# Optimal transport and statistical inference

Nguyen Xuan Long

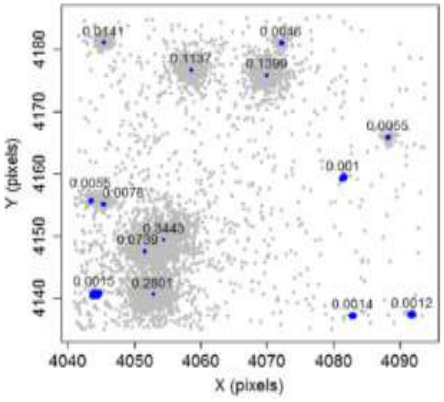
University of Michigan

Vietnam Institute for Advanced Studies in Mathematics
Hanoi, 7/6/2022

Acknowledgement: Aritra Guha, Jiacheng Zhu, Dat Do, Ding Zhao, Yun Wei, Nhat Ho, Yang Chen, Bach Viet Do, Sunrit Chakraborty

# Outline

# Data are vast, diverse, complex



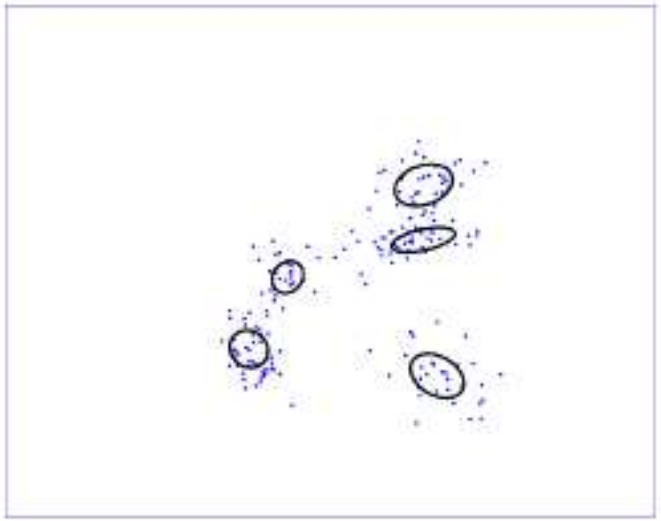


(a) Observations of photon sources near the center of the Orion Nebula from the *Chandra X-ray Observatory* (Jones et al, 2015)

(b) Trajectories of traffic primitives extracted from sensors-equipped vehicles driven in and around Ann Arbor, Michigan (Guha et al, 2020)

(c) Latent topics extracted from text corpus

# From data to statistical models



each dot represents a data point (photon source, car's signal, text doc)
we may assume

$$\text{data} \sim i.i.d.\ \ P(x|\theta)$$

- data are samples of a random variable $X$
- $P$ is a probability distribution on some observable domain
- unknown $\theta$ parameterizes $P$

# Distances on space of distributions

All statistical learning algorithms involve making movements on some (explicit or implicit) space of probability measures

- traditional notions of distance assume existence of density functions (typically wrt Lebesgue measure on Euclidean domain, or counting measure on discrete domain)

Standard distances in statistics, information theory, learning theory

- total variation: $D_{TV}(p,q) = \frac{1}{2}\int |p(x) - q(x)| d\mu$
- Hellinger:

$$h(p,q) = \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_{L^2}$$

- relative entropy (KL divergence):

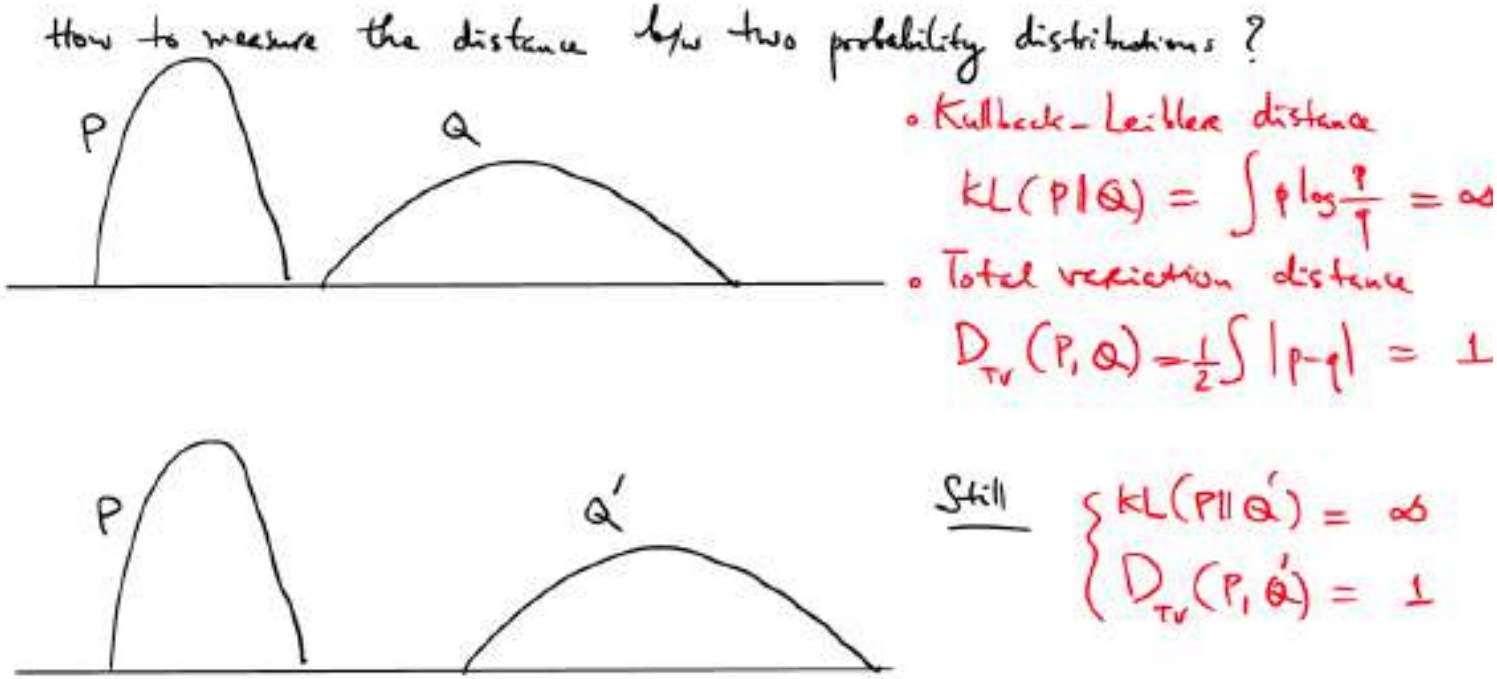
$$K(p,q) = \int p(x) \log(p(x)/q(x)) d\mu$$

- we also know

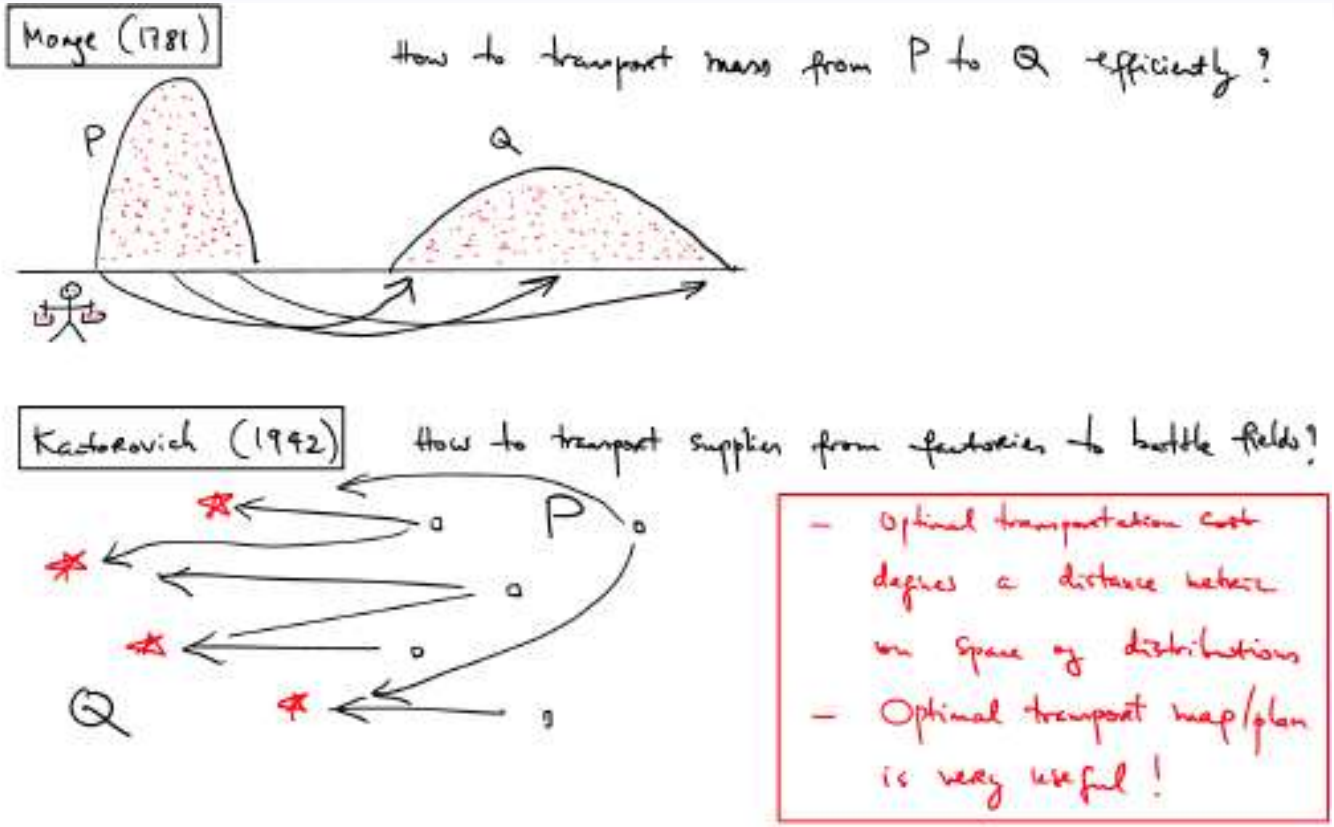
$$h^2 \leq D_{TV} \leq \sqrt{2}h \leq \sqrt{K}.$$

# Outline

There may be some metric structure in the supports of $P$ and $Q$, but they may be disjoint




Ideally, we want $0 < d(P,Q) < d(P,Q') < \infty$




Monge problem Find a map $x \mapsto T(x)$ s.t. if $X \sim P$ then $Y := T(X) \sim Q$. That is, find $T$ such that the pushforward measure satisfies $T_{\#}P = Q$.

Kantorovich problem enlarges Monge's into a solvable problem: find a "stochastic map" , i.e., a coupling of $P$ and $Q$ that minimizes expected cost of transportation

# Basic definition and facts

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two complete and separable metric probability spaces. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a non-negative lower-semicontinuous cost function.

Lemma Let $\Pi(\mu, \nu)$ be the space of all couplings of $\mu, \nu$, i.e., all joint distributions on $\mathcal{X} \times \mathcal{Y}$ that admit marginal distributions $\mu$ and $\nu$. Then there exists a coupling $\mu \in \Pi(\mu, \nu)$ that minimizes the total cost

$$\mathbb{E}_\pi c(X, Y) = \int c(x, y) d\pi(x, y).$$

## Kantorovich duality

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) = \max_{(\psi,\phi)\in L^1(\mu)\times L^1(\nu):\phi+\psi\leq c} \int \psi(x)d\mu(x) + \phi(y)d\nu(y).$$

(i) $\geq$ is trivial; $=$ is due to convex optimization in Banach spaces

(ii) dual formulation allows precise characterization of the optimal $\pi$ and corresponding optimal $\phi, \psi$: assume that the optimal cost is finite, then the support of $\pi$ satisfies $\psi(x) + \phi(y) = c(x,y)$ almost surely.

## Kantorovich duality

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) = \max_{(\psi,\phi)\in L^1(\mu)\times L^1(\nu):\phi+\psi\leq c} \int \psi(x)d\mu(x) + \phi(y)d\nu(y).$$

(i) $\geq$ is trivial; $=$ is due to convex optimization in Banach spaces

(ii) dual formulation allows precise characterization of the optimal $\pi$ and corresponding optimal $\phi, \psi$: assume that the optimal cost is finite, then the support of $\pi$ satisfies $\psi(x) + \phi(y) = c(x,y)$ almost surely.

(iii) when $c$ is continuous, then the support of $\pi$ is $c$-cyclic monotone, i.e., the set of pairs $(x_n, y_n)$ such that

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

hold for all such pairs and all $N$ (with the convention $y_{N+1} = y_1$).

# Two useful properties

(I) Wasserstein metric: Let $(X, d)$ be a Polish metric space, $r \in [1, \infty)$. For any two pm's $\mu, \nu$ on $\mathcal{X}$, the Wasserstein metric of order $r$ is given given by

$$\begin{aligned} W_r(\mu, \nu) &= \left( \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^r d\pi(x, y) \right)^{1/r} \\ &= \inf \left\{ [\mathbb{E} d(X, Y)^r]^{1/r}, \;\; X \sim \mu, Y \sim \nu \right\}. \end{aligned}$$

# Two useful properties

(I) Wasserstein metric: Let $(X, d)$ be a Polish metric space, $r \in [1, \infty)$. For any two pm's $\mu, \nu$ on $\mathcal{X}$, the Wasserstein metric of order $r$ is given given by

$$\begin{aligned} W_r(\mu, \nu) &= \left( \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^r d\pi(x, y) \right)^{1/r} \\ &= \inf \left\{ [\mathbb{E} d(X, Y)^r]^{1/r}, \;\; X \sim \mu, Y \sim \nu \right\}. \end{aligned}$$

We can define properly distance metric on distribution of any "complex" data populations, as long as we have a metric $d$ on the data instances available

(II) Brenier-Rachev-Ruschendorf theorem: Let $c(x, y) = \|x - y\|^2$ in $\mathbb{R}^n$. $\mu$ and $\nu$ two pm's with bounded second moments. If $\mu$ is absolute continuous wrt the Lebesgue measure, then there is a unique optimal coupling of $\pi$ of $x, y$, under which $y$ is uniquely determined by $x$ almost surely. In fact, for some lsc convex function $\psi$,

$$y \in \partial\psi(x) \text{ almost surely.}$$

In other words, the Monge transport plan $x \mapsto T(x) = \partial\psi(x)$ exists, and the optimal coupling $\pi$ is characterized as the pushforward measure from the source distribution $\mu$:

$$\begin{aligned} \nu &= T_{\#}\mu \\ \pi &= (Id, T)_{\#}\mu. \end{aligned}$$

See Villani (2008) and Ambrosio-Gigli-Savare (2005) for much recent advances in analysis, PDEs, differential geometry associated with OT

# Implications to statistical learning/ inference

(1) we can define distance between any data populations, which inherit the metric $d$ of data instances

(2) if we know the source distribution for $X \sim \mu$, we can perhaps model the target $Y \sim \nu$ via a (optimal transport) map $X \mapsto Y = T(X)$, i.e., $\nu$ is a pushforward measure of $\mu$ by the map $T$:

$$\nu = T_{\#}(\mu)$$

- the question of "learning" a target distribution becomes the learning of transport map $T$, e.g., Wasserstein-GAN (Arjovsky et al, 2017)

(3) while all this seems nice, optimal transport is only useful if the metric on the data $d$ is meaningful, and map $T$ can be effectively learned
- we will illustrate this in the following "domain adaptation problem"

# Outline

# Domain adaptation problems

1. If I learned to drive in Arizona, can I adapt my experience to driving in California? Answer: yes.

   How about driving in Hanoi?
   - here, driving experience = samples of time series of driving trajectories

2. If a robot knows how to pick up an object, how can it be taught to bend and pick up efficiently?

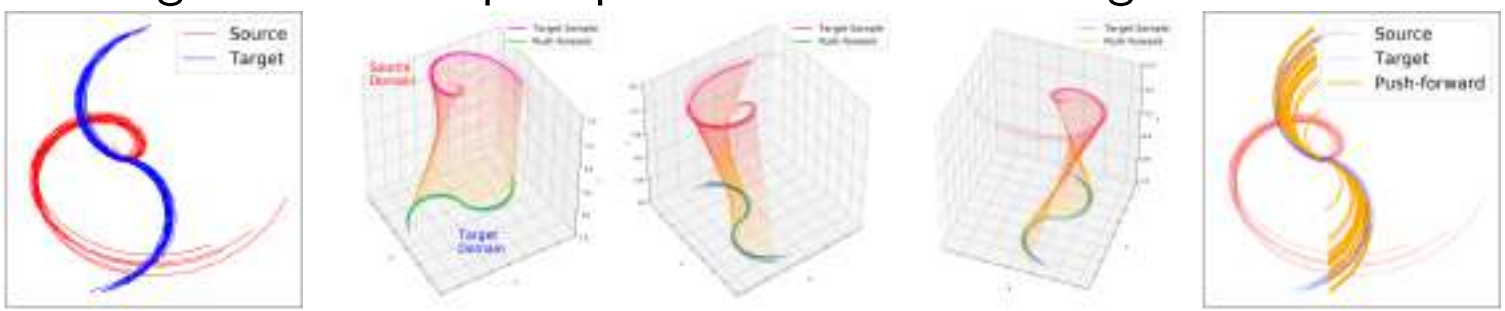
> General problem If we know/ can learn well distribution $\mu$, how can we adapt $\mu$ to learn another distribution $\nu$

# Our approach: functional optimal transport

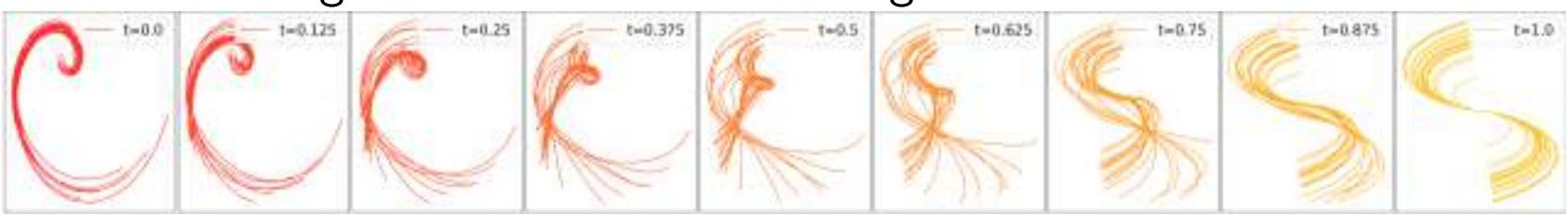
Main ideas: (Zhu, Guha, Do, Xu, Nguyen and Zhao, 2021)

- each data point is a realization of a random function
- given source and target distributions $\mu$ and $\nu$ on functions
- learn the transport map from space of compact linear operators

Pushing forward sampled paths for source to target distribution




Geodesic linking source distribution to target distribution



# Learning OT map in function space

Let $H_1, H_2$ be Hilbert space of functions endowed with Borel probability measures $\mu_1$ and $\mu_2$, resp.
Let $\mathcal{B}_{HS}(H_1, H_2)$ be the space of Hilbert-Schmidt operators, i.e., a Hilbert space of linear operators endowed with the scalar product

$$\langle A, B \rangle_{HS} = \sum_{i=1}^{\infty} \langle AU_i, BU_i \rangle_{H_2}$$

where $(U_i)_{i=1}^{\infty}$'s form a complete orthonormal basis of $H_1$.

Consider the optimization problem

$$\inf_{T \in \mathcal{B}_{HS}(H_1, H_2)} J(T) := W_2^2(T_{\#}\mu, \nu) + \eta \| T \|_{HS}^2 \tag{1}$$

$\eta > 0$ is a regularization parameter.

## Lemma

*Assume that $\mu$ and $\nu$ have bounded second moments:*

$$E_{f_1 \sim \mu} \|f_1\|_{H_1}^2 < \infty, \quad E_{f_2 \sim \nu} \|f_2\|_{H_2}^2 < \infty \qquad (A.1)$$

*then the objective (1) is a strictly convex function, which admits a unique minimizer.*

In practice, given i.i.d. samples

$$f_{11}, f_{12}, \ldots, f_{1n_1} \sim \mu,$$

$$f_{21}, f_{22}, \ldots, f_{2n_2} \sim \nu,$$

the empirical version of our optimization problem becomes:

$$\inf_{T \in \mathscr{B}_{HS}} \hat{J}_n(T), \quad \hat{J}_n(T) := W_2^2(T_{\#}\hat{\mu}_{n_1}, \hat{\nu}_{n_2}) + \eta \| T \|_{HS}^2, \tag{2}$$

where $\hat{\mu}_{n_1} = \frac{1}{n_1} \sum_{l=1}^{n_1} \delta_{f_{1l}}$ and $\hat{\nu}_{n_2} = \frac{1}{n_2} \sum_{k=1}^{n_2} \delta_{f_{2k}}$ are the empirical measures, and $n = (n_1, n_2)$.

Moreover, restrict $T$ to a $K_1 \times K_2$ dimensional subspace of $\mathscr{B}_{HS}$

## Lemma

*Under assumption (A.1), the following hold.*

- *For any fixed* $C_0 > 0$,

$$\sup_{\|T\| \leq C_0} |\hat{J}_n(T) - J(T)| \xrightarrow{P} 0 \quad (n \to \infty). \tag{3}$$

- *For any* $n, K$, $\hat{J}_n$ *has a unique minimizer* $\hat{T}_{K,n}$ *over* $B_K$. *Moreover, there exists a finite constant* $D$ *such that* $P(\sup_K \|\hat{T}_{K,n}\| < D) \to 1$ *as* $n \to \infty$.

It then follows that

## Theorem

*The minimizer of Eq.* (2) *for* $\hat{T}_{K,n} \in B_K$ *is a consistent estimate for the minimizer of Eq.* (1). *Specifically,* $\hat{T}_{K,n} \xrightarrow{P} T_0$ *as* $K_1, K_2, n_1, n_2 \to \infty$.

# Additional considerations

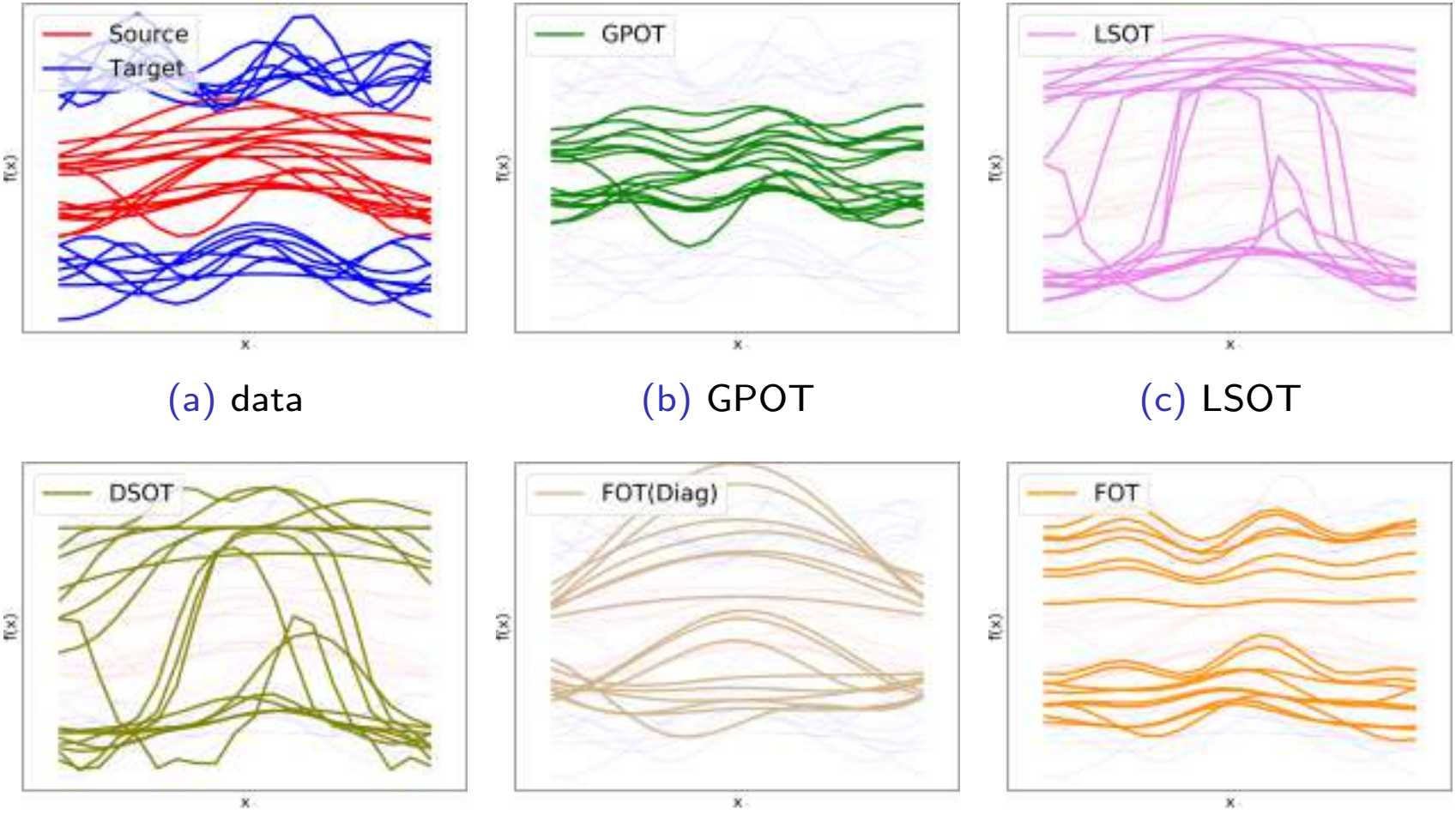
(i) in practice, sampled functions $(f_{1l})_{l=1}^{n_1}$, $(f_{2l})_{l=1}^{n_2}$ are observed only at a finite number of design points $d_1, d_2$, resp.

(ii) the HS operator is approximated by dimension truncation (to spaces of $K_1 \times K_2$ matrices)

(iii) consistency theory can be extended under the regime that $n_1, n_2, K_1, K_2 \to \infty$, and $d \to \infty$ suitably

- see the paper (Zhu et al, 2021) for details.

# FOT vs non-functional approaches




(a) data (b) GPOT (c) LSOT

(d) DSOT (e) FOT with diagonal $\Lambda$ (f) FOT

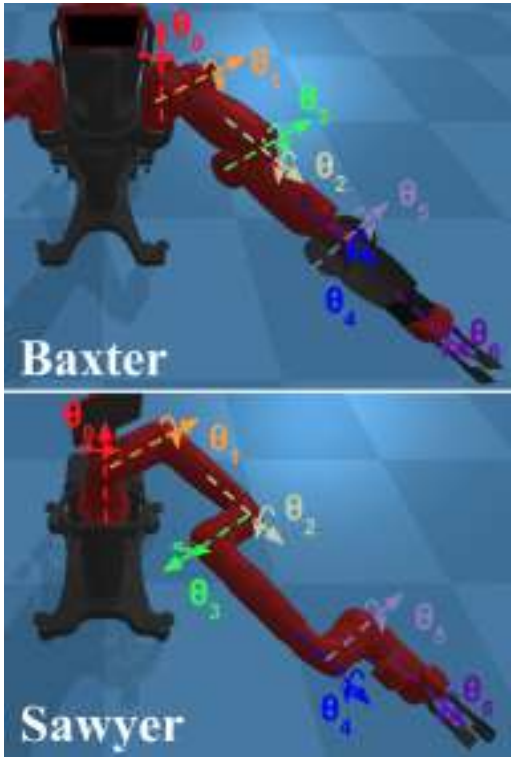
Figure: Pushforward measures of functions obtained by various approaches on mixtures of sinusoidal functions data

# FOT vs non-functional approaches: quantitative comparison

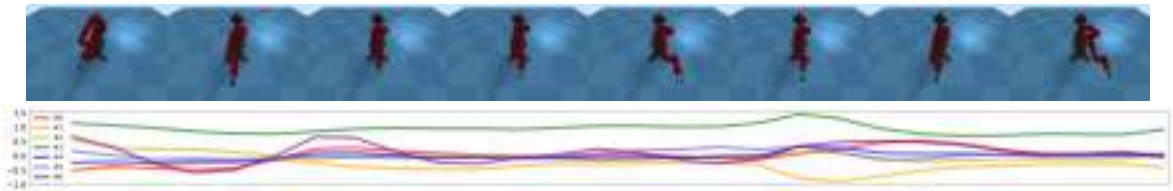
| Method | $1 \to 1$ | $1 \to 2$ | $2 \to 1$ | $2 \to 2$ | $2 \to 3$ |
|---|---|---|---|---|---|
| GPOT | 17.560 | 12.895 | 15.263 | 61.561 | 39.159 |
| LSOT | 133.434 | 94.229 | 117.832 | 929.108 | 663.461 |
| DSOT | 6.871 | 13.226 | 9.679 | 46.521 | 41.009 |
| FOT | **2.873** | **11.982** | **3.316** | **44.071** | **32.547** |

Table: Quantitative comparison on the mixture of sinusoidal functions data. The maps obtained by FOT method achieved the best performance under the Wasserstein distance objective.

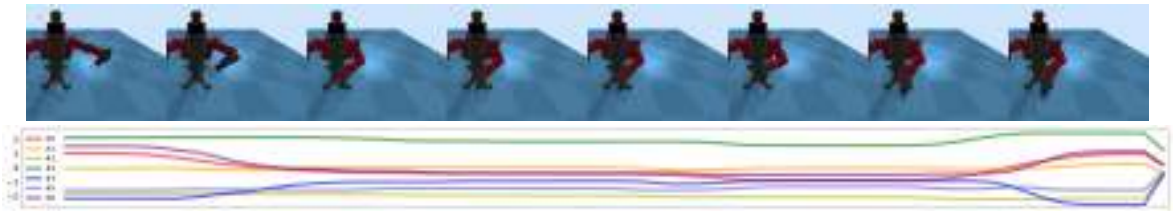
# Domain adaptation for robotic arm movements



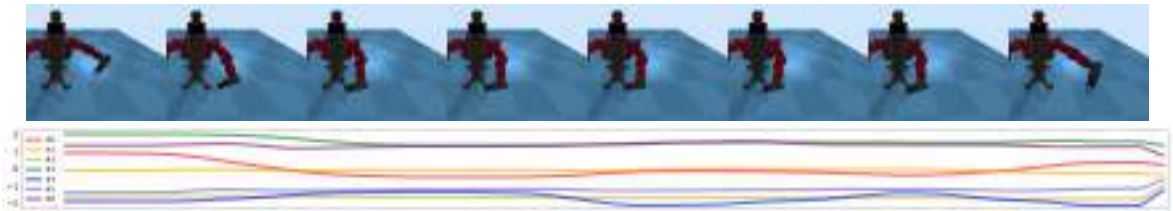

(a) The arm of the Baxter robot and the Sawyer robot used in MIME dataset and Roboturk dataset. They share a similar structure, 7 joints and one end effector.



(b) Source motion: "Roboturk-bins-Bread" by Sawyer.



(c) Target motion: "MIME Picking (left-hand)" by Baxter.



(d) The pushforward motion of the transport map looks like Baxter's but inherits trait of Sawyer's motions.

# Augmenting deep learning models

Semi-supervised learning

Time series motion prediction tasks with FOT augmentation
(use FOT to generate more samples in target domain for training with deep learning predictive models)

| Method | LSTM | ANP | RANP | MAML* | TL* | $FOT_{LSTM}$ | $FOT_{ANP}$ | $FOT_{RANP}$ | $FOT_{MAML}$ | $FOT_{TL}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| R1→M1 | 2.0217 | 1.3261 | 1.9874 | 0.0307 | 0.5743 | 0.0271 | 0.0963 | 0.0687 | 0.0165 | 0.0277 |
| R1→M2 | 1.6821 | 1.0951 | 1.5681 | 0.0374 | 0.7083 | 0.0414 | 0.1642 | 0.1331 | 0.0191 | 0.0446 |
| R2→M1 | 1.3963 | 0.6642 | 1.7256 | 0.0327 | 0.2491 | 0.0277 | 0.0951 | 0.0696 | 0.0202 | 0.0906 |
| R2→M2 | 1.1952 | 0.6307 | 1.3659 | 0.0477 | 0.4020 | 0.0331 | 0.1620 | 0.1554 | 0.0167 | 0.0406 |

Table: MSE error results of different predictive models.
R1: Roboturk-bins-bread, R2: Roboturk-pegs-RoundNut, M1:MIME1-Pour-left, M2: MIME12-Picking-left.

Our story so far: using the pushforward map $T$ to model a target distribution via

$$\nu = T_{\#}\mu$$

in a complex *data domains*, but there are huge technical challenges, when

- incorporating domain knowledge of the support of $\nu$, and $T$,
- and when the interest is *not* data distribution but on distribution on meaningful quantities related to it

This leads to optimal transport on distributions on the space of quantities of interest. Thus, the starting point is with probability models via latent *random* variables/structures.

# Outline

Let's talk about latent structured models!

# Observations of the Orion constellation

ancient data

an ancient model

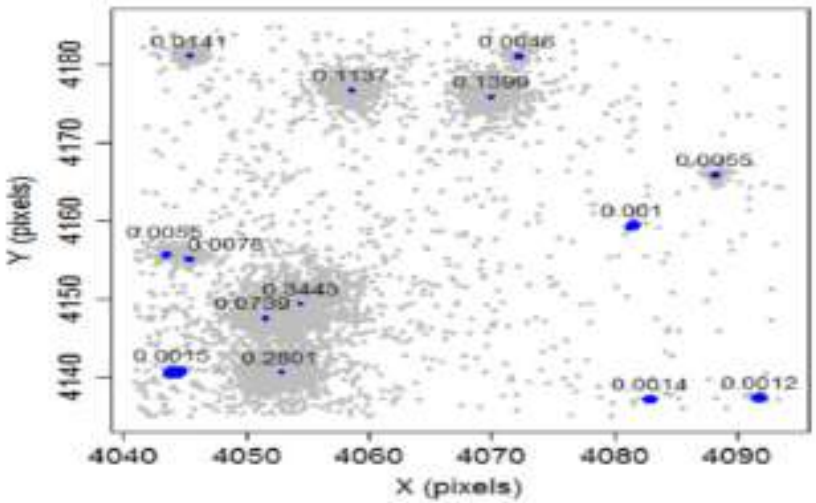
# Observations of the Orion constellation

ancient data

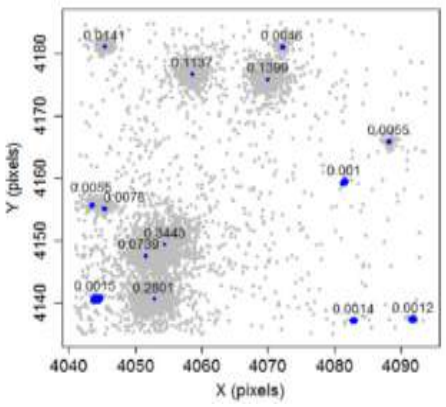
an ancient model

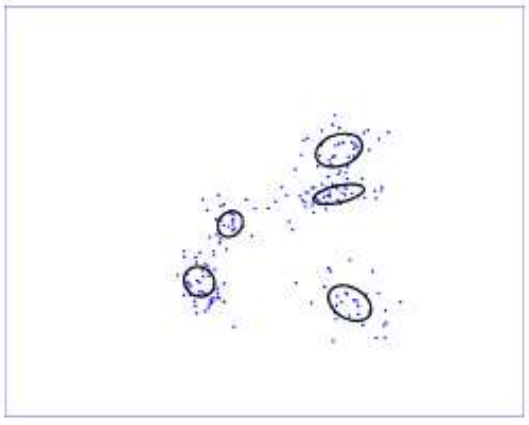
Orion nebula

X-ray data via Chandra observatory



Consider a situation where the quantity of interest is not data distribution itself: photon sources in the Orion Nebula



Assume the pixel locations are sample from a mixture distribution

$$X_1, \ldots, X_n \sim \sum_{j=1}^{K} p_j f(x|\theta_j)$$

there are $K$ photon sources; $\theta_j$ represents information about the arrival time, location, energy level of photon source $j$

- $p_j$: the probability that the photon comes from source $j$ for certain type of stellar and inter-galatic events of interest (exploded stars, star birth, etc)
- $f$ is a probability kernel which captures distribution of observations (e.g., King profile for the location observations)

The mixture distribution used to describe the observed locations has very little scientific meaning:

$$P_G = \sum_{j=1}^{K} p_j f(x|\theta_j)$$

Of interest to astrophysicists and astronomers is the mixing measure

$$G = \sum_{j=1}^{K} p_j \delta_{\theta_j}$$

Here, optimal transport continues to play a fundamental role in characterizing the learning behavior of quantities of interest, where the underlying metric structure of support is derived from the rich structure of the probabilistic models

# Wasserstein metric on $G$

How to define a metric on the space of $\theta$'s:

- suppose $\theta \in \mathbb{R}^d$, we may take $d(\theta, \theta') := \|\theta - \theta'\|_r^r$
- better yet, make use of Hellinger distance:

$$d(\theta, \theta') \quad := \quad h(f(\cdot|\theta), f(\cdot|\theta'))$$

- this results in the following Wasserstein metric

$$W(G, G') = \inf_{\pi \in \Pi(G, G')} \int d(\theta, \theta') d\pi$$

This is called "composite distance" in Nguyen (2013)

# Statistical learning methods

Statistical formulation: given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_G$ for some "truth"

$$G = G_0 = \sum_{j=1}^{K_0} p_j^0 \delta_{\theta_j^0}$$

To learn $G$, we can either apply

- maximum likelihood estimate via the EM algorithm:

$$\hat{G} := \text{argmax}_G \sum_{i=1}^{n} \log p_G(X_i)$$

- Bayesian method: place a prior distribution on $G$, and apply Bayes formula to obtain the posterior dist. $\Pi(G|X_1, \ldots, X_n)$.

Theoretical questions: in what sense does the estimate $\hat{G}$, or the posterior distribution for $G$ converge to the truth $G_0$ in the Wasserstein space?

# Statistical learning methods

Statistical formulation: given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_G$ for some "truth"

$$G = G_0 = \sum_{j=1}^{K_0} p_j^0 \delta_{\theta_j^0}$$

To learn $G$, we can either apply

- maximum likelihood estimate via the EM algorithm:

$$\hat{G} := \text{argmax}_G \sum_{i=1}^{n} \log p_G(X_i)$$

- Bayesian method: place a prior distribution on $G$, and apply Bayes formula to obtain the posterior dist. $\Pi(G|X_1, \ldots, X_n)$.

Theoretical questions: in what sense does the estimate $\hat{G}$, or the posterior distribution for $G$ converge to the truth $G_0$ in the Wasserstein space?

A key to these questions is the derivation of inverse bounds

# A simpler problem: deconvolution problem

Example: we are interested in the distribution of signal $\theta_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, given i.i.d. noisy observations of the form

$$\begin{aligned} X_i &= \theta_i + \epsilon_i, \\ \epsilon_i &\sim \text{i.i.d.} \quad f, \quad \epsilon_i \perp \theta_i. \end{aligned}$$

Suppose that $\theta_i \sim \text{i.i.d. } G_0$, then $X_i \sim \text{i.i.d. } f * G_0$.

# A simpler problem: deconvolution problem

Example: we are interested in the distribution of signal $\theta_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, given i.i.d. noisy observations of the form

$$\begin{aligned} X_i &= \theta_i + \epsilon_i, \\ \epsilon_i &\sim \text{i.i.d. } f, \quad \epsilon_i \perp \theta_i. \end{aligned}$$

Suppose that $\theta_i \sim \text{i.i.d. } G_0$, then $X_i \sim \text{i.i.d. } f * G_0$.

An inverse bound is an inequality of the type

$$\boxed{W_2(G_0, G) \leq \Phi(V(f * G_0, f * G))}$$

where

- LHS is an optimal transport distance e.g., $W_1, W_2, \ldots$ on distribution of the latent $\theta$
- $V$ is the total variation distance of data populations $f * G_0$ and $f * G$
- $\Phi : [0, \infty) \to [0, \infty)$ is a strictly increasing function, $\Phi(0) = 0$.

# why inverse bounds are important

- Inverse bounds are useful for deriving rates of convergence of the latent $\theta$ distributions, once the convergence in distribution of data population $X$ has been established
- The opposite direction of the inverse bound is typically easy
  - let $(\theta, \theta_0)$ be an optimal coupling of $G$ and $G_0$
  - then $(X, X_0)$ where $X = \theta + \epsilon$, $X_0 = \theta_0 + \epsilon$ represent a coupling of $f * G$ and $f * G_0$, so

$$W_2^2(f * G, f * G_0) \leq \mathbb{E}\|X - X_0\|^2 = W_2^2(G, G_0).$$

  - to obtain bound for $V(f * G, f * G_0)$, by an application of Jensen's inequality one gets

$$\begin{aligned} \|f * G - f * G_0\|_{L^1} &\leq \mathbb{E}_{\theta,\theta_0}\|f(\cdot - \theta) - f(\cdot - \theta_0)\|_{L^1} \\ &\leq C W_1(G, G_0), \end{aligned}$$

    where $C$ is an "integrated" Lipschitz constant of the density function $f(\cdot - \theta)$ wrt $\theta$.
  - similar upper bounds on any $f$-divergence using the same technique (Nguyen, 2013)

# Inverse bounds for convolutional models $X = \theta + \epsilon$

- let $f$ be a pdf on $\mathbb{R}^d$ that is symmetric around 0, and the Fourier transform of $f$ satisfies $\tilde{f}(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$.
- Borrowing from deconvolution literature (cf. Fan (1991)) say
  - $f$ is ordinary smooth with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}^{-2} d\omega \lesssim (1/\delta)^{2d\beta}$ as $\delta \to 0$,
  - $f$ is supersmooth with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}^{-2} d\omega \lesssim \exp(2d\delta^{-\beta})$ as $\delta \to 0$.

# Inverse bounds for convolutional models $X = \theta + \epsilon$

- let $f$ be a pdf on $\mathbb{R}^d$ that is symmetric around 0, and the Fourier transform of $f$ satisfies $\tilde{f}(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$.
- Borrowing from deconvolution literature (cf. Fan (1991)) say
  - $f$ is ordinary smooth with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}^{-2} d\omega \lesssim (1/\delta)^{2d\beta}$ as $\delta \to 0$,
  - $f$ is supersmooth with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}^{-2} d\omega \lesssim \exp(2d\delta^{-\beta})$ as $\delta \to 0$.
- for any pair of $G$ and $G'$ whose support lie in a bounded subset of $\mathbb{R}^d$, then (Nguyen, 2013)
  - if $f$ is ordinary smooth, then for any $m < 4/(4 + (2\beta + 1)d)$, for some constant $C(d, \beta, m)$,

$$\boxed{W_2^2(G, G') \leq C(d, \beta, m) V(f * G, f * G')^m.}$$

  - if $f$ is supersmooth, then there is $C(d, \beta) > 0$ such that

$$\boxed{W_2^2(G, G') \leq C(d, \beta)[-\log V(f * G, f * G')]^{-2/\beta}.}$$

# Beyond convolutional model

- let kernel $f$ be the Gaussian pdf on $\mathbb{R}$: $f(x|\mu,\nu)$.
- let $G$ be a distribution on the bivariate parameter $(\mu,\nu) \in \mathbb{R} \times \mathbb{R}_+$.
- then the marginal distribution of the observed $X$ is the mixture distribution

$$X \sim P_G := \int f(\cdot|\mu,\nu) dG(\mu,\nu)$$

- Open question: obtain an inverse bound of the form

$$W_2(G_0, G) \leq \Phi(V(P_{G_0}, P_G))$$

where the optimal transport distance is defined in a natural way for the distribution on $\mathbb{R} \times \mathbb{R}_+$, under the metric, e.g.:

$$d((\mu_1,\nu_1),(\mu_2,\nu_2)) := |\mu_1 - \mu_2| + |\nu_1^{-1} - \nu_2^{-1}|.$$

Where are the "great" models in our time?

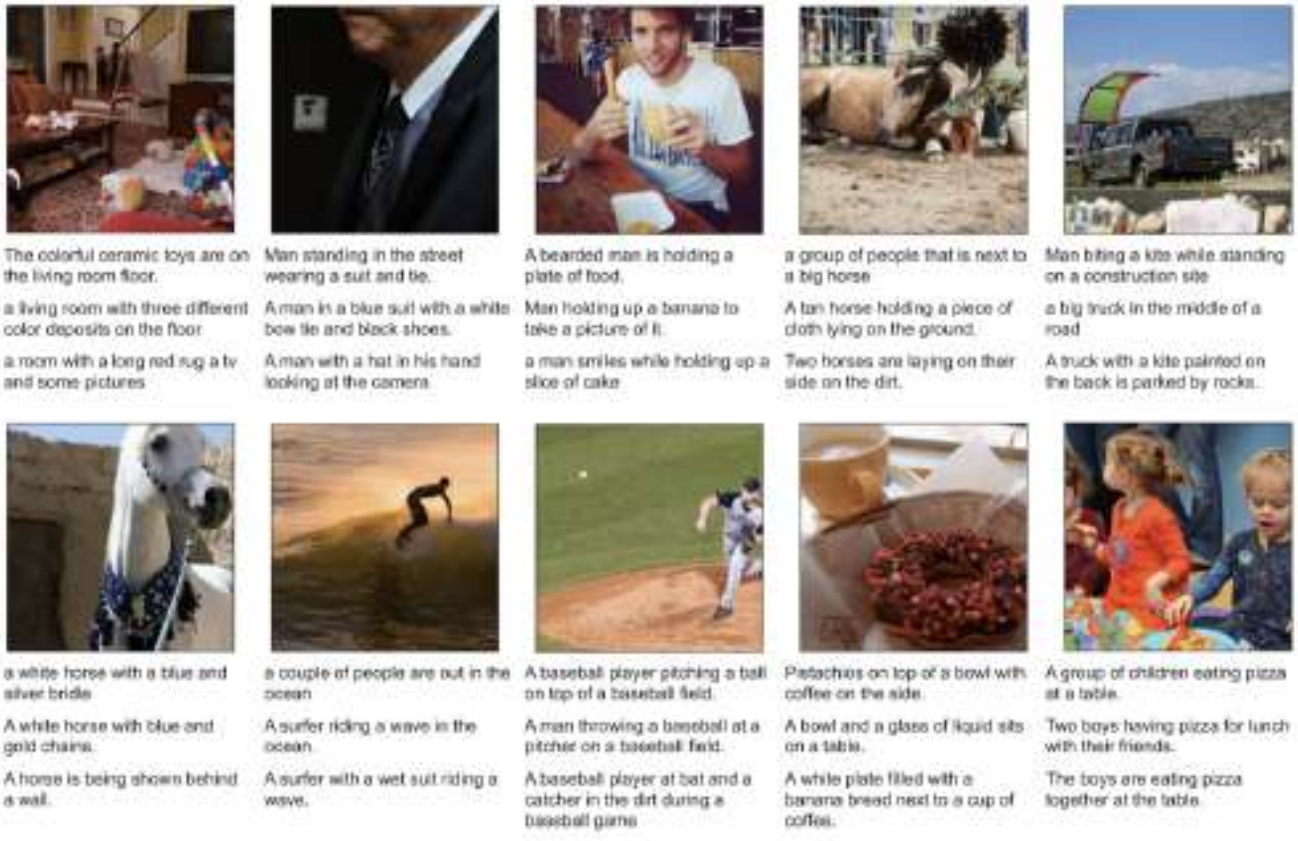

Gato: a "multi-modal, multi-task, multi-embodiment generalist agent"

Where are the "great" models in our time?


Gato: a "multi-modal, multi-task, multi-embodiment generalist agent"
"...Gato can sense and act with different embodiments across a wide range of environments using a single neural network with the same set of weights. (It) was trained on 604 distinct tasks..."

Deepmind's publication "A Generalist Agent", 5/19/2022

# Nando de Freitas: "The Game is Over!" But is it?




We heard similar claims before: including but not restricted to neural nets circa 1950s, 1980s (convolutional neural nets), 2010s (cnn on steroid), 2020s (now with transformers)

Despite truly impressive demos, AI models are still unreliable in domains where reliability really matters!

These models don't understand their domain, and neither do we understand if and when they work! But we always try to make progress at creating and (hopefully) understanding them.

Despite truly impressive demos, AI models are still unreliable in domains where reliability really matters!

These models don't understand their domain, and neither do we understand if and when they work! But we always try to make progress at creating and (hopefully) understanding them.

"great models are temporary, mathematical principles are forever"

*more to the point*, I mean the mathematics which helps to justify the presence of latent (hidden) variables, the mechanism for memory and attention, and the statistical/computational theory which helps to explain and achieve the emergence of such representations

# From data assumptions to models of populations

de Finetti's Theorem: If $(X_j)_{j=1}^{\infty}$ is an infinite exchangeable sequence of random variables, i.e.,

$$(X_1, \ldots, X_N) \stackrel{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(N)}) \quad \forall N, \forall \pi$$

then there exists a random probability $P$ such that

$$X_1, X_2, \ldots | P \overset{i.i.d.}{\sim} P$$

- Mixtures of product distributions for $N$-sequence $X_1, \ldots, X_N$: conditionally given some $\theta$, the $X_j$ are i.i.d.

$$P_{G,N}(X_1 \in A_1, \ldots, X_N \in A_N) = \int \prod_{n=1}^{N} P_\theta(X_n \in A_n | \theta) G(d\theta)$$

  - kernel $P_\theta$ known and uniquely parameterized by $\theta \in \Theta$
  - $G$ is de Finetti mixing measure on $\Theta$;
    $G$ characterizes heterogeneity of underlying data population

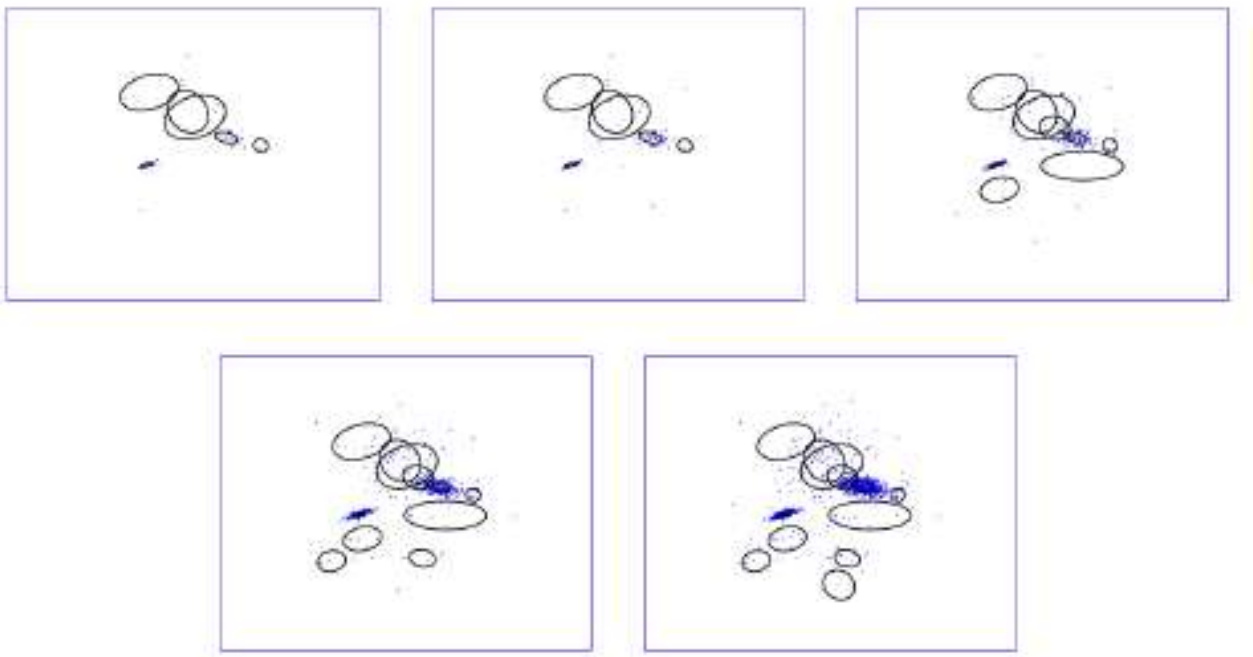
# Compositional data structures

Data are often composed of a collection of dependent populations

- there are multiple hospitals, each hospital has many patients
- there are different animals, each animal carry a set of genes
- different countries, each of which is organized into regions, each of which is organized into counties, with residents in each of them
- "activity recognition problem": a collection of computer users, each user is associated with a collection of computer related activities (organized by days), each day has a collection of activities (apps run)
- a collection of text corpora, each text corpus is a collection of documents, each document is a collection of words
- a database of images divided by groups, each image is a collection of image patches, each patch a collection of pixels or other specific computer vision elements

# Exchangeable collection of data sets

Each data set modeled via a mixture model
$\Longrightarrow$ they are coupled to enable "borrowing of strength"

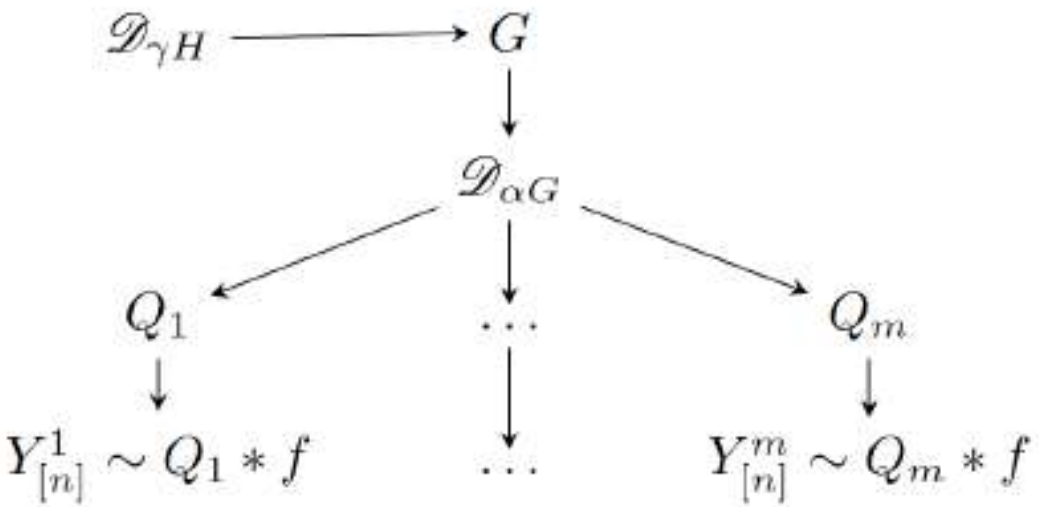


[courtesy M. Jordan's slides]

This gives rise naturally to a hierarchical model

# A hierarchical model setting

$m$ groups of data, each of which is given an $n$-sample




A key point here: $\mathscr{D}_{\alpha G}$ represents a distribution on the space of distributions

# Metrics on Bayesian hierarchies of distributions

Need a notion of distance between, say $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$

Recall: for $G, G' \in \mathcal{P}(\Theta)$, space of Borel probability measures on $\Theta$,

$$W_r(G, G') := \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

$\mathcal{T}(G, G')$ is the space of all couplings of $G, G'$.

# Metrics on Bayesian hierarchies of distributions

Need a notion of distance between, say $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$

Recall: for $G, G' \in \mathcal{P}(\Theta)$, space of Borel probability measures on $\Theta$,

$$W_r(G, G') := \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

$\mathcal{T}(G, G')$ is the space of all couplings of $G, G'$.

Distance between measures of measures in Bayesian hierarchy:

Let $\mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta))$ (the space of Borel probability measures on $\mathcal{P}(\Theta)$). Define Wasserstein distance between $\mathcal{D}, \mathcal{D}'$

$$W_r(\mathcal{D}, \mathcal{D}') := \inf_{\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')} \left[ \int W_r^r(G, G') \, d\mathcal{K}(G, G') \right]^{1/r}.$$

$\mathcal{T}(\mathcal{D}, \mathcal{D}')$ is the space of all couplings of $\mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta))$

# A lemma and a "theorem"

(Nguyen, Bernoulli 2016)

## Lemma

(a) *Let* $G, G' \in \mathcal{P}(\Theta)$, *and* $\mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta))$ *such that* $\int P d\mathcal{D} = G$ *and* $\int P d\mathcal{D}' = G'$. *For* $r \geq 1$, *if* $W_r(\mathcal{D}, \mathcal{D}')$ *is finite then*

$$W_r(\mathcal{D}, \mathcal{D}') \geq W_r(G, G').$$

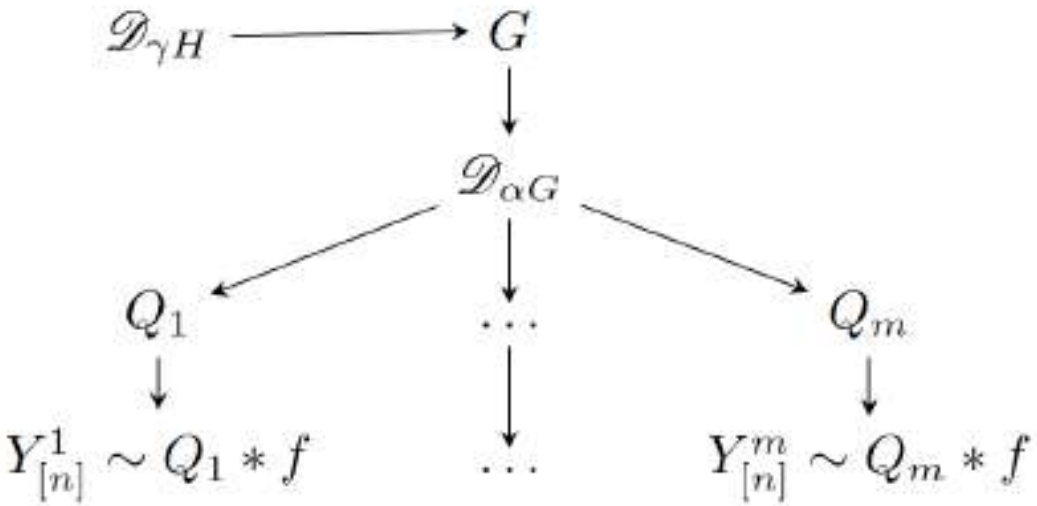
(b) *If* $\mathcal{D} = \mathscr{D}_{\alpha G}$ *and* $\mathcal{D}' = \mathscr{D}_{\alpha G'}$ *(same* $\alpha$*), then*

$$W_r(\mathcal{D}, \mathcal{D}') = W_r(G, G').$$

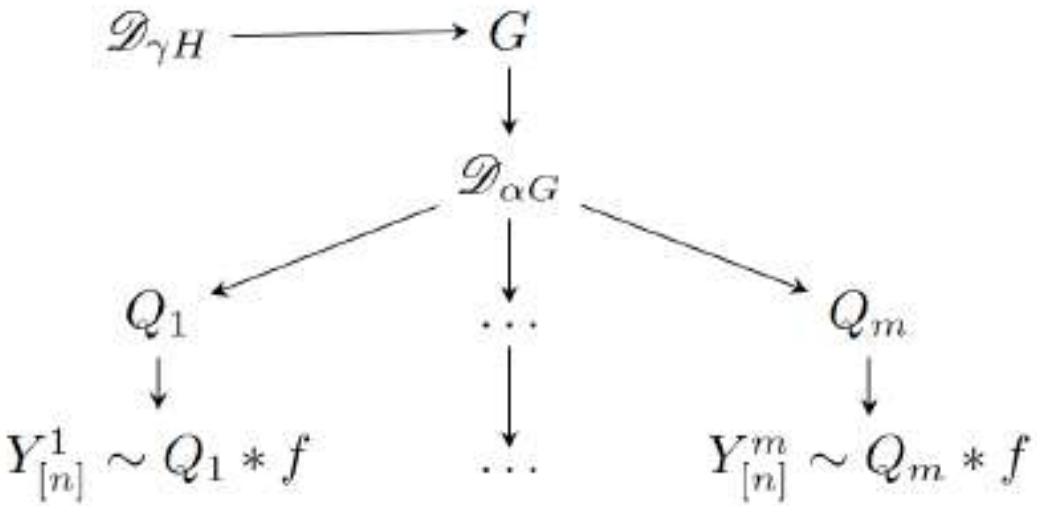
"Theorem"

- when the number of groups $m$ increases and the sample size $n$ increases suitably, the posterior distribution of $G$ contracts to true $G_0$ under the above Wasserstein metric
- individual group admits improved posterior contraction due to the "borrowing of information" from other groups

# Setting and assumptions



# Setting and assumptions




## Setting

- Let $\Theta$ be a bounded subset of $\mathbb{R}^d$.
- True Dirichlet base measure $G_0 \in \mathcal{P}(\Theta)$ is atomic.
- Given $m$ Dirichlet processes $Q_1, \ldots, Q_m$ drawn from $\mathscr{D}_{\alpha G}$, for $G = G_0$.
- For each process $Q_i$ there is an $n$-sample from the mixture dist $Q_i * f$

## Assumptions

(A1) For some $r \geq 1, C_1 > 0$, $h(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r$ and $K(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r \ \forall \theta, \theta' \in \Theta$.

(A2) There holds $M = \sup_{\theta,\theta' \in \Theta} \chi(f(\cdot|\theta), f(\cdot|\theta')) < \infty$.

(A3) $G$ is endowed with Dirichlet prior $\mathscr{D}_{\gamma H}$, where $H \in \mathcal{P}(\Theta)$ is non-atomic.

# Posterior concentration theorem (Nguyen, 2016)

As $n \to \infty$ and $m \to \infty$, the posterior distribution of Dirichlet base measure $G$ concentrates to $G_0$ at the rate

$$\epsilon_{m,n} \asymp \left( \frac{n^{3d} \log m}{m} \right)^{1/(2d+2)} + A(\delta_n).$$

# Posterior concentration theorem (Nguyen, 2016)

As $n \to \infty$ and $m \to \infty$, the posterior distribution of Dirichlet base measure $G$ concentrates to $G_0$ at the rate

$$\epsilon_{m,n} \asymp \left( \frac{n^{3d} \log m}{m} \right)^{1/(2d+2)} + A(\delta_n).$$

$\delta_n \to 0$ is the demixing rate — the rate of estimating mixing measure $Q$ from an $n$-sample of a mixture density $Q * f$ (obtaining this rate is the earlier focus)

# Posterior concentration theorem (Nguyen, 2016)

As $n \to \infty$ and $m \to \infty$, the posterior distribution of Dirichlet base measure $G$ concentrates to $G_0$ at the rate

$$\epsilon_{m,n} \asymp \left( \frac{n^{3d} \log m}{m} \right)^{1/(2d+2)} + A(\delta_n).$$

$\delta_n \to 0$ is the demixing rate — the rate of estimating mixing measure $Q$ from an $n$-sample of a mixture density $Q * f$ (obtaining this rate is the earlier focus)

function $A$ depends on the geometric structure of the support of $G_0$

(ii) If $f$ is supersmooth with parameter $\beta$, then it suffices to set

$$\frac{m}{\log m(\log n)^{\alpha^*(2d+2)/\beta}} \lesssim n^{3d} \lesssim \frac{m}{\log m}.$$

In particular, if $n$ satisfies $n^{3d}(\log n)^{\alpha^*(2d+2)/\beta} \asymp \frac{m}{\log m}$, then we obtain the concentration rate $\epsilon_{m,n} \asymp (\log n)^{-\alpha^*/\beta} \asymp (\log m)^{-\alpha^*/\beta}$.

(iii) Requirements of the type $n_1(m) \leq n \leq n_2(m)$ appear crucial in deriving posterior concentration rates in hierarchical models. It is an interesting open question to establish the concentration behavior (or the lack thereof) for the full range of $n$.

(i) if $G_0$ has a finite and unknown number of support points on a bounded subset of $\mathbb{R}^d$, then

$$A(\delta_n) \asymp \delta_n^{\alpha^*/(\alpha^*+1)}.$$

where $\alpha^* = \inf_{\theta \in \text{spt } G_0} \alpha G_0(\{\theta\})$,

(i) if $G_0$ has a finite and unknown number of support points on a bounded subset of $\mathbb{R}^d$, then

$$A(\delta_n) \asymp \delta_n^{\alpha^*/(\alpha^*+1)}.$$

where $\alpha^* = \inf_{\theta \in \text{spt } G_0} \alpha G_0(\{\theta\})$,

(ii) if $G_0$ has infinite and supersparse support on $\mathbb{R}^d$,

$$A(\delta_n) \asymp \exp -[\log(1/\delta_n)]^{1/(1 \vee \gamma_0 + \gamma_1)}$$

(iii) if $G_0$ has infinite and ordinary sparse support on $\mathbb{R}^d$,

$$A(\delta_n) \asymp [\log(1/\delta_n)]^{-1/(\gamma_0+\gamma_1)}$$

(i) if $G_0$ has a finite and unknown number of support points on a bounded subset of $\mathbb{R}^d$, then

$$A(\delta_n) \asymp \delta_n^{\alpha^*/(\alpha^*+1)}.$$

where $\alpha^* = \inf_{\theta \in \text{spt } G_0} \alpha G_0(\{\theta\})$,

(ii) if $G_0$ has infinite and supersparse support on $\mathbb{R}^d$,

$$A(\delta_n) \asymp \exp -[\log(1/\delta_n)]^{1/(1 \vee \gamma_0 + \gamma_1)}$$

(iii) if $G_0$ has infinite and ordinary sparse support on $\mathbb{R}^d$,

$$A(\delta_n) \asymp [\log(1/\delta_n)]^{-1/(\gamma_0+\gamma_1)}$$

(iv) **Finite admixtures:** if $G_0$ has $k < \infty$ support points, $k$ known, then we obtain a parametric rate:

$$\epsilon_{m,n} \asymp [\log(mn)/m]^{1/2} + [(\log n)^{1/2}/n^{1/4}]^{\alpha^*}.$$

# Geometric sparsity of support

- Sparse covering number: covering $\epsilon$-balls that are separated by $O(\epsilon)$ in distance
- $G_0$ is *supersparse* with non-negative parameters $(\gamma_0, \gamma_1)$, if its support admits *sparse*-covering number $K(\epsilon) \lesssim [\log(1/\epsilon)]^{\gamma_0}$, and the measure on such covering balls is at least $g(\epsilon) \gtrsim [\log(1/\epsilon)]^{-\gamma_1}$.
  - let $\Theta = [0, 2]$, $G_0$ is supported on $S = \{1/2^k | k \in \mathbb{N}, k \geq 1\} \cup \{0\}$, and $G_0(\{1/2^k\}) \propto k^{-\gamma_1}$ for any $k \in \mathbb{N}$ and some $\gamma_1 > 1$; then $G_0$ is a supersparse measure with parameters $\gamma_0 = 1$ and $\gamma_1$.

# Geometric sparsity of support

- Sparse covering number: covering $\epsilon$-balls that are separated by $O(\epsilon)$ in distance
- $G_0$ is *supersparse* with non-negative parameters $(\gamma_0, \gamma_1)$, if its support admits *sparse*-covering number $K(\epsilon) \lesssim [\log(1/\epsilon)]^{\gamma_0}$, and the measure on such covering balls is at least $g(\epsilon) \gtrsim [\log(1/\epsilon)]^{-\gamma_1}$.
  - let $\Theta = [0, 2]$, $G_0$ is supported on $S = \{1/2^k | k \in \mathbb{N}, k \geq 1\} \cup \{0\}$, and $G_0(\{1/2^k\}) \propto k^{-\gamma_1}$ for any $k \in \mathbb{N}$ and some $\gamma_1 > 1$; then $G_0$ is a supersparse measure with parameters $\gamma_0 = 1$ and $\gamma_1$.
- $G_0$ is *ordinary sparse* with parameters $(\gamma_0, \gamma_1)$ if $K(\epsilon) \lesssim (1/\epsilon)^{\gamma_0}$, and $g(\epsilon) \gtrsim \epsilon^{\gamma_1}$.

# Geometric sparsity of support

- Sparse covering number: covering $\epsilon$-balls that are separated by $O(\epsilon)$ in distance
- $G_0$ is *supersparse* with non-negative parameters $(\gamma_0, \gamma_1)$, if its support admits *sparse*-covering number $K(\epsilon) \lesssim [\log(1/\epsilon)]^{\gamma_0}$, and the measure on such covering balls is at least $g(\epsilon) \gtrsim [\log(1/\epsilon)]^{-\gamma_1}$.
  - let $\Theta = [0, 2]$, $G_0$ is supported on $S = \{1/2^k | k \in \mathbb{N}, k \geq 1\} \cup \{0\}$, and $G_0(\{1/2^k\}) \propto k^{-\gamma_1}$ for any $k \in \mathbb{N}$ and some $\gamma_1 > 1$; then $G_0$ is a supersparse measure with parameters $\gamma_0 = 1$ and $\gamma_1$.
- $G_0$ is *ordinary sparse* with parameters $(\gamma_0, \gamma_1)$ if $K(\epsilon) \lesssim (1/\epsilon)^{\gamma_0}$, and $g(\epsilon) \gtrsim \epsilon^{\gamma_1}$.
- Ordinary sparse measures are studied in fractal geometry: $\gamma_0$ is the Hausdorff dimension of the support, while $\gamma_1$ is the packing dimension.
  - if $\Theta = [0, 1]$, then the Hausdorff measure on the Cantor set is ordinary sparse with $\gamma_0 = \gamma_1 = \log 2 / \log 3$.

# Summarizing (this section)

- Optimal transport provides the first (and only so far) posterior contraction analysis of a hierarchical Bayesian model (Nguyen, Bernoulli 2016)
  - recent advances via harmonic analysis of mixture of product distributions (Wei & Nguyen, Annals 2022)

- It also connects to a notion of barycenter among a collection of measures (Agueh and Carlier, 2012), and gave rise to optimal transport based multi-level clustering methods (Ho et al, ICML 2017, Huynh et al, JMLR 2021)

- Full treatment of inverse bound for this general hierarchy remains elusive!
  - but there have been promising recent contributions from Nhat Ho, Aritra Guha, Yun Wei, Dat Do, Linh Do, Sunrit Chakraborty

  - for more details see my second lecture at VIASM on Thursday

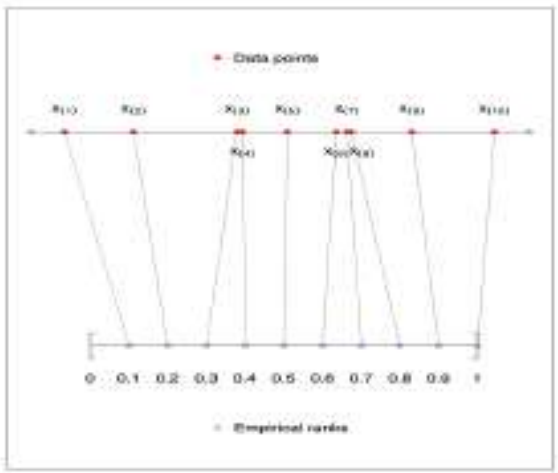
# Outline

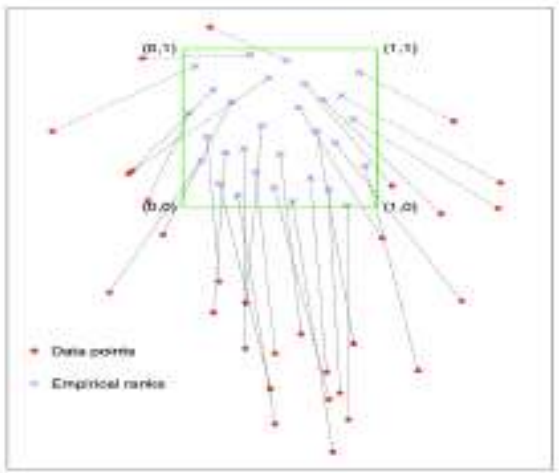
- Story so far: Optimal transport distance based inference on space of *distributions of data*, space of *distribution of quantities of interest*, distributions of *distributions*, and so on

- Other abstract notion of dependence can be quantified too:
  - multivariate rank and quantiles
  - independence
  - exchangeability, partial exchangebility

# Multivariate ranking

Given data sample $X_1, \ldots, X_n \in \mathbb{R}^d$, how to rank them?




Define empirical measure $\mu_n = \sum_{i=1}^n \delta_{X_i}$. Let $\nu_n$ be a discrete approximation of $\mathrm{Unif}[0,1]^d$ (supported on a regular lattice). Then, the empirical rank function $F$ is one which solves

$$F = \mathrm{argmin}_F \int \|x - F(x)\|^2 d\mu_n$$

such that $F_{\#}\mu_n = \nu_n$.

Deb and Sen (2019)

# Independence

Given observrvations of two random variables $X$ and $Y$ on Polish spaces $\mathcal{X}, \mathcal{Y}$, resp. Are $X$ and $Y$ independent or not?

Let $X \sim \mu$; $Y \sim \nu$; and $(X, Y) \sim \gamma$

- Shannon's mutual information:
  $I(X, Y) = K(\gamma, \mu \otimes \nu) = \int \log(d\gamma / d(\mu \otimes \nu)) d\gamma$ provided $\gamma \ll \mu \otimes \nu$ (and $I(X, Y) = +\infty$ otherwise).
  Clearly $I(X, Y) = 0$ if and only if $X$ and $Y$ are independent.
- OT based dependence: (Nies et al, 2021; Wiesel, 2021)

$$\tau(X, Y) := T_c(\gamma, \mu \otimes \nu) = \inf_{\pi \in \Pi(\gamma, \mu \otimes \nu)} \int c((x, y), (x', y')) d\pi$$

$$\tau^Y(X, Y) := \int T_{c_Y}(\gamma_x, \nu) \mu(dx)$$

where $c, c_Y$ is some notion of cost on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{Y}$, resp. $\gamma_x$ denotes the conditional distribution of $Y$ given $X = x$ under the joint distribution $\gamma$ (i.e., disintegration of $\gamma$ wrt $x$)

Example: $\xi \sim \mathrm{Unif}[0,1]$ and $\zeta = f_n(\xi) \in \mathrm{Unif}[0,1]$ for zigzag functions $f_n$ with $n$ linear segments. $n = 1$ in (a), $n = 8$ in (b).

Note that $I(\xi, \zeta) = \infty$ in both cases for it fails to account for the metric structure of the support.

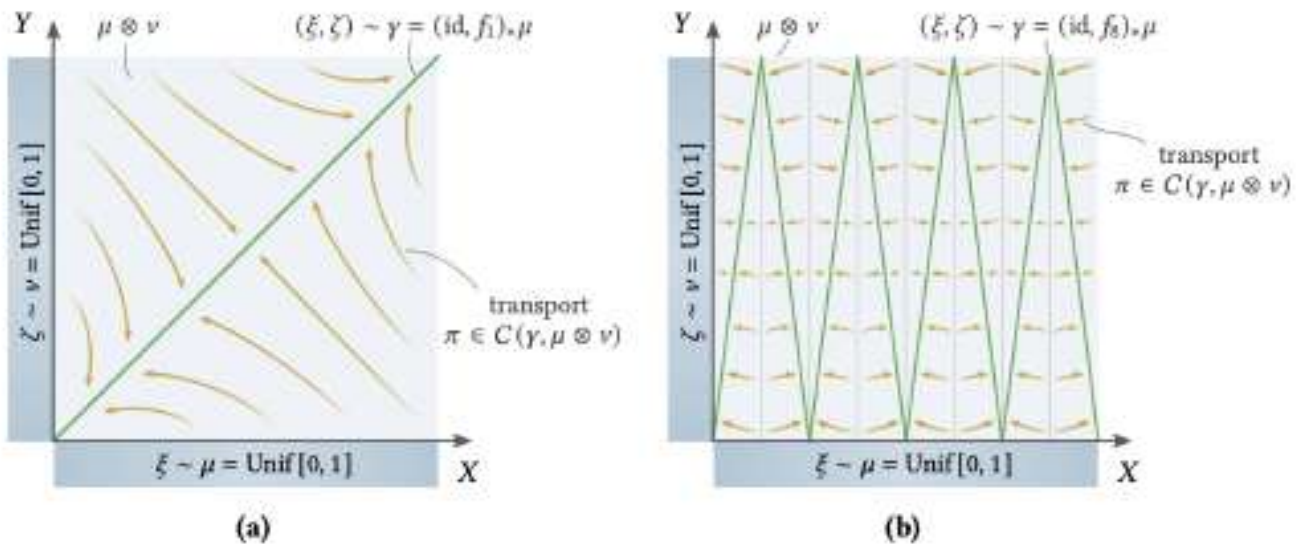



Illustration of Nies et al (2021)

# Distance to exchangeability

Given two exchangeable populations

$$X_1, \ldots, X_n | P_1 \overset{iid}{\sim} P_1; \quad Y_1, \ldots, Y_n | P_1 \overset{iid}{\sim} P_2$$

Can we tell whether $P_1 \perp P_2$, or $P_1 = P_2$ (almost surely)?

Catalano et al (2021):

- assume $(P_1, P_2) = (T(\mu_1), T(\mu_2))$, where $\mu_1 \overset{d}{=} \mu_2$
  - $(\mu_1, \mu_2)$ random measures with joint independent increments
  - $T$ is a generic map (e.g., normalization, exponential, kernel mixtures)
  - the above assumption covers a large range of model for two-sample problems
- the problem can be formulated as measuring the distance between complete random measures, which in turn boils down to the distance of the corresponding Lévy measures

# Summary

- OT is a useful device for quantifying the space of distributions of data, distribution of latent quantities of interest, distributions of distributions, abstract notions of dependence
    - pushforward measures may be used as a useful way for modeling distributions, or to characterize convergence behavior of algorithms
- many open questions surrounding the mathematics behind the presence of latent-variable representation of complex models, and the statistical and computational theory for achieving and explicating the emergence of such representation
    - OT framework provides a promising approach

# Selected References


- J. Zhu, A. Guha, D. Do, M. Xu, X. Nguyen and D. Zhao. Functional optimal transport: map estimation and domain adaptation for functional data. arXiv:2102.03895.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. Annals of Statistics, 2013.
- X. Nguyen. Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. Bernoulli, 2016.
- Y. Wei and X. Nguyen. Convergence of de Finetti's mixing measures in latent structure models for exchangeable sequences. arXiv:2004.05542. Annals of Statistics, to appear.
- D. Do, N. Ho and X. Nguyen. Beyond black box densities: parameter learning for the deviated components. arXiv:2202.02651.