The foundations of neural variational Wasserstein PDEs and its operator learning

Andrew Gracyk*

Fall 2025

Abstract

These are reading course notes emphasizing an understanding of gradient flow structure, focusing on the continuity equation, and its theoretical underpinnings that dominate the machine learning optimal transport archetype. We study neural variational PDEs, optimal transport, and connections with these areas to the operator learning archetype of machine learning. In particular, we will study background regarding Wasserstein gradient flows, Wasserstein geometry, and various sub-topics, such as JKO-algorithms.

Contents

1 Deriving the continuity equation

1.1 First derivation

Let $V \subseteq \mathbb{R}^d$ be a closed, connected, and bounded volume in Euclidean space and $\rho : \mathcal{V} \times [0, T] \to \mathbb{R}^+$ be a mass density. Let $v : \mathcal{V} \times [0, T] \to \mathbb{R}^d$ be a vector field. We have

$$\text{total mass} = \int_V \rho dV. \tag{1}$$

It is well known via flux theory in calculus that the change of mass through the surface is given by

$$\text{change of mass in the volume} = \int_{\partial V} \rho v \cdot n dS. \tag{2}$$

The mass is conserved in the volume, and we get

$$\partial_t \int_V \rho dV = -\int_{\partial V} \rho v \cdot n dS. \tag{3}$$

*Purdue Mathematics; a major thanks to Rongjie Lai for supervising this reading course

This is generally a well-known identity in engineering (see [5]). Here, our surface integral is outward-oriented. We will assume basic regularity properties such as smoothness on ρ, and since our domain is sufficiently nice, we can exchange differentiation and integration. Hence,

$$\int_V \partial_t \rho dV + \int_{\partial V} \rho v \cdot n dS = 0. \tag{4}$$

We cannot combine the integrals into one equation yet due to the first being an integral over Euclidean space and the second a surface integral. Notice the criteria of the divergence theorem are satisfied (we will allow v to be sufficiently smooth), hence

$$\int_V \partial_t \rho dV + \int_V \nabla \cdot (\rho v) dV = \int_V (\partial_t \rho + \nabla \cdot (\rho v)) dV = 0. \tag{5}$$

Just because an integral is zero, that does not imply the integrand is zero. Moreover, ρ is not a (compactly supported) test function.

There are some ways to conclude the proof. For example, we have $\partial_t \rho + \nabla \cdot (\rho v)$ if we permit V to be arbitrary, and if the equation must hold for all V.

Let us justify the negative sign in equation 3 intuitively with a special case. Suppose V is a hypercube and suppose v is outward oriented for each face. Thus, mass is moving outside the volume. But the integral is negative since the inner product is positive (assuming v is not zero) and rho is nonnegative. But mass is leaving the surface, so the left is clearly negative. Thus, the negative sign is needed.

In the next subsection, we will establish an argument in the sense of distributions.

1.2 Second derivation

Let m be a vector-valued measure such that $dm = \rho dx$. Here, ρ is also vector-valued. Let ϕ be a C^∞ compactly supported test function. It is known by integration by parts in the theory of distributions

$$\int \phi d(\nabla \cdot m) = -\int \nabla \phi \cdot dm. \tag{6}$$

There is an equivalence

$$\int \phi (\nabla \cdot \rho) dx = -\int \nabla \phi \cdot \rho dx. \tag{7}$$

By the conservation of mass equation we saw on the previous slide, but in the weak form

$$\partial_t \int \phi \rho_s dx = -\int \phi (\nabla \cdot \rho_s v) dx. \tag{8}$$

Here, $\rho = \rho_s v$ is the vector field multiplied by scalar density.

It is a known fact from PDE theory

$$\partial_t \int \phi d\alpha - \int \nabla \phi \cdot v d\alpha = 0 = \partial_t \int \phi d\alpha - \int \phi \nabla \cdot v d\alpha = 0 \implies \partial_t \alpha + \nabla \cdot (v\alpha) = 0. \tag{9}$$

The second equality is due to integration by parts. This is because ϕ is any (smooth compactly supported) function. We will denote $d\alpha = \rho_s dx$. Thus,

$$0 = -\int \phi (\nabla \cdot \rho_s v) dx - \int \nabla \phi \cdot \rho dx \tag{10}$$

$$= \partial_t \int \phi \rho_s dx - \int \nabla \phi \cdot v \rho_s dx \tag{11}$$

$$\implies \partial_t \rho_s + \nabla \cdot (v \rho_s) = 0, \tag{12}$$

and we have the result.

2 JKO conditions with continuity equation equivalence

We show the JKO conditions as in [4] are equivalent to a continuity equation.

Consider the density under the flow map $\rho(t, T(t,x))$. By the chain rule,

$$\frac{d}{dt}\rho(t,T(t,x)) = \partial_t\rho + \nabla\rho \cdot \partial_t T(t,x). \tag{13}$$

We are aware that $\partial_t T = v$, so that gives

$$\frac{d}{dt}\rho = \partial_t\rho + \nabla\rho \cdot v. \tag{14}$$

Now, consider the following:

$$\frac{d}{dt}\rho(t,T(t,x)) = \frac{d}{dt}\frac{\rho^k(x)}{\det|J(t,x)|} = -\frac{\rho^k(x)}{(\det|J(t,x)|)^2}\cdot\partial_t\det|J(t,x)| \tag{15}$$

$$= -\frac{\rho^k(x)}{\det|J(t,x)|}\cdot\frac{\partial_t\det|J(t,x)|}{\det|J(t,x)|} = -\frac{\rho^k(x)}{\det|J(t,x)|}\partial_t\log\det|J(t,x)| \tag{16}$$

$$= -\rho(t,T(t,x))\partial_t\log\det|J(t,x)| = -\rho(t,T(t,x))\mathrm{div}(v). \tag{17}$$

Thus,

$$-\rho(t,T(t,x))\mathrm{div}(v) = \partial_t\rho + \nabla\rho\cdot v \implies \partial_t\rho + \nabla\cdot(\rho v) = 0, \tag{18}$$

as desired. The equivalence of $\partial_t\log\det|J(t,x)| = \mathrm{div}(v)$ is due to Jacobi's formula.

3 Derivation of the JKO algorithm

3.1 The Euclidean case

First, we discuss the Euclidean case.

Consider the optimization problem

$$\min_x\left\{\frac{1}{2\tau}||x-x_k||_2^2 + \mathcal{E}(x)\right\}. \tag{19}$$

Note that the solution existence is dependent on conditions, such as convexity. A local minima to this problem can be found by taking the gradient

$$\nabla\left[\frac{1}{2\tau}||x-x_k||_2^2 + \mathcal{E}(x)\right] = \frac{1}{\tau}(x-x_k) + \nabla\mathcal{E}(x) = 0. \tag{20}$$

As a side remark, the first term derivative can be seen more clearly using

$$\nabla\langle x-x_k, x-x_k\rangle = [\partial_j(\sum_i (x_i - x_{k,i})^2)]_j = [2(x_j - x_{k,j})]_j. \tag{21}$$

Here, $[\cdot]_j$ denotes vector concatenation. Rearranging our equation, we see

$$\frac{x-x_k}{\tau} - \nabla\mathcal{E}(x), \tag{22}$$

which is exactly the Euler scheme for the gradient flow equation

$$\partial_t x(t) = -\nabla\mathcal{E}(x(t)). \tag{23}$$

4 The Wasserstein case

The Wasserstein gradient flow PDE is

$$\partial_t \rho = \nabla \cdot (\rho \nabla(\mathcal{E}'(\rho))). \tag{24}$$

We have denoted $\mathcal{E}'$ the first variation of the functional. Note that [2] is a good reference for this section. Similarly as before, consider the optimization problem

$$\min_{\rho} \left\{ \frac{1}{2\tau} \int |x - T(x)| d\rho + \mathcal{E}(\rho) \right\}. \tag{25}$$

We cannot take the gradient of this optimization problem, as we did in the Euclidean case. Instead, the equivalent is the first variation, so we have a solution to this problem takes the form

$$\delta(\frac{1}{2\tau} \int |x - T(x)| d\rho + \mathcal{E}(\rho)) = \text{constant}. \tag{26}$$

A minima occurs when the first variation is constant, not zero [2]. It is known $T(x) = x - \nabla\phi(x)$ for some convex ϕ [6][3] (page 4) [2]. It is also known the first variation of the Wasserstein distance is [2]

$$\delta(\frac{1}{2\tau} W_2^2(\rho, \rho_k)) = \frac{1}{\tau}\phi. \tag{27}$$

Thus,

$$\frac{1}{\tau}\phi(x) + \mathcal{E}'(\rho_k)(x) = \text{constant} \implies \frac{1}{\tau}\nabla\phi(x) + \nabla\mathcal{E}'(\rho_k)(x) = 0. \tag{28}$$

Equivalenty,

$$\frac{1}{\tau}(x - T(x)) = -\nabla\mathcal{E}'(\rho_k)(x). \tag{29}$$

In the continuum limit, we get

$$v = -\nabla\mathcal{E}'(\rho_k), \tag{30}$$

which yields our Wasserstein gradient flow when substituted into 24.

5 Wasserstein geometry

In this section, we discuss the Riemannian metric of Wasserstein space. Recall the Riemannian metric is defined as

$$g|_p : \text{Tan}_p M \times \text{Tan}_p M \to \mathbb{R}^*, g_{ij} = g(\partial_{x^i}|_p, \partial_{x^j}|_p). \tag{31}$$

We have used nonstandard notation here, and we will denote $\mathbb{R}^* = \mathbb{R}^+$ corresponding to $i = j$ and $\mathbb{R}^* = \mathbb{R}$ corresponding to $i \neq j$.

Consider the set of Borel measures on M such that

$$\mathcal{B}_2(M) = \left\{ \mu : \text{Borel sets on } M \to \mathbb{R}^+ \cup \{0\} : \int_M d(x, x_0)^2 d\mu(x) < \infty, x_0 \in M, \int_M d\mu(x) = 1 \right\}. \tag{32}$$

We remark this notation is also a bit nonstandard, but is very clear, so we will use it. As a remark, recall that $M = \mathbb{R}^d$ is indeed a valid Riemannian manifold. Our continuity equation stems from this set.

The Riemannian metric on Wasserstein space is

$$g_W(\nabla\phi, \nabla\psi) = \int_M \langle \nabla\phi, \nabla\psi \rangle d\mu(x). \tag{33}$$

Here, $\mu = \mu_0 \in \mathcal{B}_2(M)$ is the initial measure corresponding to a continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t \nabla\Psi) = 0, \tag{34}$$

where the divergence is defined in the sense of distributions (or we can define the gradient of the measure using the Radon-Nikodym derivative, which is the density), and

$$\nabla\phi, \nabla\psi \in \left\{ \nabla\Psi : \Psi \text{ is a plausible potential} \right\} = \Gamma, \Gamma^* = \overline{\left\{ \nabla\Psi : \Psi \in \mathcal{C}(M) \right\}}^{L^p(\mu;X)} : \tag{35}$$

are two velocity fields in the continuity equations corresponding to two distinct tangent elements. Γ is a generalized set of admissible potentials, and Γ^* is a more prototypical example. Here, the closure is taken to ensure the set is compact. $\mathcal{C}$ is a set of admissible smooth functions. We refer to [1], chapter 8 for discussion of this. In particular, it is known [1]

$$v_t \in \overline{\left\{ j_q(\nabla\Psi) : \Psi \in \text{Cyl}(X) \right\}}^{L_p(\mu_t;X)} \tag{36}$$

and j_q is the map $j_q(v) = ||v||_2^{q-2} v$.

6 JKO variational equivalence

Consider the OT problem [4]

$$\begin{cases} (\rho, v) = \arg\min \int_0^1 \int_{\mathbb{R}^d} \rho ||v||_2^2 dx dt + 2\Delta t \mathcal{E}(\rho(1,\cdot)) \\ \text{subject to } \partial_t \rho + \nabla(\rho v) = 0, \rho(0,x)\rho_0. \end{cases} \tag{37}$$

Equivalenty, this can be written

$$\begin{cases} \min_v \int_0^1 \int_{\mathbb{R}^d} \rho^n(x) ||v(T(t,x))||_2^2 dx dt + 2\Delta t \mathcal{E}(T(1,\cdot)\#\rho^n) \\ \text{subject to } \frac{d}{dt} T = v, T(0,x) = x. \end{cases} \tag{38}$$

We discuss why these are equivalent. We already proved in section 2 why the Lagrangian flow map condition $\frac{d}{dt}T = v$ is equivalent to a continuity equation.

It is also known

$$\rho(\tau,\cdot) = T(\tau,\cdot)\#\rho^n. \tag{39}$$

There is a known equivalence [7] (chapter 1)

$$\int_Y \varphi(y) d\nu(y) = \int_X \varphi(T(x)) d\mu(x) \tag{40}$$

when $T\#\mu = \nu$. This implies $\rho(1,\cdot) = T(1,\cdot)\#\rho^n$. The result follows from these facts.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition, 2008. ISBN 978-3-7643-8722-8. doi: 10.1007/978-3-7643-8722-8. URL https://doi.org/10.1007/978-3-7643-8722-8.

[2] Abdul Fatir Ansari. Introduction to gradient flows in the 2-wasserstein space. URL https://abdulfatir.com/blog/2020/Gradient-Flows/.

[3] Gang Bao and Yixuan Zhang. Optimal transportation for electrical impedance tomography, 2022. URL https://arxiv.org/abs/2210.16082.

[4] Wonjun Lee, Li Wang, and Wuchen Li. Deep jko: time-implicit particle methods for general nonlinear gradient flows, 2023. URL https://arxiv.org/abs/2311.06700.

[5] J. Gordon Leishman. *Conservation of Mass: Continuity Equation*. 2021. URL https://eaglepubs.erau.edu/introductiontoaerospaceflightvehicles/chapter/conservation-of-mass-continuity-equation/.

[6] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows, 2021. URL https://arxiv.org/abs/2106.00736.

[7] Cédric Villani. Optimal transport, old and new. URL https://www.ceremade.dauphine.fr/~mischler/articles/VBook-O&N.pdf.