CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 29, 2021

Yale University

Last time

- Wasserstein distance
- Optimal transport
- Displacement interpolation as geodesic

Today: Otto calculus

References

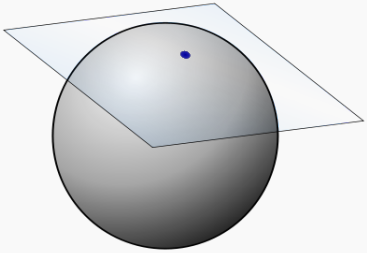
- Villani, *Topics in Optimal Transportation*, Springer, 2003
- Villani, *Optimal Transport: Old and New*, Springer, 2008
- Ambrosio, Gigli & Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Springer, 2005

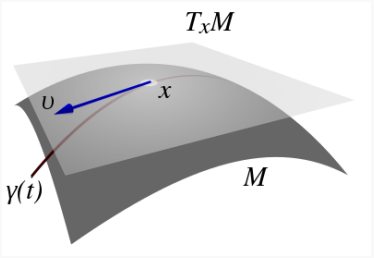
Motivation

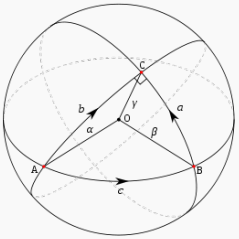
Review: Distance on manifold

Given a manifold $\mathcal{X}$ with *metric* $g_x \colon \mathrm{T}_x\mathcal{X} \times \mathrm{T}_x\mathcal{X} \to \mathbb{R}$

$$\|v\|_x^2 = \langle v, v \rangle_x = g_x(v, v)$$

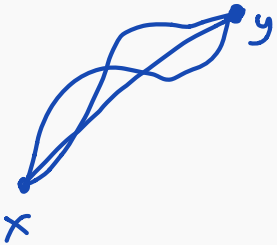






Distance between two points $x, y \in \mathcal{X}$ is

$$d(x, y)^2 = \inf_{\substack{(X_t) \\ X_0 = x,\, X_1 = y}} \int_0^1 \|\dot{X}_t\|_{X_t}^2 \, dt$$



Shortest curve is **geodesic**: $d(x, X_t) = t\, d(x, y)$

Wasserstein distance

$$\mathcal{P}(\mathbb{R}^n) = \left\{ \text{probability distributions } \rho \text{ on } \mathbb{R}^n \text{ with } \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$$

Wasserstein distance:

$$W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|^2]$$

where $\Pi(\rho, \nu)$ is set of couplings $(X, Y) \sim \pi$ such that $X \sim \rho$, $Y \sim \nu$

Wasserstein distance

$\mathcal{P}(\mathbb{R}^n) = \left\{ \text{probability distributions } \rho \text{ on } \mathbb{R}^n \text{ with } \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$

Wasserstein distance:

$$W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|^2]$$

where $\Pi(\rho, \nu)$ is set of couplings $(X, Y) \sim \pi$ such that $X \sim \rho$, $Y \sim \nu$

Brenier: There is unique convex $\phi \colon \mathbb{R}^n \to \mathbb{R}$ s.t. $(\nabla \phi)_{\#} \rho = \nu$ and

$$W_2(\rho, \nu)^2 = \mathbb{E}_\rho[\|X - \nabla \phi(X)\|^2]$$

Wasserstein distance

$\mathcal{P}(\mathbb{R}^n) = \left\{ \text{probability distributions } \rho \text{ on } \mathbb{R}^n \text{ with } \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$

Wasserstein distance:

$$W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|^2]$$

where $\Pi(\rho, \nu)$ is set of couplings $(X, Y) \sim \pi$ such that $X \sim \rho$, $Y \sim \nu$

Brenier: There is unique convex $\phi \colon \mathbb{R}^n \to \mathbb{R}$ s.t. $(\nabla \phi)_{\#}\rho = \nu$ and

$$W_2(\rho, \nu)^2 = \mathbb{E}_\rho[\|X - \nabla \phi(X)\|^2]$$

McCann: *Geodesic* is displacement interpolation $\rho_t = (T_t)_{\#}\rho$ induced by linear interpolation $T_t(x) = (1 - t)x + t\,\nabla \phi(x)$

From distance to metric

Wasserstein W_2 distance:

$$W_2(\rho,\nu)^2 = \inf_{\pi\in\Pi(\rho,\nu)} \mathbb{E}_\pi[\|X-Y\|^2]$$

Question: What metric $\|\phi\|_\rho^2$ for $\phi \in \mathsf{T}_\rho\mathcal{P}$ generates W_2 distance?

$$W_2(\rho,\nu)^2 = \inf_{\substack{(\rho_t)\\ \rho_0=\rho,\,\rho_1=\nu}} \int_0^1 \|\dot\rho_t\|_{\rho_t}^2\, dt$$

Wasserstein metric

Tangent vector $\phi \in \mathsf{T}_\rho \mathcal{P}$ is $\phi = -\nabla \cdot (\rho \nabla u)$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Wasserstein metric:

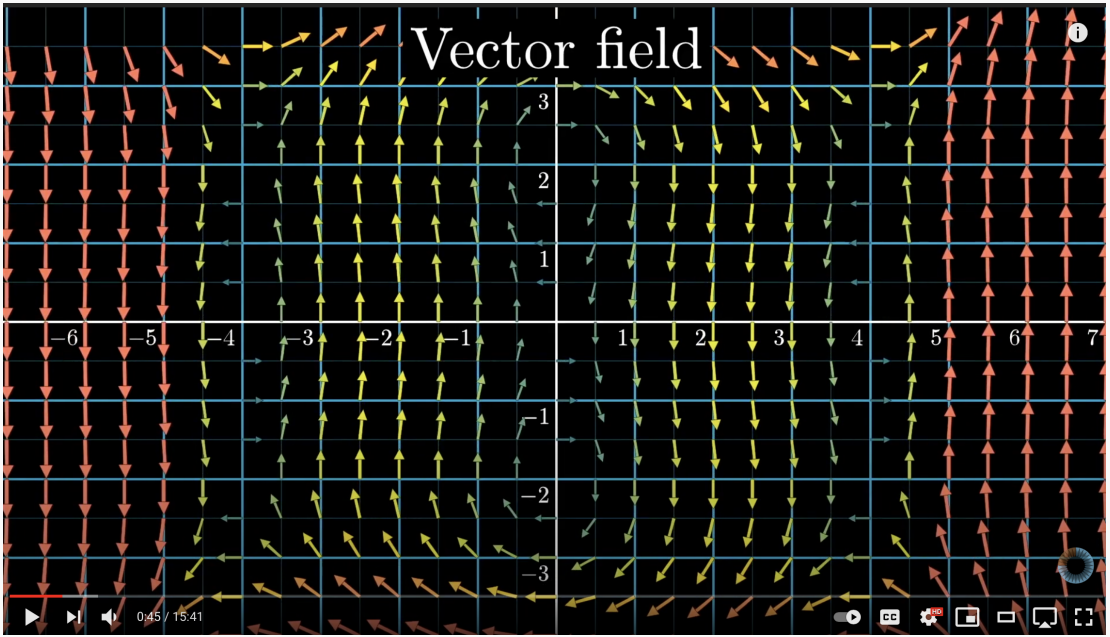
$$\|\phi\|_\rho^2 = \int_{\mathbb{R}^n} \|\nabla u(x)\|^2 \, \rho(x) \, dx = \mathbb{E}_\rho[\|\nabla u(X)\|^2]$$

Theorem: Wasserstein metric generates W_2 distance

$$W_2(\rho, \nu)^2 = \inf_{\substack{(\rho_t) \\ \rho_0 = \rho,\, \rho_1 = \nu}} \int_0^1 \|\dot{\rho}_t\|_{\rho_t}^2 \, dt$$

Divergence

Divergence



3Blue1Brown, *Divergence and curl: The language of Maxwell's equations, fluid flow, and more*, https://www.youtube.com/watch?v=rB83DpBJQsE, 2018

Divergence

Given a vector field $v : \mathbb{R}^n \to \mathbb{R}^n$

$$v(x) = (v_1(x), \ldots, v_n(x))$$

The **divergence** $\nabla \cdot v : \mathbb{R}^n \to \mathbb{R}$ is

$$(\nabla \cdot v)(x) = \sum_{i=1}^{n} \frac{\partial v_i(x)}{\partial x_i}$$

Divergence

Given a vector field $v : \mathbb{R}^n \to \mathbb{R}^n$

$$v(x) = (v_1(x), \ldots, v_n(x))$$

The **divergence** $\nabla \cdot v : \mathbb{R}^n \to \mathbb{R}$ is

$$(\nabla \cdot v)(x) = \sum_{i=1}^{n} \frac{\partial v_i(x)}{\partial x_i}$$

- Measures net change in flow defined by v
- $\nabla \cdot v = \langle \nabla, v \rangle$ where $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n})$
- $\nabla \cdot v = \mathrm{Tr}(\nabla v)$ where $\nabla v(x)$ is Jacobian matrix

$$\left(\frac{\partial v_i(x)}{\partial x_j} \right)_{i,j=1}^{n}$$

Example: Laplacian

Suppose v is the **gradient** of $f : \mathbb{R}^n \to \mathbb{R}$

$$v(x) = \nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

The Jacobian is the **Hessian** of f

$$\nabla(\nabla f)(x) = \nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \, \partial x_j} \right)_{i,j=1}^{n}$$

The divergence is the **Laplacian** of f

$$(\nabla \cdot \nabla f)(x) = \Delta f(x) = \mathrm{Tr}(\nabla^2 f(x)) = \sum_{i=1}^{n} \frac{\partial^2 f(x)}{\partial x_i^2}$$

Integration by parts

For any $v\colon \mathbb{R}^n \to \mathbb{R}^n$, $h\colon \mathbb{R}^n \to \mathbb{R}$ (smooth and compact support):

$$\int_{\mathbb{R}^n} h(x)\,(\nabla \cdot v)(x)\,dx = -\int_{\mathbb{R}^n} \langle \nabla h(x), v(x) \rangle\,dx$$

Integration by parts

For any $v\colon \mathbb{R}^n \to \mathbb{R}^n$, $h\colon \mathbb{R}^n \to \mathbb{R}$ (smooth and compact support):

$$\int_{\mathbb{R}^n} h(x)\,(\nabla \cdot v)(x)\,dx = -\int_{\mathbb{R}^n} \langle \nabla h(x), v(x) \rangle\,dx$$

- Divergence is *adjoint* of gradient:

$$\langle h, \nabla \cdot v \rangle_{L^2(\mathbb{R}^n \to \mathbb{R})} = \langle -\nabla h, v \rangle_{L^2(\mathbb{R}^n \to \mathbb{R}^n)}$$

Integration by parts

For any $v\colon \mathbb{R}^n \to \mathbb{R}^n$, $h\colon \mathbb{R}^n \to \mathbb{R}$ (smooth and compact support):

$$\int_{\mathbb{R}^n} h(x)\,(\nabla \cdot v)(x)\,dx = -\int_{\mathbb{R}^n} \langle \nabla h(x), v(x)\rangle\,dx$$

- Divergence is *adjoint* of gradient:

$$\langle h, \nabla \cdot v\rangle_{L^2(\mathbb{R}^n \to \mathbb{R})} = \langle -\nabla h, v\rangle_{L^2(\mathbb{R}^n \to \mathbb{R}^n)}$$

- Divergence has 0 integral:

$$\int_{\mathbb{R}^n} 1 \cdot (\nabla \cdot v)(x)\,dx = -\int_{\mathbb{R}^n} \langle \nabla 1, v(x)\rangle\,dx = 0$$

$\nabla 1 = 0$

Divergence-free vector field

A vector field $v\colon \mathbb{R}^n \to \mathbb{R}^n$ is **divergence-free** if

$$\nabla \cdot v = 0$$

$$(\nabla \cdot v)(x) = 0 \quad \forall\, x \in \mathbb{R}^n$$

- Incompressible flow
- Orthogonal to gradient: if $\nabla \cdot v = 0$

then $$\int_{\mathbb{R}^n} \langle \nabla h(x), v(x) \rangle \, dx = -\int_{\mathbb{R}^n} h(x) \underbrace{(\nabla \cdot v)(x)}_{=0} \, dx = 0$$

Helmholtz decomposition

Theorem: (Helmholtz decomposition)

Any vector field $v\colon \mathbb{R}^n \to \mathbb{R}^n$ can be written uniquely as the sum of a gradient and a divergence-free vector field:

$$v = \nabla u + w$$

for some $u\colon \mathbb{R}^n \to \mathbb{R}$ and $w\colon \mathbb{R}^n \to \mathbb{R}^n$ with $\nabla \cdot w = 0$

Helmholtz decomposition

Theorem: (Helmholtz decomposition)

Any vector field $v\colon \mathbb{R}^n \to \mathbb{R}^n$ can be written uniquely as the sum of a gradient and a divergence-free vector field:

$$v = \nabla u + w$$

for some $u\colon \mathbb{R}^n \to \mathbb{R}$ and $w\colon \mathbb{R}^n \to \mathbb{R}^n$ with $\nabla \cdot w = 0$

$v(x) = Ax$

- Linear vector field: Any matrix $A \in \mathbb{R}^{n\times n}$ can be written uniquely as

$$A = A_{\text{sym}} + A_{\text{ant}}$$

where $A_{\text{sym}} = \frac{1}{2}(A + A^\top)$ is symmetric and $A_{\text{ant}} = \frac{1}{2}(A - A^\top)$ is anti-symmetric

Continuity equation

Continuity equation

$v(x) = (v_1(x), \dots, v_n(x))$

Suppose $X_t \in \mathbb{R}^n$ follows dynamics of a vector field $v : \mathbb{R}^n \to \mathbb{R}^n$

$x = (x_1, \dots, x_n) \in \mathbb{R}^n$

ODE: $$\dot{X}_t = v(X_t)$$

$\forall i: \quad \frac{d}{dt} X_{t,i} = v_i(X_t)$

If $X_0 \sim \rho_0$ is random, then $X_t \sim \rho_t$ is random

The density $\rho_t \in \mathcal{P}(\mathbb{R}^n)$ follows the **continuity equation**

$\rho_t : \mathbb{R}^n \to \mathbb{R}$

PDE: $$\frac{\partial \rho_t}{\partial t}(x) = -\nabla \cdot (\rho_t v)(x) \iff \frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$$

$\frac{d}{dt} \rho_t(x)$

Note: Also for time-dependent vector field $v_t(x)$

Continuity equation

$\frac{\partial \rho_t}{\partial t} : \mathbb{R}^n \to \mathbb{R}$ $\qquad \rho_t v : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto \rho_t(x)\, v(x)$

Lemma: If $\dot{X}_t = v(X_t)$, then $\dfrac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$ $\qquad \nabla \cdot (\rho_t v) : \mathbb{R}^n \to \mathbb{R}$

Proof: For $h : \mathbb{R}^n \to \mathbb{R}$, compute time derivative of

$$\mathbb{E}[h(X_t)] = \int_{\mathbb{R}^n} \rho_t(x)\, h(x)\, dx$$

(1.) RHS: $\dfrac{d}{dt} \displaystyle\int_{\mathbb{R}^n} \rho_t(x)\, h(x)\, dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t(x)}{\partial t}\, h(x)\, dx$

(2.) LHS: since $\dot{X}_t = v(X_t)$

$$\Rightarrow \frac{d}{dt} h(X_t) = \langle \nabla h(X_t), \dot{X}_t \rangle = \langle \nabla h(X_t), v(X_t) \rangle$$

Since $\mathbb{E}$ is linear:

$$\frac{d}{dt}\mathbb{E}[h(X_t)] = \mathbb{E}\left[\langle \nabla h(X_t),\, v(X_t)\rangle\right]$$

$$= \int_{\mathbb{R}^n} \langle \nabla h(x),\, v(x)\rangle\, \rho_t(x)\, dx$$

integration by parts:

$$= -\int_{\mathbb{R}^n} h(x)\, \nabla\cdot(\rho_t v)(x)\, dx$$

$\therefore$ ① = ②

$$\Rightarrow \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}(x)\, h(x) = -\int_{\mathbb{R}^n} \nabla\cdot(\rho_t v)(x)\, h(x)\, dx$$

$$\Leftrightarrow \int_{\mathbb{R}^n} \left(\frac{\partial \rho_t}{\partial t}(x) + \nabla\cdot(\rho_t v)(x)\right) h(x) = 0$$

true for arbitrary $h : \mathbb{R}^n \to \mathbb{R}$

$$\Rightarrow \quad \frac{\partial \rho_t}{\partial t}(x) + \nabla\cdot(\rho_t v)(x) = 0$$

which is continuity eq. □

Continuity equation

Lemma: If $\dot{X}_t = v(X_t)$, then $\dfrac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$

Proof: For test function $h\colon \mathbb{R}^n \to \mathbb{R}$, compute time derivative of

$$\mathbb{E}[h(X_t)] = \int_{\mathbb{R}^n} \rho_t(x)\, h(x)\, dx$$

1. For RHS:

$$\frac{d}{dt} \int_{\mathbb{R}^n} \rho_t(x)\, h(x)\, dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}(x)\, h(x)\, dx$$

2. For LHS, since $\dot{X}_t = v(X_t)$

$$\begin{aligned}\frac{d}{dt} h(X_t) &= \langle \nabla h(X_t), \dot{X}_t \rangle \\ &= \langle \nabla h(X_t), v(X_t) \rangle\end{aligned}$$

Since $\mathbb{E}$ is linear and using integration by parts

$$\begin{aligned}\frac{d}{dt} \mathbb{E}[h(X_t)] &= \mathbb{E}[\langle \nabla h(X_t), v(X_t) \rangle] \\ &= \int_{\mathbb{R}^n} \langle \nabla h(x), v(x) \rangle \, \rho_t(x) \, dx \\ &= -\int_{\mathbb{R}^n} h(x) \, (\nabla \cdot (\rho_t v))(x) \, dx\end{aligned}$$

Both are equal: $\frac{d}{dt}\mathbb{E}[h(X_t)] = \frac{d}{dt}\int_{\mathbb{R}^n} \rho_t(x)\, h(x)\, dx$. Therefore

$$\int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}(x)\, h(x)\, dx = -\int_{\mathbb{R}^n} h(x)\, (\nabla \cdot (\rho_t v))(x)\, dx$$

Since h is arbitrary, this implies

$$\frac{\partial \rho_t}{\partial t}(x) = -(\nabla \cdot (\rho_t v))(x)$$

which is the continuity equation. $\square$

Conservation of mass

Continuity equation: $\dot{X}_t = v(X_t) \quad \Rightarrow \quad \dfrac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$

implies **conservation of mass**:

$$\int_{\mathbb{R}^n} \rho_t(x)\, dx = \int_{\mathbb{R}^n} \rho_0(x)\, dx$$

Conservation of mass

Continuity equation: $\dot{X}_t = v(X_t) \quad \Rightarrow \quad \frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$

implies **conservation of mass**:

$$\int_{\mathbb{R}^n} \rho_t(x)\, dx = \int_{\mathbb{R}^n} \rho_0(x)\, dx$$

Proof:

$$\frac{d}{dt} \int_{\mathbb{R}^n} \rho_t(x)\, dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t(x)}{\partial t}\, dx = -\int_{\mathbb{R}^n} \nabla \cdot (\rho_t v)(x)\, dx = 0$$

Example: Gradient flow

If X_t follows **gradient flow**:

$$\dot{X}_t = -\nabla f(X_t) \qquad v = -\nabla f$$

then its density ρ_t follows the continuity equation

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) \quad = -\nabla \cdot (\rho_t v)$$

Time-dependent formulation of W_2

Kinetic energy

Suppose $X_t \in \mathbb{R}^n$ follows dynamics generated by $v_t \colon \mathbb{R}^n \to \mathbb{R}^n$

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

Let $X_t \sim \rho_t$, so $\rho_t \colon \mathbb{R}^n \to \mathbb{R}$ follows the continuity equation

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$$

Kinetic energy

Suppose $X_t \in \mathbb{R}^n$ follows dynamics generated by $v_t \colon \mathbb{R}^n \to \mathbb{R}^n$

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

Let $X_t \sim \rho_t$, so $\rho_t \colon \mathbb{R}^n \to \mathbb{R}$ follows the continuity equation

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$$

At time $t \geq 0$, define **kinetic energy**:

$$E(t) = \int_{\mathbb{R}^n} \rho_t(x) \, \|v_t(x)\|_2^2 \, dx = \mathbb{E}_{\rho_t}\left[\|v_t\|^2\right]$$

Action

Action of dynamics $\rho = (\rho_t)_{0 \le t \le 1}$ generated by $v = (v_t)_{0 \le t \le 1}$

$$\mathcal{A}(\rho, v) = \int_0^1 E(t)\, dt = \int_0^1 \int_{\mathbb{R}^n} \rho_t(x)\, \|v_t(x)\|_2^2\, dx\, dt$$

$X_0 \sim \rho_0$

$\dot{X}_t = v_t(X_t) \iff \frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$

Time-dependent formulation of W_2

$$W_2(\rho_0, \rho_1)^2 = \inf_{\pi \in \Pi(\rho_0, \rho_1)} \mathbb{E}[\|X - Y\|^2]$$

Theorem: (Benamou-Brenier formula)

$$W_2(\rho_0, \rho_1)^2 = \inf_{(\rho, v) \in V(\rho_0, \rho_1)} \mathcal{A}(\rho, v)$$

where $V(\rho_0, \rho_1) = \{(\rho_t, v_t)_{0 \le t \le 1}$ from ρ_0 to ρ_1 via $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)\}$

- See [Villani 2003, Theorem 8.1]

Wasserstein metric

Tangent space

$$\text{Let } \mathcal{P}(\mathbb{R}^n) = \left\{ \rho \colon \mathbb{R}^n \to \mathbb{R}_{\geq 0} \mid \int_{\mathbb{R}^n} \rho(x)\, dx = 1,\ \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$$

- Tangent space $\mathsf{T}_\rho \mathcal{P}$ contains $\phi \colon \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \phi(x)\, dx = 0$

Tangent space

$$\text{Let } \mathcal{P}(\mathbb{R}^n) = \left\{ \rho \colon \mathbb{R}^n \to \mathbb{R}_{\geq 0} \mid \int_{\mathbb{R}^n} \rho(x)\, dx = 1,\ \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$$

- Tangent space $\mathsf{T}_\rho \mathcal{P}$ contains $\phi \colon \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \phi(x)\, dx = 0$
- Want $\phi \in \mathsf{T}_\rho \mathcal{P}$ to represent velocity $\frac{\partial \rho_t}{\partial t}$ at $\rho_0 = \rho$ from motion $\dot{X}_t = v_t(X_t)$, so ρ_t follows continuity equation $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$
- So $\phi \in \mathsf{T}_\rho \mathcal{P}$ is of the form $\phi = -\nabla \cdot (\rho v)$ for some $v \colon \mathbb{R}^n \to \mathbb{R}^n$

Tangent space

Let $\mathcal{P}(\mathbb{R}^n) = \left\{ \rho \colon \mathbb{R}^n \to \mathbb{R}_{\geq 0} \mid \int_{\mathbb{R}^n} \rho(x)\,dx = 1,\ \mathbb{E}_\rho[\|X\|^2] < \infty \right\}$

- Tangent space $\mathsf{T}_\rho\mathcal{P}$ contains $\phi \colon \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \phi(x)\,dx = 0$
- Want $\phi \in \mathsf{T}_\rho\mathcal{P}$ to represent velocity $\frac{\partial \rho_t}{\partial t}$ at $\rho_0 = \rho$ from motion $\dot{X}_t = v_t(X_t)$, so ρ_t follows continuity equation $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$
- So $\phi \in \mathsf{T}_\rho\mathcal{P}$ is of the form $\phi = -\nabla \cdot (\rho v)$ for some $v \colon \mathbb{R}^n \to \mathbb{R}^n$
- Not unique: if v is a solution, then so is $v + \frac{w}{\rho}$ for any $\nabla \cdot w = 0$
- Which one? Choose v that minimizes *kinetic energy* $\mathbb{E}_\rho[\|v\|^2]$

if $\nabla \cdot (\rho v) = -\phi$ and $\nabla \cdot w = 0$

then $\nabla \cdot \left(\rho \left(v + \frac{w}{\rho}\right)\right) = \nabla \cdot (\rho v) + \nabla \cdot w = \nabla \cdot (\rho v) = -\phi$

Wasserstein metric

Let $\phi \in \mathsf{T}_\rho \mathcal{P}$, so $\phi \colon \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \phi(x)\, dx = 0$

Wasserstein metric:

$$\|\phi\|_\rho^2 = \inf_{\substack{v \colon \mathbb{R}^n \to \mathbb{R}^n \\ \nabla \cdot (\rho v) = -\phi}} \int_{\mathbb{R}^n} \rho(x)\, \|v(x)\|_2^2\, dx = \mathbb{E}_\rho \left[\|v\|^2 \right]$$

Minimizer is gradient

Lemma: Minimizer is a *gradient* $v^* = \nabla u$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Minimizer is gradient

Lemma: Minimizer is a *gradient* $v^* = \nabla u$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Proof: Let v^* be minimizer. For any w with $\nabla \cdot w = 0$ and $\eta \neq 0$:

$$\int_{\mathbb{R}^n} \rho(x) \|v^*(x)\|_2^2 \, dx \leq \int_{\mathbb{R}^n} \rho(x) \left\| v^*(x) + \eta \frac{w(x)}{\rho(x)} \right\|_2^2 dx$$

Expand and simplify, divide by η, and let $\eta \to 0$:

$$\int_{\mathbb{R}^n} \langle v^*(x), w(x) \rangle dx = 0$$

Since w is arbitrary with $\nabla \cdot w = 0$, this implies $v^* = \nabla u$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$. by Helmholtz decomposition. □

Wasserstein metric

$\phi : \mathbb{R}^n \to \mathbb{R}$ with $\int_{\mathbb{R}^n} \phi(x)\, dx = 0$

Tangent function $\phi \in \mathsf{T}_\rho \mathcal{P}$ is $\phi = -\nabla \cdot (\rho \nabla u)$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Wasserstein metric:

$$\|\phi\|_\rho^2 = \mathbb{E}_\rho[\|\nabla u\|^2] = \int_{\mathbb{R}^n} \rho(x)\, \|\nabla u(x)\|_2^2\, dx$$

Wasserstein metric

Tangent function $\phi \in \mathsf{T}_\rho \mathcal{P}$ is $\phi = -\nabla \cdot (\rho \nabla u)$ for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Wasserstein metric:

$$\|\phi\|_\rho^2 = \mathbb{E}_\rho[\|\nabla u\|^2] = \int_{\mathbb{R}^n} \rho(x) \, \|\nabla u(x)\|_2^2 \, dx$$

Benamou-Brenier formula $\Rightarrow$ metric generates W_2 distance

$$W_2(\rho, \nu)^2 = \inf_{\substack{(\rho_t) \\ \rho_0 = \rho,\, \rho_1 = \nu}} \int_0^1 \|\dot{\rho}_t\|_{\rho_t}^2 \, dt$$

Inner product

For $\phi_1, \phi_2 \in \mathsf{T}_\rho \mathcal{P}$ with

$$\phi_1 = -\nabla \cdot (\rho \nabla u_1)$$
$$\phi_2 = -\nabla \cdot (\rho \nabla u_2)$$

Wasserstein inner product:

$$g_s(\phi_1, \phi_2) = \langle \phi_1, \phi_2 \rangle_\rho = \mathbb{E}_\rho[\langle \nabla u_1, \nabla u_2 \rangle]$$
$$= \int_{\mathbb{R}^n} \rho(x) \langle \nabla u_1(x), \nabla u_2(x) \rangle \, dx$$

Optimization on $\mathcal{P}(\mathbb{R}^n)$

Optimization on $\mathcal{P}(\mathbb{R}^n)$

Given functional $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$, want to optimize

$$\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} F(\rho)$$

Optimization on $\mathcal{P}(\mathbb{R}^n)$

Given functional $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$, want to optimize

$$\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} F(\rho)$$

- Want to run **gradient flow**:

  $$\dot{\rho}_t = -\operatorname{grad} F(\rho_t)$$

  or **gradient descent**:

  $$\rho_{k+1} = \operatorname{Exp}_{\rho_k}(-\eta \operatorname{grad} F(\rho_k))$$

  or proximal method, accelerated method, ...

Optimization on $\mathcal{P}(\mathbb{R}^n)$

Given functional $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$, want to optimize

$$\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} F(\rho)$$

- Want to run **gradient flow**:

  $$\dot{\rho}_t = -\operatorname{grad} F(\rho_t)$$

  or **gradient descent**:

  $$\rho_{k+1} = \operatorname{Exp}_{\rho_k}(-\eta \operatorname{grad} F(\rho_k))$$

  or proximal method, accelerated method, ...

- Want to understand convexity, smoothness, gradient domination, ...

Examples of Functionals

$$F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$$

1. **Potential energy:**

$$F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx = \mathbb{E}_\rho[f] \quad \text{for some } f : \mathbb{R}^n \to \mathbb{R}$$

$$\text{or } F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, \|\nabla f(x)\|^2\, dx = \mathbb{E}_\rho\left[\|\nabla f\|^2\right]$$

Examples of Functionals

$$F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$$

1. **Potential energy:**

$$F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx = \mathbb{E}_\rho[f]$$

2. **Internal energy:**

$$F(\rho) = \int_{\mathbb{R}^n} U(\rho(x))\, dx$$

- Entropy: $H(\rho) = \int_{\mathbb{R}^n} \rho(x)\, \log \rho(x)\, dx = \mathbb{E}_\rho[\log \rho]$

3. **Interaction energy:** $= \frac{1}{2}\,\mathbb{E}[W(X-Y)]$, $X, Y \sim \rho$ iid.

$$F(\rho) = \frac{1}{2}\int_{\mathbb{R}^n \times \mathbb{R}^n} W(x-y)\,\rho(x)\,\rho(y)\,dx\,dy$$

- Variance: $\mathrm{Var}(\rho) = \frac{1}{2}\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2\,\rho(x)\,\rho(y)\,dx\,dy$

Functionals

3. **Interaction energy:**

$$F(\rho) = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} W(x-y)\, \rho(x)\, \rho(y)\, dx\, dy$$

- Variance: $\mathrm{Var}(\rho) = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2\, \rho(x)\, \rho(y)\, dx\, dy$

4. Functionals involving derivatives

$$F(\rho) = \int_{\mathbb{R}^n} V(\rho(x), \nabla \rho(x))\, dx$$

- Fisher information: $J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2] = \int_{\mathbb{R}^n} \frac{\|\nabla \rho(x)\|^2}{\rho(x)}\, dx$

$= \frac{\nabla \rho}{\rho}$

Otto Calculus

Review: Gradient on manifold

The **gradient** of $f : \mathcal{X} \to \mathbb{R}$ at $x \in \mathcal{X}$ is the tangent vector

$$\operatorname{grad} f(x) \in \mathsf{T}_x\mathcal{X}$$

that gives directional derivative:

$$\langle \operatorname{grad} f(x), v \rangle_x = \frac{d}{dt} f(X_t)\Big|_{t=0}$$

where X_t is geodesic from $X_0 = x$ along direction $\dot{X}_0 = v$

L^2 variation

The L^2 **variation** of $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is $\frac{\delta F}{\delta \rho} : \mathbb{R}^n \to \mathbb{R}$ given by

$$\frac{\delta F}{\delta \rho}(x) = \frac{\partial F(\rho)}{\partial \rho(x)}$$

$\rho \in \mathcal{P}(\mathbb{R}^n)$

$\rho : \mathbb{R}^n \to \mathbb{R}$

$\rho = (\rho(x) : x \in \mathbb{R}^n)$

L^2 variation

The L^2 **variation** of $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is $\frac{\delta F}{\delta \rho}\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\frac{\delta F}{\delta \rho}(x) = \frac{\partial F(\rho)}{\partial \rho(x)}$$

- Potential energy: $F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$

$$\frac{\delta F}{\delta \rho}(x) = f(x)$$

L^2 variation

The L^2 **variation** of $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is $\frac{\delta F}{\delta \rho}\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\frac{\delta F}{\delta \rho}(x) = \frac{\partial F(\rho)}{\partial \rho(x)}$$

- Potential energy: $F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$

$$\frac{\delta F}{\delta \rho}(x) = f(x)$$

- Internal energy: $F(\rho) = \int_{\mathbb{R}^n} U(\rho(x))\, dx$

$$\frac{\delta F}{\delta \rho}(x) = U'(\rho(x))$$

Gradient

$T_\rho \mathcal{P} = \{ \phi = -\nabla \cdot (\rho \nabla u) \text{ for some } u : \mathbb{R}^n \to \mathbb{R} \}$

Otto calculus: Rule for computing derivatives in W_2 metric

Lemma: The *gradient* of $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ at ρ is

$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right)$$

Gradient

Lemma:

$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \, \nabla \frac{\delta F}{\delta \rho} \right)$$

Proof: For any geodesic ρ_t from $\rho_0 = \rho$ with $\dot{\rho}_0 = -\nabla \cdot (\rho \nabla u)$

$$\begin{aligned}
\frac{d}{dt} F(\rho_t)\Big|_{t=0} &= \int_{\mathbb{R}^n} \frac{\delta F(\rho)}{\delta \rho(x)} \, \dot{\rho}_0(x) \, dx \\
&= -\int_{\mathbb{R}^n} \frac{\delta F(\rho)}{\delta \rho(x)} \, \nabla \cdot (\rho \nabla u)(x) \, dx \\
&= \int_{\mathbb{R}^n} \rho(x) \left\langle \nabla \frac{\delta F(\rho)}{\delta \rho(x)}, \, \nabla u \right\rangle \, dx \\
&= \langle \operatorname{grad} F(\rho), \, \dot{\rho}_0 \rangle_\rho
\end{aligned}$$

Potential energy

Gradient of potential energy

Potential energy:

$$F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx = \mathbb{E}_\rho[f]$$

L^2 variation:

$$\frac{\delta F}{\delta \rho}(x) = f(x)$$

Gradient of potential energy:

$$\operatorname{grad} F(\rho) = -\nabla \cdot (\rho\, \nabla f)$$

Gradient flow of potential energy

Potential energy: $F(\rho) = \mathbb{E}_\rho[f]$

Gradient flow of potential energy:

$$\dot{\rho}_t = -\mathrm{grad}\, F(\rho_t) = \nabla \cdot (\rho_t \nabla f)$$

$$\Leftrightarrow \quad \frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

Gradient flow of potential energy

Gradient flow of potential energy $F(\rho) = \mathbb{E}_\rho[f]$:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

This is the continuity equation of the **gradient flow** of f:

$$\dot{X}_t = -\nabla f(X_t)$$

Gradient flow of $F(\rho) = \mathbb{E}_\rho[f]$ is *implemented* by gradient flow of f