SUPPLEMENT: UNIVERSAL APPROXIMATION UNDER CONSTRAINTS IS POSSIBLE WITH TRANSFORMERS

Anonymous authors
Paper under double-blind review

A PRECISE TRANSFORMER COMPLEXITIES AND APPROXIMATION RATES

This section records the exact approximation rates, or equivalently the precise model complexities, of the transformer networks implemented in our quantitative constrained universal approximation results. The rates are simply those recorded in Table 1 but with explicit constants.

Table 2 makes use of the following notation. We denote the diameter of the compact set K by $\operatorname{diam}(K) \triangleq \max_{y_1,y_2\in K} \|y_1 - y_2\|$. Furthermore, $k > 0$ in Table 2 is a universal constant independent of ϵ_K, ϵ_f, f, K, n, m, and of d. The big $\mathscr{O}$ notation used in Table 2 masks any constants not depending on these quantities.

Network	$\hat{\mathcal{E}}$	$\hat{\mathcal{D}}$
Depth	$\mathscr{O}(m^{\frac{1}{s}}(1+\epsilon_f^{\frac{2n}{3(kn+1)}-\frac{2n}{kn+1}}))$	$\mathscr{O}\left((m^{\frac{1}{s}}N^{\frac{3}{2}}(\operatorname{Lip}(\Phi)\operatorname{diam}(K)+2\epsilon_f)(1-\frac{\epsilon_K^{-1}}{4})^2(1+\frac{m^{\frac{1}{s}}}{4}))^{\frac{2m}{s}}\right)$
Width	$\mathscr{O}(m^{\frac{1}{s}}(4n+10))$	$\mathscr{O}(m^{\frac{1}{s}}+N+2)$
N	-	$\mathscr{O}\left(\left(\frac{km^{\frac{2}{s}}2^{\frac{9}{2}}\operatorname{Lip}(\Phi)(\operatorname{diam}(K)+\epsilon_f)}{\sqrt{m^{\frac{1}{s}}+1}\epsilon_K}\right)^{\frac{m}{s}}\right)$
Q	-	$\mathscr{O}\left((\epsilon_K^{-1}\operatorname{Lip}(\Phi)\operatorname{diam}(K)m^{\frac{5}{2s}})^{\frac{m}{s}}\right)$

Table 2: Complexities of encoder-decoder Network with P-attention for $0 < \epsilon_K \leq 1$.

Remark A.1. *If $\epsilon_K > 1$ then the above rates hold but the term $(1-\frac{\epsilon_K^{-1}}{4})^2$ in $\hat{\mathscr{D}}$'s depth estimate must be replaced by $(1-\frac{\epsilon_K^{-1}}{4})(1-\epsilon_K^{-1})$ in order for the rates to remain valid.*

B CAN THE PROBABILISTIC TRANSFORMER NETWORK BE TRAINED?

The purpose of this appendix is to answer the following questions:

(i) *Are our probabilistic transformer networks trainable and, if so, how?*

(ii) *How do probabilistic transformer networks perform on a toy non-convex problem?*

We first affirm (i) by describing a potential training algorithm for our model. Then, we address (ii) on a toy non-convex problem whose objective is to learn (randomly generated) functions taking values on the standard 2-sphere in $\mathbb{R}^3$. Our code is available at code is available at Anonymized (2021).

B.1 A TRAINING ALGORITHM

Our analysis only has practical implications as we can affirmatively answer the following question:

"Is the probabilistic transformer network $\hat{F}$ of Theorem 2.7 trainable?"

Our theoretical analysis motivates the following training procedure, whose steps we briefly explain.

Step 1 - Get Particles: We assume that the user is has access to a subroutine Generate which generates particles on K. This is always possible, for example, by randomly by sampling the available training outputs $\{y_t\}_{t=1}^T$; for example, this is what is done in our toy implementations. However, if one has access to additional structure, such as a K-supported probability measure (Miolane et al., 2020) then, samples can be drawn therefrom, or if K is equipped with a meaningful metric, then the randomized partition procedure such as (Bartal, 1999; Pulat, 1989) can be deployed.

Step 2 - Train Model: When $m > 1$, the Wasserstein distance is costly to evaluate numerically (Pele & Werman, 2009; Cuturi, 2013; Kolouri et al., 2019; Sommerfeld et al., 2019). A variety of approximate or regularized transport distances have been introduced to manage this problem but only approximately. However, in the context of Theorem 2.7 and Algorithm 3 we are always interested in distances to the pointmass and therefore, $\mathscr{W}_1$ has the following exceptional *closed-form* expression which bypasses these computational issues:

$$\sum_{t=1}^{T} \mathcal{W}_1(\delta_{f(x_t)}, \text{P-attention}(\hat{f}(x_t), Y)) = \sum_{t=1}^{T}\sum_{n=1}^{N} \|y_t - Y_n\| \, [\text{Softmax}_N \circ \hat{f}(x_t)]_n. \tag{11}$$

Remark B.1 (Exceptional Closed-Form for $\mathcal{W}_1(\delta_{f(x_t)}, \text{P-attention}(\hat{f}(x_t), Y))$ in (11))**.** *A derivation of the closed-form identity 11 is in Lemma C.4.*

Step 2 - Prediction: In the case where K is a Riemannian manifold, Since Fréchet means are readily implemented in a variety of packages (Miolane et al., 2020), we assume that the user has access to a subroutine Fréchet Mean which takes an $N \times Q$-matrix of weights $(w_{n,q})_{n,q=1}^{N,Q}$ and an $N \times Q \times 1$-array and computes the Fréchet mean (9). By Corollary 2.11, once the network $\hat{F}$ is trained, its outputs generate points on K via the Fréchet mean (9); i.e.: $\overline{\hat{F}(x)} = \text{argmin}_{y \in K} \sum_{n,q=1}^{N,Q} w_{n,q} d_g^2(y, y_{n,q})$.

Algorithm 1: Training Probabilistic Transformers for Exact Constraint Satisfaction

Input: Training Data $\{(x_t, y_t)\}_{t=1}^T \subseteq \mathbb{R}^n \times K$

Output: Probabilistic Transformer Network: $\hat{F} \triangleq \sum_{n=1}^{N} \sum_{q=1}^{Q} [\text{Softmax}_N \circ f(\cdot)]_n w_{n,q} \delta_{y_{n,q}}$

1 **Get Particles:** Use Generate K to generate $y_1, \ldots, y_S \in K$
for $n = 1, \ldots, N$ **do**
  $\{s_q\}_{q=1}^Q \leftarrow \text{argsort}_Q \{\|y_s - Y_n\|\}_{s=1}^S$
  $\{y_{n,q}\}_{q=1}^Q \leftarrow \{y_{s_q}\}_{q=1}^Q$
endfor

2 **Train Model:**
Get Labels: **for** $t \leq T$ **do**
  for $n \leq N$ **do**
    $(L_t)_n \leftarrow I(\|y_t - Y_n\| \leq \min_{m=1,\ldots,N} \|y_t - Y_m\|)$
  endfor
endfor

3 $\hat{f} \leftarrow \text{argmin}_{\hat{f}} \sum_{t=1}^{T} \sum_{n=1}^{N} \|(L_t)_n - [\text{Softmax}_N \circ \hat{f}(x_t)]_n\|^2$

return $\hat{F}(\cdot) \triangleq \sum_{n=1}^{N} \text{Softmax}_N \circ \hat{f}(\cdot)_n \delta_{Y_n}$

Remark B.2 (Prediction)**.** *Predictions can be made using $\hat{F}$ by either applying an expectation, in which case classical tranformer networks of Vaswani et al. (2017) are recovered, using the Fréchet mean as a final layer if K is a geodesically convex subset of a Riemannian submanifold of $\mathbb{R}^m$, or taking the most-likely particle if noting more is known of K other than its point-set.*

We now address question (ii).

B.2 Performance on a Toy Non-Convex Problem

Let $K \subseteq \mathbb{R}^2$ be a 2-dimensional sphere in $\mathbb{R}^3$. Let a, b, c be independently drawn from a uniform distribution on $[0, 1]$ and let A be a 2×10^3 random matrix with i.i.d. standard Gaussian entries.

Let $f = \tilde{f}(Ax)$ be the random K-valued function where $u_i = ax_i^2 + bx_i + c$, for $i = 1, 2$, and $\tilde{f}(u) \triangleq (\cos(\tilde{f}(u)_1)\sin(\tilde{f}(u)_2)), \sin(\tilde{f}(u)_1)\sin(\tilde{f}(u)_2)), \cos(\tilde{f}(u)_1)\cos(\tilde{f}(u)_2)))$. Therefore, A projects $\mathbb{R}^{10^3}$ onto the latent low-dimensional space $\mathbb{R}^2$ and $\tilde{f}$ sends data in $\mathbb{R}^2$ to a point on the sphere obtain by a random polynomial transformation of its spherical coordinates (which is a non-convex constraint set).

We independently repeat this experiment 500-times, generating a random f each time and generating $1k$ training inputs $\{x_t\}_{t=1}^{10^3} \subseteq [0,1]^{10^3}$ (resp. 100 testing inputs) by independently and uniformly sampling from $[0,1]^{10^3}$ and producing $1k$ corresponding training (resp. 100 testing) outputs $\{f(x_t)\}_{t=1}^{10^3} \subset K$. For each independent experiment, a probabilistic transformer network $\mathbb{P}$-Trans.) is trained using Algorithm 3, and benchmarked against a deep feedforward network (MLP) and a classical transformer network (Trans.). Table 3 reports the average and standard deviation, across all experiments, of the test-set MSE and the distance to the constraint set (d_K) of the test-set predictions for each learning model.

Figure 5 shows that, high emphasis is placed on constraint satisfaction ($\lambda \in [0, 0.75]$) then the $\mathbb{P}$-Trans.+Fréchet model outperforms the benchmark models. As the emphasis parameter λ approaches the critical value of $\approx .75$ then, the MSE dominates the constraint satisfaction metric d_K and the $\mathbb{P}$-Trans.+Fréchet's larger average test MSE is larger than that of the MLP and Trans. models. This validates the error terms ϵ_K and the factor $k\,\mathrm{Lip}(\Phi^{-1})d$ in Theorem 2.7 (ii), reflected in Table 3, which is due to the decoder network $\hat{\mathcal{D}}$ in f approximating a random projection of $\mathbb{R}^3$ onto K.

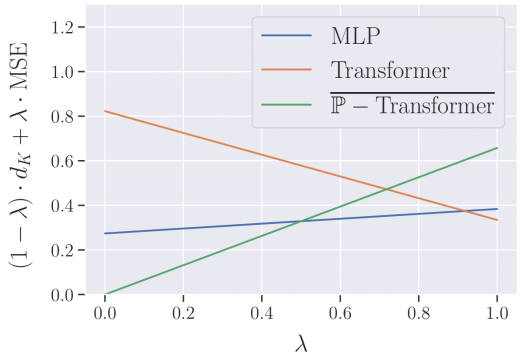

Figure 5: Performance for varying importance on constraint satisfaction vs. MSE.

Therefore, we find that our probabilistic transformer network is both implementable, and that, as expected, it offers good predictive performance even while enforcing non-convex constrained. In other words, we have obtained positive answers to the natural questions (i) and (ii) posed at the start of Appendix B.

	d_K		MSE	
	Mean	Std	Mean	Std
MLP	0.274	0.106	0.384	0.034
Transformer	0.822	0.100	0.334	0.009
$\mathbb{P}$-Transformer	0.000	0.000	0.657	0.059

Table 3: Performance metrics across all 500 experiments.

Table 3 emphasizes that classical transformer networks are not built to handle non-convex constraints. Indeed, the poor performance of the transformer network, with respect to the d_K, is due to most its predictions lying inside the sphere (which is hollow).

Further study into training algorithms for our model, and detailed ablation of the model parameters are topics of focus in forthcoming research. Nevertheless, we have obtained an affirmative answer

to both our questions (i) and (ii). Namely, we have shown that our probabilistic transformer morel is trainable via a simple procedure such as Algorithm 3.

B.3 Examining the impact of K's Geometry on Transformer Networks

To gain further insight into how probabilistic transformers encode geometric priors, we will examine the impact of perturbations to K on the probabilistic transformer's approximation capabilities. We consider toy illustrations beginning with the convex setting before moving on to the fully non-convex setting where no projections, charts, or even a Riemannian structure is available.

We further underline that step 3 in Algorithm 3 may be performed in a variety of ways and, unlike the previous experiments, all networks in this section are obtained by randomizing their hidden weights and only training their final layer. Theoretical guarantees for this approach has become relatively well understood (Louart et al., 2018; Gonon et al., 2020a;b). Implicitly, our examples also show that probabilistic transformers can equally be integrated into domains where randomized models such as extreme learning machines (ELMs) are typical; e.g. in the reservoir computing Lukoševičius & Jaeger (2009); Grigoryeva & Ortega (2018; 2019).

This section's primary goal is to experimentally validate the main quantitative claim made implicitly in our main result; i.e. Theorem 2.7. Namely, we verify that:

"The model complexities in Table 1 are independent of K's geometry."

That is, the approximation quality of any optimized probabilistic transformer network only depends on the involved dimensions. Expressed another way, we empirically validate our result that the probabilistic transformer networks can encode any geometric prior with the model complexity agnostic of K's geometry.

Accordingly, all model architectures' hyperparameters are kept fixed across all experiments. Each experiment reports the probabilistic transformer's MSE relative to the benchmark MLP model $\left(\frac{\text{MSE}}{\text{MSE-MLP}}\right)$.

Our result is validated upon observing that the probabilistic transformer model's $\frac{\text{MSE}}{\text{MSE-MLP}}$ is of the same order across all experiments. In other words: *probabilistic transformers can approximate a K-valued function while simultaneously encoding K's geometry with the same efficiency as an MLP trained only to approximate f that ignores K's geometry.*

Each toy experiment is trained on a dataset of 900 instances and tested on a dataset of 100 instances. We maintain the coloring scheme of Figure 5 in all our subsequent plots, namely the MLP is colored in blue, the Transformer is colored in orange, and the $\mathbb{P}$-Transformer is colored in green.

B.3.1 Convex Constraints

Our first set of examples concern the case where K is a *convex* constraint set, as studied in Corollaries 2.9. In each instance, we generate a random target function f, mapping $\mathbb{R}^2$ to K. The function f is defined via the following two-step procedure:

$$f : \mathbb{R} \xrightarrow{\tilde{f}} \mathbb{R}^2 \xrightarrow{P_K} K;$$

where $\tilde{f} : \mathbb{R}^2 \ni x \mapsto Ax$ is a random rotation matrix re-scaled by a factor of 1.5 with the angle sampled uniformly from $[0, 2\pi]$ and where $P_K : \mathbb{R}^2 \ni x \mapsto \operatorname{argmin}_{y \in K} \|y - x\|$ is the *metric projection* onto the K, which exist in this context by (Motzkin, 1935). Figures 6 and 9 illustrate this two step transformation by first representing the uniformly generated input data in violet then, illustrating their images under $\tilde{f}$ in light green, and finally plotting their value under f in dark green.

We perform our illustrations in the visualizable two dimensional case where K is either the square $[-1, 1]^2$ or the disk $\{y \in \mathbb{R}^2 : \|y\| \leq 1\|$. This is because the projection operators (P_K) are readily interpretable from their closed-form formulations (Bauschke & Combettes, 2011). Respectively, these are given by $P_K(x) = (\min\{\max\{x_i, -1\}, 1\})_{i=1}^2$ and $P_K(x) := \frac{x}{\max\{1, \|x\|\}}$.

Figures 7 and 10 demonstrate the constraint set (K) in green and the particles, which populate Y's rows, in violet. These are generated randomly by first sampling uniformly from $[-2, 2]^2$ and then projecting each sample onto K via P_K. Figures 7 and 10 illustrate the role of the particles

defining the probabilistic attention mechanism, defined in (3); namely, they identify the points in K on which any output may *approximately lie*. Thus, the role of the encoder and decoder networks can be summarized as learning to classify which input is closest to which particle.

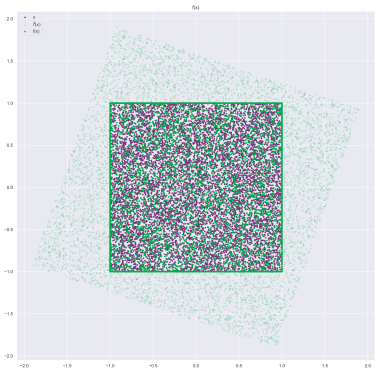

Figure 6: $x \mapsto f(x)$.

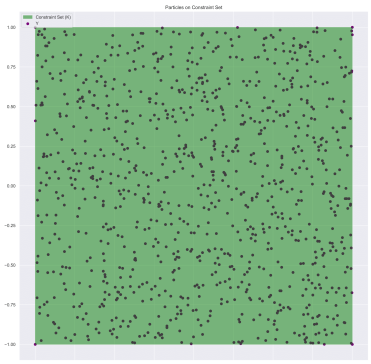

Figure 7: Particles (Y) on Constraint Set (K).

Therefore, at high-level, the probabilistic attention mechanism 3 quantizes the constraint set K. The (simplified) classical Attention mechanism of (4) implements a (convex) interpolation between the particles quantizing K and an analogous statement is true of Riemannian analogue (Section 2.3.2). For general K, however, such interpolations within K can be impossible or unclear how to implement them. Nevertheless, the probabilistic attention mechanism never faces such a difficulty since it explicitly "interpolates" in $\mathcal{P}_1(K)$ and not on K directly.

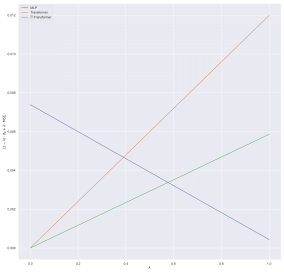

Figure 8: Performance: MSE vs. d_K.

	$\frac{\text{MSE}}{\text{MSE MLP}}$	MSE	d_K
MLP	1	4.35e-04	7.39e-03
Transformer	4.81	1.20e-02	0.00e+00
$\mathbb{P}$-Transformer	4.01	5.86e-03	0.00e+00

Table 4: Performance Metrics

We obtain analogous results to the 500 experiments performed in the case where K is geodesically-convex in Section B.2. Just as in Figure 5, Figures 8 and 11 show that the transformer can simultaneously encode K and approximate f, wheras the MLP cannot. More precisely, in each case, if at-least roughly equal importance is placed on predictive accuracy (MSE) and constraint satisfaction (d_K) then, the transformer models offer the best performance. This is equally reflected in the test set performance metrics of Tables B.3.1 and B.3.1 which are consistent with the findings of Table 3.

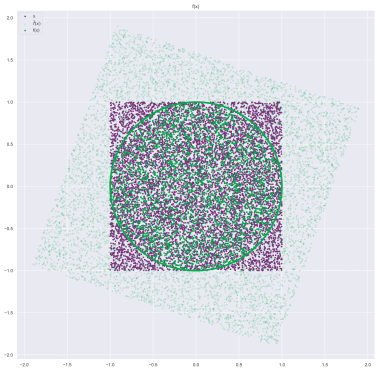

Figure 9: $x \mapsto f(x)$.

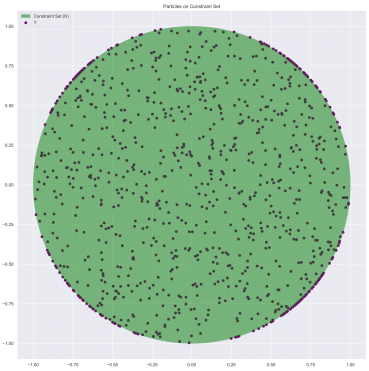

Figure 10: Particles (Y) on the constraint set (K).

At times, when K's geometric is sufficiently simple we the transformer can outperform the probabilistic transformer. This is not surprising, since Corollary 2.9 guaranteed that the transformer universal in this setting an exactly implements the K's geometry. Nevertheless, in both instances, the MLP cannot compete when encoding the geometric prior defined by the constraint set K.

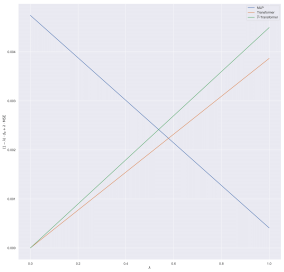

Figure 11: Performance: MSE vs. d_K.

	$\frac{\text{MSE}}{\text{MSE MLP}}$	MSE	d_K
MLP	1	4.09e-04	4.75e-03
Transformer	2.98	3.87e-03	0.00e+00
$\mathbb{P}$-Transformer	3.41	4.50e-03	6.11e-19

Table 5: Performance Metrics

B.3.2 Fully Non-Convex Constraints

Let us study the milieu in which probabilistic transformer is the only universal approximator capable of constraint satisfaction (unlike the case where K is convex and we showed that the transformer filled this role). Specifically, we consider the case where K does not admit a single chart (since it has non-trivial homotopy), nor is there a well-defined projection operator of some $\mathbb{R}^m$ onto K.

Analogously to the convex situation investigated in Section B.3.1, we define

$$f : \mathbb{R} \xrightarrow{\tilde{f}} \mathbb{R} \xrightarrow{\rho} K;$$

where $\tilde{f}(x) \triangleq \sum_{i=0}^{5} \beta_i x^i$ is a (random) quintic polynomial function with $\beta_i \sim N(0,1)$, and the constraint set's geometry is defined by K; where $\rho : \mathbb{R} \to \mathbb{R}^2$. In this experiment, we also allow the training data to be perturbed by multivariate Gaussian noise with variance 10^{-1}.

Similarly to Figures 6 and 9, in Figures 12 and 15, we use a color coded visualization method to understand f. Sample points from $[-10, 10]$ and label them with a color gradient ranging from pink to blue such that pinkish points are close to -10 and blueish points are a nearer to 10. The image $(f(x)$ of each input (x) on K is illustrated using the same colour as x did. This coloring helps us visualize the winding of f around K.

Nevertheless, as in the convex case, we can generate particles on K by first sampling from $[-10, 10]$ and then mapping them onto K using ρ. Thus, even if no chart or projection operator is available, we can easily build probabilistic attention mechanisms.

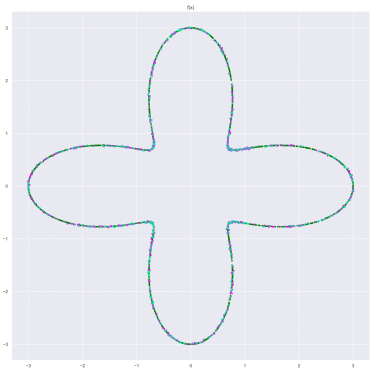

Figure 12: $x \mapsto f(x)$.

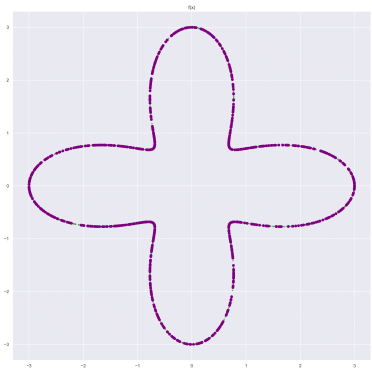

Figure 13: Particles (Y) on Constraint Set (K).

In Figures 12 and 13, the map defining K's geometry is $\rho(y) \triangleq (2\cos(y)^2 + 1) \cdot (\cos(y/3), \sin(y/3))$. Figure 14 and Table B.3.2 show that our probabilistic transformer network's performance is "robust to changes of geometric priors", in the sense that the relative performance of our models is entirely analogous to the above experiments where K was convex or it was a geodesically convex patch on a Riemannian manifold.

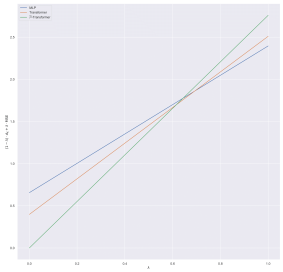

Figure 14: Performance: MSE vs. d_K.

	$\frac{\text{MSE}}{\text{MSE MLP}}$	MSE	d_K
MLP	1	2.40e+00	6.56e-01
Trans.	1.01	2.51e+00	3.98e-01
$\mathbb{P}$-Trans.	1.05	2.76e+00	0.00e+00

Table 6: Performance Metrics

We complete our discussion by considering an instance where K's geometry is both non-convex and it is not a differentiable manifold (due to the self-intersecting point). This last toy example is illustrated in Figures 15 and 16 in which case K's geometry is the image of the map $\rho(y) \triangleq \Phi(\text{sinc}(y+1)(\cos(y/2), \sin(y/2)))$ where Φ is a randomly generated invertible feedforward network with invertible square weight matrices and $\tanh$ activation function (i.e.: a random homeomorphism on $\mathbb{R}^2$).

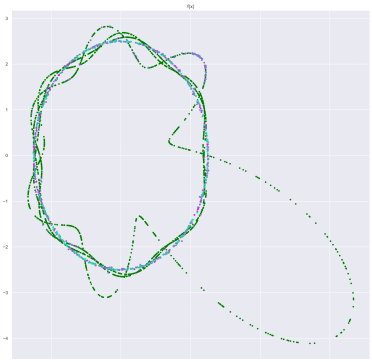

Figure 15: $x \mapsto f(x)$.

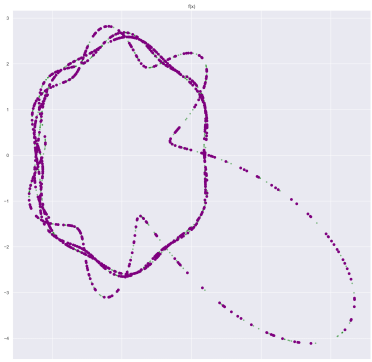

Figure 16: Particles (Y) on Constraint Set (K).

We conclude our study examining the impact of K's geometry on the probabilistic transformer's performance by noting that the probabilistic transformer's relative performance is analogous to its performance in the previous experiments. Figure 17 and Table B.3.2 reaffirm that the probabilistic transformer outperforms the MLP and the transformer network when the mixed objective of optimizing the MSE and the distance to the constraint set.

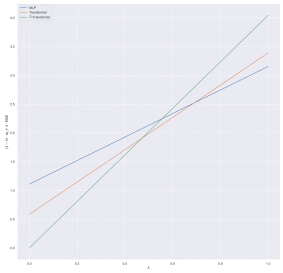

Figure 17: Performance: MSE vs. d_K.

	$\frac{\text{MSE}}{\text{MSE MLP}}$	MSE	d_K
MLP	1	3.15e+00	1.11e+00
Trans.	1.01	3.39e+00	5.84e-01
$\mathbb{P}$-Trans.	1.09	4.04e+00	0.00e+00

Table 7: Performance Metrics

This appendix showed that the probabilistic transformer is implementable, that it can indeed approximate functions while exactly encoding constraints, and that its performance doesn't degrade for more complicated geometries. In conclusion, probabilistic transformers can generically and canonically encode geometric priors without sacrificing the expressibility of more familiar deep learning models.

C Proofs

In what follows, we denote the set of couplings of two probability measures $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^n)$ on $\mathbb{R}^n$ by $\mathrm{Cpl}(\mu, \nu)$. I.e. these are Borel product measures π on $\mathbb{R}^n \times \mathbb{R}^n$ with respective marginals μ and ν. We begin by deriving some useful lemmata.

C.1 Lemmata

This section records lemmata that will be frequently be used throughout this paper's proofs. The lemmata's proofs are deferred until Section C.2.1 of this Appendix.

Note. ***For the reader interested in convex constraints:*** *We recognize that the results where K is convex follow the more general results where K is a geodesically convex subset of some embedded*

submanifold of $\mathbb{R}^m$. Nevertheless, so as to provide a self-contained reading to those focused on classical transformers or on convex constraints, independent proofs for both of these two cases.

C.1.1 LEMMATA IN THE CASE WHERE K IS CONVEX

The results are especially useful for results pertaining to convex constraint sets.

Lemma C.1 (Collapsing Measure-Valued Estimates for Convex Constraint Sets)**.** *Let $K \subseteq \mathbb{R}^m$ be non-empty, compact, and convex. Let $F \in C(\mathbb{R}^n, \mathcal{P}_1(K))$ and $f \in C(\mathbb{R}^n, K)$. For every $x \in \mathbb{R}^n$, the following hold:*

(i) ***Convex Constraints Hold:*** $\mathbb{E}_{Y\sim F(x)}[Y] \in K$,

(ii) ***Non-Expansive Distance:*** $\left\| f(x) - \mathbb{E}_{Y\sim F(x)}[Y] \right\| \leq \mathscr{W}_1\left(\delta_{f(x)}, F(x)\right)$.

Moreover, let $\epsilon > 0$ and some non-empty compact $C \subset \mathbb{R}^n$ be non-empty and compact. If $\max_{x\in C} \mathscr{W}_1(\delta_{f(x)}, F(x)) \leq \epsilon$ then, in addition we have that:

$$\max_{x\in C} \left\| f(x) - \mathbb{E}_{Y\sim F(x)}[Y] \right\| \leq \epsilon. \tag{12}$$

The next lemma, though immediate, is still helpful to write down explicitly as it clearly relates P-attention to Attention. For any $N \in \mathbb{N}_+$, we denote the standard N-simplex by Δ_N; i.e.:

$$\Delta_N \triangleq \left\{ w \in [0,1]^N : \sum_{n=1}^{N} w_n = 1 \right\}.$$

Lemma C.2 (An Identity: P-attention as implicit Attention)**.** *Let $\{y_{n,q}\}_{n=1,\ldots,N,q=1,\ldots,Q} \subseteq K \subset \mathbb{R}^m$, let Y be an $N \times 1 \times m$-array with $Y_n = Q^{-1}\sum_{q=1}^{Q} y_{n,q}$, and let $f(x) \in C(\mathbb{R}^n, \Delta_N)$. Then:*

$$\text{Attention}\left(f(x)|Y\right) = \mathbb{E}_{X\sim \text{P-attention}(F(x),Y)}\left[X\right].$$

C.1.2 LEMMATA IN THE CASE WHERE K IS A CLOSED GEODESIC BALL

We now consider the analogue of Lemma C.1, in the case where K is geodesically convex of controlled radius[1]. A K subset of (M, g) is called *geodesically convex* if for every two points $y_1, y_2 \in K$ there is a unique geodesic (Riemannian distance minimizing curve) joining y_1 to y_2.

Lemma C.3. *Let (M, g) be a connected Riemannian manifold with sectional curvatures uniformly bounded-above by $C \geq 0$ and which is complete as a metric space under d_g. Fix $y_0 \in M$,*

$$0 < \rho < 2^{-1} \min\left\{ \text{inj}_g(y_0), \frac{\pi}{\sqrt{C}} \right\} \tag{13}$$

(where, following Afsari (2011), $\frac{1}{\sqrt{C}} \triangleq \infty$ whenever $C \leq 0$), and let K be a non-empty, compact, and geodesically convex subset of $\overline{B(y_0, \rho)}$. Then, the "Fréchet mean" function:

$$\begin{aligned} \mathcal{P}_1(K) &\to \overline{B(y_0,\rho)} \\ \mathbb{P} &\mapsto \text{argmin}_{y\in K}\ \mathbb{E}_{Y\sim\mathbb{P}}\left[d_g^2(y, Y)\right], \end{aligned} \tag{14}$$

is a well-defined (i.e.: single-valued) and non-expansive (i.e.: 1-Lipschitz) function. Furthermore, if $\mathbb{P}$ is finitely-supported then:

$$\overline{\mathbb{P}} \in K. \tag{15}$$

C.2 EXCEPTIONAL CLOSED-FORM FOR WASSERSTEIN DISTANCE

The following result is folklore in the optimal transport community. Since its statement is difficult to track down, we record the statement and derive its proof here, for a self-contained reading.

Lemma C.4 (Closed-Form Expression for Wasserstein Distance to Pointmass)**.** *Let $K \subseteq \mathbb{R}^m$ be non-empty and compact, let y be in K, and let $\mathbb{P} \in \mathcal{P}_1(K)$. Then:*

$$\mathscr{W}_1(\mathbb{P}, \delta_y) = \mathbb{E}_{Y\sim\mathbb{P}}[\|Y - y\|].$$

[1] We use the terminology controlled in direct analogy with (Kratsios & Papon, 2021, Theorem 10).

C.2.1 PROOFS OF LEMMATA

Proof of Lemma C.1. Fix $\mu \triangleq F(x)$. We first show (ii). Since $\mathbb{R}^n$ is a Banach space, then (Bru et al., 1993) implies that there exists a unique contracting barycenter map on $\mathcal{P}_1(\mathbb{R}^n)$; i.e.: a Lipschitz map $\beta_{\mathbb{R}^n} : \mathcal{P}_1(\mathbb{R}^n) \to \mathbb{R}^n$ satisfying $\beta_{\mathbb{R}^n}(\delta_x) = x$ for all $x \in \mathbb{R}^n$, whose Lipschitz constant $\mathrm{Lip}(\beta_{\mathbb{R}^n})$ is at most 1. Moreover, the result guarantees that the barycenter map is linear an given by the Bochner integral (i.e. the usual vector-valued expectation of a $\mathbb{R}^n$-valued random-vector):

$$\beta_{\mathbb{R}^n} : \mathcal{P}_1(\mathbb{R}^n) \ni \mu \mapsto \mathbb{E}_{Y\sim\mu}[Y] \in \mathbb{R}^n. \tag{16}$$

Therefore, we conclude that:

$$\begin{aligned} \|f(x) - \mathbb{E}_{Y\sim\mu}[Y]\| =& \|f(x) - \beta_{\mathbb{R}^n}(F(x))\| \\ =& \|\beta_{\mathbb{R}^n}(\delta_{f(x)}) - \beta_{\mathbb{R}^n}(F(x))\| \\ \leq& \mathscr{W}_1\left(\delta_{f(x)}, F(x)\right). \end{aligned} \tag{17}$$

This gives (ii). Furthermore, if the right-hand side of (17) is upper-bounded by a constant $\epsilon > 0$, uniformly over C, then so must be the left-hand side. This gives (12).

We now show (i). Since $K \subset \mathbb{R}^n$, we may view $\mathcal{P}_1(K)$ as a subspace of $\mathcal{P}_1(\mathbb{R}^n)$. Thus, $\beta_{\mathbb{R}^n}|_{\mathcal{P}_1(K)}$ satisfies $\beta_{\mathbb{R}^n}(\delta_x) = x$ for all $x \in K$. Moreover, we may view (16) as a map on $\mathcal{P}_1(K)$. Therefore, if $\mu \in \mathcal{P}_1(K)$ then, any $\mathbb{R}^n$-valued random-vector Y with law μ, by definition, μ-a.s. takes values in K. Since K is convex, and $\mu \in \mathcal{P}_1(K)$ (i.e. $\mathbb{E}_{Y\sim\mu}[|Y|] < \infty$) then the formulation of Jensen's inequality given in (Dudley, 2002, Theorem 10.2.6) guarantees that

$$\mathbb{E}_{Y\sim\mu}[Y] \in K. \tag{18}$$

Hence, we may refine (16) to state: $\beta_{\mathbb{R}^n}|_{\mathcal{P}_1(K)}$ is 1-Lipschitz and satisfies

$$\beta_{\mathbb{R}^n}|_{\mathcal{P}_1(K)} : \mathcal{P}_1(K) \ni \mu \mapsto \mathbb{E}_{Y\sim\mu}[Y] \in K. \tag{19}$$

Thus (i) holds. □

Proof of Lemma C.2. Follows directly from the linearity of integration and the fact that integration against a pointmass is just point-evaluation. □

Proof of Lemma C.4. By definition of the Wasserstein distance between $\mathbb{P}$ and δ_y we have that:

$$\mathscr{W}_1(\mathbb{P}, \delta_y) = \inf_{\pi\in\mathrm{Cpl}(\mathbb{P},\delta_y)} \mathbb{E}_{(Y_1,Y_2)\sim\pi}[\|Y_1 - Y_2\|]. \tag{20}$$

Since $\mathbb{P} \otimes \delta_y \in \mathrm{Cpl}(\mathbb{P}, \delta_y)$ (e.g. see (Villani, 2009, Page 6)) then it is enough to show that if π is a coupling in $\mathrm{Cpl}(\mathbb{P}, \delta_y)$ then $\pi = \mathbb{P} \otimes \delta_y$. We show this now.

Let $B_1, B_2 \subseteq K$ be Borel and let $\pi \in \mathrm{Cpl}(\mathbb{P}, \delta_y)$. If $y \in B_2$, then $\mathbb{P}(B_1) = \pi(B_1 \times K) \geq \pi(B_1 \times B_2) \geq \pi(B_1 \times \{y\})$. Therefore, $1 - \mathbb{P}(B_1) \geq \pi(K \times \{y\}) - \pi(B \times \{y\})$; thus, $\pi(B_1 \times \{y\}) \leq \mathbb{P}(B_1)$. Therefore, $\mathbb{P}(B_1) \leq \pi(B_1 \times \{y\}) \leq \pi(B_1 \times B_2)$; whence, $\pi(B_1 \times B_2) = \nu(B_2)\delta_y(B_1) \overset{\text{(def)}}{=} \nu \otimes \delta_y(B_2 \times B_1)$. Now, suppose that $y \notin B_2$ then, $\pi(B_1 \times B_2) \leq \pi(K \times B_2) = \delta_y(B_2) = 0$. We have show that $\pi = \nu \otimes \delta_y$. Hence, (20) reduces to:

$$\begin{aligned} \mathscr{W}_1(\mathbb{P}, \delta_y) =& \inf_{\pi\in\mathrm{Cpl}(\mathbb{P},\delta_y)} \mathbb{E}_{(Y_1,Y_2)\sim\pi}[\|Y_1 - Y_2\|] \\ =& \mathbb{E}_{(Y_1,Y_2)\sim\mathbb{P}\otimes\delta_y}[\|Y_1 - Y_2\|] \\ =& \mathbb{E}_{Y_1\sim\mathbb{P}}[\mathbb{E}_{Y_2\sim\delta_y}[\|Y_1 - Y_2\|]] \qquad (21) \\ =& \mathbb{E}_{Y_1\sim\mathbb{P}}[\|Y_1 - y\|]; \qquad (22) \end{aligned}$$

where we have applied the Fubini-Tonelli Theorem (see (Kallenberg, 2021, Theorem 1.27)) in (21) and the definition of a pointmass to derive (22). □

Proof of Lemma C.3. We first observe that, since (M, g) is connected and complete as a metric space, then by the Hopf-Rinow Theorem ((Jost, 2017, Theorem 1.7.1)) (M, g) is a complete Riemannian manifold (sometimes also called a geodesically complete Riemannian manifold; see (Jost, 2017, Definition 1.7.1).

Fix a $\mathbb{P} \in \mathcal{P}_1(K)$. Since K is compact, then $\mathbb{P} \in \mathcal{P}_2(K)$. The completeness of (M, g) (as a Riemannian manifold) and the facts that K is a non-empty geodesically convex subset of $B(y_0 0, \rho)$ (where ρ satisfies (13)) implies that the conditions of (Afsari, 2011, Theorem 2.1) are met; whence, the "Fréchet mean function" of (14) is well-defined function from $\mathcal{P}_1(K)$ to $\overline{B(y_0, \rho)}$. It remains to show that it is 1-Lipschitz. Since $\rho < \mathrm{inj}_g(y_0)$, then the remark on (Jost, 2017, Page 299) implies that (Jost, 2017, Theorem 6.9.2) holds; therefore, for any $y_1, y_2, y_3 \in B(p, \rho)$ the map:

$$[0, 1] \ni t \mapsto d^2(\gamma_{[y_1, y_2]}(t), \gamma_{[y_1, y_3]}(t)) \in [0, \infty),$$

is convex. We may now conclude our proof by arguing analogously to (Sturm, 2003, Theorem 6.3's proof) . Fix $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_1(K)$ and let $\pi \in \mathcal{P}(K \times K)$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Then, applying Jensen's inequality, we have that:

$$d_g(\bar{P}, \bar{Q}) \leq \int d^2(y_1, y_2)\, \pi(d(y_1, y_2)). \tag{23}$$

Since we have just showed that right-hand side of (23) holds for any such π. Consequently, taking the infimum over all such π implies that:

$$d_g(\bar{P}, \bar{Q}) \leq \mathscr{W}_1(\mathbb{P}, \mathbb{Q}).$$

Thus, (14) is 1-Lipschitz.

For the last claim, suppose that $\mathbb{P} \in \mathcal{P}_1(K)$ is finitely supported. Since K is geodesically convex and since $\mathbb{P}$ is finitely supported then (Afsari, 2011, Theorem 3.4 (i)) implies that $\bar{\mathbb{P}}$ is an element of the smallest closed geodesically convex subset $C_{\mathbb{P}}$ containing the support of $\mathbb{P}$; since K itself is itself closed and geodesically convex then we infer that $C_{\mathbb{P}} \subseteq K$. Thus, (14) takes values in K. □

C.3 Proof of Theorem 2.2

Proof of Theorem 2.2. Since $K \subseteq \mathbb{R}^n$ is non-empty and compact, for each $x \in \mathbb{R}^n$ the set C_x is closed and has non-empty intersection with K, thus each $C_x \cap K$ is compact. Thus, the map $\varphi : \mathbb{R}^n \ni x \mapsto C_x \cap K \in 2^{\mathbb{R}^m}$ is a non-empty and compact-valued multifunction. Moreover, by (Aliprantis & Border, 2006, 18.4 Lemma) φ is weakly-measurable since C is and so is the correspondence $\mathbb{R}^n \ni x \mapsto K \in 2^{\mathbb{R}^m}$. Thus, the hemicontinuity of φ and the assumptions made on L are such that the Measurable Maximum Theorem (see (Aliprantis & Border, 2006, Theorem 18.19)) applies; whence, the "optimality" sets

$$\mathcal{O}(x) \triangleq \mathrm{argmax}_{y \in \varphi(x)} -L(x, y) \in \mathbb{R} = \mathrm{argmin}_{y \in C_x \cap K} L(x, y), \tag{24}$$

are a well-defined for each $x \in \mathbb{R}^n$ and, there exists a Borel measurable function $S : \mathbb{R}^n \to \mathbb{R}^m$ satisfying the "optimal selection condition":

$$S(x) \in \mathcal{O}(x) \qquad (\forall x \in \mathbb{R}^n). \tag{25}$$

Since $\mathbb{R}^n$ is a complete and separable metric space and since $\mathbb{P}$ is a Borel probability measure on $\mathbb{R}^n$, then by (Klenke, 2014, Theorem 13.6), $\mathbb{P}$ is a Radon measure on $\mathbb{R}^n$. Since S is Borel measurable, $\mathbb{R}^n$ and $\mathbb{R}^m$ are locally-compact and second-countable topological spaces, and since $\mathbb{P}$ is a Radon measure on $\mathbb{R}^n$, then Lusin's Theorem (as formulated in (Klenke, 2014, Excersize 13.1.3)) implies that, for every $\epsilon \in (0, 1]$, there is a compact subset $\mathcal{X}_\epsilon \subseteq \mathbb{R}^n$ on which $S|_{\mathcal{X}_\epsilon}$ is continuous and $\mathbb{P}(\mathcal{X}_\epsilon) \geq 1 - \epsilon$. By (Villani, 2009, Point 5 - Page 99), the map $\mathbb{R}^m : y \mapsto \delta_y \in \mathcal{P}(\mathbb{R}^m)$ is an isometry. In particular, the map $\mathbb{R}^m : y \mapsto \delta_y \in \mathcal{P}(\mathbb{R}^m)$ is continuous. Hence, $S^\star : \mathcal{X}_\epsilon \ni x \mapsto \delta_{S(x)} \in \mathcal{P}_1(\mathbb{R}^m)$ is continuous. However, by construction, $S(x) \in \varphi(x) \subseteq K$; thus, $S^\star$ defines a map with codomain $\mathcal{P}_1(K)$.

Therefore, (Kratsios, 2021, Theorem 3) implies that there exists an $\hat{F}$ of the form

$$\hat{F} : \mathbb{R}^n \ni x \mapsto \sum_{n=1}^{N} [\mathrm{Softmax}_N(\hat{f}(x))]_n \frac{1}{Q} \sum_{q=1}^{Q} \delta_{k_{n,q}} \in \mathcal{P}_1(\mathbb{R}^m), \tag{26}$$

satisfying:

$$\max_{x \in \mathcal{X}_\epsilon} \mathscr{W}_1 \left(\hat{F}(x), S^\star(x) \right) < \epsilon. \tag{27}$$

Grouping the sums $\sum_{n=1}^{N}$ and $\sum_{q=1}^{Q}$ and the weights $Q^{-1}[\mathrm{Softmax}_N(\hat{f}(x))_n]$ in (26), we may rewrite (26) in the form (6).

By construction, for each $x \in \mathcal{X}_\epsilon$ we have that $S(x) \in \mathcal{O}(x)$. Thus, (25) implies that:

$$\max_{x\in\mathcal{X}_\epsilon} \mathscr{W}_1\left(\hat{F}(x), \inf_{y^\star\in\mathcal{O}(x)} \delta_{y}^{\star}\right) \leq \max_{x\in\mathcal{X}_\epsilon} \mathscr{W}_1\left(\hat{F}(x), S^\star(x)\right) < \epsilon. \tag{28}$$

This gives (ii).

Now, by construction, each $y_1, \ldots, y_{N,Q} \in K$. Therefore, for each $x \in \mathbb{R}^n$, $\hat{F}(x)$ is supported in K and, moreover, $\hat{F}(x) \in \mathcal{P}_1(K)$. Thus, (i) holds. □

C.4 Proof of Theorem 2.7

We make use of the following notation during Theorem 2.7's proof. For $d \leq m$, $d \in \mathbb{N}_+$, we denote $p_d^m : \mathbb{R}^m \ni (x_1, \ldots, x_m) \to (x_1, \ldots, x_n) \in \mathbb{R}^d$ and similarly, $\iota_d^m : \mathbb{R}^d \ni (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n, 0, \ldots, 0) \in \mathbb{R}^m$. Before proceeding, we also emphasize the following identities: if $x_1, \ldots, x_n \in \mathbb{R}$ then $\iota_d^m \circ p_d^m(x_1, \ldots, x_n, 0, \ldots, 0) = (x_1, \ldots, x_n, 0, \ldots, 0)$ and conversely, $p_m^d \circ \iota_d^m$ is the identity on $\mathbb{R}^d$.

Proof of Theorem 2.7. Let $k \in \mathbb{N}_+$, let $\mathcal{X} \subseteq [0,1]^n$ be non-empty, and let $f \in C_{tr}^k(\mathcal{X}, K)$. Since $f \in C_{tr}^k(\mathcal{X}, K)$ then, there exists a k-times continuously differentiable $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ such that: for every $x \in \mathcal{X}$, we have that:

$$\boldsymbol{f}(x) = f(x). \tag{29}$$

Note that $\mathcal{X} \subseteq [0,1]^n$. We begin by building our encoder to approximate $p_d^m \circ \Phi^{-1} \circ \boldsymbol{f} \in C([0,1]^n, \mathbb{R}^m)$. By Assumption 2.4 and the fact that $f(\mathcal{X}) \subseteq K$ we have that, for each $x \in \mathcal{X}$:

$$0 \leq \inf_{y\in K} L(x,y) \leq L(x, f(x)) \leq l\left(\|f(x) - f(x)\|\right) = l(0) = 0.$$

Therefore, by (29), for each $x \in \mathcal{X}$, we know that $\{\boldsymbol{f}(x)\} \subseteq \operatorname*{argmin}_{y\in K} L(x,y)$. In particular, for each $x \in \mathcal{X}$ we have that $\operatorname*{argmin}_{y\in K} L(x,y)$ is non-empty and therefore, for any $\hat{F} : [0,1]^n \to \mathcal{P}_1(K)$ we may compute:

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{F}(x), \operatorname*{argmin}_{y\in K} L(x,y)\right) &\overset{(\mathrm{def})}{=} \sup_{x\in\mathcal{X}} \inf_{y^\star\in\operatorname*{argmin}_{y\in K} L(x,y)} \mathscr{W}_1\left(\hat{F}(x), \delta_{y^\star}\right) \\
&\leq \sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{F}(x), \delta_{f(x)}\right) \\
&= \sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{F}(x), \delta_{\boldsymbol{f}(x)}\right).
\end{aligned} \tag{30}$$

Therefore, it is enough to construct models $\hat{\mathcal{D}}$ and $\hat{\mathcal{E}}$ such that the composite model $\hat{F} = \hat{\mathcal{D}} \circ \hat{\mathcal{E}}$ controls the approximation error on the right-hand side of (30). The remainder and bulk of the proof is devoted to precisely this task.

NB, by Assumption 2.5 we have that $\boldsymbol{f}(x) = \Phi^{-1} \circ \iota_d^m \circ (p_d^m \circ \Phi \circ \boldsymbol{f})(x)$. Now, since p_d^m is a linear map between finite-dimensional normed spaces then, is is analytic, and therefore it is smooth. Moreover, by hypothesis, both Φ and Φ^{-1} are both also smooth. Thus, the map $p_d^m \circ \Phi \circ \boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^d$ is k-times continuously differentiable.

Since $p_d^m \circ \boldsymbol{f}$ is k-times continuously differentiable then, by (Kratsios & Papon, 2021, Proposition 17), $[0,1]^n$ is efficient for $p_d^m \circ \Phi \circ \boldsymbol{f}$ (in the sense of (Kratsios & Papon, 2021, Definition 16)); thus, (Kratsios & Papon, 2021, Corollary 43) (activation function parameter α set to $\alpha = 0$. Thus, σ_0 is non-affine, continuous, and piecewise linear) implies that there is a $\hat{\mathcal{E}} \in \mathcal{NN}_{n,d}^{\sigma_0}$ satisfying:

$$\sup_{x\in[0,1]^n} \|p_d^m \circ \Phi \circ \boldsymbol{f}(x) - \hat{\mathcal{E}}(x)\| < \epsilon_f, \tag{31}$$

Furthermore, $\hat{\mathcal{E}}$ also satisfies the following quantitative estimates:

(i-$\mathscr{E}$) $\hat{\mathcal{E}}$ has width $d \leq W \leq d(4n+10)$,

(ii-$\mathscr{E}$) $\hat{\mathcal{E}}$ has depth of the order $\mathscr{O}(d + d\epsilon_f^{\frac{2n}{3(kn+1)} - \frac{2n}{kn+1}})$,

(iii-$\mathscr{E}$) The number of trainable parameters determining $\hat{\mathcal{E}}$ are of the order $\mathscr{O}(d(d^2+1)\epsilon_f^{-\frac{2n}{3(kn+1)}})$.

Since $\mathcal{X} \subseteq [0,1]^n$ and $f(\mathcal{X}) = \boldsymbol{f}(\mathcal{X}) \subseteq K$ then, Assumption 2.5 implies that $p_d^m \circ \Phi \circ f(\mathcal{X}) \subseteq p_d^m(\Phi(K)) \subseteq \mathbb{R}^d$. This together with (31) implies that:

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \|\hat{\mathcal{E}}(x) - p_d^m(\Phi(K))\| &\overset{(\text{def})}{=} \sup_{x\in\mathcal{X}} \inf_{y\in p_d^m(\Phi(K))} \|\hat{\mathcal{E}}(x) - y\| \\
&\leq \sup_{x\in\mathcal{X}} \inf_{y\in p_d^m\circ\Phi\circ f(\mathcal{X})} \|\hat{\mathcal{E}}(x) - y\| \\
&\leq \sup_{x\in\mathcal{X}} \|\hat{\mathcal{E}}(x) - p_d^m \circ \Phi \circ f(x)\| \\
&= \sup_{x\in\mathcal{X}} \|\hat{\mathcal{E}}(x) - p_d^m \circ \Phi \circ \boldsymbol{f}(x)\| \\
&\leq \epsilon_f.
\end{aligned} \tag{32}$$

Thus, (32) indicates that $\hat{\mathcal{E}}$ need not take values in $p_d^m(\Phi(K))$ but, it does take values in the following closed and bounded subset of $\mathbb{R}^m$:

$$\Phi(K)_{\epsilon_f} \triangleq \{y \in \mathbb{R}^m : \|y - p_d^m(\Phi(K))\| \leq \epsilon_f\}$$

By (Munkres, 2000, Theorem 26.5), $\Phi(K)$ is compact since K is compact and since Φ is continuous. Thus, the Heine-Borel Theorem (see (Munkres, 2000, Theorem 27.3)) implies that $\Phi(K)_{\epsilon_f}$ is compact as it is closed and bounded (because $\Phi(K)$ is closed and bounded). Thus, we can approximate functions from $p_d^m(\Phi(K))_{\epsilon_f}$ to $\mathcal{P}_1(K)$ uniformly using the main result of Kratsios (2021). Specifically, we will approximate a random project (in the sense [2]of (Ohta, 2009, Definition 3.1)) of $\mathbb{R}^d$ onto $p_d^m(\Phi(K))$, uniformly on the compact subset $\Phi(K)_{\epsilon_f}$ of $\mathbb{R}^d$.

To this end, we make the following observation on the bi-Lipschitz regularity of Φ, when restricted to $K \subseteq \mathbb{R}^m$. Since K is non-empty and compact, then $\Phi|_K : K \to \mathbb{R}^d$ is Lipschitz, as it is at-least once continuously differentiable. Since K is compact, and Φ is continuous, then by (Munkres, 2000, Theorem 26.5) $\Phi(K)$ is also compact. Therefore, since Φ^{-1} is also at-least once continuously differentiable then, $\Phi^{-1}|_{\Phi(K)_{\epsilon_f}} : \Phi(K)_{\epsilon_f} \to \mathbb{R}^m$ is Lipschitz. Hence, $\Phi|_K : K \to \Phi(K) \subseteq \mathbb{R}^d$ is bi-Lipschitz[3]. In particular, $\Phi(K)_{\epsilon_f}$ and $\Phi(K)$ have diameter at-most:

$$\operatorname{diam}(\Phi(K)) \leq \operatorname{Lip}(\Phi)\operatorname{diam}(K) \text{ and } \operatorname{diam}(\Phi(K)_{\epsilon_f}) \leq \operatorname{Lip}(\Phi)\operatorname{diam}(K) + 2\epsilon_f. \tag{33}$$

We may therefore apply (Heinonen, 2001, Theorem 12.1), as $p_d^m(\Phi(K))$ has a (finite) doubling constant $\lambda(p_d^m(\Phi(K)))$ since it is a subset of $\mathbb{R}^d$. More precisely, we have that:

$$\lambda(p_d^m(\Phi(K))) = \lambda(\mathbb{R}^d) = 2^d, \tag{34}$$

where the first inequality in (34) follows from (Robinson, 2011, Lemma 9.6 (i)) and the second in (34) from (Robinson, 2011, Lemma 9.2).

Therefore, we may apply (Bruè et al., 2021, Theorem 3.2) to conclude that there exists a Lipschitz map $\Pi : \mathbb{R}^d \to \mathcal{P}_1(p_d^m(\Phi(K)))$ such that, for all $y \in p_d^m(\Phi(K))$, Π satisfies:

$$\Pi_y = \delta_y. \tag{35}$$

Moreover, the same result bounds Π's Lipschitz constant, denoted by $\operatorname{Lip}(\Pi)$, by $k\log(\lambda(p_d^m(\Phi(K))))$ where, k is an absolute constant; i.e. it does not depend on $\mathbb{R}^n$, $\mathbb{R}^d$, ϵ, or

[2]The author of this first paper on the subject calls such maps Lipschitz stochastic retracts. The terminology "random projection" was later adopted by other authors, such as Ambrosio & Puglisi (2020) and Bruè et al. (2021) in connection with the work of Lee & Naor (2005) and the Johnson & Lindenstrauss (1984)'s Lemma.

[3]A map $f : \mathbb{R}^n \to \mathbb{R}^m$ is bi-Lipschitz (see (Heinonen, 2001, page 78)) if there are constants $c, C > 0$ such that, for every $x_1, x_2 \in \mathbb{R}^n$ the estimate holds: $c\|x_1 - x_2\| \leq \|f(x_1) - f(x_2)\| \leq C\|x_1 - x_2\|$.

on $\lambda(p_d^m(\Phi(K)))$. Consequently, k does not depend on n, d, ϵ, or on κ_K. Combining this with (34), Π's Lipschitz constant is bounded as follows:

$$\mathrm{Lip}(\Pi) \leq k \log_2(\lambda(p_d^m(\Phi(K)))) \leq kd. \tag{36}$$

Since Π is 1-Lipschitz (continuous), $\Phi(K)_{\epsilon_f}$ is compact and $p_d^m(\Phi(K))$ is compact then by (Kratsios, 2021, Theorem 3), there exists a $\hat{D} \in \mathcal{NN}_{d,N}^{\sigma}$ and $y_{1,1}, \ldots, y_{N,Q} \in p_d^m(\Phi(K))$ such that the "probabilistic decoder network" $\hat{\mathcal{D}}_0$ defined by:

$$\hat{\mathcal{D}}_0 : \mathbb{R}^d \ni x \mapsto \sum_{k=1}^{N} \text{P-attention}\left(\hat{D}(x), Y\right) \in \mathcal{P}_1(p_d^m(\Phi(K)));$$

where, Y is the $N \times Q \times m$-array with $Y_{n,q} = y_{n,q}$, and $\hat{\mathcal{D}}_0$ satisfies:

$$\sup_{y \in p_d^m \circ \Phi(K_{\epsilon_f})} \mathscr{W}_1\left(\Pi(y), \hat{\mathcal{D}}_0(y)\right) \leq \epsilon_K,; \tag{37}$$

where we have set the activation function parameter α to $\alpha = 1$. Thus, σ_1 is smooth and non-polynomial; in which case (Kratsios, 2021, Theorem 3 and Example 7) and the estimates in (33) and in (36) also implies $\hat{\mathcal{D}}_0$ satisfies the following "complexity estimates":

(i-$\mathscr{D}$) $Q \leq 8(\epsilon_K^{-1} \mathrm{Lip}(\Phi) \operatorname{diam}(K) d^{\frac{5}{2}})^d$,

(ii-$\mathscr{D}$) $N \leq \left(\frac{kd^2 2^{\frac{9}{2}} \mathrm{Lip}(\Phi)(\operatorname{diam}(K)+\epsilon_f)}{\sqrt{d+1}\epsilon_K}\right)^d$

(iii-$\mathscr{D}$) $\hat{\mathcal{D}}_0$ has depth at most $\mathscr{O}\left((dN^{\frac{3}{2}}(\mathrm{Lip}(\Phi)\operatorname{diam}(K) + 2\epsilon_f)(1-4^{-1}\epsilon_K^{-1})(1-\epsilon_K^{-1})(1+4^{-1}d))^{2d}\right)$.

Here, we denote the Lipschitz constant of $\Phi|_K$ by $\mathrm{Lip}(\Phi)$. We have also used Jung's Theorem (Jung, 1910) and the fact that $\frac{d^2}{2(d-1)(d+1)} < d^{\frac{5}{2}}$ allows us to simplify the estimate in (Kratsios, 2021, Theorem 3 (ii)) to simplify the expression in (i-$\mathcal{D}$) and in (iii-$\mathscr{D}$).

Since $\Phi^{-1} \circ i_d^m : p_d^m(\Phi(K)) \to K$ is Lipschitz and since ι_d^m is 1-Lipschitz then, $(\Phi^{-1} \circ i_d^m)_\# : \mathcal{P}_1(p_d^m(\Phi(K))) \to \mathcal{P}_1(K)$ is also Lipschitz with Lipschitz-constant at most $\mathrm{Lip}(\Phi^{-1})$. Let $\hat{\mathcal{D}}(\cdot) \triangleq (\Phi^{-1} \circ i_d^m)_\# \hat{\mathcal{D}}_0(\cdot)$. Thus, (37) implies:

$$\begin{aligned}\sup_{y \in \Phi(K)_{\epsilon_f}} \mathscr{W}_1\left((\Phi^{-1} \circ i_d^m)_\#(\Pi(y)), \hat{\mathcal{D}}(y)\right) &\leq \mathrm{Lip}(\Phi^{-1}) \sup_{y \in \Phi(K)_{\epsilon_f}} \mathscr{W}_1\left(\Pi(y), \hat{\mathcal{D}}_0(y)\right) \\ &\leq \mathrm{Lip}(\Phi^{-1})\epsilon_K,\end{aligned} \tag{38}$$

Moreover, the injectivity of $\Phi^{-1} \circ i_d^m$ implies that $\hat{\mathcal{D}}$ has the following simple expression:

$$\hat{\mathcal{D}} : \mathbb{R}^d \ni x \mapsto \sum_{k=1}^{N} \text{P-attention}\left(\hat{D}(x), \tilde{Y}\right) \in \mathcal{P}_1(K),$$

where $\tilde{Y}$ is the $N \times Q \times m$-array with $Y_{n,q} = \Phi^{-1} \circ \iota_d^m(y_{n,q})$.

Therefore, by (30), we have the following preliminary estimate:

$$\sup_{x \in \mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{D}} \circ \hat{\mathcal{E}}(x), \operatorname*{argmin}_{y \in K} L(x,y)\right) \leq \sup_{x \in \mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{D}} \circ \hat{\mathcal{E}}(x), \delta_{f(x)}\right) \tag{39}$$

$$\leq \sup_{x \in \mathcal{X}} \Big[\mathscr{W}_1\left(\hat{\mathcal{D}} \circ \hat{\mathcal{E}}(x), (\Phi^{-1} \circ \iota_d^m)_\# \circ \Pi \circ \hat{\mathcal{E}}(x)\right) \tag{40}$$

$$+ \mathscr{W}_1\left((\Phi^{-1} \circ \iota_d^m)_\# \circ \Pi \circ \hat{\mathcal{E}}(x), (\Phi^{-1} \circ \iota_d^m)_\# \circ \Pi \circ f(x)\right) \tag{41}$$

$$+ \mathscr{W}_1\left((\Phi^{-1} \circ \iota_d^m)_\# \circ \Pi \circ f(x), (\Phi^{-1} \circ \iota_d^m)_\# \circ \delta_{f(x)}\right)\Big]. \tag{42}$$

To conclude the proof, we must first bound term (40). Since we found that $\hat{\mathscr{E}}(\mathcal{X}) \cup f(\mathcal{X}) \subseteq \mathcal{X}_\epsilon$ then, utilizing (38) we compute:

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), (\Phi^{-1}\circ\iota_d^m)_\#\circ\Pi\circ\hat{\mathcal{E}}(x)\right) &= \sup_{x\in\mathcal{X}} \mathscr{W}_1\left((\Phi^{-1}\circ\iota_d^m)_\#\circ\hat{\mathcal{D}}_0\circ\hat{\mathcal{E}}(x), (\Phi^{-1}\circ\iota_d^m)_\#\circ\Pi\circ\hat{\mathcal{E}}(x)\right)\\
&\leq \operatorname{Lip}(\Phi^{-1})1 \sup_{y\in\mathcal{X}_\epsilon} \mathscr{W}_1\left(\hat{\mathcal{D}}_0(\hat{\mathcal{E}}(x)), \Pi(\hat{\mathcal{E}}(x))\right)\\
&\leq \operatorname{Lip}(\Phi^{-1}) \sup_{y\in\mathcal{X}_\epsilon} \mathscr{W}_1\left(\hat{\mathcal{D}}_0(y), \Pi(y)\right)\\
&\overset{(38)}{\leq} \operatorname{Lip}(\Phi^{-1})\epsilon_K;
\end{aligned} \tag{43}$$

where $\operatorname{Lip}(\Phi^{-1})$ denotes the Lipschitz constant of Φ on $\Phi(K)_{\epsilon_f}$. For the second term, i.e. (41), we combine the fact that Π is Lipschitz with $\operatorname{Lip}(\Pi)$ given in (36) and our estimate on f obtained in (32) to find that:

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \mathscr{W}_1\left((\Phi^{-1}\circ i_d^m)_\#\circ\Pi\circ\hat{\mathcal{E}}(x), (\Phi^{-1}\circ i_d^m)_\#\circ\Pi\circ f(x)\right)\\
&\quad\leq \operatorname{Lip}(\Phi^{-1}) \sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\Pi\circ\hat{\mathcal{E}}(x), \Pi\circ f(x)\right)\\
&\quad\leq \operatorname{Lip}(\Phi^{-1}) \sup_{x\in\mathcal{X}} \operatorname{Lip}(\Pi)\mathscr{W}_1\left(\hat{\mathcal{E}}(x), f(x)\right)\\
&\quad\overset{(36)}{\leq} k\operatorname{Lip}(\Phi^{-1})d \sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{E}}(x), f(x)\right)\\
&\quad\overset{(29)+(31)}{\leq} k\operatorname{Lip}(\Phi^{-1})d\epsilon_f
\end{aligned} \tag{44}$$

The third term, i.e. (42), we use the random projection property of Π on K defined in (35) and the assumption that $f(\mathcal{X}) \subseteq K$. This is done as follows:

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \mathcal{W}_1\left((\Phi^{-1}\circ\iota_d^m)_\#\circ\Pi\circ f(x), (\Phi^{-1}\circ\iota_d^m)_\#\delta_{f(x)}\right) &\leq \operatorname{Lip}(\Phi^{-1}\circ\iota_d^m)\,\mathcal{W}_1\left(\Pi\circ f(x), f(x)\right)\\
&= \sup_{x\in\mathcal{X}} \operatorname{Lip}(\Phi^{-1})\operatorname{Lip}(\iota_d^m)\,\mathcal{W}_1\left(\Pi\circ f(x), f(x)\right)\\
&= \sup_{x\in\mathcal{X}} \operatorname{Lip}(\Phi^{-1})\,\mathcal{W}_1\left(\Pi\circ f(x), f(x)\right)\\
&\overset{\because f(x)\in K\ +(35)}{\leq} \sup_{y\in K} \operatorname{Lip}(\Phi^{-1})\,\mathcal{W}_1\left(\Pi(y), y\right)\\
&= 0
\end{aligned} \tag{45}$$

Therefore, incorporating (43), (44), and (45), we may control the right-hand of (39) with the following upper-bound:

$$\sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \operatorname*{argmin}_{y\in K} L(x,y)\right) \leq \epsilon_K + k\operatorname{Lip}(\Phi^{-1})d\epsilon_f + 0. \tag{46}$$

Relabelling $k\operatorname{Lip}(\Phi^{-1})$ as k and the $\Phi^{-1}\circ i_d^m(y_{k,q})$ as $y_{k,q}$, and incorporating the rate $d \in \mathscr{O}(m^{\frac{1}{s}})$ (implied by Assumption 2.5) into the complexity estimates (i)-(iii) and (i-$\mathcal{D}$)-(iii-$\mathcal{D}$) yields the rates of Table 2 and, the explicit rates of Table 2. Thus the proof is complete. □

D Proof of Corollaries

This appendix contains proofs of the paper's main corollaries.

D.1 PROOF OF COROLLARY 2.3

Proof of Corollary 2.3. Let $(\Omega, \mathcal{F}, \nu)$ be a probability space on which the random-field $(Y^x)_{x\in\mathbb{R}^n}$, satisfying $Y^x \sim F(x)$, is defined. Consider the non-empty-valued correspondence[4] $\mathscr{O}(x)$ is defined as in (24). We continue where Theorem 2.2's proof left off. For any $x \in \mathcal{X}_\epsilon$, we now compute the concrete lower-bound on $\inf_{y^\star\in\mathcal{O}(x)} \mathscr{W}_1(\hat{F}(x), \delta_{y^\star})$.

$$\begin{aligned}
\inf_{y^\star\in\mathscr{O}(x)} \mathscr{W}_1(\hat{F}(x), \delta_{y^\star}) &= \inf_{y\in C_x\cap K} \inf_{\pi\in \mathrm{Cpl}(\hat{F}(x),\delta_y)} \int_{(u,v)\in\mathbb{R}^n\times\mathbb{R}^n} \|u-v\|\pi(d(u,v)) \\
&= \inf_{y^\star\in\mathscr{O}(x)} \int_{(u,v)\in K\times K} \|u-v\| \left(\hat{F}(x)\otimes\delta_y(d(u,v))\right) \\
&= \inf_{y^\star\in\mathscr{O}(x)} \int_{u\in K}\int_{v\in K} \|x-v\|\hat{F}(x)(du)\delta_y(dv) \\
&= \inf_{y^\star\in\mathscr{O}(x)} \int_{u\in K} \|x-y^\star\|\hat{F}(x)(du) \\
&\overset{\text{(def)}}{=} \inf_{y^\star\in\mathscr{O}(x)} \mathbb{E}_{Y^x\sim\hat{F}(x)}\left[\|Y^x-y^\star\|\right] \\
&\geq \mathbb{E}_{Y^x\sim\hat{F}(x)}\left[\operatorname*{ess\text{-}inf}_{y^\star\in\mathscr{O}(x)} \|Y^x-y^\star\|\right] \qquad (47) \\
&\overset{\text{(def)}}{=} \mathbb{E}_{Y^x\sim\hat{F}(x)}\left[\left\|Y^x - \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)\right\|\right]. \qquad (48)
\end{aligned}$$

In more detail: The first equality is just the definition of the Wasserstein distance. The second equality follows from the fact that for any $y^\star \in \mathbb{R}^n$ (and in particular any such $y^\star$ in $C_x \cap K$) the product measure $\hat{F}(x) \otimes \delta_{y^\star}$ is the only coupling of $\hat{F}(x)$ with $\delta_{y^\star}$ (see Lemma C.4) and the facts that, by (i), $\hat{F}(x)$ is supported in K and, by definition, $\delta_{y^\star}$ is also supported in K. The third equality follows the Fubini-Tonelli Theorem (see (Kallenberg, 2021, Theorem 1.27)) since all involved quantities are integrable over the compact set $K \times K$. The inequality (47) follows from Fatou's Lemma (see (Kallenberg, 2021, Lemma 1.20)) and the fact that the *essential-infimum* lower-bounds the infimum. The final equality is just the definition of the distance from $Y^x(\omega)$ to the optimality set $\mathscr{O}(x)$ (for each $\omega \in \Omega$). Combining the upper-bound on the right-hand side of (28) with the lower-bound in (48) yields the result. □

D.2 PROOF OF COROLLARY 2.8

Proof of Corollary 2.8. We continue with the notation of Theorem 2.7, and specifically with the following estimate derived in (46):

$$\sup_{x\in\mathcal{X}} \mathscr{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \delta_{f(x)}\right) \leq \epsilon_K + k\,\mathrm{Lip}(\Phi^{-1})d\epsilon_f. \qquad (49)$$

Combining (49), the monotonicity of integration, Assumption 2.4, Jensen's inequality (applicable due to the concavity of l), and from Lemma C.4, we deduce the following estimate: for each $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{E}_{Y^x\sim\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\left[L(x,Y^x)\right] &\leq \mathbb{E}_{Y^x\sim\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\left[l(\|f(x)-Y^x\|)\right] \\
&\leq l\left(\mathbb{E}_{Y^x\sim\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\left[\|f(x)-Y^x\|\right]\right) \\
&\overset{\text{(C.4)}}{=} l\left(\mathscr{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \delta_{f(x)}\right)\right) \\
&\overset{\text{(49)}}{\leq} l(\epsilon_K + k\,\mathrm{Lip}(\Phi^{-1})d\epsilon_f). \qquad (50)
\end{aligned}$$

□

[4] Also called multifunction, multivalued function, or set-valued function.

D.3 PROOF OF COROLLARY 2.9

Proof of Corollary 2.9. Let $L : \mathbb{R}^n \times \mathbb{R}^n \ni (x,y) \mapsto \|f(x)-y\| \in [0,\infty)$. Then, for each fixed $x \in \mathbb{R}^n$, the strict convexity of $y \mapsto L(x,y)$ and the assumption that $f(x) \in K$ imply that $\{f(x)\} = \operatorname{argmin}_{y\in K} L(x,y)$. Thus, for each $x \in \mathcal{X}$ and each $\mathbb{P} \in \mathcal{P}_1(K)$, we have that:

$$\mathcal{W}_1\left(\mathbb{P}, \delta_{f(x)}\right) = \mathcal{W}_1\left(\mathbb{P}, \operatorname*{argmin}_{y\in K} L(x,y)\right). \tag{51}$$

Since $f \in C^k_{tr}(\mathcal{X}, K)$ and K is non-empty satisfying Assumption 2.5, and compact Theorem 2.7 implies that for each $\epsilon_K, \epsilon_f > 0$ there exist a $\hat{\mathcal{D}}$ and a $\hat{\mathcal{E}}$ as in Table 2 satisfying the estimate:

$$\mathcal{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \operatorname*{argmin}_{y\in K} L(x,y)\right) \leq \epsilon_K + kd\epsilon_f. \tag{52}$$

Define the map $\beta : \mathcal{P}_1(K) \ni \mathbb{P} \mapsto \mathbb{E}_{Y\sim\mathbb{P}}[Y] \in \mathbb{R}^m$. By Lemma C.1 (i), β takes values in K and according to Lemma C.1 (ii) it is 1-Lipschitz. Notice also that $\beta(\delta_y) = y$ for each $y \in K$ and, in particular, $\beta(\delta_{f(x)} = f(x))$. These observations together with (51) and (52) imply that:

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \|f(x) - \mathbb{E}_{Y^x\sim\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}[Y^x]\| &= \sup_{x\in\mathcal{X}} \|\beta(\delta_{f(x)}) - \beta(\hat{\mathcal{D}}\circ\hat{\mathcal{E}})(x)\| \\
&\leq \sup_{x\in\mathcal{X}} 1\cdot \mathscr{W}_1\left(\delta_{f(x)}, \hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)\right) \\
&\overset{(52)}{\leq} \sup_{x\in\mathcal{X}} 1\cdot \mathscr{W}_1\left(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \operatorname*{argmin}_{y\in K} L(x,y)\right) \\
&\overset{\text{Thm.2.7}}{\leq} \epsilon_K + kd\epsilon_f.
\end{aligned}$$

Whence, (i) and (ii) hold. □

D.4 PROOF OF COROLLARY 2.11

Proof of Corollary 2.11. Let $L : \mathbb{R}^n \times \mathbb{R}^n \ni (x,y) \mapsto \|f(x)-x\| \in [0,\infty)$, note that L satisfies Assumption 2.4, and that for each $x \in \mathcal{X}$ we have that $\operatorname{argmin}_{y\in K} L(x,y) = \{f(x)\}$. By Theorem 2.7, for every $f \in C^k_{tr}(\mathcal{X}, K)$ and for every $\epsilon > 0$, there exist a $\hat{\mathcal{D}}$ and a $\hat{\mathcal{E}}$ as in Table 2 satisfying $\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x) \in \mathcal{P}_1(K)$ for each $x \in \mathbb{R}^n$ and satisfying the uniform estimate:

$$\max_{x\in\mathcal{X}} \mathscr{W}_1(\delta_{f(x)}, \hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)) \leq \epsilon_K + k\operatorname{Lip}(\Phi^{-1})d\epsilon_f. \tag{53}$$

Since K satisfies Assumption 2.10 then, Lemma C.3 applies. Therefore, (14) Theorem 2.7 imply:

$$\begin{aligned}
\max_{x\in\mathcal{X}} d_g\left(f(x), \overline{\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\right) &= \max_{x\in\mathcal{X}} d_g\left(\overline{\delta_{f(x)}}, \overline{\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\right) \\
&\overset{(14)}{\leq} \max_{x\in\mathcal{X}} 1\,\mathscr{W}_1(\delta_{f(x)}, \hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)) \\
&\overset{\text{Thm.2.7}}{\leq} \epsilon_K + k\operatorname{Lip}(\Phi^{-1})d\epsilon_f
\end{aligned}$$

Furthermore, (15) and the fact that $\mathcal{P}_1(K) \ni \mathbb{P} \mapsto \overline{\mathbb{P}} \in K$ is a left-inverse of the map $K \ni y \mapsto \delta_y$ imply that: for every $x \in \mathbb{R}^n$ it follows that:

$$\overline{\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)} \in K.$$

This concludes the proof. □

D.4.1 DISCUSSION: THEOREM 2.7 VS. COROLLARY 2.8

The modulus of continuity ω in Assumption 2.4 does not enter into the estimate in Theorem 2.7 (ii) but it does appear in the estimate of Corollary 2.8. This is because[5]:

$$\mathscr{W}_1(\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x), \operatorname{argmin}_{y\in K} L(x,y)) = \inf_{y\in\operatorname{argmin}_{y\in K} L(x,y)} \mathbb{E}_{Y^x\sim\hat{\mathcal{D}}\circ\hat{\mathcal{E}}(x)}\left[\|Y^x - y\|\right]. \tag{54}$$

[5]The right-most expression in (54) is justified in Lemma C.4; see Corollary 2.8's proof.

Thus, the right-hand side is controlled by the of (54) the average (in Y^x) worst-case (in x) Euclidean average distance between Y^x and the optimality set $\operatorname{argmin}_{y\in K} L(x,y)$; whereas, the estimate in Corollary 2.8 is controlling the average (in Y^x) worst-cast (in x) loss $L(x,Y^x)$. In other words, Corollary 2.8 controls the optimal value of L on K and Theorem 2.7 approximates the optimal prediction.

E Further Corollaries to the Deep Maximum Theorem

This brief appendix contains additional corollaries of Theorem 2.2 which were not included in our manuscript's main body. The intent here is to show how our "Deep Maximum Theorem" simplifies in the convex case, a similar result can be derived for the geodesically convex case.

Corollary E.1 (Deep Maximum Theorem: Convex Case)**.** *Assume the context of Theorem 2.2. Let $\{Y^x\}_{x\in\mathbb{R}^n}$ be an $\mathbb{R}^m$-valued random field with $X^x \sim \hat{F}$. If each $C_x \cap K$ is a convex set and L is strictly convex then, $\mathbb{R}^n \ni x \mapsto \mathbb{E}[Y^x] \in \mathbb{R}^m$ has the following representation:*

$$\mathbb{E}[Y^x] = \operatorname{Attention}(\hat{f}(x), Y), \tag{55}$$

where $Y = (\sum_{q=1}^{Q} \frac{1}{Q} y_{k,q})_{k=1}^{N}$ is an $N \times m$-matrix. Moreover, the following hold:

(i) ***Constraint Satisfaction:*** *$\mathbb{E}[Y^x] \in K$ for each $x \in \mathbb{R}^n$,*

(ii) ***Probable Optimality:*** *$\max_{x\in\mathcal{X}_\epsilon} \|\mathbb{E}[Y^x] - y^\star(x)\| \leq \epsilon$,*

where $y^\star(x)$ is the well-defined and unique minimizer of $L(x,\cdot)$ on $C_x \cap K$.

Proof of Corollary E.1. First we note that since each $C_x \cap K$ is a non-empty, compact, and convex subset of $\mathbb{R}^n$ and since L is strictly convex and bounded-below on $K \cap C_x$ (since it is continuous and $K \cap C_x$ is compact) then it must have a unique minimizer (see Planiden & Wang (2016)). Thus, $y^\star(x)$ exists and is uniquely defined for each $x \in \mathbb{R}^n$.

Consider the setting of Theorem 2.2 and suppose further that K is convex. Then, we may apply Lemma C.1. Thus, in the notation of Theorem 2.2, for each $x \in \mathcal{X}_\epsilon$ and every $y^\star \in \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)$ we have the estimate:

$$\left\|\mathbb{E}_{Y\sim\hat{F}(x)}[Y] - \mathbb{E}_{\tilde{Y}\sim\delta_{y^\star}}[\tilde{Y}]\right\| \leq \mathscr{W}_1\left(\hat{F}(x), y^\star\right). \tag{56}$$

Applying the estimate: $\max_{x\in\mathcal{X}_\epsilon} \inf_{y^\star\in \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)} \mathscr{W}_1(\hat{F}(x), \delta_{y^\star}) \leq \epsilon$ obtained in Theorem 2.2 to the right-hand side of (56), and noting that $\mathbb{E}_{Y\sim\delta_{y^\star}}[Y] = y^\star$ yields:

$$\begin{aligned}
\max_{x\in\mathcal{X}_\epsilon} \inf_{y^\star\in \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)} \left\|\mathbb{E}_{Y\sim\hat{F}(x)}[Y] - y^\star\right\| &= \max_{x\in\mathcal{X}_\epsilon} \inf_{y^\star\in \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)} \left\|\mathbb{E}_{Y\sim\hat{F}(x)}[Y] - \mathbb{E}_{\tilde{Y}\sim\delta_{y^\star}}[\tilde{Y}]\right\| \\
&\leq \max_{x\in\mathcal{X}_\epsilon} \inf_{y^\star\in \operatorname*{argmin}_{y\in C_x\cap K} L(x,y)} \mathscr{W}_1\left(\hat{F}(x), y^\star\right) \\
&\leq \epsilon.
\end{aligned}$$

This gives the second part of the statement.

Since $\operatorname{supp}(\hat{F}(x)) \subseteq K$ then, any $\mathbb{R}^n$-valued random-vector distributed according to $\hat{F}(x)$, $\hat{F}(x)$-a.s. takes values in K. Thus, $\hat{F}(x)\,(X \in K) = \mathbb{E}_{X\sim\hat{F}(x)}[I_K(X)] = 1$. This gives the first claim. □

For completeness, we include the deterministic analogue of Corollary E.1 when K is a geodesically convex subset of a complete connected Riemannian submanifold (M, g) of $\mathbb{R}^m$ satisfying Assumption 2.10. The result is a qualitative generalization of Corollary 2.11.

Corollary E.2 (Deep Maximum Theorem: Riemannian Case)**.** *Assume the context of Theorem 2.2 and suppose that Assumption 2.10 holds. Suppose also that for each $x \in [0,1]^n$ there exists a unique $y(x) \in C_x \cap K$ minimizing L; i.e.:*

$$L(x, y(x)) = \inf_{y \in C_x \cap K} L(x, y),$$

moreover, assume that $x \mapsto y(x)$ is continuous on $[0,1]^n$. Then, the function:

$$[0,1]^n \ni x \mapsto \overline{\text{P-attention}(\hat{f}(x), Y)}, \tag{57}$$

is well-defined; moreover, the following hold:

(i) ***Constraint Satisfaction:*** $\overline{\text{P-attention}(\hat{f}(x), Y)} \in K$ *for each* $x \in \mathbb{R}^n$,

(ii) ***Probable Optimality:*** $\max_{x \in \mathcal{X}_\epsilon} d_g \left(\overline{\text{P-attention}(\hat{f}(x), Y)}, y^\star(x) \right) \leq \epsilon,$

where $y^\star(x)$ is the well-defined and unique minimizer of $L(x, \cdot)$ on $C_x \cap K$.

The proof of Corollary E.2 is nearly identical to that of Corollary 8.

Proof of Corollary E.2. Consider the setting of Theorem 2.2 and suppose further that K satisfies Assumption 2.10. Then, we may apply Lemma C.3. Thus, in the notation of Theorem 2.2, for each $x \in \mathcal{X}_\epsilon$ and every $y^\star \in \operatorname{argmin}_{y \in C_x \cap K} L(x, y)$ we have the estimate:

$$d_g \left(\overline{\hat{F}(x)}, \overline{\delta_{y^\star}} \right) \leq \mathscr{W}_1 \left(\hat{F}(x), y^\star \right). \tag{58}$$

Applying the estimate: $\max_{x \in \mathcal{X}_\epsilon} \inf_{y^\star \in \operatorname{argmin}_{y \in C_x \cap K} L(x,y)} \mathscr{W}_1(\hat{F}(x), \delta_{y^\star}) \leq \epsilon$ obtained in Theorem 2.2 to the right-hand side of (58), and noting that $\overline{\delta_{y^\star}} = y^\star$ yields:

$$\begin{aligned}
\max_{x \in \mathcal{X}_\epsilon} \inf_{y^\star \in \operatorname{argmin}_{y \in C_x \cap K} L(x,y)} d_g \left(\overline{\hat{F}(x)}, y^\star \right) =& \max_{x \in \mathcal{X}_\epsilon} \inf_{y^\star \in \operatorname{argmin}_{y \in C_x \cap K} L(x,y)} d_g \left(\mathbb{E}_{Y \sim \hat{F}(x)}[Y], \mathbb{E}_{\tilde{Y} \sim \delta_{y^\star}}[\tilde{Y}] \right) \\
\leq & \max_{x \in \mathcal{X}_\epsilon} \inf_{y^\star \in \operatorname{argmin}_{y \in C_x \cap K} L(x,y)} \mathscr{W}_1 \left(\hat{F}(x), y^\star \right) \\
\leq & \epsilon.
\end{aligned}$$

This gives (ii). Lastly, (i) follows from (15) in Lemma C.3. □

REFERENCES

Bijan Afsari. Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis: A hitchhiker's guide*. Springer, Berlin, third edition, 2006.

Luigi Ambrosio and Daniele Puglisi. Linear extension operators between spaces of Lipschitz maps and optimal transport. *Journal für die Reine und Angewandte Mathematik*, 764:1–21, 2020.

Anonymized. Pytorch implementation of attend-to-constraints, 2021. URL https://drive.google.com/file/d/1vryYsUmHt0fok3Mrje6oN9Tjs2UmpgkA/view.

Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis*. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, 2009.

Michel Baes, Calypso Herrera, Ariel Neufeld, and Pierre Ruyssen. Low-rank plus sparse decomposition of covariance matrices using neural network parametrization. *IEEE Transaction on Neural Networks and Learning Systems*, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Yair Bartal. On approximating arbitrary metrices by tree metrics. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, pp. 161–168. ACM, New York, 1999.

Basel Committee on Banking Supervision. Fundamental review of the trading book: outstanding issues, February 2015. https://www.bis.org/bcbs/publ/d305.pdf.

Basel Committee on Banking Supervision. Minimum capital requirements for market risk, February 2019. https://www.bis.org/bcbs/publ/d457.pdf.

Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. ISBN 978-1-4419-9466-0. doi: 10.1007/978-1-4419-9467-7. URL https://doi.org/10.1007/978-1-4419-9467-7. With a foreword by Hédy Attouch.

Claude Berge. *Éspaces Topologiques (Topological Spaces)*. Dunod, 1963.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

Silvère Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3): 1055–1070, 2009.

Silvère Bonnabel, Anne Collard, and Rodolphe Sepulchre. Rank-preserving geometric means of positive semi-definite matrices. *Linear Algebra and its Applications*, 438(8):3202–3216, 2013.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.08708*, 2021. URL http://arxiv.org/abs/2104.13478.

Bernard Bru, Henri Heinich, and Jean-Claude Lootgieter. Distances de Lévy et extensions des théoremès de la limite centrale et de Glivenko-Cantelli. *Publ. Inst. Statist. Univ. Paris*, 37(3-4): 29–42, 1993.

Alexander Brudnyi and Yuri Brudnyi. *Methods of geometric analysis in extension and trace problems. Volume 1*, volume 102 of *Monographs in Mathematics*. Birkhäuser/Springer Basel AG, Basel, 2012a.

Alexander Brudnyi and Yuri Brudnyi. *Methods of geometric analysis in extension and trace problems. Volume 2*, volume 103 of *Monographs in Mathematics*. Birkhäuser/Springer Basel AG, Basel, 2012b.

Elia Bruè, Simone Di Marino, and Federico Stra. Linear Lipschitz and C^1 extension operators through random projection. *Journal of Functional Analysis*, 280(4):108868, 2021.

Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. *Integer Programming*, volume 271 of *Graduate Texts in Mathematics*. Springer, Cham, 2014.

Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann. Discrete-time signatures and randomness in reservoir computing. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2292–2300, 2013.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

Meng Ding and Guoliang Fan. Multilayer joint gait-pose manifolds for human gait motion modeling. *IEEE Transactions on Cybernetics*, 45(11):2413–2424, 2014.

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015. doi: 10.1109/MSP.2015.2398954.

Richard M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Charles L. Fefferman. A sharp form of Whitney's extension theorem. *Annals of Mathematics*, 161 (1):509–577, 2005.

Thomas Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, 105(2):171–185, 2013.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10:215–310, 1948.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Risk bounds for reservoir computing. *Journal of Machine Learning Research*, 21(240):1–61, 2020a. URL http://jmlr.org/papers/v21/19-902.html.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *arXiv preprint arXiv:2002.05933*, 2020b.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *J. Mach. Learn. Res.*, 19:Paper No. 24, 40, 2018.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Differentiable reservoir computing. *J. Mach. Learn. Res.*, 20:Paper No. 179, 62, 2019.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Juha Heinonen. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York, 2001.

Ludger Holters, Björn Bahl, Maike Hennen, and André Bardow. Playing Stackelberg games for minimal cost for production and utilities. In *ECOS 2018-Proceedings of the 31st International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems*, pp. 36–36. University of Minho, 2018.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Network*, 2(5):359–366, July 1989.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemp. Math.*, pp. 189–206. American Mathematical Society, RI, 1984.

Jürgen Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, Heidelberg, seventh edition, 2017.

Heinrich W. E. Jung. Über die Cremonasche Transformation der Ebene. *J. Reine Angew. Math.*, 138:255–318, 1910. ISSN 0075-4102. doi: 10.1515/crll.1910.138.255.

Olav Kallenberg. *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition, 2021.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Machine Learning Research*, volume 125, pp. 2306–2327. PMLR, 09–12 Jul 2020.

Achim Klenke. *Probability theory: A comprehensive course*. Universitext. Springer, London, second edition, 2014.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Anastasis Kratsios. Universal regular conditional distributions. *arXiv preprint:2105.07743*, 2021. URL https://arxiv.org/abs/2105.07743.

Anastasis Kratsios and Eugene Bilokopytov. Non-Euclidean universal approximation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Anastasis Kratsios and Leonie Papon. Universal approximation theorems for differentiable geometric deep learning, 2021. URL https://arxiv.org/abs/2101.05390.

James R. Lee and Assaf Naor. Extending Lipschitz functions via random metric partitions. *Inventiones Mathematicae*, 160(1):59–95, 2005.

Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *arXiv:2104.08708*, 2021. URL https://arxiv.org/abs/2104.08708.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the Fréchet mean. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28(2):1190–1248, 2018. ISSN 1050-5164. doi: 10.1214/17-AAP1328. URL https://doi.org/10.1214/17-AAP1328.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009. ISSN 1574-0137.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2015.

Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.

Theodore Samuel Motzkin. *Sur quelques propriétés caractéristiques des ensembles bornés non convexes*. Bardi, 1935.

James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, 2000. Second edition.

Shin-ichi Ohta. Extending Lipschitz and Hölder maps between metric spaces. *Positivity*, 13(2): 407–425, 2009.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. *International Conference on Learning Representations (ICLR)*, 2021.

Ofir Pele and Michael Werman. Fast and robust Earth Mover's distances. In *Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV)*, pp. 460–467, 2009.

Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148 (4):1567–1581, 2020.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica, 1999*, 8:143–195, 1999.

Chayne Planiden and Xianfu Wang. Most convex functions have unique minimizers. *Journal of Convex Analysis*, 23(3):877–892, 2016.

Pakize Simin Pulat. On the relation of max-flow to min-cut for generalized networks. *European Journal of Operational Research*, 39(1):103–107, 1989.

Michael Puthawala, Konik Kothari, Matti Lassas, Ivan Dokmanić, and Maarten de Hoop. Globally injective ReLU networks. *arXiv:2006.08464*, 2020. URL https://arxiv.org/abs/2105.07743.

Prajit Ramachandran, Barret Zoph, and Quoc Le. Searching for activation functions. In *Proceedings of the International Conference of Learning Representations (ICLR)*, 2018.

Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *arXiv:2102.11436*, 2021. URL https://arxiv.org/abs/2102.11436.

James C. Robinson. *Dimensions, embeddings, and attractors*, volume 186 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2011.

Denis Rosset, Felipe Montealegre-Mora, and Jean-Daniel Bancal. RepLAB: A computational/numerical approach to representation theory. In *Quantum Theory and Symmetries*, pp. 643–653. Springer, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021a.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33(4):1005–1036, 03 2021b.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces*, volume 338 of *Contemp. Math.*, pp. 357–390. Amer. Math. Soc., Providence, RI, 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. On the regularity of attention. *arXiv:2102.05628*, 2021. URL https://arxiv.org/abs/2102.05628.

Steven Weinberg. Implications of dynamical symmetry breaking. *Physical Review D*, 13(4):974, 1976.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.

Dmitry Yarotsky. Elementary superexpressive activations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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

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

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, 2016.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.