Lecture 3

Q: format of exercises/solutions?

Recall:

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

"normalizing"

Problem: Given a reference measure $\mu \in P(X)$ about which we know everything, and given a target measure $\nu \in P(Y)$, 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"

Suppose X, Y are Polish spaces.
complete, separable metric spaces

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

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

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

Lemma: Narrow convergence is metrizable.

Pf: Exercise 4.

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

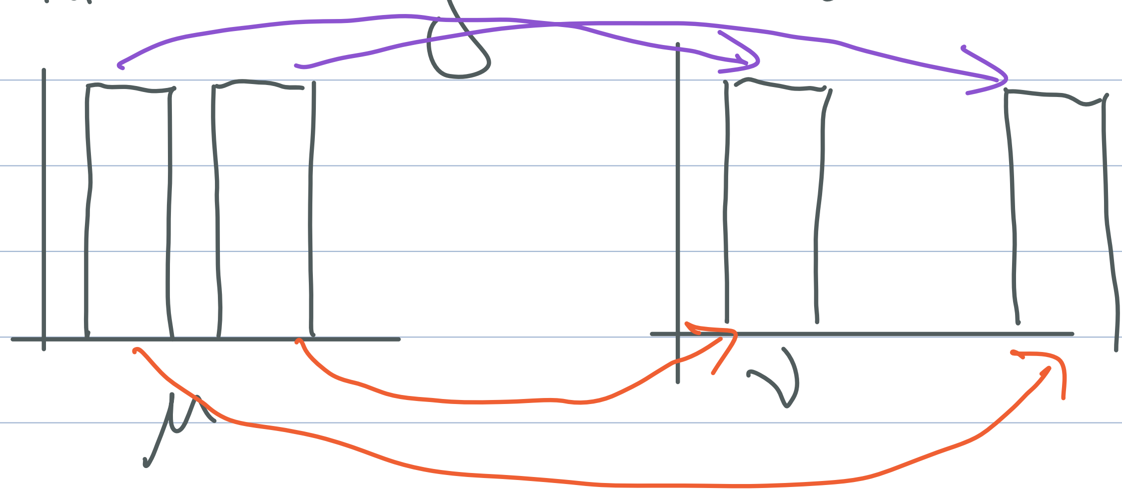
① Drawing new samples from ν
$\iff$ generating new candidate images of shoes

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

(see Kobyzev, et. al.)

$d = m$.

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

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 $D: \mathcal{P}(Y) \times \mathcal{P}(Y) \to [0,+\infty)$ 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.

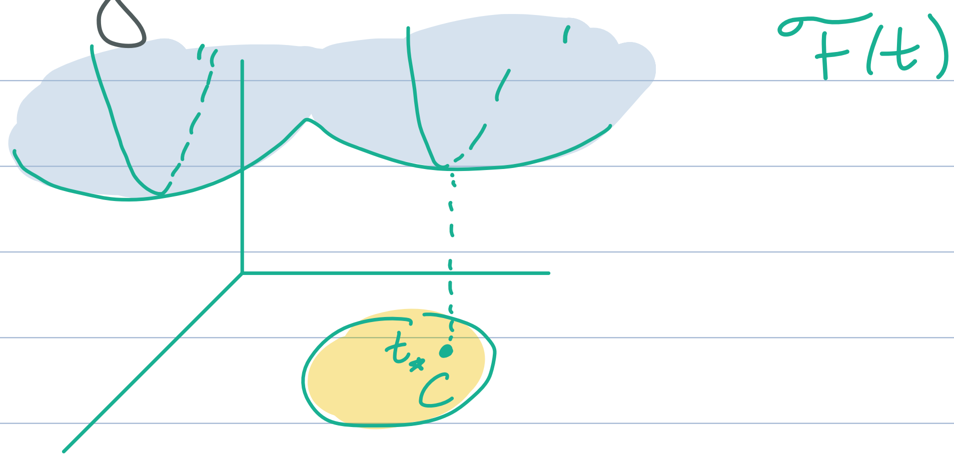
Throughout the course, we'll see many optimization problems of this form:

objective function

$$\min_{t \in C} \tilde{F}(t)$$

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.

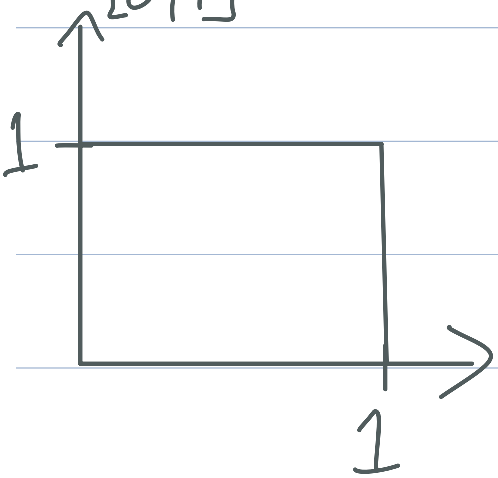
... next file

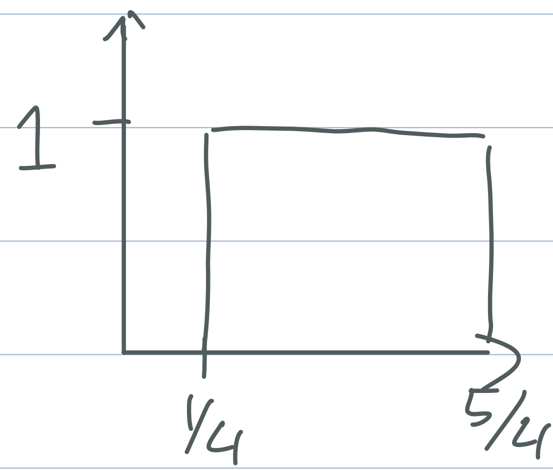
Difficulty #2: Solutions may not be unique.

That is, given $\mu, \nu \in \mathcal{P}(X)$, there may exist multiple, distinct optimal transport maps.

Ex: "books on a shelf"

$\mathbb{1}_{[0,1]}$





$d\mu(x) = \mathbb{1}_{[0,1]}(x)\, d\lambda(x)$ $\qquad\qquad$ ν

Consider $t_0(x) = x + \frac{1}{4}$ "shift all right"

$$t_1(x) = \begin{cases} x+1 & \text{if } x \in [0, 1/4) \\ x & \text{otherwise} \end{cases}$$ "first book to end"

Exercise 6: Show t_0, t_1 are both transp maps from μ to ν that require same amt of effort.

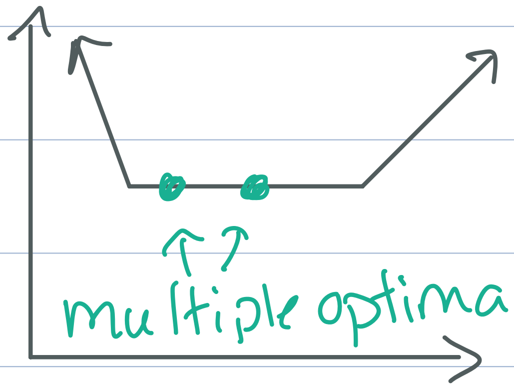
Fact (will prove late): In fact, both are optimal transport maps.

Potential solution to nonuniqueness of optima: modify notion of effort to make obj fn "strictly" convex...

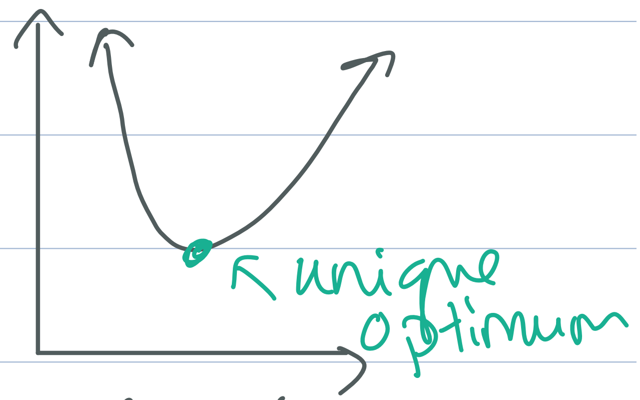
Monge's original problem on $\mathbb{R}$

p-Wasserstein, $p>1$

$$\min_{t: t\#\mu=\nu} \int |t(x)-x|\,d\mu(x) \Rightarrow \min_{t: t\#\mu=\nu} \left(\int |t(x)-x|^p\,d\mu(x)\right)^{1/p}$$



convex



strictly convex

Recall basic convexity facts:

vector space X

$C \subseteq X$ is convex if $\forall\, x_0, x_1 \in C$,

$$x_\alpha := (1-\alpha)x_0 + \alpha x_1 \in C, \quad \forall \alpha \in [0,1]$$

$f: C \to \mathbb{R} \cup \{+\infty\}$ is...

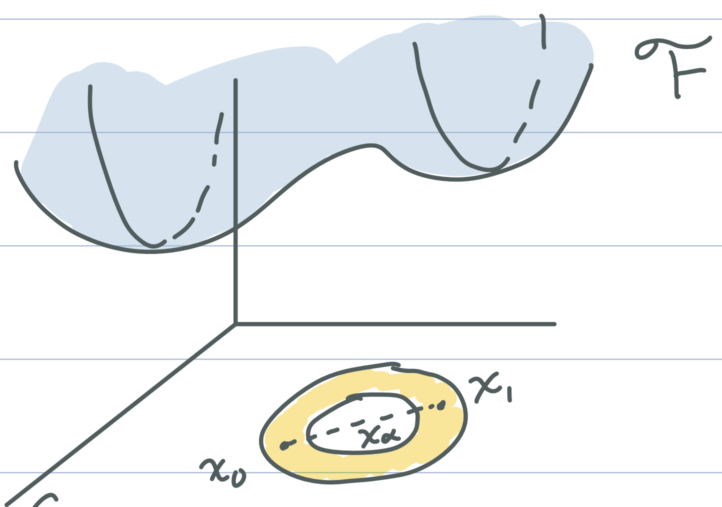
- <u>convex</u> if $f(x_\alpha) \leq (1-\alpha) f(x_0) + \alpha f(x_1)$
- <u>concave</u> $\geq$
- <u>strictly convex</u> $<$

... for all $x_0, x_1 \in C$, $\alpha \in (0,1)$.

If f is <u>convex</u> and <u>concave</u>, it is <u>affine linear</u>.

Difficulty #3: 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.



Exercise 6: For Monge's problem, linear perturbations of $\{t : t\#\mu = \nu\}$ kick us out of the constraint set.

Solution: consider transport plans.

How can we get around the difficulties of Monge's problem?

Relax the problem.

Leonid Kantorovich, 1942

"On the translocation of masses"

Notation:

Projection maps:

$\pi_X : X \times Y \to X$, $\pi_X(x,y) = x$

$\pi_Y : X \times Y \to Y$, $\pi_Y(x,y) = y$

$A \in \mathcal{B}(X)$

Marginals: For $\gamma \in \mathcal{P}(X \times Y)$, define **first marginal** $\pi_X \# \gamma(A) = \gamma(\pi_X^{-1}(A)) = \gamma(A \times Y)$

second marginal $\pi_Y \# \gamma$

Def (transport plan): Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, the set of **transport plans** from μ to ν is

$$\Gamma(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times Y) : \pi_X \# \gamma = \mu, \pi_Y \# \gamma = \nu\}$$

We will use transport plans as a new way to model rearranging mass in μ to look like ν. For $A \in \mathcal{B}(X)$, $B \in \mathcal{B}(Y)$,

$\gamma(A \times B) =$ amt of mass from $\mu(A)$ that is sent to $\nu(B)$.

How do transport plans relate to transport maps?

Notation: $id : X \to X$, $id(x) = x$

Lemma: Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, if $t\#\mu = \nu$, then

$$\gamma := (id \times t)\#\mu \in \Gamma(\mu, \nu).$$

$id \times t : X \to X \times Y$, $(id \times t)(x) = (x, t(x))$

Pf: Exercise 7

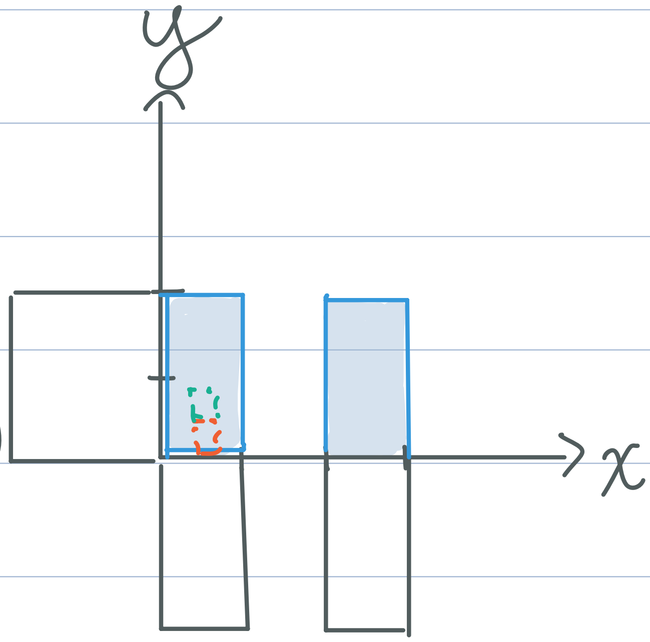
Visualizing transport plans

Ex: For μ, ν as below, consider the transport plan "where all mass starting at location x_0 in μ is distributed evenly in ν."

$$d\gamma(x,y) = \frac{1}{2}\left(1_{[0,1]\times[0,1]} + 1_{[2,3]\times[0,1]}\right)(x,y)\,d\lambda^2(x,y)$$

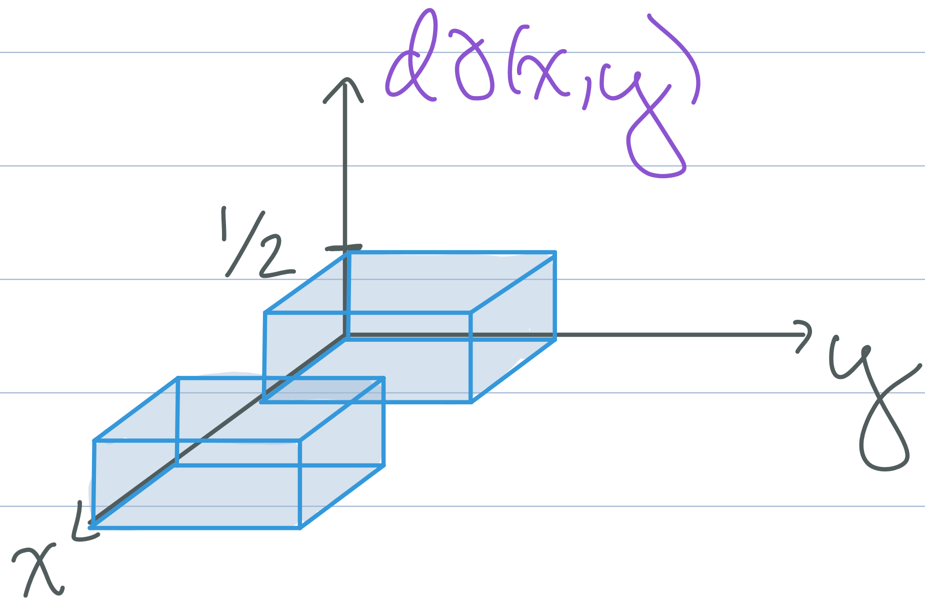
Bird's eye view:

$$d\nu(y) = 1_{[0,1]}(y)\,d\lambda(y)$$



$$d\mu(x) = \frac{1}{2}\left(1_{[0,1]} + 1_{[2,3]}\right)(x)\,d\lambda(x)$$

Side view:



This is a special case of the fact that...

For any $\mu, \nu \in \mathbb{P}(X)$, the transport plan $\gamma = \mu \otimes \nu \in \Gamma(\mu, \nu)$

$$\mu \otimes \nu (A \times B) = \mu(A)\nu(B)$$

"takes mass from any location x_0 in μ and distributes it across ν, in proportion to the amount of mass ν assigns to each location."

Moral: (i) For any μ, ν, $\Gamma(\mu, \nu) \neq \emptyset$.

(ii) transport plans can "split mass"

Ex: For $\mu = \frac{1}{2}(\mathbb{1}_{[0,1]} + \mathbb{1}_{[2,3]})$, $\nu = \frac{1}{2}(\mathbb{1}_{[0,2]})$,

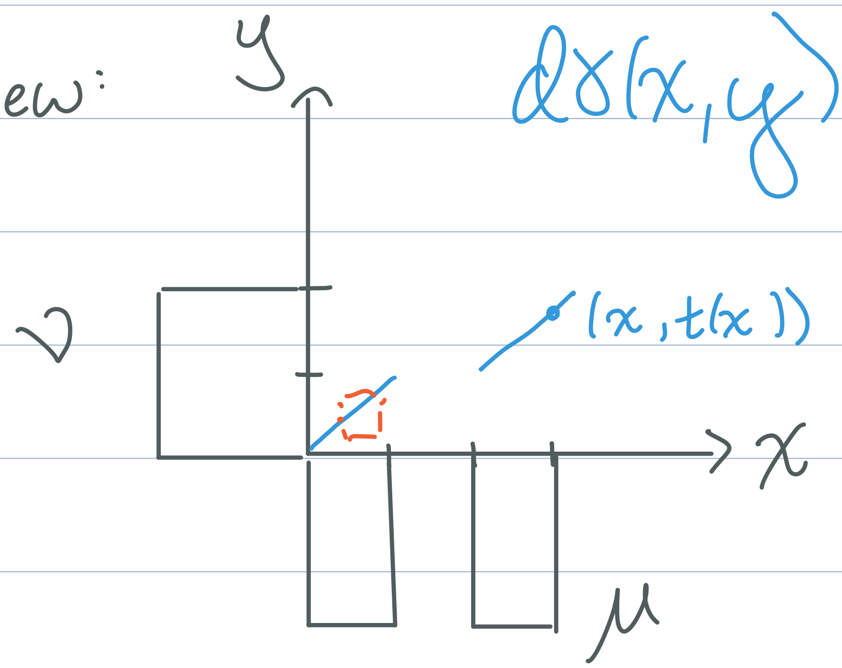
consider the transport map

$$t(x) = \begin{cases} x & \text{if } x \in [0,1] \\ x-1 & \text{otherwise} \end{cases}$$

Then $t \# \mu = \nu$, so by lemma,

$$\gamma := (id \times t) \# \mu \in \Gamma(\mu, \nu).$$

Bird's eye view:

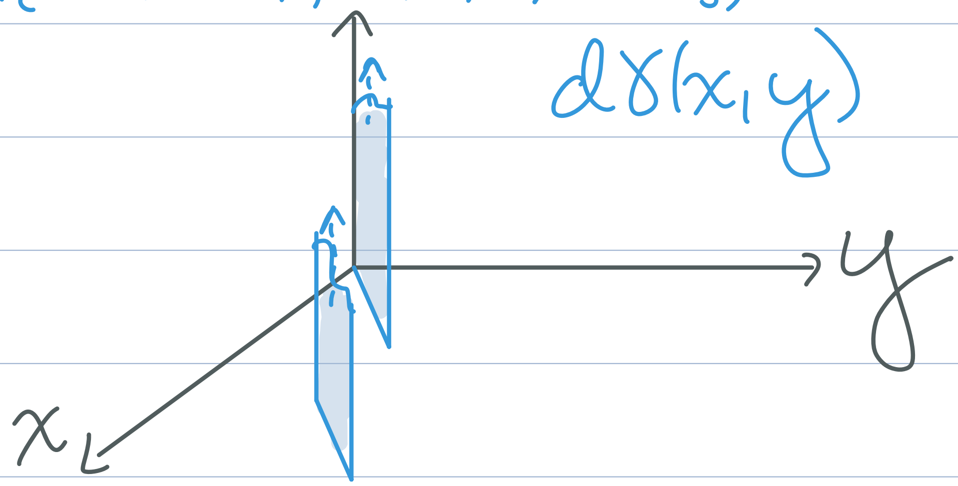


γ is the uniform probability measure supported on

$$\{(x, t(x)) : x \in [0,1] \cup [2,3]\}$$

$$\gamma(A\times B) = \mu\left((\mathrm{id}\times t)^{-1}(A\times B)\right)$$
$$= \mu\left(\{x\in X : (x,t(x))\in A\times B\}\right)$$

Side view:



Foreshadowing: When $\mu \ll \lambda^1$, we will see that γ is an optimal transport plan from μ to ν iff it is supported on $\{(x,t(x)) : x\in\mathbb{R}\}$ for an increasing function $t(x)$. $c(x,y) = d(x,y)^p,\ p\geq 1$

Using transport plans, we can now state...

Kantorovich's Optimal Transport Problem

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

$c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ lower semicts

$$\min_{\gamma : \gamma \in \Gamma(\mu,\nu)} \overbrace{\int_{X \times Y} c(x,y)\, d\gamma(x,y)}^{\mathbb{K}_c(\gamma)}$$

cost of moving mass from location x to y

If γ_* attains the minimum, we will call it an <u>optimal transport plan</u>.