

Optimal Transportation Map

David Gu

Yau Mathematics Science Center
Tsinghua University
Computer Science Department
Stony Brook University

gu@cs.stonybrook.edu

August 6, 2020

Duality Theories

Monge Problem

Assume Ω and Σ are two domains in the Euclidean space, \mathbb{R}^d , μ and ν are two probability measures on Ω and Σ respectively, $\mu \in \mathcal{P}(\Omega)$, $\nu \in \mathcal{P}(\Sigma)$, such that they have equal total measure:

$$\mu(\Omega) = \nu(\Omega). \quad (1)$$

Definition (Measure-preserving Map)

A mapping $T : \Omega \rightarrow \Sigma$ is called *measure preserving*, if for any Borel set $B \subset \Sigma$,

$$\int_{T^{-1}(B)} d\mu = \int_B d\nu, \quad (2)$$

and is denoted as $T_\# \mu = \nu$. T pushes μ forward to ν .

Monge Problem

Suppose the density functions of μ and ν are given by $f : \Omega \rightarrow \mathbb{R}$ and $g : \Sigma \rightarrow \mathbb{R}$, namely

$$d\mu = f(x_1, x_2, \dots, x_d) dx_1 \wedge dx_2 \wedge \dots \wedge dx_d,$$

$$d\nu = g(y_1, y_2, \dots, y_d) dy_1 \wedge dy_2 \wedge \dots \wedge dy_d,$$

and $T : \Omega \rightarrow \Sigma$ is C^1 and measure-preserving,

$$f(x_1, \dots, x_d) dx_1 \wedge \dots \wedge dx_d = g(T(x)) dy_1 \wedge \dots \wedge dy_d.$$

then T satisfies the Jacobi equation:

Definition (Jacobi Equation)

$$\det DT(x) = \frac{f(x)}{g \circ T(x)} \tag{3}$$

Monge Problem

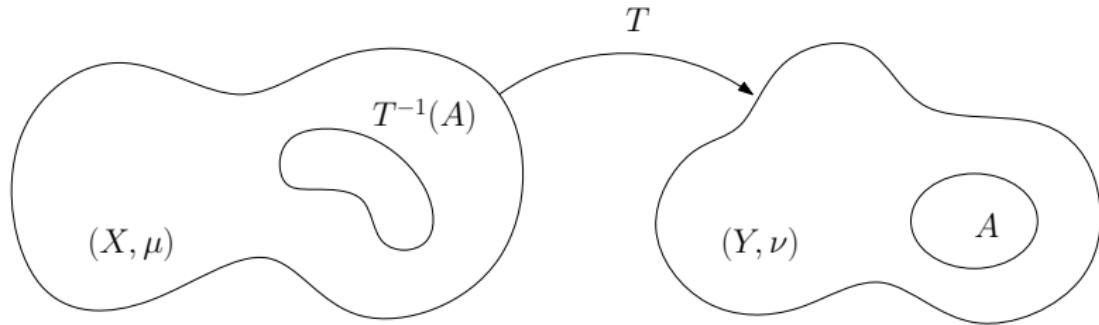


Figure: Measure-preserving map.

Monge Problem

Definition (Transportation Cost)

Given a cost function $c : \Omega \times \Sigma \rightarrow \mathbb{R}$, the total transportation cost for a map $T : \Omega \rightarrow \Sigma$ is defined as

$$\mathcal{C}(T) := \int_{\Omega} c(x, T(x)) d\mu(x).$$

Problem (Monge)

Among all the measure-preserving mappings, $T : \Omega \rightarrow \Sigma$ and $T_{\#}\mu = \nu$, find the one with the minimal total transportation cost,

$$MP : \quad \min \left\{ \int_{\Omega} c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\}. \quad (4)$$

Monge Problem

Definition (Optimal Transportation Map)

The solution to the Monge problem is called an optimal transportation map between (Ω, μ) and (Σ, ν) .

Suppose Ω coincides with Σ

Definition (Wasserstein Distance)

The total cost of the optimal transportation map $T : \Omega \rightarrow \Sigma$, $T_{\#}\mu = \nu$, is called the Wasserstein distance between μ and ν .

Suppose the cost is the square of the Euclidean distance
 $c(x, y) = |x - y|^2$, then the Wasserstein distance is defined as

$$\mathcal{W}_2^2(\mu, \nu) := \inf \left\{ \int_{\Omega} |x - T(x)|^2 d\mu(x) : \quad T_{\#}\mu = \nu \right\}.$$

Kantorovich Problem

Transportation Plan

Kantorovich relax the transportation map to transportation scheme, or transportation plan, which is represented by a joint probability distribution $\rho : \omega \times \Sigma \rightarrow \mathbb{R}$, $\rho(x, y)$ represents how much mass is transported from the source point x to the target point y .

Marginal Distribution

The marginal distribution of ρ equals to μ and ν , namely we have the condition

$$(\pi_x)_\# \rho = \mu, \quad (\pi_y)_\# \rho = \nu, \quad (5)$$

where the projection maps

$$\pi_x(x, y) = x, \quad \pi_y(x, y) = y.$$

Kantorovich Problem

Transportation map vs. Transportation plan

Transportation map is a special case of transportation plan, namely a transportation map $T : \Omega \rightarrow \Sigma$ induces a transportation plan

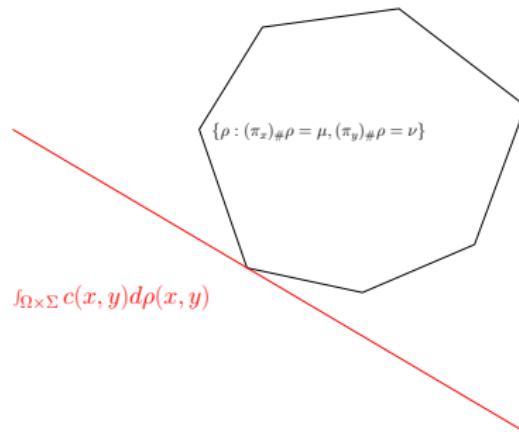
$$(Id, T)_\# \mu = \rho. \quad (6)$$

Kantorovich Problem

Problem (Kantorovich)

Find a transportation plan with the minimal total transportation cost,

$$KP : \quad \min \left\{ \int_{\Omega \times \Sigma} c(x, y) d\rho(x, y) : (\pi_x)_\# \rho = \mu, (\pi_y)_\# \rho = \nu \right\}. \quad (7)$$



Kantorovich Problem

Linear Programming

Kantorovich problem is to find a minimal value of a linear function defined on a convex polytope, so the solution exists. KP can be solved using linear programming method, such as simplex, interior point or ellipsoid algorithms.

Kantorovich Problem

In general situation, the support of a transportation plan ρ covers all the $\Omega \times \Sigma$. If the transportation map T exists, the support of $(Id, T)_\# \mu$ has 0 measure in $\Omega \times \Sigma$. KP doesn't discover the intrinsic structure, it is highly inefficient to compute optimal transportation map.

Kantorovich Dual Problem

Denote $\Pi(\mu, \nu) = \{\rho : (\pi_x)_\# \rho = \mu, (\pi_y)_\# \rho = \nu\}$. We consider the constraint $\rho \in \Pi(\mu, \nu)$. we have

$$\sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu - \int_{\Omega \times \Sigma} (\varphi(x) + \psi(y)) d\rho = \begin{cases} 0 & \rho \in \Pi(\mu, \nu), \\ +\infty & \rho \notin \Pi(\mu, \nu), \end{cases} \quad (8)$$

where the superimum is taken among all bounded continuous functions, $\varphi \in C_b(\Omega)$ and $\psi \in C_b(\Sigma)$.

Kantorovich Problem

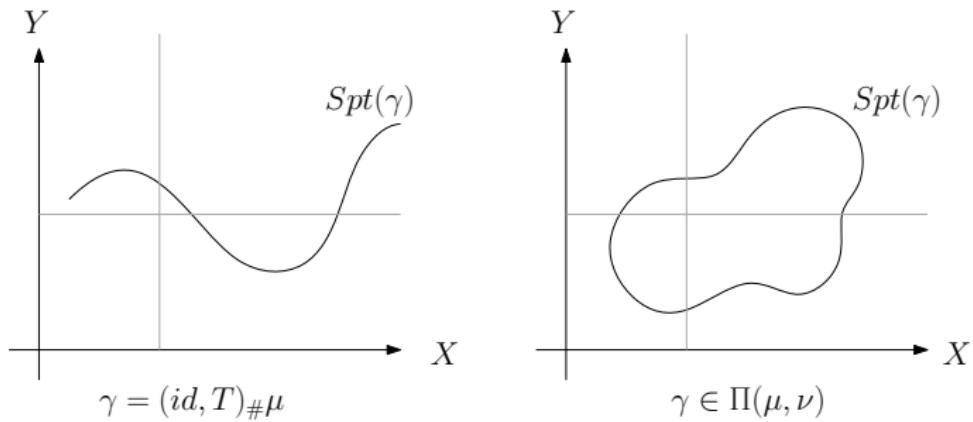


Figure: Caption

Kantorovich Dual Problem

We use this as a generalized Lagrange multiplier in (KP), and rewrite (KP) as

$$\min_{\rho} \int_{\Omega \times \Sigma} c d\rho + \sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu - \int_{\Omega \times \Sigma} (\varphi(x) + \psi(y)) d\rho \quad (9)$$

Under suitable conditions, such as Rockafella's conditions, we can exchange sup and inf

$$\sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu + \inf_{\rho} \int_{\Omega \times \Sigma} (c(x, y) - (\varphi(x) + \psi(y))) d\rho. \quad (10)$$

We can rewrite the infimum in ρ as a constraint on φ and ψ :

$$\inf_{\rho \geq 0} \int_{\Omega \times \Sigma} (c - \varphi \oplus \psi) d\rho = \begin{cases} 0 & \varphi \oplus \psi \leq c \text{ on } X \times Y \\ -\infty & \varphi \oplus \psi > c \end{cases}$$

where $\varphi \oplus \psi$ denotes the function $\varphi \oplus \psi(x, y) := \varphi(x) + \psi(y)$.

Kantovorich Dual Problem

This leads to the dual optimization problem.

Problem (Dual)

Given $\mu \in \mathcal{P}(\Omega)$ and $\nu \in \mathcal{P}(\Sigma)$ and the cost function $c : \Omega \times \Sigma \rightarrow [0, +\infty)$, we consider the problem

$$(DP) \quad \max \left\{ \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu : \varphi \in C_b(\Omega), \psi \in C_b(\Sigma) : \varphi \oplus \psi \leq c \right\}. \quad (11)$$

From the condition $\varphi \oplus \psi \leq c$, we obtain $\sup DP \leq \min KP$,

$$\int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu = \int_{\Omega \times \Sigma} \varphi \oplus \psi d\rho \leq \int_{\Omega \times \Sigma} cd\rho$$

This is valid for all admissible pairs (φ, ψ) and every admissible ρ .

Cyclic Monotonicity

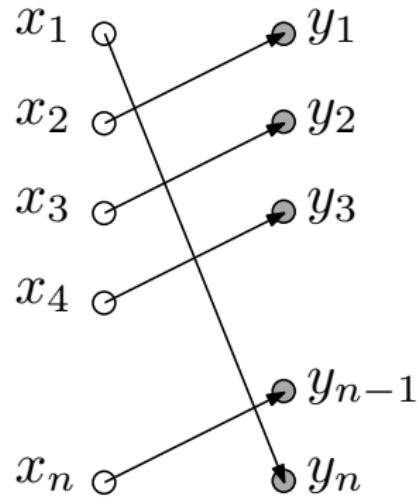
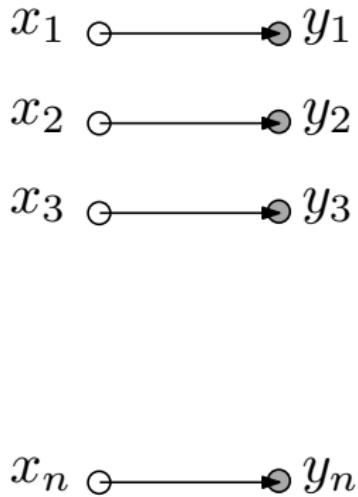


Figure: Cyclic monotonicity.

Cyclic Monotonocity

Theorem (Cyclic Monotonocity)

Suppose $\rho : \Omega \times \Sigma$ is an optimal transpportation plan for (Ω, μ) and (Σ, ν) , for any set of pair of points:

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k) \subset \text{Supp}(\rho),$$

then we have the following inequality

$$\sum_{i=1}^k c(x_i, y_i) \leq \sum_{i=1}^k c(x_i, y_{\sigma(i)}),$$

where σ is a permuation of $1, 2, \dots, k$.

The cyclic monotonocity can be applied to prove the equivalence between Kantorovich problem and Kantorovic dual problem.

Cyclic Monotonicity

Theorem

If ρ is an optimal transport plan for the continuous cost c , then its support $spt(\rho)$ is cyclic monotonous.

Theorem ($\max(DP) = \min(KP)$)

Suppose that Ω and Σ are Polish spaces and that $c : \Omega \times \Sigma \rightarrow \mathbb{R}$ is uniformly continuous and bounded. Then the problem (DP) admits a solution (φ, φ^c) and we have $\max(DP) = \min(KP)$.

Definition (c-transform)

Given $\varphi \in L^1(\Omega)$, and the cost function $c : \Omega \times \Sigma \rightarrow \mathbb{R}$, the c-transform of φ is defined as $\varphi^c : \Sigma \rightarrow \mathbb{R}$,

$$\varphi^c(y) := \inf_{x \in \Omega} c(x, y) - \varphi(x), \quad (12)$$

The optimization of Kantorovich functional is equivalent to replace the Kantorovich potentials (φ_n, ψ_n) by the c-transforms of the other, namely

$$(\varphi, \psi) \rightarrow (\varphi, \varphi^c) \rightarrow (\varphi^{cc}, \varphi^c) \rightarrow (\varphi^{cc}, \varphi^{ccc}) \cdots$$

c-transform

Geometrically, if we fix a point $x \in \Omega$, then we get a supporting surface $\Gamma_x : \Sigma \rightarrow \mathbb{R}$,

$$\Gamma_x(y) := c(x, y) - \varphi(x),$$

the graph of the c-transform $\varphi^c(y)$ is the envelope of all these supporting surfaces.

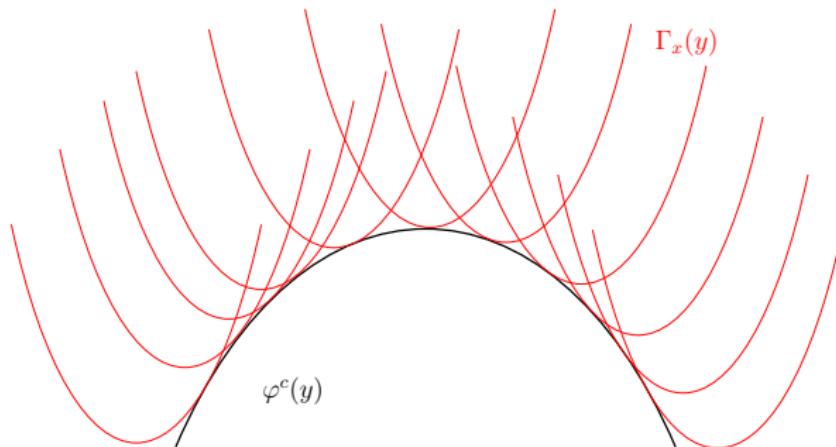


Figure: Geometric interpretation of c-transform.

Dual Problem

By utilizing c-transform, we obtain

Problem (Dual Problem)

Given $\mu \in \mathcal{P}(\Omega)$, $\nu \in \mathcal{P}(\Sigma)$, the dual problem is

$$DP : \max_{\varphi \in C_b(\Omega)} \left\{ \int_{\Omega} \varphi(x) d\mu(x) + \int_{\Sigma} \varphi^c(y) \nu(y) \right\}. \quad (13)$$

Monge-Ampere Equation

Suppose the cost function is a strictly convex function, satisfying the condition $c(x, y) = c(x - y)$, then

$$D_x c(x, y) - D\varphi(x) = 0,$$

we obtain $D_x c(y - x) = D\varphi(x)$,

$$T(x) = y = x - (Dc)^{-1}(D\varphi(x)),$$

Theorem (Brenier)

Given $\mu \in \mathcal{P}(\Omega)$ and $\nu \in \mathcal{P}(\Sigma)$, and the cost function $c(x, y) = \frac{1}{2}|x - y|^2$, the optimal transportation map is the gradient of a function $u : \Omega \rightarrow \mathbb{R}$, $T(x) := \nabla u(x)$.

Brenier Problem

Problem (Brenier)

Find a convex function $u : \Omega \rightarrow \mathbb{R}$, satisfying the Monge-Ampère equation,

$$\det \left(\frac{\partial^2 u(x)}{\partial x_i \partial x_j} \right) = \frac{f(x)}{g \circ \nabla u(x)}. \quad (14)$$

Brenier Problem

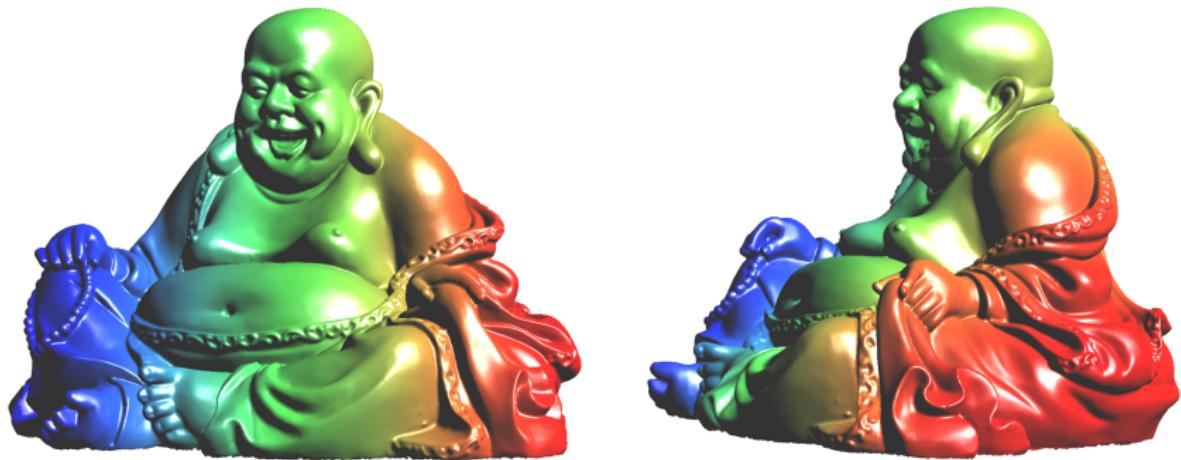


Figure: Buddha surface.

Brenier Problem



Figure: Optimal transportation map.

Brenier Problem

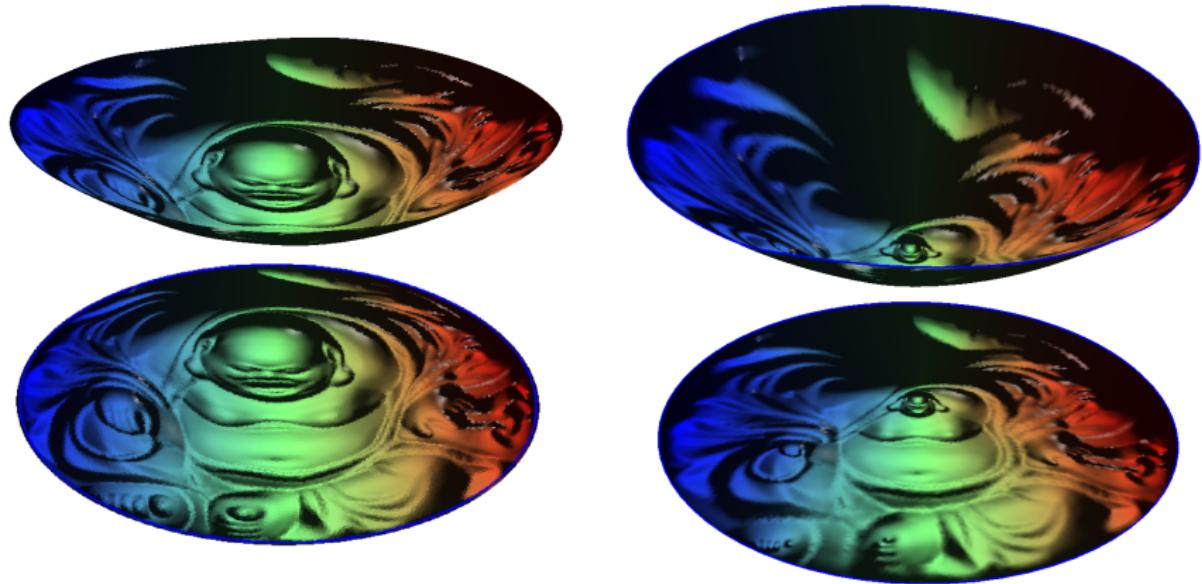


Figure: Brenier potential.

Fluid View

Consider a flow field of special gas. At each time $t \in [0, 1]$, at point $x \in \Omega$, the density of the gas is $\rho(x, t)$. For Lagrangian point of view, the trajectory of each particle (molecule) is a curve, denoted as $\gamma_x(t)$, with initial position and velocity

$$\gamma_x(0) = x, \quad \gamma'_x(t) = \mathbf{v}(\gamma_x(t)).$$

Diffeomorphism

Suppose the trajectories of different particles intersect at some time t , then globally there will be shock waves in the flow field. If there is no shocks, then at each time $t \in [0, 1]$, the initial position x of each particle is mapped to the current position $\gamma_x(t)$, this gives a global diffeomorphism

$$g_t := g(\cdot, t) : x \mapsto \gamma_x(t),$$

at time t , the velocity of the particle is $\gamma'_x(t) = \mathbf{v}(\gamma_x(t))$, the global velocity field is denoted as $\mathbf{v}(x, t)$, then the diffeomorphism satisfies the ODE:

$$\frac{d}{dt}g(x, t) = \mathbf{v}(g(x, t), t). \quad (15)$$

Diffeomorphism

Given a smooth velocity field $\mathbf{v}(x, t)$, we can get the diffeomorphism group $g(x, t)$. Namely, if $\mathbf{v}(x, t)$ is smooth enough, no shock waves will appear,

$$\partial_t \log \det \left[\frac{\partial g(x, t)}{\partial x} \right] = \nabla \cdot \mathbf{v}(g(x, t), t). \quad (16)$$

McCann Interpolation

If the cost function is strictly convex, under Lagrangian point of view, all particles move with uniform speed in a straight line, their trajectories are

$$g_t(x) = (1 - t)x + t(\nabla c)^{-1}(\nabla \varphi), \quad (17)$$

where φ is the optimal Kantorovich potential, this is called McCann interpolation.

McCann's interpolation gives geodesics in Wasserstein space. One can show that

$$\mathcal{W}_c((g_s)_\# \mu, (g_t)_\# \nu) = |s - t| \mathcal{W}_c(\mu, \nu), \quad \forall s, t \in [0, 1].$$

Time Dependent Optimal Transport Problem

Given a differential cost function $c(\mathbf{v})$, defined on velocity vector, then the cost for a trajectory is

$$\mathcal{C}[g_t(x)] := \int_0^1 c(\dot{g}_t(x)) dt.$$

Problem (Time Dependent Optimal Transport)

Find a flow field connecting μ and ν , that minimizes the total cost of all trajectories:

$$\inf \left\{ \int_{\Omega} \mathcal{C}[g_t(x)] d\mu(x) : g_0 = Id, (g_1)_\# \mu = \nu \right\}. \quad (18)$$

Time Dependent Optimal Transport Problem

From mass-conservation law, we get the continuity equation:

$$\frac{d}{dt}\rho(x, t) + \nabla \cdot (\rho(x, t)\mathbf{v}(x, t)) = 0,$$

By McCann interpolation, the velocity field satisfies the Euler equation:

$$\frac{d}{dt}\mathbf{v}(x, t) + \mathbf{v}(x, t) \cdot \nabla \mathbf{v}(x, t) = 0.$$

The cost function information is implied by the initial condition

$$\mathbf{v}(x, 0) = (\nabla c)^{-1}(\nabla \varphi),$$

where φ is the optimal Kantorovich potential.

Benamou-Brenier Problem

Consider all flow fields connecting μ and ν , denote $(\rho(x, t), \mathbf{v}(x, t))$ as (ρ_t, \mathbf{v}_t) , then

$$\Gamma(\mu, \nu) := \left\{ (\rho_t, \mathbf{v}_t) : \frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \mathbf{v}_t), \quad \rho_0 = \mu, \quad \rho_1 = \nu \right\}$$

Given any time $t \in [0, 1]$, the Kinetic energy of the velocity field $\mathbf{v}(x, t)$ is defined as

$$E(\mathbf{v}_t) := \frac{1}{2} \int_{\Omega} \rho(x, t) \|\mathbf{v}(x, t)\|^2 dx.$$

Benamou-Brenier Problem

Problem (Benamou-Brenier)

Find the flow field in $\Gamma(\mu, \nu)$, that minimizes the total kinetic energy,

$$BB : \quad \min \left\{ \frac{1}{2} \int_0^1 \int_{\Omega} \rho_t \|\mathbf{v}_t\|^2 dx dt : (\rho_t, \mathbf{v}_t) \right\} \quad (19)$$

Benamou-Brenier Problem

Use variational approach, assume $\rho_t \mathbf{w}_t$ is divergence free,

$$\frac{1}{2} \frac{d}{d\varepsilon} \int_0^1 \int_{\Omega} \rho_t \langle \mathbf{v}_t + \varepsilon \mathbf{w}_t, \mathbf{v}_t + \varepsilon \mathbf{w}_t \rangle dx dt = \int_0^1 \int_{\Omega} \langle \mathbf{v}_t, \varepsilon \mathbf{w}_t \rangle dx dt = 0,$$

by Hodge decomposition theorem, \mathbf{v}_t is orthogonal to all divergence free vector fields, so $\mathbf{v}_t = \nabla u_t$, where $u_t : \Omega \rightarrow \mathbb{R}$ is a family of functions. This can be obtained by McCann interpolation.

Benamou-Brenier Problem

Problem (Benamou-Brenier)

$$\mathcal{W}_2(\mu, \nu) := \min \left\{ \int_0^1 \left\| \frac{\partial \rho}{\partial t} \right\|_{\rho(t)}^2 dt, \rho_0 = \mu, \rho_1 = \nu, -\nabla \cdot (\rho \nabla u) = \frac{\partial \rho}{\partial t} \right\}$$

where

$$\left\| \frac{\partial \rho}{\partial t} \right\|_{\rho(t)}^2 = \int_{\Omega} \rho |\nabla u|^2,$$

Otto's Interpretation

Given two geodesics $\rho_1(t), \rho_2(t) \subset \mathcal{P}(\Omega)$, $\rho_1(0) = \rho_2(0) = \rho$, the tangent vector at $\rho \in \mathcal{P}(\Omega)$,

$$\frac{\partial \rho_1}{\partial t} = -\nabla \cdot (\rho \nabla \varphi_1)$$

$$\frac{\partial \rho_2}{\partial t} = -\nabla \cdot (\rho \nabla \varphi_2)$$

the Riemannian inner product is

$$\left\langle \frac{\partial \rho_1}{\partial t}, \frac{\partial \rho_2}{\partial t} \right\rangle_{\rho} = \int_{\Omega} \rho \langle \nabla \varphi_1, \nabla \varphi_2 \rangle dx.$$

Entropy Flow

Definition (Entropy)

Given a probability measure $\rho \in \mathcal{P}(\Omega)$, its entropy is defined as

$$\text{Ent}(\rho) := \int_{\Omega} \rho \log \rho dx.$$

Principle of maximum entropy

The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with largest entropy, in the context of precisely stated prior data.

Entropy Flow

Consider a path $\rho(t) \subset \mathcal{P}(\Omega)$,

$$\frac{d}{dt} \text{Ent}(\rho(t)) = \int_{\omega} \left(\dot{\rho} \log \rho + \rho \frac{\dot{\rho}}{\rho} \right) dx = \int_{\Omega} (1 + \log \rho) \dot{\rho} dx.$$

By continuity equation $\dot{\rho} = -\nabla \cdot (\rho \mathbf{v})$, assume $\Omega = \mathbb{R}^d$, hence

$$\int_{\Omega} \dot{\rho} dx = - \int_{\Omega} \nabla \cdot (\mathbf{v} \rho) dx = - \int_{\partial \Omega} \rho \mathbf{v} dx = 0.$$

We obtain

$$\frac{d}{dt} \text{Ent}(\rho(t)) = \int_{\Omega} \log \rho \dot{\rho} dx = - \int_{\Omega} \log \rho \nabla \cdot (\rho \mathbf{v}),$$

Entropy Flow

At the same time

$$\nabla \cdot (\rho \log \rho \mathbf{v}) = \log \rho \nabla \cdot (\rho \mathbf{v}) + \langle \nabla \log \rho, \rho \mathbf{v} \rangle,$$

We obtain

$$\begin{aligned}\frac{d}{dt} \text{Ent}(\rho(t)) &= - \int_{\Omega} \log \rho \nabla \cdot (\rho \mathbf{v}) \\ &= \int_{\Omega} \langle \nabla \log \rho, \rho \mathbf{v} \rangle - \int_{\partial \Omega} \rho \log \rho \mathbf{v} \\ &= \int_{\Omega} \langle \nabla \log \rho, \mathbf{v} \rangle \rho dx.\end{aligned}$$

This shows the Wasserstein gradient of Entropy is $\nabla \log \rho$.

Entropy Flow

In order to reduce the entropy, we let $\mathbf{v} = -\nabla \log \rho$, plug into the continuity equation:

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t \mathbf{v}_t) = 0,$$

hence

$$\frac{\partial \rho_t}{\partial t} - \nabla \cdot (\rho_t \nabla \log \rho_t) = 0$$

$$\frac{\partial \rho_t}{\partial t} - \nabla \cdot \left(\rho_t \frac{\nabla \rho_t}{\rho_t} \right) = 0$$

$$\frac{\partial \rho_t}{\partial t} - \Delta_t = 0$$

This shows Wasserstein gradient flow of entropy equals to the classical heat flow.

Entropy Flow

We let $\mathbf{v} = -\nabla \log \rho$, plug into the continuity equation:

$$\frac{d}{dt} \text{Ent}(\rho(t)) = \int_{\Omega} \langle \nabla \log \rho, \mathbf{v} \rangle \rho dx = - \int_{\Omega} \frac{|\nabla \rho|^2}{\rho} dx = -4 \int_{\Omega} |\nabla \sqrt{\rho}|^2 dx.$$

This gives the dissipation speed of the entropy. Let t go to infinity, ρ_∞ becomes a uniform distribution.