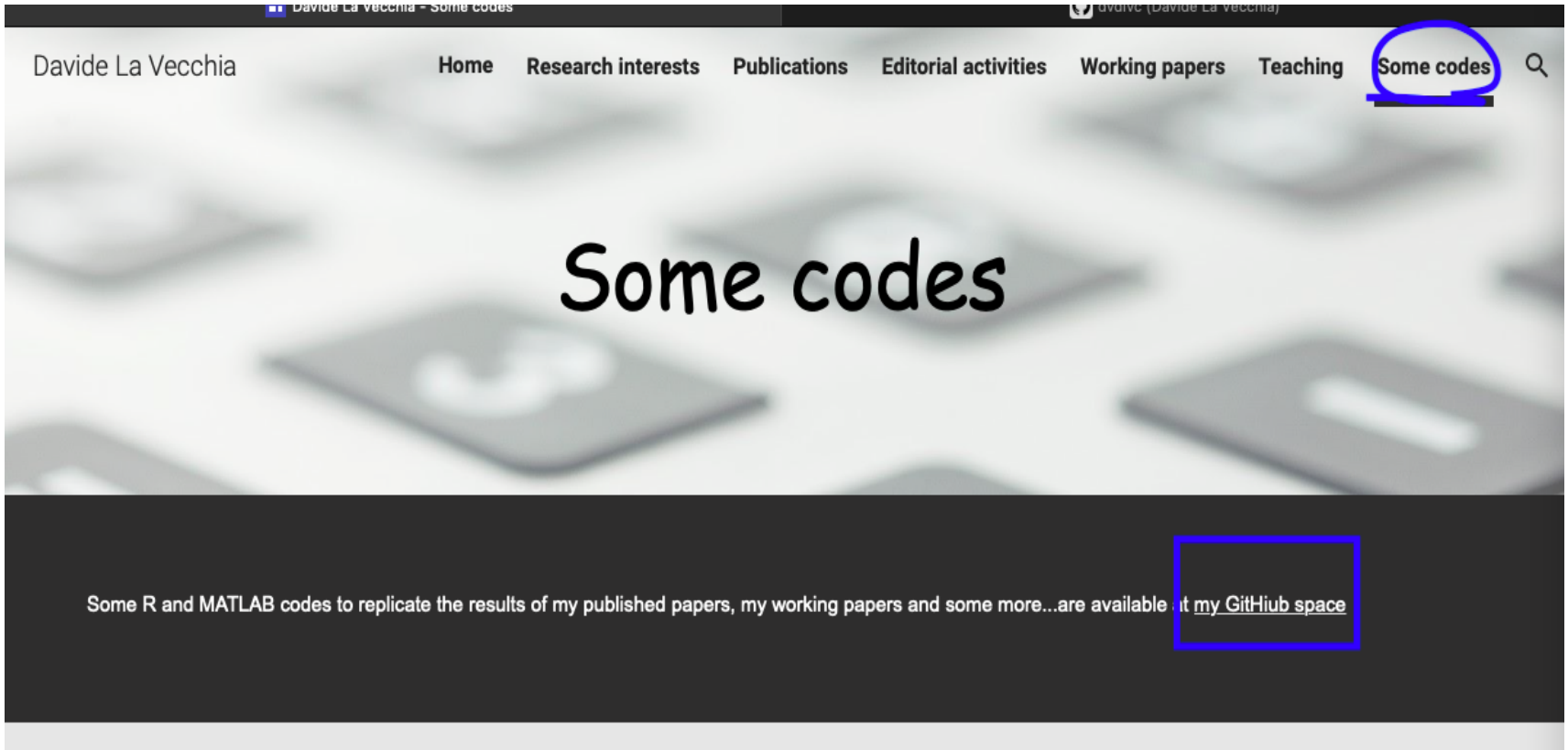
ROBOT: Statistics, Machine Learning, and generative Artificial Intelligence

Davide La Vecchia
Research Institute for Statistics and Information Science

Torino, Aug 2025

Codes (Python, Matlab and R) related to most of the examples in this talk are available on my GitHub:



An historical perspective

Looking at the issue of finding the best way to move given piles of sand to fill up given holes of the same total volume, **Gaspard Monge** (1746-1818) formulated a **mathematical problem** that in modern jargon reads as:

An historical perspective

Looking at the issue of finding the best way to move given piles of sand to fill up given holes of the same total volume, **Gaspard Monge** (1746-1818) formulated a **mathematical problem** that in modern jargon reads as:

Let α and β denote two probability measures over (for simplicity) $(\mathbb{R}^d, \mathcal{B}^d)$, for $d \geq 1$. Let $c : \mathbb{R}^{2d} \to \mathbb{R}$ be a Borel-measurable cost function such that $c(\mathbf{x}, \mathbf{y})$ represents the cost of transporting $\mathbf{x}$ to $\mathbf{y}$. Then, find a measurable transport map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves

$$\inf_{\mathcal{T} \in M} \int_{\mathbb{R}^d} c[\mathbf{x}, \mathcal{T}(\mathbf{x})] \, \mathrm{d}\alpha \tag{1}$$

where

$$M := \{\mathcal{T} : \mathbf{X} \to \mathbf{Y}\},$$

with $\mathbf{X} \sim \alpha$, $\mathbf{Y} \sim \beta$. The map $\mathcal{T}\#\alpha = \beta$ does the push forward of α to β.

An historical perspective

Looking at the issue of finding the best way to move given piles of sand to fill up given holes of the same total volume, **Gaspard Monge** (1746-1818) formulated a **mathematical problem** that in modern jargon reads as:

Let α and β denote two probability measures over (for simplicity) $(\mathbb{R}^d, \mathcal{B}^d)$, for $d \geq 1$. Let $c : \mathbb{R}^{2d} \to \mathbb{R}$ be a Borel-measurable cost function such that $c(\mathbf{x}, \mathbf{y})$ represents the cost of transporting $\mathbf{x}$ to $\mathbf{y}$. Then, find a measurable transport map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves

$$\inf_{\mathcal{T} \in M} \int_{\mathbb{R}^d} c[\mathbf{x}, \mathcal{T}(\mathbf{x})] \, \mathrm{d}\alpha \tag{1}$$

where

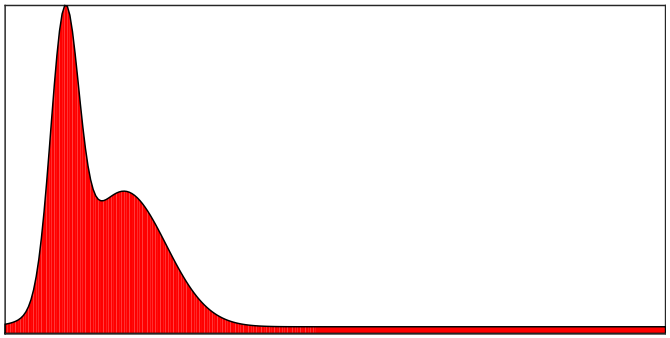
$$M := \{\mathcal{T} : \mathbf{X} \to \mathbf{Y}\},$$

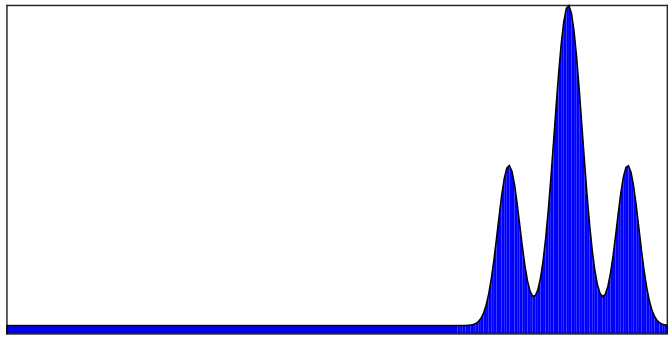
with $\mathbf{X} \sim \alpha$, $\mathbf{Y} \sim \beta$. The map $\mathcal{T}\#\alpha = \beta$ does the push forward of α to β.

$\Rightarrow$ The map solution to (1) is called the optimal transportation map.

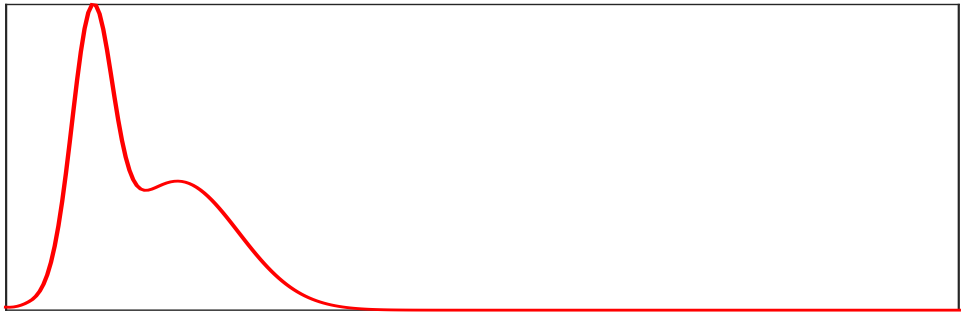
An historical perspective

To visualize...





We need to find the best (i.e. minimizing the cost given by the Euclidean distance) $\mathcal{T}$ that transform the red into blue:



An historical perspective

Monge's problem remained open until the 1940s, when it was revisited by **Leonid Vitaliyevitch Kantorovich** (1912-1986; Nobel Prize in Economics in 1975) for the economic problem of optimal allocation of resources; see e.g. Villani (2008), Santambrogio (2015), Galichon (2016).

An historical perspective

Monge's problem remained open until the 1940s, when it was revisited by **Leonid Vitaliyevitch Kantorovich** (1912-1986; Nobel Prize in Economics in 1975) for the economic problem of optimal allocation of resources; see e.g. Villani (2008), Santambrogio (2015), Galichon (2016).

In the Kantorovich primal problem, the objective is to find the optimal transportation plan γ, which solves

$$\inf_{\gamma \in \Gamma(\alpha,\beta)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\gamma(\mathbf{x}, \mathbf{y}), \tag{2}$$

where the infimum is over all coupling $(\mathbf{X}, \mathbf{Y})$ of (α, β), belonging to $\Gamma(\alpha, \beta)$, the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$, satisfying

$$\gamma(A \times \mathbb{R}^d) = \alpha(A) \text{ and } \gamma(\mathbb{R}^d \times B) = \beta(B),$$

for measurable sets $A, B \subset \mathbb{R}^d$.

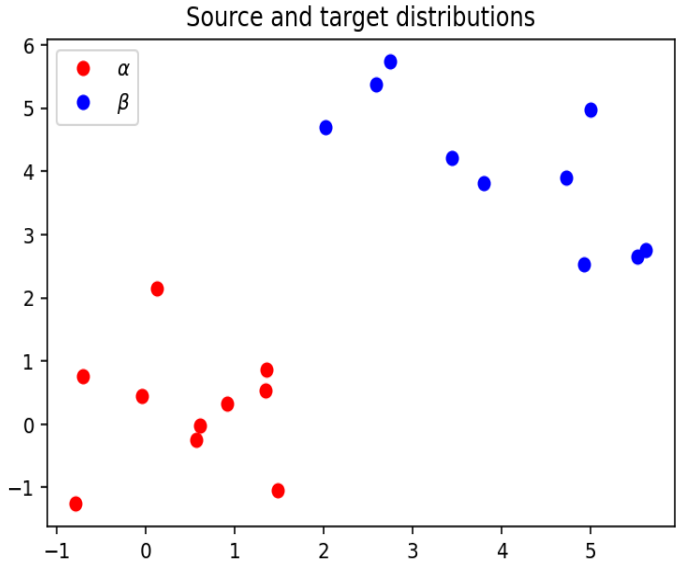
A visual intuition

- Solving the optimal transport problem (2) with $c = d^p$, so called **ground distance**, introduces (it lifts the ground distance) a distance between α and β:

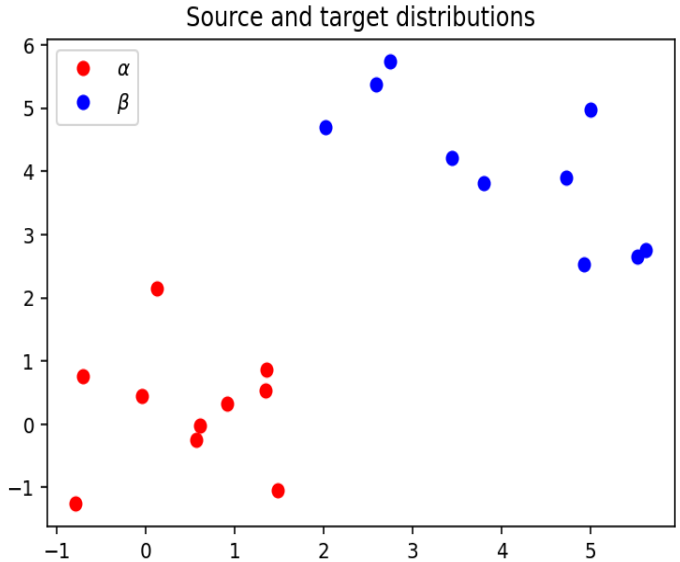
$$W_p(\alpha, \beta) = \left(\inf_{\gamma \in \Gamma(\alpha, \beta)} \int d^p(x, y) \, \mathrm{d}\gamma(x, y) \right)^{1/p}, \tag{3}$$

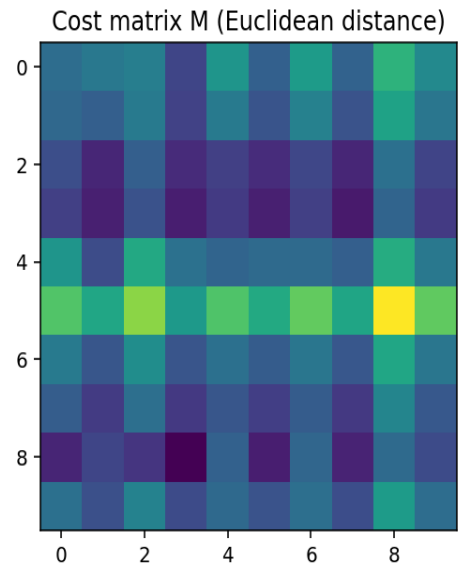
  which is the Wasserstein distance of order p, with $p \geq 1$.
- Solving Kantorovich problem with the quadratic cost ($p = 2$) yields L_2-**optimal assignment**. The solution is an optimal transportation plan induced by the optimal map $\mathcal{T}$, which can be expressed as the gradient of a convex function.
- Setting $p = 1$ we have W_1, also known as the Earth Mover's Distance (EMD), see e.g. `Python Optimal Transport` (POT) library

A visual intuition (W_1)

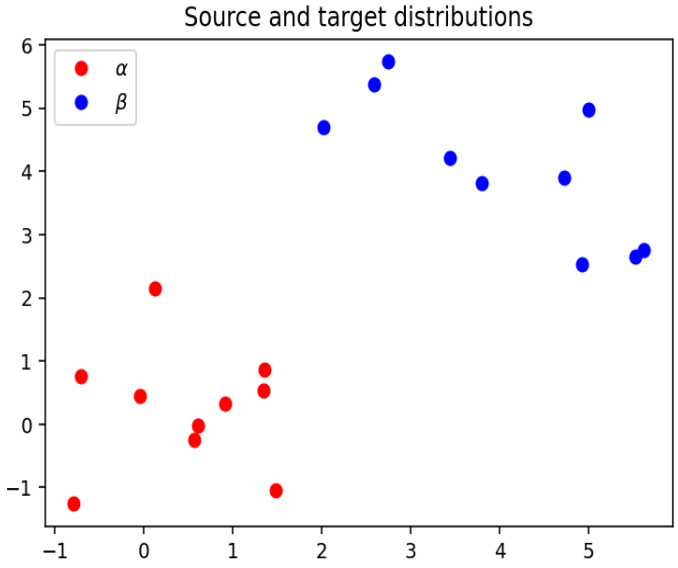


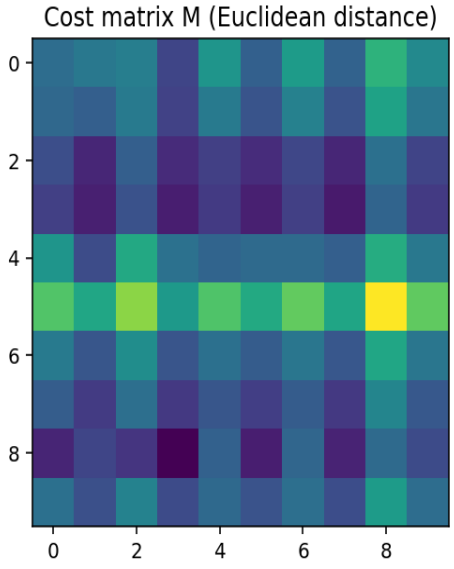
A visual intuition (W_1)

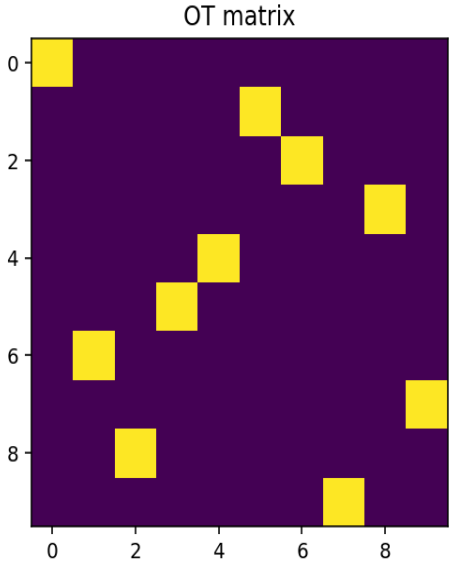




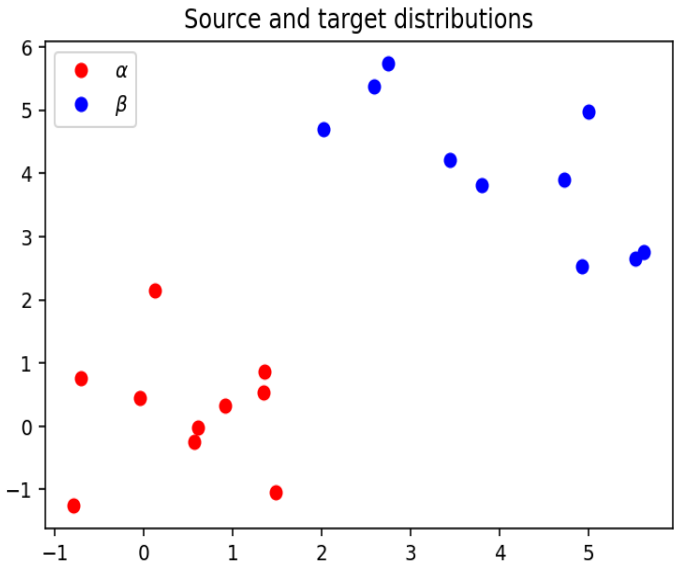
A visual intuition (W_1)

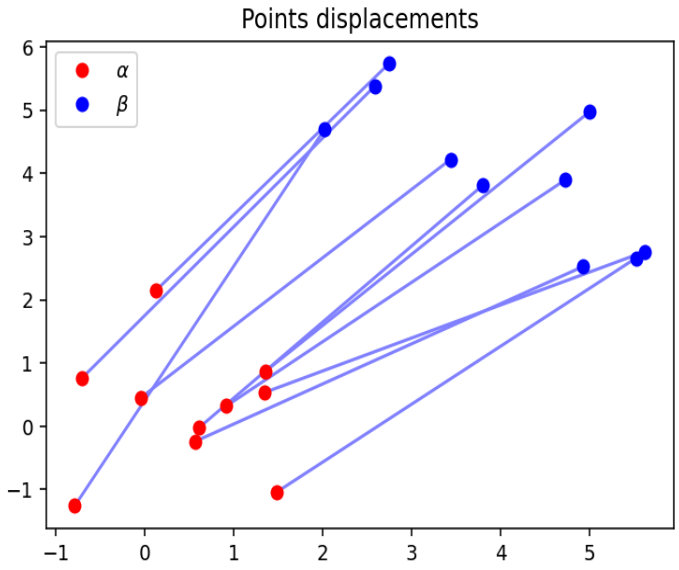


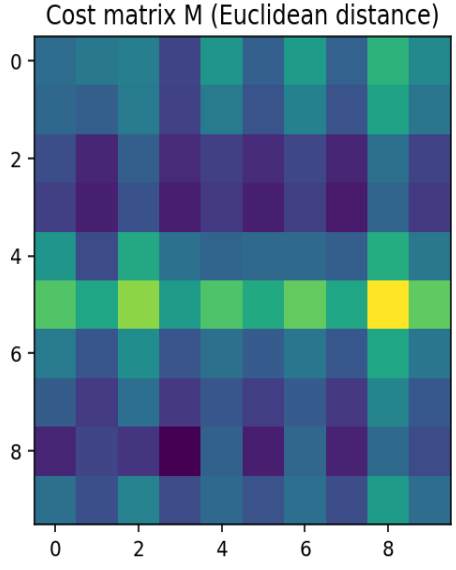


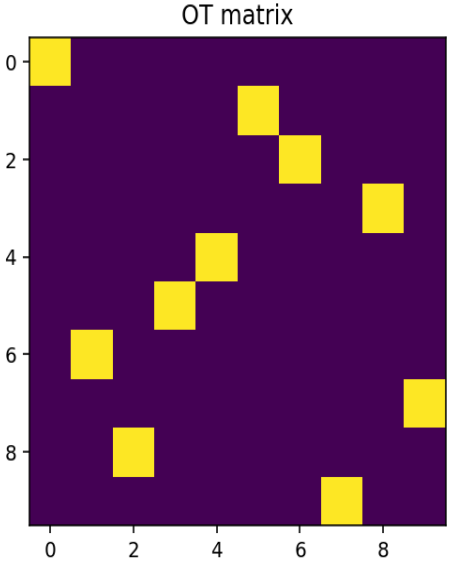


A visual intuition (W_1)









OT in action: some examples

Images analysis in ML (pics taken with mobile phone in the garden....during the COVD-19 lockdown)

OT in action: some examples

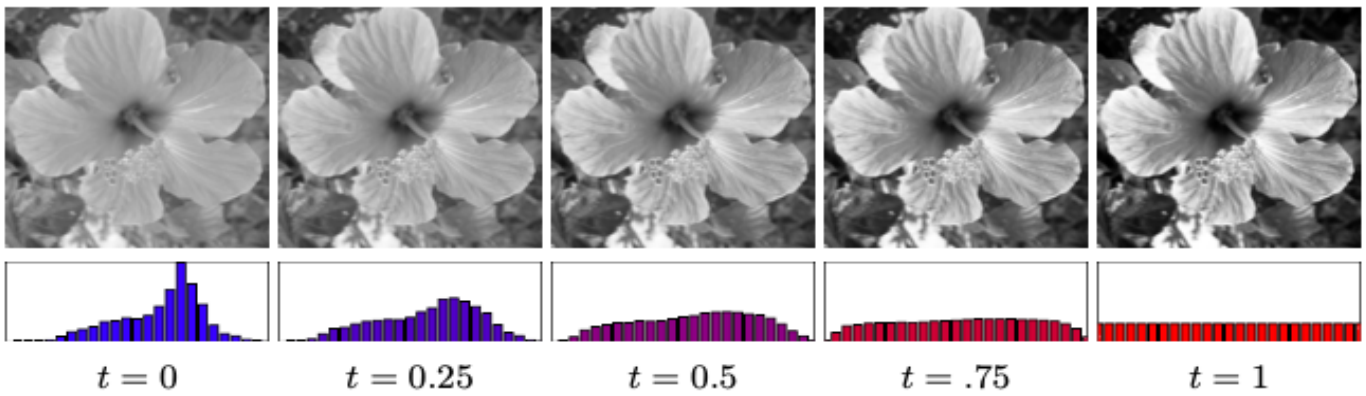
Images analysis in ML (pics taken with mobile phone in the garden....during the COVD-19 lockdown)

Original picture

The histogram in blue below the flower image represents the distribution of pixel intensities in the grayscale version of the image– pixel intensity typically ranges from 0 (black) to 1 (white)if normalized, or from 0 to 255 in standard 8-bit format

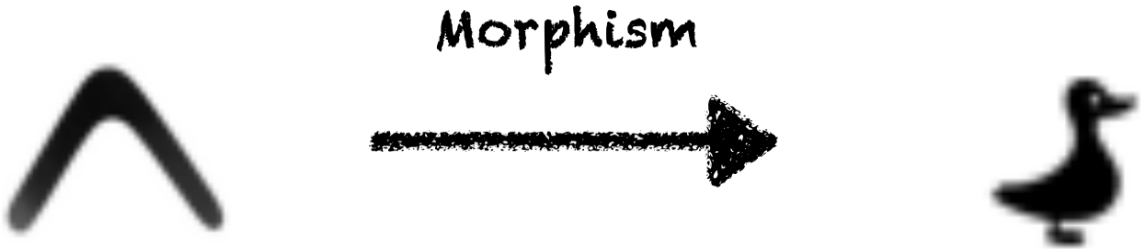
OT in action: some examples

As in Peyré & Cuturi (2019)

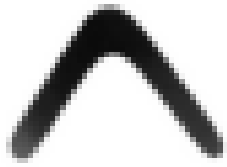


OT in action: some examples

Shapes analysis (morphism) in ML (a way to entertain my kids during COVID-19 lockdown)



Q. How to do it?



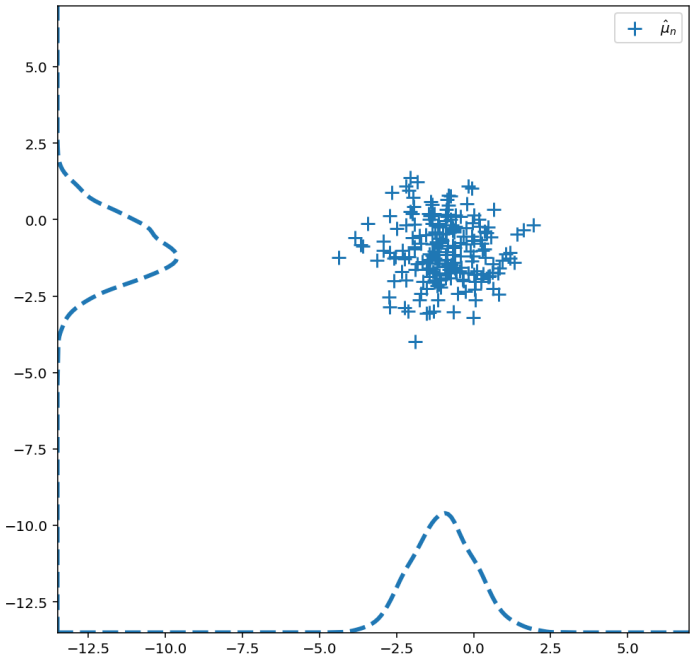
OT in action: some examples

Minimum distance estimation: given a parametric model

$$\mathcal{M}_\theta = \{\mu_\theta,\ \theta \in \mathbb{R}^k,\ k \geq 1\}.$$

Labelling as $\hat{\mu}_n$ the empirical measure, we search a Minimum Kantorovich Estimator (MKE), defined as

$$\hat{\theta}_n = \arg\min W_p(\hat{\mu}_n, \mu_\theta),$$



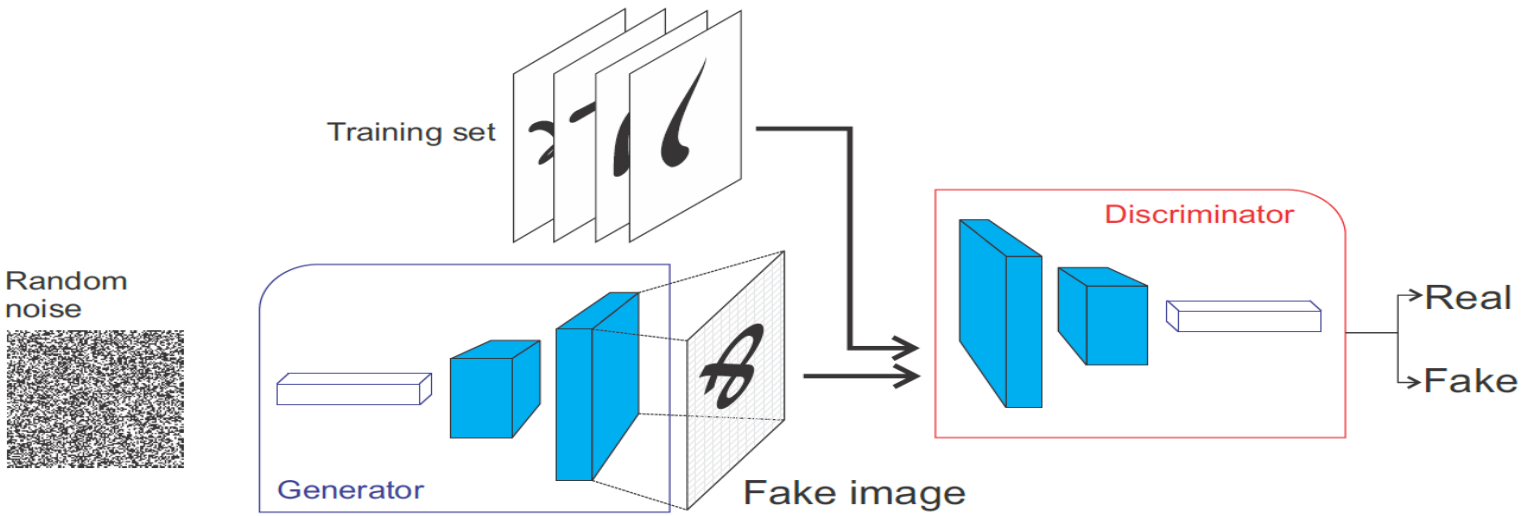
OT in action: some examples

With the development of neural networks and increased computational power, **deep generative modeling** has emerged as a major focus in Artificial Intelligence (AI). Its applications began in core Machine Learning (ML) areas such as text, image, and shape analysis.

OT in action: some examples

With the development of neural networks and increased computational power, **deep generative modeling** has emerged as a major focus in Artificial Intelligence (AI). Its applications began in core Machine Learning (ML) areas such as text, image, and shape analysis.

Architecture of Generative Generative Adversarial Networks (GAN) by Ian J. Goodfellow, Yoshua Bengio et al.



OT in action: some examples

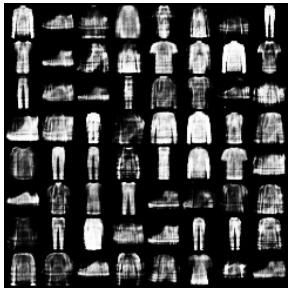
For instance, **Wasserstein Generative Adversarial Networks (WGAN)** are widely-applied to create images.

OT in action: some examples

For instance, **Wasserstein Generative Adversarial Networks (WGAN)** are widely-applied to create images. As an example, let us consider this image which from Fashion-MNIST database — Zalando's article images and each image is 28×28 pixels

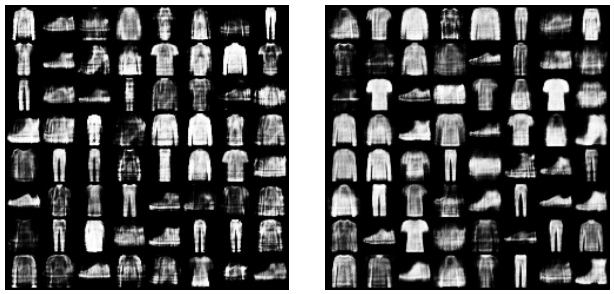
OT in action: some examples

Using different types of the OT problem specifications (more to come) one get:



OT in action: some examples

Using different types of the OT problem specifications (more to come) one get:

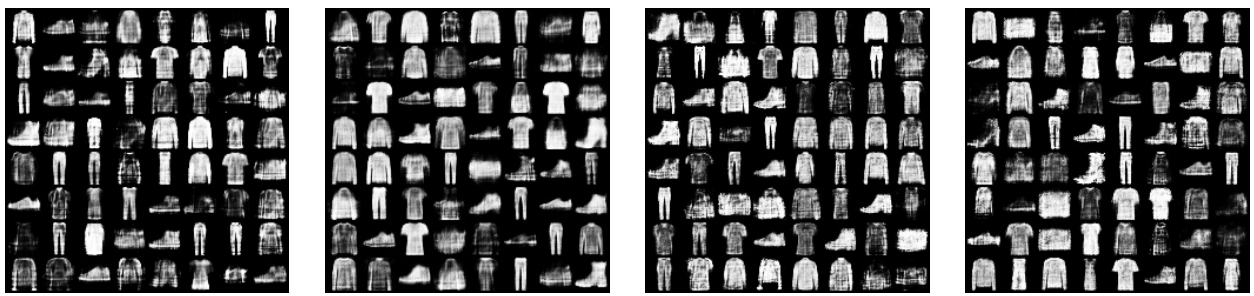


OT in action: some examples

Using different types of the OT problem specifications (more to come) one get:

OT in action: some examples

Using different types of the OT problem specifications (more to come) one get:



Domain adaptation

In supervised learning, **Domain Adaptation (DA)** addresses distribution mismatch between training samples Ω_s and test samples Ω_t.

Domain adaptation

In supervised learning, **Domain Adaptation (DA)** addresses distribution mismatch between training samples Ω_s and test samples Ω_t.

Aim (transfer learning)

Transfer knowledge from a labeled source domain to an unlabeled target domain, even if data distributions differ—e.g., due to varying acquisition conditions in vision tasks.

Domain adaptation

In supervised learning, **Domain Adaptation (DA)** addresses distribution mismatch between training samples Ω_s and test samples Ω_t.

Aim (transfer learning)

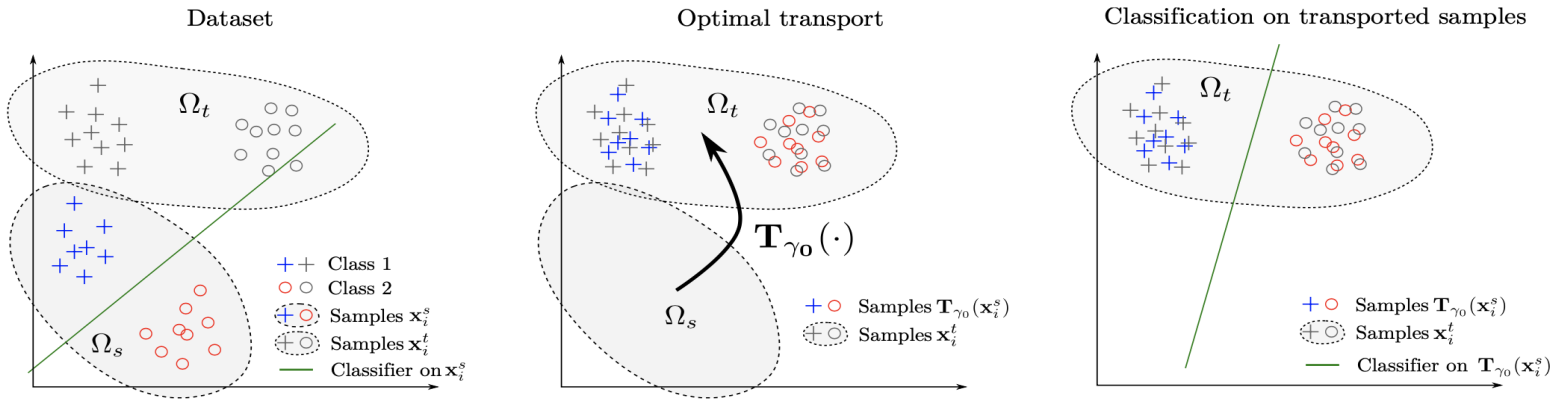
Transfer knowledge from a labeled source domain to an unlabeled target domain, even if data distributions differ—e.g., due to varying acquisition conditions in vision tasks.

Solution

(i) transform data so that we align source and target distributions; (ii) use the label information available in the source domain to learn a classifier in the transformed domain, which can be applied to the target domain.

Domain adaptation

From Coutry et al. (2014, 2017):



Left (Before Adaptation): A classifier trained on source domain data performs poorly on target data due to distribution mismatch.

Middle (Transport): Estimate a **nonlinear optimal transport map** T_{γ_0} that aligns the source and target distributions by minimizing a transport cost.

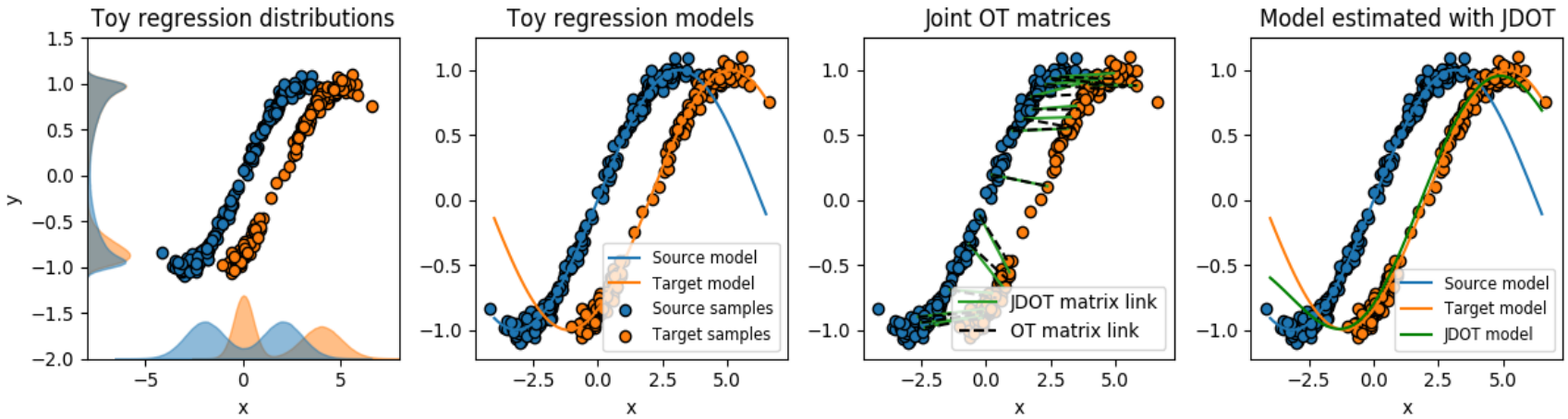
Right (After Adaptation): Transported labeled source samples are used to train a new classifier that generalizes well on the target domain.

Remark

The OT-based method handle both (linear and nonlinear) regression and classification problems.

Domain adaptation

From Coutry et al. (2017):



Remark

Y is a **sinus transformation** *of the X, but X in* Ω_t *is a shifted version of the X in* Ω_s*. In spite of this shift, the application of OT allows recover a model (green line) which is very similar to the theoretical one (orange line).*

A comment and a natural question

C: Everything works beautifully!

A comment and a natural question

C: Everything works beautifully!

Q: Why do we need to introduce robustness?

Some papers of mine

Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

Some papers of mine

Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

- *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (with. M. Hallin and H. Liu), **Bernoulli**, 2020
- *Center-outward R-estimation for semiparametric VARMA models*, (with. M. Hallin and H. Liu), **Journal of the American Statistical Association**, 2022

Some papers of mine

Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

- *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (with. M. Hallin and H. Liu), **Bernoulli**, 2020
- *Center-outward R-estimation for semiparametric VARMA models*, (with. M. Hallin and H. Liu), **Journal of the American Statistical Association**, 2022
- *On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective*, (with E. Ronchetti and A. Ilievski), **Statistical Science**, 2023

Some papers of mine

Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

- *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (with. M. Hallin and H. Liu), **Bernoulli**, 2020
- *Center-outward R-estimation for semiparametric VARMA models*, (with. M. Hallin and H. Liu), **Journal of the American Statistical Association**, 2022
- *On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective*, (with E. Ronchetti and A. Ilievski), **Statistical Science**, 2023
- *Discussion of "Robust Distance Covariance" by S. Leyder, J. Raymaekers, and PJ Rousseeuw*, (with M. Hallin, H. Liu, and X. Xu), **International Statistical Review**, 2025+

Some papers of mine

Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

- *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (with. M. Hallin and H. Liu), **Bernoulli**, 2020
- *Center-outward R-estimation for semiparametric VARMA models*, (with. M. Hallin and H. Liu), **Journal of the American Statistical Association**, 2022
- *On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective*, (with E. Ronchetti and A. Ilievski), **Statistical Science**, 2023
- *Discussion of "Robust Distance Covariance" by S. Leyder, J. Raymaekers, and PJ Rousseeuw*, (with M. Hallin, H. Liu, and X. Xu), **International Statistical Review**, 2025+
- *Inference via robust optimal transportation: theory and methods*, (with Y. Ma, H. Liu, and M. Lerasle), **International Statistical Review**, 2025+

Some papers of mine

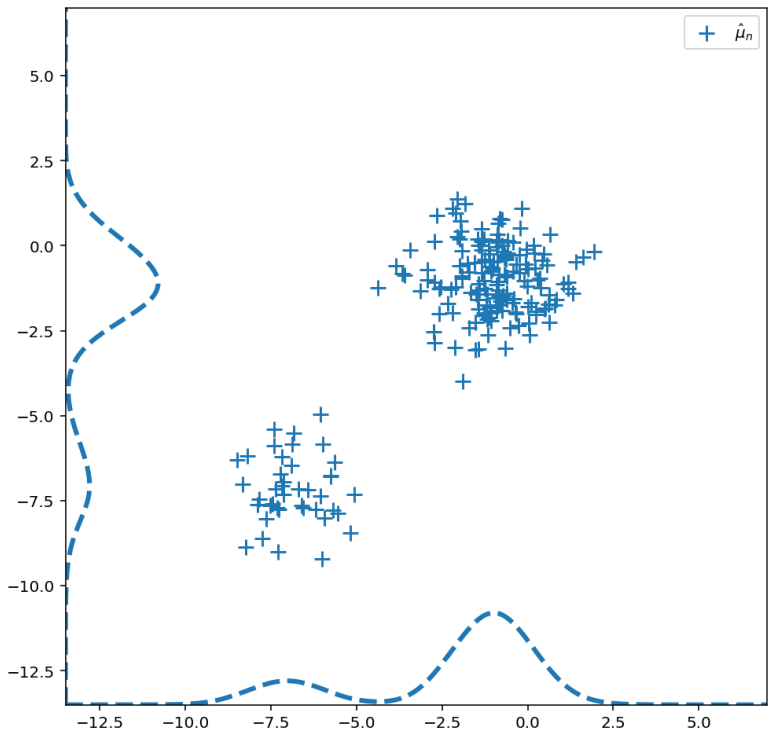
Building on these ideas and techniques, I've been developing some robust inference procedures, during the time frame 2020-2025:

- *Rank-based testing for semiparametric VAR models: a measure transportation approach*, (with. M. Hallin and H. Liu), **Bernoulli**, 2020
- *Center-outward R-estimation for semiparametric VARMA models*, (with. M. Hallin and H. Liu), **Journal of the American Statistical Association**, 2022
- *On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective*, (with E. Ronchetti and A. Ilievski), **Statistical Science**, 2023
- *Discussion of "Robust Distance Covariance" by S. Leyder, J. Raymaekers, and PJ Rousseeuw*, (with M. Hallin, H. Liu, and X. Xu), **International Statistical Review**, 2025+
- *Inference via robust optimal transportation: theory and methods*, (with Y. Ma, H. Liu, and M. Lerasle), **International Statistical Review**, 2025+
- ...

The sensitivity of OT to outliers

Remark

The ability to lift the ground distance d^p is one of the perks of W_p and it makes it a suitable tool in statistics and ML: W_1 and W_2 are widely-applied in many scientific areas: Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

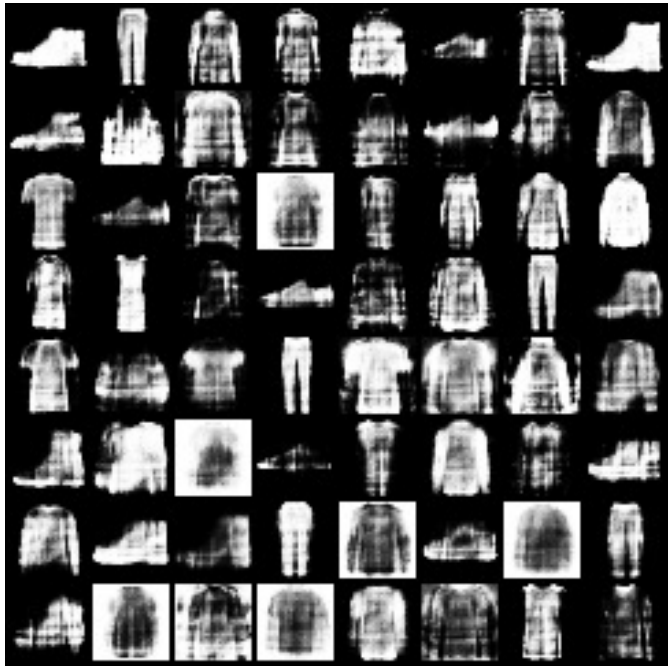


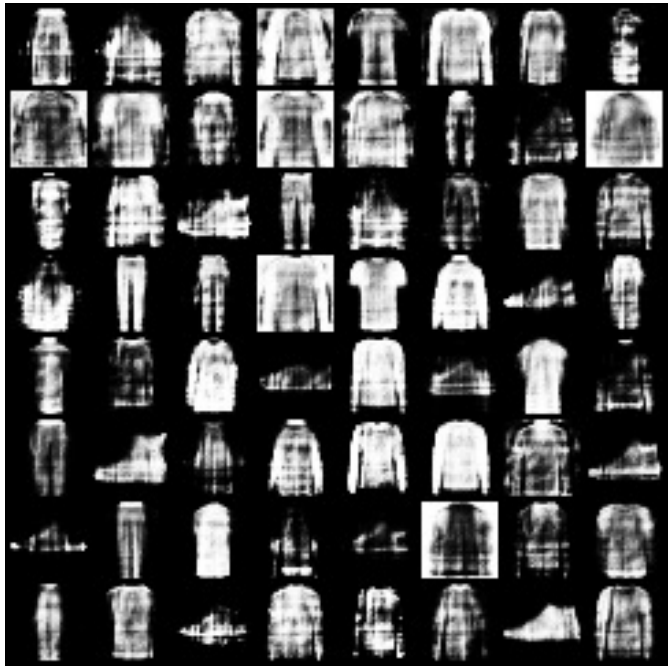
The sensitivity of OT to outliers

To the Fashion MNIST dataset, we add contaminated images

and we apply the WGAN ...

The sensitivity of OT to outliers





Remark

We clearly see that the WGAN generates images containing outliers: the algorithm does not distinguish clean from contaminated images and it learns how to creates anomalous records!

The sensitivity of OT to outliers

Q.: Why should we care about it?

[1] Adversarial attacks in ML involve creating carefully crafted inputs that are designed to fool a machine learning model into making incorrect predictions— often a human can do better.

The sensitivity of OT to outliers

Q.: Why should we care about it?

Shariff et al. (2016) focus on facial biometric systems, which are widely used in surveillance: define and investigate a novel class of adversarial attacks[1] that are physically realizable and allow an attacker to evade recognition or impersonate another individual.

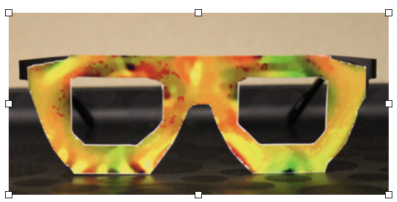
[1] Adversarial attacks in ML involve creating carefully crafted inputs that are designed to fool a machine learning model into making incorrect predictions— often a human can do better.

The sensitivity of OT to outliers

Q.: Why should we care about it?

Shariff et al. (2016) focus on facial biometric systems, which are widely used in surveillance: define and investigate a novel class of adversarial attacks[1] that are physically realizable and allow an attacker to evade recognition or impersonate another individual.

Consider a home owner who uses a face recognition system as a security feature. We can generate an adversarial eyeglass that can be printed and placed on a real eyeglass frame to fool face recognition models



[1] Adversarial attacks in ML involve creating carefully crafted inputs that are designed to fool a machine learning model into making incorrect predictions— often a human can do better.

The aim of the talk

Q.: How to fix (some of) these issues?

The aim of the talk

Q.: How to fix (some of) these issues?

I am going to mention some developments of OT theory to derive novel, robust inference procedures for data analysis.

The aim of the talk

Q.: How to fix (some of) these issues?

I am going to mention some developments of OT theory to derive novel, robust inference procedures for data analysis.

Aim

The resulting techniques:

The aim of the talk

Q.: How to fix (some of) these issues?

I am going to mention some developments of OT theory to derive novel, robust inference procedures for data analysis.

Aim

The resulting techniques:

- *are based on the novel concept of robust Wasserstein distance* $(W^{(\lambda)}, \lambda > 0)$ *between measures and do not need finite moments*

The aim of the talk

Q.: How to fix (some of) these issues?

I am going to mention some developments of OT theory to derive novel, robust inference procedures for data analysis.

Aim

The resulting techniques:

- *are based on the novel concept of robust Wasserstein distance* $(W^{(\lambda)}, \lambda > 0)$ *between measures and do not need finite moments*
- *yield novel concentration inequalities and mean convergence rates*

The aim of the talk

Q.: How to fix (some of) these issues?

I am going to mention some developments of OT theory to derive novel, robust inference procedures for data analysis.

Aim

The resulting techniques:

- *are based on the novel concept of robust Wasserstein distance* $(W^{(\lambda)}, \lambda > 0)$ *between measures and do not need finite moments*
- *yield novel concentration inequalities and mean convergence rates*
- *can be applied to many problems in Stats, ML and AI, e.g. Minimum Distance Estimation in parametric models, Generative Adversarial Networks (GAN) and domain adaptation (DA)*

ROBust Optimal Transport (ROBOT)

ROBOT problem is defined in Mukherjee et al. (2021) who introduce a TV-regularized OT re-formulation and define a novel version of the OT problem:

$$\inf \left\{ \int c_\lambda(x,y) \, \mathrm{d}\gamma(x,y) : \gamma \in \Gamma(\mu,\nu) \right\}, \tag{4}$$

which is similar to the original OT problem, but the cost function

$$c(x,y) = d(x,y)$$

is replaced by

$$c_\lambda = \min\{c, 2\lambda\}$$

that is bounded from above by $2\lambda \in \mathbb{R}_+$.

ROBust Optimal Transport (ROBOT)

We take over from Mukherjee et al. (2021) and we prove that, similarly to OT, for $c_\lambda(x, y)$,

$$W^{(\lambda)}(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int c_\lambda(x, y) \, \mathrm{d}\gamma(x, y) \right\}. \tag{5}$$

In Ma et a. (2025) we prove that to this OT problem we can attach a notion of Robust Wasserstein distance, which is such that, if $W_1(\mu, \nu)$ exists, we have

$$\lim_{\lambda \to \infty} W^{(\lambda)}(\mu, \nu) = W_1(\mu, \nu).$$

Given a class of parametric models $\{\mu_\theta, \theta \in \Theta \subset \mathbb{R}^k\}$, to this distance, we associate the *minimum robust Wasserstein estimator* (MRWE)

$$\hat{\theta}_n^\lambda = \operatorname*{argmin}_{\theta \in \Theta} \underbrace{W^{(\lambda)}(\mu_\theta, \hat{\mu}_n)}_{\text{loss function}},$$

where $\hat{\mu}_n$ is the empirical measure.

Remark

ROBust Optimal Transport (ROBOT)

- *Intuitively, the* **consistency** *can be conceptualized as follows. The empirical measure converges to* $\mu_\star$*:*

$$W^{(\lambda)}(\hat{\mu}_n, \mu_\star) \to 0$$

as $n \to \infty$*. Therefore,*

$$\hat{\theta}_n^\lambda = \arg\min W^{(\lambda)}(\hat{\mu}_n, \mu_\theta)$$

converges to

$$\theta_\star = \arg\min W^{(\lambda)}(\mu_\star, \mu_\theta)$$

Remark

ROBust Optimal Transport (ROBOT)

- *Intuitively, the* **consistency** *can be conceptualized as follows. The empirical measure converges to* $\mu_\star$*:*

$$W^{(\lambda)}(\hat{\mu}_n, \mu_\star) \to 0$$

as $n \to \infty$*. Therefore,*

$$\hat{\theta}_n^\lambda = \arg\min W^{(\lambda)}(\hat{\mu}_n, \mu_\theta)$$

converges to

$$\theta_\star = \arg\min W^{(\lambda)}(\mu_\star, \mu_\theta)$$

- *The boundedness of the* c_λ *implies* **robustness** *to outliers and existence of the estimator even if* $\mu_\star$ *does not admit finite moments of any order.*

Remark

ROBust Optimal Transport (ROBOT)

- *Intuitively, the* **consistency** *can be conceptualized as follows. The empirical measure converges to* $\mu_\star$*:*

$$W^{(\lambda)}(\hat{\mu}_n, \mu_\star) \to 0$$

as $n \to \infty$*. Therefore,*

$$\hat{\theta}_n^\lambda = \arg\min W^{(\lambda)}(\hat{\mu}_n, \mu_\theta)$$

converges to

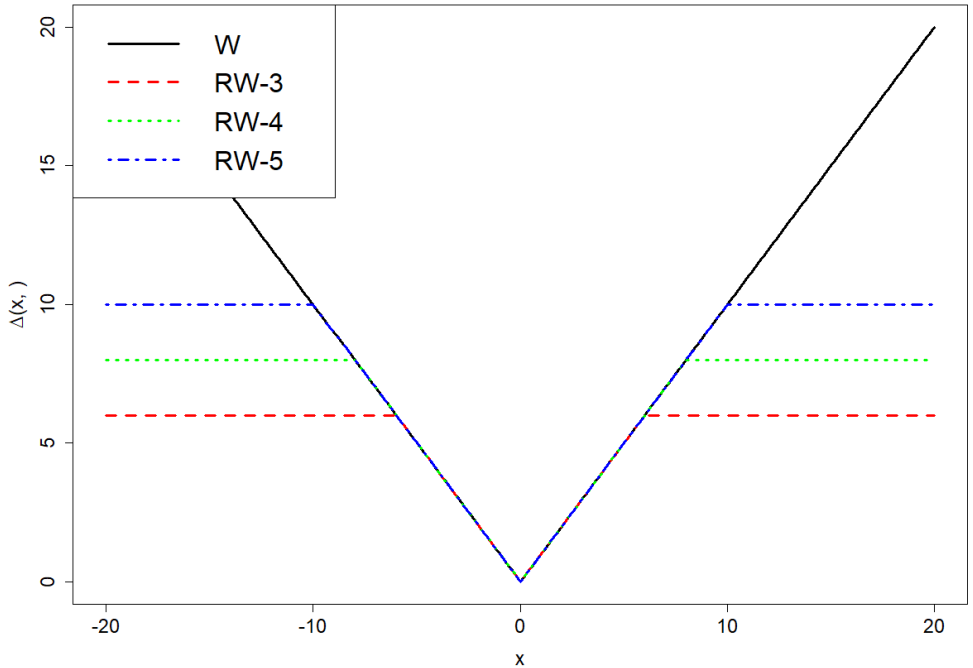
$$\theta_\star = \arg\min W^{(\lambda)}(\mu_\star, \mu_\theta)$$

- *The boundedness of the* c_λ *implies* **robustness** *to outliers and existence of the estimator even if* $\mu_\star$ *does not admit finite moments of any order.*

- *We derive a* **data-driven, non-asymptotic selection criterion for the tuning constant** λ*: we resort on a concentration inequality for* $W^{(\lambda)}$ *to control the stability of its distribution in the presence of contamination.*

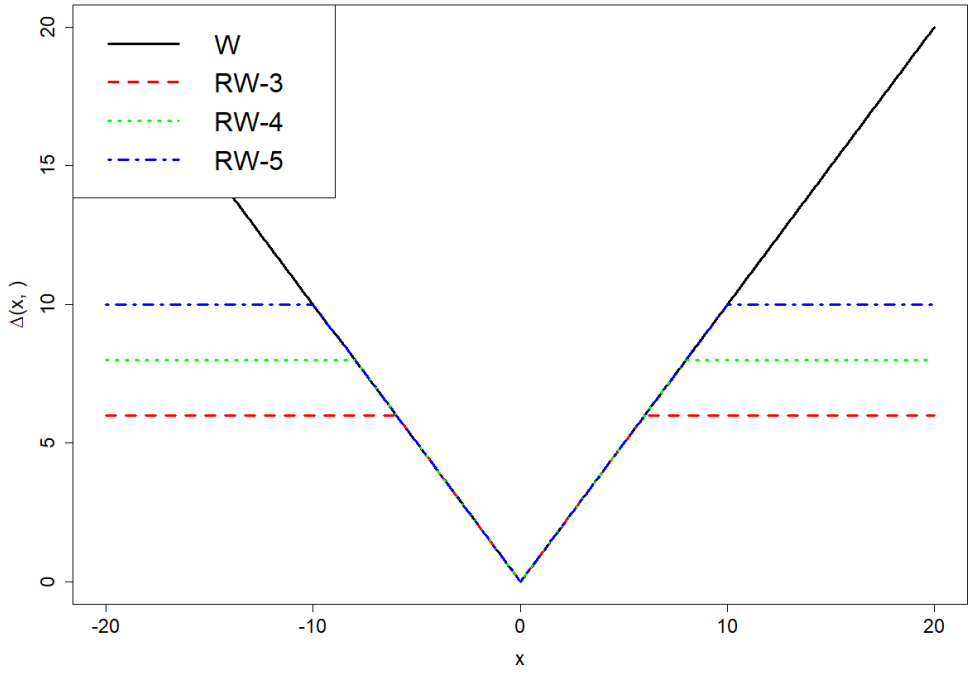
ROBust Optimal Transport (ROBOT)

As far as robustness is concerned, plotting the loss function $W^{(\lambda)}$ and W_1 yields



ROBust Optimal Transport (ROBOT)

As far as robustness is concerned, plotting the loss function $W^{(\lambda)}$ and W_1 yields



Remark

The cost c_λ determines, in the language of robust statistics, the so-called "hard rejection": it bounds the influence of outlying values (to be contrasted with the behavior of Huber loss, which downweights outliers to preserve efficiency at the reference model); see Ronchetti (2022).

Minimum Distance Estimation: parametric model $\mathcal{M}_\theta$

Synthetic data: illustrate the performance of MERWE in estimation of a (location) parameter in the univariate setting. We consider the sum of log-normal r.v.s, with and w/o ε-contamination and different n

Minimum Distance Estimation: parametric model $\mathcal{M}_\theta$

Synthetic data: illustrate the performance of MERWE in estimation of a (location) parameter in the univariate setting. We consider the sum of log-normal r.v.s, with and w/o ε-contamination and different n

	$n = 100$				$n = 200$				$n = 1000$			
	BIAS		MSE		BIAS		MSE		BIAS		MSE	
	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE
$\varepsilon = 0.1, \eta = 1$	0.049	0.092	0.003	0.010	0.042	0.093	0.002	0.012	0.037	0.086	0.002	0.008
$\varepsilon = 0.1, \eta = 4$	0.035	0.090	0.002	0.012	0.029	0.097	0.001	0.016	0.013	0.100	≈ 0	0.018
$\varepsilon = 0.2, \eta = 1$	0.071	0.157	0.008	0.028	0.086	0.178	0.009	0.033	0.081	0.172	0.007	0.031
$\varepsilon = 0.2, \eta = 4$	0.046	0.204	0.003	0.045	0.035	0.203	0.002	0.043	0.017	0.195	≈ 0	0.038
$\varepsilon = 0$	0.036	0.034	0.002	0.002	0.022	0.022	0.001	0.001	0.012	0.010	≈ 0	≈ 0

Minimum Distance Estimation: parametric model $\mathcal{M}_\theta$

Synthetic data: illustrate the performance of MERWE in estimation of a (location) parameter in the univariate setting. We consider the sum of log-normal r.v.s, with and w/o ε-contamination and different n

	$n = 100$				$n = 200$				$n = 1000$			
	BIAS		MSE		BIAS		MSE		BIAS		MSE	
	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE
$\varepsilon = 0.1, \eta = 1$	0.049	0.092	0.003	0.010	0.042	0.093	0.002	0.012	0.037	0.086	0.002	0.008
$\varepsilon = 0.1, \eta = 4$	0.035	0.090	0.002	0.012	0.029	0.097	0.001	0.016	0.013	0.100	≈ 0	0.018
$\varepsilon = 0.2, \eta = 1$	0.071	0.157	0.008	0.028	0.086	0.178	0.009	0.033	0.081	0.172	0.007	0.031
$\varepsilon = 0.2, \eta = 4$	0.046	0.204	0.003	0.045	0.035	0.203	0.002	0.043	0.017	0.195	≈ 0	0.038
$\varepsilon = 0$	0.036	0.034	0.002	0.002	0.022	0.022	0.001	0.001	0.012	0.010	≈ 0	≈ 0

Remark

- *In small samples $n = 100$, the MERWE has smaller bias and MSE than the MEWE, in all settings. Similar results are available in moderate samples, $n = 200$*

Minimum Distance Estimation: parametric model $\mathcal{M}_\theta$

Synthetic data: illustrate the performance of MERWE in estimation of a (location) parameter in the univariate setting. We consider the sum of log-normal r.v.s, with and w/o ε-contamination and different n

	$n = 100$				$n = 200$				$n = 1000$			
	BIAS		MSE		BIAS		MSE		BIAS		MSE	
	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE
$\varepsilon = 0.1, \eta = 1$	0.049	0.092	0.003	0.010	0.042	0.093	0.002	0.012	0.037	0.086	0.002	0.008
$\varepsilon = 0.1, \eta = 4$	0.035	0.090	0.002	0.012	0.029	0.097	0.001	0.016	0.013	0.100	≈ 0	0.018
$\varepsilon = 0.2, \eta = 1$	0.071	0.157	0.008	0.028	0.086	0.178	0.009	0.033	0.081	0.172	0.007	0.031
$\varepsilon = 0.2, \eta = 4$	0.046	0.204	0.003	0.045	0.035	0.203	0.002	0.043	0.017	0.195	≈ 0	0.038
$\varepsilon = 0$	0.036	0.034	0.002	0.002	0.022	0.022	0.001	0.001	0.012	0.010	≈ 0	≈ 0

Remark

- *In small samples $n = 100$, the MERWE has smaller bias and MSE than the MEWE, in all settings. Similar results are available in moderate samples, $n = 200$*
- *For $n = 1000$, MERWE and MEWE have similar performance when $\varepsilon = 0$ (no contamination), whilst the MERWE still has smaller MSE for $\varepsilon > 0$. This implies that the MERWE maintains good efficiency with respect to MEWE at the reference model.*

Robust WGAN (RWGAN)

We propose RWGAN-1 and RWGAN-2, which are two RWGAN deep learning models: both approaches are based on dual version of ROBOT; Ma et al. 2025. We compare these two methods with routinely-applied Wasserstein GAN (WGAN) and with the robust WGAN introduced by Balaji et al 2020.

Using **syntetic data**, we study the robustness of RWGAN-1 and RWGAN-2. We consider reference samples generated from a simple model, which includes some outliers:

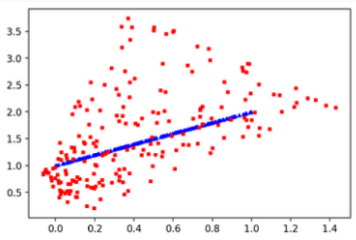
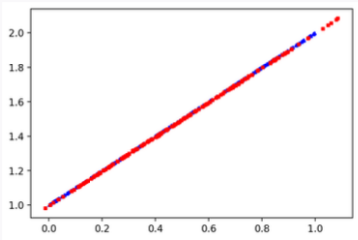
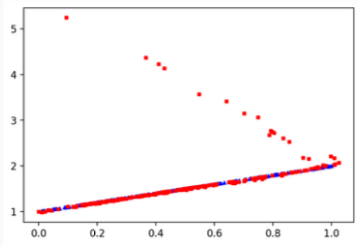
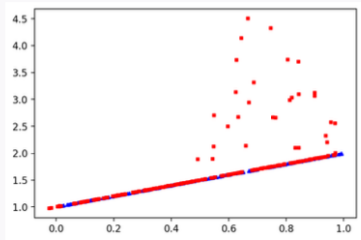
$$\begin{aligned} X_{i_1}^{(n)} \sim \mathrm{U}(0,1), X_{i_2}^{(n)} = X_{i_1}^{(n)} + 1, \\ X_i^{(n)} = (X_{i_1}^{(n)}, X_{i_2}^{(n)}), i = 1, 2, \ldots, n_1, \\ X_i^{(n)} = (X_{i_1}^{(n)}, X_{i_2}^{(n)} + \eta), i = n_1 + 1, n_1 + 2, \ldots, n, \end{aligned} \tag{6}$$

with η representing the size of outliers. We set $n = 1000$ and try four different settings by changing values of $\varepsilon = (n - n_1)/n$ and η.

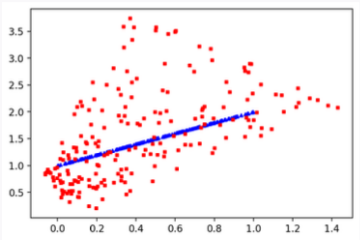
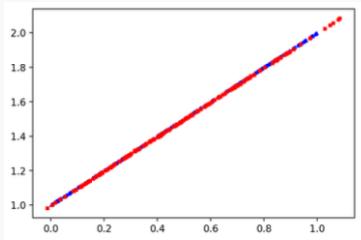
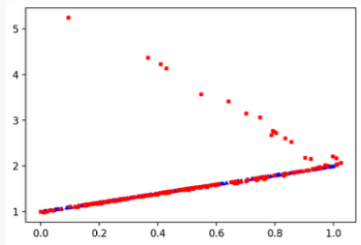
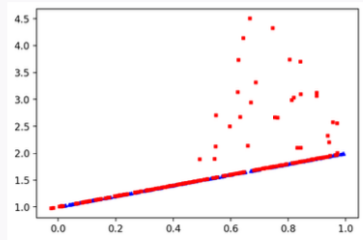
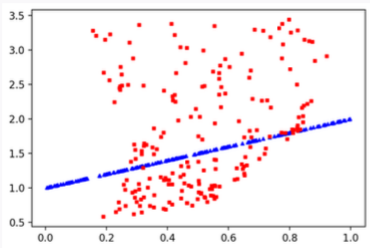
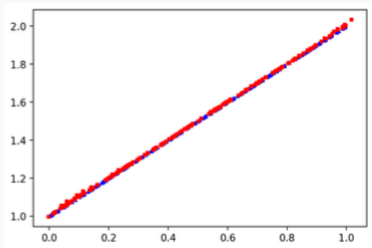
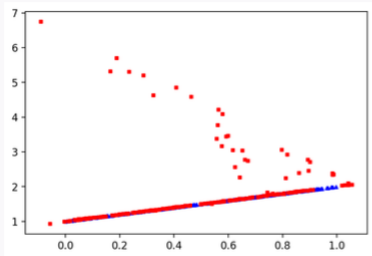
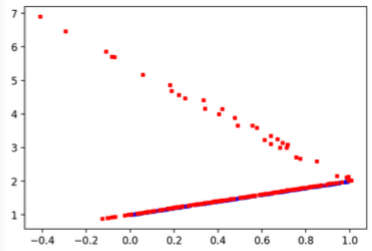
Robust WGAN (RWGAN)

WGAN RWGAN-1 RWGAN-2 RWGAN-B

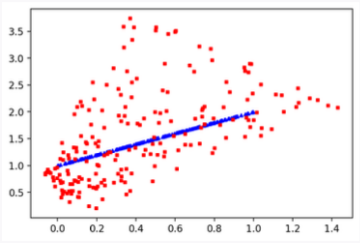
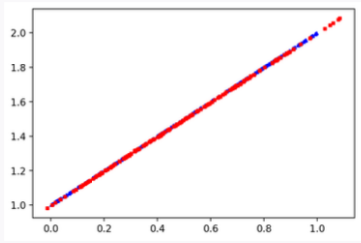
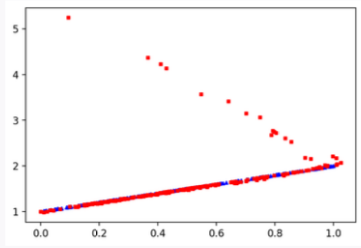
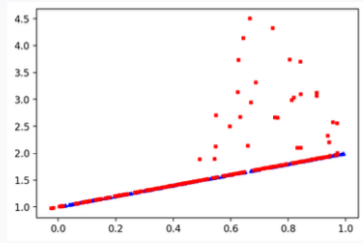
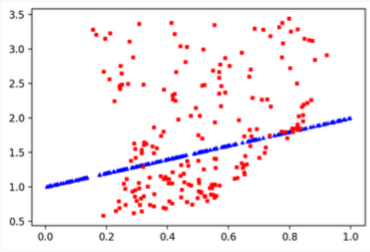
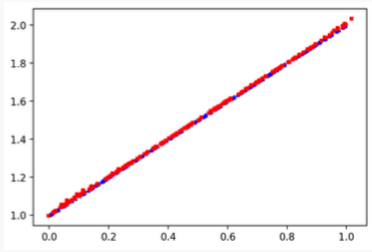
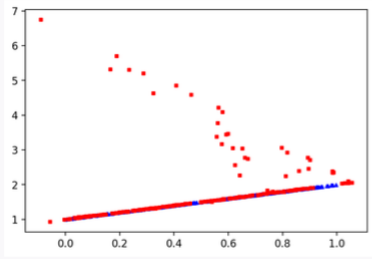
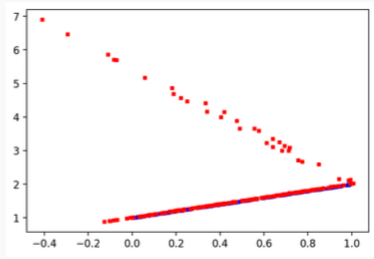
Robust WGAN (RWGAN)

	WGAN	RWGAN-1	RWGAN-2	RWGAN-B
10%	 (a) $W^1 = 0.5864$	 (b) $W^1 = 0.0514$	 (c) $W^1 = 0.1560$	 (d) $W^1 = 0.1771$

Robust WGAN (RWGAN)

	WGAN	RWGAN-1	RWGAN-2	RWGAN-B
10%	 (a) $W^1 = 0.5864$	 (b) $W^1 = 0.0514$	 (c) $W^1 = 0.1560$	 (d) $W^1 = 0.1771$
20%	 (i) $W^1 = 0.5646$	 (j) $W^1 = 0.0470$	 (k) $W^1 = 0.2938$	 (l) $W^1 = 0.3229$

Robust WGAN (RWGAN)

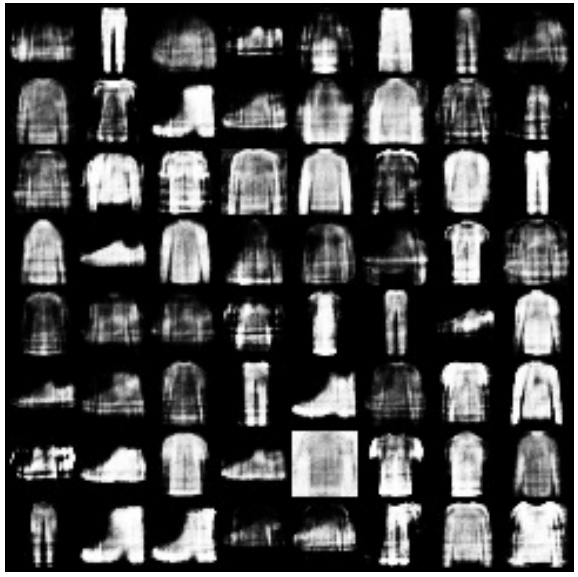
	WGAN	RWGAN-1	RWGAN-2	RWGAN-B
10%	 (a) $W^1 = 0.5864$	 (b) $W^1 = 0.0514$	 (c) $W^1 = 0.1560$	 (d) $W^1 = 0.1771$
20%	 (i) $W^1 = 0.5646$	 (j) $W^1 = 0.0470$	 (k) $W^1 = 0.2938$	 (l) $W^1 = 0.3229$

Remark

WGAN is greatly affected by outliers. Differently, RWGAN-1 performs better than its competitors: it generates data that agree with the uncontaminated distribution, even when the proportion and size of outliers are large.

Robust WGAN (RWGAN)

...and for the Fashion MNIST dataset:


Remark

RWGAN remains stable even in the presence of outliers.

Take home message

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$
- We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$
- We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects
- We derive some concentration inequalities for $W^{(\lambda)}$ and illustrate their practical relevance for the selection of λ

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$
- We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects
- We derive some concentration inequalities for $W^{(\lambda)}$ and illustrate their practical relevance for the selection of λ
- We use $W^{(\lambda)}$ to define minimum distance estimators and provide their statistical guarantees, explaining that our novel estimators are outlier-resistant and well-defined even if the underlying model has heavy-tails

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$
- We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects
- We derive some concentration inequalities for $W^{(\lambda)}$ and illustrate their practical relevance for the selection of λ
- We use $W^{(\lambda)}$ to define minimum distance estimators and provide their statistical guarantees, explaining that our novel estimators are outlier-resistant and well-defined even if the underlying model has heavy-tails
- We derive the dual form of the ROBOT and illustrate its applicability to **ML and AI problems** (generative adversarial networks and domain adaptation)

Take home message

In the paper Ma et al. (2025):

- We consider a robust version of the primal OT problem (ROBOT) and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, which depends on a tuning parameter $\lambda > 0$
- We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects
- We derive some concentration inequalities for $W^{(\lambda)}$ and illustrate their practical relevance for the selection of λ
- We use $W^{(\lambda)}$ to define minimum distance estimators and provide their statistical guarantees, explaining that our novel estimators are outlier-resistant and well-defined even if the underlying model has heavy-tails
- We derive the dual form of the ROBOT and illustrate its applicability to **ML and AI problems** (generative adversarial networks and domain adaptation)
- Extension to Variational Auto-Encoders? Derivation of other concentration inequalities? Curse of dimensionality?