Math 260J: Optimal Transport

Prof. Katy Craig

Discord server

Website: http://web.math.ucsb.edu/~kcraig/math/260J_W22.html

Course goals

- What is the optimal transport problem? How does duality help us solve this problem? What type of geometry does it induce?
- What is a Wasserstein gradient flow? What is the relationship between convexity of an energy and well-posedness of the PDE characterizing the flow?
- Interplay between convex analysis, PDE, probability, functional analysis, geometry, and optimization.
- Expository writing to general scientific audience → OT Wiki (essential skill for job applications, grants, interdisciplinary papers, ...)

Optimal Transport

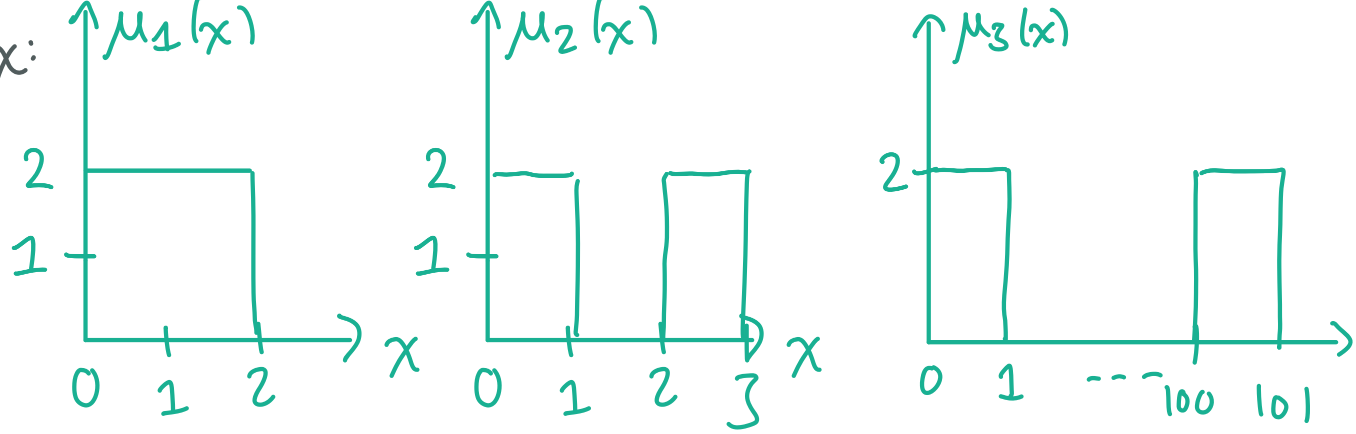
Gaspard Monge, 1781
"On cuttings and embankments"

Q: How can we rearrange the dirt in μ to look like ν in the most efficient way?

Q': Why do we care?

A': The amount of effort it takes to rearrange one pile of dirt to look like another provides a notion of distance that is useful in PDE, geometry, statistics, machine learning, ...

Ex:

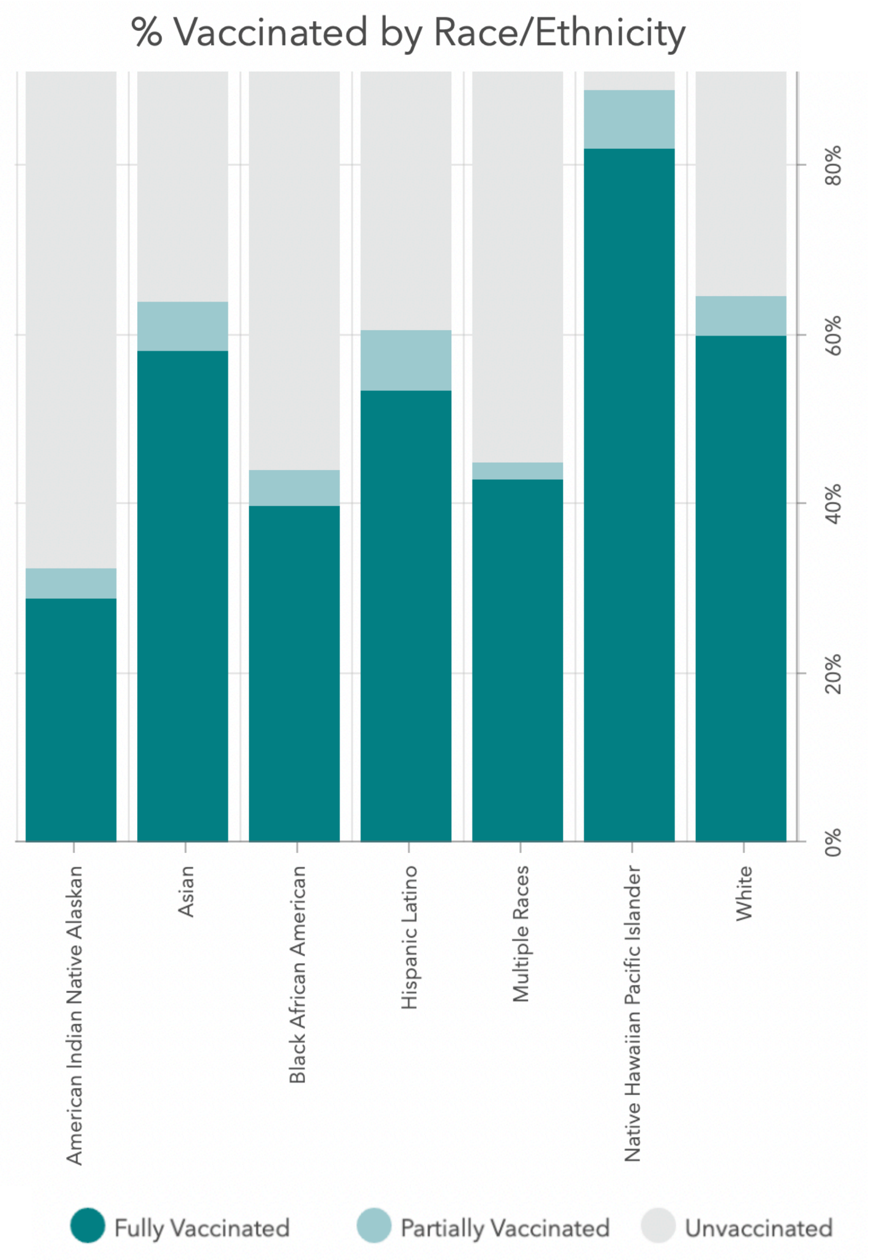


If we measure distance in the "usual way" (L^p norms, statistical divergences, ...)

$$\int |\mu_1(x) - \mu_2(x)|\, dx = \int |\mu_1(x) - \mu_3(x)|\, dx = 4$$

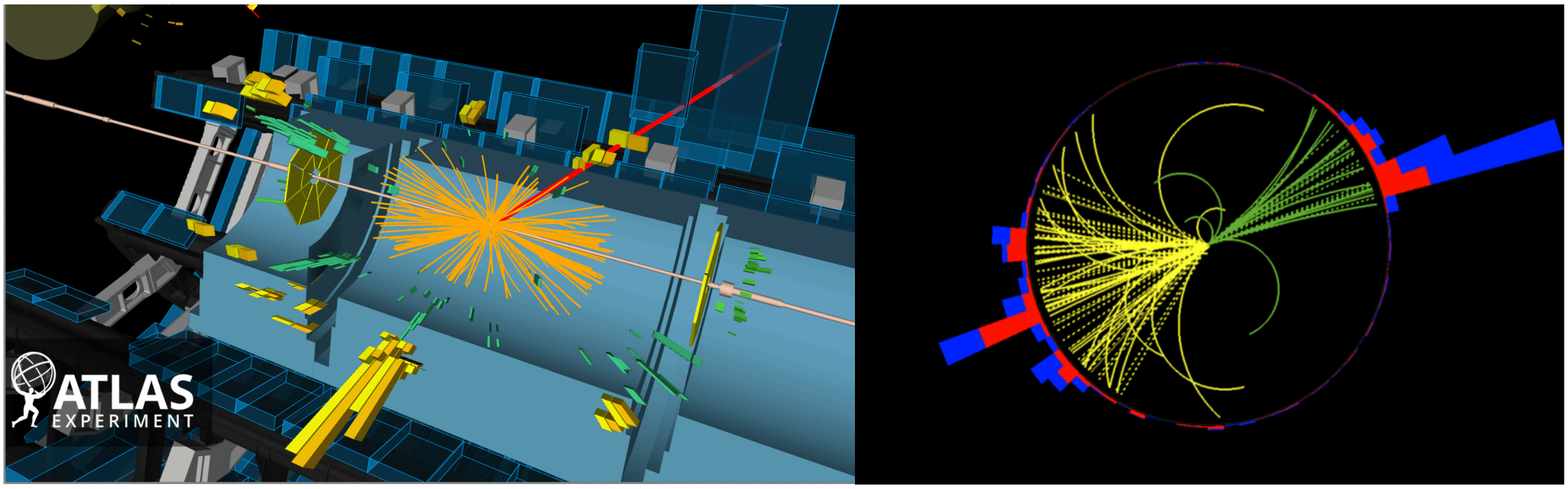
Moral: common notions of distance between functions do not allow independent variable with a spatial interpretation.

For certain data sets, this makes sense:



source: publichealthsbc.org

For other data sets, disregarding the spatial interpretation of independent variable throws away important information:



Optimal transport provides a notion of distance between
[functions, data distributions, measures]
that preserves the spatial interpretation of the independent variable.

Over the past 20 years, this has had an enormous impact in:

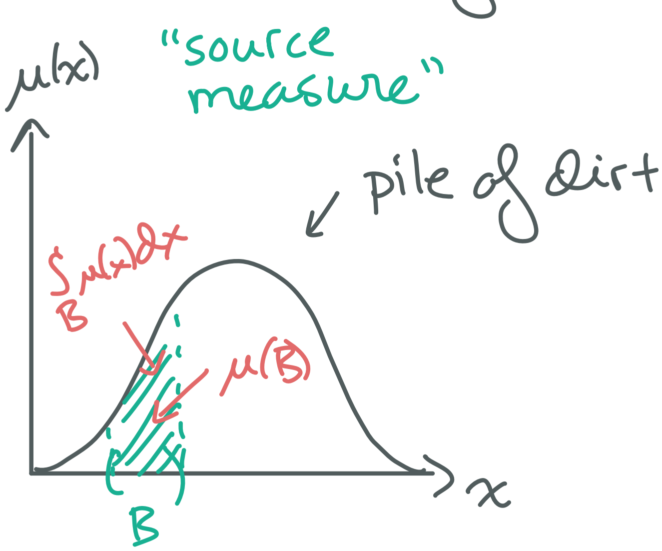
① PDE: two Fields medals
Villani (2010), Figalli (2018)

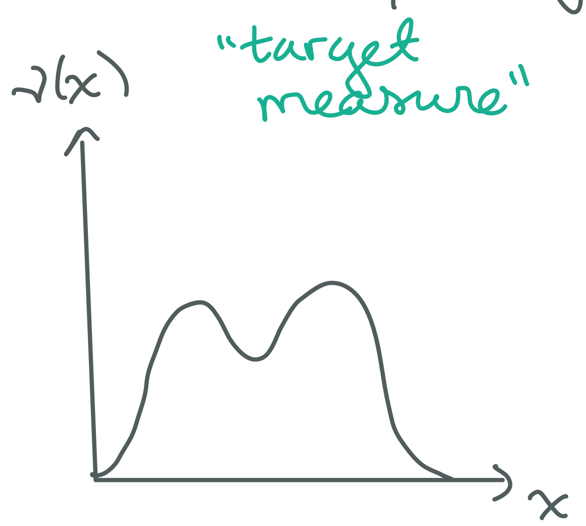
② Geometry: novel characterization of Ricci curvature in terms of convexity of entropy

③ Statistics: sampling

④ Machine learning: 2-layer neural networks, normalizing flows, ...

Back to original optimal transport problem...





Q: How can we rearrange the dirt in μ to look like ν in the most efficient way?

We will represent the piles of dirt by measures.

$(\mathcal{X}, d)$ metric space, e.g. $(\mathbb{R}^d, |\cdot|)$
$(\mathcal{Y}, d)$

where the dirt lives

$\mathcal{B}(\mathcal{X})$ Borel σ-algebra

smallest σ-algebra containing all open sets

closed under countable unions + complements

$\mathcal{M}(\mathcal{X})$ finite (Borel) measures on $\mathcal{X}$

functions $\mu : \mathcal{B}(\mathcal{X}) \to [0, +\infty)$ s.t. $\mu(\emptyset) = 0$, $\mu(\underbrace{\bigcup_{i=1}^{\infty} B_i}_{\text{disjoint}}) = \sum_{i=1}^{\infty} \mu(B_i)$.

Given $\mu \in \mathcal{M}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{X})$
$\mu(B)$ = amt of dirt in the pile μ that lies in B.

Important Notational Abuse:

How does this relate to the pictures we were drawing earlier?

if $\lambda(B) = 0 \Rightarrow \mu(B) = 0$

- Let $(\mathcal{X}, d) = (\mathbb{R}^d, |\cdot|)$

"μ is absolutely continuous w.r.t. Lebesgue measure"

- Recall: if $\mu \ll \lambda$, $\exists\, f \in L^1(\lambda)$ s.t. $d\mu = f\, d\lambda$

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

"f is the Radon-Nikodym derivative"
$f = \frac{d\mu}{d\lambda}$

- To avoid doubling the number of symbols, we commit the following notational abuse:

  instead of $f(x)$, write $\mu(x)$, so whenever we have $\mu \ll \lambda$, $d\mu(x) = \mu(x)dx$

- A functional analysis perspective: consider $C_b(X) = \{\varphi : X \to \mathbb{R} : \varphi$ is bounded and continuous$\}$.

  Any $\mu \in \mathcal{M}(X)$ induces a bounded linear functional $C_b(X)$ via

$$\langle \mu, \varphi \rangle = \int_X \varphi \, d\mu$$

- linear $\langle \mu, \alpha\varphi + \beta\psi \rangle = \int_X \alpha\varphi + \beta\psi \, d\mu = \alpha\int_X \varphi \, d\mu + \beta\int_X \psi \, d\mu$

$$= \alpha\langle \mu, \varphi \rangle + \beta\langle \mu, \psi \rangle$$

- bounded $|\langle \mu, \varphi \rangle| \le \|\varphi\|_\infty \int_X d\mu$

Thus $\mathcal{M}(X) \subseteq (C_b(X))^*$.

Similarly, any $f \in L^1(\mathbb{R}^d)$ induces a bdd linear functional $C_b(\mathbb{R}^d)$ via $\langle f, \varphi \rangle = \int_{\mathbb{R}^d} f(x)\varphi(x)dx$

Thus $L^1(\mathbb{R}^d) \subseteq (C_b(\mathbb{R}^d))^*$

The notational abuse for $\mu \ll \lambda$, $d\mu(x) = \mu(x)dx$, is using the same symbol for identical elements in $(C_b(\mathbb{R}^d))^*$.

Given $\mu \in \mathcal{M}(\mathbb{R}^d)$, $B \in \mathcal{B}(\mathbb{R}^d)$, $\mu \ll \lambda$, $\mu(B) = \int_B \mu(x)\,dx =$ amt. of dirt in B.

Finally, to have any hope of rearranging μ to look like ν, we must have $\mu(X) = \nu(X)$.

total amt of dirt in μ

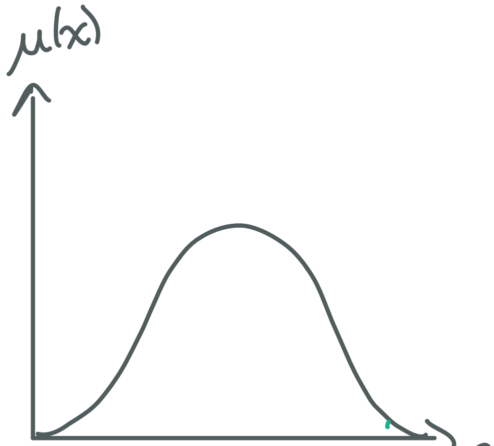
WLOG, suppose $\mu(X) = \nu(X) = 1$, that is μ and ν are probability measures.

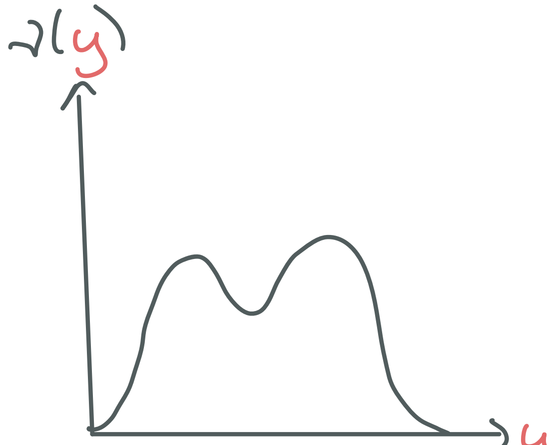
Thus, we will represent piles of dirt as probability measures. What does it mean to "rearrange one to look like another"?

Def (transport map): Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a measurable function $t: X \to Y$, we say that t transports μ to ν if

$$\nu(B) = \mu(t^{-1}(B)), \quad \forall B \in \mathcal{B}(Y).$$

We call ν the pushforward of μ under t, writing $\nu = t\#\mu$, and we will call t a transport map from μ to ν.





x