SPLAT REGRESSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a highly expressive class of function approximators called Splat Regression Models. Model outputs are mixtures of heterogeneous and anisotropic bump functions, termed splats, each weighted by an output vector. The power of splat modeling lies in its ability to locally adjust the scale and direction of each splat, achieving both high interpretability and accuracy. Fitting splat models reduces to optimization over the space of mixing measures, which can be implemented using Wasserstein-Fisher-Rao gradient flows. As a byproduct, we recover the popular Gaussian Splatting methodology as a special case, providing a unified theoretical framework for this state-of-the-art technique that clearly disambiguates the inverse problem, the model, and the optimization algorithm. Through numerical experiments, we demonstrate that the resulting models and algorithms constitute a flexible and promising approach for solving diverse approximation, estimation, and inverse problems involving low-dimensional data.

1 INTRODUCTION

In the recent history of the deep machine learning discipline, certain problem areas have enjoyed "inflection points" wherein the attainable performance and scale have rapidly improved by many orders of magnitude. These inflection points are often directly coupled to the discovery of the *right* model architecture for the problem at hand. To name a few examples: deep convolutional networks and ResNets for image classification (Krizhevsky et al., 2012; He et al., 2016); the U-Net architecture for image segmentation and generation (Ronneberger et al., 2015; Song & Ermon, 2019); and the transformer architecture for language modeling Vaswani et al. (2017). This motivates the search for new parsimonious architectures for different problem domains, and in this work, we target low-dimensional modeling problems such as those lying at the intersection of computational science and machine learning.

We introduce a new candidate that we call the 'Splat Regression Model.' In its simplest form, the model can be written as,

$$f(x) = \sum_{i=1}^{k} v_i \, \mathcal{N}(x; b_i, A_i A_i^T) \qquad v_i \in \mathbb{R}^p, \quad b_i \in \mathbb{R}^d, \quad A_i \in \mathbb{R}^{d \times d}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is the Gaussian density function. This rather simple architecture can be seen as a two-layer neural network with an atypical activation function, or alternatively, as a generalization of the classical Nadaraya-Watson estimator for nonparametric regression to use heterogeneous mixture weights. Toward the goal of developing a general theory for these heterogeneous mixture models, we abstract them into the form $f_\mu(x) := \mathbb{E}_{v \sim \mu_x}[v]$, where μ_x are conditionals of a probability distribution $(x, v) \sim \mu$. We develop a theoretical framework for optimizing these functions and understanding their expressivity. We further demonstrate that the splat regression model is an *extremely performant* architecture for low dimensional approximation, regression, and physics-informed fitting problems, typically outperforming Kolmogorov-Arnold Network (KAN) Liu et al. (2025) and Multilayer Perceptron (MLP) models by 10-100x even with far fewer parameters.

Another major success story of large-scale machine learning is in the computer graphics literature on solving *Novel View Synthesis*, where the goal is to learn a 3D scene from 2D snapshots annotated by camera position. A major breakthrough came with the introduction of Neural Radiance Fields Mildenhall et al. (2021a). A few years later, the so-called "3D Gaussian Splatting" methodology took over as a premier architecture for graphics and reconstruction problems. There are parallels

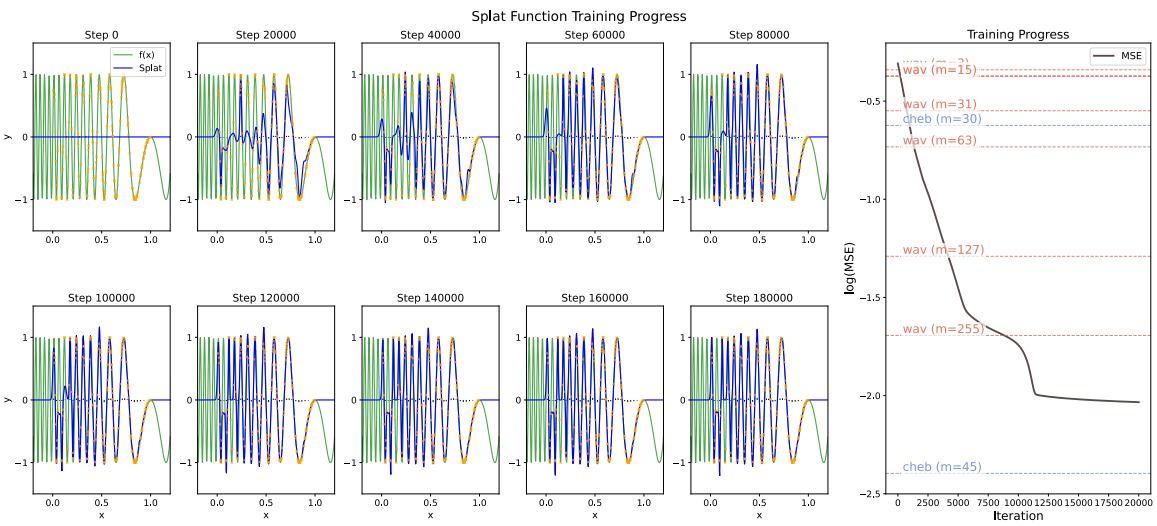

Figure 1: A representative approximation problem for the function $f^*(x) = \sin(20\pi x(2-x))$, $d = p = 1$. We fit a $k = 30$ splat model using least squares with $n = 200$ noiseless samples and we compare to the performance of Chebyshev polynomial interpolation and Haar wavelet approximation. By learning an 'adaptive grid' interpolation, the splat regression model drastically outperforms a Haar wavelet approximations, and competes with the gold-standard Chebyshev polynomial interpolation. *(Left)*. Training iterates of a $k = 30$ splat model (blue) as it fits f^* (green) by minimizing squared error with respect to $n = 200$ uniform samples (orange). *(Right)*. Validation MSE of the splat model over training. We also show the validation MSE of a Chebyshev approximation with $m = 30, 45$ gridpoints and of a Haar wavelet approximation at scales 2^l, $l = 1, \ldots, 8$.

between developments in Novel View Synthesis and recent progress in physics-informed learning, which has arguably not yet enjoyed its inflection point. An essential component in NeRF modeling is the use of sinusoidal positional encoding, which drastically improves the modeling capabilities across different 'scales' of variation in the target modeling problem, and which is paralleled by the use of positional encoding schemes in physics informed PDE-fitting (Tancik et al., 2020; Zeng et al., 2024), to moderate success. In both cases, rendering solutions by pointwise evaluation of an MLP across the spatial domain can be prohibitively slow. Recognizing that Novel View Synthesis is inherently a spatial inverse problem, **a major goal of our work is to replicate the successes of 3D Gaussian Splatting across a wide variety of physical modeling and inverse problems**. We intend for this work to provide an instruction manual for deploying and training splat regression models in these settings. Our main contributions are the following.

1. We introduce the Splat Regression Model. We prove some preliminary structural theorems about the it and we show that it is a universal approximator.
2. We develop principled optimization algorithms for gradient-based training, which are crucial to our empirical success. To do this, we leverage an interpretation of the splat model parameters as a hierarchical 'distribution over distributions,' and we invoke the theory of Wasserstein-Fisher-Rao gradient flows to compute gradient updates in parameter space.
3. We recover 3D Gaussian Splatting as an instance of splat regression modeling. In Example 2, we detail a clean formulation of it, where different aspects of the pipeline are split into clear, 'modular' parts.
4. We test the performance of splat modeling in a few representative modeling problems including low-dimensional regression (Figure 2) and physics-informed fitting.

2 Related Work

Novel View Synthesis. The Novel View Synthesis in computer graphics has a long history. Input data is typically sourced using Structure From Motion (SfM) methods (Ullman, 1979; Özyeşil et al.,

2017). In the idealized setting, scenes are rendered using the Radiative Transfer Equation (Chandrasekhar, 2013). Much work has gone into developing fast approximations to the RTE, such as the widely used α-blending technique introduced in (Porter & Duff, 1984; Carpenter, 1984). For 3D scene representation, Zwicker et al. (2002b) introduced the "elliptical weighted average" method, although these ideas were not applied at scale for about two decades (Kerbl et al., 2023a). Major interest in this problem was sparked with the release of Neural Radiance Fields Mildenhall et al. (2021b), an approach based on optimizing MLP parameters by differentiating through (an approximation of) the RTE.

Physics-informed Modeling. Following the introduction of PINNs (Raissi et al., 2019), much work has gone into developing ML methods for solving computational science problems. As documented by Krishnapriyan et al. (2021), PINN training can have surprising failure modes and requires careful tuning and hand-crafted architectures. More recently, adding positional encoding schemes Tancik et al. (2020); Huang et al. (2021) through either RBF or sinusoidal encoding was shown to improve performance. Alternatively, the Kolmogorov-Arnold Network architecture Liu et al. (2025); Rigas & Papachristou (2025) was introduced as a competitor, and some comparisons between KAN and RBF interpolation methods are explored in Li (2024). One way to view splat regression modeling is as learning an 'adaptive interpolation grid,' and our experimental results suggest that 'smart positional encoding is all you need' for low-dimensional modeling problems.

Mean-field Theory of Two-layer Networks. Wasserstein-Fisher-Rao gradient flows can be used to study the optimization of two-layer ReLU neural networks. This approach was pioneered by Chizat & Bach (2018), who prove a conditional global convergence result for two-layer network optimization. Further work in this direction includes (Mei et al., 2019a;b). The main difference between optimizing ReLU networks and Splat models is that our model has a fundamentally different geometry of its parameter space, since first-layer output neurons are represented as elements of a Bures-Wasserstein manifold. This is also a major difference between our work and Chewi et al. (2025b), which introduces a modified ReLU network architecture whose neurons can be 'superpositions' of ReLU functions.

3 Heterogeneous Mixtures and Splat Regression Models

We now state a few essential definitions and basic properties of splat regression models. Readers who are uninterested in the abstract formulation may skip to equation 1, which is the concrete object that appears in algorithms.

We write $\mathcal{L}^s(\mathbb{R}^d;\mathbb{R}^p)$ to denote the functions $f:\mathbb{R}^d\to\mathbb{R}^p$ for which $\int \|f(x)\|_s^s\,dx<\infty$, and $C_b^s(\mathbb{R}^d)$ to denote the real-valued bounded s-times continuously differentiable functions on $\mathbb{R}^d$. For a function $\phi:\mathbb{R}^d\to\mathbb{R}$, we write $\nabla\phi(x)$ its gradient, $\nabla^2\phi(x)$ its Hessian, and $\Delta\phi(x)=\sum_{i=1}^d \partial^2_{x_i}\phi(x)$ its Laplacian. For $\psi:\mathbb{R}^d\to\mathbb{R}^p$ we write $D\psi(x)$ its Jacobian and (when $p=d$) we write $\operatorname{div}\psi(x)=\sum_{i=1}^d \partial_{x_i}\phi^{(i)}(x)$ its divergence. We denote by $\mathcal{P}(\Omega)$ the space of probability distributions on Ω, by $\mathcal{P}_s(\Omega)$ the distributions with $s\geq 1$ finite moments, and by $\mathcal{P}_{ac}(\Omega)$ the distributions with a density. For $\mu\in\mathcal{P}(\Omega)$ we write $\mathbb{E}_{X\sim\mu}[f(x)]=\int f(x)\,\mu(dx)$ and, if μ has a density, we denote it as $\mu(x):\Omega\to\mathbb{R}$. Last, we denote by $T_{\#}\mu$ the pushforward of μ, that is, the distribution of $T(X)$ when $X\sim\mu$. We write $\mathsf{id}(x)=x$ the identity map, $\mathsf{id}_{\#}\mu=\mu$.

Our general philosophy is to investigate the functions $f\in\mathcal{L}^s(\mathbb{R}^p;\mathbb{R}^d)$ that can be represented by a probabilistic coupling $\mu\in\mathcal{P}(\mathbb{R}^p\times\mathbb{R}^d)$ which dictates, at each $(v,x)\in\mathbb{R}^p\times\mathbb{R}^d$, how much the vector v contributes to the value of $f(x)$. This makes sense whenever μ satisfies a "density in x, moment in v" assumption.

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 a measure $\nu\in\mathcal{P}(\mathbb{R}^p)$ (the 'mixture components') and a density $k(v,x)>0$ (the 'heterogeneous mixture weights') so that*

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

Having a density in x is required to be able to make sense of $f_\mu(x)$ pointwise. For each x, $k(\cdot,x)$ determines which $v\in\operatorname{supp}(\nu)$ (and in what proportions) contribute to the value of $f_\mu(x)$. We

propose to parametrize $k(\cdot, x)$ as a mixture of *shifted-and-scaled copies of a simple base distribution* ρ. We envision ρ as an isotropic 'bump' function and call each individual copy a *splat* in homage to the computer graphics literature where an instance of this model was first widely popularized. We call ρ the *mother splat* in analogy to the *mother wavelets* used in wavelet theory (Mallat, 2008).

Definition 1 (Splat Measures and Splat Models)**.** Let $\rho \in \mathcal{P}_{ac}(\mathbb{R}^d)$ have zero mean and identity covariance, and let $\rho_{A,b}$ be the distribution of $AX+b$ when $X \sim \rho$. We say that μ is a *splat measure* if it is a heterogeneous mixture such that for each $v \in \operatorname{supp} \nu$, there exists $w_v \in \mathcal{P}(\mathbb{R}^{d\times d} \times \mathbb{R}^d)$ so that

$$k(v, x) = \mathbb{E}_{(A,b)\sim w_v}[\rho_{A,b}(x)] \qquad \mu(dv, x) = \nu(dv)\, k(v, x)$$

and the associated splat model $f_\mu(x)$ is equal to,

$$f_\mu(x) = \mathbb{E}[v\rho_{A,b}(x)] \qquad v \sim \nu \quad A, b \sim w_v.$$

If the support of the joint distribution of the (v, A, b) has $k < \infty$ elements, we say μ is a *k-splat measure*, f_μ is a *k-splat model*, and we write

$$f_\mu(x) = \frac{1}{k}\sum_{i=1}^{k} v_i\, \nu(v_i)\, \rho(A_i^{-1}(x - b_i))\, |\det A_i^{-1}| \qquad \{(v_i, A_i, b_i)\}_{i=1}^k = \operatorname{supp}(v, A, b). \tag{1}$$

It may seem counterintuitive that $k(v, x)$ is defined through a v-dependent distribution w_v instead of through a more 'x-centric' expression. One benefit is that it simplifies the model's dependence on x allowing us to prove a few desirable properties of splat models as function approximators. Perhaps more importantly, it also leads to a natural interpretation of the parameters ν, $(w_v)_{v\in\operatorname{supp}\nu}$ as elements of a 'pseudo-Riemannian' manifold, allowing us to perform gradient-based training of k-splat models. We now discuss both of these.

3.1 Regularity and Universal Approximation

It is helpful to have some basic sufficient conditions for $f_\mu(x)$ to be continuous and to have continuous derivatives.

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*

$$\begin{aligned}\omega(\epsilon) &= \sup\left\{\left|\mathbb{E}_{\mu(\cdot,x)}[v] - \mathbb{E}_{\mu(\cdot,y)}[v]\right| : \|x - y\| \le \epsilon\right\} \\ &\le \sup\left\{W_1(\mu(\cdot, x), \mu(\cdot, y)) : \|x - y\| \le \epsilon\right\}.\end{aligned}$$

   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 $s \ge 0$ bounded derivatives. Then $f_\mu(x)$ also has s bounded derivatives.*

It is also helpful to understand the expressiveness of the class of splat models, and in particular, to understand what kinds of functions can be approximated by k-splat models for finite k. As we show in Appendix A, the finite splat model (equation 1) is already in the scope of Cybenko's early result on universal approximation of real-valued functions by two-layer neural networks.

Proposition 3 (Corollary of Cybenko (1989), Definition 1 and Theorem 1)**.** *Let Ω be a compact subset of $\mathbb{R}^d$ 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*

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

We also prove in Appendix A a quantitative bound on the number of splats required to approximate any Lipschitz function on a 'nice' domain using any 'nice' mother splat.

Theorem 3 (Informal – quantitative universal approximation) **.** For any bounded, 'nice' $\Omega \subseteq \mathbb{R}^d$, any bounded, Lipschitz $f : \Omega \to \mathbb{R}^p$, and any bounded, Lipschitz mother splat with finite first moment, there exists for each $\epsilon > 0$ a k-splat measure μ with $k \lesssim \epsilon^{-2(d+2)}$ and,

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

To put this rate into perspective, we prove in Appendix A the following lower bound on the error of uniform approximation over the family of 1-Lipschitz functions.

Proposition 4 (Uniform approximation lower bound)**.** *Set $\mathcal{F} = \{f : [0,1]^d \to \mathbb{R} : f$ 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.

This rate is tight: given $f \in \mathcal{F}$, one may create a splat model approximation using the piecewise constant approximation of f on sets of the form $\bigotimes_{i=1}^d [s_i\epsilon, (s_i+1)\epsilon)$ where each $(s_i)_{i=1\ldots d}$ has $1 \le s_i \le d$ and where $\epsilon = k^{-1/d}$. The piecewise constant approximation is representable by a splat model with mother splat $\rho = [-c, c)^d$ where $c > 0$ is so that ρ has identity covariance, and the number of pieces in the piecewise approximation is indeed $\epsilon^{-d} = k$. Taken together, Theorem 3 and Proposition 4 imply that *there exists a mother splat that achieves the minimax optimal rate $\epsilon \sim k^{-1/d}$ of approximation of Lipschitz functions*, and also, that *for any 'nice' mother splat, the approximation rate is at worst $\epsilon \sim k^{-1/2(d+2)}$*. However, much like the corresponding universal approximation rates for multilayer perceptrons, these worst case approximation bounds typically do not describe the number of parameters required to achieve a high quality fit for real data. We discuss further in Section 4.

3.2 Geometry of Splat Models

Ignoring some technical caveats, the space $\mathcal{P}(\mathbb{R}^d)$ of probability distributions can be viewed as a manifold by assigning it a distance metric such as $W_2(\mu, \nu)$, the 2-Wasserstein distance, or $H(\mu, \nu)$, the Hellinger distance. In both cases the distance induces a metric in the geometric sense, and a tangent space structure, which together can be used to construct gradient-descent-like algorithms over the space of probability distributions. In Wasserstein space, moving between points $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ is accomplished by applying "infinitesimal" transport maps and the geodesics follow the straight line interpolations of optimal transport maps, $\mu_t = (t\, T_{\mu_0 \to \mu_1} + (1-t)\,\mathrm{id})_{\#}\mu_0$. Under the geometry of the Hellinger metric, known as the Fisher-Rao or Information geometry Amari (1983), moving between points is accomplished by "mass teleportation," which means directly scaling the density up or down, pointwise over the domain, in a mass-preserving way. Its geodesics are $\mu_t(x) = (\alpha_t\sqrt{\mu_0}(x) + \beta_t\sqrt{\mu_1}(x))^2$, where α_t, β_t are spherical linear interpolation coordinates, as can be seen from the $\mathcal{L}^2$ form of $H^2(\mu_0, \mu_1) = \int (\sqrt{\mu_0}(x) - \sqrt{\mu_1}(x))^2\, dx$ for densities $\mu_0, \mu_1 \in \mathcal{P}_{ac}(\mathbb{R}^d)$. Finally, it is also possible to consider the geometry that arises when allowing movements by transport and teleportation at the same time, which is the *Wasserstein-Fisher-Rao geometry* (Chewi et al., 2025a).

The Wasserstein and Fisher-Rao geometric perspectives are widely used in the design and analysis of probabilistic algorithms, such as for dynamical sampling (Chewi et al., 2025a), variational inference (Lambert et al., 2022), barycentric interpolation (Gouic et al., 2019; Chewi et al., 2020), and lineage tracing (Schiebinger et al., 2019). We apply both to design a principled gradient-based training method for splat regression. As a byproduct, we recover the heuristic methods used to optimize Gaussian Splat models, which has a clean interpretation as regularized risk minimization via Wasserstein-Fisher-Rao gradient descent. We first explain the Wasserstein structure of the space of splats, then the Wasserstein-Fisher-Rao structure of the space of splat measures.

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

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

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.*

In the context of Definition 1, the parameters $\nu, (w_v)_{v\in\operatorname{supp}\nu}$ are simply a conditioning decomposition of a joint distribution belonging to $\mathcal{S}_{p,d} := \mathcal{P}(\mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d))$. We abuse notation and write a splat measure $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$ interchangeably as a measure in $\mathcal{S}_{p,d}$ (the abuse is, "$\mu(dv, x) = w_v(x)\nu(dv) = \mu(dv, dw)$," and now when we say 'splat measure' we mean the latter). The finite splat measures are the elements of $\mathcal{S}_{p,d}$ with discrete finite support.

Relative to particle methods, lifting to $\mathcal{S}_{p,d}$ allows one to implement 'smoothed particles' on $\mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$, thus providing a computationally tractable way to flow measures that are everywhere positive and absolutely continuous. We comment in Appendix B on the relationship between gradient flows in $\mathcal{S}_{p,d}$ and in $\mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$. This generalizes the concept of 'Gaussian particles' introduced in (Lambert et al., 2022), because particles are represented by any affine pushforward of an arbitrary density ρ. Such a particle representation was also hinted at in the computer graphics literature on 'volume splatting' (Zwicker et al., 2002a), and has also appeared in early work on hydrodynamics simulations (Gingold & Monaghan, 1977).

We endow $\mathcal{S}_{p,d}$ with the Wasserstein and/or Fisher-Rao metric, similar to the 'Wasserstein over Wasserstein' approach taken in (Lambert et al., 2022; Bonet et al., 2025). After stating our results, we give Examples 1, 2 that illustrate these objects concretely in two practical settings.

Definition 2 (First variation)**.** Let $\mathcal{F} : \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}^p) \to \mathbb{R}$ be a functional. Its first variation (when it exists) is the function $\delta\mathcal{F}[f](x)$ which satisfies for every $f, \xi \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}^p)$,

$$\partial_\epsilon \mathcal{F}(f + \epsilon\xi) \mid_{\epsilon=0} = \int_{\mathbb{R}^d} \langle \xi(x), \delta\mathcal{F}[f](x)\rangle \, dx.$$

It is convenient to express the Wasserstein gradient $\mathbb{W}_\mu \mathcal{F}(f_\mu) : \mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d) \to \mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d)$ in the global coordinate system $(v, A, b) \cong (v, \rho_{A,b})$. This coordinate system is convenient for writing the gradient $\mathbb{W}_\mu \mathcal{F}(f_\mu) = (\mathbb{W}_v, \mathbb{W}_A, \mathbb{W}_b)\mathcal{F}(f_\mu)$, so that the particle dynamics

$$\dot{v}_t = -\mathbb{W}_v \mathcal{F}(f_\mu)(v, A, b) \qquad \dot{A}_t = -\mathbb{W}_A \mathcal{F}(f_\mu)(v, A, b) \qquad \dot{b}_t = -\mathbb{W}_b \mathcal{F}(f_\mu)(v, A, b).$$

implement a Wasserstein gradient flow in $\mathcal{S}_{p,d}$. Our main theorem is the following.

Theorem 1 (Informal – Wasserstein-Fisher-Rao gradient of $\mu \mapsto \mathcal{F}(f_\mu)$)**.** *Let* $\mu \in \mathcal{S}_{p,d} := \mathcal{P}(\mathbb{R}^p \times \mathsf{BW}_\rho(\mathbb{R}^d))$ *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 by,*

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

and the Wasserstein gradient is given by,

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

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

The condition on ρ is required to apply integration by parts without a boundary term, but when this does not hold (such as when $\rho(x) = \mathbf{1}\{\|x\|_2 \le 1\}$) one can derive the appropriate corrections on a case-by-case basis. We conclude with a few concrete examples of $\mathcal{F}$ which are relevant to applications.

Example 1 (Empirical risk minimization)**.** Suppose $x_1, x_2, \ldots, x_n \sim \mathcal{U}(\Omega)$ are i.i.d. samples, where $\Omega = [0, 1]^d$ for simplicity, and set $y_i = f^*(x_i)$ with

$$\mathcal{F}(f) := \frac{1}{n}\sum_{i=1}^n L(f(x_i), y_i)$$

for a loss function $L(\hat{y}, y)$. To calculate the variation

$$\partial_{\epsilon=0}\mathcal{F}(f+\epsilon\xi) = \partial_{\epsilon=0}\left\{\frac{1}{n}\sum_{i=1}^{n} L(f(x_i)+\epsilon\xi(x_i), x_i)\right\} = \frac{1}{n}\sum_{i=1}\langle \xi(x_i), \nabla_{\hat{y}} L(f(x_i), x_i)\rangle$$

and so $\delta\mathcal{F}[f](x_i) = \frac{2}{n}\sum_{i=1}^{n}\nabla_{\hat{y}}L(f(x_i), x_i)$, which is otherwise undefined for $x \in \mathbb{R}^d$ outside the sample. We can use importance sampling to estimate the gradients,

$$\mathbb{E}_{X\sim\rho_{A,b}}\left[\delta\mathcal{F}[f](X)(\ldots)\right] \approx \frac{1}{n}\sum_{i=1}^{n}\rho_{A,b}(x_i)\delta\mathcal{F}[f](x_i)(\ldots).$$

This estimator is unbiased because $\{x_i\}_{i=1}^n \sim \mathcal{U}([0,1]^d)$. For samples drawn from an unknown distribution, one could instead estimate $\delta\mathcal{F}[f](x)$ by interpolating the available data, then approximating the average over $X \sim \rho_{A,b}$ using either Monte-Carlo or a quadrature rule.

Example 2 (Inverse problems and physics-informed training)**.** We can also take

$$\mathcal{F}(f) = \frac{1}{2}\|\mathcal{A}[f](x) - g(x)\|^2_{\mathcal{L}^2(\Omega)} \tag{2}$$

where $\mathcal{A}$ is a known integro-differential operator, g is a known forcing function, and we again take $\Omega = [0,1]^d$ for simplicity. To calculate the variation,

$$\begin{aligned}\partial_{\epsilon=0}\mathcal{F}(f+\epsilon\xi) &= \partial_{\epsilon=0}\left\{\frac{1}{2}\int \|\mathcal{A}[f+\epsilon\xi](x) - g(x)\|_2^2\right\}\\ &= \int \langle \mathcal{A}[f](x) - g(x), D_f\mathcal{A}[\xi](x)\rangle\, dx\\ &= \int \langle (D_f\mathcal{A})^*\mathcal{A}[f](x) - (D_f\mathcal{A})^*[g](x), \xi(x)\rangle\, dx.\end{aligned}$$

where $D_f\mathcal{A}$ is the linearization of $\mathcal{A}$ at f and $(D_f\mathcal{A})^*$ is its adjoint. So $\delta\mathcal{F}[f] = (D_f\mathcal{A})^*\mathcal{A}[f](x) - \mathcal{A}^*[g](x)$. Two illustrative instances of this setup are,

1. *A 'Physics-informed' splat solver for the Poisson equation.* Take $p = 1$, $\mathcal{A}[f_\mu] = \Delta f_\mu$. The Fisher-Rao gradient is,

$$\nabla^{\mathsf{FR}}_\mu\mathcal{F}(f_\mu)(v, A, b) = \int_\Omega v(\Delta f(x) - g(x))\,\Delta\rho_{A,b}(x)\,dx.$$

   In the same way, the coordinates of the Wasserstein gradient can be expressed as integrals with respect to $\Delta\rho(x)$ and $\nabla(\Delta\rho)(x)$. For simple ρ, these functions can be precomputed to accelerate gradient computations, and the integral approximated by Monte-Carlo.

2. *Splat regression modeling for Novel View Synthesis (NVS).* NVS can be viewed as an inverse problem whose forward operator $\mathcal{A}$ is the *Radiative Transfer Equation* (RTE) (Chandrasekhar, 2013), (Zwicker et al., 2002a, Equation (7)), whose unknown parameters are: the 'emission function,' $s : \mathbb{R}^3 \times S^2 \to \mathbb{R}$, and the 'extinction function' $\sigma : \mathbb{R}^3 \to \mathbb{R}$. The emission $s(x, v)$ is the amount of light radiating from $x \in \mathbb{R}^d$ outwards in direction v, and the extinction $\sigma(x)$ represents the degree to which point x 'occludes' points behind it. Following (Kerbl et al., 2023a), we parametrize splat models $\sigma(x) = g_\nu(x)$ and we parameterize $s(x, v)$ as,

$$s(x, \cdot) = \sum_{i=1}^{p} f_\mu^{(i)}(x)\phi_i(v)$$

   where $\{\phi_i\}_{i=1}^p$ are spherical harmonic basis functions and $f_\mu : \mathbb{R}^3 \to \mathbb{R}^p$ is another splat model ($p \approx 20$). Given these two fields, rendering the scene involves evaluating the *Radiative Transfer Equation*.

$$\mathcal{A}[s, \sigma](x, v) = \int_0^\infty s(x+tv, v)\sigma(x+tv)\exp\left(-\int_0^t \sigma(x+sv)\,ds\right)\,dt. \tag{3}$$

   The RTE is typically evaluated using a discrete approximation called α-blending (Zwicker et al., 2002b; Kerbl et al., 2023a) and splat parameters are trained via SGD with particle birth-death dynamics to minimize residuals (equation 2).

Finally, we remark that the performance of Gaussian splatting in large-scale Novel View Synthesis depends heavily on the use of many training heuristics, such as initialization of splat locations Kerbl et al. (2023b), selective noising of splat locations during training Kheradmand et al. (2024), and on 'pruning strategies' to move spurious splats either by deleting them or teleporting them elsewhere in space (Kerbl et al., 2023b; Hanson et al., 2025). By providing a principled optimization perspective on fitting splat models, our results pave the way to interpret these heuristics through the lens of *regularized risk minimization*. For instance, in a sampling context, the stochastic *Unadjusted Langevin Algorithm* can be interpreted as Wasserstein gradient descent of the functional $\mathcal{F}(\mu) = D_{\mathrm{KL}}(\mu \mid \mu_{\text{target}})$ divergence, which enjoys strong convexity whenever μ_{target} is strongly log concave. We anticipate that the selective noising heuristic introduced by Kheradmand et al. (2024) can be interpreted as adding a convex entropic regularizer to the objective (2), improving convergence. Similarly, particle birth-death dynamics are a well-known discretization algorithm for Fisher-Rao gradient flows Lu et al. (2023). With an appropriate discretization, Theorem 1 thus prescribes a pruning criterion that is guaranteed (in the continuous-time limit) to decrease the loss. While it is outside the scope of the present manuscript, principled regularization methods for splats is a very promising direction for future work.

4 Experiments

In this section, we demonstrate the power of the splat regression modeling by directly comparing it on approximation, regression, and physics-informed modeling problems.

4.1 Multiscale Approximation

First, we show in Figure 1 a representative 1D example of training a splat regression model, which we compare to two standard function approximation algorithms: *chebyshev polynomial interpolation* and *wavelet basis projection*.

In the left subpanel, we show snapshots of splat training arranged from right to left, top to bottom. We train a $k = 30$ splat model initialized with $v_i = 0, b_i = (i-1)/k$, and $A_i = 1/2k$ for $i = 1 \dots k$. Training is performed by least squares fitting via Wasserstein gradient descent with learning rate 10^{-4} and no momentum. The loss function is $\mathcal{L}(f_\mu) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$ for $n = 200$ noiseless samples $x_i \sim \mathcal{U}([0,1])$. In the right subpanel, we show the *validation log-MSE* of the splat model. Anecdotally, the exponentially fast convergence which appears in Figure 1 (right) is robust to different initializations and functions. We show in Appendix C, Figure 5 a version of this experiment where the $\{b_i\}_{i=1}^k$ are initialized as a k-point Chebyshev grid as well as a version where data is sampled from a discontinuous sawtooth function.

We select this example as a simple one dimensional picture of how splat models are well suited to fit 'multi-scale' features of the observed data. For comparison, horizontal lines indicate the validation MSE of the Chebyshev and wavelet approximations. The splat model with $k = 30$ splats significantly outperforms the Chebyshev interpolation with $m = 30$ interpolation nodes. Setting $m = 45$ in order to control for the total number of model parameters ($(3 \times k = 90)$ for splat vs. $(2 \times m = 90)$ for Chebyshev), we see that the splat model achieves slightly worse approximation error. Relative to wavelet approximation, the splat model outperforms a Haar wavelet projection with $m = 255$ parameters (corresponding to level $l = 8$ of the basis hierarchy). We find these results extremely encouraging, particularly given that very simple optimization and initialization schemes are enough to successfully train the splat model.

4.2 Regression with KAN, MLP, and SRM

We compare the performance of splat regression modeling to Kolmogorov-Arnold networks Liu et al. (2025) and to fully connected MLP architectures on a 2D regression task. We sample $x_i \sim \mathcal{U}([0,1]^2)$ i.i.d. and set $y_i = f(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 0.01)$ and,

$$f(x, y) = \sin(3\pi\sqrt{x})\cos(3\pi y).$$

Each model is trained using 3000 iterations using Adam with learning rate 10^{-4}.

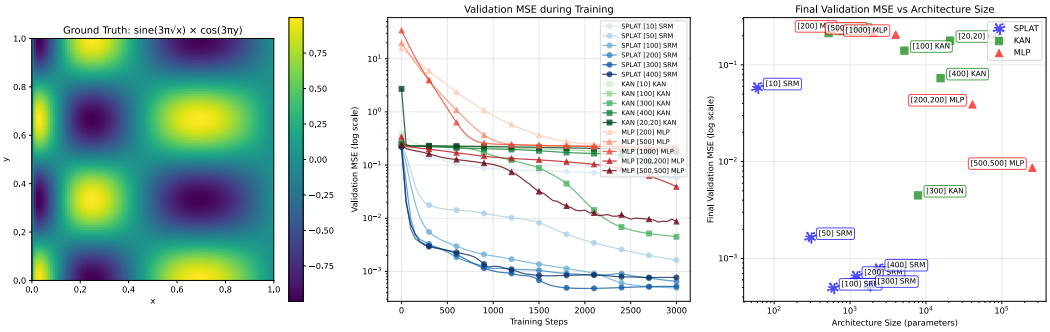

Figure 2: We compare splat regression, Kolmogorov-Arnold Networks Liu et al. (2025), and fully connected Multi-layer Perceptron in a noisy regression task. We observe that splat models achieve order of magnitude lower fitting error while using a small fraction of the parameters of MLP and KAN networks.

We attribute the improved performance of splat models to their spatially localized nature, which can be viewed as a learned positional encoding scheme. However, due to their expressivity, splat models may be more susceptible to overfitting, requiring regularization to achieve good fits.

4.3 Physics-informed Modeling

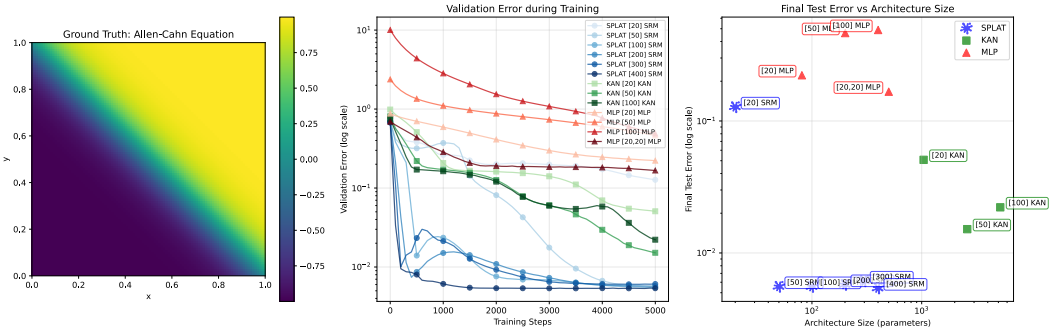

Figure 3: We compare splat regression, Kolmogorov-Arnold Networks Liu et al. (2025), and fully connected Multi-layer Perceptron in a physics informed regression task. Models are fit to solve the Allen-Cahn equation on $[0,1]^2$. *(Left)*. True solution under this parameter regime. *(Middle)*. Validation error for each model class as a function of the number of training iterations. *(Right)*. Validation error relative to total number of model parameters. Among the test pool, a $k = 50$ splat model outperforms all KAN and MLP architectures by an order of magnitude while using significantly fewer parameters.

We further test the splat regression model in a two dimensional physics-informed learning problem where the goal is to estimate the solution of the *Allen Cahn* equation,

$$\epsilon^2 \Delta u(x) + u(x) - u^3(x) = 0 \qquad x \in [0,1]^d.$$

This equation is a well known example of a PDE whose solutions can have 'boundary interfaces' that converge to discontinuities as $\epsilon \to 0$. We take $\epsilon = 0.1$ and train each model to minimize a weighted sum of the loss,

$$\mathcal{L}(u_\theta) = \frac{1}{n_{\text{int}}} \sum_{i=1}^{n_{\text{int}}} \|\mathcal{D}u_\theta(x_i) - f(x_i)\|_2^2 + \frac{1}{n_{\text{bdy}}} \sum_{j=1}^{n_{\text{bdy}}} \|u_\theta(z_j) - u^*(z_j)\|_2^2$$

where $x_i \sim \mathcal{U}([0,1]^d)$ and where $z_j \sim \mathcal{U}(\partial[0,1]^d)$ for $i = 1, \ldots, 10^5$ and $j = 1, \ldots, 5 \cdot 10^4$. both i.i.d. We observe that the splat equation converges very quickly and to a highly accurate solution, even when limited to use orders of magnitude fewer parameters the KAN and MLP models.

5 CONCLUSION

We have shown that splat models are highly effective in solving a variety of representative low dimensional modeling problems. We believe that these results are extremely promising from a practical perspective. Furthermore, we have drawn a novel connection between 3D Gaussian Splatting and Wasserstein-Fisher-Rao gradient flows, which we hope will lead to many symbiotic interactions between the optimal transport, computer graphics, and statistics communities.

There is much future work to be done. First, as we discuss in Section 3.2, splat models are highly expressive and therefore susceptible to overfitting, warranting a larger scale computational study beyond the scope of our work. The splatting community has developed many heuristics for regularizing and pruning splat models, which we expect will be instrumental to large-scale splat regression modeling. As a second direction, we remark that the proposed model can be interpreted as a new *neural network layer*, opening the door to compositions of splat models ('deep splat networks') and/or integrations into existing deep architectures. We look forward to investigating these directions in followup work.

REPRODUCIBILITY STATEMENT

The Jax/Python source code used to replicate our experiments will be made publicly available in a open source repository upon publication. The repository will include detailed instructions for setting up the environment, training the models, and evaluating the results. We will also release all the information required to recreate the datasets and baseline comparisons used in this work, including random seeds used in experiments.

ETHICS STATEMENT

The research presented in this paper introduces a novel trainable architecture for scientific modeling and general-purpose regression tasks, with the primary goal of advancing machine learning methodology. We use standard benchmark datasets and do not engage with sensitive personal data or high-risk application areas. We have considered the societal and environmental impacts and conclude that our work adheres to established ethical guidelines, presenting no immediate societal risks.

REFERENCES

Shun-Ichi Amari. A foundation of information geometry. *Electronics and Communications in Japan (Part I: Communications)*, 66(6):1–10, 1983. doi: https://doi.org/10.1002/ecja.4400660602. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecja.4400660602. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecja.4400660602.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 4-431-55977-9.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2. ed edition, 2008. ISBN 978-3-7643-8722-8.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, January 2000. doi: 10.1007/s002110050002. Publisher: Springer Science and Business Media LLC.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. ISSN 0723-0869. doi: https://doi.org/10.1016/j.exmath.2018.01.002. URL https://www.sciencedirect.com/science/article/pii/S0723086918300021.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing Datasets with Wasserstein over Wasserstein Gradient Flows. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=I1OHPb4zWo.

Loren Carpenter. The A-buffer, an antialiased hidden surface method. *ACM SIGGRAPH Computer Graphics*, 18(3):103–108, 1984. Publisher: ACM New York, NY, USA.

S. Chandrasekhar. *Radiative Transfer*. Dover Books on Physics. Dover Publications, 2013. ISBN 978-0-486-31845-5. URL https://books.google.com/books?id=1YHCAgAAQBAJ.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1276–1304. PMLR, July 2020. URL https://proceedings.mlr.press/v125/chewi20a.html.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, 2025a.

Sinho Chewi, Philippe Rigollet, and Yuling Yan. Gaussian mixture layers for neural networks. *arXiv preprint 2508.04883*, 2025b.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2nd edition, 2010. ISBN 978-0-8218-4974-3.

R. A. Gingold and J. J. Monaghan. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly Notices of the Royal Astronomical Society*, 181(3):375–389, December 1977. ISSN 0035-8711. doi: 10.1093/mnras/181.3.375. URL https://doi.org/10.1093/mnras/181.3.375. _eprint: https://academic.oup.com/mnras/article-pdf/181/3/375/3104055/mnras181-0375.pdf.

Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 2019. URL https://api.semanticscholar.org/CorpusID:199405372.

Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. PUP 3D-GS: Principled Uncertainty Pruning for 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 5949–5958, June 2025. URL https://pup3dgs.github.io/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xinquan Huang, Tariq Alkhalifah, and Chao Song. A modified physics-informed neural network with positional encoding. In *First International Meeting for Applied Geoscience & Energy Expanded Abstracts*, pp. 2480–2484. 2021. doi: 10.1190/segam2021-3584127.1. URL https://library.seg.org/doi/abs/10.1190/segam2021-3584127.1. _eprint: https://library.seg.org/doi/pdf/10.1190/segam2021-3584127.1.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998. doi: 10.1137/s0036141096303359. Publisher: Society for Industrial & Applied Mathematics (SIAM).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023a. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), July 2023b. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25*, pp. 1097–1105, 2012.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8. event-place: New Orleans, LA, USA.

Ziyao Li. Kolmogorov-Arnold Networks are Radial Basis Function Networks, 2024. URL https://arxiv.org/abs/2405.06721. _eprint: 2405.06721.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–Arnold Networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth–death dynamics for sampling: global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731, September 2023. doi: 10.1088/1361-6544/acf988. URL https://dx.doi.org/10.1088/1361-6544/acf988. Publisher: IOP Publishing.

Stphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., USA, 3rd edition, 2008. ISBN 0-12-374370-2.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT 2019)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2388–2464. PMLR, 2019a. URL https://proceedings.mlr.press/v99/mei19a.html.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT 2019)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2388–2464. PMLR, 2019b. URL https://proceedings.mlr.press/v99/mei19a.html.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021a. ISSN 0001-0782. doi: 10.1145/3503250.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021b. ISSN 0001-0782. doi: 10.1145/3503250. URL https://doi.org/10.1145/3503250. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Klas Modin. Geometry of Matrix Decompositions Seen Through Optimal Transport and Information Geometry. *The Journal of Geometric Mechanics*, 9:335–390, 2016. URL https://api.semanticscholar.org/CorpusID:119636228.

Thomas Porter and Tom Duff. Compositing digital images. *ACM SIGGRAPH Computer Graphics*, 18(3):253–259, 1984. Publisher: ACM New York, NY, USA.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. Publisher: Elsevier.

Spyros Rigas and Michalis Papachristou. jaxKAN: A unified JAX framework for Kolmogorov-Arnold Networks. *Journal of Open Source Software*, 10(108):7830, 2025. doi: 10.21105/joss.07830.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. 2015. URL http://www.math.toronto.edu/mccann/assignments/477/Santambrogio15.pdf.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176, 2019. URL https://api.semanticscholar.org/CorpusID:72333759.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32*, 2019.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6. event-place: Vancouver, BC, Canada.

S. Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. doi: 10.1098/rspb.1979.0006.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems 30*, 2017.

Cédric Villani. *Topics in Optimal Transport*. American Mathematical Society, Providence, RI, 2003.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-49802-9. doi: 10.1017/9781108627771. URL https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E.

Chengxi Zeng, Tilo Burghardt, and Alberto M. Gambaruto. RBF-PINN: Non-Fourier Positional Embedding in Physics-Informed Neural Networks. *ArXiv*, abs/2402.08367, 2024.

M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002a.

M. Zwicker, H. Pfister, J. van Baar, and M. Gross. EWA Splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002b. Publisher: IEEE Section: 223.

Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, 2017. doi: 10.1017/S096249291700006X.