Optimal Transport for Transfer Learning and Algorithmic Fairness Problems Arising in High-Energy Physics

Mikael Kuusela

Department of Statistics and Data Science,
Carnegie Mellon University

Institute of Advanced Study,
University of Geneva,
Geneva, Switzerland

October 4, 2023

Joint work with: Tudor Manole, Patrick Bryant, John Alison,
Purvasha Chakravarti and Larry Wasserman

Special thanks to Larry for contributing some of the following materials!

Hypothesis testing for discovery of new physics

Search for new phenomena at the LHC usually boils down to testing for the presence of a signal distribution over a background of known physics:

- Known physics: $p_b(x)$
- New signal: $p_s(x)$
- Nature: $q(x) = (1 - \lambda)p_b(x) + \lambda p_s(x)$

Want to test $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$

If we reject H_0 at high enough significance level, then we could proceed to claim discovery of new physics

Classifier-based tests

Over the past 20 years or so, the high-energy physics community has developed an impressive statistical machinery for performing these tests

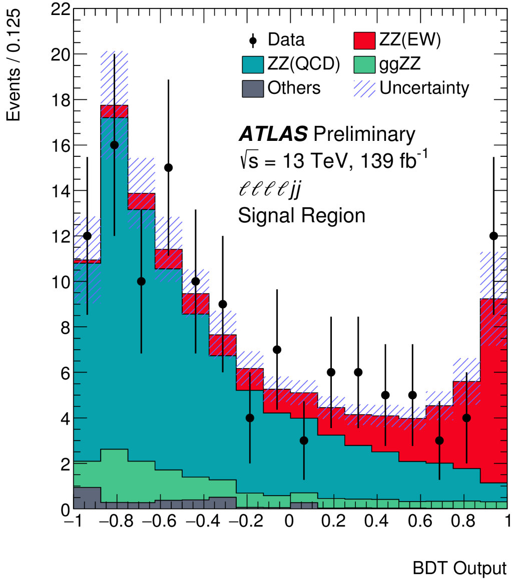
Relevant datasets:

$$
\begin{aligned}
\text{Training background:} \quad & \mathcal{X} = \{X_1, \ldots, X_{m_b}\}, & X_i \sim p_b \\
\text{Training signal:} \quad & \mathcal{Y} = \{Y_1, \ldots, Y_{m_s}\}, & Y_i \sim p_s \\
\text{Experimental data:} \quad & \mathcal{W} = \{W_1, \ldots, W_n\}, & W_i \sim q = (1-\lambda)p_b + \lambda p_s
\end{aligned}
$$

Basic idea:

1. Train a supervised classifier to separate $\mathcal{X}$ from $\mathcal{Y}$
2. Use the classifier output to test for the presence of signal in $\mathcal{W}$

Classifier output



Several options for the test:

- Counting experiment in the highest purity output bin
- Cut on the classifier output; test using the resulting signal-enriched sample
- LRT: Use the connection of the classifier output to the likelihood ratio
- ...

Problem 1: Data-Driven Di-Higgs Background Modeling

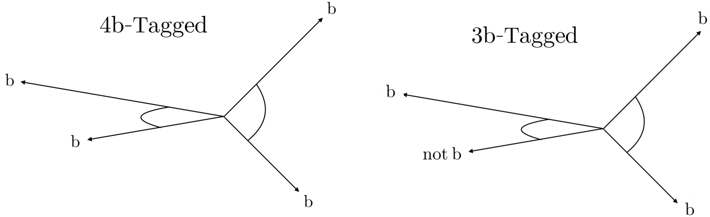
Two similar distributions P_{3b} and P_{4b} over a 16-dimensional space

Sample space $\mathcal{X} = C \bigcup S$, C = control region, S = signal region, $C \cap S = \emptyset$

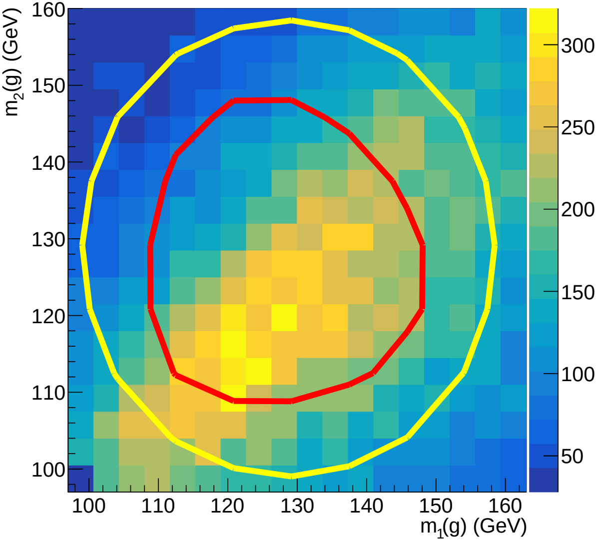
Given: a sample $X_1, \ldots, X_n \sim P_{3b}$ and a sample $Y_1, \ldots, Y_m \sim P_{4b}(\cdot|C)$

Goal: estimate $P_{4b}(\cdot|S)$

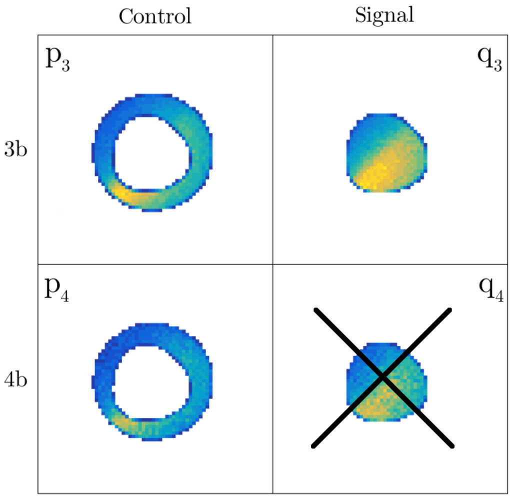
The problem is ill-posed; we will have to make (reasonable) assumptions.



Control and Signal Regions

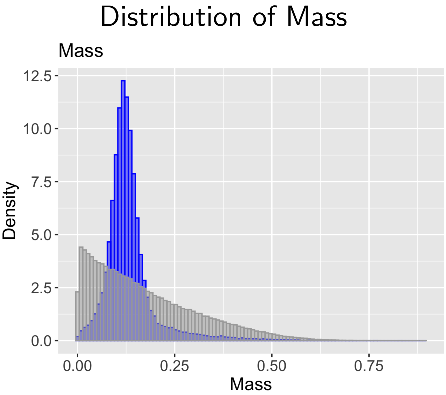


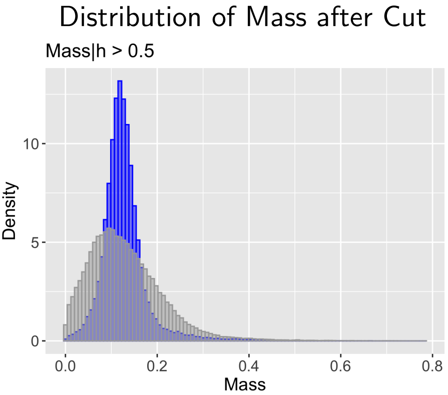
Control and Signal Regions, 3b vs. 4b



This is a transfer learning problem either in the vertical or in the horizontal direction

Problem 2: Decorrelating signal vs. background classifiers





Mass is a *protected variable*

$\rightarrow$ This is an example of an algorithmic fairness problem

Introduction: What is Optimal Transport?

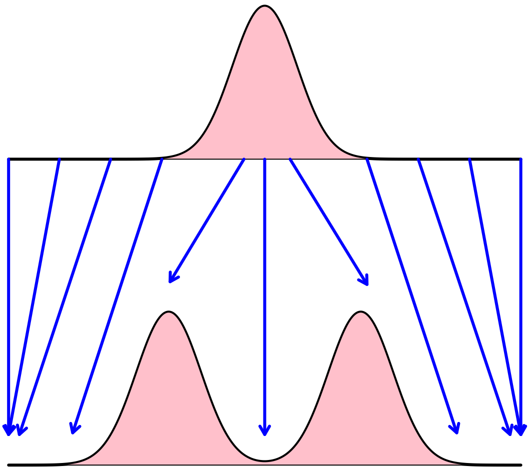
We have two probability distributions P_0 and P_1.

Goal: Define an "optimal map" that transforms P_0 into P_1.

This enables us to:

- Define a distance based on transport (Wasserstein distance)
- Define a path (geodesic) between P_0 and P_1 in the space of distributions (morphing)
- Define a shape-preserving notion of "averages" of distributions (barycenter)

Optimal Transport (Monge 1781)



Optimal Transport (Monge Version)

Let $X \sim P_0$.

Find T to minimize

$$\mathbb{E}\Big[||X - T(X)||^p\Big] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.

Can replace the Euclidean distance $\|\cdot\|$ with any valid distance metric.

For now, assume that the minimizer exists. Then the minimizer T^* is called the optimal transport map from P_0 to P_1.

Common choices: $p = 2$ or $p = 1$.

Wasserstein distance

The pth Wasserstein distance between P_0 and P_1 is defined as:

$$W_p(P_0, P_1) = \left(\int ||x - T^*(x)||^p dP_0(x) \right)^{1/p}$$

where T^* is the optimal transport map.

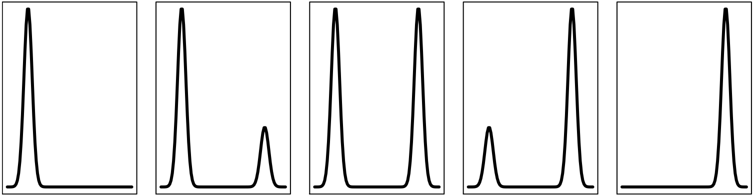
Defines a metric on the space of (nearly) all distributions.

W_1 is called the Earth Mover's Distance

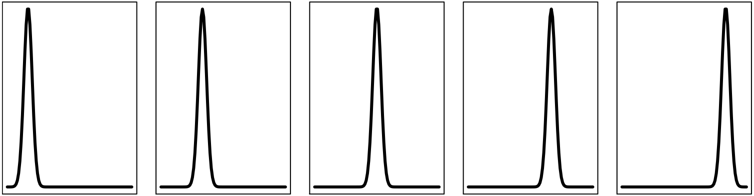
Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W_p is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1, there is a shortest path (geodesic) between them.
- For $0 \leq s \leq 1$, let P_s be the distribution of the random variable $(1 - s)X + sT(X)$ where $X \sim P_0$.
- Then $(P_s : 0 \leq s \leq 1)$ is the desired geodesic.
  Length of the path $= W_p(P_0, P_1)$.

Euclidean Path between Two Gaussians



Geodesic Path between Two Gaussians



Geodesic Path between Two Mixtures

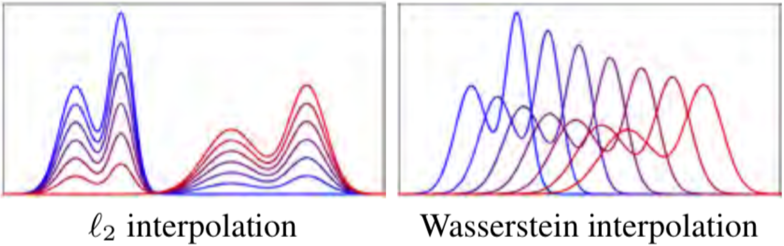


Image credit: Bonneel, Peyre and Cuturi (2016)

Geodesic Path Between Two Images

Image credit: Bauer, Joshi and Modin (2015)

Barycenters

Given $P_1, \ldots, P_N$, what is the "average" of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

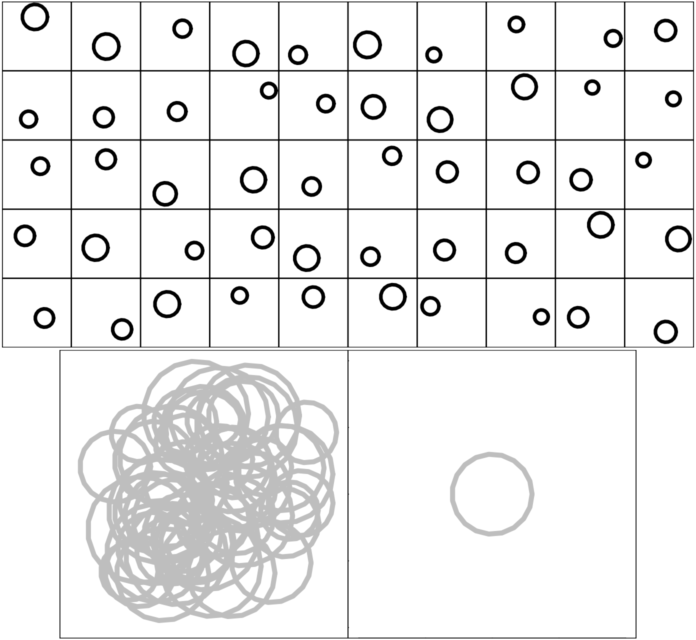
Same problem as before: this does not necessarily look like any of the P_j's.

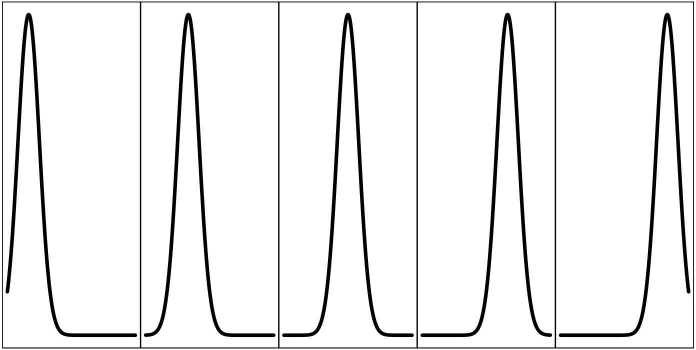
Wasserstein barycenter: find P to minimize

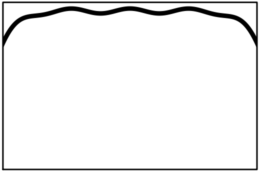
$$\sum_j W_2^2(P, P_j).$$

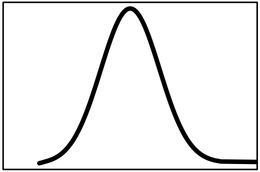
This is the barycenter and it is shape preserving.

Weighted version of this gives us the ability to morph between multiple distributions.

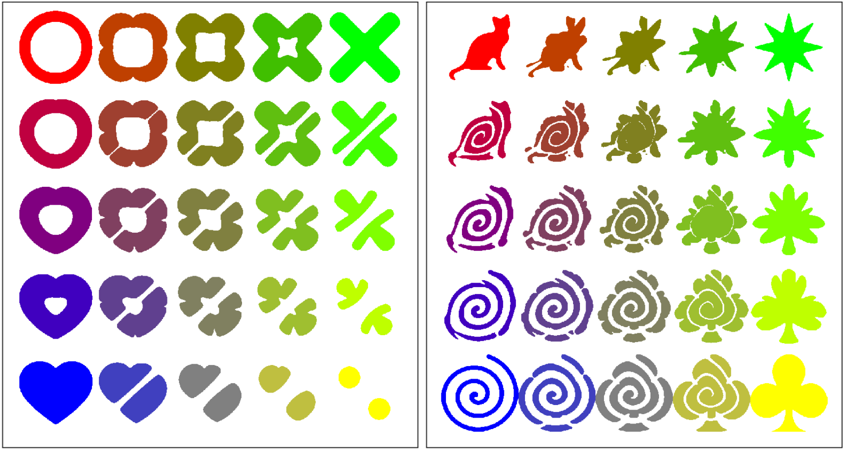








Example from Peyre and Cuturi (2019)



Optimal Transport (Kantorovich Version)

An important detail that we have ignored so far:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$ Gaussian.

Solution: Kantoravich relaxation

Take mass at x, and split it into many small pieces.

Optimal Transport (Kantorovich Version)

Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.

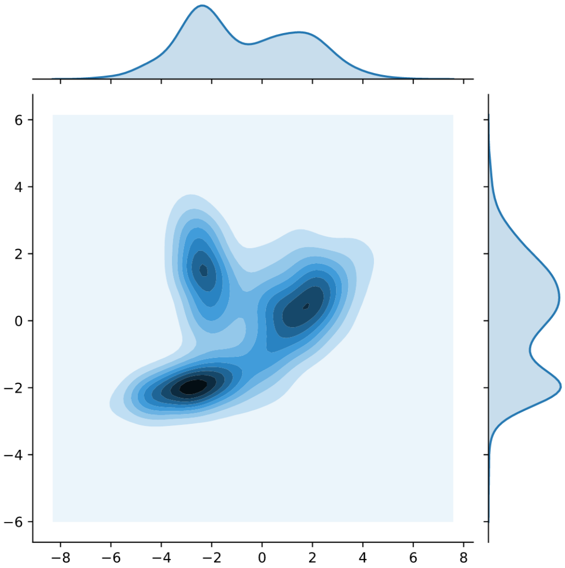
Find J (optimal transport plan / optimal coupling) to minimize

$$\mathbb{E}_J[||X - Y||^p] = \int ||x - y||^p \, dJ(x, y)$$

Again, this defines a distance

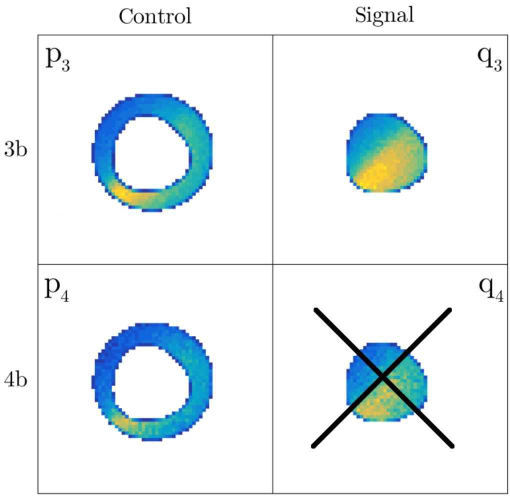
$$W_p(P, Q) = \left(\inf_J \int ||x - y||^p dJ(x, y) \right)^{1/p}$$

called the Wasserstein distance, as before.



Joint distribution J with a given X marginal and a given Y marginal.
Image credit: Wikipedia.

Control and Signal Regions, 3b vs. 4b



This is a transfer learning problem either in the vertical or in the horizontal direction

Three Methods

1. Density ratio

Estimate $\frac{p_{3b}(x)}{p_{4b}(x)}$ over C using a classifier and apply out-of-sample in S.

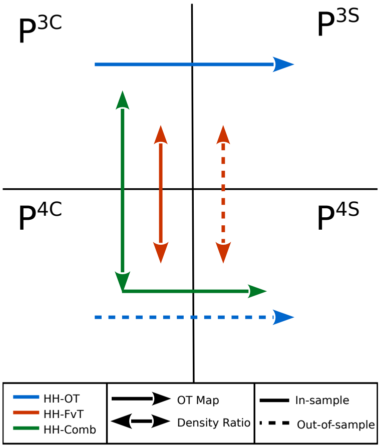
2. Optimal transport

Use P_{3b} to find a map T that optimally transports mass from C to S. Apply the map to P_{4b} using a nearest-neighbor look-up in C.

3. Combination

Use the classifier to reweight P_{3b} to look like P_{4b} in C. Then apply T to the weighted sample to transport C to S.

EMD used as the ground metric when computing T (double optimal transport)



Energy Mover's Distance (EMD)

Proposed by Komiske, Metodiev and Thaler (2019).

A jet is described by (p_T, η, ϕ, m), where p_T = transverse momentum, η = pseudorapidity, ϕ = azimuthal angle and m = mass.

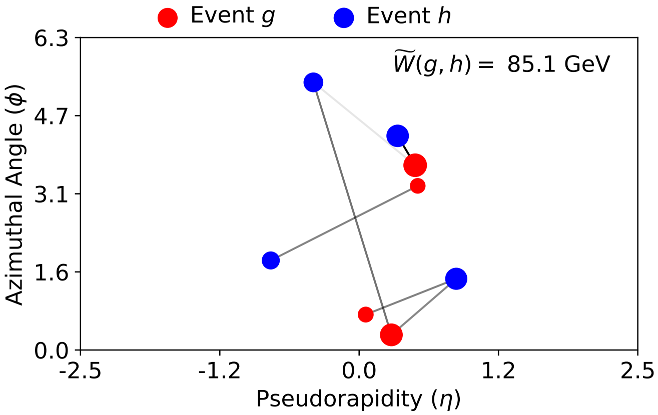
In our case, an event $\mathcal{E}$ contains 4 jets. We treat it as a measure:

$$\mathcal{E} = \sum_{i=1}^{4} p_{T,i}\delta_i,$$

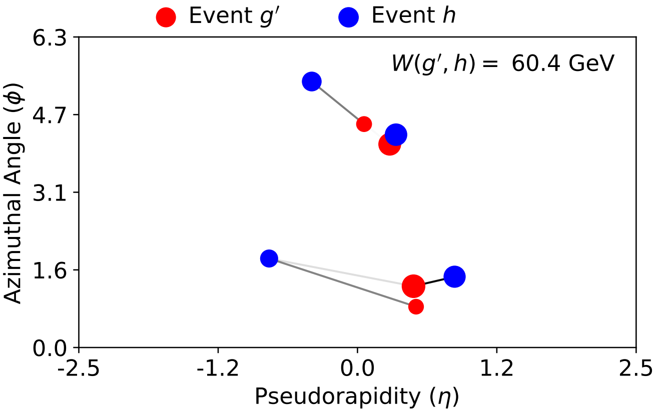
where δ_i is a point mass at (η_i, ϕ_i, m_i).

The Energy Mover's Distance (EMD) between two events $\mathcal{E}_1$ and $\mathcal{E}_2$ is defined as the (modified) Wasserstein distance between these two measures.

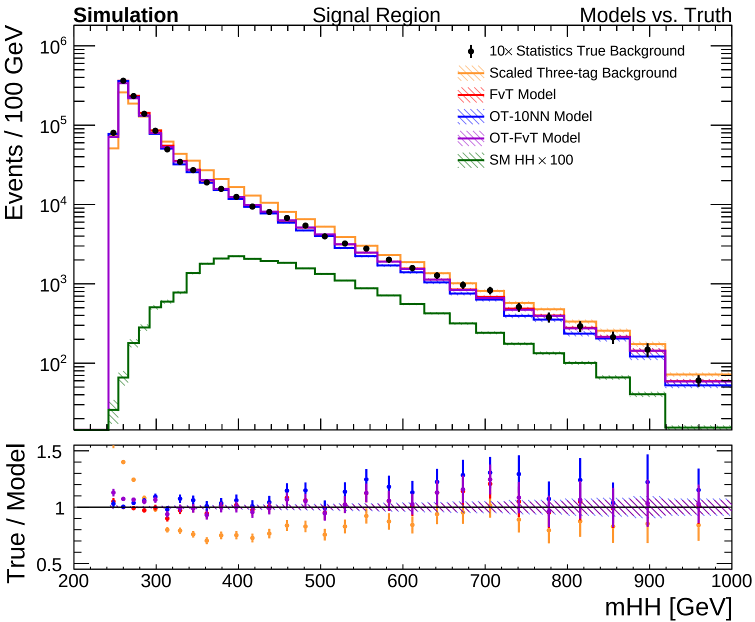
Energy Mover's Distance (EMD)



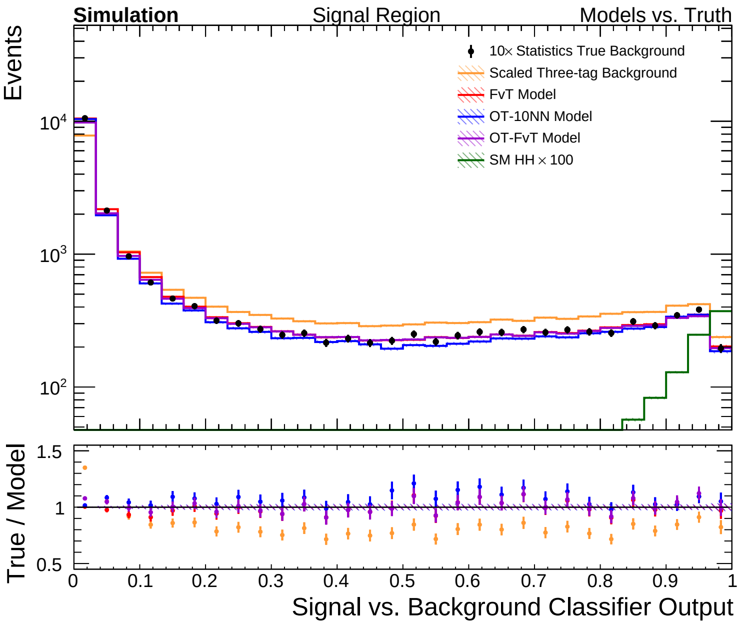
Energy Mover's Distance (EMD)



Results: m_{HH}



Results: Signal-versus-background classifier output



Results: Classifier AUCs

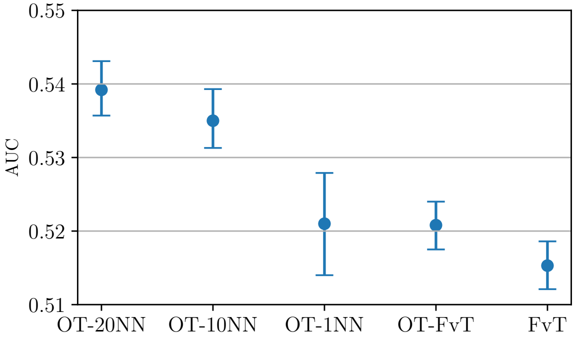


Figure: AUCs for a classifier trained to separate the background models from the actual $4b$ background sample. For $3b$-tagged data, we obtain AUC 0.5843, with variability interval [0.5812,0.5874].

For more information, see: T. Manole, P. Bryant, J. Alison, M. Kuusela, and L. Wasserman. Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport. arXiv:2208.02807, 2022.

Optimal transport for decorrelation

Setting: features X, protected variable $m(X)$ (e.g., invariant mass) on the background data

Problem: classifier h trained to separate signal from background based on X will not preserve the distribution of $m(X)$

Idea: train h as usual, then apply optimal transport to map $h(X)$ so that $T(h(X))$ is independent of $m(X)$ on the background data

Optimal transport for decorrelation

- Objective: $\min_T (T(h(X)) - h(X))^2$ subject to $T(h(X))$ independent of $M = m(X)$, given $X \sim p_b$ (i.e., $T(h) \perp\!\!\!\perp M | X \sim p_b$)

- That is, we want

$$P(T(h(X)) \leq t | M, X \sim p_b) = P(T(h(X)) \leq t | X \sim p_b)$$

(i.e., $T(h)|M \stackrel{d}{=} T(h)|X \sim p_b$)

- Additionally, we want

$$P(T(h(X)) \leq t | X \sim p_b) = P(h(X) \leq t | X \sim p_b)$$

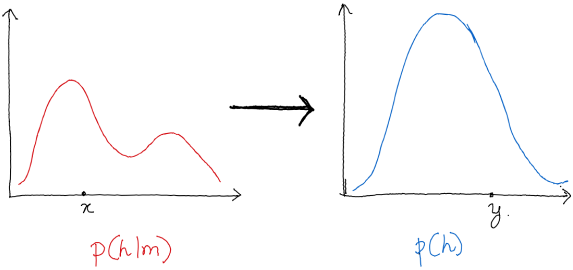
(i.e., $T(h) \stackrel{d}{=} h | X \sim p_b$)

$$\arg\min_T (T(h(X)) - h(X))^2 \;\; s.t. \;\; T(h)|M \stackrel{d}{=} h|X \sim p_b$$

Optimal transport for decorrelation

$$\arg\min_{T} \left(T(h(X)) - h(X)\right)^2 \;\; s.t. \;\; T(h)|M \overset{d}{=} h|X \sim p_b$$

Solution: the conditional optimal transport map T_a from $p(h(X)|M = a, X \sim p_b)$ to the marginal $p(h(X)|X \sim p_b)$.



Optimal transport for decorrelation

$h(X)$ is univariate so there exists a closed form solution to optimal transport problem:

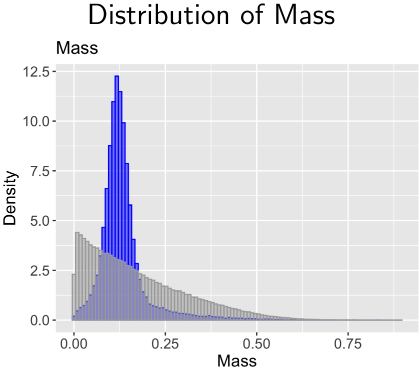
$$T_a(h(X)) = G^{-1}(F_{h|M}(h(X)|M = a)),$$

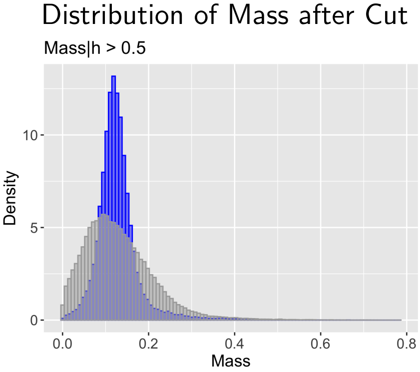
where G is the marginal cdf of $h(X)$ and $F_{h|M}$ is the conditional distribution of $h(X)$ given $m(X) = a$ and X is from the background distribution

Solution is found by estimating G and $F_{h|M}$

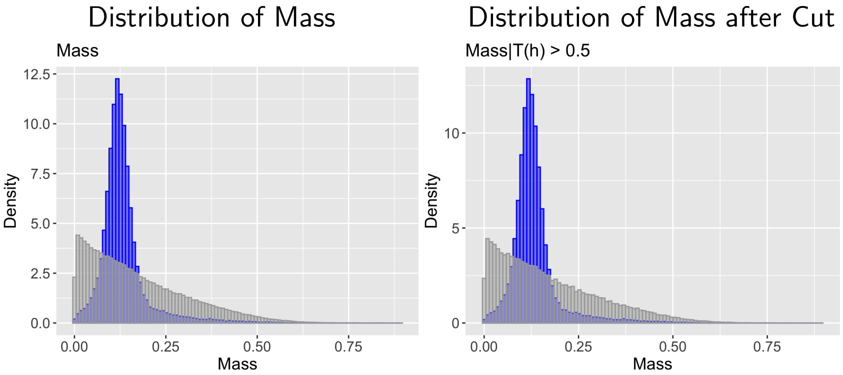
We call this CDOT (Classifier Decorrelated through Optimal Transport)

Sculpting problem solved!

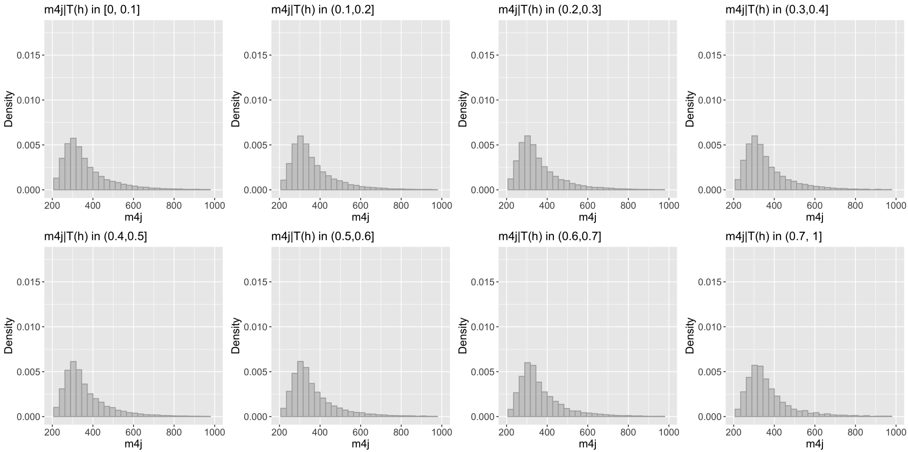




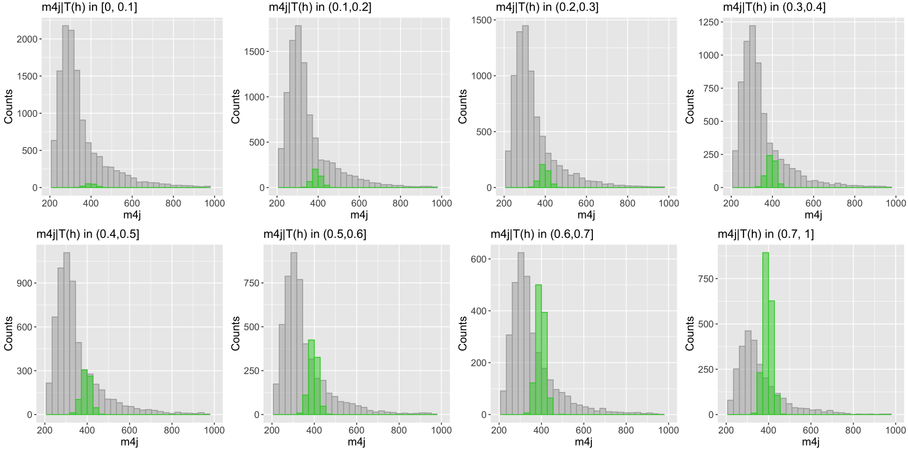
Sculpting problem solved!



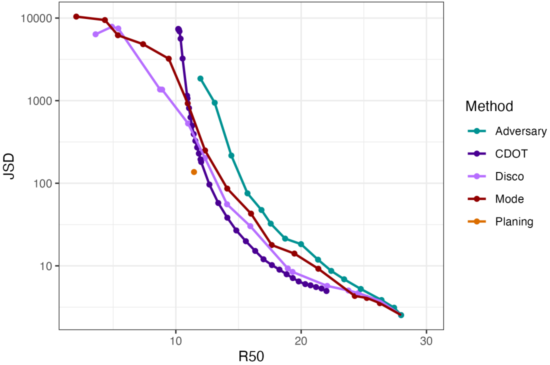
Optimal transport for decorrelation



Optimal transport for decorrelation



WTagging dataset: comparison



CDOT achieves superior signal-to-background ratio for strongly decorrelated classifiers.

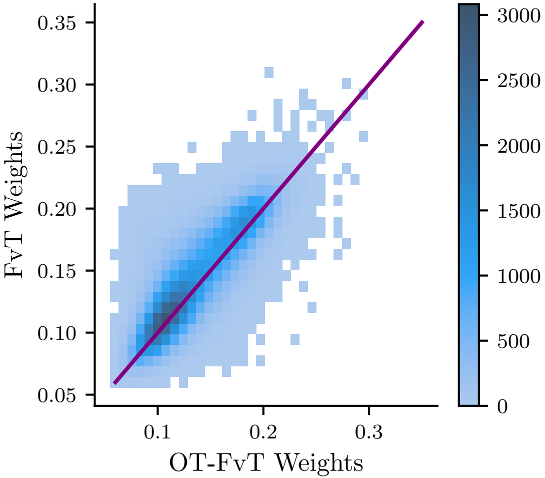
Original figure without CDOT taken from the MoDe [Kitouni et al. (2010.09745)] paper.

Conclusions

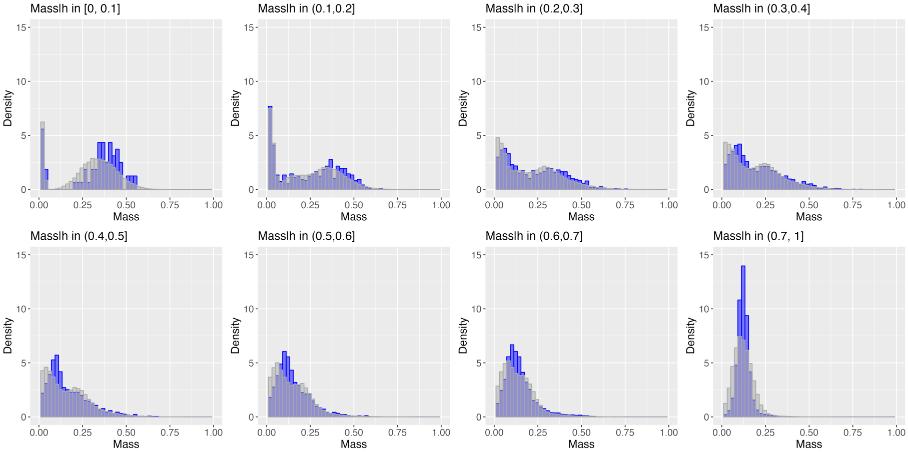
- Optimal transport provides an appealing tool for morphing between distributions, measuring the distance between distributions and computing averages of distributions
- Well-established mathematical theory; surge of interest in statistics / data science / machine learning in recent years; increasing interest in HEP as well
- We have found optimal transport to be a useful tool for solving background estimation (transfer learning) and decorrelation (algorithmic fairness) problems in HEP
- Many other possible applications of optimal transport in HEP and beyond

Backup

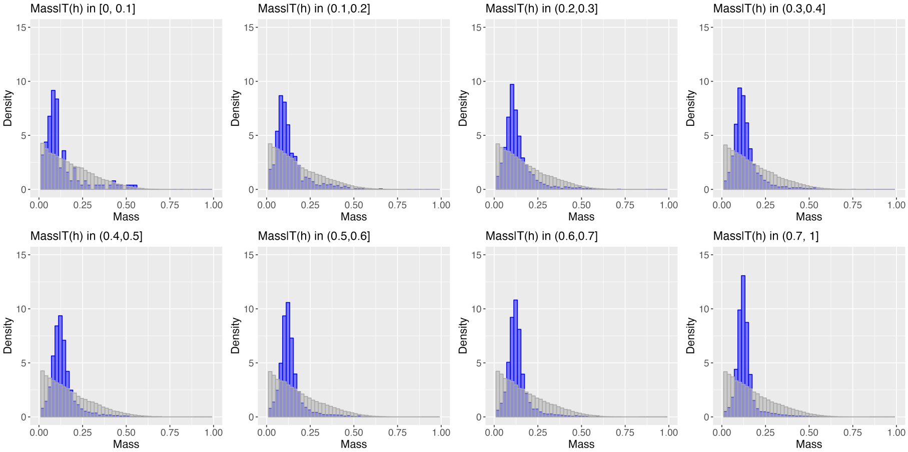
Results: Classifier weights vs. OT weights



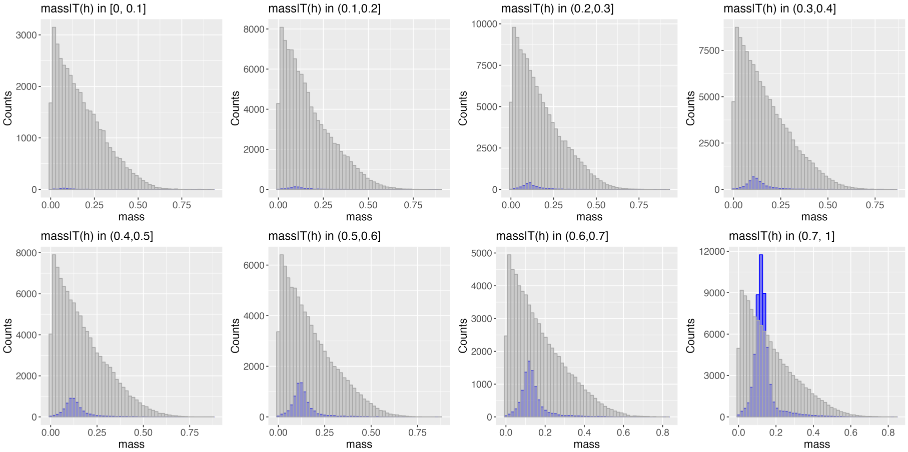
WTagging dataset: before OT transformation



WTagging dataset: after OT transformation



WTagging dataset: after OT transformation



Finding the Transport Map: One-Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$
- $W_p(P_0, P_1) = \big(\int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds\big)^{1/p}$
- The morphing — geodesic linking F_0 and F_1 — is

$$F_s = \left[(1-s)F_0^{-1} + sF_1^{-1}\right]^{-1}$$

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

$$\widehat{F}_0(s) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq s)$$

$$\widehat{F}_1(s) = \frac{1}{m} \sum_{i=1}^{m} I(Y_i \leq s)$$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$, $Y \sim N(\mu_1, \Sigma_1)$

Then:

$$T(x) = \mu_1 + \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}(x - \mu_0)$$

$$W_2^2(P_0, P_1) = ||\mu_0 - \mu_1||^2 + B(\Sigma_0, \Sigma_1)^2$$

where

$$B(\Sigma_0, \Sigma_1) = \mathrm{trace}(\Sigma_0) + \mathrm{trace}(\Sigma_1) - 2\mathrm{trace}\big[(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\big].$$

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\}, \quad X_i \in \mathbb{R}^d$
- $\mathcal{Y} = \{Y_1, \ldots, Y_n\}, \quad Y_i \in \mathbb{R}^d$
- $T : X_i \to Y_{\pi(i)}$, where the permutation π minimizes

$$\sum_i ||X_i - Y_{\pi(i)}||^2$$

  over all permutations π.
- Hungarian algorithm: $O(n^3)$ computing time

OT in HEP: Horizontal Morphing

Linear interpolation of histograms

A.L. Read[1]

University of Oslo, Department of Physics, P.O. Box 1048, Blindern, 0316 Oslo, Norway

Received 19 October 1998

Abstract

A prescription is defined for the interpolation of probability distributions that are assumed to have a linear dependence on a parameter of the distributions. The distributions may be in the form of continuous functions or histograms. The prescription is based on the weighted mean of the inverses of the cumulative distributions between which the interpolation is made. The result is particularly elegant for a certain class of distributions, including the normal and exponential distributions, and is useful for the interpolation of Monte Carlo simulation results which are time-consuming to obtain.

The p.d.f. $\bar{f}(x)$ is obtained by inverting the cumulative distributions in Eqs. (4) and (5), substituting these results in Eq. (6),

$$\bar{F}^{-1}(y) = aF_1^{-1}(y) + bF_2^{-1}(y), \tag{7}$$

deriving this with respect to y and solving for the interpolated p.d.f. $\bar{f}(x)$,

$$\bar{f}(x) = \frac{f_1(x_1)f_2(x_2)}{af_2(x_2) + bf_1(x_1)}. \tag{8}$$

This is the Wasserstein geodesic between 1D distributions!

Density Ratios and Classifiers

In general, given two densities p and q and samples

$$X_1, \ldots, X_n \sim p$$

$$Y_1, \ldots, Y_n \sim q$$

Give labels:

$$\begin{array}{c|cccccc} & X_1 & \ldots & X_n & Y_1 & \ldots & Y_n \\ Z & 1 & \ldots & 1 & 0 & \ldots & 0 \end{array}$$

Classifier ψ:

$$\psi(u) = P(Z = 1|u) = \frac{p}{p+q}$$

and so

$$\frac{p}{q} = \frac{\psi}{1-\psi}.$$