Reminders: HW 2 due Friday
Comments on Project proposals:

How to deal with outliers
Density level sets
Metric measure spaces / Wasserstein distance
Distance to measure.

$$d_I(PH(R(X)), PH(R(Y))) \leq d_H(X, Y)$$

long H_1
1 H_0

H_1 abt half length
short extra H_0

long H_1
long extra H_0.

Problem: because sampling single point can drastically change PH.

Common setup:

$$X \overset{\text{unif}}{\sim} M + N(0, 1)$$

$M \subset \mathbb{R}^d$

$N(0,1)$: has tails; some probability of sampling outlier

What we often want is to capture topology of dense subsets, ignore outliers.

E.g. Carlsson & de Silva's idea of using kNN-based codensity filter.

→ what is k? what is threshold for cut?

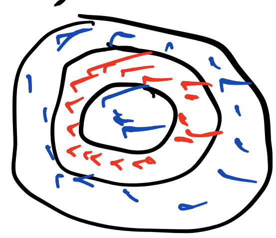
---

one solution for low dimensions:

don't do Rips. Do density level sets $f^{-1}([a,+\infty])$

$X \overset{\text{unif}}{\sim} S^1 + N(0,\sigma^2)$

pdf "f"



$PH(f)$: will have robust H_1 feature

if we have ground truth f, this is straightforward.

problem: we need to estimate pdf f from samples. Convergence is exponential in dimension

i.e. $X \subseteq \mathbb{R}^d$ need $O(\frac{1}{\epsilon}^d)$ samples for $O(\epsilon)$ accurate estimation of f.

in low dimensions ($d=1,2,3$) this is tractable.

---

can adapt stability results from levelset persistence

$$d_I(PH(f), PH(g)) \le \|f-g\|_\infty$$

so if f is pdf, g is $\hat{f}$ and estimator converges uniformly as $n \to \infty$, then barcodes converge.

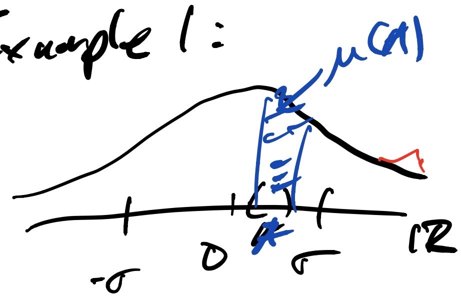
↑ assume Euclidean.

In higher dimensions ($d \geq 4$) need something else.
Metric measure spaces are good definition to work with.

Def: A metric measure space is a triple
$(\underset{\text{set}}{X}, \overset{\text{metric}}{d}, \underset{\text{measure}}{\mu})$ $\quad$ (X,d) is metric space

μ is a measure in sense of prob. measure.
mass of Borel subset $A \subseteq X$ is $\mu(A)$
and $A \cap B = \emptyset \Rightarrow \mu(A \cup B) = \underbrace{\mu(A) + \mu(B)}_{\text{(countably) additive}}$

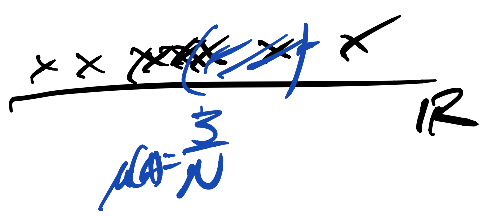
Example 1:



pdf f $N(0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma}$

$\mu_f(A) = \int_A f(x)\,dx$

Example 2: $X \overset{\text{iid}}{\sim} N(0,\sigma^2)$, N samples



empirical measure:
$\mu_X(A) = \frac{1}{N} |X \cap A|$
$= \#\{x \in A \mid x \in X\}$
$= \frac{1}{N} \int_A \sum_x \delta_x(y)\,dy$ $\leftarrow$ dirac delta

in what sense are measures close?
i.e. we might expect $\mu_X \to \mu_f$ if $X \sim f$, $N \to \infty$.

Note: positive measures have $\mu(A) \geq 0 \ \forall A$

probability measures have $\mu(X) = 1$

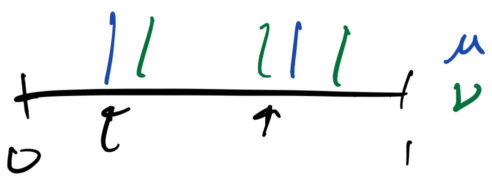
$\mu(\mathbb{R}^n)$

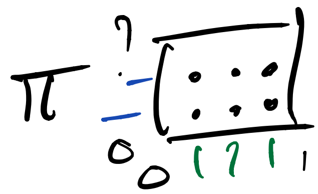
Transport plan: let μ, ν be measures on X, $\mu(X) = \nu(X)$ (same mass). A <u>transport plan</u> btw. μ and ν is a measure π on $\underline{X \times X}$ s.t. $\pi(A \times X) = \mu(A)$, $\pi(X \times B) = \nu(B)$

$$\begin{tikzcd} & X \overset{\pi}{\times} X \arrow[dl] \arrow[dr] & \\ \mu & & \nu \end{tikzcd}$$

idea is that all mass in μ is transported to mass in ν

Example:





$\mu(\cdot) = \frac{1}{6}$

In discrete setting, can encode in a Assignment matrix i.e.

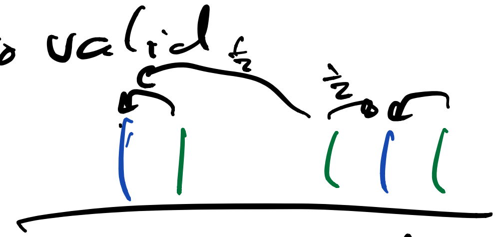
$$C = \begin{array}{c} \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix} \\ \begin{matrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{matrix} \end{array} \begin{matrix} \frac{1}{2} \\ \frac{1}{2} \end{matrix} \qquad \frac{1}{2}\delta_x$$

sum rows: projection onto μ

sum cols: projection onto ν

There are many possible transport plans eg

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$ also valid



k transport plans: $\sum_{i=1}^{k} w_i C_i \qquad \sum_{i=1}^{k} w_i = 1, \ w_i \geq 0$

is also a transport plan.

$c_i \geq 0$

The p^{th} order cost of a transport plan π is $C_p(\pi) = \left(\int_{X \times X} d_{\cancel{X}}(x,y)^p \, d\pi(x,y) \right)^{1/p}$

$\leq d(x,y) \, d\pi(x,y)$

Note: $C_1(\pi)$ is "Earth mover's cost" : ie sum of cost to move 1 unit of mass 1-unit of distance

Let $\Pi(\mu,\nu)$ denote set of all transport plans btw μ and ν.

Def: the pth order <u>Wasserstein</u> distance btw. μ and ν is

$$W_p(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} C_p(\pi)$$

<u>If</u> measures have finite pth moments, $W_p(\mu,\nu) < \infty$

Comment 1: There is a generalization of Wasserstein distance to barcodes. W_∞ is bottleneck distance.

Comment 2: There is a notion of Gromov-Wasserstein distance.

Comment 3: Finding π which minimizes C_p is subject of <u>optimal Transport</u>. Important in applications. eg. supply logistics.

Distance to measure (DTM)

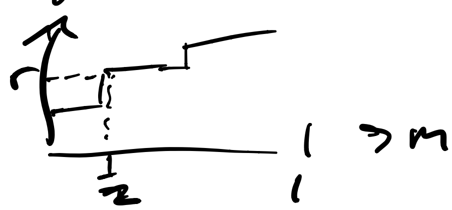
Let μ be a prob. measure on (X, d) and $m \in (0,1]$ be mass parameter. define distance $d_{\mu,m}$ to measure μ as

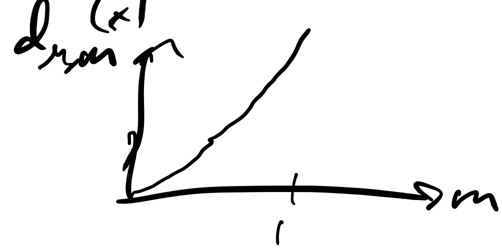
$$d_{\mu,m} : x \in X \longmapsto \sqrt{\frac{1}{m}\int_0^m \delta_{\mu,\ell}(x)^2 \, d\ell}$$

where $\delta_{\mu,m} : x \longmapsto \inf_{r \geq 0} \{\mu(\overline{B}(x;r)) > m\}$

$\delta_{\mu,\frac{1}{2}-\varepsilon}(x) = r$

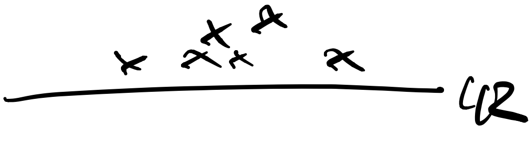
$\delta_{\mu,m}(x)$





fix m

μ: empirical measure



- points w/ small $d_{\mu,m}$ are "more important" than points w/ large DTM.

- m is a parameter that can be tuned

  $m \to 0$ distance to closest point

  $m \to 1$ smooth, may lose structure

idea: create sub-levelset filtration w/ $d_{\mu,m}$.

Buchet et al 2016 (Oudot Ch. 5.6)

Thm 3.1:

$$d_B(PH(d_{\mu,m}), PH(d_{\nu,m})) \le \frac{1}{\sqrt{m}} W_2(\mu,\nu)$$

Proof: ...

A measure ν is a submeasure of μ if $\nu(B) \le \mu(B)$ $\forall B$. Let $Sub_m(\mu)$ be submeasures w/ mass m.

Prop: Let μ be a prob. measure on X, and $m \in [0,1]$. Then

$$d_{\mu,m}(x) = \min_{\nu \in Sub_m(\mu)} \frac{1}{\sqrt{m}} W_2(m\delta_x, \nu)$$

← mass m conc. at x.

Pf: ν is measure of mass $m \Rightarrow \exists$ transport plan betw. ν and $m\delta_x$

$$W_2(m\delta_x, \nu)^2 = \int_X d_X(y,x)^2 \, d\nu(y)$$

let $d_x : X \to \mathbb{R} : d_x(y) = d(x,y)$

let $\nu_x : \nu_x(I) = \nu(d_x^{-1}(I)) \; \forall I \subseteq \mathbb{R}$

Let $F_\nu(y) = \nu([0,y))$; $F_\nu^{-1} : m \mapsto \inf\{t \in \mathbb{R} \mid F_\nu(t) > m\}$

$\Rightarrow F_{\nu_x}^{-1}(m) = \delta_{\nu,m}(x)$

$\shortparallel$

$\inf\{t \in \mathbb{R} \mid \nu_x([0,t)) > m\}$

$\shortparallel$

$\inf\{t \in \mathbb{R} \mid \nu(\{y \mid d(x,y) < t\}) > m\}$

$\shortparallel$

$\inf\{r \mid \mu(\overline{B}(x;r)) > m\}$

$\shortparallel$

$\delta_{\nu,m}(x)$

$$\int_X d_x(y,x)^2 \, d\nu(y) = \int_{\mathbb{R}^+} t^2 \, d\nu_x(t) =$$

$$= \int_0^m F_{\nu_x}^{-1}(l)^2 \, dl$$

ν submeasure of $\mu \Rightarrow F_{\nu_x}(t) \leq F_{\mu_x}(t) \; \forall t > 0$

$\Rightarrow F_{\nu_x}^{-1}(l) \geq F_{\mu_x}^{-1}(l)$

$$\Rightarrow W_2(m\delta_x ; \nu)^2 \geq \int_0^m F_{\mu_x}(l)^2 \, dl = \int_0^m \delta_{\mu,l}(x)^2 \, dl$$

$\shortparallel$

$m \, d_{\mu,m}(x)^2$

inequality is tight for set of submeasures

$R_{\mu,m}(x) \subseteq \mathrm{Sub}_m(\mu)$

$$\rotatebox{90}{=}$$

$$\{\nu \mid \mathrm{supp}(\nu) \subseteq \overline{B}(x, \delta_{\mu,m}(x)),$$
$$\nu(B(x, \delta_{\mu,m}(x))) = \mu(B(x, \delta_{\mu,m}(x))\}$$

Existence of such a measure

$\mu_{x,m} = \mu(\overline{B}(x, \delta_{\mu,m}(x)))$, rescale mass on $\partial B(x, \delta_{\mu,m}(x))$ if too much mass

$$\Rightarrow W_2(m\delta_x, \mu_{x,m})^2 = \frac{1}{m} d_{\mu,m}(x)^2$$

$$\frac{1}{\sqrt{m}} W_2(m\delta_x, \mu_{x,m}) = d_{\mu,m}(x) \quad \square$$

Thm: let μ, ν be prob. measures on $(\mathcal{X}, d)$, $m \in (0,1]$ mass param. Then

$$\| d_{\mu,m} - d_{\nu,m} \|_\infty \leq \frac{1}{\sqrt{m}} W_2(\mu, \nu)$$

Pf: $\sqrt{m}\, d_{\mu,m}(x) = W_2(m\delta_x, \mu_{x,m})$

Let π be OT plan μ to ν

$$\int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,y)^2 \, d\pi(x,y) = W_2(\mu, \nu)^2$$

Consider submeas $\mu_{x,m}$. $\exists\ \tilde{\pi}$ submeas. of π that transports $\mu_{x,m}$ to $\tilde{\nu}$ submeas of ν

$\Rightarrow W_2(\mu_{x,m}; \tilde{\nu}) \le W_2(\mu, \nu)$

again, $\sqrt{m}\, d_{\nu,m}(x) \le W_2(m\delta_x; \tilde{\nu}) \Rightarrow$

$$\sqrt{m}\, d_{\nu,m}(x) \le W_2(m\delta_x, \tilde{\nu}) \le \quad \text{Δ-ineq.}$$

$$W_2(m\delta_x, \mu_{x,m}) + W_2(\tilde{\nu}, \mu_{x,m})$$

$$\le \sqrt{m}\, d_{\mu,m}(x) + W_2(\nu, \mu)$$

Similarly,

$$\sqrt{m}\, d_{\mu,m} \le \sqrt{m}\, d_{\nu,m}(x) + W_2(\nu, \mu)$$

$$\Rightarrow \|d_{\mu,m} - d_{\nu,m}\|_\infty \le W_2(\mu, \nu)$$

Now, can apply levelset stability result.

$$d_B(PH(d_{\mu,m}), PH(d_{\nu,m})) \le \|d_{\mu,m} - d_{\nu,m}\|_\infty \le W_2(\mu,\nu)$$

$\square$.

outliers don't change measure much.

measure assigned to outlier is $\frac{1}{N}$

$W_1(\mu,\nu) \sim \frac{1}{N} r$

Problem: this function is defined on all of $\mathbb{R}^d$. Want simplicial construction.

Oudot Ch 5.6.