

---

# GeONet: a neural operator for learning the Wasserstein geodesic (Supplementary Material)

---

Andrew Gracyk<sup>1</sup>

Xiaohui Chen<sup>2</sup>

<sup>1</sup>Department of Statistics, University of Illinois at Urbana-Champaign

<sup>2</sup>Department of Mathematics, University of Southern California

## A TRAINING ALGORITHM

## B DERIVATION OF PRIMAL-DUAL OPTIMALITY CONDITIONS FOR DYNAMICAL OT PROBLEM

The primal-dual analysis is a standard technique in the optimization literature such as in analyzing certain semidefinite programs [Chen and Yang, 2021]. Recall the Benamou-Brenier fluid dynamics formulation of the static optimal transport problem

$$\min_{(\mu, \mathbf{v})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|\mathbf{v}(x, t)\|_2^2 \mu(x, t) \, dx \, dt \quad (20)$$

$$\text{subject to } \partial_t \mu + \operatorname{div}(\mu \mathbf{v}) = 0, \quad (21)$$

$$\mu(\cdot, 0) = \mu_0, \quad \mu(\cdot, 1) = \mu_1. \quad (22)$$

Here, equation (21) is referred to as the *CE* (CE), preserving the unit mass of the density flow  $\mu_t = \mu(\cdot, t)$ . We write the Lagrangian function for any flow  $(\mu_t)_{t \in [0,1]}$  initializing from  $\mu_0$  and terminating at  $\mu_1$  as

$$L(\mu, \mathbf{v}, u) = \int_0^1 \int_{\mathbb{R}^d} \left[ \frac{1}{2} \|\mathbf{v}\|_2^2 \mu + (\partial_t \mu + \operatorname{div}(\mu \mathbf{v})) u \right] \, dx \, dt, \quad (23)$$

where  $u := u(x, t)$  is the dual variable for (CE). To find the optimal solution  $\mu^*$  for the minimum kinetic energy (20), we study the saddle point optimization problem

$$\min_{(\mu, \mathbf{v}) \in (\text{CE})} \max_u L(\mu, \mathbf{v}, u), \quad (24)$$

where the minimization over  $(\mu, \mathbf{v})$  runs over all flows satisfying (CE) such that  $\mu(\cdot, 0) = \mu_0$  and  $\mu(\cdot, 1) = \mu_1$ . Note that if  $\mu \notin (\text{CE})$ , then by scaling with arbitrarily large constant, we see that

$$\max_u \int_0^1 \int_{\mathbb{R}^d} (\partial_t \mu + \operatorname{div}(\mu \mathbf{v})) u \, dx \, dt = +\infty. \quad (25)$$

Thus,

$$\min_{(\mu, \mathbf{v}) \in (\text{CE})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|\mathbf{v}\|_2^2 \mu \, dx \, dt = \min_{(\mu, \mathbf{v})} \max_u L(\mu, \mathbf{v}, u) \quad (26)$$

$$\geq \max_u \min_{(\mu, \mathbf{v})} L(\mu, \mathbf{v}, u), \quad (27)$$

where the minimization over  $(\mu, \mathbf{v})$  is unconstrained. Using integration-by-parts and suitable decay for vanishing boundary as  $\|x\|_2 \rightarrow \infty$ , we have

$$L(\mu, \mathbf{v}, u) = \int_0^1 \int_{\mathbb{R}^d} \left[ \frac{1}{2} \|\mathbf{v}\|_2^2 \mu - \mu \partial_t u - \langle \mathbf{v}, \nabla u \rangle \mu \right] dx dt + \int_{\mathbb{R}^d} [\mu(\cdot, 1)u(\cdot, 1) - \mu(\cdot, 0)u(\cdot, 0)] dx.$$

Now, we fix  $\mu$  and  $u$ , and minimize  $L(\mu, \mathbf{v}, u)$  over  $\mathbf{v}$ . The optimal velocity vector is  $\mathbf{v}^* = \nabla u$ , and we have

$$\max_u \min_{\mu} L(\mu, \mathbf{v}^*, u) = \int_0^1 \int_{\mathbb{R}^d} \left[ - \left( \frac{1}{2} \|\nabla u\|_2^2 + \partial_t u \right) \mu \right] dx dt + \int_{\mathbb{R}^d} [u(\cdot, 1)\mu_1 - u(\cdot, 0)\mu_0] dx, \quad (28)$$

for any flow  $\mu_t$  satisfying the boundary conditions  $\mu(\cdot, 0) = \mu_0$  and  $\mu(\cdot, 1) = \mu_1$ . If  $\frac{1}{2} \|\nabla u\|_2^2 + \partial_t u \neq 0$ , then by the same scaling argument above, we have

$$\min_{\mu} \int_0^1 \int_{\mathbb{R}^d} \left[ - \left( \frac{1}{2} \|\nabla u\|_2^2 + \partial_t u \right) \mu \right] dx dt = -\infty \quad (29)$$

because  $\mu$  is unconstrained (except for the boundary conditions). Then we deduce that

$$\min_{(\mu, \mathbf{v}) \in (\text{CE})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|\mathbf{v}\|_2^2 \mu \geq \max_{u \in (\text{HJ})} \left\{ \int_{\mathbb{R}^d} u(\cdot, 1)\mu_1 - \int_{\mathbb{R}^d} u(\cdot, 0)\mu_0 \right\}, \quad (30)$$

where  $u \in (\text{HJ})$  means that  $u$  solves the *HJ equation* (HJ)

$$\partial_t u + \frac{1}{2} \|\nabla u\|_2^2 = 0. \quad (31)$$

From (30), we see that the duality gap is non-negative, and it is equal to zero if and only if  $(\mu^*, u^*)$  solves the following system of PDEs

$$\begin{cases} \partial_t \mu + \operatorname{div}(\mu \nabla u) = 0, & \partial_t u + \frac{1}{2} \|\nabla u\|_2^2 = 0, \\ \mu(\cdot, 0) = \mu_0, & \mu(\cdot, 1) = \mu_1. \end{cases} \quad (32)$$

PDEs in (32) are referred to as the Karush–Kuhn–Tucker (KKT) conditions for the Wasserstein geodesic problem.

## C METRIC GEOMETRY STRUCTURE OF THE WASSERSTEIN SPACE AND GEODESIC

In this section, we review some basic facts on the metric geometry properties of the Wasserstein space and geodesic. We first discuss the general metric space  $(X, d)$ , and then specialize to the Wasserstein (metric) space  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$  for  $p \geq 1$ . Furthermore, we connect to the fluid dynamic formulation of optimal transport. Most of the materials are based on the reference books [Burage et al., 2001, Ambrosio et al., 2008, Santambrogio, 2015].

### C.1 GENERAL METRIC SPACE

*Definition C.1* (Absolutely continuous curve). Let  $(X, d)$  be a metric space. A curve  $\omega : [0, 1] \rightarrow X$  is *absolutely continuous* if there is a function  $g \in L^1([0, 1])$  such that for all  $t_0 < t_1$ , we have

$$d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(\tau) d\tau. \quad (33)$$

Such curves are denoted by  $\text{AC}(X)$ .

*Definition C.2* (Metric derivative). If  $\omega : [0, 1] \rightarrow X$  is a curve in a metric space  $(X, d)$ , the *metric derivative* of  $\omega$  at time  $t$  is defined as

$$|\omega'| (t) := \lim_{h \rightarrow 0} \frac{d(\omega(t+h), \omega(t))}{|h|}, \quad (34)$$

if the limit exists.

The following theorem generalizes the classical Rademacher theorem from a Euclidean space into any metric space in terms of the metric derivative.

*Theorem C.3 (Rademacher).* If  $\omega : [0, 1] \rightarrow X$  is Lipschitz continuous, then the metric derivative  $|\omega'| (t)$  exists for almost every  $t \in [0, 1]$ . In addition, for any  $0 \leq t < s \leq 1$ , we have

$$d(\omega(t), \omega(s)) \leq \int_t^s |\omega'|(\tau) d\tau. \quad (35)$$

Theorem C.3 tells us that absolutely continuous curve  $\omega$  has a metric derivative well-defined almost everywhere, and the “length” of the curve  $\omega$  is bounded by the integral of the metric derivative. Thus, a natural definition of the length of a curve in a general metric space is to take the best approximation over all possible meshes.

*Definition C.4 (Curve length).* For a curve  $\omega : [0, 1] \rightarrow X$ , we define its *length* as

$$\text{Length}(\omega) := \sup \left\{ \sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1})) : n \geq 1, 0 = t_0 < t_1 < \dots < t_n = 1 \right\}. \quad (36)$$

Note that if  $\omega \in \text{AC}(X)$ , then

$$d(\omega(t_k), \omega(t_{k+1})) \leq \int_{t_k}^{t_{k+1}} g(\tau) d\tau \quad (37)$$

so that

$$\text{Length}(\omega) \leq \int_0^1 g(\tau) d\tau < \infty, \quad (38)$$

i.e., the curve  $\omega$  is of bounded variation.

*Lemma C.5.* If  $\omega \in \text{AC}(X)$ , then

$$\text{Length}(\omega) = \int_0^1 |\omega'|(\tau) d\tau. \quad (39)$$

*Definition C.6 (Length space and geodesic space).* Let  $\omega : [0, 1] \rightarrow X$  be a curve in  $(X, d)$ .

1. The space  $(X, d)$  is a *length space* if

$$d(x, y) = \inf \{ \text{Length}(\omega) : \omega(0) = x, \omega(1) = y, \omega \in \text{AC}(X) \}. \quad (40)$$

2. The space  $(X, d)$  is a *geodesic space* if

$$d(x, y) = \min \{ \text{Length}(\omega) : \omega(0) = x, \omega(1) = y, \omega \in \text{AC}(X) \}. \quad (41)$$

*Definition C.7 (Geodesic).* Let  $(X, d)$  be a length space.

1. A curve  $\omega : [0, 1] \rightarrow X$  is said to be a *constant-speed geodesic* between  $\omega(0)$  and  $\omega(1)$  if

$$d(\omega(t), \omega(s)) = |t - s| \cdot d(\omega(0), \omega(1)), \quad (42)$$

for any  $t, s \in [0, 1]$ .

2. If  $(X, d)$  is further a geodesic space, a curve  $\omega : [0, 1] \rightarrow X$  is said to be a *geodesic* between  $x_0 \in X$  and  $x_1 \in X$  if it minimizes the length among all possible curves such that  $\omega(0) = x_0$  and  $\omega(1) = x_1$ .

Note that in a geodesic space  $(X, d)$ , a constant-speed geodesic is indeed a geodesic. In addition, we have the following equivalent characterization of the geodesic in a geodesic space.

*Lemma C.8.* Let  $(X, d)$  be a geodesic space,  $p > 1$ , and  $\omega : [0, 1] \rightarrow X$  a curve connecting  $x_0$  and  $x_1$ . Then the following statements are equivalent.

1.  $\omega$  is a constant-speed geodesic.
2.  $\omega \in \text{AC}(X)$  such that for almost every  $t \in [0, 1]$ , we have

$$|\omega'| (t) = d(\omega(0), \omega(1)). \quad (43)$$

3.  $\omega$  solves

$$\min \left\{ \int_0^1 |\tilde{\omega}'|^p dt : \tilde{\omega}(0) = x_0, \tilde{\omega}(1) = x_1 \right\}. \quad (44)$$

## C.2 WASSERSTEIN SPACE

Since the Wasserstein space  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$  for  $p \geq 1$  is a metric space, the following definition specializes Definition C.2 to the Wasserstein metric derivative.

**Definition C.9** (Wasserstein metric derivative). Let  $\{\mu_t\}_{t \in [0,1]}$  be an absolutely continuous curve in the Wasserstein (metric) space  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ . Then the *metric derivative* at time  $t$  of the curve  $t \mapsto \mu_t$  with respect to  $W_p$  is defined as

$$|\mu'|_p(t) := \lim_{h \rightarrow 0} \frac{W_p(\mu_{t+h}, \mu_t)}{|h|}. \quad (45)$$

For  $p = 2$ , we write  $|\mu'|_p(t) := |\mu'|_2(t)$ .

In the rest of this section, we consider probability measures  $\mu_t$  that are absolutely continuous with respect to the Lebesgue measure on  $\mathbb{R}^d$  and we use  $\mu_t$  to denote the probability measure, as well as its density, when the context is clear.

**Theorem C.10.** Let  $p > 1$  and assume  $\Omega \in \mathbb{R}^d$  is compact.

**Part 1.** If  $\{\mu_t\}_{t \in [0,1]}$  is an absolutely continuous curve in  $W_p(\Omega)$ , then for almost every  $t \in [0, 1]$ , there is a velocity vector field  $\mathbf{v}_t \in L^p(\mu_t)$  such that

1.  $\mu_t$  is a weak solution of the CE  $\partial_t \mu_t + \operatorname{div}(\mu_t \mathbf{v}_t) = 0$  in the sense of distributions (cf. the definition in (51) below);
2. for almost every  $t \in [0, 1]$ , we have

$$\|\mathbf{v}_t\|_{L^p(\mu_t)} \leq |\mu'|_p(t), \quad (46)$$

where  $\|\mathbf{v}_t\|_{L^p(\mu_t)}^p = \int_{\Omega} \|\mathbf{v}_t\|_2^p d\mu_t$ .

**Part 2.** Conversely, if  $\{\mu_t\}_{t \in [0,1]}$  are probability measures in  $\mathcal{P}_p(\Omega)$ , and for each  $t \in [0, 1]$  we suppose  $\mathbf{v}_t \in L^p(\mu_t)$  and  $\int_0^1 \|\mathbf{v}_t\|_{L^p(\mu_t)} dt < \infty$  such that  $(\mu_t, \mathbf{v}_t)$  solves the CE, then we have

1.  $\{\mu_t\}_{t \in [0,1]}$  is an absolutely continuous curve in  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ ;
2. for almost every  $t \in [0, 1]$ ,

$$|\mu'|_p(t) \leq \|\mathbf{v}_t\|_{L^p(\mu_t)}. \quad (47)$$

As an immediate corollary, we have the following dynamical representation of the Wasserstein metric derivative.

**Corollary C.11.** If  $\{\mu_t\}_{t \in [0,1]}$  is an absolutely continuous curve in  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ , then the velocity vector field  $\mathbf{v}_t$  given in Part 1 of Theorem C.10 must satisfy

$$\|\mathbf{v}_t\|_{L^p(\mu_t)} = |\mu'|_p(t). \quad (48)$$

Corollary C.11 suggests that  $\mathbf{v}_t$  can be viewed as the *tangent vector field* of the curve  $\{\mu_t\}_{t \in [0,1]}$  at time point  $t$ . Moreover, Corollary C.11 suggests the following (Euclidean) gradient flow for tracking particles in  $\mathbb{R}^d$ : let  $y(t) := y_x(t)$  be the trajectory starting from  $x \in \mathbb{R}^d$  (i.e.,  $y(0) = x$ ) that evolves according the ordinary differential equation (ODE)

$$\frac{d}{dt} y(t) = \mathbf{v}_t(y(t)). \quad (49)$$

The dynamical system (49) defines a flow  $Y_t : \Omega \rightarrow \Omega$  of vector field  $\mathbf{v}_t$  on  $\Omega$  via

$$Y_t(x) = y(t). \quad (50)$$

Then, it is straightforward to check that the pushforward measure flow  $\mu_t := (Y_t)_\# \mu_0$  and the chosen velocity vector field  $\mathbf{v}_t$  in the ODE (49) is a weak solution of the CE  $\partial_t \mu_t + \operatorname{div}(\mu_t \mathbf{v}_t) = 0$  in the sense that

$$\frac{d}{dt} \int_{\Omega} \phi dt = \int_{\Omega} \langle \nabla \phi, \mathbf{v}_t \rangle d\mu_t, \quad (51)$$

for any  $\mathcal{C}^1$  function  $\phi : \Omega \rightarrow \mathbb{R}$  with compact support.

*Theorem C.12 (Constant-speed Wasserstein geodesic).* Let  $\Omega \in \mathbb{R}^d$  be a convex subset and  $\mu, \nu \in \mathcal{P}_p(\Omega)$  for some  $p > 1$ . Let  $\gamma$  be an optimal transport plan under the cost function  $\|x - y\|_p^p$ . Define

$$\begin{aligned}\pi_t &: \Omega \times \Omega \rightarrow \Omega, \\ \pi_t(x, y) &= (1 - t)x + ty,\end{aligned}$$

as the linear interpolation between  $x$  and  $y$  in  $\Omega$ . Then, the curve  $\mu_t = (\pi_t)_\# \gamma$  is a constant-speed geodesic in  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$  connecting  $\mu_0 = \mu$  and  $\mu_1 = \nu$ .

If  $\mu$  has a density with respect to the Lebesgue measure on  $\mathbb{R}^d$ , then there is an optimal transport map  $T$  from  $\mu$  to  $\nu$  [Brenier, 1991]. According to Theorem C.12, we obtain *McCann's interpolation* [McCann, 1997] in the Wasserstein space as

$$\mu_t = [(1 - t)\text{id} + tT]_\# \mu, \quad (52)$$

which is a constant-speed geodesic in  $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ .  $\text{id}$  is the identity function in  $\mathbb{R}^d$ .

To sum up, we collect the following facts about the geodesic structure and dynamical formulation of the OT problem. Let  $p > 1$ , and  $\Omega \subset \mathbb{R}^d$  be a convex subset (either compact or have no mass escaping at infinity).

1. The metric space  $(\mathcal{P}_p(\Omega), W_p)$  is a geodesic space.
2. For  $\mu, \nu \in \mathcal{P}_p(\Omega)$ , a constant-speed geodesic  $\{\mu_t\}_{t \in [0,1]}$  in  $(\mathcal{P}_p(\Omega), W_p)$  between  $\mu$  and  $\nu$  (i.e.,  $\mu_0 = \mu$  and  $\mu_1 = \nu$ ) must satisfy  $\mu_t \in \text{AC}(\mathcal{P}_p(\Omega))$  and

$$|\mu'|_p(t) = W_p(\mu(0), \mu(1)) = W_p(\mu, \nu) \quad (53)$$

for almost every  $t \in [0, 1]$ .

3. The above  $\mu_t$  solves

$$\min \left\{ \int_0^1 |\tilde{\mu}'|_p(t) dt : \tilde{\mu}(0) = \mu, \tilde{\mu}(1) = \nu, \tilde{\mu} \in \text{AC}(\mathcal{P}_p(\Omega)) \right\}. \quad (54)$$

4. The above  $\mu_t$  solves the Benamou-Brenier problem

$$W_p^p(\mu, \nu) = \min \left\{ \int_0^1 \|\mathbf{v}_t\|_{L^p(\tilde{\mu}_t)}^p dt : \tilde{\mu}(0) = \mu, \tilde{\mu}(1) = \nu, \partial_t \tilde{\mu}_t + \text{div}(\tilde{\mu}_t \mathbf{v}_t) = 0 \right\}, \quad (55)$$

and  $\mu_t = \mu(\cdot, t)$  defines a constant-speed geodesic in  $(\mathcal{P}_p(\Omega), W_p)$ .

## D ENTROPIC REGULARIZATION

Our GeONet is compatible with entropic regularization, which is closely related to the Schrödinger bridge problem and stochastic control [Chen et al., 2016]. Specifically, the entropic-regularized GeONet (ER-GeONet) solves the following fluid dynamic problem:

$$\begin{aligned}\min_{(\mu, \mathbf{v})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|\mathbf{v}(x, t)\|_2^2 \mu(x, t) dx dt \\ \text{subject to } \partial_t \mu + \text{div}(\mu \mathbf{v}) + \varepsilon \Delta \mu = 0, \quad \mu(\cdot, 0) = \mu_0, \quad \mu(\cdot, 1) = \mu_1.\end{aligned} \quad (56)$$

Applying the same variational analysis as in the unregularized case  $\varepsilon = 0$  (cf. Appendix B), we obtain the KKT conditions for the optimization (56) as the solution to the following system of PDEs:

$$\partial_t \mu + \text{div}(\mu \nabla u) = -\varepsilon \Delta \mu, \quad (57)$$

$$\partial_t u + \frac{1}{2} \|\nabla u\|_2^2 = \varepsilon \Delta u, \quad (58)$$

with the boundary conditions  $\mu(\cdot, 0) = \mu_0, \mu(\cdot, 1) = \mu_1$  for  $\varepsilon > 0$ . Note that (58) is a parabolic PDE, which has a unique smooth solution  $u^\varepsilon$ . The term  $\varepsilon \Delta u$  effectively regularizes the (dual) HJ equation in (7). In the zero-noise limit as  $\varepsilon \downarrow 0$ , the solution of the optimal entropic interpolating flow (56) converges to solution of the Benamou-Brenier problem (4) in the sense of the method of vanishing viscosity [Mikami, 2004, Evans, 2010].

## E GRADIENT ENHANCEMENT

In practice, we may fortify the base method by adding extra residual terms of the differentiated PDEs to our loss function of GeONet. Such gradient enhancement technique has been used to strengthen PINNs [Yu et al., 2022], which improves the efficiency as fewer data points are needed to be sampled from  $U(\Omega) \times U(0, 1)$ , and prediction accuracy as well.

The motivation behind gradient enhancement stems from minimizing the residual of a differentiated PDE. We turn our attention to PDEs of the form

$$\begin{cases} \mathcal{F}(x, t, \partial_{x_1} u, \dots, \partial_{x_d} u, \partial_{x_1 x_1} u, \dots, \partial_{x_d x_d} u, \dots, \partial_t u, \lambda) = 0 & \text{on } \Omega \times [0, 1], \\ u(\cdot, 0) = u_0, \quad u(\cdot, 1) = u_1 & \text{on } \Omega, \end{cases} \quad (59)$$

for domain  $\Omega \subseteq \mathbb{R}^d$ , parameter vector  $\lambda$ , and boundary conditions  $u_0, u_1$ . One may differentiate the PDE function  $\mathcal{F}$  with respect to any spatial component to achieve

$$\frac{\partial}{\partial x_\ell} \mathcal{F}(x, t, \partial_{x_1} u, \dots, \partial_{x_d} u, \partial_{x_1 x_1} u, \dots, \partial_{x_d x_d} u, \dots, \partial_t u, \lambda) = 0. \quad (60)$$

The differentiated PDE is additionally equal to 0, similar to what we see in a PINN setup. If we substitute a neural network into the differentiated PDE of (60), what remains is a new residual, just as we saw with the neural network substituted into the original PDE. Minimizing this new residual in the related loss function characterizes the gradient enhancement method.

We utilize the same loss function in (16), but we add the additional terms

$$\mathcal{L}_{\text{GE,cty}} = \sum_{\ell=1}^d \gamma_\ell \mathbb{E}[\| \frac{\partial}{\partial x_\ell} (\frac{\partial}{\partial t} \mathcal{C}_\phi + \text{div}(\mathcal{C}_\phi \nabla \mathcal{H}_\psi)) \|_{L^2(\Omega \times (0,1))}^2], \quad (61)$$

$$\mathcal{L}_{\text{GE,HJ}} = \sum_{\ell=1}^d \omega_\ell \mathbb{E}[\| \frac{\partial}{\partial x_\ell} (\frac{\partial}{\partial t} \mathcal{H}_\psi + \frac{1}{2} \|\nabla \mathcal{H}_\psi\|_2^2) \|_{L^2(\Omega \times (0,1))}^2], \quad (62)$$

where the variables and neural networks that also appeared in (16) are the same. Here  $\gamma_\ell$  and  $\omega_\ell$  are positive weights. The summation is taken in order to account for the gradient enhancement of each spatial component of  $x \in \Omega$ .

## F SPECIALIZED ARCHITECTURES

### F.1 MODIFIED MULTI-LAYER PERCEPTRON

Here we outline the forward pass of the modified multi-layer perceptron used throughout the experiments as presented in Wang et al. [2021b]. Let  $\sigma$  denote an activation function (at least twice differentiable to allow automatic differentiation of the networks to satisfy the PDEs),  $X$  as neural network design input,  $W^i$  the weights of the neural network at index  $i$ , and  $b^i$  the bias at layer  $i$ . Here,  $X$  can refer to either branch or trunk inputs, as this architecture is used for both.

The forward pass is given by

$$U = \sigma(W^1 X + b^1), \quad V = \sigma(W^2 X + b^2) \quad (63)$$

$$H^1 = \sigma(W^{h,1} X + b^{h,1}) \quad (64)$$

$$Z^k = \sigma(W^{z,k} H^k + b^{z,k}) \quad (65)$$

$$H^k = (1 - Z^{k-1}) \odot U + Z^{k-1} \odot V \quad (66)$$

$$\mathcal{N}_\theta = W^\ell H^\ell + b^\ell, \quad (67)$$

where  $k \in \{1, \dots, \ell\}$ ,  $\odot$  is an element-wise product, and  $\mathcal{N}_\theta$  is the neural network final output, either a branch or a trunk.

### F.2 FOURIER FEATURE ARCHITECTURE

We outline the Fourier feature architecture of Wang et al. [2021b]. We embed trunk input  $y = (x, t)$  in a higher-dimensional space by taking transformations of the form

$$U = (\cos(2\pi B_x y), \sin(2\pi B_x y))^T \quad (68)$$

and passing them into trunk input. Alternatively, we consider the more elaborated architecture of Wang et al. [2021a], which requires passing in  $x, t$  into distinct Fourier embeddings of the form of  $U$ , and using separate layers for each. An element-wise product is taken before the last layer. We used this for our experiments of 4.2, but generally found the Fourier feature architecture of passing in  $y = (x, t)$  to formulate  $U$  as effective as well.

## G HYPERPARAMETER SETTINGS AND TRAINING DETAILS

We discuss training characteristics of GeONet based on the primary experiments. An unmodified Adam optimizer was chosen for all branch, trunk neural networks with a learning rate starting from  $5e-4$ . All layers share the same width. We use tanh activation for all neural networks. Coefficients  $\alpha_1, \alpha_2, \beta_0, \beta_1$  were computed after examining errors. Coefficients were selected in the range  $[0.05, 20]$ . Neural network depths refer to  $\ell$  in each modified MLP. Training is done on a NVIDIA T4 GPU.

Table 4: Architecture and training details in our Gaussian mixture experiments of Section 4 and Appendix H.

Hyperparameter	1D Gaussians	2D Gaussians
No. of initial conditions $(\mu_0, \mu_1)$	20,000	5,000
$m$ (branch input dimension)	100	576
Branch width	150	200
Branch depth	7	7
Trunk width	100	150
Trunk depth	7	7
$p$ (dimension of outputs)	800	800
Batch size	2,000	2,000
Final training time	$\sim 2$ hrs	$\sim 2$ hrs
Final training loss	$\sim 1.5e-4$	$\sim 1.8e-5$
$\alpha_1, \alpha_2, \beta_0, \beta_1$	0.5, 0.25, 1, 1	0.5, 0.25, 1, 1

Table 5: Architecture and training details in our empirical Gaussians and encoded MNIST experiments of Section 4 and Appendix H.

Hyperparameter	Empirical Gaussians	Encoded MNIST
No. of initial conditions $(\mu_0, \mu_1)$	1,000	30,000
$m$ (branch input dimension)	625	32
Branch width	100	150
Branch depth	7	7
Trunk width	100	100
Trunk depth	5	7
$p$ (dimension of outputs)	200	200
Batch size	1,000	1,000
Final training time	$\sim 2$ hrs	$\sim 4$ hrs
Final training loss	$\sim 7.0e-4$	$\sim 2.0e-2$
$\alpha_1, \alpha_2, \beta_0, \beta_1$	0.5, 0.25, 1, 1	1, 1, 1, 1

## H TRAINING AND PERFORMANCE

### H.1 UNIVARIATE AND BIVARIATE GAUSSIAN MIXTURE EXPERIMENTS

**Performance.** Our baseline results were collected by deploying GeONet on the identity geodesic in Table 2. The baseline identity geodesic provides a benchmark for comparing and interpreting the errors across different setups. The univariate cases were evaluated upon a 100 point mesh, and the bivariate upon a  $40 \times 40$  mesh, except in the zero-shot super-resolution case, in which the grid is refined and previously specified. From Table 2, we can draw the following observations. The loss boundary conditions (19) allow greater precision for  $t = 0, 1$ , which suggests that a lack of data-enforced conditions along the inner region of the time continuum would cause greater error. Errors for predicting the univariate Gaussian trivial identity geodesic in the intermediate  $t = 0.25, 0.5, 0.75$  are uniformly smaller than other in-distribution setups since the former is an easier task. In the bivariate experiment, we found that error quickly rises as variance decreases, which is equivalent to a task of learning more complicated geodesics. We did not find lower variance drastically affects performance in the univariate experiment, suggesting GeONet and potentially physics-informed DeepONets in general are less effective as the dimension increases. We did not find the number of Gaussians in the mixtures drastically affected results, but naturally more complicated geodesics induce greater error, which is to be expected. We found bivariate errors are similar to the random case as in the identity case, suggesting there is some notion of base neural operator error, which may not exist with simpler data.

### H.2 GAUSSIAN EMPIRICAL DENSITIES

**Training.** 3000 point cloud particles were sampled from mixtures composed of 3 Gaussians for source  $\mu_0$  and target  $\mu_1$ . 2D histograms were constructed to turn particle data into empirical densities, with bins ranging from  $-7$  to  $7$ . Domain  $\Omega = [0, 5] \times [0, 5]$  was discretized into a  $25 \times 25$  point domain and assigned for the histograms’ locations used as GeONet spatial input. A batch size of 1,000 was chosen. We take  $p = 200$ ,  $\alpha_1 = 0.5$ ,  $\alpha_2 = 0.25$ ,  $\beta_0 = \beta_1 = 1$ , which can be altered to impose strength of the boundary and physics terms accordingly. We employ the Fourier feature network architecture of Wang et al. [2021a] for trunk networks. We take matrix  $B_v$  with elements sampled in  $\mathcal{N}(0, \sigma_v^2)$ , subsequently taking  $(\cos(2\pi B_v v), \sin(2\pi B_v v))^T$  as input for a fully-connected network, where  $v$  is either space or time input. Our architecture for this experiment is fully outlined in F.2. Empirically, we found low variance necessary, and we chose  $\sigma = 0.5$  for both  $v = x, t$  for both continuity and trunk branches.

**Performance.** In this experiment, GeONet correctly captures the translocation of mass and overall geodesic behavior. The other methods are more suited for point clouds but yield high errors in learning the geodesic. GeONet tends to slightly regularize the solution by smoothing them, in which GeONet has trouble learning precision that comes with particle samples.

### H.3 MNIST EXPERIMENT

**Training.** As described in section 4, to learn the geodesic, we ensure all values within the encoded representation are nonnegative, meaning we can shift all encoded representations by some arbitrary constant. We choose 10 for this. This constant can be deducted in later stages to ensure the valid representation is met. We normalize the data so that the density conditions are satisfied before GeONet input. A domain of  $[0, 5]$  was divided into an equispaced mesh of 32 points for the encoded representation. This domain is rather arbitrary and is chosen simply for DeepONet input purposes, which can be modified as seen fit. 30,000 encoded pairs were chosen to train GeONet and the pre-trained autoencoder, the entirety of MNIST. We used a batch size of 1,000. Additional details are found in Appendix G.

Table 6:  $L^1$  error of GeONet on 50 test pairings of encoded MNIST. All values are multiplied by  $10^{-2}$ . Error was calculated upon the geodesic in both the shifted and ambient/original space.

Test setting	GeONet $L^1$ error on encoded MNIST data				
	$t = 0$	$t = 0.25$	$t = 0.5$	$t = 0.75$	$t = 1$
Encoded, identity	$0.923 \pm 0.213$	$0.830 \pm 0.166$	$0.825 \pm 0.165$	$0.834 \pm 0.173$	$0.931 \pm 0.215$
Encoded, random	$1.62 \pm 0.333$	$2.14 \pm 1.22$	$2.78 \pm 1.62$	$2.11 \pm 1.17$	$1.54 \pm 0.282$
Ambient, identity	$26.7 \pm 11.2$	$34.0 \pm 6.88$	$35.3 \pm 8.32$	$36.4 \pm 9.77$	$34.0 \pm 13.2$
Ambient, random	$32.1 \pm 16.6$	$58.2 \pm 15.0$	$68.1 \pm 18.8$	$56.4 \pm 14.3$	$24.7 \pm 10.7$

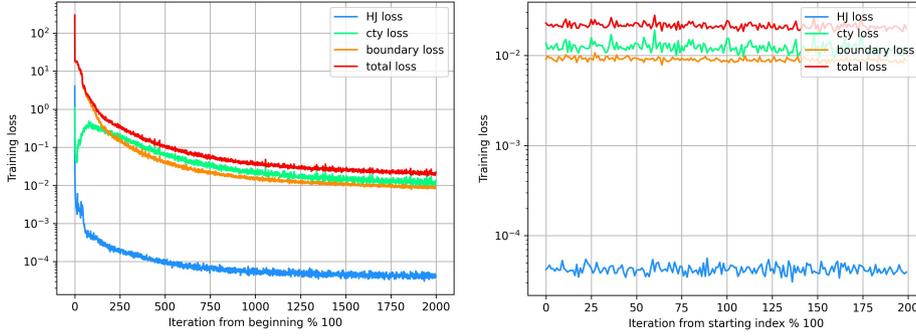


Figure 8: We examine iterations of the Adam optimizer in the total and late training on a log scale. We examine late training in order to observe oscillatory behavior between the continuity and HJ loss to see if they adversarially compete in late training. We do not observe this pattern, and the continuity loss greatly surpasses the HJ loss in value. These graphs were created using the encoded MNIST experiment.

**Performance.** GeONet performs well in this experiment. Scaling the physics-informed term by a constant of less than one did not prove necessary in this experiment to ensure all loss terms are met to a sufficient degree. As before, boundary terms are uniformly smaller, likely since these terms are known and included in the loss function to be minimized. The same error metric is used as in the synthetic experiments but with normalization, making the  $L^1$  error relative. We remark OOD generalization is omitted because the distribution of the encoded data is not known. We also remark the decoded images, being the geodesic returned to its original state, do not directly translate to a geodesic performed upon an original pair of images. NaN values are omitted in the error computations, which are possible in the POT solutions due to the irregularity of the initial conditions.

**Regularization.** Classical geodesic algorithms require a small regularization parameter in order to be computed. This affects the synthetic experiments trivially, but we found this regularization induces greater in the MNIST experiment. This is to be considered when evaluating the errors, and true error is likely smaller between GeONet and the reference geodesics computed with POT than what is displayed. This regularization acts as a form of "smoothing" of the solutions.

## I GEONET ERROR FOR ADDITIONAL ERROR METRICS

Table 7: We list mean and standard deviations of error of GeONet on 50 random  $\mu_0 \neq \mu_1$  samples for alternative error metrics, being  $L^2$  error and the Wasserstein-1 distance. We remark we use sliced Wasserstein distance for the 2D case, as this metric is computationally feasible for higher dimensional cases. We perform this for random Gaussian mixture pairings. All values are multiplied by  $10^{-2}$  by those of the table.

Experiment	GeONet alternative metric error for random Gaussian mixtures		
	$t = 0$	$t = 0.25$	$t = 0.5$
1D, $L^2$	$5.19 \pm 1.74$	$6.91 \pm 4.81$	$7.28 \pm 5.39$
1D, Wasserstein	$0.352 \pm 0.116$	$0.364 \pm 0.178$	$0.403 \pm 0.228$
2D, $L^2$	$6.93 \pm 0.883$	$7.72 \pm 1.23$	$8.11 \pm 1.30$
2D, Wasserstein	$0.245 \pm 0.0329$	$0.264 \pm 0.0316$	$0.275 \pm 0.0447$

Experiment	$t = 0.75$	$t = 1$
1D, $L^2$	$6.49 \pm 4.36$	$4.81 \pm 1.58$
1D, Wasserstein	$0.386 \pm 0.166$	$0.347 \pm 0.101$
2D, $L^2$	$7.79 \pm 1.14$	$6.87 \pm 1.05$
2D, Wasserstein	$0.267 \pm 0.0338$	$0.246 \pm 0.0356$

## J 3D GAUSSIANS FIGURE

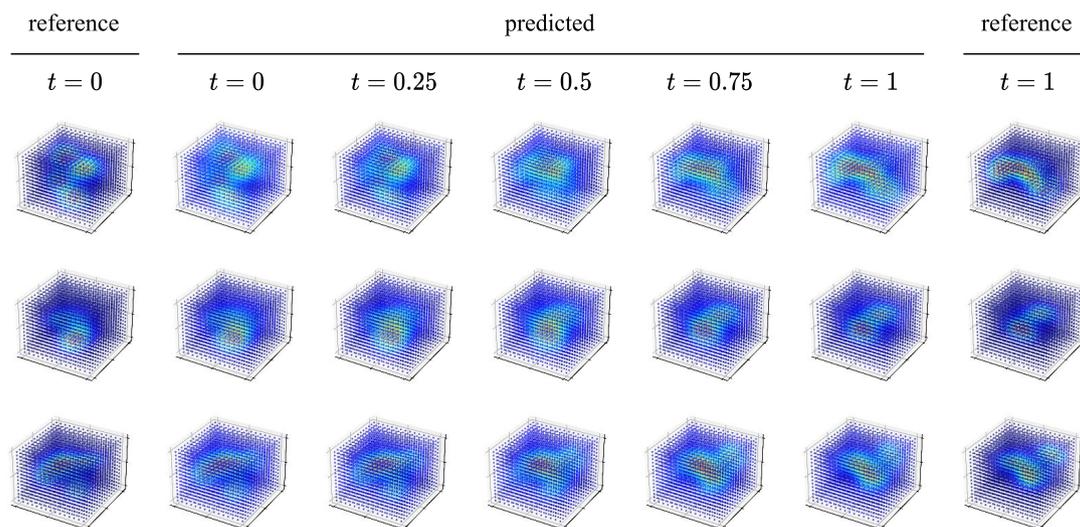
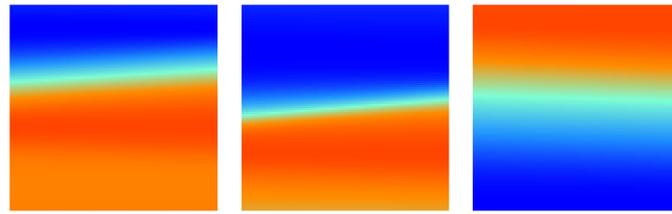
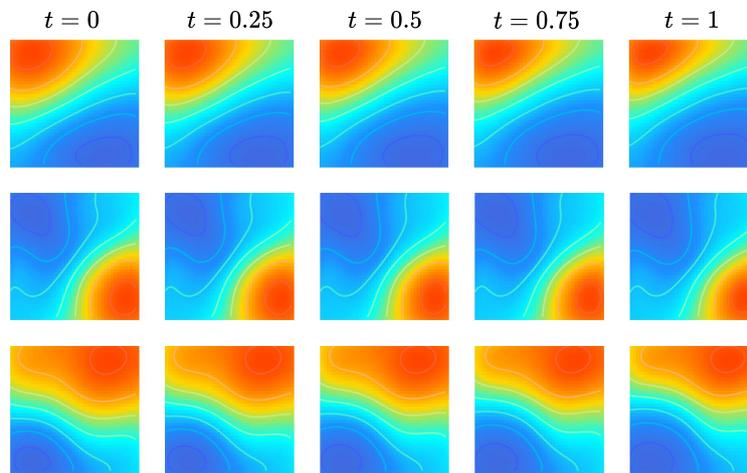


Figure 9: We illustrate GeONet on 3D Gaussians.

## K SAMPLE HJ GRAPHS



(a)



(b)

Figure 10: We present sample HJ equations for (a) three univariate Gaussian mixtures and (b) three bivariate Gaussian mixtures from the primary experiments performed in Section 4. The univariate HJ samples at certain times are the vertical cross-sections of the graphs, and the bivariate samples are given at certain times.