Foundations of continuity equations for their neural applications

Andrew Gracyk*

Fall 2025

Abstract

These are reading course notes in conjunction with Professor Rongjie Lai at Purdue University 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 JKO variational equivalence and the Lebesgue integral formulation

Consider the OT problem Lee et al. (2023)

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

Equivalenty, this can be written

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

*Purdue University. A major thanks to Rongjie Lai for supervising this reading course; in particular, much of these notes is from his optimal transport course

We discuss why these are equivalent. We already proved in section 5 why the Lagrangian flow map condition $\frac{d}{dt}T = v$ is equivalent to a continuity equation.

It is also known

$$\rho(\tau, \cdot) = T(\tau, \cdot)\#\rho^n. \tag{3}$$

There is a known equivalence Villani (chapter 1)

$$\int_Y \varphi(y)d\nu(y) = \int_X \varphi(T(x))d\mu(x) \tag{4}$$

when $T\#\mu = \nu$. This implies $\rho(1, \cdot) = T(1, \cdot)\#\rho^n$. The result follows from these facts. In particular, we can take $\varphi = ||v||_2^2$ and we see

$$\iint\limits_{[0,1]\times\mathbb{R}^d} \rho(t,x)||v(t,x)||_2^2 dxdt = \iint\limits_{[0,1]\times\mathbb{R}^d} ||v(t,x)||_2^2 d\rho dt \tag{5}$$

$$= \iint\limits_{[0,1]\times\mathbb{R}^d} ||v(t,T(t,x))||_2^2 d\rho^n dt = \iint\limits_{[0,1]\times\mathbb{R}^d} \rho^n(x)||v(t,T(t,x))||_2^2 dxdt. \tag{6}$$

Here, we have assumed our measures are absolutely continuous with respect to Lebesgue measure. In particular, it can be noted that Lebesgue measure coincides with Riemann integration if we are allowed the assumption of Riemann integrability, which we will indeed. As a reminder, we say a measure Γ is absolutely continuous with respect to Lebesgue measure λ^*, or $\Gamma \ll \lambda^*$, if there exists a density ρ such that

$$\Gamma(A) = \int_A f d\lambda^*, \tag{7}$$

where A is any measurable set on the Lebesgue σ-algebra.

2 Deriving the continuity equation

2.1 A more prototypical engineering 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

$$\text{total mass} = \int_V \rho dV. \tag{8}$$

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

$$\text{change of mass in the volume} = \int_{\partial V} \rho v \cdot n dS. \tag{9}$$

The mass is conserved in the volume, and we get

$$\partial_t \int_V \rho dV = -\int_{\partial V} \rho v \cdot n dS. \tag{10}$$

This is generally a well-known identity in engineering (see Leishman (2021)). 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,

$$\int_V \partial_t \rho dV + \int_{\partial V} \rho v \cdot n dS = 0. \tag{11}$$

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

$$\int_V \partial_t \rho dV + \int_V \nabla \cdot (\rho v) dV = \int_V (\partial_t \rho + \nabla \cdot (\rho v)) dV = 0. \tag{12}$$

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 10 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.

2.2 Continuation equation definition in the weak sense

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

$$\int \phi d(\nabla \cdot m) = -\int \nabla \phi \cdot dm. \tag{13}$$

There is an equivalence

$$\int \phi (\nabla \cdot \rho) dx = -\int \nabla \phi \cdot \rho dx. \tag{14}$$

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

$$\partial_t \int \phi \rho_s dx = -\int \phi (\nabla \cdot \rho_s v) dx. \tag{15}$$

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

It is a known fact from PDE theory

$$\partial_t \int \phi d\alpha - \int \nabla \phi \cdot v d\alpha = 0 = \partial_t \int \phi d\alpha - \int \phi \nabla \cdot v d\alpha = 0 \implies \partial_t \alpha + \nabla \cdot (v\alpha) = 0. \tag{16}$$

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,

$$0 = -\int \phi (\nabla \cdot \rho_s v) dx - \int \nabla \phi \cdot \rho dx \tag{17}$$

$$= \partial_t \int \phi \rho_s dx - \int \nabla \phi \cdot v \rho_s dx \tag{18}$$

$$\implies \partial_t \rho_s + \nabla \cdot (v \rho_s) = 0, \tag{19}$$

and we have the result.

2.3 Theorem deriving the continuity equation

Let $T_t : \mathbb{R}^n \to \mathbb{R}^n$, $0 \leqslant t \leqslant 1$, be a locally Lipschitz family of diffeomorphisms (isomorphism between manifolds) with $T_0 = \text{id}$. Let $v(t, x)$ be the velocity field associated with the trajectories $T_t = X(t, x)$,

$$\partial_t X(t, x) = v(t, X(t, x)). \tag{20}$$

Let α_0 be a probability distribution on $\Omega \subseteq \mathbb{R}^n$ and $\alpha(t, \cdot) = T_t \# \alpha_0$. Then $\alpha(t, \cdot)$ is the unique solution of the following linear transport equation:

$$\begin{cases} \partial_t \alpha + \nabla \cdot (\alpha v) = 0, 0 < t < 1 \\ \alpha(0, \cdot) = \alpha_0. \end{cases} \tag{21}$$

Proof. We previously saw the weak formulation of the continuity equation is such that

$$\partial_t \int \phi \rho_s dx = -\int \phi (\nabla \cdot \rho_s v) dx \tag{22}$$

in the previous section. Note that this formulation is the conservation of mass, which roughly follows from the definition of flux. This conservation of mass equation has heavy flavor from engineering fields, but we will mostly accept it as true. We will adopt $\alpha = \rho_s dx$ to be our measure with respect to Lebesgue measure, i.e. $\alpha(A) = \int_A \rho_s dx$, so ρ_s is a scalar density. Hence, if we can check for $\varphi \in C_0^\infty(\Omega)$

$$\partial_t \int \phi d\alpha = -\int \phi(\nabla \cdot v) d\alpha, \tag{23}$$

then we have the continuity equation holds. In particular, most of our proof will be attempting to boil down what is given to this form. We will begin with the equivalence Villani (chapter 1)

$$\int_\Omega \varphi(y) dT_t \# \alpha_0 = \int_\Omega \varphi(T_t(x)) d\alpha_0. \tag{24}$$

Note that it is not necessarily the case the domain need be the same in this formulation, but we will operate so that they are equal. Moreover, it is known

$$\frac{\partial}{\partial t}(\varphi \circ T_t) = \nabla \varphi(T_t) \cdot \frac{\partial T_t}{\partial_t} = \nabla \varphi(T_t) \cdot v(t, T_t(x)). \tag{25}$$

The equivalence is by definition. As a reminder, this is a Euclidean inner product. Now, for all $h > 0$, we compute

$$\frac{1}{h}\left(\int_\Omega \varphi d \underbrace{\alpha(t+h, \cdot)}_{= T_{(t+h)} \# \alpha_0} - \int_\Omega \varphi d \underbrace{\alpha(t, \cdot)}_{T_t \# \alpha_0} \right) = \int_\Omega \frac{\varphi(T_{t+h}(x)) - \varphi(T_t(x))}{h} d\alpha_0, \tag{26}$$

which follows from 24. If we let $h \to 0$, then the above formulation is precisely a derivative, hence

$$\lim_{h \searrow 0} \int_\Omega \frac{\varphi(T_{t+h}(x)) - \varphi(T_t(x))}{h} d\alpha_0 = \frac{d}{dt} \int_\Omega \varphi d\alpha(t, \cdot), \tag{27}$$

where we have permitted regularity conditions. Moreover,

$$\frac{d}{dt} \int_\Omega \varphi d\alpha(t, \cdot) = \int_\Omega \nabla \varphi \cdot v d\alpha_0. \tag{28}$$

This follows from this equivalence. In particular,

$$\lim_{h \searrow 0} \int_\Omega \frac{\varphi(T_{t+h}(x)) - \varphi(T_t(x))}{h} d\alpha_0 = \int (\partial_t \phi(T_t(x)) d\alpha_0 = \int_\Omega \nabla \varphi(T_t) \cdot v(t, T_t(x)) d\alpha_0. \tag{29}$$

Note that is precisely conservation of mass.

There is more to prove. We will show uniqueness, i.e. if $\alpha_0 = 0$, then $\alpha(t, \cdot) = 0$ for all $0 < t < 1$. This shows uniqueness because we can take $\alpha_1 - \alpha_2 = 0$, and the result follows from well-posedness of the PDE. We examine the duality method. Assume that we can construct a function $\varphi(t, x)$ satisfying

$$\begin{cases} \frac{\partial \varphi}{\partial t} = -v \cdot \nabla \varphi \\ \varphi(s, \cdot) = \varphi_s, \ \forall s < 1 \text{ fixed.} \end{cases} \tag{30}$$

Here, φ is an arbitrary test function

$$\frac{d}{dt} \int_\Omega \varphi(t, x) d\alpha(t, \cdot) \overset{\text{chain rule}}{=} \int_\Omega \frac{\partial \varphi}{\partial t} d\alpha(t, \cdot) + \int_\Omega \varphi(t, x) d \frac{\partial \alpha}{\partial t} \tag{31}$$

$$= \underbrace{\int_\Omega (-v \cdot \nabla \varphi) d\alpha(t, \cdot)}_{\text{follows by assumption}} + \underbrace{\int_\Omega \varphi(t, x) d(-\nabla \cdot (v\alpha))}_{\text{follows from continuity equation}} \tag{32}$$

$$= \int_\Omega (-v \cdot \nabla \varphi) d\alpha + \underbrace{\int_\Omega \nabla \varphi \cdot v d\alpha}_{\text{we refer to 13, which is an integration by parts for divergence of measures}} \tag{33}$$

$$= 0. \tag{34}$$

From this, we can conclude

$$\int_{\mathbb{R}^n} \varphi(s,x)d\alpha(s,\cdot) = \int_{\mathbb{R}^n} \varphi(0,x)d\alpha_0 = 0 \text{ since } \alpha_0 \text{ is the trivial measure,} \tag{35}$$

since the time derivative is zero, thus constant in time. Since φ was arbitrary, we conclude $\alpha(s,\cdot) = 0$, $0 \leqslant s \leqslant 1$.

From 30, we have using this and the chain rule

$$\partial_t\varphi + v\cdot\nabla\varphi = 0 \implies \frac{d}{dt}\varphi(t,X(t,x)) = 0 \tag{36}$$

$$\implies \varphi(t,T_t(x)) = \varphi(s,T_s(x)) \tag{37}$$

$$\implies \varphi(t,x) = \varphi(s,T_s\circ T_t^{-1}(x)). \tag{38}$$

□

Remark. If the trajectory has randomness, i.e.

$$dX(t,x) = v(t,X(t,x))dt + \sqrt{2\epsilon}dW \tag{39}$$

is obeyed, then the corresponding density follows

$$\partial_t\rho + \nabla\cdot(v\rho) = \epsilon\Delta\rho \tag{40}$$

$$\text{or equivalently} \tag{41}$$

$$\partial_t\rho + \sum_i \frac{\partial}{\partial x_i}(v_i\rho) = \epsilon\sum_i \frac{\partial^2}{\partial^2 x_i^2}\rho, \tag{42}$$

which is a Fokker-Planck equation with nonzero diffusion ($\epsilon \neq 0$).

3 The Benamou-Brenier formulation

If the Monge map exists, then the quadratic Wasserstein distance

$$W_2^2(\alpha_0,\alpha_1) = \min_{T \text{ such that } T\#\alpha_0=\alpha_1} \int_\Omega ||x - T(x)||dx \tag{43}$$

has the equivalent formulation

$$A^2(\alpha_0,\alpha_1) = \begin{cases} \min_{\alpha,v} \iint_{[0,1]\times\mathbb{R}^n} ||v||_2^2 d\alpha dt \\ \partial_t\alpha + \nabla\cdot(\alpha v) = 0 \\ \alpha(0,\cdot) = \alpha_0 \\ \alpha(1,\cdot) = \alpha_1. \end{cases} \tag{44}$$

Proof. Let us consider the class of functions

$$C(\alpha_0,\alpha_1) = \left\{ (\alpha,v) \middle| \partial_t\alpha + \nabla\cdot(\alpha v) = 0, \alpha(0,\cdot) = \alpha_0, \alpha(1,\cdot) = \alpha_1 \right\}. \tag{45}$$

Let $(\alpha,v) \in C(\alpha_0,\alpha_1)$. Consider the problem

$$\begin{cases} \frac{\partial X(t,x)}{\partial t} = v(t,X(t,x)) \\ X(0,x) = x, x \sim \alpha_0 \\ T_t : x \in \mathbb{R}^n \to X(t,x) \in \mathbb{R}^n \\ \alpha(t,\cdot) = T_t\#\alpha_0 \text{ satisfies } \partial_t\alpha + \nabla\cdot(\alpha v) = 0. \end{cases} \tag{46}$$

Thus, we have

$$\iint_{[0,1]\times\mathbb{R}^n} ||v(t,x)||_2^2 d\alpha(t,\cdot)dt \tag{47}$$

$$\overset{\text{pushforward relation}}{=} \iint_{[0,1]\times\mathbb{R}^n} ||v(t,x)||_2^2 dT_t\#\alpha_0 dt \tag{48}$$

$$\overset{\text{Villani}}{=} \iint_{[0,1]\times\mathbb{R}^n} ||v(t,T_t(x))||_2^2 d\alpha_0 dt \tag{49}$$

$$= \iint_{[0,1]\times\mathbb{R}^n} ||v(t,X(t,x))||_2^2 d\alpha_0 dt \tag{50}$$

$$\overset{\text{Fubini's (assume regularity)}}{=} \iint_{[0,1]\times\mathbb{R}^n} ||v(t,X(t,x))||_2^2 dt d\alpha_0 \tag{51}$$

$$\overset{\text{Jensen's inequality}}{\geqslant} \int_{\mathbb{R}^n} \Big|\Big| \int_{[0,1]} v(t,X(t,x))dt \Big|\Big|_2^2 d\alpha_0 \tag{52}$$

$$= \int_{\mathbb{R}^n} \Big|\Big| \int_{[0,1]} \frac{\partial X(t,x)}{dt} dt \Big|\Big|_2^2 d\alpha_0 \tag{53}$$

$$= \int_{\mathbb{R}^n} \Big|\Big| X(1,x) - X(0,x) \Big|\Big|_2^2 d\alpha_0 \tag{54}$$

$$= \int_{\mathbb{R}^n} \Big|\Big| T_1(x) - X(0,x) \Big|\Big|_2^2 d\alpha_0 \tag{55}$$

$$\overset{\text{since Wasserstein distance is optimal}}{\geqslant} W_2^2(\alpha_0,\alpha_1) \tag{56}$$

$$\implies A_2^2(\alpha_0,\alpha_1) \geqslant W_2^2(\alpha_0,\alpha_1). \tag{57}$$

Now, let $X(t,x) = x + t(T^*(x) - x), v(t,x) = T^*(x) - x$, then the two inequalities become equality and we have the result. In particular, Jensen's inequality attains equality if the function to be integrated is independent of the variable integration (in other words, almost surely constant) Pratt (2014). Notice that $v(t,x) = T^*(x) - x$ is independent of time.

4 Coupling with the Hamilton-Jacobi equation

Consider the Lagrangian

$$\mathcal{L}(\rho,v;\phi,u) = \frac{1}{2}\iint_{[0,1]\times\Omega} ||v||_2^2 \rho dx dt + \iint_{[0,1]\times\Omega} (\partial_t\rho + \text{div}(\rho v))\phi dx dt - \iint_{[0,1]\times\Omega} \rho u dx dt. \tag{58}$$

This is a weak formulation of Lagrange multipliers, where the typical multiplication with a Lagrange multiplier is modified into an analog with a product with a compactly supported function, and $u(t,x) \geqslant 0$. Now, notice

$$= \frac{1}{2}\iint_{[0,1]\times\Omega} ||v||\rho dx dt + \iint_{\Omega\times[0,1]} \partial_t\rho\phi dt dx + \iint_{[0,1]\times\Omega} \text{div}(\rho v)\phi dx dt - \iint_{[0,1]\times\Omega} \rho u dx dt \tag{59}$$

$$\overset{\text{integration by parts}}{=} \frac{1}{2}\iint_{[0,1]\times\Omega} ||v||\rho dx dt + \int_\Omega (\phi\rho|_0^1 - \int_0^1 \rho\partial_t\phi)dx + \iint_{[0,1]\times\Omega} (-\rho v)\cdot\nabla\phi dx dt - \iint_{[0,1]\times\Omega} \rho u dx dt. \tag{60}$$

From standard Lagrange multiplier theory, we have a critical point at

$$\frac{\partial\mathcal{L}}{\partial\rho} = 0, \frac{\partial\mathcal{L}}{\partial v} = 0, \frac{\partial\mathcal{L}}{\partial\phi} = 0, \frac{\partial\mathcal{L}}{\partial u} = 0. \tag{61}$$

Taking such first variations, we will set these equal to zero. *Note:* in section 6.2, we saw a critical point to the optimization was done with respect to a constant, not zero. This was because our optimization in 6.2 was done w.r.t. ρ, so when the first variation was taken, this annihilated the term involving ρ, leaving the constant multiplier. In particular, this notion is consistent with the first variation is zero, since it is really just

$$\frac{\partial\mathcal{E}}{\partial\rho} - \lambda = 0, \tag{62}$$

before moving the constant to the other side. Returning to our proof, we get

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \rho} = \frac{1}{2}||v||_2^2 - \phi_t - v \cdot \nabla \phi - u = 0 \iff \frac{1}{2}||v||_2^2 - \phi_t - v \cdot \nabla \phi = u \geqslant 0 \\ \frac{\partial \mathcal{L}}{\partial v} = \rho v - \rho \nabla \phi = 0 \overset{\rho > 0}{\implies} v = \nabla \phi \\ \partial_t \rho + \operatorname{div}(\rho v) = 0 \\ \rho(0,x) = \rho_0, \rho(1,x) = \rho_1 \\ v \cdot n(t,x) = 0. \end{cases} \tag{63}$$

We will not extensively derive where all of these come from, but just know they are first variation calculations. If $\rho > 0$,

$$\begin{cases} \frac{1}{2}||\nabla \phi||_2^2 - \phi_t - \nabla \phi \cdot \nabla \phi = 0 \implies \boxed{\partial_t \phi + \frac{1}{2}||\nabla \phi||_2^2 = 0} \\ \partial_t \rho + \operatorname{div}(\rho \nabla \phi) = 0 \\ \rho(0,\cdot) = \rho_0 \\ \rho(1,\cdot) = \rho_1 \\ v \cdot n = 0, x \in \partial \Omega. \end{cases} \tag{64}$$

The boxed equation is the Hamilton-Jacobi equation. In general, a Hamilton-Jacobi equation is any equation of the form

$$-\partial_t S = H(x, \nabla_x S, t), \tag{65}$$

which is the form our equation that is boxed is.

Let us denote $m(t,x) = p(t,x)v(t,x)$. Our optimization problem is equivalently

$$\begin{cases} \min_{\rho,m} \frac{1}{2} \iint_{[0,1]\times\Omega} \frac{||m(t,x)||_2^2}{\rho(t,x)} dx dt \\ \text{such that} \\ \partial_t \rho(t,x) + \operatorname{div}(m(t,x)) = 0 \\ \rho(0,x) = \rho_0(x), \rho(1,x) = \rho_1(x) \\ m \cdot n(t,x) = 0, x \in \partial\Omega. \end{cases} \tag{66}$$

The integral equivalence is due to $||m||_2^2 = |\rho|||v||_2^2 = \rho||v||_2^2$, which is precisely the integrand of our original optimization (aside from the Lagrange multiplier terms). Note that last condition $m \cdot n$ is equivalent to $v \cdot n$ only because it is equal to zero. If it was nonzero, it would not be the same because the weighting ρ is a scalar multiplier that affects the weighting. The argument is similar to how a compactly supported test function has annihilating properties, because the density ρ can "scale v however across its domain." We remark ρ is not allowed to be anything, so the argument is different, but it has some similarities to compactly supported test arguments.

We will define the function

$$\theta(a,b) = \begin{cases} ||b||_2^2, /a \ a > 0 \\ 0, \ a = b = 0 \\ +\infty, a = 0, b \neq 0. \end{cases} \tag{67}$$

Note that θ is convex (it is basically a quadratic with discontinuities addressed) and lower semicontinuous (we will not give the exact definition, but not that it basically means "the lower half of all discontinuous parts contain all limit points"), i.e.

$$\liminf_{x \to x_0} f(x) \geqslant f(x_0). \tag{68}$$

For example, consider a jump discontinuity with an open point above and a closed point below. The lim inf is the lower point, and so is $f(x_0)$ in this scenario.

Let us consider a transformation

$$\theta : p(t,x)m(t,x) \in \mathbb{R}^d \to \theta(p(t,x), m(t,x)) \in \mathbb{R}_{\geqslant 0}. \tag{69}$$

Due to the definition of θ, we arrive at the same optimization problem of 66 but rewritten with the double integral

$$\min_{\rho,m} \frac{1}{2} \iint_{[0,1]\times\Omega} \theta(p(t,x), m(t,x))dxdt. \tag{70}$$

I will skip the proof that θ is lower semicontinuous and convex. This is more or less an exercise in analysis and let us remain focused in more OT/PDEs areas for now. Since θ is convex, it can be noted that any optimization problem of the form

$$\min_{\rho_1,\dots,\rho_n} \iint \dots \int \Theta(\rho_1(t,x), \dots, \rho_n(t,x))dxdt \tag{71}$$

has local minima are global minima if Θ is convex. It does not imply uniqueness, and see Gigili (2016) for a basic $\mathbb{R} \to \mathbb{R}$ example with infinitely many solutions (but they are all the same minima!).

4.1 Proximal gradient descent

Let us discuss proximal gradient descent and its qualities before returning to our optimization discussion. Proximal gradient descent is the iterative algorithm

$$x^{k+1} = \text{Prox}_{\tau g}\Big(x^k - \tau\nabla f(x^k)\Big), \qquad \text{Prox}_{\tau g}\Big(x\Big) := \arg\min_u \Big\{\tau g(u) + \frac{1}{2}||u-x||_2^2\Big\}. \tag{72}$$

The goal of this method is that is satisfies the optimization problem

$$\min_x f(x) + g(x) \tag{73}$$

where

- f is convex, smooth with Lipschitz constant L
- g convex but not necessarily smooth .

We say u_x is in the subdifferential $u_x \in \partial g(x)$ if

$$g(y) \geqslant g(x) + \langle u_x, y-x\rangle \ \forall y. \tag{74}$$

A subdifferential is an extension of the derivative for any function which is not typically differentiable but convex. For example, the subdifferential of $f(x) = |x|$ at $x = 0$ is $[-1, 1]$, meaning these are all possible "slopes" at $x = 0$ satisfying the subdifferential criterion. It can be noted ∂g is monotone, i.e. $\langle u_x - u_y, x-y\rangle \geqslant 0$, $u_x \in \partial g(x), u_y \in \partial g(y)$. The proof of this following from

$$g(y) \geqslant g(x) + \langle u_x, x-y\rangle \tag{75}$$
$$g(x) \geqslant g(y) + \langle u_y, y-x\rangle \tag{76}$$
$$\implies g(y) + g(x) + \leqslant g(x) + g(y)\langle u_x - u_y, x-y\rangle \tag{77}$$
$$\implies 0 \leqslant \langle u_x - u_y, x-y\rangle. \tag{78}$$

Here, we have swapped the sign of $y - x$ by adding the second equation.

Now we will Prox is contractive. Note: contractive properties are very useful in optimization analysis because a contractive property can be used to prove something is convergent, i.e. $||f(x^{k+1}) - f(x^k)|| \leqslant c^k||x_0 - x^k||$ where $c < 1$. Clearly the right-hand side will tend to 0 if $c < 1$ and the norm is finite. We will prove

$$||\text{Prox}_g(x) - \text{Prox}_g(y)|| \leqslant ||x-y||. \tag{79}$$

Proof. Consider

$$x' = \text{Prox}_g(x) = \min_u g(u) + \frac{1}{2}||u-x||_2^2 \tag{80}$$
$$\implies 0 \in \partial g(x) + x' - x = \{y + (x'-x) : y \in \partial g(x)\} \tag{81}$$
$$\implies x - x' \in \partial g(x'). \tag{82}$$

Similarly, we can repeat this argument and get $y - y' \in \partial g(y')$. It follows from the monotone property

$$\langle u_{x,u_x \in \partial g(x')} - u_{y,u_y \in \partial g(y')}, x' - y' \rangle \geqslant 0 \tag{83}$$

$$\implies \langle (x - x') - (y - y'), x' - y' \rangle \geqslant 0 \tag{84}$$

$$\implies ||x - y||||x' - y'|| \overset{\text{Cauchy-Schwarz}}{\geqslant} \langle x - y, x' - y' \rangle \geqslant ||x' - y'||_2^2 \tag{85}$$

$$\implies ||x - y|| \geqslant ||x' - y'|| = ||\text{Prox}_g(x) - \text{Prox}_g(y)||, \tag{86}$$

giving the result.

□

Returning to proximal gradient descent,

$$x^{k+1} = \text{Prox}_{\tau g}(x^k - \tau \nabla f(x^k)) \tag{87}$$

$$= \text{Prox}_{\tau g}\Big[(I - \tau \nabla f) \circ (x^k)\Big]. \tag{88}$$

We use the $\circ$ notation to emphasize the above is well-defined, since I is a matrix and ∇f is a vector. Observe

$$x^* = \arg\min_x f(x) + g(x) \tag{89}$$

$$\implies 0 \in \tau \nabla f(x^*) + \tau \partial g(x^*) = \{\tau u_x + \tau \nabla f(x^*) : g(x^*) \geqslant \langle u_x, x^* - x \rangle\} \tag{90}$$

$$\in \tau \nabla f(x^*) + \tau(\text{anything in the subdifferential}) \tag{91}$$

$$\implies x^* - \tau \nabla f(x^*) \in x^* + \tau \partial g(x^*) \tag{92}$$

$$\implies (I - \tau \nabla f(x))(x^*) \in (I + \tau \partial g)(x^*) \tag{93}$$

$$\implies x^* \in \underbrace{(I + \tau \partial g)^{-1}}_{=\text{Prox}_{\tau g}} \circ (I - \tau \nabla f)(x^*). \tag{94}$$

In particular, this shows the fixed point of proximal gradient descent yields the minimizer. Note that convergence behavior depends on the Jacobian

$$J\Big((I - \tau \nabla f)(x)\Big). \tag{95}$$

Example 1. Consider the example $g(x) = |x| := \sum_i |x_i|$. Also,

$$\min_x \frac{1}{2}||Ax - b||^2 + \mu|x| \tag{96}$$

$$\text{Prox}_{\tau g}(y) = \arg\min_x \tau\mu|x| + \frac{1}{2}||x - y||^2 \tag{97}$$

$$= \left\{ \text{sign}(y_i) \cdot \max\Big\{|y_i| - \tau\mu, 0\Big\} \right\}_i. \tag{98}$$

This is the soft thresolding operator, and can be found by taking the partial ∂_i with respect to $\tau\mu x_i + \frac{1}{2}(x_i - y_i)^2$. Hence,

$$x^{k+1} = \text{Prox}_{\tau g}\Big(x^k - \underbrace{\tau A^T(Ax^k - b)}_{=\tau \nabla f(x) = \tau \nabla \frac{1}{2}||Ax - b||^2}\Big). \tag{99}$$

Putting them together, we get

$$x^{k+1} = \left\{ \text{sign}\Big(x^k - \tau A^T(Ax^k - b)\Big)_i \cdot \max\Big\{\Big|\Big(x^k - \tau A^T(Ax^k - b)\Big)_i\Big| - \tau\mu, 0\Big\} \right\}_i \tag{100}$$

Example 2. Consider

$$\min_x \frac{1}{2}||Ax - b||^2 \text{ such that } x \in \Omega. \tag{101}$$

Here, $\Omega \subseteq \mathbb{R}^d$ is convex. Let us define

$$\chi_\Omega(x) = \begin{cases} \infty, x \notin \Omega \\ 0, x \in \Omega. \end{cases} \tag{102}$$

Our optimization problem can be rewritten to ensure the function is infinite off the domain with

$$\min_{x \in \mathbb{R}^d} \underbrace{\frac{1}{2}||Ax-b||^2}_{f(x)} + \underbrace{\chi_\Omega(x)}_{g(x)}. \tag{103}$$

We can check these functions satisfy the criteria for f, g established earlier. Now,

$$\text{Prox}_{\tau g}(y) = \arg\min_x \chi_\Omega(x) + \frac{1}{2}||x-y||^2 \tag{104}$$

$$= \arg\min_{x \in \Omega} \frac{1}{2}||x-y||^2 \tag{105}$$

$$= \text{Proj}_\Omega(y), \tag{106}$$

and so

$$x^{k+1} = \text{Proj}_\Omega(x^k - A^T(Ax^k - b)). \tag{107}$$

4.2 Proximal algorithms in OT

Let us return to the OT problem we saw earlier. We established the problem

$$\begin{cases} \min_{\rho,m} \frac{1}{2} \iint_{[0,1]\times\Omega} (||m||^2/\rho) dxdt \\ \text{such that} \\ C(\rho_0, \rho_1) := \begin{cases} \partial_t \rho + \text{div}(m) = 0 \\ \rho(0,x) = \rho_0 \\ \rho(1,x) = \rho_1 \\ m \cdot n = 0, x \in \partial\Omega. \end{cases} \end{cases} \tag{108}$$

Subject to constraints, our OT problem can be written using a similar χ function,

$$\min_{\rho,m} \frac{1}{2} \iint_{[0,1]\times\Omega} \frac{||m||^2}{\rho_{\geqslant 0}} dxdt + \chi_{C(\rho_0,\rho_1)}(m,\rho). \tag{109}$$

Let us begin by simplifying this example. Let us consider a linearly constrained problem

$$\min_x f(x) + \chi_{Ax=b}(x) \iff \min_{Ax=b} f(x). \tag{110}$$

Applying proximal gradient descent (recall gradient descent will attempt to find a minima),

$$x^{k+1} = \text{Prox}_{\tau\chi_{Ax=b}}(\underbrace{x^k - \tau\nabla f(x^k)}_{=y^k}) \tag{111}$$

$$= \arg\min_u \tau\chi_{Au=b}(u) + \frac{1}{2}||u-y^k||_2^2 \tag{112}$$

$$= \arg\min_u \frac{1}{2}||u-y^k||_2^2 \text{ such that } Au = b. \tag{113}$$

Let us consider the method of Lagrange multipliers,

$$\mathcal{L}(u;\phi) = \frac{1}{2}||u-y^k||_2^2 + \langle b - Au, \phi\rangle \tag{114}$$

$$0 = \frac{\partial \mathcal{L}}{\partial u} = u - y^k - A^T\phi \tag{115}$$

$$0\frac{\partial \mathcal{L}}{\partial \phi} = b - Au. \tag{116}$$

Let us multiply the first equation equal to zero by A, and we will substitute in the second. As a remark, we assume A is wide so that AA^T can have an inverse (if it is tall, it cannot, since it would involve a mapping from a low to high dimension which can never be full rank for linear transformations). Notice

$$\underbrace{Au}_{=b} - Ay^k - AA^T\phi = 0 \tag{117}$$

$$\implies \phi = (AA^T)^{-1}(b - Ay^k) \tag{118}$$

$$\implies Au = Ay^k + AA^T(AA^T)^{-1}(b - Ay^k) \tag{119}$$

$$\implies u = y^k + A^T(AA^T)^{-1}(b - Ay^k). \tag{120}$$

In particular, "removing A" is justified. Note that A is not injective as it has nontrivial kernel since A is wide. Note that we do not have surjective is possible, since A maps from a high to low (or equal) dimension. The reason "removing A" is justified is by use of the first variation condition $u - y^k - A^T\phi$. Note in general taking away A cannot always be done without injectivity.

5 JKO conditions with continuity equation equivalence

We show the JKO conditions as in Lee et al. (2023) are equivalent to a continuity equation.

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

$$\frac{d}{dt}\rho(t, T(t, x)) = \partial_t\rho + \nabla\rho \cdot \partial_t T(t, x). \tag{121}$$

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

$$\frac{d}{dt}\rho = \partial_t\rho + \nabla\rho \cdot v. \tag{122}$$

Now, consider the following:

$$\frac{d}{dt}\rho(t, T(t, x)) = \frac{d}{dt}\frac{\rho^k(x)}{\det|J(t, x)|} = -\frac{\rho^k(x)}{(\det|J(t, x)|)^2} \cdot \partial_t \det|J(t, x)| \tag{123}$$

$$= -\frac{\rho^k(x)}{\det|J(t, x)|} \cdot \frac{\partial_t \det|J(t, x)|}{\det|J(t, x)|} = -\frac{\rho^k(x)}{\det|J(t, x)|}\partial_t \log\det|J(t, x)| \tag{124}$$

$$= -\rho(t, T(t, x))\partial_t \log\det|J(t, x)| = -\rho(t, T(t, x))\mathrm{div}(v). \tag{125}$$

Thus,

$$-\rho(t, T(t, x))\mathrm{div}(v) = \partial_t\rho + \nabla\rho \cdot v \implies \partial_t\rho + \nabla \cdot (\rho v) = 0, \tag{126}$$

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

6 Derivation of the JKO algorithm

6.1 The Euclidean case

First, we discuss the Euclidean case.

Consider the optimization problem

$$\min_x \left\{ \frac{1}{2\tau}||x - x_k||_2^2 + \mathcal{E}(x) \right\}. \tag{127}$$

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

$$\nabla\left[\frac{1}{2\tau}||x - x_k||_2^2 + \mathcal{E}(x)\right] = \frac{1}{\tau}(x - x_k) + \nabla\mathcal{E}(x) = 0. \tag{128}$$

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

$$\nabla\langle x - x_k, x - x_k\rangle = [\partial_j(\sum_i (x_i - x_{k,i})^2)]_j = [2(x_j - x_{k,j})]_j. \tag{129}$$

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

$$\frac{x - x_k}{\tau} - \nabla\mathcal{E}(x), \tag{130}$$

which is exactly the Euler scheme for the gradient flow equation

$$\partial_t x(t) = -\nabla\mathcal{E}(x(t)). \tag{131}$$

6.2 The Wasserstein case

The Wasserstein gradient flow PDE is

$$\partial_t \rho = \nabla \cdot (\rho\nabla(\mathcal{E}'(\rho))). \tag{132}$$

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

$$\min_\rho \left\{ \frac{1}{2\tau} \int |x - T(x)| d\rho + \mathcal{E}(\rho) \right\}. \tag{133}$$

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

$$\delta(\frac{1}{2\tau} \int |x - T(x)| d\rho + \mathcal{E}(\rho)) = \text{constant}. \tag{134}$$

The reason it is constant and not zero is because ρ is constrained, so the constant follows from the method of Lagrange multipliers since the first variation of

$$\delta(\lambda(\int \rho - 1)) = \lambda. \tag{135}$$

It is known $T(x) = x - \nabla\phi(x)$ for some convex ϕ Mokrov et al. (2021)Bao and Zhang (2022) (page 4) Ansari. It is also known the first variation of the Wasserstein distance is Ansari

$$\delta(\frac{1}{2\tau} W_2^2(\rho, \rho_k)) = \frac{1}{\tau}\phi. \tag{136}$$

Thus,

$$\frac{1}{\tau}\phi(x) + \mathcal{E}'(\rho_k)(x) = \text{constant} \implies \frac{1}{\tau}\nabla\phi(x) + \nabla\mathcal{E}'(\rho_k)(x) = 0. \tag{137}$$

Equivalenty,

$$\frac{1}{\tau}(x - T(x)) = -\nabla\mathcal{E}'(\rho_k)(x). \tag{138}$$

In the continuum limit, we get

$$v = -\nabla\mathcal{E}'(\rho_k), \tag{139}$$

which yields our Wasserstein gradient flow when substituted into 132.

7 Wasserstein geometry

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

$$g|_p : \mathrm{Tan}_p M \times \mathrm{Tan}_p M \to \mathbb{R}^*, g_{ij} = g(\partial_{x^i}|_p, \partial_{x^j}|_p). \tag{140}$$

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 Takatsu

$$\mathcal{P}_2(M) = \left\{ \mu : \text{Borel sets on } M \to \mathbb{R}^+ \cup \{0\} : \int_M d(x, x_0)^2 d\mu(x) < \infty, x_0 \in M, \int_M d\mu(x) = 1 \right\}. \tag{141}$$

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 Ay (2024) Craig and contributors

$$\left\langle \frac{\partial \phi}{\partial t_1}, \frac{\partial \psi}{\partial t_1} \right\rangle_\mu = g_W(\nabla\phi, \nabla\psi) = \int_M \langle \nabla\phi, \nabla\psi \rangle d\mu(x). \tag{142}$$

Recall call that an induced metric is an inner product of tangent vectors, so the above formulation is consistent with such an inner product definition. Here, ϕ, ψ are solutions to a continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t \nabla \Psi) = 0. \tag{143}$$

The inner product is in such a way that ϕ, ψ are tangent vectors located at μ. Such holds in the sense of distributions Ambrosio et al. (2008)

$$\iint_{[0,1]\times\mathbb{R}^d} (\partial_t \varphi + \langle v_t, \nabla\varphi \rangle) d\mu_t dt = 0, \quad \varphi \text{ is a suitable compactly supported test function,} \tag{144}$$

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). Note that the tangent bundle is

$$\mathrm{Tan}_\mu \mathcal{P}(M) := \overline{\left\{ \nabla\varphi : \varphi \in \mathcal{P}_2(M) \right\}}^{L^2(\mu, X)}. \tag{145}$$

Also, our vector field is such that

$$v_t \in \left\{ \nabla\Psi : \Psi \text{ is a plausible potential} \right\} = \Gamma, \Gamma^* = \overline{\left\{ \nabla\Psi : \Psi \in \mathcal{C}(M) \right\}}^{L^p(\mu; X)}. \tag{146}$$

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 Ambrosio et al. (2008), chapter 8 for discussion of this. In particular, it is known (page Ambrosio et al. (2008)

$$v_t \in \overline{\left\{ j_q(\nabla\Psi) : \Psi \in \mathrm{Cyl}(X) \right\}}^{L_p(\mu_t; X)} \tag{147}$$

and j_q is the map $j_q(v) = ||v||_2^{q-2} v$. Here Cyl denotes the cylindrical class of functions

References

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.

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

Nihat Ay. Information geometry of the otto metric. *Information Geometry*, 2024. doi: 10.1007/s41884-024-00149-w. URL https://doi.org/10.1007/s41884-024-00149-w.

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

Katy Craig and contributors. Formal riemannian structure of the wasserstein metric. URL https://www.otwiki.xyz/RiemannianStructureOriginal.html.

Gigili. The solution of a convex optimization problem is unique, 2016. URL https://math.stackexchange.com/questions/1925527/the-solution-of-a-convex-optimization-problem-is-unique.

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.

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

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.

Rob Pratt. When jensen's inequality is equality, 2014. URL https://math.stackexchange.com/questions/628386/when-jensens-inequality-is-equality. Math Stack Exchange.

Asuka Takatsu. Wasserstein geometry of gaussian measures. URL https://www2.sonycsl.co.jp/person/nielsen/infogeo/Seminar/gauss-Takatsu.pdf.

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