A WELL-POSEDNESS, REGULARITY, AND UNIVERSAL APPROXIMATION

Here we consolidate the proofs that are related to understanding structural properties of splat measures $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d))$ and splat functions. This includes Propositions 1, 2 on the well-posedness and differentiability of splat functions. We also provide the proof of the two universal approximation theorems, Proposition 3 and Theorem 3. We restate these claims for reading convenience.

We dedicate the rest of the section to proving Theorem 1, which we split into a few steps.

Proposition 1 (Heterogeneous Mixtures). Suppose that $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$ is of the form $\mu(dv, dx) = \mu(dv, x)\, dx$ and where $\mathbb{E}_{(v,x)\sim\mu}[\|v\|_s^s] < \infty$, for $s \geq 1$. Then the function $f_\mu(x) := \mathbb{E}_{v\sim\mu(\cdot,x)}[v]$ belongs to $\mathcal{L}^s(\mathbb{R}^d; \mathbb{R}^p)$, and furthermore, there exists $\nu \in \mathcal{P}(\mathbb{R}^p)$ (the 'mixture components') and a density $k(v, x) > 0$ (the 'heterogeneous mixture weights') so that

$$\mu(dv, x) = \nu(dv)\, k(v, x) \qquad f_\mu(x) = \mathbb{E}_{v\sim\nu}[vk(v, x)].$$

Proof. If $\mathbb{E}_{(v,x)\sim\mu}[\|v\|_s^s]$ then by Jensen's inequality,

$$\|f_\mu\|^s_{\mathcal{L}^s(\mathbb{R}^d;\mathbb{R}^p)} = \int \left\| \int v\, \mu(dv, x) \right\|_s^s dx \leq \iint \|v\|_s^s\, \mu(dv, x)\, dx < \infty$$

so $f_\mu \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}^p)$. Now take ν to be the v-marginal $((v, x) \mapsto v)_\# \mu$. By the disintegration theorem (Ambrosio et al., 2008) there exists a family of measures $k_v \in \mathcal{P}(\mathbb{R}^d)$ so that $\mu(dv, x)dx = k_v(dx)\nu(dv)$. We must check that each k_v has a density, and for this it is sufficient to check that for any $\phi \in \mathcal{L}^1_b(\mathbb{R}^p)$ and any sequence $\psi_k \in \mathcal{L}^1(\mathbb{R}^d)$ that converges $\phi_k \overset{k\to\infty}{\longrightarrow} 0$ in $\mathcal{L}^1$,

$$\begin{aligned}\int \phi(v) \left(\int \psi_k(x)\, k_v(dx) \right) \nu(dv) &= \iint \phi(v)\psi_k(x)\mu(dv, dx) \\ &\leq \sup_{v\in\mathrm{supp}(\nu)} \phi(v) \cdot \|\psi_k\|_{\mathcal{L}^1(\mathbb{R}^d)} \\ &\overset{k\to\infty}{\longrightarrow} 0.\end{aligned}$$

by Hölder's inequality. We have shown that for ν-almost every v, any sequence converging to zero in $\mathcal{L}^1(\mathbb{R}^d)$ also converges to zero in $\mathcal{L}^1(\mathbb{R}^d; k_v)$. The proof follows by identifying $k(v, x)$ as the density of $k_v(dx)$. □

Next we prove the sufficient conditions on μ for $f_\mu(\cdot)$ to be continuous and/or differentiable.

Proposition 2 (Sufficient conditions for regularity). Let $\mu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$.

1. The map $f_\mu(\cdot)$ has uniform modulus of continuity
$$\begin{aligned}\omega(\epsilon) &= \sup\left\{ \left| \mathbb{E}_{\mu(\cdot,x)}[v] - \mathbb{E}_{\mu(\cdot,y)}[v] \right| : \|x - y\| \leq \epsilon \right\} \\ &\leq \sup\left\{ W_1(\mu(\cdot, x), \mu(\cdot, y)) : \|x - y\| \leq \epsilon \right\}.\end{aligned}$$
In particular, $f_\mu(\cdot)$ is absolutely continuous whenever $(\mu(\cdot, x))_{x\in\mathbb{R}^d}$ is a W_1-absolutely continuous measure-valued process on $\mathbb{R}^d$.
2. Suppose μ is a splat measure whose mother splat ρ has $s \geq 0$ bounded derivatives. Then $f_\mu(x)$ also has s bounded derivatives.

Proof. The first property is by definition of the Wasserstein-1 distance. For the second since $\rho \in C^s_b(\mathbb{R}^d)$ and for any $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, $x \mapsto Ax + b$ is $C^\infty_b(\mathbb{R}^d)$, the function $(v, x) \mapsto v\,[(A(\cdot) + b)_\#\rho](x)$ has s bounded and continuous derivatives in x and is dominated by the ν-integrable function $v \mapsto C_s v \|\partial_s \rho\|_\infty \sup_{v\in\mathrm{supp}(\nu)} \|A\|^s_{\mathrm{op}}$ where C_s is a universal constant. Derivatives of $f_\mu(\cdot)$ are therefore given by differentiation under the integral. □

Next, we show that under a wide range of ρ, splat models satisfy the conditions of Cybenko's theorem on universal approximation by two-layer neural networks. We restate the universal approximation theorem and check that its conditions hold.

Theorem 2 (Definition 1 and Theorem 1, Cybenko (1989)). *We say that σ is discriminatory if for any measure $\mu \in I_n := [0,1]^n$,*

$$\int_{I_n} \sigma(y^T x + \theta)\, d\mu(x) = 0$$

for all $y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ implies that $\mu = 0$. If σ is any continuous discriminatory function, then the finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $C_b^0(I_n)$.

Corollary 1. *Let $\Omega \subseteq \mathbb{R}^d$ compact and suppose ρ is a continuous density with marginal $\rho_1(x_1) = \int_{\mathbb{R}^{d-1}} \rho(x_1, x_2 \dots x_d)\, dx$. Then for any $f \in C_b^0(\Omega; \mathbb{R})$ and $\epsilon > 0$, there exists a k-splat measure μ (with mother splat ρ) such that*

$$\sup_{x \in \Omega} |f(x) - f_\mu(x)| < \epsilon.$$

Proof. By scaling we may assume $\Omega \subseteq [0,1]^d$. We claim that ρ_1 is a discriminatory function and that the approximator $G(x)$ with $\sigma = \rho_1$ and with parameters $\{(\alpha_j, y_j, \theta_j) : j = 1 \dots k\}$, which is guaranteed by Theorem 2 for some $k \gg 1$, can be realized as a splat model. It is easy to see the latter: to realize $G(x)$ as a splat, take $A_{v_j} = e_1 y_j^T$, $b_{v_j} = e_1 \theta_j$, and $v_j = \alpha_j$.

Let μ be a nonzero measure on $[0,1]^d$ and let $s \in \mathrm{supp}(\mu)$. Assume for contradiction that for all $y \in \mathbb{R}^d, \theta \in \mathbb{R}$,

$$0 = \int_{[0,1]^d} \rho_1(y^T x + \theta)\, \mu(dx) \geq t\, P_{x \sim \mu}(\rho_1(y^T x + \theta) > t).$$

Thus $\mu(B_{y,\theta}(t)) = 0$ for all sets $B_{y,\theta}(t) := \{x \in [0,1]^d : \rho_1(y^T x + \theta) > t\}$ and all $t > 0$. By continuity, the set $\rho_1^{-1}((t, \infty))$ is open, so that by affine rescaling we can construct from the basis $\{B_{y,\theta}(t) : y \in \mathbb{R}^d, t, \theta \in \mathbb{R}\}$ the entire Borel σ-algebra on $[0,1]^d$, which would imply $\mu = 0$. □

Finally, we give a more constructive proof of a universal approximation theorem on sets $\Omega \subset \mathbb{R}^d$ whose uniform measure $\mathcal{U}(\Omega)$ has a finite *Poincaré constant*. This constant is the smallest $C_\Omega > 0$ for which any $f \in H^1(\Omega)$ satisfies the inequality

$$\operatorname*{var}_{X \sim \mathcal{U}(\Omega)}(f(X)) \leq C \|\nabla f\|^2_{\mathcal{L}^2(\Omega)}.$$

This is satisfied by most 'nice' sets, namely whenever Ω is an open connected set with a Lipschitz boundary (Evans, 2010, §5.8 Theorem 1).

Theorem 3. *Let $\Omega \subseteq \mathbb{R}^d$ compact and suppose $f \in C_b^1(\Omega; \mathbb{R}^p)$. Suppose further that the measure $\omega = \mathrm{Unif}(\Omega)$ has finite Poincaré constant $C_\Omega < \infty$. For $\epsilon > 0$ there exists a k-splat measure μ at most $k \leq \frac{C_\Omega}{2} \epsilon^{-2(d+2)} \left(L_f B_\rho \epsilon + B_f L_\rho\right)^2$ splats and for which*

$$\sup_{x \in \Omega} \|f(x) - f_\mu(x)\| < (1 + L_f M_\rho)\epsilon.$$

Where the constants are:

$$B_\rho = \|\rho\|_\infty \qquad B_f = \|f\|_\infty$$
$$L_\rho = \sup_{x,y \in \mathbb{R}^d} \frac{|\rho(x) - \rho(y)|}{\|x - y\|_2} \qquad L_f = \sup_{x,y \in \Omega} \frac{\|f(x) - f(y)\|_2}{\|x - y\|_2}$$
$$M_\rho = \mathbb{E}_{z \sim \rho}[\|z\|_2]$$

Proof. Set $f^\epsilon(x) = \mathbb{E}_{z \sim \rho}[f(x + \epsilon z)]$, then

$$\|f(x) - f^\epsilon(x)\|_2 \leq \int \|f(x) - f(x + \epsilon z)\|_2\, \rho(dz) \leq L_f M_\rho \epsilon$$

We now show that, for $k \gg 1$ also to be determined, there exist a k-splat measure μ so that

$$\sup_{x\in\Omega} \|f_\mu(x) - f^\epsilon(x)\|_2 \le \epsilon.$$

Take random $x_1, \dots, x_k \sim \omega^{\otimes k}$ and take $\nu = \frac{1}{k}\sum_{i=1}^k \delta_{f(x_i)}$, $k(v,x) = \epsilon^{-d}\rho(x/\epsilon)$, and $\mu(dv,x)dx = \nu(dv)k(v,x)$. Then we have

$$\begin{aligned} P\left(x_1 \dots x_k \mapsto \sup_{x\in\Omega} \|f_\mu(x) - f^\epsilon(x)\|_2 > \epsilon\right) &\le \frac{\mathrm{Var}\left(x_1 \dots x_k \mapsto \sup_{x\in\Omega} \|f_\mu(x) - f^\epsilon(x)\|_2\right)}{\epsilon^2} \\ &\le \frac{1}{\epsilon^2}\sum_{i=1}^k \mathbb{E}[\mathrm{Var}_i(x_1 \dots x_k \mapsto \sup_{x\in\Omega} \|f_\mu(x) - f^\epsilon(x)\|_2)] \\ &= \frac{1}{\epsilon^2}\sum_{i=1}^k \mathbb{E}[\mathrm{Var}_i(x_1 \dots x_k \mapsto \sup_{\substack{x\in\Omega \\ r\in S^{p-1}}} \langle r, f_\mu(x) - f^\epsilon(x)\rangle] \end{aligned}$$

where $\mathrm{Var}_i(f(x_1, \dots, x_k))$ is the variance of $f(\cdot)$ holding $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k$ fixed and resampling $x_i \sim \omega$ independently. In advance of applying Poincaré's inequality, invoke the envelope theorem to see,

$$\begin{aligned} \nabla_{x_i}\left\{x_1 \dots x_k \mapsto \sup_{\substack{x\in\Omega \\ r\in S^{p-1}}} \langle r, f_\mu(x) - f^\epsilon(x)\rangle\right\} &= (r^*)^T D_{x_i}\left\{\frac{1}{k}\sum_{i=1}^k f(x_i)\epsilon^{-d}\rho((x-x_i)/\epsilon)\right\} \\ &= \frac{1}{k\epsilon^d}(r^*)^T Df(x_i)\rho((x-x_i)/\epsilon) \\ &\quad - \frac{1}{k\epsilon^{d+1}}\langle r^*, f(x_i)\rangle(\nabla\rho)((x-x_i)/\epsilon) \end{aligned}$$

where x^*, r^* attain the supremum. Its norm is bounded by

$$\|\dots\| \le \frac{1}{k\epsilon^{d+1}}\left(L_f B_\rho \epsilon + B_f L_\rho\right)$$

where $\|f\|_\infty \coloneqq \sup_{x\in\Omega}\|f(x)\|_2$, and $C_\rho \coloneqq \sup_{x,y\in\mathbb{R}^d} |\rho(x)-\rho(y)|/\|x-y\|_2$, so that by Poincaré inequality,

$$P(x_1, \dots, x_k \mapsto \sup_{x\in\Omega} |f_\mu(x) - f^\epsilon(x)| > \epsilon) \le \frac{C_\Omega}{\epsilon^{2(d+2)}k}\left(L_f B_\rho \epsilon + B_f L_\rho\right)^2.$$

Consequently, for $k = \frac{1}{2C_\Omega}\epsilon^{-2(d+2)}\left(L_f B_\rho \epsilon + B_f L_\rho\right)^2$, it occurs with probability at least one half (over $x_1, \dots, x_k \sim \omega^{\otimes k}$) that $\sup_{x\in\Omega}\|f_\mu(x) - f^\epsilon(x)\|_2 \le \epsilon$. By the triangle inequality

$$\sup_{x\in\Omega}\|f_\mu(x) - f(x)\|_2 \le \sup_{x\in\Omega}\|f_\mu(x) - f^\epsilon(x)\|_2 + \|f^\epsilon(x) - f(x)\|_2 \le (1 + L_f M_\rho)\epsilon.$$

□

Theorem 4. (Uniform approximation lower bound). Set $\mathcal{F} = \{f : [0,1]^d \to \mathbb{R} : f \text{ is 1-Lipschitz}\}$ and let $\mathcal{S}_{1,d}^{(k)}$ be the family of k-splat models from $\mathbb{R}^p$ to $\mathbb{R}$. Then we have,

$$\sup_{f\in\mathcal{F}} \inf_{\hat{f}\in\mathcal{S}_{1,d}^{(k)}} \|f - \hat{f}\|_\infty \gtrsim k^{-1/d}$$

where the inequality holds up to universal constants.

Proof. We invoke some well known facts about the metric entropy of $\mathcal{F}$ and of the parametric class $\mathcal{S}_{1,d}^{(k)}$ Wainwright (2019). The metric entropy (in uniform norm) of $\mathcal{F}$ scales as $H(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \sim \epsilon^{-d}$, whereas by parameter counting, one has $H(\epsilon, \mathcal{S}_k, \|\cdot\|_\infty) \sim k\log(\epsilon^{-1})$. Uniform approximation is therefore impossible unless $H(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim H(\epsilon, \mathcal{S}_k, \|\cdot\|_\infty)$, which forces the rate $k \gtrsim \epsilon^{-d}$ when $\epsilon < 1$. □

B GRADIENT FLOWS ON MEASURE SPACES

We provide in this section some of the mathematical background on Wasserstein-Fisher-Rao calculus that is required to state our main results. We then give the proofs of Theorem 1, Proposition 5, and some helpful calculation rules for computing gradients with respect to splat measures.

B.1 REVIEW OF WASSERSTEIN AND FISHER-RAO CALCULUS

First, we define the Hellinger and the 'static' Wasserstein distances.

Definition 3 (Wasserstein Distance (Villani, 2003))**.** Let $\mathcal{M}$ be a Riemannian manifold (endowed with the canonical volume measure) and let $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ be measures with finite p-th moments. The p-*Wasserstein* distance is,

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int d_{\mathcal{M}}(X, Y)^p \, \pi(dX, dY)$$

where $d_{\mathcal{M}}(\cdot, \cdot)$ is the geodesic distance on $\mathcal{M}$ and $\Gamma(\mu, \nu)$ is the set of all joint couplings whose marginals are $\pi_X = \mu$, $\pi_Y = \nu$. We denote by $\mathcal{W}_p(\mathcal{M})$ the space $\mathcal{P}_p(\mathcal{M})$ endowed with the W_p metric.

Definition 4 (Hellinger Distance (Amari, 1983))**.** Let $\mathcal{M}$ be as in 3 and let $\mu, \nu \in \mathcal{P}(\mathcal{M})$ be measures with finite second moments. Setting $\lambda = \frac{1}{2}(\mu + \nu)$, the Hellinger distance is

$$H^2(\mu, \nu) = \int \left(\sqrt{\frac{d\mu}{d\lambda}}(x) - \sqrt{\frac{d\nu}{d\lambda}}(x) \right)^2 \lambda(dx)$$

where $\frac{d\mu}{d\lambda}(x)$ is the Radon-Nikodym derivative. Here, λ can be replaced by any measure $\lambda' \gg \mu, \nu$. We denote by $\mathcal{H}(\mathcal{M})$ the space $\mathcal{P}(\mathcal{M})$ endowed with the H metric.

For our purposes we only need to consider rather two rather simple manifolds: the Euclidean space (with the canonical $\|\cdot\|_2$ metric) and the Bures-Wasserstein space $\mathsf{BW}_\rho(\mathbb{R}^d)$. We postpone the proof of Proposition 5, where we formally define the Bures-Wasserstein manifold, and proceed with the review. For further reference and for proofs of the background theorems stated here, see wonderful reference (Santambrogio, 2015) as well as the more contemporary (Chewi et al., 2025a), which emphasizes a statistical perspective.

In their seminal 1998 paper, Jordan, Kinderlehrer, and Otto (Jordan et al., 1998) introduced an iterative 'minimizing movements' scheme as a variational approach to solving the Fokker-Plank equation in fluid dynamics, leading them (largely by accident) to discover the geometric perspective on the so-called 'static' Wasserstein distance. Shortly afterwards, Benamou & Brenier (2000) developed the following elegant formulation.

Theorem 4 (Benamou–Brenier Benamou & Brenier (2000))**.** *Given two probability measures $\mu_0, \mu_1 \in \mathcal{W}_2(\mathcal{M})$, it holds that*

$$W_2^2(\mu_0, \mu_1) \tag{4}$$

$$= \inf_{\tilde{v}_t} \left\{ \int_0^1 \mathbb{E}_{X \sim \mu_t} \left\| \tilde{v}_t(X) \right\|^2 dt : \partial_t \mu_t + \operatorname{div}(\mu_t \tilde{v}_t) = 0, \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \right\}. \tag{5}$$

where $\|\tilde{v}_t(x)\|^2 := \langle \tilde{v}_t(x), \tilde{v}_t(x) \rangle_x$ is the tangent space norm at $x \in \mathcal{M}$. Moreover, the optimizer $\{v_t\}$ induces a curve $\{\mu_t\}$ with the following properties.

1. *At time $t = 0$, $v_0(x) = T_{\mu_0 \to \mu_1}(x) - x$ where $T_{\mu_0 \to \mu_1}(x)$ is the optimal transportation map from μ_0 to μ_1.*
2. *$(v_t)_{t \geq 0}$ has zero acceleration, in the sense that $v_t(X_t(x)) = T_{\mu_0 \to \mu_1}(x) - x$ is constant in time.*
3. *For times $t > 0$, $\mu_t = \mathsf{Law}\left[X + t(T_{\mu_0 \to \mu_1}(X) - X)\right]$ with $X \sim \mu_0$, and the joint law of $(X, X + t(T_{\mu_0 \to \mu_1}(X) - X))$ is the optimal coupling of (μ_0, μ_t).*

The integrand of equation 4 is called the 'kinetic energy' of the ensemble μ_t and the minimizing curves $(\mu_t)_{t\in[0,1]}$ are the geodesics of $\mathcal{W}_2(\mathcal{M})$. One can show by analyzing the Euler-Lagrange equations of equation 4 that the tangent vectors $(v_t)_{t\in[0,1]}$ have the form $v_t(x) = \nabla\phi_t(x)$. Conversely, for any curve of measures $(\nu_t)_{t\in[0,1]}$ that is *absolutely continuous* in the following sense, there exists a ϕ_t whose gradient weakly solves the transport equation $\partial_t\nu_t = -\operatorname{div}_{\mathcal{M}}(\nu_t\nabla\phi_t)$. Evidently ϕ_t itself is the solution of a Poisson equation with positive semidefinite coefficents, which already enjoys strong existence, uniqueness, and regularity theorems Evans (2010).

Definition 5 (Absolute continuity)**.** Let $(\mu_t)_{t\in[0,1]} \in \mathcal{W}_2(\mathcal{M})$ be a curve of measures. The *metric derivative* $|\dot{\mu}|(t)$ is defined,

$$|\dot{\mu}|(t) = \lim_{h\to 0} \frac{W_2(\mu_{t+h}, \mu_h)}{h}$$

and μ_t is *absolutely continuous* (in W_2 metric) at $t \in [0,1]$ if $|\dot{\mu}|(t) < \infty$.

Theorem 5. *If $(\mu_t)_{t\in[0,1]}$ is absolutely continuous, then there exists a function $\phi_t : \mathcal{M} \to \mathbb{R}$ which weakly solves*

$$\partial_t\mu_t + \operatorname{div}_{\mathcal{M}}(\mu_t\nabla\phi_t) = 0$$

and moreover for which $\nabla\phi_t \in \mathcal{L}^2(\mu_t)$ for every $t \in [0,1]$.

The tangent spaces of $T_\mu\mathcal{W}_2(\mathcal{M})$ are defined in the usual way as the linear span of the tangent vector of any absolutely continuous curve passing through μ_t. It is customary to identify the tangent vector $\partial_t\mu_t = -\operatorname{div}_{\mathcal{M}}(\mu_t\nabla\phi_t)$ with its drift function $\nabla\phi_t$, since the metric $\langle\cdot,\cdot\rangle_\mu$ has a simple expression in terms of the drifts.

Theorem 6. *The space $\mathcal{W}_2(\mathcal{M})$ is a Riemannian manifold with tangent space structure*

$$T_\mu\mathcal{W}_2(\mathcal{M}) = \overline{\{v \in \mathcal{L}^2(\mu) :\ \exists\,\phi : C_c^\infty(\mathcal{M}) \to \mathbb{R}, \nabla\phi = v\}}_{\mathcal{L}^2(\mu)}$$

with inner product $\langle u, v\rangle_\mu = \mathbb{E}_{X\sim\mu}\langle u(X), v(X)\rangle_X$, where $\langle\cdot,\cdot\rangle_X$ is the metric on $T_X\mathcal{M}$.

Finally, the Wasserstein gradient $\nabla\mathcal{F}(\mu)$ is just the representer of $\xi \mapsto \partial_{\epsilon=0}\mathcal{F}(\mu + \epsilon\xi)$ in $T_\mu\mathcal{W}_2(\mathcal{M})$.

Definition 6 (Wasserstein Gradient)**.** Let $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, the Wasserstein gradient at μ is (when it exists) the function $\nabla\mathcal{F}(\mu) \in T_\mu\mathcal{W}_2(\mathcal{M})$ such that for every $\xi \in T_\mu\mathcal{W}_2(\mathcal{M})$,

$$\partial_{\epsilon=0}\mathcal{F}(\mu_\epsilon) = \langle\delta\mathcal{F}, \xi\rangle_\mu = \int \langle\delta\mathcal{F}(X), \xi(X)\rangle_X\, \mu(dX).$$

where $(\mu_t)_{t\in(-\delta,\delta)}$ is any curve with tangent vector ξ at $\mu_0 = \mu$.

We turn our attention to the Fisher-Rao geometry, which is more straightforward to explain. A standard reference for this material is the book Amari (2016). The 'admissible' tangent vectors in this geometry are those belonging to the absolutely continuous curves,

Definition 7 (Fisher-Rao absolute continuity)**.** A curve $(\mu_t)_{t\in[0,1]}$ is *absolutely continuous* at $t \in [0,1]$ if its Hellinger metric derivative is finite:

$$|\dot{\mu}|(t) \coloneqq \lim_{\epsilon\to 0} \frac{H(\mu_{t+\epsilon}, \mu_t)}{\epsilon} < \infty.$$

Theorem 7. *If $(\mu_t)_{t\in[0,1]} \in \mathcal{H}(\mathcal{W})$ is absolutely continuous at $t \in [0,1]$, then there exists $\alpha_t : \mathcal{M} \to \mathbb{R}$ such that $\partial_t\mu_t = \alpha_t\mu_t$, and where $\mathbb{E}_{X\sim\mu_t}[\alpha_t(X)] = 0$.*

The Fisher-Rao geometry is essentially based on the fact that for any probability density $\mu(x) : \mathcal{M} \to \mathbb{R}$, the square root density $\sqrt{\mu}(x)$ is an element of the unit sphere $S(\mathcal{L}^2(\mathcal{M})) \coloneqq \{v \in \mathcal{L}^2(\mathcal{M}) : \|v\| = 1\}$, since it has $\int(\sqrt{\mu}(x))^2\,dx = 1$ trivially. We endow it with the pullback metric on $\mathcal{P}(\mathcal{M})$ induced by the map $\iota : \mu \mapsto \sqrt{\mu} \in S(\mathcal{L}^2(\mathcal{M}))$. What that means concretely is:

- The geodesic connecting μ_0, μ_1 is the spherical linear interpolant connecting $\sqrt{\mu_0}, \sqrt{\mu_1}$:

$$\sqrt{\mu_t} = \left(\frac{\sin((1-t)\theta)}{\sin(\theta)}\right)\sqrt{\mu_0} + \left(\frac{\sin(t\theta)}{\sin(\theta)}\right)\sqrt{\mu_1}, \qquad \cos(\theta) = \langle\sqrt{\mu_0}, \sqrt{\mu_1}\rangle$$

- Each tangent space $T_\mu\mathcal{H}(\mathcal{M})$ is by definition isomorphic to $T_{\sqrt{\mu}}\mathcal{H}(\mathcal{M})$. The isomorphism map is the pushforward or 'differential' $d\iota$, which is given by

$$d\iota(\xi) \coloneqq \partial_\epsilon\, \iota(\mu + \epsilon\xi)\Big|_{\epsilon=0} = \frac{\xi}{2\sqrt{\mu}}$$

  and so the inner product on $T_\mu\mathcal{H}(\mathcal{M})$ is

$$\langle \xi, \psi\rangle_\mu = \langle d\iota(\xi), d\iota(\psi)\rangle_{\sqrt{\mu}} = \frac{1}{4}\int \frac{\xi\psi}{\mu}(x)\, dx.$$

- By absolute continuity we may write $\xi = \alpha\mu$, $\psi = \beta\mu$ for densities $\alpha, \beta : \mathcal{M} \to \mathbb{R}$, and the Fisher-Rao metric can be rewritten as

$$\langle \alpha\mu, \beta\mu\rangle = \frac{1}{4}\int \alpha(x)\beta(x)\mu(dx).$$

  It is customary to identify the tangent vectors with their densities and to write

$$T_\mu\mathcal{H}(\mathcal{M}) = \overline{\{\alpha \in C_c^\infty(\mathcal{M}) : \mathbb{E}_\mu[\alpha(X)] = 0\}}^{\mathcal{L}^2(\mu)}$$

  with metric $\langle\cdot,\cdot\rangle_\mu = \frac{1}{4}\langle\cdot,\cdot\rangle_{\mathcal{L}^2_\mu(\mathcal{M})}$.

- The Fisher-Rao gradient of a functional $\mathcal{F}$ is (when it exists) the function $\nabla^{\mathsf{FR}}\mathcal{F} \in T_\mu\mathcal{H}(\mathcal{M})$ that satisfies, for $\alpha \in T_\mu\mathcal{H}(\mathcal{M})$,

$$\partial_\epsilon \mathcal{F}(\mu + \epsilon\alpha\mu)\Big|_{\epsilon=0} = \langle \nabla^{\mathsf{FR}}\mathcal{F}(\mu), \alpha\rangle_\mu.$$

Finally, the Wasserstein-Fisher-Rao geometry is the one whose tangent spaces are direct sums of the Wasserstein and the Fisher-Rao tangent spaces. In other words, Wasserstein-Fisher-Rao tangent vectors have the form

$$\partial_t\mu_t + \mathrm{div}_{\mathcal{M}}(\mu_t\nabla\phi_t) = \alpha_t\mu_t, \qquad \nabla\phi_t \in T_\mu\mathcal{W}(\mathcal{M}),\ \alpha_t \in T_\mu\mathcal{H}(\mathcal{M}).$$

and as usual we identify each tangent vector with its coefficients $(\nabla\phi_t, \alpha_t)$. The Wasserstein-Fisher-Rao metric is

$$\langle(\nabla\phi_t, \alpha_t), (\nabla\psi_t, \beta_t)\rangle_\mu = \frac{1}{4}\int \alpha_t(x)\beta_t(x)\,\mu(dx) + \int \langle\nabla\phi_t(x), \nabla\psi_t(x)\rangle\,\mu(dx),$$

and its tangent spaces are

$$T_\mu\mathcal{W}(\mathcal{M}) \oplus T_\mu\mathcal{H}(\mathcal{M}) = \overline{\{(a, v) : \exists\, \alpha, \phi \in C_c^\infty(\mathcal{M}) \to \mathbb{R},\ v = \nabla\phi\}}^{\mathcal{L}^2(\mu)},$$

and the Wasserstein-Fisher-Rao gradient of $\mathcal{F}$ is $\nabla^{\mathsf{WFR}}\mathcal{F}(\mu) = (\mathbb{W}\mathcal{F}(\mu), \nabla^{\mathsf{FR}}\mathcal{F}(\mu))$.

B.2 Proofs of Proposition 5 and Theorem 1

The proof of Proposition 5 follows exactly the same steps as the case $\rho = \mathcal{N}(0, I_d)$.

Proposition 5 (Splats are a generalized Bures-Wasserstein manifold) **.** Let $\rho \in \mathcal{P}_{ac}(\mathbb{R}^d)$ be a centered isotropic mother splat. We denote the set of all splats as,

$$\mathsf{BW}_\rho(\mathbb{R}^d) \coloneqq \left\{(A(\cdot) + b)_\#\,\rho : A \in \mathbb{R}^{d\times d}, b \in \mathbb{R}^d\right\}.$$

Then $\mathsf{BW}_\rho(\mathbb{R}^d)$ is a geodesically convex subset of $\mathcal{W}_2(\mathbb{R}^d)$, and on this space the Wasserstein metric reduces to the *Bures-Wasserstein metric* (Modin, 2016; Bhatia et al., 2019),

$$W_2^2(\rho_{A,b}, \rho_{R,s}) = \|b - s\|_2 + \|A\|_F^2 + \|R\|_F^2 - 2\|A^T R\|_* \qquad A, R \in \mathbb{R}^{d\times d} \quad b, s \in \mathbb{R}^d$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_*$ is the nuclear norm.

Proof. Fix $\rho_{A_0,b_0} = (A_0(\cdot)+b_0)_{\#}\rho$ and $\rho_{A_1,b_1} = (A_1(\cdot)+b_1)_{\#}$. By Brenier's uniqueness theorem (Chewi et al., 2025a, Theorem 1.16), the map

$$T(x) = b_1 + A_0^{-1}(A_0A_1A_1^TA_0)^{1/2}A_0^{-1}(x-b_0)$$

is the optimal transport map $T_{\#}\rho_{A_0,b_0} = \rho_{A_1,b_1}$ as it is the gradient of a convex function. It follows that the geodesic connecting ρ_0 to ρ_1 is $\rho_t = ((1-t)\mathsf{id} + tT)_{\#}\rho_0 \in \mathsf{BW}_\rho(\mathbb{R}^d)$. As ρ is centered and isotropic, the distance $W_2(\rho_0,\rho_1)$ is

$$\begin{aligned} W_2(\rho_0,\rho_1) &= \mathbb{E}_\rho\left[\|(A_0X+b_0) - T(A_0X+b_0)\|^2\right] \\ &= \mathbb{E}_\rho\left[\|b_0 - b_1 + (A_0 - A_0^{-1}(A_0A_1A_1^TA_0)^{1/2})X\|\right] \\ &= \|b_0-b_1\|^2 + \|A_0 - A_0^{-1}(A_0A_1A_1^TA_0)^{1/2}\Sigma_X^{1/2}\|_F^2 \end{aligned}$$

where $\Sigma_X = \mathrm{Cov}_\rho(X) = I_d$. Thus $W_2(\rho_0,\rho_1) = W_2(\mathcal{N}(b_0, A_0A_0^T), \mathcal{N}(b_1, A_1A_1^T))$ is the Bures-Wasserstein metric Bhatia et al. (2019). □

We now proceed to the proof of Theorem 1. We develop these calculations at a formal level – the reader who wishes to understand an entirely rigorous proof may compare to Ambrosio et al. (2008, Chapter 11).

Theorem 1 (Wasserstein-Fisher-Rao gradient of $\mu \mapsto \mathcal{F}(f_\mu)$)**.** Let $\mu \in \mathcal{S}_{p,d} \coloneqq \mathcal{P}(\mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d))$ and let $\mathcal{F} : \mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p) \to \mathbb{R}$ be a functional. Assume that ρ has a sub-exponential density. Then the Fisher-Rao gradient is given (if it exists) by the formula,

$$\nabla_\mu^{\mathsf{FR}}\mathcal{F}(f_\mu)(v,A,b) = \mathbb{E}_{X\sim\rho_{A,b}}[\langle\delta\mathcal{F}(X),v\rangle] - \mathbb{E}_{v,A,b\sim\mu}[\mathbb{E}_{X\sim\rho_{A,b}}[\langle\delta\mathcal{F}(X),v\rangle].$$

and the Wasserstein gradient is given (if it exists) by the formula,

$$\begin{aligned} \mathbb{W}_v\mathcal{F}(f_\mu)(v,A,b) &= \mathbb{E}_{X\sim\rho_{A,b}}[\delta\mathcal{F}(X)] \\ \mathbb{W}_A\mathcal{F}(f_\mu)(v,A,b) &= -\mathbb{E}_{X\sim\rho_{A,b}}\left[\langle\delta\mathcal{F}(X),v\rangle\left(I_d + \nabla_x\log\rho_{A,b}(X)(X-b)^T\right)A^{-T}\right] \\ \mathbb{W}_b\mathcal{F}(f_\mu)(v,A,b) &= -\mathbb{E}_{X\sim\rho_{A,b}}\left[\langle\delta\mathcal{F}(X),v\rangle\nabla_x\log\rho_{A,b}(X)\right] \end{aligned}$$

where the argument of $\delta\mathcal{F}(\cdot) = \delta\mathcal{F}[f](\cdot)$ was suppressed.

Proof. We begin by calculating the Fisher-Rao gradient. Along the way, we develop a 'chain rule,' which may aid in future calculations. Specifically we view $\mu \mapsto f_\mu$ as a mapping between manifolds, $f_{(\cdot)} : \mathcal{H}(\mathcal{S}_{p,d}) \to \mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p)$, and we calculate its *differential* $df_\mu : T_\mu\mathcal{H}(\mathcal{S}_{p,d}) \to \mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p)$. For $\alpha \in T_\mu\mathcal{H}(\mathcal{S}_{p,d})$,

$$d^{FR}f_\mu[\alpha](x) = \partial_\epsilon f_{\mu+\epsilon\alpha\mu}\big|_{\epsilon=0} = \int_{\mathcal{S}_{p,d}} v\,\alpha(v,A,b)\rho_{A,b}(x)\,\mu(dv,dA,db).$$

Invoking the definition of the first variation, our chain rule takes the form

$$\begin{aligned} \partial_\epsilon\mathcal{F}(f_{\mu+\epsilon\alpha}) &= \langle\delta\mathcal{F}[f_\mu], df_\mu[\alpha]\rangle_{\mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p)} \\ &= \iint\langle\delta\mathcal{F}[f_\mu](x),v\rangle\,\alpha(v,A,b)\,\rho_{A,b}(x)\,dx\,\mu(dv,dA,db) \\ &= \int\alpha(v,A,b)\left(\mathbb{E}_{X\sim\rho_{A,b}}[\langle\delta\mathcal{F}[f_\mu](X),v\rangle]\right)\mu(dv,dA,db). \end{aligned}$$

We identify the Fisher-Rao gradient by matching the integrand to the definition.

Computing the Wasserstein gradient follows largely the same steps. One checks the following auxiliary calculations.

$$\begin{aligned} \nabla_A\left\{\rho\left(A^{-1}(x-b)\right)|\det A|^{-1}\right\} &= -|\det A|^{-1}A^{-T}(\nabla\rho)(A^{-1}(x-b))(x-b)^TA^{-T} \\ &\quad - \rho\left(A^{-1}(x-b)\right)|\det A|^{-1}A^{-T} \\ &= -\rho_{A,b}(x)\left(I + \nabla_x\log\rho_{A,b}(x)(x-b)^T\right)A^{-T} \\ \nabla_b\left\{\rho\left(A^{-1}(x-b)\right)|\det A|^{-1}\right\} &= -(\nabla\rho)(A^{-1}(x-b))|\det A|^{-1} \\ &= -\nabla_x\log\rho_{A,b}(x)\,\rho_{A,b}(x). \end{aligned}$$

Now fix $u \in T_\mu \mathcal{W}_2(\mathcal{S}_{p,d})$. We calculate the differential $d^W f_\mu[u]$ as,

$$\begin{aligned} d^W f_\mu[u](x) &= -\int_{\mathcal{S}_{p,d}} v\rho_{A,b}(x)\ \mathrm{div}_{\mathcal{S}_{p,d}}(u\mu)(dv, dA, db) \\ &= \int \langle \nabla_{v,A,b} \{v\rho_{A,b}(x)\}, u(v,A,b)\rangle\ \mu(dv, dA, db) \\ &= \int u_v(v,A,b)\rho_{A,b}(x)\,\mu(dv, dA, db) \\ &\quad + \int v\,\langle \nabla_A \{\rho_{A,b}\}(x), u_A(v,A,b)\rangle\ \mu(dv, dA, db) \\ &\quad + \int v\,\langle \nabla_b \{\rho_{A,b}\}(x), u_b(v,A,b)\rangle\ \mu(dv, dA, db) \end{aligned}$$

where $u = (u_v, u_A, u_b)$ are the coordinate function representations of u. Thus, setting $\xi = -\mathrm{div}_{\mathcal{S}_{p,d}}(u\mu)$,

$$\begin{aligned} \partial_\epsilon \mathcal{F}(f_{\mu+\epsilon\xi}) &= \langle \delta\mathcal{F}[f_\mu], df_\mu[u]\rangle_{\mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p)} \\ &= \iint \langle \delta\mathcal{F}[f_\mu](x), u_v(v,A,b)\rangle \rho_{A,b}(x)\,dx\,\mu(dv, dA, db) \\ &\quad + \iint \langle \delta\mathcal{F}[f_\mu](x), v\rangle\,\langle \nabla_A \{\rho_{A,b}\}(x), u_A(v,A,b)\rangle\ dx\,\mu(dv, dA, db) \\ &\quad + \iint \langle \delta\mathcal{F}[f_\mu](x), v\rangle\,\langle \nabla_b \{\rho_{A,b}\}(x), u_b(v,A,b)\rangle\ dx\,\mu(dv, dA, db). \end{aligned}$$

From this we see,

$$\begin{aligned} \mathbb{W}_v \mathcal{F}(f_\mu)(v,A,b) &= \int \delta\mathcal{F}[f_\mu](x)\,\rho_{A,b}(x)\,dx \\ \mathbb{W}_A \mathcal{F}(f_\mu)(v,A,b) &= -\int \langle \delta\mathcal{F}[f_\mu](x), v\rangle (I + \nabla_x \log \rho_{A,b}(x)(x-b)^T) A^{-T}\,\rho_{A,b}(x)\,dx \\ \mathbb{W}_b \mathcal{F}(f_\mu)(v,A,b) &= -\int \langle \delta\mathcal{F}[f_\mu](x), v\rangle \nabla_x \log \rho_{A,b}(x)\,\rho_{A,b}(x)\,dx. \end{aligned}$$

□

There is an interesting relationship between the geometry of $\mathcal{S}_{p,d}$ and that of the space $\mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$. Intuitively, splats in $\mathsf{BW}_\rho(\mathbb{R}^d)$ are like 'smoothed particles,' this relationship bears out concretely when calculating the Wasserstein and Fisher-Rao gradients in $\mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$.

Theorem 8 (Wasserstein-Fisher-Rao gradient of $\mu \mapsto \mathcal{F}(f_\mu)$ for $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$ and let $\mathcal{F} : \mathcal{L}^2(\mathbb{R}^d;\mathbb{R}^p) \to \mathbb{R}$ be a functional. Assume that ρ has a sub-exponential density. Then the Fisher-Rao gradient is given (if it exists) by the formula,*

$$\nabla^{\mathsf{FR}}_\mu \mathcal{F}(f_\mu)(v,x) = \langle \delta\mathcal{F}(x), v\rangle - \mathbb{E}_{X,V\sim\mu}[\langle \delta\mathcal{F}(X), V\rangle].$$

and the Wasserstein gradient is given (if it exists) by the formula,

$$\begin{aligned} \mathbb{W}_v \mathcal{F}(f_\mu)(v,x) &= \delta\mathcal{F}(x) \\ \mathbb{W}_x \mathcal{F}(f_\mu)(v,x) &= -v^T D_x \mathcal{F}(x) \end{aligned}$$

where the argument of $\delta\mathcal{F}(\cdot) = \delta\mathcal{F}[f](\cdot)$ was suppressed.

Comparing to the Wasserstein-Fisher-Rao gradients in Theorem 1, we see that the gradients have the same form as Theorem 8, but they are averaged with respect to the splat density. The $\mathbb{W}_b\mathcal{F}$ has the same relationship as can be seen by integrating by parts,

$$\mathbb{W}_b \mathcal{F}(f_\mu)(v,A,b) = -\mathbb{E}_{X\sim\rho_{A,b}}[\langle v, \delta\mathcal{F}(X)\rangle \nabla_x \log \rho_{A,b}(x)] = \mathbb{E}_{X\sim\rho_{A,b}}[v^T D_x \delta\mathcal{F}(X)].$$

These formulas follow from the following expressions for $d^W f_\mu$, $d^{FR} f_\mu$, which can be derived by the same arguments as for Theorem 1.

$$\langle \delta\mathcal{F}[f_\mu], d^W f_\mu(u) \rangle := \int \langle \delta\mathcal{F}[f_\mu] \otimes v^T D_x \delta\mathcal{F}(f_\mu), u(v,x) \rangle \, \mu(dv, dx)$$

$$\langle \delta\mathcal{F}[f_\mu], d^{FR} f_\mu(\alpha) \rangle := \int \langle \delta\mathcal{F}[f_\mu], v \rangle \alpha(v,x) \, \mu(dv, dx).$$

C Experimental Details and Additional Experiments

C.1 Wall clock runtimes

We report the average wall clock runtime to train each of the models in Figures 2 run on a 2020 Macbook Pro M1. We also include each model's parameter count and final MSE for comparison.

Model Type	Parameter count	Wall Time (s)	Final MSE
KAN	521	24.2853	0.2129
KAN	5201	121.9557	0.1404
KAN	7861	930.1017	0.0045
KAN	15601	358.9623	0.0728
KAN	20801	529.0655	0.1781
MLP	801	7.6694	0.2266
MLP	2001	8.7870	0.2165
MLP	4001	10.5255	0.2052
MLP	41001	12.1382	0.0390
MLP	252501	22.5081	0.0087
SPLAT	60	51.9213	0.0575
SPLAT	300	104.6052	0.0016
SPLAT	600	149.7740	0.0005
SPLAT	1200	243.9303	0.0006
SPLAT	1800	295.8554	0.0005
SPLAT	2400	378.7603	0.0008

Although they have significantly fewer parameters, training splat models has a more expensive wall clock time. We use autodifferentiation to compute gradients with respect to the splat model parameters, and the appearance of A^{-1} and $\det A^{-1}$ in equation equation 1 are particularly in our non-optimized implementation. We expect that a faster implementation with hardcoded gradient formulas will be significantly faster. Further, as can be seen in Figures 2 and 3, the splat models converge in far fewer iterations than KAN and MLP, so the reported run times include the time required to take many unnecessary training steps for the splat models.

C.2 Additional Approximation Experiments

The following plots show versions of our function approximation experiment in two different settings. The first shows the same experiment, but for fitting a sawtooth target function. The second shows the same experiment, but with $\{b_i\}_{i=1}^k$ initialized as a k-point chebyshev grid.

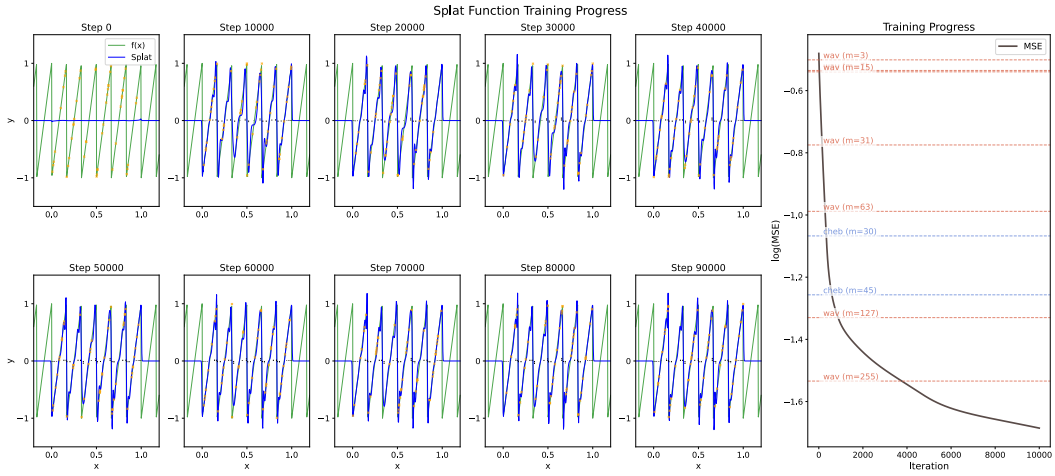

Figure 4: Fitting a sawtooth function in the setting of 1. This is a much harder function to fit with interpolation methods. Perhaps surprisingly, splats outperforms the Haar wavelet decomposition, which can exactly fit vertical discontinuities.

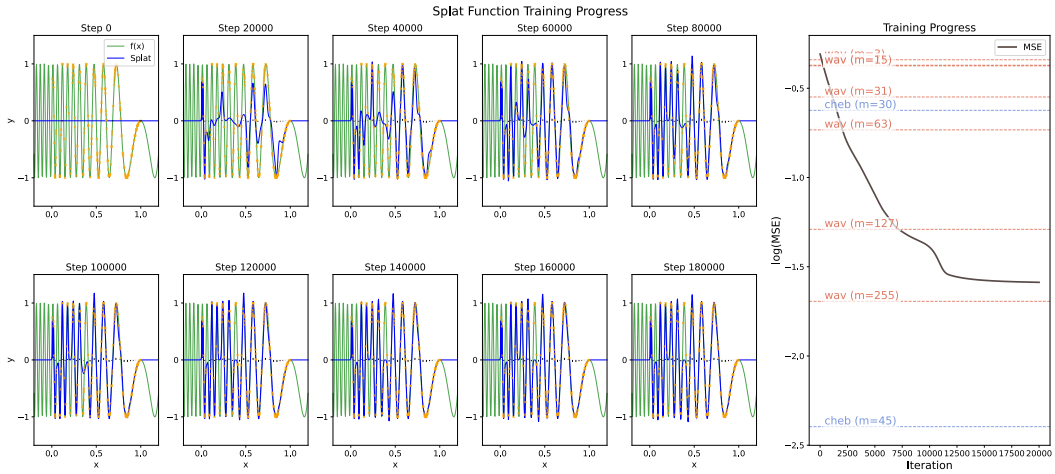

Figure 5: In the setting of 1, we test the effect of Chebyshev initialization, which is slightly less performant.