An introduction to optimal transport (OT bootcamp)

Brendan Pass (U. Alberta)

June 20, 2022

Goal of these lectures

- Introduce some basic concepts from optimal transportation theory.
- Focus on ideas (rather than technical details) and on building intuition (with diagrams, non-rigorous proof sketches, etc.)
- Briefly cover a few topics requested by speakers.

Very incomplete list of references

- C. Villani. *Topics in optimal transportation.* AMS, 2003.
- C. Villani. *Optimal transport: old and new.* Springer, 2009.
- F. Santambrogio. *Optimal transport for applied mathematicians.* Birkhauser, 2015.
- G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science* Now Publishers, 2019.
- A. Galichon. *Optimal Transport Methods in Economics* Princeton University Press, 2019.

Very incomplete list of references

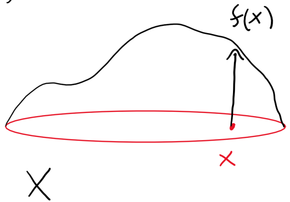
- C. Villani. *Topics in optimal transportation.* AMS, 2003.
- C. Villani. *Optimal transport: old and new.* Springer, 2009.
- F. Santambrogio. *Optimal transport for applied mathematicians.* Birkhauser, 2015.
- G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science* Now Publishers, 2019.
- A. Galichon. *Optimal Transport Methods in Economics* Princeton University Press, 2019.

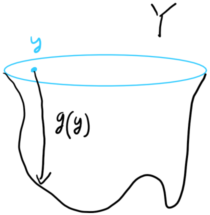
Origins of optimal transport

- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?

$d\mu(x) = f(x)\,dx$

$d\nu(y) = g(y)\,dy$





$$\int_X f(x)\,dx = \int_Y g(y)\,dy = 1$$

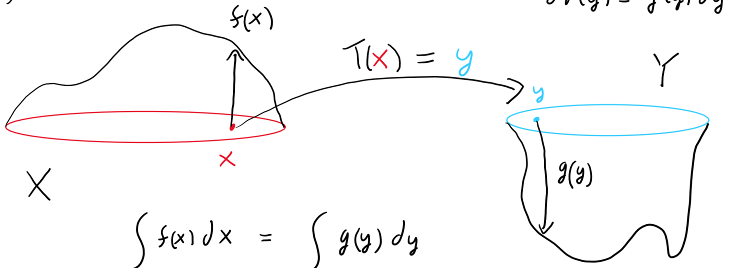
Origins of optimal transport

- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?

$d\mu(x) = f(x)\,dx$

"transport map"

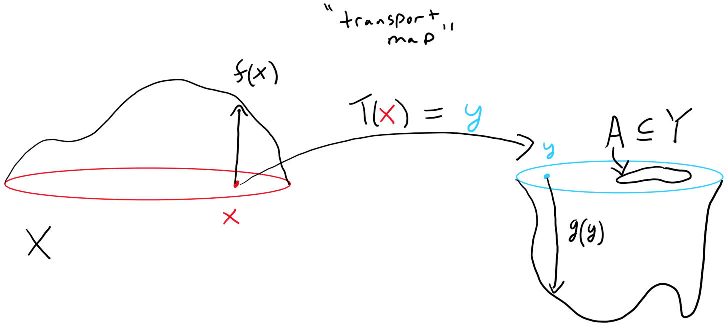
$d\nu(y) = g(y)\,dy$



$$\int_X f(x)\,dx = \int_Y g(y)\,dy = 1$$

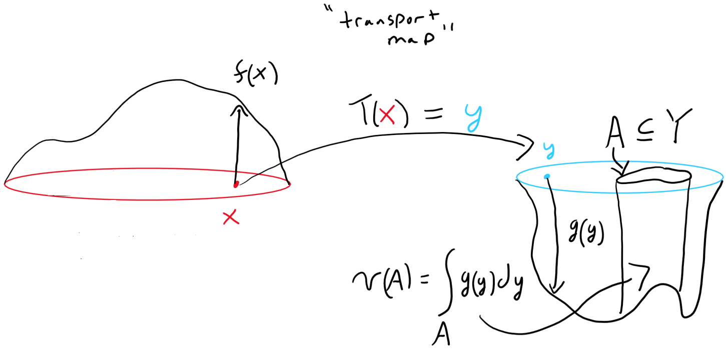
Origins of optimal transport

- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?

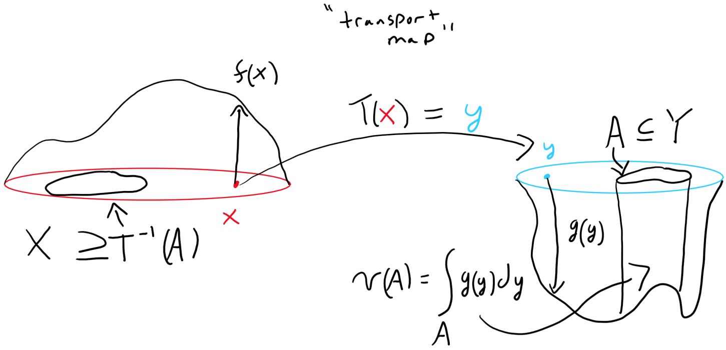


Origins of optimal transport

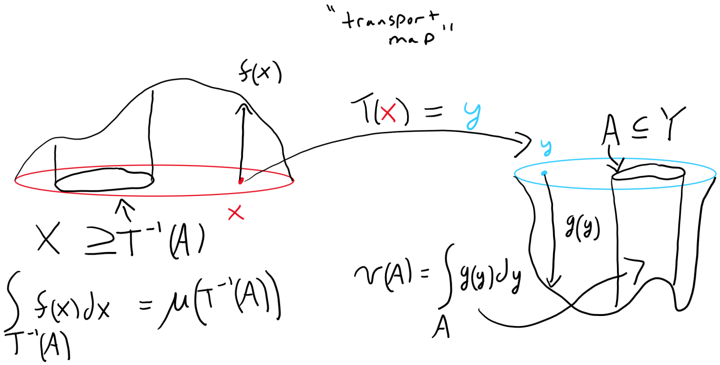
- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?



- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?

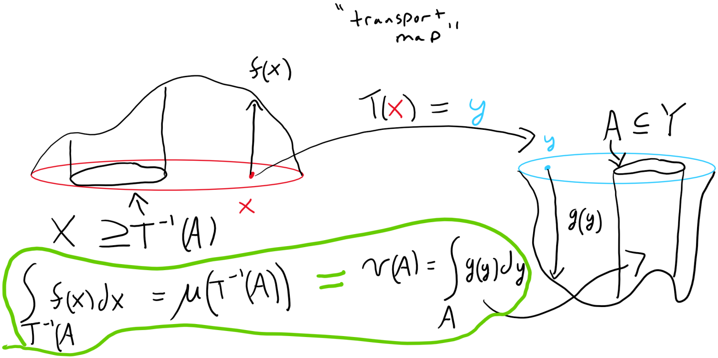


- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?



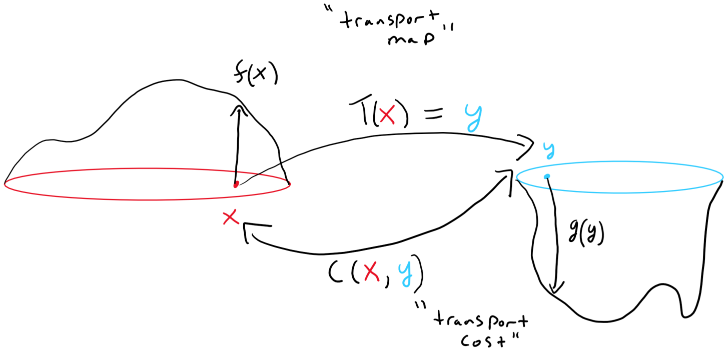
Origins of optimal transport

- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?

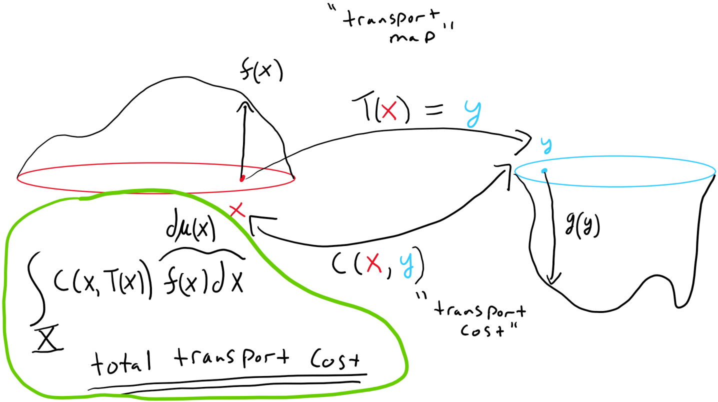


Origins of optimal transport

- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?



- Gaspard Monge (1781): How do I fill a hole with dirt as efficiently as possible?



- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on bounded domains $X, Y \subseteq \mathbb{R}^n$, we say a map $T : X \to Y$ **pushes** μ **forward to** ν, and write $T_{\#}\mu = \nu$, if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \subseteq Y$. We sometimes call these $T's$ **transport maps**.

Monge's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on bounded domains $X, Y \subseteq \mathbb{R}^n$, we say a map $T : X \to Y$ **pushes** μ **forward to** ν, and write $T_{\#}\mu = \nu$, if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \subseteq Y$. We sometimes call these $T's$ **transport maps**.
- Note: if $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$, and T is a diffeomorphism (ie, $1-1$, onto, smooth with a smooth inverse), this means T satisifes the change of variables equation $f(x) = |detDT(X)|g(T(x))$.

Monge's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on bounded domains $X, Y \subseteq \mathbb{R}^n$, we say a map $T : X \to Y$ **pushes** μ **forward to** ν, and write $T_{\#}\mu = \nu$, if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \subseteq Y$. We sometimes call these $T's$ **transport maps**.
- Note: if $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$, and T is a diffeomorphism (ie, $1-1$, onto, smooth with a smooth inverse), this means T satisifes the change of variables equation $f(x) = |detDT(X)|g(T(x))$.
- Given a cost function $c(x, y)$, **Monge's optimal transport problem** is to minimize:

$$\int_X c(x, T(x))d\mu(x)$$

among all T such that $T_{\#}\mu = \nu$.

Monge's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on bounded domains $X, Y \subseteq \mathbb{R}^n$, we say a map $T : X \to Y$ **pushes** μ **forward to** ν, and write $T_\# \mu = \nu$, if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \subseteq Y$. We sometimes call these $T's$ **transport maps**.
- Note: if $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$, and T is a diffeomorphism (ie, $1-1$, onto, smooth with a smooth inverse), this means T satisifes the change of variables equation $f(x) = |detDT(X)|g(T(x))$.
- Given a cost function $c(x,y)$, **Monge's optimal transport problem** is to minimize:

$$\int_X c(x, T(x))d\mu(x)$$

among all T such that $T_\# \mu = \nu$.

- Example costs: $c(x,y) = |x-y|, |x-y|^2 \ldots$.

Monge's optimal transport problem

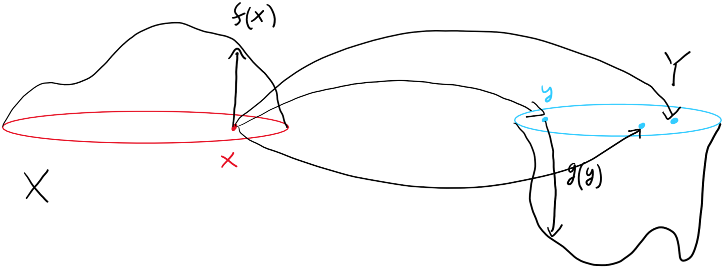
- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on bounded domains $X, Y \subseteq \mathbb{R}^n$, we say a map $T : X \to Y$ **pushes μ forward to ν**, and write $T_\#\mu = \nu$, if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \subseteq Y$. We sometimes call these $T's$ **transport maps**.
- Note: if $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$, and T is a diffeomorphism (ie, $1-1$, onto, smooth with a smooth inverse), this means T satisifes the change of variables equation $f(x) = |detDT(X)|g(T(x))$.
- Given a cost function $c(x, y)$, **Monge's optimal transport problem** is to minimize:

$$\int_X c(x, T(x))d\mu(x)$$

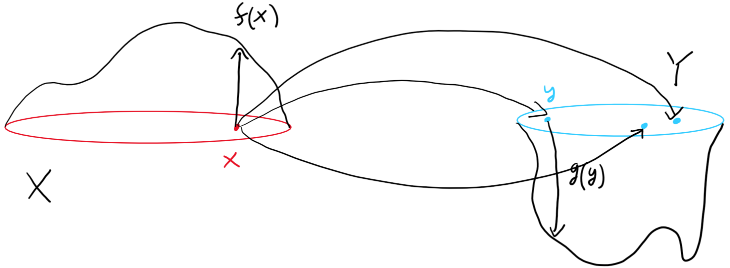
among all T such that $T_\#\mu = \nu$.
- Example costs: $c(x, y) = |x-y|, |x-y|^2 \dots$
- Challenging to analyze (lacks linearity, compactness...)

Kantorovich's optimal transport problem

Leonid Kantorovich 1942: instead of sending all the mass at the source point x to target point $y = T(x)$, allow **splitting**, so that the mass may be divided among **several** (or even infinitely many) target points.



Leonid Kantorovich 1942: instead of sending all the mass at the source point x to target point $y = T(x)$, allow **splitting**, so that the mass may be divided among **several** (or even infinitely many) target points.

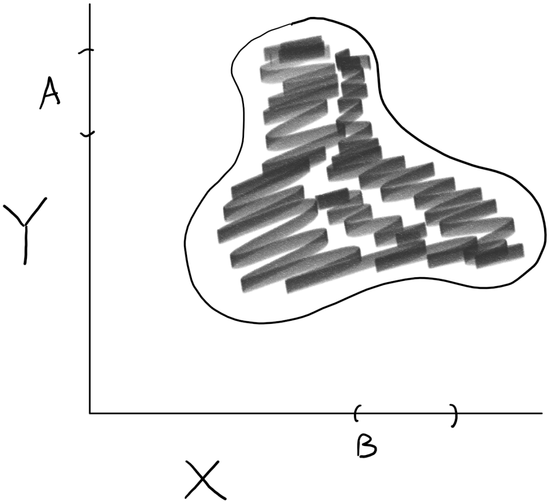


Intuitively, think of denoting the amount of mass moved from x to y by $\gamma({\color{red}x}, {\color{blue}y})$.

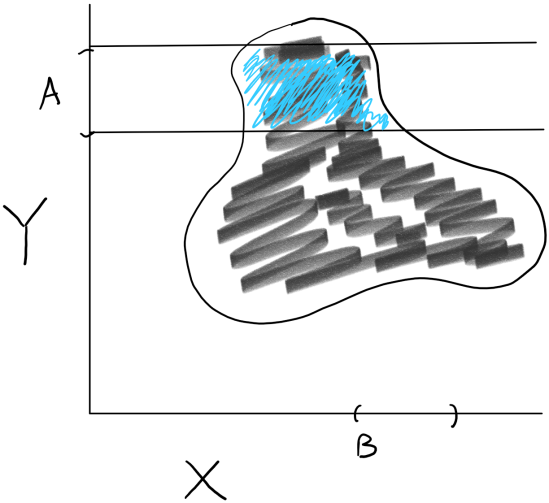
Kantorovich's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on domains $X, Y \subseteq \mathbb{R}^n$, we say a probability measure γ on $X \times Y$, has **marginals** μ and ν if $\gamma(B \times Y) = \mu(B)$ and $\gamma(X \times A) = \nu(A)$ for all $A \subseteq Y$ and $B \subseteq X$. We will sometimes call such γ's **transport plans**. We denote the set of all transport plans by $\Gamma(\mu, \nu)$.

Marginals

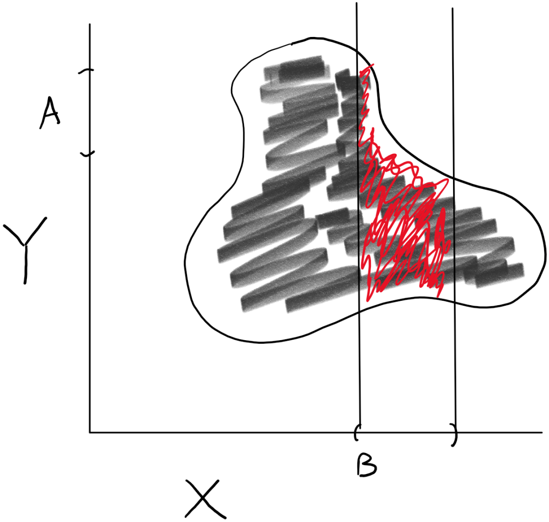


Marginals



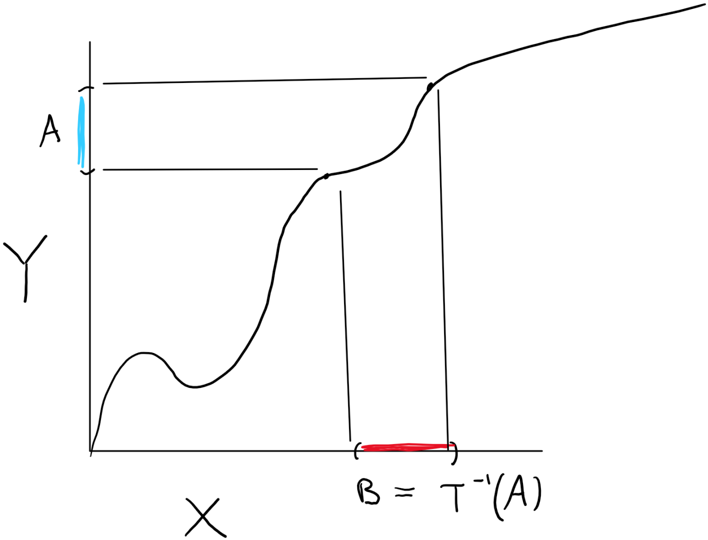
$\gamma(X \times A) = \nu(A)$

Marginals



$\gamma(B \times Y) = \mu(B)$

Marginals for a Monge type transport plan (transport map)



$$\nu(A) = \gamma(X \times A) = \gamma(B \times Y) = \mu(B) = \mu(T^{-1}(A))$$

Kantorovich's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on domains $X, Y \subseteq \mathbb{R}^n$, we say a probability measure γ on $X \times Y$, has **marginals** μ and ν if $\gamma(B \times Y) = \mu(B)$ and $\gamma(X \times A) = \nu(A)$ for all $A \subseteq Y$ and $B \subseteq X$. We will sometimes call such γ's transport plans. We denote the set of all **transport plans** by $\Gamma(\mu, \nu)$.
- Given a cost function $c(x, y)$, **Kantorovich's optimal transport problem** is to minimize:

$$\int_{X\times Y} c(x,y)d\gamma(x,y)$$

among all $\gamma \in \Gamma(\mu, \nu)$.

Kantorovich's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on domains $X, Y \subseteq \mathbb{R}^n$, we say a probability measure γ on $X \times Y$, has **marginals** μ and ν if $\gamma(B \times Y) = \mu(B)$ and $\gamma(X \times A) = \nu(A)$ for all $A \subseteq Y$ and $B \subseteq X$. We will sometimes call such γ's transport plans. We denote the set of all **transport plans** by $\Gamma(\mu, \nu)$.
- Given a cost function $c(x, y)$, **Kantorovich's optimal transport problem** is to minimize:

  $$\int_{X \times Y} c(x, y) d\gamma(x, y)$$

  among all $\gamma \in \Gamma(\mu, \nu)$.
- Linear minimization over a convex set. Under mild conditions, there exists a solution (continuity-compactness).

Kantorovich's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on domains $X, Y \subseteq \mathbb{R}^n$, we say a probability measure γ on $X \times Y$, has **marginals** μ and ν if $\gamma(B \times Y) = \mu(B)$ and $\gamma(X \times A) = \nu(A)$ for all $A \subseteq Y$ and $B \subseteq X$. We will sometimes call such γ's transport plans. We denote the set of all **transport plans** by $\Gamma(\mu, \nu)$.
- Given a cost function $c(x, y)$, **Kantorovich's optimal transport problem** is to minimize:

$$\int_{X \times Y} c(x, y) d\gamma(x, y)$$

  among all $\gamma \in \Gamma(\mu, \nu)$.
- Linear minimization over a convex set. Under mild conditions, there exists a solution (continuity-compactness).
- We will call the smallest closed set S such that $\gamma(S) = 1$ the **support** of γ, denoted by $spt(\gamma)$.

Kantorovich's optimal transport problem

- Given probability measures $\mu(x)$ (the source) and $\nu(y)$ (the target) on domains $X, Y \subseteq \mathbb{R}^n$, we say a probability measure γ on $X \times Y$, has **marginals** μ and ν if $\gamma(B \times Y) = \mu(B)$ and $\gamma(X \times A) = \nu(A)$ for all $A \subseteq Y$ and $B \subseteq X$. We will sometimes call such γ's transport plans. We denote the set of all **transport plans** by $\Gamma(\mu, \nu)$.
- Given a cost function $c(x, y)$, **Kantorovich's optimal transport problem** is to minimize:

$$\int_{X \times Y} c(x, y) d\gamma(x, y)$$

among all $\gamma \in \Gamma(\mu, \nu)$.
- Linear minimization over a convex set. Under mild conditions, there exists a solution (continuity-compactness).
- We will call the smallest closed set S such that $\gamma(S) = 1$ the **support** of γ, denoted by $spt(\gamma)$.
  - Ex: For Monge type γ, the graph of T is the support (for a continuous T).

Structure of solution: c-monotonicity

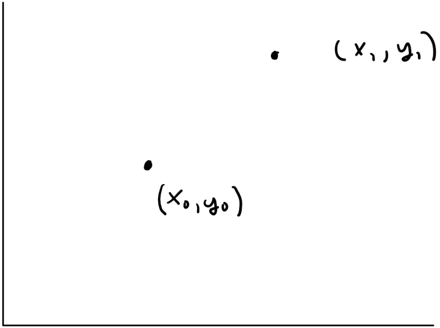
We say a set $S \subseteq X \times Y$ is c-**monotone** if for any (x_0, y_0), $(x_1, y_1) \in S$ we have

$$c(x_0, y_0) + c(x_1, y_1) \leq c(x_0, y_1) + c(x_1, y_0)$$

We say a set $S \subseteq X \times Y$ is c-**monotone** if for any (x_0, y_0), $(x_1, y_1) \in S$ we have

$$c(x_0, y_0) + c(x_1, y_1) \leq c(x_0, y_1) + c(x_1, y_0)$$

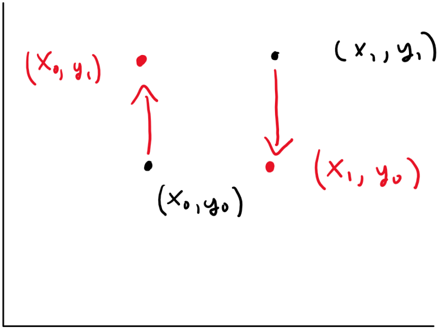
- The support of optimal transport plans is always c-monotone.



We say a set $S \subseteq X \times Y$ is c-**monotone** if for any (x_0, y_0), $(x_1, y_1) \in S$ we have

$$c(x_0, y_0) + c(x_1, y_1) \leq c(x_0, y_1) + c(x_1, y_0)$$

- The support of optimal transport plans is always c-monotone.

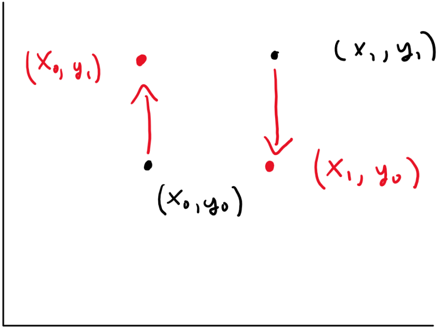


Structure of solution: c-monotonicity

We say a set $S \subseteq X \times Y$ is c-**monotone** if for any (x_0, y_0), $(x_1, y_1) \in S$ we have

$$c(x_0, y_0) + c(x_1, y_1) \leq c(x_0, y_1) + c(x_1, y_0)$$

- The support of optimal transport plans is always c-monotone.



- For $c(x, y) = |x - y|^2$, this amounts to $(x_1 - x_0) \cdot (y_1 - y_0) \geq 0$.

Structure of solution: c-cyclical monotonicity

- A set $S \subseteq X \times Y$ is c-**cyclically monotone** if for any finite collection of points $(x_0, y_0), (x_1, y_1), ...(x_N, y_N) \in S$ we have

$$\sum_{i=0}^{N} c(x_i, y_i) \leq \sum_{i=0}^{N} c(x_i, y_{i+1}) + c(x_N, y_0)$$

Structure of solution: c-cyclical monotonicity

- A set $S \subseteq X \times Y$ is c-**cyclically monotone** if for any finite collection of points $(x_0, y_0), (x_1, y_1), ...(x_N, y_N) \in S$ we have

$$\sum_{i=0}^{N} c(x_i, y_i) \leq \sum_{i=0}^{N} c(x_i, y_{i+1}) + c(x_N, y_0)$$

- The support of optimal transport plans is always c-cyclically monotone. In fact, this is a characterization.

Structure of solution: c-cyclical monotonicity

- A set $S \subseteq X \times Y$ is c-**cyclically monotone** if for any finite collection of points $(x_0, y_0), (x_1, y_1), ...(x_N, y_N) \in S$ we have

$$\sum_{i=0}^{N} c(x_i, y_i) \leq \sum_{i=0}^{N} c(x_i, y_{i+1}) + c(x_N, y_0)$$

- The support of optimal transport plans is always c-cyclically monotone. In fact, this is a characterization.
- For $c(x,y) = |x-y|^2$, this is

$$\sum_{i=0}^{N} x_i \cdot (y_i - y_{i+1}) + x_N \cdot (y_N - y_0) \geq 0$$

This is known simply as cyclical monotonicity.

One dimensional optimal transport

- Suppose $n = 1$: $X, Y \subset \mathbb{R}$.

One dimensional optimal transport

- Suppose $n = 1$: $X, Y \subset \mathbb{R}$.
- Assume $\frac{\partial^2 c}{\partial x \partial y} < 0$ (e.g. $c(x, y) = (x - y)^2$).

- Suppose $n = 1$: $X, Y \subset \mathbb{R}$.
- Assume $\frac{\partial^2 c}{\partial x \partial y} < 0$ (e.g. $c(x, y) = (x - y)^2$).
- If $(x_0, y_0), (x_1, y_1)$ are in the support of an optimal γ, c-monotonicity means that:

$$\begin{aligned} 0 &\geq c(x_1, y_1) + c(x_0, y_0) - c(x_0, y_1) - c(x_1, y_0) \\ &= \int_{y_0}^{y_1} [\frac{\partial c}{\partial y}(x_1, y) - \frac{\partial c}{\partial y}(x_0, y)] dy \\ &= \int_{y_0}^{y_1} \int_{x_0}^{x_1} \frac{\partial^2 c}{\partial x \partial y}(x, y) dx dy \end{aligned}$$

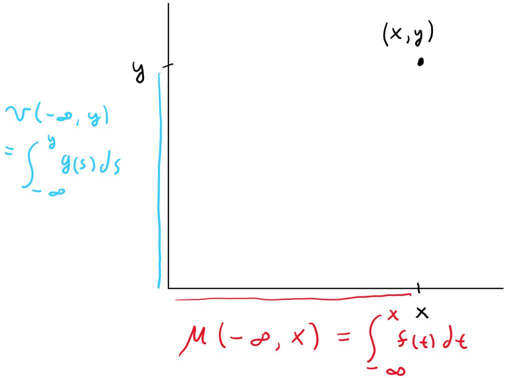
One dimensional optimal transport

- Suppose $n = 1$: $X, Y \subset \mathbb{R}$.
- Assume $\frac{\partial^2 c}{\partial x \partial y} < 0$ (e.g. $c(x,y) = (x-y)^2$).
- If $(x_0, y_0), (x_1, y_1)$ are in the support of an optimal γ, c-monotonicity means that:

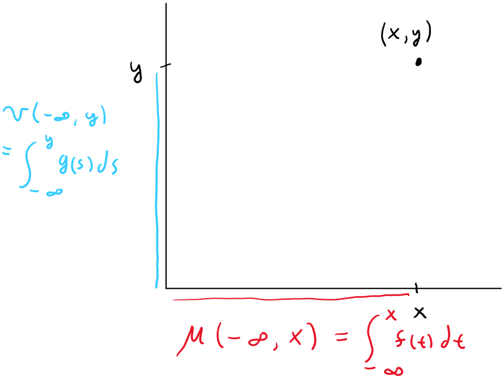
$$
\begin{aligned}
0 &\geq c(x_1, y_1) + c(x_0, y_0) - c(x_0, y_1) - c(x_1, y_0) \\
&= \int_{y_0}^{y_1} [\frac{\partial c}{\partial y}(x_1, y) - \frac{\partial c}{\partial y}(x_0, y)] dy \\
&= \int_{y_0}^{y_1} \int_{x_0}^{x_1} \frac{\partial^2 c}{\partial x \partial y}(x, y) dx dy
\end{aligned}
$$

$\implies (x_1 - x_0)(y_1 - y_0) \geq 0$. That is, γ is concentrated on a **monotone increasing** set.

One dimensional optimal transport



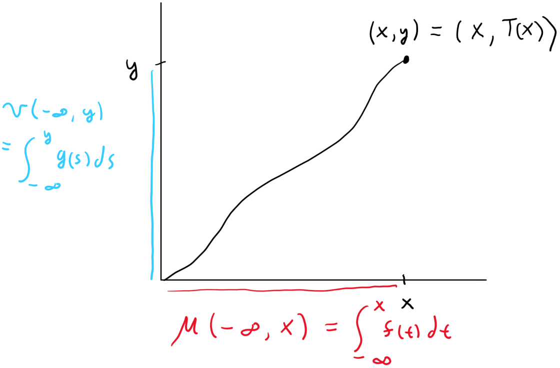
One dimensional optimal transport



For $d\mu(x) = f(x)dx$, the only transport plan with monotone increasing support is concentrated on the graph of $T : X \to Y$ defined by:

$$\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{T(x)} g(s)ds$$

One dimensional optimal transport



For $d\mu(x) = f(x)dx$, the only transport plan with monotone increasing support is concentrated on the graph of $T : X \to Y$ defined by:

$$\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{T(x)} g(s)ds$$

Remarks on one dimensional optimal transport

- The T defined above is the unique solution to Monge's optimal transport problem. The corresponding measure $\gamma = (Id, T)_{\#}\mu$ is the unique solution to Kantorovich's problem.

Remarks on one dimensional optimal transport

- The T defined above is the unique solution to Monge's optimal transport problem. The corresponding measure $\gamma = (Id, T)_{\#}\mu$ is the unique solution to Kantorovich's problem.
- The solution doesn't really depend on c (as long as $\frac{\partial^2 c}{\partial x \partial y} < 0$). Therefore, the solutions for all costs satisfying this condition (for the same marginals) are the same. This is a special feature of the one dimensional setting.

Remarks on one dimensional optimal transport

- The T defined above is the unique solution to Monge's optimal transport problem. The corresponding measure $\gamma = (Id, T)_{\#}\mu$ is the unique solution to Kantorovich's problem.
- The solution doesn't really depend on c (as long as $\frac{\partial^2 c}{\partial x \partial y} < 0$). Therefore, the solutions for all costs satisfying this condition (for the same marginals) are the same. This is a special feature of the one dimensional setting.
- For probabilistically minded people, this is $T = (F_\nu)^{-1} \circ F_\mu$, where F_ν and F_μ are the cumulative distribution functions.

Remarks on one dimensional optimal transport

- The T defined above is the unique solution to Monge's optimal transport problem. The corresponding measure $\gamma = (Id, T)_{\#}\mu$ is the unique solution to Kantorovich's problem.
- The solution doesn't really depend on c (as long as $\frac{\partial^2 c}{\partial x \partial y} < 0$). Therefore, the solutions for all costs satisfying this condition (for the same marginals) are the same. This is a special feature of the one dimensional setting.
- For probabilistically minded people, this is $T = (F_\nu)^{-1} \circ F_\mu$, where F_ν and F_μ are the cumulative distribution functions.
- Differentiating $\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{T(x)} g(s)ds$ with respect to x yields an ODE for T.

$$f(x) = T'(x)g(T(x))$$

Duality

- Kantorovich's optimal transport problem between μ and ν is dual to the problem of maximizing:

$$\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

among functions u on X and v on Y such that $u(x) + v(y) \leq c(x, y)$ for all $(x, y) \in X \times Y$

Duality

- Kantorovich's optimal transport problem between μ and ν is dual to the problem of maximizing:

$$\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

among functions u on X and v on Y such that $u(x) + v(y) \leq c(x,y)$ for all $(x,y) \in X \times Y$

- **Kantorovich duality theorem:**

$$\max_{u+v\leq c} \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y)d\gamma(x,y)$$

Duality

- Kantorovich's optimal transport problem between μ and ν is dual to the problem of maximizing:

$$\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

  among functions u on X and v on Y such that $u(x) + v(y) \leq c(x,y)$ for all $(x,y) \in X \times Y$

- **Kantorovich duality theorem:**

$$\max_{u+v\leq c} \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y)d\gamma(x,y)$$

- The "$\leq$" direction is easy to prove (just integrate both sides of the constraint $u(x) + v(y) \leq c(x,y)$ against γ).

Duality

- Kantorovich's optimal transport problem between μ and ν is dual to the problem of maximizing:

$$\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

among functions u on X and v on Y such that $u(x) + v(y) \leq c(x,y)$ for all $(x,y) \in X \times Y$

- **Kantorovich duality theorem:**

$$\max_{u+v\leq c} \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y)d\gamma(x,y)$$

- The "$\leq$" direction is easy to prove (just integrate both sides of the constraint $u(x) + v(y) \leq c(x,y)$ against γ).
- The duality is a key tool in analysis of OT problems.

Duality: sketch of proof

- Minimax theory for

$$
\begin{aligned}
H(\gamma, u, v) &= \int_{X\times Y}[c(x,y) - u(x) - v(y)]d\gamma(x,y) \\
&\quad + \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)
\end{aligned}
$$

Duality: sketch of proof

- Minimax theory for

$$\begin{aligned} H(\gamma, u, v) &= \int_{X\times Y} [c(x,y) - u(x) - v(y)] d\gamma(x,y) \\ &+ \int_X u(x) d\mu(x) + \int_Y v(y) d\nu(y) \end{aligned}$$

- For fixed γ the **unconstrained** supremum of $(u,v) \mapsto H(\gamma, u, v)$ is $\int_{X\times Y} c(x,y) d\gamma(x,y)$ **if** $\gamma \in \Gamma(\mu,\nu)$, and ∞ otherwise.
  - So $\inf_\gamma \sup_{(u,v)} H(\gamma,\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y) d\gamma(x,y)$ (the Kantorovich primal problem)

Duality: sketch of proof

- Minimax theory for

$$
\begin{aligned}
H(\gamma, u, v) \;\; &= \;\; \int_{X\times Y} [c(x,y) - u(x) - v(y)] d\gamma(x,y) \\
&\quad + \int_X u(x) d\mu(x) + \int_Y v(y) d\nu(y)
\end{aligned}
$$

- For fixed γ the **unconstrained** supremum of $(u,v) \mapsto H(\gamma, u, v)$ is $\int_{X\times Y} c(x,y)d\gamma(x,y)$ **if** $\gamma \in \Gamma(\mu,\nu)$, and ∞ otherwise.
    - So $\inf_\gamma \sup_{(u,v)} H(\gamma, \mu, \nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x,y)d\gamma(x,y)$ (the Kantorovich primal problem)
- For fixed (u,v), the **unconstrained** infimum of $\gamma \mapsto H(\gamma, u, v)$ is $\int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$ if $c(x,y) \geq u(x) + v(y)$ everywhere and $-\infty$ otherwise.

Duality: sketch of proof

- Minimax theory for

$$H(\gamma,u,v) = \int_{X\times Y}[c(x,y)-u(x)-v(y)]d\gamma(x,y) + \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

- For fixed γ the **unconstrained** supremum of $(u,v)\mapsto H(\gamma,u,v)$ is $\int_{X\times Y} c(x,y)d\gamma(x,y)$ **if** $\gamma\in\Gamma(\mu,\nu)$, and ∞ otherwise.
  - So $\inf_\gamma \sup_{(u,v)} H(\gamma,\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)}\int_{X\times Y} c(x,y)d\gamma(x,y)$ (the Kantorovich primal problem)
- For fixed (u,v), the **unconstrained** infimum of $\gamma\mapsto H(\gamma,u,v)$ is $\int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)$ if $c(x,y)\geq u(x)+v(y)$ everywhere and $-\infty$ otherwise.
  - So $\sup_{(u,v)}\inf_\gamma H(\gamma,\mu,\nu) =$ $\sup_{u+v\leq c}\int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)$ (the dual problem)

- Minimax theory for

$$H(\gamma,u,v) = \int_{X\times Y}[c(x,y)-u(x)-v(y)]d\gamma(x,y) + \int_X u(x)d\mu(x) + \int_Y v(y)d\nu(y)$$

- For fixed γ the **unconstrained** supremum of $(u,v)\mapsto H(\gamma,u,v)$ is $\int_{X\times Y}c(x,y)d\gamma(x,y)$ **if** $\gamma\in\Gamma(\mu,\nu)$, and ∞ otherwise.
    - So $\inf_\gamma \sup_{(u,v)} H(\gamma,\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)}\int_{X\times Y}c(x,y)d\gamma(x,y)$ (the Kantorovich primal problem)
- For fixed (u,v), the **unconstrained** infimum of $\gamma\mapsto H(\gamma,u,v)$ is $\int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)$ if $c(x,y)\geq u(x)+v(y)$ everywhere and $-\infty$ otherwise.
    - So $\sup_{(u,v)}\inf_\gamma H(\gamma,\mu,\nu) = \sup_{u+v\leq c}\int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)$ (the dual problem)
- Applying a minimax theorem, $\inf_\gamma\sup_{(u,v)}H(\gamma,\mu,\nu) = \sup_{(u,v)}\inf_\gamma H(\gamma,\mu,\nu)$ gives duality.

More on duality: key facts

- Suppose that $u(x)$ and $v(y)$ solve the dual problem, and $\gamma(x,y)$ solves the primal. Then, since $u(x)+v(y)\leq c(x,y)$ **everywhere**, but

$$\begin{aligned}\int_{X\times Y}[u(x)+v(y)]d\gamma(x,y) &= \int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)\\ &= \int_{X\times Y} c(x,y)d\gamma(x,y)\end{aligned}$$

we must have

$$u(x)+v(y)-c(x,y)=0$$

γ -almost everywhere.

More on duality: key facts

- Suppose that $u(x)$ and $v(y)$ solve the dual problem, and $\gamma(x,y)$ solves the primal. Then, since $u(x)+v(y)\leq c(x,y)$ **everywhere**, but

$$\begin{aligned}\int_{X\times Y}[u(x)+v(y)]d\gamma(x,y) &= \int_X u(x)d\mu(x)+\int_Y v(y)d\nu(y)\\ &= \int_{X\times Y} c(x,y)d\gamma(x,y)\end{aligned}$$

  we must have

$$u(x)+v(y)-c(x,y)=0$$

  γ -almost everywhere.

- This means that $x\mapsto u(x)-c(x,\bar{y})$ is **maximized** at any $x=\bar{x}$ such that $(\bar{x},\bar{y})\in spt(\gamma)$ (that is, any $\bar{x}$ that gets transported to $\bar{y}$ by γ). So

$$\nabla u(\bar{x})=\nabla_x c(\bar{x},\bar{y}) \text{ and } D^2u(\bar{x})\leq D^2_{xx}c(\bar{x},\bar{y})$$

Brenier's Theorem

Brenier's Theorem: Suppose that μ is absolutely continuous with respect to Lebesgue measure and that $c(x,y) = \frac{|x-y|^2}{2}$. Then the solution γ to the Kantorovich problem is unique and concentrated on the graph of a function $T : X \to Y$. Furthermore, $T(x) = \nabla\phi(x)$ for a convex function ϕ, and T is the unique solution to the Monge problem.

Brenier's Theorem: Suppose that μ is absolutely continuous with respect to Lebesgue measure and that $c(x, y) = \frac{|x-y|^2}{2}$. Then the solution γ to the Kantorovich problem is unique and concentrated on the graph of a function $T : X \to Y$. Furthermore, $T(x) = \nabla\phi(x)$ for a convex function ϕ, and T is the unique solution to the Monge problem.

Remarks:

- In $n = 1$ dimension, convexity means that $T'(x) = \phi''(x) \geq 0$, so that T is monotone increasing – we recover our earlier result.
- A non-trivial part of the theorem is that there exists a (unique) convex function ϕ so that $\nabla\phi_{\#}\mu = \nu$. This fact by itself has many applications.

Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x, y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x, y) = x - y$$

Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x,y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x,y) = x - y$$

- So we must have $y = x - \nabla u(x) = \nabla(\frac{|x|^2}{2} - u)(x) := \nabla\phi(x) := T(x).$

Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x, y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x, y) = x - y$$

- So we must have $y = x - \nabla u(x) = \nabla(\frac{|x|^2}{2} - u)(x) := \nabla \phi(x) := T(x).$
- Now, $D^2 u(x) \leq D^2_{xx} c(x, y) = I$, so $D^2 \phi(x) = D^2(\frac{|x|^2}{2} - u)(x) = I - D^2 u(x) \geq 0.$ So ϕ is convex.

Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x,y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x,y) = x - y$$

- So we must have $y = x - \nabla u(x) = \nabla(\frac{|x|^2}{2} - u)(x) := \nabla \phi(x) := T(x)$.
- Now, $D^2 u(x) \leq D^2_{xx} c(x,y) = I$, so $D^2\phi(x) = D^2(\frac{|x|^2}{2} - u)(x) = I - D^2 u(x) \geq 0$. So ϕ is convex.
- What about uniqueness? If γ_0 and γ_1 both solve Kantorovich's problem, so does $\gamma_{1/2} = \frac{1}{2}[\gamma_0 + \gamma_1]$, by linearity.

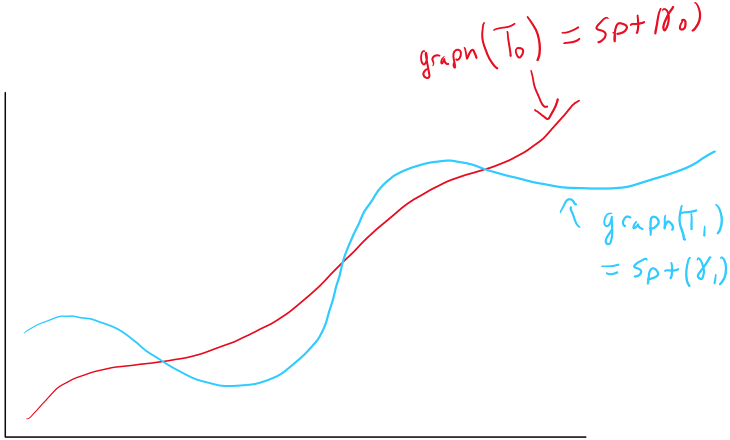
Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x,y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x,y) = x - y$$

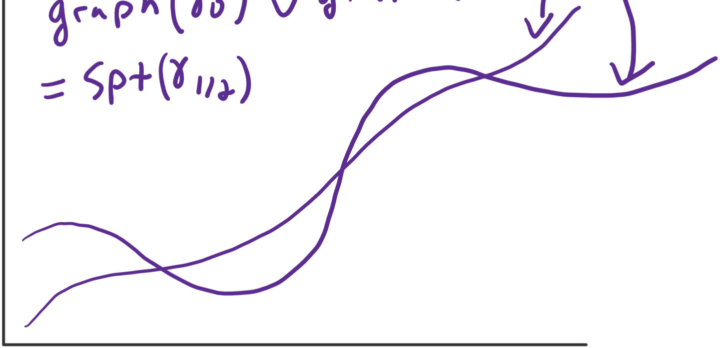
- So we must have $y = x - \nabla u(x) = \nabla(\frac{|x|^2}{2} - u)(x) := \nabla \phi(x) := T(x)$.
- Now, $D^2 u(x) \leq D^2_{xx} c(x,y) = I$, so $D^2 \phi(x) = D^2(\frac{|x|^2}{2} - u)(x) = I - D^2 u(x) \geq 0$. So ϕ is convex.
- What about uniqueness? If γ_0 and γ_1 both solve Kantorovich's problem, so does $\gamma_{1/2} = \frac{1}{2}[\gamma_0 + \gamma_1]$, by linearity.
- The argument above shows that γ_0 and γ_1 concentrate on graphs of functions T_0 and T_1. $\gamma_{1/2}$ then concentrates on the union of theses two graphs.

Graphical supports and their union



Graphical supports and their union

$\mathrm{graph}(\gamma_0) \cup \mathrm{graph}(\gamma_1)$
$= \mathrm{spt}(\gamma_{1/2})$



Brenier's Theorem: sketch of proof

- Let γ solve the primal problem and $u(x)$, $v(y)$ solve the dual. We have, for any $(x, y) \in spt(\gamma)$,

$$\nabla u(x) = \nabla_x c(x, y) = x - y$$

- So we must have $y = x - \nabla u(x) = \nabla(\frac{|x|^2}{2} - u)(x) := \nabla \phi(x) := T(x)$.
- Now, $D^2 u(x) \leq D^2_{xx} c(x, y) = I$, so $D^2 \phi(x) = D^2(\frac{|x|^2}{2} - u)(x) = I - D^2 u(x) \geq 0$. So ϕ is convex.
- What about uniqueness? If γ_0 and γ_1 both solve Kantorovich's problem, so does $\gamma_{1/2} = \frac{1}{2}[\gamma_0 + \gamma_1]$, by linearity.
- The argument above shows that γ_0 and γ_1 concentrate on graphs of functions T_0 and T_1. $\gamma_{1/2}$ then concentrates on the **union** of theses two graphs.
- But $\gamma_{1/2}$ concentrates on a graph, too (using the argument above again), which is impossible unless $T_0 = T_1$ in which case $\gamma_0 = \gamma_1$.

Remarks

- The fact that the solution is graphical and unique doesn't really rely on c being quadratic. The same conclusions can be drawn for more general costs satisfying the **twist condition**, which is injectivity of $y \mapsto \nabla_x c(x,y)$ for each fixed x.

Remarks

- The fact that the solution is graphical and unique doesn't really rely on c being quadratic. The same conclusions can be drawn for more general costs satisfying the **twist condition**, which is injectivity of $y \mapsto \nabla_x c(x,y)$ for each fixed x.
- For $c(x,y) = \frac{|x-y|^2}{2}$, note that $DT(x) = D^2\phi(x)$, so $\frac{f(x)}{g(\nabla\phi(x))} = |det(DT(x))| = detD^2\phi(x)$. That is, ϕ solves a Monge-Ampere equation.

Remarks

- The fact that the solution is graphical and unique doesn't really rely on c being quadratic. The same conclusions can be drawn for more general costs satisfying the **twist condition**, which is injectivity of $y \mapsto \nabla_x c(x,y)$ for each fixed x.
- For $c(x,y) = \frac{|x-y|^2}{2}$, note that $DT(x) = D^2\phi(x)$, so $\frac{f(x)}{g(\nabla\phi(x))} = |det(DT(x))| = detD^2\phi(x)$. That is, ϕ solves a Monge-Ampere equation.
- Instead of using duality, one could use Rockafellar's theorem (a set $S \subseteq \mathbb{R}^n \times \mathbb{R}^n$ is cyclically monotone if any only if it is contained in the subdifferential of a convex funtion).

Brenier map: examples

- If μ is uniform on a ball, $B(0,1)$ and ν uniform on the corresponding sphere, $\partial B(0,1)$, then

- If μ is uniform on a ball, $B(0,1)$ and ν uniform on the corresponding sphere, $\partial B(0,1)$, then

$$\phi(x) = |x| \text{ and } \nabla\phi(x) = \frac{x}{|x|}$$

Brenier map: examples

- If μ is uniform on a ball, $B(0,1)$ and ν uniform on the corresponding sphere, $\partial B(0,1)$, then

$$\phi(x) = |x| \text{ and } \nabla\phi(x) = \frac{x}{|x|}$$

- If μ and ν are Gaussians, $\mu = \mathcal{N}(0, I)$, $\nu = \mathcal{N}(\bar{y}, \Sigma)$, (so $f(x) = \frac{e^{-|x|^2/2}}{\sqrt{(2\pi)^n}}$ and $g(y) = \frac{e^{-(y-\bar{y})^T\Sigma^{-1}(y-\bar{y})/2}}{\sqrt{(2\pi)^n det(\Sigma)}}$) then

Brenier map: examples

- If μ is uniform on a ball, $B(0,1)$ and ν uniform on the corresponding sphere, $\partial B(0,1)$, then

$$\phi(x) = |x| \text{ and } \nabla\phi(x) = \frac{x}{|x|}$$

- If μ and ν are Gaussians, $\mu = \mathcal{N}(0,I)$, $\nu = \mathcal{N}(\bar{y},\Sigma)$, (so $f(x) = \frac{e^{-|x|^2/2}}{\sqrt{(2\pi)^n}}$ and $g(y) = \frac{e^{-(y-\bar{y})^T\Sigma^{-1}(y-\bar{y})/2}}{\sqrt{(2\pi)^n det(\Sigma)}}$) then

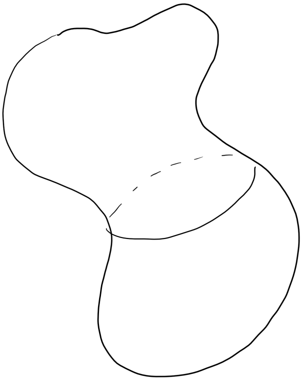
$$\phi(x) = \bar{y}\cdot x + \frac{1}{2}x^T\Sigma^{1/2}x, \text{ and } \nabla\phi(x) = \bar{y} + \Sigma^{1/2}x$$

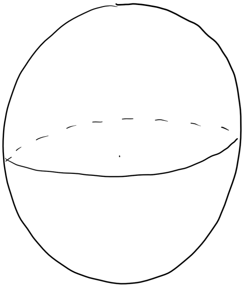
Isoperimetric inequality: The surface area of any set $M \subseteq \mathbb{R}^n$ is greater than or equal to the surface area of a ball with the same volume.

$$Vol(M) = Vol(B_R(0)) \implies S(M) \geq S(B_R(0))$$

Isoperimetric inequality

M





$B_R(0)$

Application: isoperimetric inequality

Isoperimetric inequality: The surface area of any set $M \subseteq \mathbb{R}^n$ is greater than or equal to the surface area of a ball with the same volume.

$$Vol(M) = Vol(B_R(0)) \implies S(M) \geq S(B_R(0))$$

Application: isoperimetric inequality

Isoperimetric inequality: The surface area of any set $M \subseteq \mathbb{R}^n$ is greater than or equal to the surface area of a ball with the same volume.

$$Vol(M) = Vol(B_R(0)) \implies S(M) \geq S(B_R(0))$$

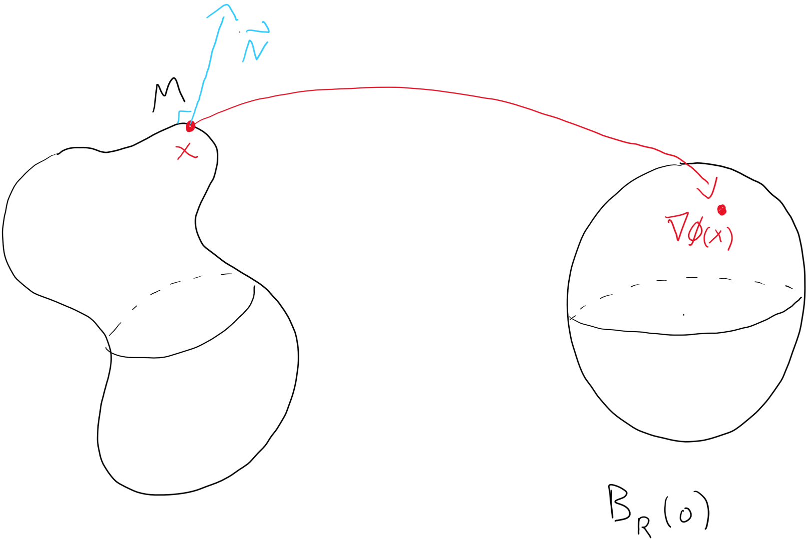
Proof:

- Take $f(x) = \chi_M$, $g(y) = \chi_{B_R(0)}$.
- $\nabla\phi(x)$ the Brenier map
  $\implies det(D^2\phi(x)) = f(x)/g(\nabla\phi(x)) = 1$ (change of variables).
- Arithmetic mean dominates geometric mean (as ϕ is convex, $D^2\phi$ has positive eigenvalues)
  $\implies det^{1/n}(D^2\phi(x)) \leq \frac{1}{n} tr(D^2\phi(x)) = \frac{1}{n}\Delta\phi(x)$

Proof sketch:

$$
\begin{aligned}
\frac{1}{n}S(B_R(0))R = Vol(B_R(0)) &= Vol(M) \\
&= \int_M 1 d^n x \\
&= \int_M det^{1/n}(D^2\phi(x))dx \\
&\leq \int_M \frac{1}{n}\Delta\phi(x)dx \\
&= \frac{1}{n}\int_{\partial M} \nabla\phi(x)\cdot\vec{N}d^{n-1}S(x)
\end{aligned}
$$

Isoperimetric inequality



Proof sketch:

$$
\begin{aligned}
\frac{1}{n}S(B_R(0))R = Vol(B_R(0)) &= Vol(M) \\
&= \int_M 1 d^n x \\
&= \int_M det^{1/n}(D^2\phi(x))dx \\
&\leq \int_M \frac{1}{n}\Delta\phi(x)dx \\
&= \frac{1}{n}\int_{\partial M} \nabla\phi(x)\cdot \vec{N} d^{n-1}S(x) \\
&\leq \frac{1}{n}\int_{\partial M} R d^{n-1}S(x) \\
&= \frac{1}{n}S(M)R
\end{aligned}
$$

Comments on the proof

Comments on the proof

- The optimal transport proof is pretty **simple**; everything in the proof is first or second year mathematics (*except* Brenier's theorem)!

Comments on the proof

- The optimal transport proof is pretty **simple**; everything in the proof is first or second year mathematics (*except* Brenier's theorem)!
- We prove an inequality about surfaces/curves/bodies in $\mathbb{R}^n$ by working with simple inequalities under the integral sign (geometric-arithmetic mean, Cauchy-Schwartz on $\mathbb{R}^n$).

Comments on the proof

- The optimal transport proof is pretty **simple**; everything in the proof is first or second year mathematics (*except* Brenier's theorem)!
- We prove an inequality about surfaces/curves/bodies in $\mathbb{R}^n$ by working with simple inequalities under the integral sign (geometric-arithmetic mean, Cauchy-Schwartz on $\mathbb{R}^n$).
- This is a **common theme** in applications of optimal transport in geometry.

The Wasserstein distance

- Optimal transport can be used to derive a metric on the space of probability measures, which we call the Wasserstein distance:

$$W_2(\mu,\nu) := \sqrt{\min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} |x-y|^2 d\gamma(x,y)}$$

The Wasserstein distance

- Optimal transport can be used to derive a metric on the space of probability measures, which we call the Wasserstein distance:

$$W_2(\mu,\nu) := \sqrt{\min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} |x-y|^2 d\gamma(x,y)}$$

- This is useful in a variety of applications when we want to compare two distributions of mass and the underlying distance plays a role.

The Wasserstein distance

- Optimal transport can be used to derive a metric on the space of probability measures, which we call the Wasserstein distance:

$$W_2(\mu,\nu) := \sqrt{\min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} |x-y|^2 d\gamma(x,y)}$$

- This is useful in a variety of applications when we want to compare two distributions of mass and the underlying distance plays a role.
- It works with discrete measures; in fact, $x \mapsto \delta_x$ isometrically embeds $\mathbb{R}^n$ into the space of probability measures.

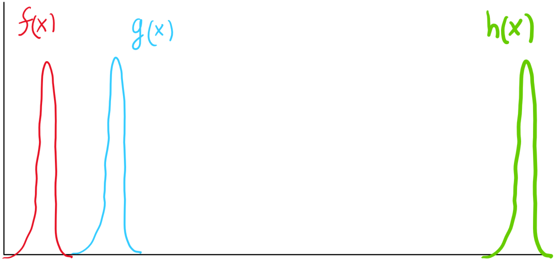
The Wasserstein distance

- Optimal transport can be used to derive a metric on the space of probability measures, which we call the Wasserstein distance:

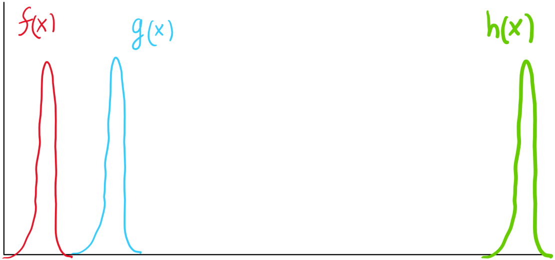
$$W_2(\mu,\nu) := \sqrt{\min_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} |x-y|^2 d\gamma(x,y)}$$

- This is useful in a variety of applications when we want to compare two distributions of mass and the underlying distance plays a role.
- It works with discrete measures; in fact, $x \mapsto \delta_x$ isometrically embeds $\mathbb{R}^n$ into the space of probability measures.
- In one dimension, $W_2^2(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^2 dt$, where F_μ and F_ν are the cdfs (ie, we compare μ and ν via their quantiles.)

Wasserstein distance



Wasserstein distance



Three probability measures: $du(x) = f(x)dx$, $d\nu(x) = g(x)dx$, $d\sigma(x) = h(x)dx$. Note that

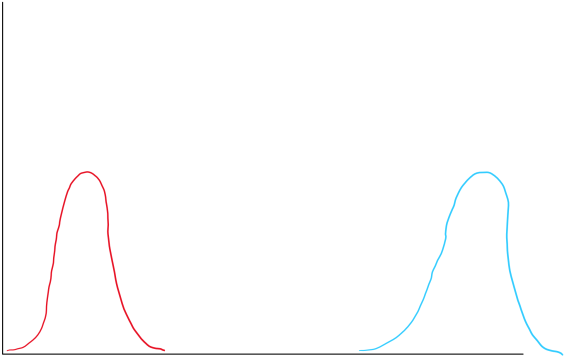
$$||f - g||_{L^2} = ||f - h||_{L^2}$$

but

$$W_2(\mu, \nu) << W_2(\mu, \sigma)$$

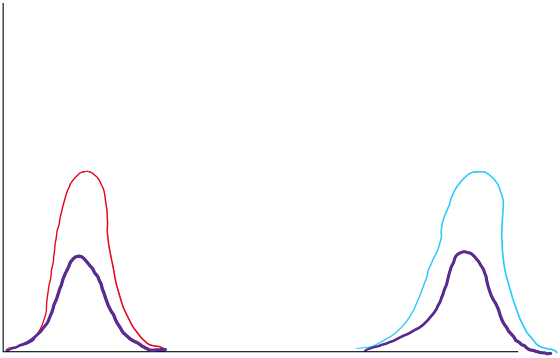
- Displacement interpolants are a natural way to interpolate (or average) between two probability measures, respecting the underlying geometry.
- The displacement interpolant between μ_0 and μ_1 is the curve of measures $\mu_t := ((1-t)I + t\nabla\phi)_{\#}\mu_0$, where $\nabla\phi$ is the Brenier (optimal transport) map between μ_0 and μ_1.

Displacement interpolation vs linear averages



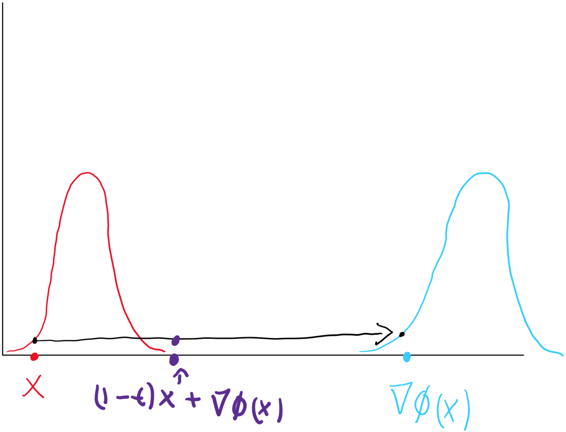
Probability measures μ_0 and μ_1.

Displacement interpolation vs linear averages



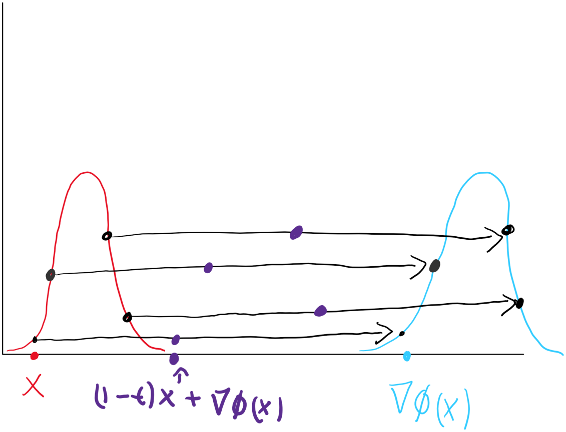
Probability measures μ_0 and μ_1. The linear interpolant $\mu_t = (1-t)\mu_0 + t\mu_1$

Displacement interpolation vs linear averages



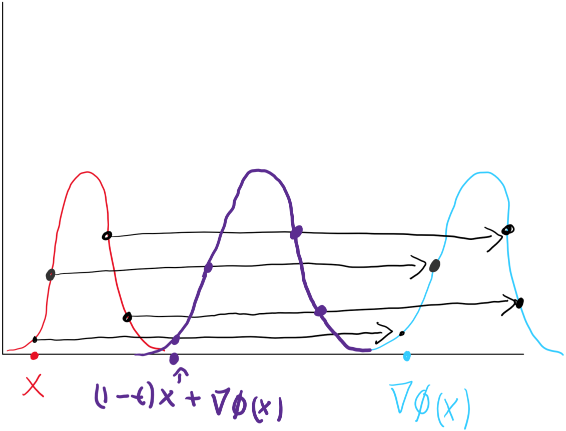
Probability measures μ_0 and μ_1.

Displacement interpolation vs linear averages



Probability measures μ_0 and μ_1.

Displacement interpolation vs linear averages



Probability measures μ_0 and μ_1. The displacement interpolant $\mu_t = ((1-t)Id + t\nabla\phi)_{\#}\mu_0$

Comments on displacement interpolation

- In one dimension, optimal transport matches equal quantiles.

Comments on displacement interpolation

- In one dimension, optimal transport matches equal quantiles.
- Dispacement interpolation then interpolates between quantiles:
  - If μ_0 and μ_1 have cdfs F_0 and F_1, the cdf of μ_t is

$$F_t = [(1-t)F_0^{-1} + tF_1^{-1}]^{-1}$$

Comments on displacement interpolation

- In one dimension, optimal transport matches equal quantiles.
- Dispacement interpolation then interpolates between quantiles:
  - If μ_0 and μ_1 have cdfs F_0 and F_1, the cdf of μ_t is

$$F_t = [(1-t)F_0^{-1} + tF_1^{-1}]^{-1}$$

- In general, displacement interpolation tends to preserve shapes/geometric features better than other interpolation methods.

Dynamic formulation of optimal transport (Benamou-Brenier)

- There is also a way to view the Wasserstein distance through action minimizing curves in the space of probability measures.

Dynamic formulation of optimal transport (Benamou-Brenier)

- There is also a way to view the Wasserstein distance through action minimizing curves in the space of probability measures.
- For an absolutely continuous probability measure at time t, $d\mu_t(x) = f_t(x)dx$ with density $f_t(x)$, we let $v_t(x)$ be the velocity of a particle at point x and time t.

Dynamic formulation of optimal transport (Benamou-Brenier)

- There is also a way to view the Wasserstein distance through action minimizing curves in the space of probability measures.
- For an absolutely continuous probability measure at time t, $d\mu_t(x) = f_t(x)dx$ with density $f_t(x)$, we let $v_t(x)$ be the velocity of a particle at point x and time t.
  - The divergence $\nabla \cdot (v_t(x)f_t(x))$ tells us how much mass is moving away from, or towards, the point x at time t.
  - Conservation of mass gives us the continuity equation: $f_t'(x) + \nabla \cdot (v_t(x)f_t(x)) = 0$.

Dynamic formulation of optimal transport (Benamou-Brenier)

- Recall that in $\mathbb{R}^n$, the squared distance $|x_0 - x_1|^2$ is given by $\min \int_0^1 |v_t|^2 dt$, where the minimum is among curves x_t joining x_0 and x_1 with velocity $v_t = x_t'$.
    - The optimal curve is the straight line $x_t = x_0 + t(x_1 - x_0)$ (so that $v_t = x_1 - x_0$ is constant).

Dynamic formulation of optimal transport (Benamou-Brenier)

- Recall that in $\mathbb{R}^n$, the squared distance $|x_0 - x_1|^2$ is given by $\min \int_0^1 |v_t|^2 dt$, where the minimum is among curves x_t joining x_0 and x_1 with velocity $v_t = x_t'$.
    - The optimal curve is the straight line $x_t = x_0 + t(x_1 - x_0)$ (so that $v_t = x_1 - x_0$ is constant).
- Analagously, the squared Wasserstein distance between probability measures μ_0 and μ_1 may be written

$$W_2^2(\mu_0, \mu_1) = \min \int_0^1 \int_{\mathbb{R}^n} f_t(x)|v_t(x)|^2 dxdt$$

among curves $d\mu_t = f_t dx$ joining μ_0 and μ_t, satisfying the continuity equation.

Dynamic formulation of optimal transport (Benamou-Brenier)

- Recall that in $\mathbb{R}^n$, the squared distance $|x_0 - x_1|^2$ is given by $\min \int_0^1 |v_t|^2 dt$, where the minimum is among curves x_t joining x_0 and x_1 with velocity $v_t = x_t'$.
  - The optimal curve is the straight line $x_t = x_0 + t(x_1 - x_0)$ (so that $v_t = x_1 - x_0$ is constant).
- Analagously, the squared Wasserstein distance between probability measures μ_0 and μ_1 may be written

$$W_2^2(\mu_0, \mu_1) = \min \int_0^1 \int_{\mathbb{R}^n} f_t(x)|v_t(x)|^2 dxdt$$

  among curves $d\mu_t = f_t dx$ joining μ_0 and μ_t, satisfying the continuity equation.
- The optimal velocity, $v_t(x) = \nabla\phi(x) - x$ is exactly the Brenier map.

Wasserstein gradient flows

- Certain evolution PDEs can be interpreted as gradient flows in the Wasserstein space: that is, a distribution of mass is rearranging itself so as to decrease a certain functional as quickly as possible, relative to the Wasserstein metric.

Wasserstein gradient flows

- Certain evolution PDEs can be interpreted as gradient flows in the Wasserstein space: that is, a distribution of mass is rearranging itself so as to decrease a certain functional as quickly as possible, relative to the Wasserstein metric.
- Recall the gradient flow on $\mathbb{R}^n$ for a function $F : \mathbb{R}^n \to \mathbb{R}$ is

$$\frac{dx}{dt}(t) = -\nabla F(x(t))$$

Wasserstein gradient flows

- Certain evolution PDEs can be interpreted as gradient flows in the Wasserstein space: that is, a distribution of mass is rearranging itself so as to decrease a certain functional as quickly as possible, relative to the Wasserstein metric.
- Recall the gradient flow on $\mathbb{R}^n$ for a function $F : \mathbb{R}^n \to \mathbb{R}$ is

$$\frac{dx}{dt}(t) = -\nabla F(x(t))$$

- How can this idea be adapted to Wasserstein space?

Wasserstein gradient flows

- First, we have to understand gradients. The key point from $\mathbb{R}^n$ we want to translate is that for a curves x_t, $\frac{d}{dt}(F(x(t))) = \nabla F(x(t)) \cdot x'(t)$

Wasserstein gradient flows

- First, we have to understand gradients. The key point from $\mathbb{R}^n$ we want to translate is that for a curves x_t, $\frac{d}{dt}(F(x(t))) = \nabla F(x(t)) \cdot x'(t)$
- For a curve of measures, $d\mu_t(x) = f_t(x)dx$, the classical way (working in L^2) to interpet its velocity would be as $f_t'(x)$ (the rate of change of mass at the point x).

Wasserstein gradient flows

- First, we have to understand gradients. The key point from $\mathbb{R}^n$ we want to translate is that for a curves x_t, $\frac{d}{dt}(F(x(t))) = \nabla F(x(t)) \cdot x'(t)$
- For a curve of measures, $d\mu_t(x) = f_t(x)dx$, the classical way (working in L^2) to interpet its velocity would be as $f_t'(x)$ (the rate of change of mass at the point x).
- Optimal transport is about how mass *moves*; we therefore interpret the velocity of a curve by its velocity vector at individual points, given by $f_t'(x) = -\nabla \cdot (v_t(x) f_t(x))$

Wasserstein gradient flows

- First, we have to understand gradients. The key point from $\mathbb{R}^n$ we want to translate is that for a curves x_t, $\frac{d}{dt}(F(x(t))) = \nabla F(x(t)) \cdot x'(t)$
- For a curve of measures, $d\mu_t(x) = f_t(x)dx$, the classical way (working in L^2) to interpet its velocity would be as $f_t'(x)$ (the rate of change of mass at the point x).
- Optimal transport is about how mass *moves*; we therefore interpret the velocity of a curve by its velocity vector at individual points, given by $f_t'(x) = -\nabla \cdot (v_t(x)f_t(x))$
- Therefore, for a functional $\mathcal{F}$ on the space of probability measures, we would like ot define $\nabla\mathcal{F}$ so that for each curve $d\mu_t(x) = f_t(x)dx$, we have

$$\frac{d}{dt}(\mathcal{F}(\mu_t)) =< w_t, v_t >_{L^2(\mu_t)}= \int_{\mathbb{R}^n} w_t(x) \cdot v_t(x) d\mu_t(x)$$

where $f_t' = -\nabla \cdot (v_t f_t)$ and $(\nabla\mathcal{F})(\mu_t) = -\nabla \cdot (w_t f_t)$

Wasserstein gradient flows

- For simplicity, consider functionals $\mathcal{F}$ on Wasserstein space of the form
$$\mathcal{F}(\mu) = \int_{\mathbb{R}^n} U(f(x))dx$$
where $d\mu(x) = f(x)dx$.

Wasserstein gradient flows

- For simplicity, consider functionals $\mathcal{F}$ on Wasserstein space of the form
$$\mathcal{F}(\mu) = \int_{\mathbb{R}^n} U(f(x))dx$$
where $d\mu(x) = f(x)dx$.
- For a curve $d\mu_t = f_t dx$, we have
$$\frac{d}{dt}\mathcal{F}(\mu_t) = \int_{\mathbb{R}^n} U'(f_t(x))f_t'(x)dx$$

Wasserstein gradient flows

- For simplicity, consider functionals $\mathcal{F}$ on Wasserstein space of the form
$$\mathcal{F}(\mu) = \int_{\mathbb{R}^n} U(f(x))dx$$
where $d\mu(x) = f(x)dx$.
- For a curve $d\mu_t = f_t dx$, we have
$$\frac{d}{dt}\mathcal{F}(\mu_t) = \int_{\mathbb{R}^n} U'(f_t(x))f_t'(x)dx$$
- We would like to write this as an integral against the corresponding velocity vector v_t which satisfies:
$$f_t'(x) = -\nabla \cdot (v_t(x)f_t(x))$$

Wasserstein gradient flows

$$\begin{aligned}
\frac{d}{dt}\mathcal{F}(\mu_t) &= \int_{\mathbb{R}^n} U'(f_t(x))f_t'(x)dx \\
&= -\int_{\mathbb{R}^n} U'(f_t(x))\nabla\cdot(v_t(x)f_t(x))dx \\
&= \int_{\mathbb{R}^n} \nabla(U'(f_t(x)))\cdot v_t(x)f_t(x)dx
\end{aligned}$$

Wasserstein gradient flows

$$\begin{aligned}\frac{d}{dt}\mathcal{F}(\mu_t) &= \int_{\mathbb{R}^n} U'(f_t(x))f_t'(x)dx \\ &= -\int_{\mathbb{R}^n} U'(f_t(x))\nabla\cdot(v_t(x)f_t(x))dx \\ &= \int_{\mathbb{R}^n} \nabla(U'(f_t(x)))\cdot v_t(x)f_t(x)dx\end{aligned}$$

This establishes $\nabla(U'(f_t(x)))$ as the **velocity** corresponding to the gradient. The corresponding rate of change of the density then comes from the continuity equation:

$$(\nabla\mathcal{F})(\mu_t) = -\nabla\cdot(f_t(x)\nabla(U'(f_t(x))).$$

$$\begin{aligned}\frac{d}{dt}\mathcal{F}(\mu_t) &= \int_{\mathbb{R}^n} U'(f_t(x))f_t'(x)dx \\ &= -\int_{\mathbb{R}^n} U'(f_t(x))\nabla\cdot(v_t(x)f_t(x))dx \\ &= \int_{\mathbb{R}^n} \nabla(U'(f_t(x)))\cdot v_t(x)f_t(x)dx\end{aligned}$$

This establishes $\nabla(U'(f_t(x)))$ as the **velocity** corresponding to the gradient. The corresponding rate of change of the density then comes from the continuity equation:

$$(\nabla\mathcal{F})(\mu_t) = -\nabla\cdot(f_t(x)\nabla(U'(f_t(x))).$$

So, the gradient of $\mathcal{F}$ evaluated at $d\mu(x) = f(x)dx$ is $-\nabla\cdot(f(x)\nabla(U'(f(x))).$

Wasserstein gradient flows: examples

Take the entropy: $\mathcal{F} = \int_{\mathbb{R}^n} f(x) \log(f(x))dx$ (so $U(r) = r\log(r)$).
Then its Wasserstein gradient is

$$\begin{aligned} -\nabla \cdot (f(x)\nabla(U'(f(x))) &= -\nabla \cdot (f(x)\nabla(\log(f(x)+1)) \\ &= -\nabla \cdot (f(x)(\frac{\nabla f(x)}{f(x)})) = -\Delta f(x) \end{aligned}$$

Wasserstein gradient flows: examples

Take the entropy: $\mathcal{F} = \int_{\mathbb{R}^n} f(x)\log(f(x))dx$ (so $U(r) = r\log(r)$). Then its Wasserstein gradient is

$$\begin{aligned} -\nabla\cdot(f(x)\nabla(U'(f(x))) &= -\nabla\cdot(f(x)\nabla(\log(f(x)+1)) \\ &= -\nabla\cdot(f(x)(\frac{\nabla f(x)}{f(x)})) = -\Delta f(x) \end{aligned}$$

So the Wasserstein gradient flow of the entropy is given by:

$$f'(x) = \Delta f(x)$$

The **heat equation!**

Wasserstein gradient flows: examples

Take the entropy: $\mathcal{F} = \int_{\mathbb{R}^n} f(x)\log(f(x))dx$ (so $U(r) = r\log(r)$). Then its Wasserstein gradient is

$$\begin{aligned} -\nabla\cdot(f(x)\nabla(U'(f(x))) &= -\nabla\cdot(f(x)\nabla(\log(f(x)+1)) \\ &= -\nabla\cdot(f(x)(\frac{\nabla f(x)}{f(x)})) = -\Delta f(x) \end{aligned}$$

So the Wasserstein gradient flow of the entropy is given by:

$$f'(x) = \Delta f(x)$$

The **heat equation!** (Note: it is actually easy to show that the entropy decreases for solutions of the heat equation – this show that it decreases as quickly as possible relative to the Wasserstein metric. On the other hand, it is well known that the heat equation is the gradient flow of the Dirichlet energy, $\int_X |\nabla f(x)|^2$ under the L^2 metric.)

Wasserstein gradient flows: examples

Take the entropy: $\mathcal{F} = \int_{\mathbb{R}^n} f(x)\log(f(x))dx$ (so $U(r) = r\log(r)$). Then its Wasserstein gradient is

$$\begin{aligned} -\nabla \cdot (f(x)\nabla(U'(f(x))) &= -\nabla \cdot (f(x)\nabla(\log(f(x)+1)) \\ &= -\nabla \cdot (f(x)(\frac{\nabla f(x)}{f(x)})) = -\Delta f(x) \end{aligned}$$

So the Wasserstein gradient flow of the entropy is given by:

$$f'(x) = \Delta f(x)$$

The **heat equation!** (Note: it is actually easy to show that the entropy decreases for solutions of the heat equation – this show that it decreases as quickly as possible relative to the Wasserstein metric. On the other hand, it is well known that the heat equation is the gradient flow of the Dirichlet energy, $\int_X |\nabla f(x)|^2$ under the L^2 metric.)

Many other examples (see, for example, Villani TOT p.252).

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, \ldots$, with $h = t_{i+1} - t_i$.

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, \ldots$, with $h = t_{i+1} - t_i$.
- Then $x(t_{i+1}) - x(t_i) \approx hx'(t_{i+1}) = -h\nabla F(x_{t_{i+1}})$.

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, ...,$ with $h = t_{i+1} - t_i$.
- Then $x(t_{i+1}) - x(t_i) \approx hx'(t_{i+1}) = -h\nabla F(x_{t_{i+1}})$.
- That is $x(t_{i+1}) \approx x(t_i) - h\nabla F(x_{t_{i+1}})$.

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, ...,$ with $h = t_{i+1} - t_i$.
- Then $x(t_{i+1}) - x(t_i) \approx hx'(t_{i+1}) = -h\nabla F(x_{t_{i+1}})$.
- That is $x(t_{i+1}) \approx x(t_i) - h\nabla F(x_{t_{i+1}})$.
- So $x = x(t_{i+1})$ makes the gradient of $x \mapsto \frac{1}{2}|x - x(t_i)|^2 + hF(x)$ vanish.

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, ...,$ with $h = t_{i+1} - t_i$.
- Then $x(t_{i+1}) - x(t_i) \approx hx'(t_{i+1}) = -h\nabla F(x_{t_{i+1}})$.
- That is $x(t_{i+1}) \approx x(t_i) - h\nabla F(x_{t_{i+1}})$.
- So $x = x(t_{i+1})$ makes the gradient of $x \mapsto \frac{1}{2}|x - x(t_i)|^2 + hF(x)$ vanish.
- Under suitable conditions on F, for small h, we expect choosing $x(t_{i+1})$ to minimize $\frac{1}{2}|x - x(t_i)|^2 + hF(x)$ to be a good approximation of the solution.

Jordan-Kinderlehrer-Otto scheme

- For $x'(t) = -\nabla F(x(t))$ in $\mathbb{R}^n$, think of discretizing time $t_0, t_1, ...,$ with $h = t_{i+1} - t_i$.
- Then $x(t_{i+1}) - x(t_i) \approx hx'(t_{i+1}) = -h\nabla F(x_{t_{i+1}})$.
- That is $x(t_{i+1}) \approx x(t_i) - h\nabla F(x_{t_{i+1}})$.
- So $x = x(t_{i+1})$ makes the gradient of $x \mapsto \frac{1}{2}|x - x(t_i)|^2 + hF(x)$ vanish.
- Under suitable conditions on F, for small h, we expect choosing $x(t_{i+1})$ to minimize $\frac{1}{2}|x - x(t_i)|^2 + hF(x)$ to be a good approximation of the solution.
- Similarly, on Wasserstein space, we expect choosing $\mu_{t_{i+1}}$ to minimize

$$\mu \mapsto \mathcal{F}(\mu) + \frac{1}{2h} W_2^2(\mu_{t_i}, \mu)$$

  to be close to Wasserstein gradient flow (or the corresponding PDE).