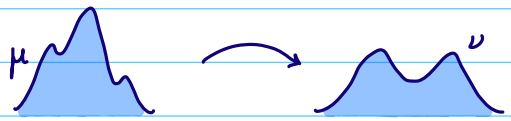
LEC. 1: WASSERSTEIN GRADIENT FLOWS

1. Basics of optimal transport

DEF. Given two probability measures μ, ν over $\mathbb{R}^d$ with finite second moment (written $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$), the 2-Wasserstein distance between μ and ν is defined to be:

$$W_2^2(\mu,\nu) := \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \|x-y\|^2 \, \gamma(dx,dy)$$

where $\mathcal{C}(\mu,\nu)$ is the set of couplings of μ and ν.



FACTS:
- For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, an optimal coupling γ always exists (not unique)
- $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space
- $W_2(\mu_n, \mu) \to 0 \iff \mu_n \to \mu$ weakly and $\int \|\cdot\|^2 d\mu_n \to \int \|\cdot\|^2 d\mu$
- $x \mapsto \delta_x$ is an isometric embedding (i.e., $W_2(\delta_x, \delta_y) = \|x-y\|$)

FUNDAMENTAL THM. OF OPTIMAL TRANSPORT

Assume μ has a density w.r.t. Lebesgue measure (written $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$). Then, the optimal transport plan between μ and ν is induced by a transport map:

$$W_2^2(\mu,\nu) = \int \|x-y\|^2 \gamma^*(dx,dy) = \int \|x - T^*(x)\|^2 \mu(dx)$$

$$\nu = T^*_{\#}\mu \iff \text{if } X \sim \mu, \text{ then } T^*(X) \sim \nu$$

Moreover, T^* is characterized as the (μ-a.e.) unique gradient of a cvx fn which pushes μ to ν. We call T^* the Brenier map.

EXAMPLES:
- $T_{\#}\delta_x = \delta_{T(x)}$
- If $\mu = N(m_1, \Sigma_1)$ and $\nu = N(m_2, \Sigma_2)$, then
  $T^*(x) = m_2 + \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}(x-m_1)$ (exercise)

2. Curves of measures

Curve in $\mathbb{R}^d$

$\dot{x}_t$ is the tangent vector to the curve

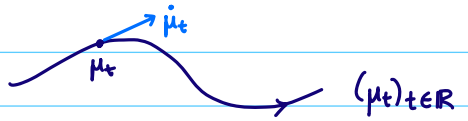
Think of $(x_t)_{t\in\mathbb{R}}$ as the trajectory of a particle

x_t = position at time t

$\dot{x}_t$ = velocity at time t

"LAGRANGIAN" perspective

Curve in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$



what is $\dot{\mu}_t$?

Think of μ_t as a collection of particles

$\mu_t(x)$ = mass density at time t, location x

$v_t(x)$ = velocity at time t, location x

v_t is the velocity vector field

"EULERIAN" perspective

Joining the two perspectives:

Now think of X_t as a random variable, $X_t \sim \mu_t$

Then, $\dot{X}_t = v_t(X_t)$

PROP. If $X_0 \sim \mu_0$ and $\dot{X}_t = v_t(X_t)$, then $\mu_t = \text{law}(X_t)$ satisfies

$$\partial_t \mu_t + \text{div}(\mu_t v_t) = 0 \qquad \text{(continuity eq.)}$$

Proof. The "weak" form of the PDE is:

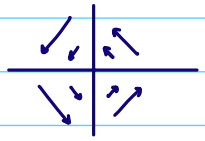
$$\partial_t \int \varphi \, d\mu_t = \int \langle \nabla\varphi, v_t \rangle \, d\mu_t, \qquad \forall \varphi \in C_c^\infty(\mathbb{R}^d)$$

But $\partial_t \int \varphi \, d\mu_t = \partial_t \mathbb{E}\varphi(X_t) = \mathbb{E}\langle \nabla\varphi(X_t), v_t(X_t)\rangle = \int \langle \nabla\varphi, v_t\rangle \, d\mu_t$. □

Q: Can we think of v_t as the tangent vector at μ_t?

Non-uniqueness: $\mu_t = N(0, I)$ for all $t \in \mathbb{R}$

$v_t = 0$ vs. v_t = rotation



Selection criterion: choose v_t to minimize the kinetic energy

$$v_t = \underset{v:\mathbb{R}^d \to \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2} \int \|v\|^2 \, d\mu_t : \quad \operatorname{div}(\mu_t v_t) = -\partial_t \mu_t \right\} \tag{*}$$

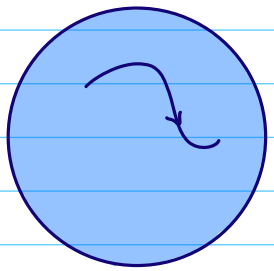
PROP. There is a unique v_t satisfying (*). It is characterized by "$v_t = \nabla \psi_t$" for some $\psi_t : \mathbb{R}^d \to \mathbb{R}$.
(More precisely, $v_t = \lim_{k \to \infty} \nabla \psi_t^{(k)}$, where $\psi_t^{(k)} \in C_c^\infty(\mathbb{R}^d)$.)

Proof. Uniqueness follows from strict cvxty of (*).
Then, note that the constraint in (*) is

$$\int \langle \nabla \varphi, v_t \rangle \, d\mu_t = \partial_t \int \varphi \, d\mu_t, \quad \forall \varphi \in C_c^\infty(\mathbb{R}^d)$$

If we replace v_t by its projection onto $\overline{\{\nabla \varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu_t)}$
then it still satisfies the constraint but has smaller norm. □

3. Elements of Riemannian geometry



A Riemannian manifold M has, at each point $p \in M$, an associated vector space — the tangent space T_pM — equipped with an inner product $\langle \cdot, \cdot \rangle_p$.

The inner product gives rise to distances:

$$d(p_0, p_1)^2 := \inf \left\{ \int_0^1 \|\dot{p}_t\|_{p_t}^2 \, dt : \text{curves } p : [0,1] \to M \text{ joining } p_0 \text{ to } p_1 \right\}$$

A minimizer $(p_t)_{t \in [0,1]}$ to this variational problem is a (const.-speed) geodesic or shortest path.

The exponential map $\exp_p : T_pM \to M$ maps $v \in T_pM$ to the endpoint of the geodesic leaving p w/ velocity v for unit time.
The logarithmic map $\log_p : M \to T_pM$ is the inverse mapping.

Given $f: \mathcal{M} \to \mathbb{R}$, the gradient $\nabla f(p)$ at $p \in \mathcal{M}$ is the element of $T_p\mathcal{M}$ s.t. for all curves $(p_t)_{t \in \mathbb{R}}$ with $p_0 = p$,

$$\partial_t f(p_t)\big|_{t=0} = \langle \nabla f(p), \dot{p}_0 \rangle_p .$$

The Hessian $\nabla^2 f(p)$ is such that for all geodesics $(p_t)_{t \in \mathbb{R}}$,

$$\partial_t^2 f(p_t)\big|_{t=0} = \nabla^2 f(p)[\dot{p}_0, \dot{p}_0] .$$

A set $\mathcal{C} \subseteq \mathcal{M}$ is (geodesically) convex if for all $p, q \in \mathcal{C}$, the geodesic joining p to q lies in $\mathcal{C}$.

The function $f: \mathcal{M} \to \mathbb{R}$ is α-convex if one of the following holds:

- for all geodesics $(p_t)_{t \in [0,1]}$,
$$f(p_t) \leq (1-t) f(p_0) + t f(p_0) - \frac{\alpha t (1-t)}{2} d(p_0, p_1)^2$$
- for all $q \in \mathcal{M}$,
$$f(q) \geq f(p) + \langle \nabla f(p), \log_p(q) \rangle_p + \frac{\alpha}{2} d(p, q)^2$$
- for all $p \in \mathcal{M}$, $v \in T_p\mathcal{M}$,
$$\nabla^2 f(p)[v, v] \geq \alpha \|v\|_p^2$$

4. Otto calculus

We now define the tangent space

$$T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) := \overline{\{\nabla \psi : \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}$$

equipped with the $L^2(\mu)$ norm: $\|\nabla \psi\|_\mu^2 := \int \|\nabla \psi\|^2 d\mu$

Q: What are the geodesics?

THM. The Benamou–Brenier formula holds:

$$W_2^2(\mu_0,\mu_1) = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t}^2 \, dt : (\mu, v) \text{ solve the cont. eq.} \right\}$$

The geodesic joining μ_0 to μ_1 is described as follows:

$$\mu_t = ((1-t)\,\mathrm{id} + t\nabla\varphi)_{\#}\mu_0, \quad \nabla\varphi = \text{Brenier map}$$

Proof. Let (μ, v) solve the cont. eq. Then,

$$\int_0^1 \|v_t\|_{\mu_t}^2 dt = \int_0^1 \mathbb{E}\|\dot{X}_t\|^2 dt \quad (\dot{X}_t = v_t(X_t))$$

$$\geq \mathbb{E}\left\| \int_0^1 \dot{X}_t \, dt \right\|^2 \quad (\text{equality if } t \mapsto \dot{X}_t \text{ is const.})$$

$$= \mathbb{E}\|X_1 - X_0\|^2$$

$$\geq W_2^2(\mu_0,\mu_1). \quad (\text{equality if } X_1 = \nabla\varphi(X_0)) \qquad \square$$

Hence, we have obtained a Riemannian structure associated with the 2-Wasserstein distance.

$$\exp_\mu(\nabla\psi) = (\mathrm{id} + \nabla\psi)_{\#}\mu$$
$$\log_\mu(\nu) = \nabla\varphi_{\mu\to\nu} - \mathrm{id}$$

Now, let $\mathcal{F}: \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ be a functional. What is the gradient?

DEF. The first variation of $\mathcal{F}$ at μ is a function $\delta\mathcal{F}(\mu): \mathbb{R}^d \to \mathbb{R}$ s.t. for all χ w/ $\int d\chi = 0$ and $\mu + \varepsilon\chi \in \mathcal{P}(\mathbb{R}^d)$ for sufficiently small $\varepsilon > 0$,

$$\lim_{\varepsilon \searrow 0} \frac{\mathcal{F}(\mu+\varepsilon\chi) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta\mathcal{F}(\mu)\, d\chi$$

(defined up to an additive const.)

How to compute $\delta\mathcal{F}(\mu)$: let $(\mu_t)_{t\in\mathbb{R}}$ be a curve, $\mu_0 = \mu$.

Compute $\partial_t \mathcal{F}(\mu_t)$ and write it in the form

$$\partial_t \mathcal{F}(\mu_t)\big|_{t=0} = \int (\cdots)\, \partial_t \mu_t\big|_{t=0} \Rightarrow (\cdots) = \delta\mathcal{F}(\mu)$$

THM. The Wasserstein gradient $\nabla_{W_2} \mathcal{F}(\mu)$ of $\mathcal{F}$ at μ is the vector field

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu).$$

Proof. By defn., $\nabla_{W_2} \mathcal{F}(\mu)$ satisfies:

- for $\partial_t \mu_t + \operatorname{div}(\mu_t \nabla \psi_t) = 0$,

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{W_2} \mathcal{F}(\mu_t), \nabla \psi_t \rangle_{\mu_t}.$$

- $\nabla_{W_2} \mathcal{F}(\mu) \in T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$ $\Leftrightarrow$ $\nabla_{W_2} \mathcal{F}(\mu)$ is a grad. vector field.

Then, note that

$$\partial_t \mathcal{F}(\mu_t) = \int \delta \mathcal{F}(\mu_t) \, \partial_t \mu_t = -\int \delta \mathcal{F}(\mu_t) \operatorname{div}(\mu_t \nabla \psi_t)$$

$$= \int \langle \nabla \delta \mathcal{F}(\mu_t), \nabla \psi_t \rangle \, d\mu_t = \langle \nabla \delta \mathcal{F}(\mu_t), \nabla \psi_t \rangle_{\mu_t}. \qquad \square$$

COR. The Wasserstein gradient flow of $\mathcal{F}$ satisfies

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \delta \mathcal{F}(\mu_t)) \quad \Leftrightarrow \quad \dot{X}_t = -\nabla \delta \mathcal{F}(\mu_t)(X_t)$$