

REFERENCES

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2 edition, 2003. (Cited on page 18)

Mohammad Al-Jarrah, Bamdad Hosseini, and Amirhossein Taghvaei. Fast filtering of non-Gaussian models using amortized optimal transport maps. *IEEE Control Systems Letters*, 2025. (Cited on page 2)

Eviatar Bach, Ricardo Baptista, Edoardo Calvello, Bohan Chen, and Andrew Stuart. Learning enhanced ensemble filters. *Journal of Computational Physics*, art. 114550, 2025. (Cited on page 2)

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of Computational Mathematics*, 7:121–157, 2021. (Cited on page 2)

Kaushik Bhattacharya, Lianghao Cao, George Stepaniants, Andrew M Stuart, and Margaret Trautner. Learning memory and material dependent constitutive laws. *The SMAI Journal of Computational Mathematics*, 12:219–267, 2026. (Cited on page 4)

Brianna Binder, Agnimitra Dasgupta, and Assad Oberai. Closed-form conditional diffusion models for data assimilation. *preprint arXiv:2603.21291*, 2026. (Cited on page 2)

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. (Cited on page 2)

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D Collins, Michael S Pritchard, and Alexander Keller. FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *preprint arXiv:2507.12144*, 2025. (Cited on page 2)

Nicolas Boullé and Alex Townsend. A mathematical guide to operator learning. In Siddhartha Mishra and Alex Townsend (eds.), *Handbook of Numerical Analysis*, volume 25, pp. 83–125. Elsevier, 2024. (Cited on page 2)

Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 1. Springer New York, NY, 2011. (Cited on pages 6, 19, and 20)

Ronald Brown. Function spaces and product topologies. *The Quarterly Journal of Mathematics*, 15 (1):238–250, 1964. (Cited on page 27)

Simone Brugiapaglia, Nicola Rares Franco, and Nicholas H Nelsen. A short tour of operator learning theory: Convergence rates, statistical limits, and open questions. *preprint arXiv:2603.00819*, 2026. (Cited on page 2)

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, 2025a. (Cited on pages 2, 8, and 28)

Edoardo Calvello, Sebastian Reich, and Andrew M Stuart. Ensemble Kalman methods: A mean-field perspective. *Acta Numerica*, 34:123–291, 2025b. (Cited on page 2)

Edoardo Calvello, Elizabeth Carlson, Nikola Kovachki, Michael N Manta, and Andrew M Stuart. Operator learning for smoothing and forecasting. *preprint arXiv:2603.20359*, 2026. (Cited on page 2)

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *International Conference on Machine Learning*, pp. 5817–5840. PMLR, 2024. (Cited on page 2)

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995. (Cited on page 2)

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. (Cited on page 2)

Tiangang Cui, Xiaodong Feng, Chenlong Pei, Xiaoliang Wan, and Tao Zhou. Amortized filtering and smoothing with conditional normalizing flows. *preprint arXiv:2604.07169*, 2026. (Cited on page 2)

Maarten V de Hoop, Nikola B Kovachki, Matti Lassas, and Nicholas H Nelsen. Extension and neural operator approximation of the electrical impedance tomography inverse map. *preprint arXiv:2511.20361*, 2025. (Cited on pages 2 and 18)

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. (Cited on page 2)

James Dugundji. An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1(3):353–367, 1951. (Cited on pages 5, 16, and 20)

Demián Fraiman. On the expressive power of contextual relations in transformers. *preprint arXiv:2603.25860*, 2026. (Cited on page 2)

Takashi Furuya, Maarten V de Hoop, and Matti Lassas. Transformers through the lens of support-preserving maps between measures. *preprint arXiv:2509.25611*, 2025a. (Cited on page 2)

Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, 2025b. (Cited on page 2)

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. (Cited on page 2)

David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer, 2001. (Cited on page 5)

Manuel Gloeckler, Michael Deistler, Christian Dietrich Weilbach, Frank Wood, and Jakob H Macke. All-in-one simulation-based inference. In *International Conference on Machine Learning*, pp. 15735–15766. PMLR, 2024. (Cited on page 2)

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024. (Cited on page 2)

Daniel Zhengyu Huang, Nicholas H Nelsen, and Margaret Trautner. An operator learning perspective on parameter-to-observable maps. *Foundations of Data Science*, 7(1):163–225, 2025. (Cited on page 4)

Vasily Ilin and Peter Sushko. DiScoFormer: Plug-in density and score estimation with transformers. In *International Conference on Machine Learning*, 2026. (Cited on page 2)

Ryotaro Kawata and Taiji Suzuki. Transformers as measure-theoretic associative memory: A statistical perspective and minimax optimality. In *International Conference on Learning Representations*, 2026. (Cited on page 2)

Jean Kossaifi, Nikola Kovachki, Zongyi Li, David Pitt, Miguel Liu-Schiaffini, Robert Joseph George, Boris Bonev, Kamyar Azizzadenesheli, Julius Berner, Valentin Duruisseaux, and Anima Anandkumar. A library for learning neural operators. *Journal of Machine Learning Research*, 27 (205):1–6, 2026. (Cited on page 28)

Nikola B Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. (Cited on page 2)

Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. In Siddhartha Mishra and Alex Townsend (eds.), *Handbook of Numerical Analysis*, volume 25, pp. 419–467. Elsevier, 2024. (Cited on page 2)

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pp. 1–11, 2023. (Cited on page 2)

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and its Applications*, 6(1):tnac001, 2022. (Cited on page 2)

Samuel Lanthaler, Zongyi Li, and Andrew M Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62(2):261–303, 2025. (Cited on pages 2, 4, 5, 16, 17, and 19)

Hugo Lavenant and Giuseppe Savaré. Continuous transformations of probability measures and their transport representations. *preprint arXiv:2604.16653*, 2026. (Cited on page 2)

Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016. (Cited on page 5)

Uriel Martínez León and Jonathan Niles-Weed. Wasserstein least squares: A canonical regression method for probability distributions. *preprint arXiv:2605.30266*, 2026. (Cited on page 2)

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4433–4445, 2022. (Cited on page 2)

Zongyi Li, Nikola B Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. (Cited on pages 2, 4, and 8)

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3D PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35836–35854, 2023. (Cited on page 2)

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. (Cited on page 2)

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. (Cited on pages 2 and 4)

Liyao Lyu, Xinyue Yu, and Hayden Schaeffer. MVNN: A measure-valued neural network for learning McKean-Vlasov dynamics from particle data. *preprint arXiv:2604.00333*, 2026. (Cited on page 2)

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via measure transport: An introduction. *Springer Books*, pp. 785–825, 2017. (Cited on page 2)

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *preprint arXiv:1411.1784*, 2014. (Cited on page 2)

Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023. (Cited on page 2)

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. (Cited on page 2)

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020. (Cited on page 2)

Emile Pierret, Valentine Tosel, Julie Delon, and Alasdair Newson. Flow matching for applied mathematicians, 2026. (Cited on page 2)

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adria Lopez Escoriza, Ruud JG Van Sloun, and Yonina C Eldar. KalmanNet: Neural network aided Kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022. (Cited on page 2)

Daniel Sanz-Alonso, Andrew Stuart, and Armeen Taeb. *Inverse problems and data assimilation*, volume 107. Cambridge University Press, 2023. (Cited on page 2)

Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, and Krikamol Muandet. Kernel conditional density operators. In *International Conference on Artificial Intelligence and Statistics*, pp. 993–1004. PMLR, 2020. (Cited on page 2)

Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015. (Cited on page 2)

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions. In *International Conference on Machine Learning*, 2009. (Cited on page 2)

Andrew M Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. (Cited on page 2)

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. (Cited on page 2)

Xin T Tong, Yanyan Wang, and Liang Yan. Latent autoencoder ensemble Kalman filter for data assimilation. *preprint arXiv:2603.06752*, 2026. (Cited on page 2)

Matthew Vandergrift, Martha White, Yury Polyanskiy, Philippe Rigollet, and Lazar Atanackovic. Measure-to-measure regression with transformers. In *The 2026 Workshop on Generative and Agentic AI for Biology*, 2026. (Cited on page 2)

George Whittle, Juliusz Ziomek, Jacob Rawling, and Michael A Osborne. Distribution transformers: Fast approximate Bayesian inference with on-the-fly prior adaptation. In *International Conference on Machine Learning*, 2026. (Cited on page 2)

Keyi Wu, Peng Chen, and Omar Ghattas. A fast and scalable computational framework for large-scale high-dimensional Bayesian optimal experimental design. *SIAM/ASA Journal on Uncertainty Quantification*, 11(1):235–261, 2023. (Cited on page 2)

Chenyu Zeng, Yanshu Zhang, Jiayi Zhou, Yuhan Wang, Zilin Wang, Yuhao Liu, Lei Wu, and Daniel Zhengyu Huang. Point cloud neural operator for parametric PDEs on complex and variable geometries. *Computer Methods in Applied Mechanics and Engineering*, 443:118022, 2025. (Cited on page 2)

Technical Appendices and Supplementary Material for:
One Operator for Many Densities: Amortized
Approximation of Conditioning by Neural Operators

A Notation

This appendix introduces notation from analysis and probability.

A.1 Analysis

Let Y be a normed vector space and let $C(X;Y)$ denote the vector space of continuous functions from X to Y. For any f, define the supremum norm

$$\|f\|_{C(X;Y)} \coloneqq \sup_{x\in X} \|f(x)\|_Y. \tag{18}$$

The norm (18) induces the metric

$$\mathsf{d}_{C(X;Y)}(f,g) \coloneqq \|f-g\|_{C(X;Y)} \tag{19}$$

for any f and g. When $Y=\mathbb{R}$, we write $C(X;\mathbb{R}) = C(X)$. When Y is a normed vector space, we equip the product space $X\times Y$ with the norm $\|(x,y)\|_{X\times Y} \coloneqq \max(\|x\|_X, \|y\|_Y)$. Hölder spaces play an important role in this work. For $\alpha\in(0,1]$ and a compact set $B\subset\mathbb{R}^k$, define the norm

$$\|f\|_{C^{0,\alpha}(B)} \coloneqq \max\left(\|f\|_{C(B)}, \sup_{\substack{(x,x')\in B\times B\\ x\neq x'}} \frac{|f(x)-f(x')|}{|x-x'|^\alpha}\right). \tag{20}$$

Then define

$$C^{0,\alpha}(B) \coloneqq \left\{f\in C(B) \,\middle|\, \|f\|_{C^{0,\alpha}(B)} < \infty\right\}. \tag{21}$$

Moreover, define the *essential infimum* as follows: fix $n\in\mathbb{N}$ and a measurable function $f\colon \mathbb{R}^n\to\mathbb{R}$. For $b\in\mathbb{R}$ write $L_b(f) \coloneqq \{x \mid f(x) < b\}$ for a sub-level set of f. Then the essential infimum is

$$\operatorname*{ess\,inf}_{x\in\mathbb{R}^n} f(x) \coloneqq \sup\{b\in\mathbb{R} \mid |L_b(f)| = 0\},$$

where $|\cdot|$ indicates the Lebesgue measure of a set.

A.2 Additional notation

Let $D\subset\mathbb{R}^m$ and $E\subset\mathbb{R}^r$ be compact sets. Then $(C(D), \|\cdot\|_{C(D)})$ and $(C(D\times E), \|\cdot\|_{C(D\times E)})$ are Banach spaces. For any compact $B\subset\mathbb{R}^k$, define the subset

$$C_{\mathrm{ac}}(B) \coloneqq \left\{\rho\in C(B) \,\middle|\, \int_B \rho(u)\,du = 1 \quad\text{and}\quad \inf_{u\in B}\rho(u)\geq 0\right\} \subset C(B). \tag{22}$$

We view $C_{\mathrm{ac}}(B)$ in (22) as the set of continuous Lebesgue probability density functions on B.

B Proofs for Section 2: Formulation

The following lemma shows that $\mathcal{M}_\delta$ and $\mathcal{M}_\delta^E$ as defined in (3) are complete. In other words, $\mathcal{X}_\delta$ and $\mathcal{X}_\delta\times E$ are closed in $C(D\times E)$ and $C(D\times E)\times E$, respectively, with respect to the supremum norm and the product norm, discussed beneath (3). In fact, we only show that $\mathcal{M}_\delta$ is closed; then $\mathcal{M}_\delta^E$ is closed as the Cartesian product of closed sets.

Lemma B.1 (Completeness)**.** *It holds that $\mathcal{M}_\delta$ as defined in (3) is a complete metric space.*

Proof. Fix $\delta>0$ and set

$$\mathcal{L}_\delta \coloneqq \left\{f\in C(D\times E) \,\middle|\, \inf_{y\in E} m_\rho(y)\geq\delta\right\}.$$

Note that we can write

$$\mathcal{X}_\delta = C_{\mathrm{ac}}(D \times E) \cap \mathcal{L}_\delta.$$

It suffices to show that both sets in the above intersection are closed in $C(D \times E)$ with respect to the topology of uniform convergence. First, we show that $\mathcal{L}_\delta$ is sequentially closed in $C(D \times E)$. Let $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{L}_\delta$ be a sequence such that $f_n \to f \in C(D\times E)$ in $C(D\times E)$ as $n \to \infty$. Fix $y \in E$. It follows that $f_n(\,\cdot\,, y) \to f(\,\cdot\,, y)$ uniformly and hence pointwise in D. Since $\sup_{n\in\mathbb{N}}\|f_n\|_{C(D\times E)} < \infty$, the dominated convergence theorem applied to $\{f_n(\,\cdot\,, y)\}_{n\in\mathbb{N}}$ gives

$$\begin{aligned} m_f(y) = \int_D f(x,y)\,dx &= \liminf_{n\to\infty} \int_D f_n(x,y)\,dx \\ &\geq \liminf_{n\to\infty} \inf_{y'\in E} \int_D f_n(x,y')\,dx \\ &\geq \delta. \end{aligned}$$

Since y is arbitrary, it follows that $\mathcal{L}_\delta$ is closed.

Next, for any $n \in \mathbb{N}$ and compact $B \subset \mathbb{R}^n$ we show that the set

$$C_{\mathrm{ac}}(B) = \left\{ f \in C(B) \,\middle|\, \int_B f(x)\ dx \quad \text{and} \quad \inf_{x\in B} f(x) \geq 0 \right\}.$$

is closed in $C(B)$. The assertion of the lemma follows by taking $B = D \times E$. To this end, let $\{\rho_n\}_{n\in\mathbb{N}} \subset C_{\mathrm{ac}}(B)$ be a sequence such that $\rho_n \to \rho \in C(B)$ in $C(B)$ as $n \to \infty$. We must show that $\rho \in C_{\mathrm{ac}}(B)$. Since ρ_n converges to ρ uniformly, it also converges pointwise. Moreover, there exists $0 < c < \infty$ such that $\sup_{n\in\mathbb{N}}\|\rho_n\|_{C(B)} \leq c$. Now fix $u \in B$. Then for all $n \in \mathbb{N}$, it holds that

$$\rho(u) \geq \rho_n(u) - |\rho_n(u) - \rho(u)| \geq -|\rho_n(u) - \rho(u)|.$$

We deduce that $\rho(u) \geq -\limsup_{n\to\infty}|\rho_n(u) - \rho(u)| = 0$ as required. Next, we see that ρ_n is pointwise uniformly bounded in n by c. An application of the dominated convergence theorem yields

$$\int_B \rho(u)\,du = \lim_{n\to\infty} \int_B \rho_n(u)\,du = 1.$$

This completes the proof. □

Remark B.2 (Uniform supremum norm upper bound)**.** A similar argument shows that for any B and any $R > 0$, the set $\mathcal{B}_R(B) := \{\rho \in C(B) \,|\, \sup_{x\in B}|\rho(x)| \leq R\}$ is closed in $C(B)$. Thus, $\mathcal{X}_\delta \cap \mathcal{B}_R(D \times E)$ is also closed in $C(D \times E)$. This set is useful because it will be shown that one formulation of distribution conditioning is Lipschitz on $\mathcal{X}_\delta \cap \mathcal{B}_R(D \times E)$.

C Proofs for Section 3: Main results

This appendix collects the proofs of the main results.

C.1 Proofs for Subsection 3.2: Operator approximation

Theorem 3.7 (Approximation of $\mathcal{G}^\star$)**.** *Fix $\delta > 0$ and let $K \subseteq \mathcal{X}_\delta$ be compact. For any $\varepsilon > 0$, there exists a neural operator Ψ with input and output spaces $\mathcal{F}_{\text{in}} = \mathcal{F}_{\text{out}} = C(D \times E)$ such that*

$$\sup_{\rho\in K} \|\mathcal{G}^\star(\rho) - \Psi(\rho)\|_{C(D\times E)} \leq \varepsilon. \tag{7}$$

Proof. By Lemma 4.1, $\mathcal{G}^\star\colon \mathcal{X}_\delta \subset C(D \times E) \to C(D \times E)$ is continuous. Since $K \subseteq \mathcal{X}_\delta$, the map $F^\star := \mathcal{G}^\star|_K\colon K \subset C(D \times E) \to C(D \times E)$ is also continuous. Since K is compact, K is closed. By the Dugundji extension theorem (Dugundji, 1951, Thm. 4.1, p. 357), there exists a continuous map $\widetilde{F}^\star\colon C(D \times E) \to C(D \times E)$ that extends $F^\star$. An application of results due to Lanthaler et al. (2025, Thm. 1 and Cor. 1, p. 267) to $\widetilde{F}^\star$ delivers a neural operator Ψ such that (7) holds. □

Theorem 3.8 (Approximation of $\Psi^\star$)**.** *For any $\delta > 0$, any compact set $K \subseteq \mathcal{X}_\delta$, and any $\varepsilon > 0$, there exists an augmented neural operator $\widetilde{\Psi}$ with input space $F_{\text{in}} = C(D \times E)$ and $\Omega_{d_{\text{vec}}} = E$ as well as output space $F_{\text{out}} = C(D)$, as in Definition 3.6, such that*

$$\sup_{(\rho,y)\in K\times E} \|\Psi^\star(\rho, y) - \widetilde{\Psi}(\rho, y)\|_{C(D)} \leq \varepsilon. \tag{8}$$

Proof. For any $y \in E$, consider the constant function

$$c_y \colon D \times E \to \mathbb{R}^r, \quad \text{where} \quad c_y(x, z) = y \quad \text{for all} \quad (x, z) \in D \times E.$$

Now define

$$\mathrm{Dom}(\widetilde{\Psi}^\star) \coloneqq \{\rho \oplus c_y \,|\, (\rho, y) \in \mathrm{Dom}(\Psi^\star)\} \subset C(D \times E; \mathbb{R}^{r+1}),$$

where we are using notation $\oplus$ to denote concatenation of functions, that is[3]

$$(\rho \oplus c_y)(x, z) \coloneqq \big(\rho(x, z), c_y(x, z)\big).$$

Now let the lift $\widetilde{\Psi}^\star$ of our operator $\Psi^\star$ be given by

$$\begin{aligned} \widetilde{\Psi}^\star \colon \mathrm{Dom}(\widetilde{\Psi}^\star) &\to C(D) \\ \rho \oplus c_y &\mapsto \Psi^\star(\rho, y). \end{aligned}$$

Since Theorem 4.4 furnishes the continuity of $\Psi^\star$ and since the mapping

$$\begin{aligned} \iota \colon E &\to C(D \times E; \mathbb{R}^r) \\ y &\mapsto c_y \end{aligned}$$

is continuous, the lift

$$\widetilde{\Psi}^\star = \Psi^\star \circ (\mathrm{id} \times \iota)$$

is continuous as a composition of continuous maps, where we have introduced the map

$$\mathrm{id} \times \iota \colon (\rho, y) \mapsto \rho \oplus c_y.$$

Now, define the lift $\widetilde{E}$ of the compact set E by

$$\widetilde{E} \coloneqq \iota(E) = \{c_y \,|\, y \in E\} \subset C(D \times E; \mathbb{R}^r).$$

Since ι is continuous and E is compact, so is $\widetilde{E}$ and hence the Cartesian product

$$K \times \widetilde{E} \subset C(D \times E; \mathbb{R}^{r+1})$$

is also compact; see Footnote 3 for a note on the implicit function space identification occurring here. Following the proof of Theorem 3.7, we can first restrict $\widetilde{\Psi}^\star$ to the compact set $K \times \widetilde{E}$ and then apply the Dugundji extension theorem to obtain a continuous extension

$$\widetilde{\Psi}^\star_0 \colon C(D \times E; \mathbb{R}^{r+1}) \to C(D)$$

satisfying

$$\widetilde{\Psi}^\star_0\big|_{K\times\widetilde{E}} = \widetilde{\Psi}^\star\big|_{K\times\widetilde{E}}. \tag{23}$$

We can finally apply a modification to a theorem of Lanthaler et al. (2025, Theorem 1, p. 267) for distinct input and output domains, discussed by Lanthaler et al. (2025, p. 270), to obtain a neural operator $\widetilde{\Psi}_0$ such that

$$\sup_{(\rho,c_y)\in K\times\widetilde{E}} \|\widetilde{\Psi}^\star_0(\rho, c_y) - \widetilde{\Psi}_0(\rho, c_y)\|_{C(D)} \leq \varepsilon.$$

[3] Here, and for the rest of the proof we identity $C(D \times E) \times C(D \times E; \mathbb{R}^r)$ with $C(D \times E; \mathbb{R}^{r+1})$ via the continuous embedding

$$\begin{aligned} J \colon C(D \times E) \times C(D \times E; \mathbb{R}^r) &\to C(D \times E; \mathbb{R}^{r+1}) \\ (\rho, c_y) &\mapsto \rho \oplus c_y \colon \big((x, z) \mapsto (\rho(x, z), c_y(x, z))\big). \end{aligned}$$

Finally, consider the vector-to-function layer V given by

$$V\colon C(D\times E)\times E\to C(D\times E;\mathbb{R}^{r+1})$$
$$(\rho,y)\mapsto(\rho,c_y)$$

and note that it satisfies Definition 3.5 with $h\equiv 1$ and $\psi=\mathrm{id}$ the constant and identity maps, respectively. Now, define the augmented neural operator $\widetilde{\Psi}$ by

$$\widetilde{\Psi}=\widetilde{\Psi}_0\circ V.$$

By (23) and the construction of $\widetilde{E}$, $\widetilde{\Psi}^\star$, and V, we have

$$\begin{aligned}\sup_{(\rho,y)\in K\times E}\|\Psi^\star(\rho,y)-\widetilde{\Psi}(\rho,y)\|_{C(D)} &= \sup_{(\rho,y)\in K\times E}\|\Psi^\star(\rho,y)-(\widetilde{\Psi}_0\circ L)(\rho,y)\|_{C(D)}\\ &= \sup_{(\rho,c_y)\in K\times\widetilde{E}}\|\widetilde{\Psi}^\star(\rho,c_y)-\widetilde{\Psi}_0(\rho,c_y)\|_{C(D)}\\ &= \sup_{(\rho,c_y)\in K\times\widetilde{E}}\|\widetilde{\Psi}_0^\star(\rho,c_y)-\widetilde{\Psi}_0(\rho,c_y)\|_{C(D)}\\ &\le\varepsilon.\end{aligned}$$

The proof is complete. □

Corollary 3.9 (Approximation on Hölder compact sets)**.** *Fix $\alpha\in(0,1]$, $R>0$, and $\delta>0$. Define*

$$\mathcal{H}_R^\alpha:=\left\{f\in C^{0,\alpha}(D\times E)\,\middle|\,\|f\|_{C^{0,\alpha}(D\times E)}\le R\right\}\quad\text{and}\quad\mathcal{S}:=\mathcal{X}_\delta\cap\mathcal{H}_R^\alpha. \tag{9}$$

For any $\varepsilon>0$, there exist a neural operator $\Psi\colon C(D\times E)\to C(D\times E)$ and an augmented neural operator $\widetilde{\Psi}\colon C(D\times E)\times E\to C(D)$ such that

$$\sup_{\rho\in\mathcal{S}}\|\mathcal{G}^\star(\rho)-\Psi(\rho)\|_{C(D\times E)}\le\varepsilon\quad\text{and}\quad\sup_{(\rho,y)\in\mathcal{S}\times E}\|\Psi^\star(\rho,y)-\widetilde{\Psi}(\rho,y)\|_{C(D)}\le\varepsilon.$$

Proof. We show that $K:=\mathcal{X}_\delta\cap\mathcal{H}_R^\alpha\subset C(D\times E)$ is compact and then apply Theorems 3.7 and 3.8. By Lemma B.1, $\mathcal{X}_\delta$ is closed in $C(D\times E)$. An argument similar to the one used to prove a result of de Hoop et al. (2025, Lem. 4.4, pp. 25–26) shows that the set $\mathcal{H}_R^\alpha$ is also closed in $C(D\times E)$. Thus, $\mathcal{X}_\delta\cap\mathcal{H}_R^\alpha$ is closed. It remains to show that $\mathcal{H}_R^\alpha\supseteq\mathcal{X}_\delta\cap\mathcal{H}_R^\alpha$ is relatively compact in $C(D\times E)$. But this is a consequence of the compactness of the embedding $C^{0,\alpha}(D\times E)\hookrightarrow C^{0,\beta}(D\times E)$ for any $0<\beta<\alpha$ (Adams & Fournier, 2003, Thm. 1.34, pp. 11–12). □

C.2 Proofs for Subsection 3.3: Unbounded domains

Theorem 3.11 (Universal approximation for functions with unbounded domain)**.** *Let $F^\star\colon L^1(\mathbb{R}^d)\to L^1(\mathbb{R}^{d'})$ be continuous. Let $\varepsilon>0$ be arbitrary. Let $K\subset L^1(\mathbb{R}^d)$ be any compact set. Then there exist $R>0$, $R'>0$, and a full space neural operator $\Psi\colon L^1(\mathbb{R}^d)\to L^1(\mathbb{R}^{d'})$ with $\Omega:=(-R,R)^d$ and $\Omega':=(-R',R')^{d'}$ truncation sets as in (3.10) such that*

$$\sup_{\rho\in K}\|F^\star(\rho)-\Psi(\rho)\|_{L^1(\mathbb{R}^{d'})}\le\varepsilon. \tag{12}$$

Proof. The proof is divided into several steps.

Step 1. *Approximate K by another compact set K_M that is bounded in L^∞.*

Step 2. *Truncate the output space $\mathbb{R}^{d'}$ to a bounded domain Ω'.*

Step 3. *Truncate the input space $\mathbb{R}^d$ to a bounded domain Ω.*

Step 4. *Apply an existing universal approximation theorem for operators between bounded domain function spaces.*

Step 5. *Combine the estimates into a single error bound.*

We now proceed with this multi-step proof procedure.

Step 1. By Lemmas C.1 and C.2, we may choose $M = M(\varepsilon) > 0$ such that

$$\sup_{\rho \in K} \|F^\star(\rho) - F^\star(T_M(\rho))\|_{L^1(\mathbb{R}^{d'})} \leq \frac{\varepsilon}{2}, \tag{24}$$

where T_M is as in (29). Let $K_M := T_M(K) \subset L^1(\mathbb{R}^d)$, which is compact because $T_M \colon L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^d)$ is continuous by Lemma C.1. Moreover, K_M is bounded in $L^\infty(\mathbb{R}^d)$ by (30a).

Step 2. Since $F^\star$ is continuous, the set $K_M^\star := F^\star(K_M) \subset L^1(\mathbb{R}^{d'})$ is compact. By a corollary to the Frechét–Kolmogorov–Riesz (FKR) compactness theorem (Brezis, 2011, Exercise 4.34(3), p. 128–129), $K_M^\star$ is uniformly tight. Thus, we can find a compact set $A' = A'(\varepsilon) \subset \mathbb{R}^{d'}$ such that

$$\sup_{g \in K_M^\star} \int_{\mathbb{R}^{d'} \setminus A'} |g(x)| \, dx \leq \frac{\varepsilon}{4}. \tag{25}$$

By the compactness of A', we may choose $R' = R'(\varepsilon)$ large enough such that $\Omega' := (-R', R')^{d'} \supset A'$. Note that Ω' is an open bounded Lipschitz domain.

Step 3. By the same FKR compactness argument, for any $\delta > 0$, we can find a compact set $A = A(\delta, \varepsilon) \subset \mathbb{R}^d$ such that

$$\sup_{h \in K_M} \int_{\mathbb{R}^d \setminus A} |h(x)| \, dx \leq \delta. \tag{26}$$

By the compactness of A, we may choose $R = R(\delta, \varepsilon)$ large enough such that $\Omega := (-R, R)^d \supset A$. Notice that $\|h - \mathcal{Z}_\Omega \mathcal{R}_\Omega h\|_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d \setminus \Omega} |h(x)| \, dx \leq \int_{\mathbb{R}^d \setminus A} |h(x)| \, dx$. Then Lemma C.2 implies that we can choose $\delta = \delta(\varepsilon)$ in (26) such that for every $h \in K_M$, it holds that

$$\|\mathcal{R}_{\Omega'} F^\star(h) - \mathcal{R}_{\Omega'} F^\star(\mathcal{Z}_\Omega \mathcal{R}_\Omega h)\|_{L^1(\Omega')} \leq \|F^\star(h) - F^\star(\mathcal{Z}_\Omega \mathcal{R}_\Omega h)\|_{L^1(\mathbb{R}^{d'})} \leq \frac{\varepsilon}{8}. \tag{27}$$

Step 4. Now define $\Psi_0 := \mathcal{R}_{\Omega'} \circ F^\star \circ \mathcal{Z}_\Omega \colon L^1(\Omega) \to L^1(\Omega')$, which is a composition of continuous maps and is hence continuous itself. The set $\mathcal{R}_\Omega(K_M) \subset L^1(\Omega)$ is compact and is also bounded in $L^\infty(\Omega)$. By the universal approximation theorem for neural operators (Lanthaler et al., 2025, Thm. 2, Sec. 2.3, p. 267, applied with $s = s' = 0$, $p = p' = 1$, $k = k' = 1$), there exists a neural operator $\Psi_{00} \colon L^1(\Omega) \to L^1(\Omega')$ such that

$$\sup_{f \in \mathcal{R}_\Omega(K_M)} \|\Psi_0(f) - \Psi_{00}(f)\|_{L^1(\Omega')} \leq \frac{\varepsilon}{8}. \tag{28}$$

Step 5. Define $\Psi := \mathcal{Z}_{\Omega'} \circ \Psi_{00} \circ \mathcal{R}_\Omega \circ T_M$. Recall that $T_M(f)(x) = t_M(f(x))$, where $t_M(z) := \min(z, M \operatorname{sgn}(z))$ because f is real-valued. A direct calculation shows that

$$z \mapsto t_M(z) = \max(-M, \min(z, M)) = (z + M)_+ - (z - M)_+ - M$$

is a shallow ReLU neural network. We claim that Ψ has the form (3.10) with operator $\mathcal{S}$ associated to function $S := t_M$. Indeed, for any $x \in \Omega$, it holds that $(\mathcal{R}_\Omega \circ T_M)(f)(x) = t_M(f(x)) = t_M(R_\Omega f(x)) = (\mathcal{S} \circ \mathcal{R}_\Omega)(f)(x)$. The claim follows.

To complete the proof, we estimate the total error by

$$\begin{aligned}
\sup_{\rho \in K} \|F^\star(\rho) - \Psi(\rho)\|_{L^1(\mathbb{R}^{d'})} &\leq \sup_{\rho \in K} \|F^\star(\rho) - F^\star(T_M(\rho))\|_{L^1(\mathbb{R}^{d'})} \\
&\quad + \sup_{h \in K_M} \|F^\star(h) - (\mathcal{Z}_{\Omega'} \circ \Psi_{00} \circ \mathcal{R}_\Omega)(h)\|_{L^1(\mathbb{R}^{d'})} \\
&\leq \frac{\varepsilon}{2} + \sup_{h \in K_M} \int_{\mathbb{R}^{d'} \setminus \Omega'} |F^\star(h)(x)| \, dx \\
&\quad + \sup_{h \in K_M} \int_{\Omega'} |F^\star(h)(x) - \Psi_{00}(\mathcal{R}_\Omega h)(x)| \, dx \\
&\leq \frac{3\varepsilon}{4} + \sup_{h \in K_M} \int_{\Omega'} |F^\star(h)(x) - \Psi_{00}(\mathcal{R}_\Omega h)(x)| \, dx.
\end{aligned}$$

We invoked Eqs. (24) and (25) in the preceding display. Moreover,

$$\begin{aligned}\sup_{h\in K_M}\int_{\Omega'}|F^\star(h)(x)-\Psi_{00}(\mathcal{R}_\Omega h)(x)|\,dx &\le \sup_{h\in K_M}\|\mathcal{R}_{\Omega'}F^\star(h)-\Psi_0(\mathcal{R}_\Omega h)\|_{L^1(\Omega')}\\ &\quad+\sup_{h\in K_M}\|\Psi_0(\mathcal{R}_\Omega h)-\Psi_{00}(\mathcal{R}_\Omega h)\|_{L^1(\Omega')}\\ &\le \frac{\varepsilon}{8}+\frac{\varepsilon}{8}=\frac{\varepsilon}{4}\end{aligned}$$

by Eqs. (27) and (28). Combining the estimates concludes the proof. □

Corollary 3.12 (Approximation of $\mathcal{G}^\star$ over functions with unbounded domain)**.** *Let $\varepsilon>0$ and $\delta>0$ be arbitrary. Let $B\subset\mathbb{R}^r$ be bounded, open, and measurable. Let $\mathcal{U}_\delta=\mathcal{U}_{\delta,B}$ be as in (10). Let $K\subset\mathcal{U}_\delta$ be compact in $L^1(\mathbb{R}^m\times\mathbb{R}^r)$. Then there exist $R>0$ and a full space neural operator $\Psi\colon L^1(\mathbb{R}^m\times\mathbb{R}^r)\to L^1(\mathbb{R}^m\times\mathbb{R}^r)$ depending on $\Omega=\Omega':=(-R,R)^{m+r}$ as in (3.10) such that*

$$\sup_{\rho\in K}\|\mathcal{G}^\star(\rho)-\Psi(\rho)\|_{L^1(\mathbb{R}^m\times B)}\le\varepsilon. \tag{13}$$

Proof. Define $\mathcal{Z}_B^{(2)}\in\mathcal{L}(L^1(\mathbb{R}^m\times B);L^1(\mathbb{R}^m\times\mathbb{R}^r))$ by zero extension in the second coordinate. Also define $F^\star:=\mathcal{Z}_B^{(2)}\circ\mathcal{G}^\star|_K\colon K\subset L^1(\mathbb{R}^m\times\mathbb{R}^r)\to L^1(\mathbb{R}^m\times\mathbb{R}^r)$, which is continuous by Lemma 4.3. By the Dugundji extension theorem (Dugundji, 1951, Thm. 4.1, p. 357), there exists a continuous map $\widetilde{F}^\star\colon L^1(\mathbb{R}^m\times\mathbb{R}^r)\to L^1(\mathbb{R}^m\times\mathbb{R}^r)$ that extends $F^\star$. With $\widetilde{F}^\star\colon L^1(\mathbb{R}^{m+r})\to L^1(\mathbb{R}^{m+r})$, application of Theorem 3.11 (taking the larger of the two radii) delivers $R>0$ and a full space neural operator $\mathcal{G}\colon L^1(\mathbb{R}^{m+r})\to L^1(\mathbb{R}^{m+r})$ as in (3.10) with $\Omega=\Omega':=(-R,R)^{m+r}$ such that

$$\sup_{\rho\in K}\|\widetilde{F}^\star(\rho)-\mathcal{G}(\rho)\|_{L^1(\mathbb{R}^{m+r})}\le\varepsilon.$$

The assertion (13) follows from the fact that

$$\|\mathcal{G}^\star(\rho)-\mathcal{G}(\rho)\|_{L^1(\mathbb{R}^m\times B)}\le\|F^\star(\rho)-\mathcal{G}(\rho)\|_{L^1(\mathbb{R}^m\times\mathbb{R}^r)}=\|\widetilde{F}^\star(\rho)-\mathcal{G}(\rho)\|_{L^1(\mathbb{R}^{m+r})}$$

for any $\rho\in K$. □

Lemma C.1 (Truncation operator)**.** *Let $U\subseteq\mathbb{R}^d$ be measurable and open. Fix $M>0$. Define the truncation operator $T_M\colon L^1(U)\to L^1(U)\cap L^\infty(U)\subseteq L^1(U)$ by*

$$T_M(f)(x):=\begin{cases}f(x), & |f(x)|\le M,\\ M\frac{f(x)}{|f(x)|}, & |f(x)|>M\end{cases} \tag{29}$$

for each $f\in L^1(U)$ and $x\in U$. Then T_M is continuous. Moreover, if $K\subset L^1(U)$ is compact, then

$$\sup_{f\in K}\|T_M(f)\|_{L^\infty(U)}\le M\quad\textit{and} \tag{30a}$$

$$\lim_{M\to\infty}\sup_{f\in K}\|f-T_M(f)\|_{L^1(U)}=0. \tag{30b}$$

Proof. The continuity of $T_M\colon L^1(U)\to L^1(U)$ follows from the fact that T_M is a Nemytskii operator associated to a 1-Lipschitz function from U into itself. The uniform bound (30a) follows by the definition (29) of T_M. For the assertion (30b), we compute

$$\begin{aligned}\int_U|f(x)-T_M(f)(x)|\,dx&=\int_U\mathbb{1}_{\{|f(x)|>M\}}\left|\frac{f(x)}{|f(x)|}\big(|f(x)|-M\big)\right|dx\\ &\le\int_U\mathbb{1}_{\{|f(x)|>M\}}|f(x)|\,dx\end{aligned}$$

for any $f\in K$ and $M>0$. Using the fact that $K\subset L^1(U)$ compact implies that K is uniformly integrable (Brezis, 2011, Exercise 4.36(1–2), Eqn. (d), p. 129), the result (30b) follows. □

Lemma C.2 (Uniform continuity near compacta)**.** *Let (X, d_X) and (Y, d_Y) be metric spaces. Let $F\colon X \to Y$ be continuous. Let $K \subset X$ be compact. Then for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $k \in K$ and all $x \in X$ satisfying $\mathsf{d}_X(k, x) \le \delta$, it holds that $\mathsf{d}_Y(F(k), F(x)) \le \varepsilon$.*

Proof. Suppose not. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$, there exists $(k_n, x_n) \in K \times X$ such that $\mathsf{d}_X(k_n, x_n) \le 1/n$ and $\mathsf{d}_Y(F(k_n), F(x_n)) > \varepsilon$. By the compactness of K in X, there exists a subsequence $\{k_{n_j}\}_{j\in\mathbb{N}}$ and a limit $k \in K$ such that $\mathsf{d}_X(k_{n_j}, k) \to 0$ as $j \to \infty$. Consequently, $\mathsf{d}_X(x_{n_j}, k) \le 1/n_j + \mathsf{d}_X(k_{n_j}, k) \to 0$. By continuity of F, we see that $\mathsf{d}_Y(F(k_{n_j}), F(x_{n_j})) \le \mathsf{d}_Y(F(k_{n_j}), F(k)) + \mathsf{d}_Y(F(k), F(x_{n_j})) \to 0$ as $j \to \infty$. This contradicts the fact that $\mathsf{d}_Y(F(k_n), F(x_n)) > \varepsilon$ for all $n \in \mathbb{N}$. □

D Proofs for Section 4: Continuity of conditioning

This appendix collects the proofs of the stability results for the conditioning operators.

D.1 Proofs for Subsection 4.1: Stability results

Lemma 4.1 ($\mathcal{G}^\star$ is locally Lipschitz)**.** *For any $p \in \mathcal{X}_\delta$ and $q \in \mathcal{X}_\delta$, it holds that*

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{C(D\times E)} \le \frac{1}{\delta}\left(1 + \frac{|D|\min\big(\|p\|_{C(D\times E)}, \|q\|_{C(D\times E)}\big)}{\delta}\right)\|p - q\|_{C(D\times E)}. \quad (14)$$

Proof. By definition of $\mathcal{G}^\star$ and the triangle inequality,

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{C(D\times E)} = \left\|\frac{p}{m_p} - \frac{q}{m_q}\right\|_{C(D\times E)} \le \left\|\frac{p}{m_p} - \frac{p}{m_q}\right\|_{C(D\times E)} + \left\|\frac{p}{m_q} - \frac{q}{m_q}\right\|_{C(D\times E)}.$$

Since both p and q belong to $\mathcal{X}_\delta \subset \mathcal{M}_\delta$, the second term on the right-hand side of the previous display is bounded above by $\delta^{-1}\|p - q\|_{C(D\times E)}$. The first term is bounded above by

$$\|p\|_{C(D\times E)}\left\|\frac{m_q - m_p}{m_p m_q}\right\|_{C(D\times E)} \le \frac{\|p\|_{C(D\times E)}}{\delta^2}\|m_p - m_q\|_{C(E)}.$$

The data marginals satisfy

$$\|m_p - m_q\|_{C(E)} = \sup_{y\in E}\left|\int_D \big(p(x,y) - q(x,y)\big)\,dx\right| \le |D|\,\|p - q\|_{C(D\times E)}.$$

The asserted stability bound follows by collecting terms and invoking symmetry. □

Lemma 4.3 ($\mathcal{G}^\star$ is Lipschitz on functions with unbounded domain)**.** *Let $\mathcal{U}_\delta$ be as in (10). For any $p \in \mathcal{U}_\delta$ and $q \in \mathcal{U}_\delta$, it holds that*

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{L^1(\mathbb{R}^m\times B)} \le \frac{2}{\delta}\|p - q\|_{L^1(\mathbb{R}^m\times\mathbb{R}^r)}. \quad (16)$$

Proof. By definition of $\mathcal{G}^\star$ and the triangle inequality,

$$\|\mathcal{G}^\star(p) - \mathcal{G}^\star(q)\|_{L^1(\mathbb{R}^m\times B)} \le \left\|\frac{p}{m_p} - \frac{p}{m_q}\right\|_{L^1(\mathbb{R}^m\times B)} + \left\|\frac{p}{m_q} - \frac{q}{m_q}\right\|_{L^1(\mathbb{R}^m\times B)}.$$

Since both p and q belong to $\mathcal{U}_\delta$, the second term on the right-hand side of the previous display is bounded above by $\delta^{-1}\|p - q\|_{L^1(\mathbb{R}^m\times\mathbb{R}^r)}$. For the first term,

$$\begin{aligned}\left\|\frac{p}{m_p} - \frac{p}{m_q}\right\|_{L^1(\mathbb{R}^m\times B)} &= \int_{\mathbb{R}^m\times B}\frac{|m_q(y) - m_p(y)|}{m_p(y)m_q(y)}p(x,y)\,dxdy\\ &= \int_B \frac{|m_q(y) - m_p(y)|}{m_q(y)}\,dy\\ &\le \frac{1}{\delta}\|m_p - m_q\|_{L^1(B)}.\end{aligned}$$

The second equality in the preceding display invoked Tonelli's theorem to exchange the order of integration. The data marginals satisfy

$$\|m_p - m_q\|_{L^1(B)} = \int_B \left| \int_{\mathbb{R}^m} \big(p(x,y) - q(x,y)\big)\, dx \right| dy \le \|p - q\|_{L^1(\mathbb{R}^m \times \mathbb{R}^r)}.$$

The asserted Lipschitz continuity follows. □

Theorem 4.4 (Continuity of $\Psi^\star$)**.** *The map $\Psi^\star$ in (4) is continuous.*

Proof. We use the sequential definition of continuity. Let $(\rho, y) \in \mathcal{X}_\delta \times E$ and $\{(\rho_n, y_n)\}_{n\in\mathbb{N}} \subset \mathcal{X}_\delta \times E$ be any sequence such that

$$\|(\rho_n, y_n) - (\rho, y)\|_{C(D\times E)\times E} = \max\big(\|\rho_n - \rho\|_{C(D\times E)}, |y_n - y|\big) \to 0$$

as $n \to \infty$. We must prove that $\Psi^\star(\rho_n, y_n) \to \Psi^\star(\rho, y)$ in $C(D)$. To this end, we use the error decomposition

$$\|\Psi^\star(\rho_n, y_n) - \Psi^\star(\rho, y)\|_{C(D)} \le \underbrace{\|\rho_n(\,\cdot\,|\,y_n) - \rho(\,\cdot\,|\,y_n)\|_{C(D)}}_{\text{(I)}} + \underbrace{\|\rho(\,\cdot\,|\,y_n) - \rho(\,\cdot\,|\,y)\|_{C(D)}}_{\text{(II)}}.$$

For the first term (I), it holds that

$$\begin{aligned}
\text{(I)} &\le \sup_{y'\in E} \|\rho_n(\,\cdot\,|\,y') - \rho(\,\cdot\,|\,y')\|_{C(D)} \\
&= \|\mathcal{G}^\star(\rho_n) - \mathcal{G}^\star(\rho)\|_{C(D\times E)} \\
&\le \frac{1}{\delta}\left(1 + \frac{|D|\min(\sup_{n'\in\mathbb{N}}\|\rho_{n'}\|_{C(D\times E)}, \|\rho\|_{C(D\times E)})}{\delta}\right)\|\rho_n - \rho\|_{C(D\times E)}
\end{aligned}$$

by Lemma 4.1. The last line of the preceding display tends to zero as $n \to \infty$.

It remains to show that term (II) converges to zero. It holds that

$$\begin{aligned}
\text{(II)} &\le \left\| \frac{\rho(\,\cdot\,, y_n)}{m_\rho(y_n)} - \frac{\rho(\,\cdot\,, y_n)}{m_\rho(y)} \right\|_{C(D)} + \left\| \frac{\rho(\,\cdot\,, y_n)}{m_\rho(y)} - \frac{\rho(\,\cdot\,, y)}{m_\rho(y)} \right\|_{C(D)} \\
&\le \frac{\|\rho\|_{C(D\times E)}|m_\rho(y_n) - m_\rho(y)|}{\delta^2} + \frac{1}{\delta}\|\rho(\,\cdot\,, y_n) - \rho(\,\cdot\,, y)\|_{C(D)}.
\end{aligned}$$

Since $\rho \in C(D \times E)$ is continuous on the compact set $D \times E$, the function ρ is in fact uniformly continuous with modulus of continuity ω_ρ. Therefore,

$$|m_\rho(y_n) - m_\rho(y)| \le \int_D |\rho(x, y_n) - \rho(x, y)|\, dx \le |D|\, \omega_\rho\big(\max(0, |y_n - y|)\big) \to 0$$

as $n \to \infty$. Similarly,

$$\|\rho(\,\cdot\,, y_n) - \rho(\,\cdot\,, y)\|_{C(D)} = \sup_{x\in D} |\rho(x, y_n) - \rho(x, y)| \le \omega_\rho(|y_n - y|) \to 0$$

by the convergence of $y_n \to y$. We deduce that (II) $\to 0$ as required. □

Corollary 4.5 (Hölder stability of $\Psi^\star$ on balls)**.** *Let $\alpha \in (0, 1]$. Fix $R > 0$ and define*

$$\mathcal{H}^\alpha_R := \big\{ f \in C^{0,\alpha}(D \times E) \,\big|\, \|f\|_{C^{0,\alpha}(D\times E)} \le R \big\}.$$

For any (p, y) and (q, y') belonging to $(\mathcal{X}_\delta \cap \mathcal{H}^\alpha_R) \times E$, it holds that

$$\|\Psi^\star(p, y) - \Psi^\star(q, y')\|_{C(D)} \le \frac{(1 + 2R)}{\delta}\left(1 + \frac{|D|R}{\delta}\right)\big\|(p, y) - (q, y')\big\|^\alpha_{C(D\times E)\times E}. \tag{17}$$

Proof. First suppose that $\|p-q\|_{C(D\times E)} \le 1$. By following the proof of Theorem 4.4, we have

$$\begin{aligned}\|\Psi^\star(p,y)-\Psi^\star(q,y')\|_{C(D)} &\le \frac{1}{\delta}\left(1+\frac{|D|R}{\delta}\right)\|p-q\|_{C(D\times E)} \\ &\quad + \frac{|D|R}{\delta^2}\omega_p(|y-y'|)+\frac{1}{\delta}\omega_p(|y-y'|) \\ &\le \frac{1}{\delta}\left(1+\frac{|D|R}{\delta}\right)\|p-q\|_{C(D\times E)} \\ &\quad + \frac{|D|R^2}{\delta^2}|y-y'|^\alpha+\frac{R}{\delta}|y-y'|^\alpha.\end{aligned}$$

Since $\|p-q\|_{C(D\times E)} \le 1$, it holds that

$$\|p-q\|_{C(D\times E)} \le \|p-q\|^\alpha_{C(D\times E)} \le \|(p,y)-(q,y')\|^\alpha_{C(D\times E)\times E}.$$

Using $|y-y'| \le \|(p,y)-(q,y')\|_{C(D\times E)\times E}$ and combining terms yield the bound

$$\begin{aligned}\|\Psi^\star(p,y)-\Psi^\star(q,y')\|_{C(D)} &\le \frac{(1+R)}{\delta}\left(1+\frac{|D|R}{\delta}\right)\big\|(p,y)-(q,y')\big\|^\alpha_{C(D\times E)\times E} \\ &\le \frac{(1+2R)}{\delta}\left(1+\frac{|D|R}{\delta}\right)\big\|(p,y)-(q,y')\big\|^\alpha_{C(D\times E)\times E}.\end{aligned}$$

We enlarged R to $2R$ in the first parenthesis of the last line.

If $\|p-q\|_{C(D\times E)} > 1$, then $\|(p,y)-(q,y')\|^\alpha_{C(D\times E)\times E} \ge \|p-q\|^\alpha_{C(D\times E)} > 1$. Using this fact, it holds that

$$\begin{aligned}\|\Psi^\star(p,y)-\Psi^\star(q,y')\|_{C(D)} &\le \|\Psi^\star(p,y)\|_{C(D)}+\|\Psi^\star(q,y')\|_{C(D)} \\ &\le \frac{2R}{\delta} \\ &< \frac{2R}{\delta}\|(p,y)-(q,y')\|^\alpha_{C(D\times E)\times E} \\ &\le \frac{(1+2R)}{\delta}\left(1+\frac{|D|R}{\delta}\right)\big\|(p,y)-(q,y')\big\|^\alpha_{C(D\times E)\times E}.\end{aligned}$$

Thus, the estimate (17) is always valid as asserted. □

D.2 Proofs for Subsection 4.2: Beyond uniform lower bounds

Lemma D.1 (Marginal continuity)**.** *The function $m_\rho\colon E\to\mathbb{R}$ is continuous for any $\rho \in C_{\mathrm{ac}}(D\times E)$.*

Proof. Since the function $\rho\colon D\times E\to\mathbb{R}$ is continuous on the compact set $D\times E$, it is in fact uniformly continuous. Take a sequence $\{y_n\}_{n\in\mathbb{N}}\subset E$ such that $y_n\to y$ as $n\to\infty$. By uniform continuity, there is some $n_0\in\mathbb{N}$ such that for all $n\ge n_0$, we have

$$\sup_{x\in D}|\rho(x,y_n)-\rho(x,y)| < \varepsilon/|D|,$$

where $|D|$ is the Lebesgue measure of D. Then for all $n\ge n_0$, we have

$$\begin{aligned}|m_\rho(y_n)-m_\rho(y)| &= \left|\int_D \rho(x,y_n)\,dx-\int_D\rho(x,y)\,dx\right| \\ &\le \int_D|\rho(x,y_n)-\rho(x,y)|\,dx \\ &\le |D|\sup_{x\in D}|\rho(x,y_n)-\rho(x,y)| \\ &< \varepsilon,\end{aligned}$$

concluding the proof. □

Lemma D.2 (Positive marginal characterization). *Let $\mathcal{X}_+ \coloneqq \bigcup_{\delta>0} \mathcal{X}_\delta$. Then it holds that*

$$\mathcal{X}_+ = \{\rho \in C_{\mathrm{ac}}(D \times E) \,|\, m_\rho > 0\}.$$

Proof. Since for any $\delta > 0$ we have

$$\mathcal{X}_\delta \subseteq \{\rho \in C_{\mathrm{ac}}(D \times E) \,|\, m_\rho > 0\},$$

we get the trivial inclusion

$$\mathcal{X}_+ \subseteq \{\rho \in C_{\mathrm{ac}}(D \times E) \,|\, m_\rho > 0\}.$$

For the other inclusion, take any $\rho \in C(D \times E)$ such that $m_\rho > 0$. By compactness of E and Lemma D.1, there is some $y_0 \in E$ such that

$$m_\rho(y_0) = \min_{y' \in E} m_\rho(y').$$

Thus, if we set $\delta \coloneqq m_\rho(y_0) > 0$, then we obtain $\rho \in \mathcal{X}_\delta \subseteq \mathcal{X}_+$ as required. □

Theorem 4.6 (Continuous extension of conditioning operators). *The operators $\mathcal{G}^\star$ and $\Psi^\star$ admit unique continuous extensions to the domains $\mathcal{X}_+$ and $\mathcal{X}_+ \times E$, denoted by $\mathcal{G}^\star_+$ and $\Psi^\star_+$, respectively.*

Proof. For any $\delta > 0$, Eqs. (4) and (6) define operators

$$\mathcal{G}^\star_\delta \colon \mathcal{X}_\delta \to C(D \times E) \quad \text{and} \quad \Psi^\star_\delta \colon \mathcal{X}_\delta \times E \to C(D).$$

Here, we are merely making the dependence on δ explicit in the notation. By Lemma 4.1 and Theorem 4.4, these operators are continuous. We can now define the extensions

$$\mathcal{G}^\star_+ \colon \mathcal{X}_+ \to C(D \times E) \quad \text{and}$$
$$\Psi^\star_+ \colon \mathcal{X}_+ \times E \to C(D)$$

by setting

$$\mathcal{G}^\star_+|_{\mathcal{X}_\delta} = \mathcal{G}^\star_\delta \quad \text{and} \quad \Psi^\star_+|_{\mathcal{X}_\delta \times E} = \Psi^\star_\delta.$$

These define valid extensions since Lemma D.2 implies that for any $\rho \in \mathcal{X}_+$, there is some $\delta > 0$ such that $\rho \in \mathcal{X}_\delta$. Therefore,

$$\mathcal{G}^\star_+(\rho) = \mathcal{G}^\star_\delta(\rho) \quad \text{and} \quad \Psi^\star_+(\rho, y) = \Psi^\star_\delta(\rho, y),$$

and the operators on the right-hand side are defined in the preceding discussion. Moreover, note that these expressions are well-defined because for any two $\delta_1 > 0$ and $\delta_2 > 0$ such that $\delta_1 > \delta_2 > 0$, Eqs. (4) and (6) yield

$$\mathcal{G}^\star_{\delta_1} = \mathcal{G}^\star_{\delta_2} \quad \text{and} \quad \Psi^\star_{\delta_1} = \Psi^\star_{\delta_2}.$$

It remains to show that these extensions are continuous and unique. Consider a sequence $\{\rho_n\}_{n\in\mathbb{N}} \subset \mathcal{X}_+$ such that

$$\rho_n \to \rho \quad \text{in} \quad C(D \times E)$$

and $\rho \in \mathcal{X}_+$. Now by a triangle inequality and an elementary inequality for the supremum we get

$$\|m_{\rho_n} - m_\rho\|_{C(E)} \leq |D| \cdot \|\rho_n - \rho\|_{C(D\times E)},$$

which combined with the assumed convergence $\rho_n \to \rho$ yields

$$m_{\rho_n} \to m_\rho \quad \text{in} \quad C(D \times E).$$

By Lemma D.2, there exists some $\delta > 0$ such that $\rho \in \mathcal{X}_\delta$, and therefore by definition

$$m_\rho \geq \delta > 0.$$

Thus, there must exist n_0 such that for all $n \geq n_0$ we get

$$\|m_{\rho_n} - m_\rho\|_{C(E)} \leq \delta/2.$$

This implies that for all $y \in E$, we have $|m_{\rho_n}(y) - m_\rho(y)| \leq \delta/2$. Rearranging, we obtain

$$m_{\rho_n}(y) \geq m_\rho(y) - \frac{\delta}{2} \geq \inf_{y\in E} m_\rho(y) - \frac{\delta}{2} \geq \frac{\delta}{2}$$

using that $m_\rho \geq \delta$. In other words, for $\delta' = \delta/2$, we have shown that for any $n \geq n_0$ we have

$$\rho_n \in \mathcal{X}_{\delta'}.$$

Now since

$$\mathcal{G}^\star_+|_{\mathcal{X}_{\delta'}} = \mathcal{G}^\star_{\delta'} \quad \text{and} \quad \Psi^\star_+|_{\mathcal{X}_{\delta'} \times E} = \Psi^\star_{\delta'},$$

continuity follows by the continuity of $\mathcal{G}^\star_{\delta'}$ and $\Psi^\star_{\delta'}$. For uniqueness, pick any $\rho \in \mathcal{X}_+$ and note that since ρ is continuous and $D \times E$ is compact, there exists some $(x_0, y_0) \in D \times E$ such that

$$\inf_{(x,y)\in D\times E} \rho(x, y) = \rho(x_0, y_0) > 0.$$

Letting $\delta'' \coloneqq \rho(x_0, y_0)$, we note that any extensions $\mathcal{G}^\star_0$ and $\Psi^\star_0$ must agree with $\mathcal{G}^\star_{\delta''}$ and $\Psi^\star_{\delta''}$ on $\mathcal{X}_{\delta''}$, which contains ρ by definition of δ''. Thus, we have

$$\mathcal{G}^\star_0(\rho) = \mathcal{G}^\star_{\delta''}(\rho) = \mathcal{G}^\star_+(\rho) \quad \text{and} \quad \Psi^\star_0(\rho, y) = \Psi^\star_{\delta''}(\rho, y) = \Psi^\star_+(\rho, y) \quad \text{for all} \quad y \in E.$$

In other words, the arbitrary extensions $\mathcal{G}^\star_0$ and $\Psi^\star_0$ agree with $\mathcal{G}^\star_+$ and $\Psi^\star_+$ at ρ, proving uniqueness. □

Lemma D.3 (Density of $\mathcal{X}_+$)**.** *Fix $\rho \in C_{\text{ac}(D\times E)}$. Then for any $\delta > 0$ and any $q \in \mathcal{X}_\delta$, the sequence*

$$\rho_n = \left(1 - \frac{1}{n}\right)\rho + \frac{1}{n}q$$

for $n \in \{1, 2, \ldots\}$ is such that

$$\rho_n \to \rho \quad \text{in} \quad C(D \times E).$$

Moreover, we have

$$\rho_n \in \mathcal{X}_{\delta/n}.$$

In particular, it follows that the space $\mathcal{X}_+$ is dense in $C_{\text{ac}}(D \times E)$ in the $C(D \times E)$ norm.

Proof. Fix $\rho \in C_{\text{ac}(D\times E)}$ and $q \in \mathcal{X}_\delta$. Now integrating over D we get

$$m_{\rho_n} = \left(1 - \frac{1}{n}\right)m_\rho + \frac{1}{n}m_\delta.$$

By definition of $\mathcal{X}_\delta$, we have $m_q(y) \geq \delta$ for all $y \in E$. Thus, for each n, we have

$$m_{\rho_n} \geq \frac{\delta}{n} > 0.$$

It follows that $\rho_n \in \mathcal{X}_{\delta/n}$ for all n. On the other hand, note that

$$\|\rho - \rho_n\|_{C(D\times E)} = \frac{\|\rho - q\|_{C(D\times E)}}{n} \leq \frac{\|\rho\|_{C(D\times E)} + \|q\|_{C(D\times E)}}{n} \xrightarrow{n\to\infty} 0$$

using that ρ and q have finite maxima as continuous functions over compact domains. We have thus shown the required uniform convergence. Last, we note that the density of $\mathcal{X}_+ = \bigcup_{\delta>0} \mathcal{X}_\delta$ in $C_{\text{ac}}(D \times E)$ follows by definition. □

Theorem 4.7 (Maximality of extension)**.** *Let $\mathcal{G}^\star_+$ and $\Psi^\star_+$ be the extension operators obtained in Theorem 4.6. There is no continuous extension of $\mathcal{G}^\star_+$ to any strict superset S of $\mathcal{X}_+$ in $C_{\text{ac}}(D \times E)$, that is, any set S satisfying $\mathcal{X}_+ \subsetneq S \subseteq C_{\text{ac}}(D \times E)$. Similarly, there is no continuous extension of $\Psi^\star_+$ to a set of the form $S' \times E$ where S' is some strict superset of $\mathcal{X}_+$ in $C_{\text{ac}}(D \times E)$.*

Proof. To see the first claim, assume such an extension exists and call it $\overline{\mathcal{G}^\star}\colon S \to C(D \times E)$. By the strict superset property of S, pick some $\rho \in S \setminus \mathcal{X}_+$. Since $\rho \notin \mathcal{X}_+$, there exists $y_* \in E$ such that $m_\rho(y_*) = 0$. Since ρ is continuous, m_ρ is a marginal and the sets D and E are assume to be closures of open sets (recall Section 2), it follows that

$$\rho(x, y_*) = 0 \quad \text{for all} \quad x \in D.$$

Now fix arbitrary $\delta > 0$ and $q \in \mathcal{X}_\delta$ and construct the sequence

$$\rho_n = \left(1 - \frac{1}{n}\right)\rho + \frac{1}{n}q$$

for $n \in \{1, 2, \ldots\}$. By Lemma D.3, we have

$$\rho_n \to \rho \quad \text{in} \quad C(D \times E).$$

By the assumed continuity of $\overline{\mathcal{G}^\star}$, it follows that

$$\overline{\mathcal{G}^\star}(\rho_n) \to \overline{\mathcal{G}^\star}(\rho) \quad \text{in} \quad C(D \times E).$$

Again by Lemma D.3, we have $\rho_n \in \mathcal{X}_{\delta/n}$ and so

$$\overline{\mathcal{G}^\star}(\rho_n) = \mathcal{G}^\star(\rho_n), \tag{31}$$

where on the right-hand side we have the $\mathcal{G}^\star$ operator defined over the set $\mathcal{X}_{\delta/n}$ via the explicit formulas of Section 2, namely

$$\mathcal{G}^\star(\rho_n) = \kappa_{\rho_n} = \frac{\rho_n}{m_{\rho_n}}.$$

The fraction is well defined because $m_{\rho_n} \geq \delta/n$. Using the definition of ρ_n and rearranging, we obtain

$$\mathcal{G}^\star(\rho_n) = \frac{\left(1 - \frac{1}{n}\right)\rho + \frac{1}{n}q}{\left(1 - \frac{1}{n}\right)m_\rho + \frac{1}{n}m_q}.$$

We can now plug-in $y_* \in E$ and any $x \in D$ to obtain

$$\mathcal{G}^\star(\rho_n)(x, y_*) = \frac{q(x, y_*)}{m_q(y_*)},$$

using the fact that $m_q(y_*) = \rho(x, y_*) = 0$. Since n was arbitrary, we can take the limit $n \to \infty$ and use the uniform convergence of (31), implying pointwise convergence, to conclude

$$\mathcal{G}^\star(\rho)(x, y_*) = \frac{q(x, y_*)}{m_q(y_*)}. \tag{32}$$

Notice, however, that in the above argument the choice $q \in \mathcal{X}_\delta$ was left arbitrary. We can thus pick $q_1 \in \mathcal{X}_\delta$ and $q_2 \in \mathcal{X}_\delta$ such that

$$\frac{q_1(x, y_*)}{m_{q_1}(y_*)} \neq \frac{q_2(x, y_*)}{m_{q_2}(y_*)}.$$

Repeating the argument that leads to (32), we obtain

$$\mathcal{G}^\star(\rho)(x, y_*) = \frac{q_1(x, y_*)}{m_{q_1}(y_*)} \neq \frac{q_2(x, y_*)}{m_{q_2}(y_*)} = \mathcal{G}^\star(\rho)(x, y_*)$$

which is a contradiction to the existence of $\overline{\mathcal{G}^\star}$. We can now show the non-existence of an extension $\overline{\Psi^\star} : S' \times E \to C(D)$ similarly. That is, we pick $\rho' \in S' \setminus C_{\text{ac}}(D \times E)$, we find y_* such that $m_\rho(y_*) = 0$, we construct a sequence ρ_n as above and use the equality

$$\overline{\Psi^\star}(\rho)(x, y_*) = \lim_{n \to \infty} \Psi^\star(\rho_n)(x, y_*) = \lim_{n \to \infty} \frac{\rho_n(x, y_*)}{m_{\rho_n}(x, y_*)}$$

to obtain an expression that depends on an arbitrary $q \in \mathcal{X}_{\delta'}$, for some $\delta' > 0$. This leads to a contradiction, exactly as above. □

E Examples and remarks

This appendix expands on aspects of the main text.

Remark E.1 (Exponential map)**.** We note that the in-context conditioning operator $\Psi^\star$ can be expressed as a composition of the kernel conditioning operator $\mathcal{G}^\star$ and the evaluation operator $\mathrm{ev}\colon C(D \times E) \times E \to C(D)$ defined by $\mathrm{ev}(f, y) = f(\,\cdot\,, y)$. In particular, it holds that

$$\Psi^\star\colon \mathrm{Dom}(\Psi^\star) \xrightarrow{\mathcal{G}^\star \times \mathrm{id}} C(D \times E) \times E \xrightarrow{\mathrm{ev}(\,\cdot\,, y)} C(D)$$

$$\Psi^\star\colon (\rho, y) \mapsto (\kappa_\rho, y) \mapsto \kappa_\rho(\,\cdot\,, y) = \frac{\rho(\,\cdot\,, y)}{m_\rho(y)}.$$

In fact, if we define the *exponential map* (Brown, 1964, Corollary 1.8)

$$J\colon C(D \times E) \to C(E; C(D))$$
$$J\colon f \mapsto (y \mapsto f(\,\cdot\,, y)),$$

then we have

$$\Psi^\star(\rho, y) = J\big(\mathcal{G}^\star(\rho)\big)(y).$$

Remark E.2 (Weakened source condition)**.** We note that Corollary 4.5 remains true in a set $\widetilde{S}$ larger than S, namely

$$\widetilde{S} \coloneqq \left\{ f \in C(D \times E) \,\middle|\, \sup_{x \in D} \|f(x, \,\cdot\,)\|_{C^{0,\alpha}(E)} \leq R \right\}.$$

Example E.3 (Loss of continuity for convergence in distribution)**.** It is harder for conditioning to be continuous in weaker topologies because the set of admissible measures becomes larger. Write $\mathscr{P}(\mathbb{R}^k)$ for the space of probability measures on $\mathbb{R}^k$ for some $k \in \mathbb{N}$. Consider the topology of convergence in distribution, denoted by $\rightsquigarrow$, also known as the topology of *weak* or *narrow* convergence. Recall that it can be metrized by many distances, including Dudley's bounded-Lipschitz metric d_{BL}, or the Wasserstein metric d_{W} for a compact state space. For every $n \in \mathbb{N}$, let

$$\nu_n \coloneqq \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1/n)} \quad \text{and} \quad \nu \coloneqq \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,0)}$$

be probability measures supported on $\mathbb{R}^2$. Let

$$\mathrm{Dom}_0(\Psi^\star) \coloneqq \big\{ \mu \in \mathscr{P}(\mathbb{R}^2) \,\big|\, \mu(\mathbb{R} \times \{0\}) > 0 \big\}.$$

For $\mu \in \mathrm{Dom}_0(\Psi^\star)$, define

$$\Psi^\star(\mu, 0)(A) \coloneqq \frac{\mu(A \times \{0\})}{\mu(\mathbb{R} \times \{0\})}$$

for every Borel set $A \subseteq \mathbb{R}$. These conditional laws are independent of the choice of regular conditional distribution because 0 is an atom of the y-marginal. Both ν_n and ν belong to $\mathrm{Dom}_0(\Psi^\star)$, and $\nu_n \rightsquigarrow \nu$ as $n \to \infty$. However,

$$\Psi^\star(\nu_n, 0) = \delta_0 \quad \text{and} \quad \Psi^\star(\nu, 0) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$$

for all n. So, the conditionals do not converge in distribution. Thus, $\Psi^\star(\,\cdot\,, 0)\colon (\mathrm{Dom}_0(\Psi^\star), \mathsf{d}_{\mathsf{BL}}) \to (\mathscr{P}(\mathbb{R}), \mathsf{d}_{\mathsf{BL}})$ is not continuous at ν.

F Numerical experiment details

This appendix describes details of the numerical experiments in Section F.1, data generation in Section F.2, as well as architectural and training details in Section F.3.

F.1 Description of experiments

The two experiments involve training on $N = 50000$ function-valued input-output pairs $\{(\rho_n, \kappa_n)\}_{n=1}^N$, given by Gaussian mixture joints ρ_n together with the corresponding conditional kernel κ_n; note that both have closed-form expressions. Both are evaluated on a finite grid of 64^2 points discretizing the domain $[-6, 6]^2$. We use an additional 1000 samples as validation during training, and an additional 1000 as test data, which is not used in training. The data generation process is described in detail in Section F.2. The *correlated Gaussian* experimental setting corresponds to $K = 1$, while the *Gaussian mixture* setting to $K = 3$, where K represents the number of components in the mixture.

We also investigate the use of our trained models on sets of kernel density estimates (KDEs) obtained over 2000 samples coming from test distributions in the Gaussian mixture class. This is done to compare our trained models against a plug-in estimator for the conditional, the latter using the KDE. We evaluate models trained on closed-form distributions on this set of KDEs, the latter viewed as functions. We describe details regarding the generation of KDE data in the following subsection.

F.2 DATA GENERATION

We describe the data generation process for the experimental settings in this paper. We begin by describing the generation mechanism for the closed-form density input-output pairs. For a given set of weights $\{w_k\}_{k=1}^K$ and a set of parameters $\{\mu_k^x, \mu_k^y, \sigma_k^x, \sigma_k^y, \xi_k\}_{k=1}^K$, consider a K-order Gaussian mixture joint distribution of the form

$$\rho(x,y) \coloneqq \sum_{k=1}^{K} w_k \mathcal{N}(x,y;\mu_k,\mathcal{C}_k), \quad \text{where} \quad \mu_k \coloneqq \begin{pmatrix} \mu_k^x \\ \mu_k^y \end{pmatrix}, \quad \text{and} \quad \mathcal{C}_k \coloneqq \begin{pmatrix} (\sigma_k^x)^2 & \xi_k \sigma_k^x \sigma_k^y \\ \xi_k \sigma_k^x \sigma_k^y & (\sigma_k^y)^2 \end{pmatrix},$$

for $\mathcal{N}$ denoting the density function of a multivariate Gaussian. A standard computation leads to the closed form Markov kernel given by

$$\kappa_\rho(x \,|\, y) = \sum_{k=1}^{K} \widetilde{w}_k(y)\, \mathcal{N}\Big(x;\ \mu_k^x + \xi_k \frac{\sigma_k^x}{\sigma_k^y}(y-\mu_k^y),\ (\sigma_k^x)^2(1-\xi_k^2)\Big),$$

where

$$\widetilde{w}_k(y) \coloneqq \frac{w_k\, \mathcal{N}\Big(y;\ \mu_k^y,\ (\sigma_k^y)^2\Big)}{\displaystyle\sum_{j=1}^{K} w_j\, \mathcal{N}\Big(y;\ \mu_j^y,\ (\sigma_j^y)^2\Big)}.$$

For the correlated Gaussian ($K = 1$) setting, the densities in the training, validation, and test sets are given by using $w_1 = 1$ and sampling independently μ_1^x and μ_1^y from $\mathrm{Uniform}(-3,3)$, σ_1^x and σ_1^y from $\mathrm{Uniform}(0.3,1.2)$, and ξ_1 from $\mathrm{Uniform}(-0.7,0.7)$. For the Gaussian mixture ($K = 3$) setting, the densities in the training, validation, and test sets are given by sampling the weights of the Gaussian mixtures independently from a $\mathrm{Dirichlet}(\mathbb{1}_K)$ distribution, μ_k^x and μ_k^y from $\mathrm{Uniform}(-3,3)$, σ_k^x and σ_k^y from $\mathrm{Uniform}(0.3,1.2)$, and ξ_k from $\mathrm{Uniform}(-0.7,0.7)$. For reproducibility, we note that all training, validation, test sets are generated using the seeds 0, 1, and 2, respectively.

In the KDE setting, for each testing instance, we randomly draw parameters defining a joint density $\rho = \rho(x,y)$ as before and then draw 2000 samples from it, constructing a Gaussian KDE with Silverman's rule-of-thumb bandwidth. We then evaluate this KDE on the same grid as the training data. This serves as the model input, while the exact conditional density $\rho(x \,|\, y) = \rho(x,y)/\int \rho(x,y)\,dx$, computed analytically, serves as the true kernel, against which we compare the model prediction. Last, the KDE plug-in estimator is computed by using the conditional formula $\widehat{\rho}(x,y)/\int \widehat{\rho}(x,y)\,dx$ for some KDE $\widehat{\rho}$ and standard numerical integration. For reproducibility, we note that these test sets are generated using the seed value 2.

F.3 ARCHITECTURAL AND TRAINING DETAILS

We report the architectural and training details for the FNO and TNO.

We use a 2D FNO from the "NeuralOperator" library (Kossaifi et al., 2026) that takes as input the joint distribution discretized over a 64×64 grid with two appended coordinate channels encoding the x- and y-grid values on the domain $[-6,6]$. The FNO we use here consists of 4 Fourier layers with 16 retained Fourier modes per spatial dimension and hidden channel width 128, producing a single-channel output which we interpret as the predicted conditional density $\rho(x \,|\, y)$. The model is trained with the Adam optimizer at an initial learning rate of 10^{-3}, with a ReduceLROnPlateau scheduler (patience 5, reduction factor 0.5), minimizing the relative L^1 loss between the predicted and exact conditional densities. We use a batch size of 64 for training.

We use a TNO from the GitHub repository of Calvello et al. (2025a) that takes as input the joint distribution discretized over a 64×64 grid with two appended coordinate channels encoding the x- and y-grid values on the domain $[-6,6]$, which serves as positional encoding. The TNOs we use consist of 12 and 24 self-attention layers for the correlated Gaussian and Gaussian mixture experiments, respectively. The layers have channel dimension 128, 8 attention heads, feedforward dimension 256, and GELU activation. This produces an output sequence of shape 64×64 which we interpret as $\rho(x \,|\, y)$. Training follows the same optimizer, loss, and batch size as the FNO.

The FNO for the Gaussian mixture experiment was trained on a Tesla V100-PCIE-16GB on our institution's HPC cluster. All other architectures were trained with one NVIDIA H200-144GB GPU, on the same cluster.