A 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

$$\min_{(\mu,\mathbf{v})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} ||\mathbf{v}(x,t)||_2^2 \, \mu(x,t) \, \mathrm{d}x \, \mathrm{d}t \tag{18}$$

$$\text{subject to } \partial_t \mu + \mathrm{div}(\mu \mathbf{v}) = 0, \tag{19}$$

$$\mu(\cdot,0) = \mu_0, \ \ \mu(\cdot,1) = \mu_1. \tag{20}$$

Here, equation (19) is referred to as the *continuity equation* (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 μ_0 and terminating at μ_1 as

$$L(\mu,\mathbf{v},u) = \int_0^1 \int_{\mathbb{R}^d} \left[\frac{1}{2} \|\mathbf{v}\|_2^2 \mu + (\partial_t \mu + \mathrm{div}(\mu\mathbf{v}))\, u \right] \, \mathrm{d}x \, \mathrm{d}t, \tag{21}$$

where $u := u(x,t)$ is the dual variable for (CE). To find the optimal solution μ^* for the minimum kinetic energy (18), we study the saddle point optimization problem

$$\min_{(\mu,\mathbf{v})\in(\text{CE})} \max_u L(\mu,\mathbf{v},u), \tag{22}$$

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 $\mu \notin$ (CE), then by scaling with arbitrarily large constant, we see that

$$\max_u \int_0^1 \int_{\mathbb{R}^d} (\partial_t \mu + \mathrm{div}(\mu\mathbf{v}))\, u \, \mathrm{d}x \, \mathrm{d}t = +\infty. \tag{23}$$

Thus,

$$\min_{(\mu,\mathbf{v})\in(\text{CE})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} ||\mathbf{v}||_2^2 \mu \, dx \, dt = \min_{(\mu,\mathbf{v})} \max_u L(\mu,\mathbf{v},u) \tag{24}$$

$$\geqslant \max_u \min_{(\mu,\mathbf{v})} L(\mu,\mathbf{v},u), \tag{25}$$

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

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

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

$$\max_u \min_\mu L(\mu,\mathbf{v}^*,u) = \int_0^1 \int_{\mathbb{R}^d} \left[-\left(\frac{1}{2} \|\nabla u\|_2^2 + \partial_t u \right) \mu \right] \, \mathrm{d}x \, \mathrm{d}t + \int_{\mathbb{R}^d} [u(\cdot,1)\mu_1 - u(\cdot,0)\mu_0] \, \mathrm{d}x, \tag{26}$$

for any flow μ_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

$$\min_\mu \int_0^1 \int_{\mathbb{R}^d} \left[-\left(\frac{1}{2} \|\nabla u\|_2^2 + \partial_t u \right) \mu \right] \, \mathrm{d}x \, \mathrm{d}t = -\infty \tag{27}$$

because μ is unconstrained (except for the boundary conditions). Then we deduce that

$$\min_{(\mu,\mathbf{v})\in(\text{CE})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} ||\mathbf{v}||_2^2 \mu \geqslant \max_{u\in(\text{HJ})} \left\{ \int_{\mathbb{R}^d} u(\cdot,1)\mu_1 - \int_{\mathbb{R}^d} u(\cdot,0)\mu_0 \right\}, \tag{28}$$

where $u \in$ (HJ) means that u solves the *Hamilton-Jacobi equation* (HJ)

$$\partial_t u + \frac{1}{2}\|\nabla u\|_2^2 = 0. \tag{29}$$

From (28), we see that the duality gap is non-negative, and it is equal to zero if and only if (μ^*, u^*) solves the following system of PDEs

$$\begin{cases} \partial_t \mu + \operatorname{div}(\mu \nabla u) = 0, \;\; \partial_t u + \frac{1}{2}\|\nabla u\|_2^2 = 0, \\ \mu(\cdot, 0) = \mu_0, \;\; \mu(\cdot, 1) = \mu_1. \end{cases} \tag{30}$$

PDEs in (30) are referred to as the Karush–Kuhn–Tucker (KKT) conditions for the Wasserstein geodesic problem.

B Metric geometry structure of the Wasserstein space and geodesic

In this section, we review some basic facts on 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 $p \geqslant 1$. 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).

B.1 General metric space

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

$$d(\omega(t_0), \omega(t_1)) \leqslant \int_{t_0}^{t_1} g(\tau)\, \mathrm{d}\tau. \tag{31}$$

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

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

$$|\omega'|(t) := \lim_{h \to 0} \frac{d(\omega(t+h), \omega(t))}{|h|}, \tag{32}$$

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

$$d(\omega(t), \omega(s)) \leqslant \int_t^s |\omega'|(\tau)\, \mathrm{d}\tau. \tag{33}$$

Theorem B.3 tells us that absolutely continuous curve ω has a metric derivative well-defined almost everywhere, and the “length" of the curve ω 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 B.4 (Curve length). For a curve $\omega : [0, 1] \to X$, we define its *length* as

$$\mathrm{Length}(\omega) := \sup \left\{ \sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1})) : n \geqslant 1, 0 = t_0 < t_1 < \ldots < t_n = 1 \right\}. \tag{34}$$

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

$$d(\omega(t_k), \omega(t_{k+1})) \leqslant \int_{t_k}^{t_{k+1}} g(\tau)\, \mathrm{d}\tau \tag{35}$$

so that

$$\text{Length}(\omega) \leqslant \int_0^1 g(\tau)\,\mathrm{d}\tau < \infty, \tag{36}$$

i.e., the curve ω is of bounded variation.

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

$$\text{Length}(\omega) = \int_0^1 |\omega'|(\tau)\,\mathrm{d}\tau. \tag{37}$$

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

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

$$d(x, y) = \inf\left\{\text{Length}(\omega) : \omega(0) = x, \omega(1) = y, \omega \in \text{AC}(X)\right\}. \tag{38}$$

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

$$d(x, y) = \min\left\{\text{Length}(\omega) : \omega(0) = x, \omega(1) = y, \omega \in \text{AC}(X)\right\}. \tag{39}$$

Definition B.7 (Geodesic). Let (X, d) be a length space.

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

$$d(\omega(t), \omega(s)) = |t - s| \cdot d(\omega(0), \omega(1)), \tag{40}$$

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

2. If (X, d) is further a geodesic space, a curve $\omega : [0,1] \to X$ 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 B.8. Let (X, d) be a geodesic space, $p > 1$, and $\omega : [0,1] \to X$ a curve connecting x_0 and x_1. Then the followings are equivalent.

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

$$|\omega'|(t) = d(\omega(0), \omega(1)). \tag{41}$$

3. ω solves

$$\min\left\{\int_0^1 |\tilde{\omega}'|^p\,\mathrm{d}t : \tilde{\omega}(0) = x_0, \tilde{\omega}(1) = x_1\right\}. \tag{42}$$

B.2 Wasserstein space

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

Definition B.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

$$|\mu'|_p(t) := \lim_{h\to 0} \frac{W_p(\mu_{t+h}, \mu_t)}{|h|}. \tag{43}$$

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

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

Theorem B.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. μ_t is a weak solution of the continuity equation $\partial_t\mu_t + \text{div}(\mu_t\mathbf{v}_t) = 0$ in the sense of distributions (cf. the definition in (49) below);
2. for almost every $t \in [0,1]$, we have
$$\|\mathbf{v}_t\|_{L^p(\mu_t)} \leqslant |\mu'|_p(t), \tag{44}$$
where $\|\mathbf{v}_t\|^p_{L^p(\mu_t)} = \int_\Omega \|\mathbf{v}_t\|_2^p \, \mathrm{d}\mu_t$.

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 $\int_0^1 \|\mathbf{v}_t\|_{L^p(\mu)} \, \mathrm{d}t < \infty$ such that $(\mu_t, \mathbf{v}_t)$ solves the continuity equation, 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 $t \in [0,1]$,
$$|\mu'|_p(t) \leqslant \|\mathbf{v}_t\|_{L^p(\mu_t)}. \tag{45}$$

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

Corollary B.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 B.10 must satisfy
$$\|\mathbf{v}_t\|_{L^p(\mu_t)} = |\mu'|_p(t). \tag{46}$$

Corollary B.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 B.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)
$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = \mathbf{v}_t(y(t)). \tag{47}$$
The dynamical system (47) defines a flow $Y_t : \Omega \to \Omega$ of vector field $\mathbf{v}_t$ on Ω via
$$Y_t(x) = y(t). \tag{48}$$
Then, it is straightforward to check that the pushforward measure flow $\mu_t := (Y_t)_\sharp\mu_0$ and the chosen velocity vector field $\mathbf{v}_t$ in the ODE (47) is a weak solution of the continuity equation $\partial_t\mu_t + \text{div}(\mu_t\mathbf{v}_t) = 0$ in the sense that
$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \phi \, \mathrm{d}t = \int_\Omega \langle \nabla\phi, \mathbf{v}_t \rangle \, \mathrm{d}\mu_t, \tag{49}$$
for any $\mathcal{C}^1$ function $\phi : \Omega \to \mathbb{R}$ with compact support.

Theorem B.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 γ be an optimal transport plan under the cost function $\|x - y\|_p^p$. Define
$$\pi_t : \Omega \times \Omega \to \Omega,$$
$$\pi_t(x, y) = (1-t)x + ty,$$
as the linear interpolation between x and y in Ω. Then, the curve $\mu_t = (\pi_t)_\sharp\gamma$ is a constant-speed geodesic in $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ connecting $\mu_0 = \mu$ and $\mu_1 = \nu$.

If μ has a density with respect to the Lebesgue measure on $\mathbb{R}^d$, then there is an optimal transport map T from μ to ν (Brenier, 1991). According to Theorem B.12, we obtain *McCann's interpolation* (McCann, 1997) in the Wasserstein space as
$$\mu_t = [(1-t)\text{id} + tT]_\sharp\mu, \tag{50}$$

which is a constant-speed geodesic in $(\mathcal{P}_p(\mathbb{R}^d), W_p)$. 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 μ and ν (i.e., $\mu_0 = \mu$ and $\mu_1 = \nu$) must satisfy $\mu_t \in \mathrm{AC}(\mathcal{P}_p(\Omega))$ and
$$|\mu'|(t) = W_p(\mu(0), \mu(1)) = W_p(\mu, \nu) \tag{51}$$
for almost every $t \in [0, 1]$.
3. The above μ_t solves
$$\min \left\{ \int_0^1 |\tilde{\mu}'|^p(t) \, \mathrm{d}t : \tilde{\mu}(0) = \mu, \tilde{\mu}(1) = \nu, \tilde{\mu} \in \mathrm{AC}(\mathcal{P}_p(\Omega)) \right\}. \tag{52}$$
4. The above μ_t solves the Benamou-Brenier problem
$$W_p^p(\mu, \nu) = \min \left\{ \int_0^1 \|\mathbf{v}_t\|_{L^p(\tilde{\mu}_t)}^p \, \mathrm{d}t : \tilde{\mu}(0) = \mu, \tilde{\mu}(1) = \nu, \partial_t \tilde{\mu}_t + \mathrm{div}(\tilde{\mu}_t \mathbf{v}_t) = 0 \right\}, \tag{53}$$
and $\mu_t = \mu(\cdot, t)$ defines a constant-speed geodesic in $(\mathcal{P}_p(\Omega), W_p)$.

C 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:

$$\begin{aligned} &\min_{(\mu,\mathbf{v})} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} ||\mathbf{v}(x,t)||_2^2 \, \mu(x,t) \, \mathrm{d}x \, \mathrm{d}t \\ &\text{subject to } \ \partial_t \mu + \mathrm{div}(\mu \mathbf{v}) + \varepsilon \Delta \mu = 0, \ \ \mu(\cdot, 0) = \mu_0, \ \ \mu(\cdot, 1) = \mu_1. \end{aligned} \tag{54}$$

Applying the same variational analysis as in the unregularized case $\varepsilon = 0$ (cf. Appendix A), we obtain the KKT conditions for the optimization (54) as the solution to the following system of PDEs:

$$\partial_t \mu + \mathrm{div}(\mu \nabla u) = -\, \varepsilon \Delta \mu, \tag{55}$$
$$\partial_t u + \frac{1}{2} \|\nabla u\|_2^2 = \ \varepsilon \Delta u, \tag{56}$$

with the boundary conditions $\mu(\cdot, 0) = \mu_0, \mu(\cdot, 1) = \mu_1$ for $\varepsilon > 0$. Note that (56) is a parabolic PDE, which has a unique smooth solution u^ε. The term $\varepsilon \Delta u$ effectively regularizes the (dual) Hamilton-Jacobi equation in (7). In the zero-noise limit as $\varepsilon \downarrow 0$, the solution of the optimal entropic interpolating flow (54) converges to solution of the Benamou-Brenier problem (4) in the sense of the method of vanishing viscosity (Mikami, 2004; Evans, 2010).

D Gradient enhancement

In this section, we fortify the base method we presented in Section 3. We present gradient enhancement, which is a technique to strengthen any standard PINN (Yu et al., 2022). This technique improves efficiency, as fewer data points are needed to be sampled from $U(\Omega) \times U(0, 1)$, and accuracy as well. We apply gradient enhancement to our proposed neural operator.

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

$$\begin{cases} \mathcal{F}\Big(x, t, \partial_{x_1} u, \ldots, \partial_{x_d} u, \partial_{x_1 x_1} u, \ldots, \partial_{x_d x_d} u, \ldots, \partial_t u, \lambda\Big) = 0 & \text{on} \quad \Omega \times [0,1], \\ u(\cdot, 0) = u_0, \quad u(\cdot, 1) = u_1 & \text{on} \quad \Omega, \end{cases} \tag{57}$$

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

$$\frac{\partial}{\partial x_\ell} \mathcal{F}\Big(x, t, \partial_{x_1} u, \ldots, \partial_{x_d} u, \partial_{x_1 x_1} u, \ldots, \partial_{x_d x_d} u, \ldots, \partial_t u, \lambda\Big) = 0. \tag{58}$$

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 (58), 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 (14), but we add the additional terms

$$\begin{aligned} \mathcal{L}_{\text{GE,cty},i} &= \frac{1}{N} \sum_{\ell=1}^{d} \gamma_\ell \Big|\Big| \ \frac{\partial}{\partial x_\ell} \Big(\frac{\partial}{\partial t} \mathcal{C}_{\phi,i} + \text{div}(\mathcal{C}_{\phi,i} \nabla \mathcal{H}_{\psi,i}) \Big) \ \Big|\Big|^2_{L^2(\Omega \times (0,1))}, & (59) \\ \mathcal{L}_{\text{GE, HJ},i} &= \frac{1}{N} \sum_{\ell=1}^{d} \omega_\ell \Big|\Big| \ \frac{\partial}{\partial x_\ell} \Big(\frac{\partial}{\partial t} \mathcal{H}_{\psi,i} + \frac{1}{2} ||\nabla \mathcal{H}_{\psi,i}||_2^2 \Big) \ \Big|\Big|^2_{L^2(\Omega \times (0,1))}, & (60) \end{aligned}$$

where the variables and neural networks that also appeared in (14) are the same. Here γ_ℓ and ω_ℓ are positive weights. The summation is taken in order to account for gradient enhancement of each spatial component of $x \in \Omega$.

E DeepONets

A challenge resides in solving the previous risk minimization problem over numerous instances of data. This challenge may be conciliated by instituting the novel architecture of the DeepONet that learns a general nonlinear operator, where one (or a pair of) neural network(s) encode(s) the input and another encodes the collocation samples. This architecture originates as a fine equivalence to the universal approximation theorem for operators.

General DeepONet. A general operator $G^\dagger$ may be approximated by an unstacked DeepONet (Chen and Chen, 1995; Lu et al., 2021)

$$G^\dagger(u_0)(x,t) \approx \sum_{k=1}^{p} \mathcal{B}_k\big(u_0(x_1), \ldots, u_0(x_m), \theta\big) \cdot \mathcal{T}_k(x, t, \xi), \tag{61}$$

where $\mathcal{B}_k, \mathcal{T}_k$ are scalar elements of output of neural networks $\mathcal{B}, \mathcal{T}$, and p is a constant denoting the number of such elements. We take $\mathcal{B}$ and $\mathcal{T}$ to be artificial neural networks parameterized by θ, ξ respectively. $\mathcal{B}, \mathcal{T}$ are known as the branch and trunk networks respectively. u_0 is the initial function in which the operator is applied, evaluated at distinct locations $x_1, \ldots, x_m$ for branch input. (x,t) is any arbitrary point in space and time in which $G^\dagger(u_0)$ may be evaluated.

Enhanced DeepONet. The above framework is restricted to one initial input function u_0. We turn our attention to the enhanced DeepONet, a DeepONet styled to act upon dual initial conditions Tan and Chen (2022). Our true operator $\Gamma^\dagger$ may be approximated using a second neural network encoder for input u_1,

$$\Gamma^\dagger(u_0, u_1)(x,t) \approx \sum_{k=1}^{p} \mathcal{B}^0_k\big(u_0(x_1), \ldots, u_0(x_m), \theta^0\big) \cdot \mathcal{B}^1_k\big(u_1(x_1), \ldots, u_1(x_m), \theta^1\big) \cdot \mathcal{T}_k(x, t, \xi). \tag{62}$$

Physics-informed DeepONet. The enhanced DeepONet may be substituted into any physics-informed framework, such as that of equation (9), taking place of the PDE solution value in the empirical loss to be minimized Wang et al. (2021). Generalization of the trained DeepONet permits any solution to the PDEs to be evaluated instantaneously given the appropriate input function(s).

F DeepONet differentiation

We found it effective to differentiate our trunk networks individually, and subsequently form the DeepONets afterwards. It can be noted the branch networks take no (x, t) input, and so no branch derivatives are taken. The product rule of calculus allows the derivatives in the continuity physics-informed term to be reformed, which can be stated as

$$\begin{aligned}\mathrm{div}(\mathcal{C}_\phi^j \nabla \mathcal{H}_\psi^j) = \sum_{j=1}^{d} \Big\{ &(\sum_{k=1}^{p} \mathcal{B}_k^{0,\mathrm{cty}} \mathcal{B}_k^{1,\mathrm{cty}} \partial_{x_j} \mathcal{T}_k^{\mathrm{cty}})(\sum_{k=1}^{p} \mathcal{B}_k^{0,\mathrm{HJ}} \mathcal{B}_k^{1,\mathrm{HJ}} \partial_{x_j} \mathcal{T}_k^{\mathrm{HJ}}) \\ &+(\sum_{k=1}^{p} \mathcal{B}_k^{0,\mathrm{cty}} \mathcal{B}_k^{1,\mathrm{cty}} \mathcal{T}_k^{\mathrm{cty}})(\sum_{k=1}^{p} \mathcal{B}_k^{0,\mathrm{HJ}} \mathcal{B}_k^{1,\mathrm{HJ}} \partial_{x_j}^2 \mathcal{T}_k^{\mathrm{HJ}}) \Big\},\end{aligned} \tag{63}$$

where the above trunk network derivatives are evaluated at $(x^j, t^j), x^j = (x_1^j, \ldots, x_d^j) \in \Omega \subseteq \mathbb{R}^d$. Similarly, the norm-gradient term in the Hamilton-Jacobi physics-informed loss can be reformulated

$$||\nabla \mathcal{H}_\psi^j||_2^2 = \sum_{j=1}^{d} (\sum_{k=1}^{p} \mathcal{B}_k^{0,\mathrm{HJ}} \mathcal{B}_k^{1,\mathrm{HJ}} \partial_{x_j} \mathcal{T}_k^{\mathrm{HJ}})^2. \tag{64}$$

We reiterate the above derivatives are computed with automatic differentiation.

G Additional simulation result for Gaussian mixtures

Here we present additional simulation result for Gaussian mixtures with $k_0 = k_1 = 5$ and $\pi_i = 0.2$ for all i, with the same loss coefficients. We choose $u_i \in [1.1, 3.9]^2$, $\sigma_{0,i}^2, \sigma_{1,i}^2 \in [0.4, 0.8]$, and covariance $\sigma_{01,i} \in [-0.4, 0.4]$. The MSE of GeONet on testing pairs is shown in Table 4.

Table 4: MSE of GeONet on 50 test data of Gaussian mixtures over a 50 × 50 mesh. All values are multiplied by 10^{-3}. We report the means and standard deviations of the MSE. Training is done with $k_0 = k_1 = 5$, $\pi_i = 0.2$ for all i, with the same loss coefficients. Training data at the boundaries has resolution 30 × 30.

	GeONet error for Gaussian mixtures				
Number of Gaussians	$\boldsymbol{t = 0}$	$\boldsymbol{t = 0.25}$	$\boldsymbol{t = 0.5}$	$\boldsymbol{t = 0.75}$	$\boldsymbol{t = 1}$
Identity $k_0 = k_1 = 5$	0.23 ± 0.23	1.70 ± 0.67	2.60 ± 1.10	1.80 ± 0.74	0.22 ± 0.21
Generic $k_0 = k_1 = 5$	0.22 ± 0.15	1.70 ± 0.92	2.70 ± 1.50	1.70 ± 0.76	0.20 ± 0.13

H Hyperparameter settings and training details

We discuss training characteristics of GeONet. We provide details of the base method. An unmodified Adam optimizer was chosen for all branch, trunk neural networks. The below widths are held for each hidden layer. The activation listed is for each hidden layer in all six neural networks. We found constant learning rate to be successful, but one can consider modifying this. Coefficients $\alpha_1, \alpha_2, \beta_0, \beta_1$ were chosen after experimentation on sub-datasets. Training is done on a NVIDIA Tesla P100 GPU.

Table 5: Architecture and training details in our experimental Section 4. Gaussian mixture details pertain to Table 4.

Hyperparameter/Characteristic	Gaussian mixtures	CIFAR-10 data
No. of initial conditions (μ_0, μ_1)	1,500	600
L (no. of collocations per epoch)	900	1,024
Branch width	180	300
Branch depth	5	4
Trunk width	120	120
Trunk depth	7	6
p (dimension of branch, trunk outputs)	120	250
Number of parameters	1,461,120	2,821,260
Batch size	9,000	10,240
Number of batches	150	60
Activation	tanh	gelu
Learning rate	5×10^{-5}	4×10^{-5}
Time per epoch	$\sim$ 22 sec	$\sim$ 12 sec
Total number of epochs	$\sim 2,500$	$\sim 4,500$
Final training time	$\sim$ 15 hrs	$\sim$ 15 hrs
Final training loss	$\sim 4 \times 10^{-6}$	$\sim 1 \times 10^{-4}$
$\alpha_1, \alpha_2, \beta_0, \beta_1$	$30, 30, 1, 1$	$10, 10, 1, 1$