Math 260R: Optimal Transport
Prof. Katy Craig
(No office hours on Friday)

Recall:

optimal transport problem

(X, d_X), (Y, d_Y) metric spaces

$\mathcal{B}(X)$ Borel σ-algebra

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

$\mathcal{P}(X)$ (Borel) probability measures on X

source measure $\mu \in \mathcal{P}(X)$

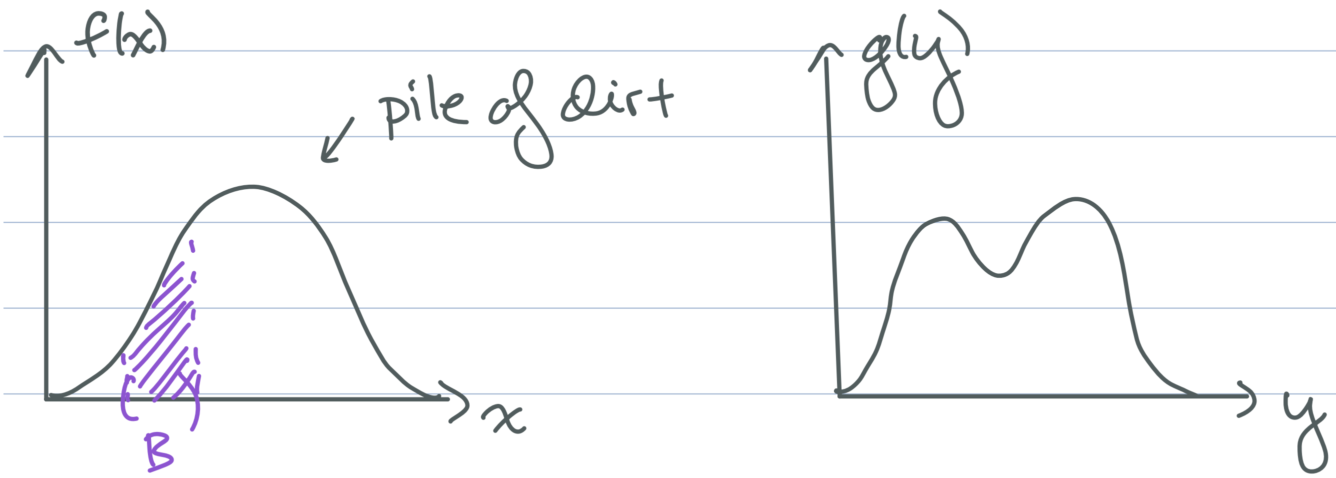
target measure $\nu \in \mathcal{P}(Y)$

For $B \in \mathcal{B}(X)$, "amount of dirt in region B" $= \mu(B)$

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

If measures have densities wrt Lebesgue, can draw pictures...

$$d\mu(x) = f(x)\, d\lambda(x) \qquad d\nu(y) = g(y)\, d\lambda(y)$$



What does it mean to "rearrange" one probability measure 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 t <u>transports μ to ν</u> if

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

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

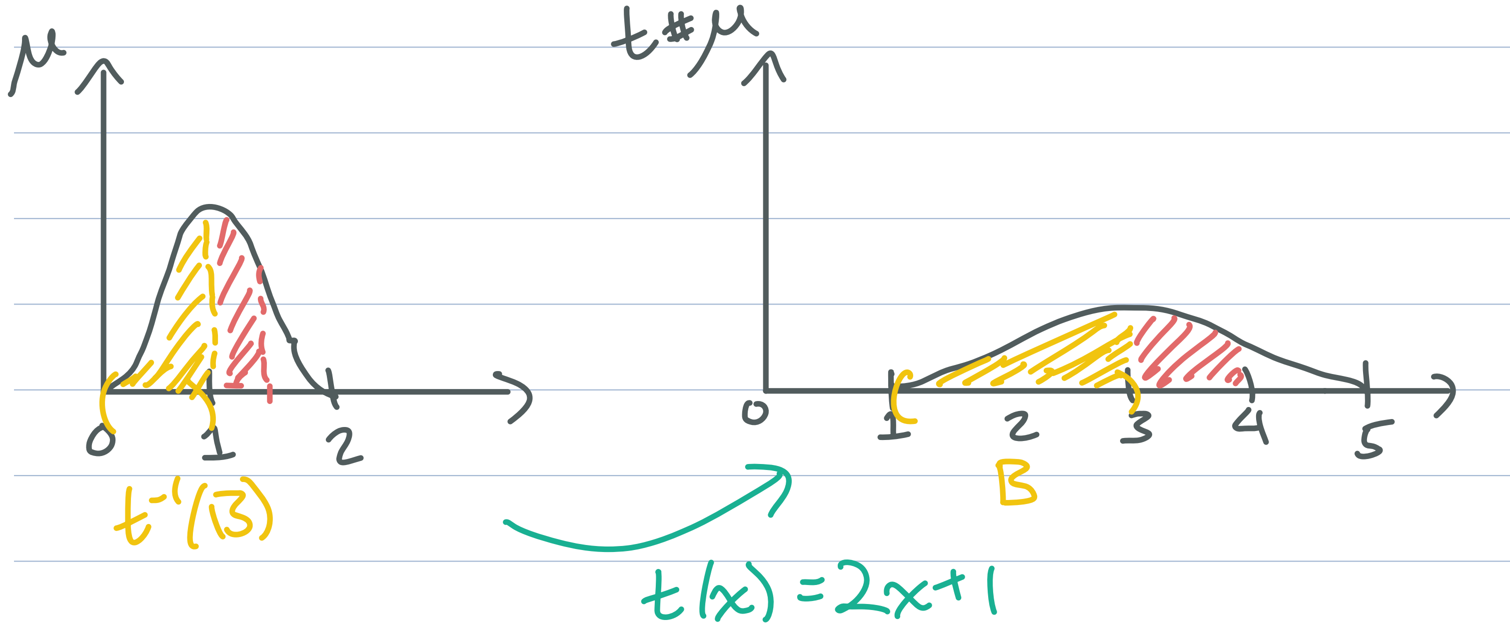
Informally, "mass starting at location x_0 in μ is sent to $t(x_0)$ in ν." See Exercise 1.

Ex (translation/dilation)

Suppose $(X, d_X) = (Y, d_Y) = (\mathbb{R}^d, |\cdot|)$
Fix $a > 0$, $b \in \mathbb{R}^d$, and let $t(x) = ax + b$.

Thus, for any $\mu \in P(\mathbb{R}^d)$, $t\#\mu$ satisfies

$$(t\#\mu)(B) = \mu(t^{-1}(B)) = \mu\left(\frac{B-b}{a}\right)$$



<u>Lemma</u> (equiv characterization of transp map): Given $\mu \in \mathcal{P}(X)$ and $t: X \to Y$ measurable,

$$t\# \mu = \nu \iff \int \varphi(t(x))\, d\mu(x) = \int \varphi(y)\, d\nu(y) \quad \forall \varphi \in L^1(\nu)$$

Pf: Exercise 2

<u>Lemma</u> (change of variables formula)

Suppose

- $f \in L^1(\lambda^d)$, $f \geq 0$, $\int f \, d\lambda^d = 1$
- μ is given by $d\mu(x) = f(x) \, d\lambda^d(x)$
- $t \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ is injective and satisfies $|\det(Dt)|(x) \neq 0 \ \forall x \in \mathbb{R}$.

Then $\nu := t\#\mu$ satisfies $d\nu(y) = g(y) \, d\lambda^d(y)$

where

$$g(y) = \begin{cases} \left(\dfrac{f}{|\det(Dt)|}\right) \circ t^{-1}(y) & \text{if } y \in t(\mathbb{R}^d) \\ 0 & \text{if } y \notin t(\mathbb{R}^d) \end{cases}$$

Pf: Exercise 3

<u>Cor</u>: Under the hypotheses of the previous lemma, if $a > 0$, $b \in \mathbb{R}^d$, and

$$t(x) = ax + b$$

then $g(y) = \frac{1}{a^d} f\left(\frac{y-b}{a}\right)$.

Application: Normalizing Flows

Reference: Kobyzev, Prince Brubaker '21

e.g. (i) μ is a uniform prob measure on some region
(ii) μ is a Gaussian

F... "normalizing"

Problem: Given a reference measure $\mu \in \mathcal{P}(X)$ about which we know everything, and given a target measure $\nu \in \mathcal{P}(Y)$, from which we have samples $\{y_i\}_{i=1}^{n}$ find $t: X \to Y$ "nice" so that $t\#\mu \approx \nu$.

satisfying hypotheses of change of variables lemma

"rearranging/flowing μ to ν"

What does it mean to "have samples"?

Suppose $\mathcal{X}, \mathcal{Y}$ are Polish spaces.
complete, separable metric spaces

Def: $C_b(\mathcal{X}) = \{\varphi : \mathcal{X} \to \mathbb{R} : \varphi \text{ is bdd, cts}\}$

Def: (narrow convergence) Given $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$, we say $\mu_n \to \mu$ <u>narrowly</u> if

$$\lim_{n\to\infty} \int_{\mathcal{X}} \varphi(x)\, d\mu_n(x) = \int_{\mathcal{X}} \varphi(x)\, d\mu(x), \quad \forall \varphi \in C_b(\mathcal{X})$$

Lemma: Narrow convergence is metrizable.

Pf: Exercise 4.

Unconventional Def: $\{y_i\}_{i=1}^{n} \subseteq \mathcal{Y}$, $n \in \mathbb{N}$, are <u>samples</u> of $\nu \in \mathcal{P}(\mathcal{Y})$ if

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i} \xrightarrow{n\to\infty} \nu \text{ narrowly.}$$

Rmk: The previous definition is equivalent to

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varphi(y_i) = \int_Y \varphi(y)\,d\nu(y), \ \forall \varphi \in C_b(Y).$$

Motivation for Problem:

① Draw new samples from ν (at least approximately)

If $\{x_j\}_{j=1}^m$ are samples of μ and $t: X \to Y$ is continuous, then $\{t(x_j)\}_{j=1}^m$ are samples of $t\#\mu$. (Exercise 5)

② (In the setting of the change of var lemma...) Find the value of the density of ν w.r.t. Lebesgue at arbitrary $y \in Y$.

Example: Fashion MNIST

70,000 28×28 grey scale fashion images.

Consider images of shoes as samples $\{y_i\}_{i=1}^n \subseteq \mathbb{R}^{28\times28}$ as samples of some unknown measure $\nu \in \mathcal{P}(\mathbb{R}^{28\times28})$

For any $B \in \mathcal{B}(\mathbb{R}^{28\times28})$,

$\nu(B)$ = proportion of shoe images in B.

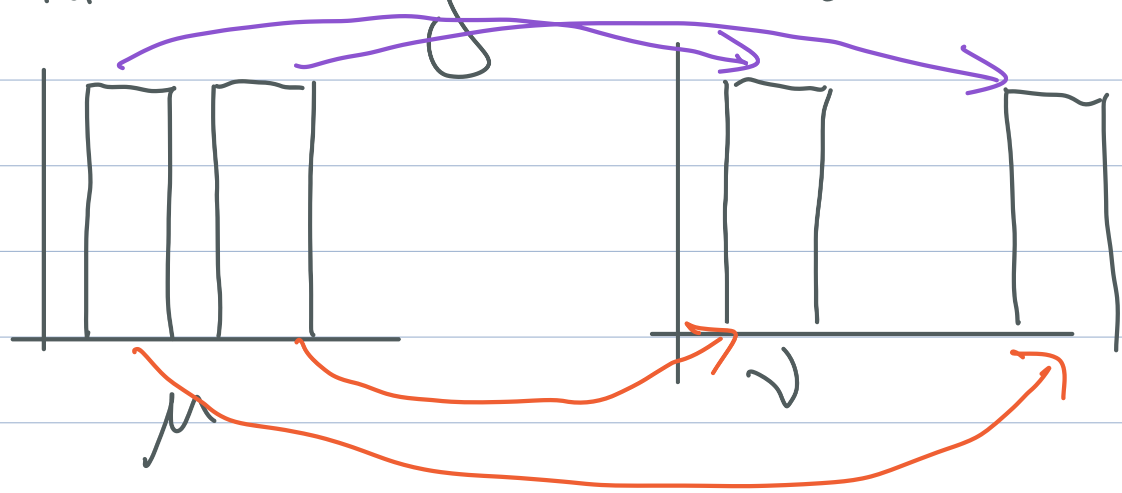
(1) Drawing new samples from ν
$\iff$ generating new candidate images of shoes

(2) Finding the value of the density of ν at some $y \in \mathbb{R}^{28 \times 28}$
$\iff$ finding relative confidence that y is a picture of a shoe

---

Challenges in Solving NF Problem:

- badly underspecified: there can exist many "nice" t s.t. $t_{\#}\mu = \nu$



- need to ensure $t\#\mu \approx \nu$ based only on knowledge of $\{y_i\}_{i=1}^n$.

Normalizing Flows Approach:

- Require t to belong to a parametric class of functions $\mathcal{T}$ that are convenient to compute, invert, and calculate Jacobian determinant

e.g. $\mathcal{T} = \{ t : \mathbb{R}^d \to \mathbb{R}^m : t(x) = Ax + b$ for $A \in M_{m \times d}(\mathbb{R}),$ $b \in \mathbb{R}^m \}$

(see Kobyzev, et. al.)

$d = m$.

maybe even... $\mathcal{T} = \{ \nabla \varphi : \varphi : \mathbb{R}^d \to \mathbb{R}$ convex $\}$

these are "optimal" transport maps (in some sense)

$D(\nu|\omega)$ = "how similar ν is to ω"

generalization of idea of metric

- Given a statistical divergence, that is $D : P(Y) \times P(Y) \to [0, +\infty)$ s.t. $D(\nu|\omega) = 0 \iff \omega = \nu$, want to solve...

$$\min_{t \in \tilde{\mathcal{T}}} D(\nu | t\#\mu)$$

... but, in practice, approximate

$$D(\nu | t\#\mu) \approx D_n\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}, t\#\mu\right)$$

and solve

$$(*)\left\{ \min_{t \in \tilde{\mathcal{T}}} D_n\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}, t\#\mu\right).\right.$$

Most important example:

If $d\nu(y) = g(y)\,d\lambda^m(y)$, $d\omega(y) = h(y)\,d\lambda^m(y)$

$$D(\nu|\omega) = KL(\nu|\omega) = \int_Y \log\left(\frac{g(y)}{h(y)}\right) d\nu(y).$$

$$= \underbrace{\int_Y \log(g(y))\,d\nu(y)}_{:= C_\nu} - \int_Y \log(h(y))\,d\nu(y).$$

$$D_n\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i} \Big| \omega\right) = C_\nu - \frac{1}{n}\sum_{i=1}^n \log(h(y_i))$$

Thus, if $\omega = t\#\mu$, solving (*) is equivalent to finding t so that $d(t\#\mu)(y) = h(y)\,d\lambda^m(y)$ makes

$$\frac{1}{n}\sum_{i=1}^n \log(h(y_i))$$ log-likelihood maximization

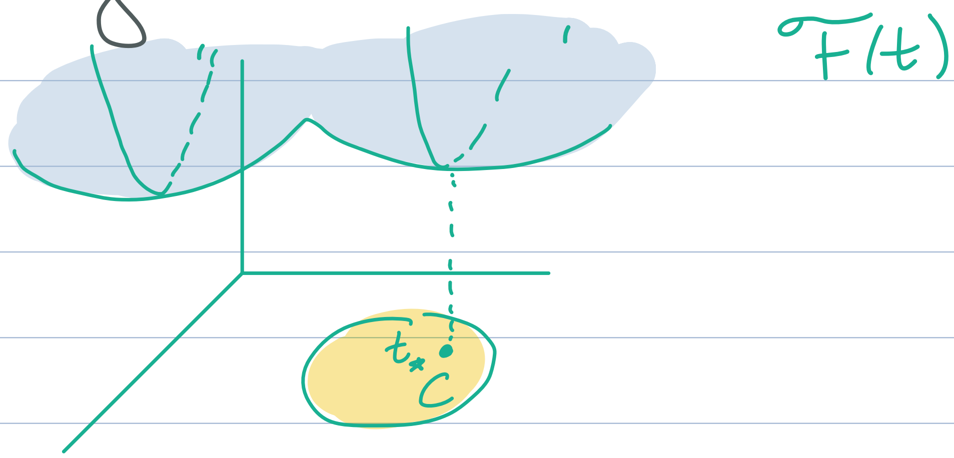
as large as possible.

Throughout the course, we'll see many optimization problems of this form:

$$\min_{t \in C} \overbrace{F(t)}^{\text{objective function}}$$

C ← constraint set

Mental image:



Monge's Optimal Transport Problem

Given $\mu, \nu \in \mathcal{P}(X)$, solve

$$\min_{\substack{t : X \to X \text{ measurable} \\ t\#\mu = \nu}} \overbrace{\int d(x, t(x))\, d\mu(x)}^{\text{effort}}$$

Unfortunately, Monge's problem is a horrible optimization problem!

Sudakov 1979, Ambrosio and Pratelli 2001
Evans and Gangbo 1999

Reasons the Monge Problem is difficult:

Difficulty #1: the constraint set can be empty.

That is, given $\mu, \nu \in \mathcal{P}(X)$, there doesn't necessarily exist t s.t. $t\#\mu = \nu$.

For example, by Exercise 1, we see that if μ is countably supported and $t\#\mu = \nu$, then ν must be countably supported.

Heuristically, the problem is that a transport map t sends all mass starting at a location x_0 to $t(x_0)$. In particular, mass cannot split.

Two potential solutions to empty constraint set:

(a) don't allow source measure to concentrate mass on "small sets" (like points)

(b) instead of considering transport maps, consider transport plans.

... Next Time!

Difficulty #2: Solutions may not be unique.

Exercise 6: $t_0 \# \mu = \nu$ and $t_1 \# \mu = \nu$, so both t_0 and t_1 belong to the constraint set, and both transport maps require the same amount of "effort"!

Fact (will show later): t_0 and t_1 are <u>both</u> optimal transport maps.

Potential solution to nonuniqueness of optima

<u>Difficulty #3</u>: The constraint set is nonconvex

Generally, in optimization, we want our constraint set C to be convex, since our normal strategy is to take an initial guess, perturb it, and see if the objective function decreases.