GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT
UNIVERSITÉ DE GENÈVE

Inference via robust optimal transportation: theory and methods

Davide La Vecchia
(with Y. Ma, H. Liu & M. Lerasle)

SMSA 2024, TU Delft, March-2022

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

Features

The resulting techniques:

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

Features

The resulting techniques:

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

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

Features

The resulting techniques:

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

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

Features

The resulting techniques:

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

I am going to introduce some developments of optimal transportation (OT) theory to derive novel, robust inference procedures for data analysis.

Features

The resulting techniques:

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

Cornell University

We gratefully acknowledge s
member insti

arXiv > math > arXiv:2301.06297

Search...

Help | Advanced Se

Mathematics > Statistics Theory

[Submitted on 16 Jan 2023 (v1), last revised 6 Dec 2023 (this version, v3)]

Inference via robust optimal transportation: theory and methods

Yiming Ma, Hang Liu, Davide La Vecchia, Metthieu Lerasle

Optimal transport (OT) theory and the related p–Wasserstein distance $(W_p, p \geq 1)$ are widely–applied in statistics and machine learning. In spite of their popularity, inference based on these tools is sensitive to outliers or it can perform poorly when the underlying model has heavy–tails. To cope with these issues, we introduce a new class of procedures. (i) 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 $\lambda > 0$. (ii) We illustrate the link between W_1 and $W^{(\lambda)}$ and study its key measure theoretic aspects. (iii) We derive some concentration inequalities for $W^{(\lambda)}$. (iii) We use $W^{(\lambda)}$ to define minimum distance estimators, we provide their statistical guarantees and we illustrate how to apply concentration inequalities for the selection of λ. (v) We derive the {dual} form of the ROBOT and illustrate its applicability to machine learning problems (generative adversarial networks and domain adaptation). Numerical exercises provide evidence of the benefits yielded by our methods.

Outline

- Monge-Kantorovich OT problem and the related Wasserstein distance $\{W_p, p \geq 1\}$
- Motivation: robustness issues of $\{W_p, p \geq 1\}$
- Robust Optimal Transport (ROBOT)
  - Dual form and Robust Wasserstein distance $\{W^{(\lambda)}, \lambda > 0\}$
  - Minimum Robust Wasserstein distance estimation
  - Implementation aspects and statistical guarantees
- Synthetic data examples
- Take home message

Monge-Kantorovich OT problem and $\{W_p, p \geq 1\}$

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:

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 ν and μ 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 $c(x, y)$ represents the cost of transporting x to y. Then, find a measurable transport map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves

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

where

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

with $X \sim \nu$, $Y \sim \mu$. The map $\mathcal{T}\#\nu = \mu$ does the push forward of ν to μ.

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 ν and μ 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 $c(x, y)$ represents the cost of transporting x to y. Then, find a measurable transport map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ that achieves

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

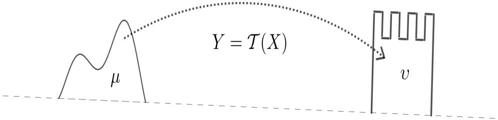
where

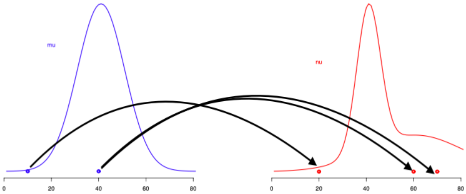
$$M := \{\mathcal{T} : X \to Y\},$$

with $X \sim \nu$, $Y \sim \mu$. The map $\mathcal{T}\#\nu = \mu$ does the push forward of ν to μ.

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

Two sketchy plots for visualisation ...





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).

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

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

where the infimum is over all coupling (X, Y) of (ν, μ), belonging to $\Gamma(\nu, \mu)$, the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$, satisfying

$$\boxed{\gamma(A \times \mathbb{R}^d) = \nu(A) \text{ and } \gamma(\mathbb{R}^d \times B) = \mu(B),}$$

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

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

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

where the infimum is over all coupling (X, Y) of (ν, μ), belonging to $\Gamma(\nu, \mu)$, the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$, satisfying

$$\boxed{\gamma(A \times \mathbb{R}^d) = \nu(A) \text{ and } \gamma(\mathbb{R}^d \times B) = \mu(B),}$$

for measurable sets $A, B \subset \mathbb{R}^d$: **one typically imposes exact marginal constraints!**

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

$$\inf_{\gamma\in\Gamma(\nu,\mu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \boxed{c(x,y)}\,\mathrm{d}\gamma(x,y), \tag{2}$$

where the infimum is over all coupling (X, Y) of (ν,μ), belonging to $\Gamma(\nu,\mu)$, the set of probability measures γ on $\mathbb{R}^d\times\mathbb{R}^d$, satisfying

Remark

Solving the optimal transport problem (2) with $\boxed{c = d^p}$*, introduces a distance between* μ *and* β*:*

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

which is the Wasserstein distance of order p $(p \geq 1)$.

Motivation

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. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

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. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

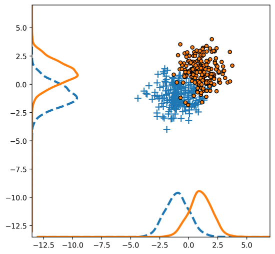
Example (Robustness issues and a preview of the solution)

Given two measure μ (target) and ν (original), OT embeds the distributions geometry: when the underlying distribution is contaminated by outliers, **the marginal constraints force OT** to transport outlying values, inducing an undesirable extra cost, which entails large changes in W_p.

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. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

Example (Robustness issues and a preview of the solution)

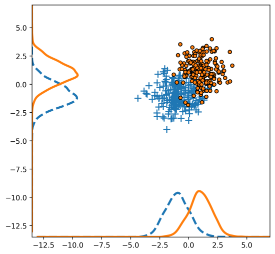
Given two measure μ (target) and ν (original), OT embeds the distributions geometry: when the underlying distribution is contaminated by outliers, **the marginal constraints force OT** to transport outlying values, inducing an undesirable extra cost, which entails large changes in W_p.

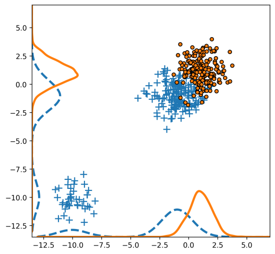


The ability to lift the ground distance d^p is one of the perks of W_p and it makes it a suitable tool in statistics and ML. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

Example (Robustness issues and a preview of the solution)

Given two measure μ (target) and ν (original), OT embeds the distributions geometry: when the underlying distribution is contaminated by outliers, **the marginal constraints force OT** to transport outlying values, inducing an undesirable extra cost, which entails large changes in W_p.

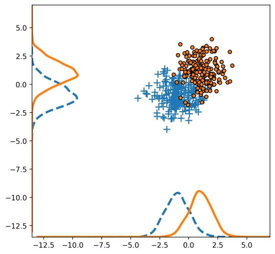




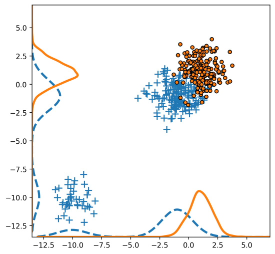
The ability to lift the ground distance d^p is one of the perks of W_p and it makes it a suitable tool in statistics and ML. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

Example (Robustness issues and a preview of the solution)

Given two measure μ (target) and ν (original), OT embeds the distributions geometry: when the underlying distribution is contaminated by outliers, **the marginal constraints force OT** to transport outlying values, inducing an undesirable extra cost, which entails large changes in W_p.



$W_1 = 3.00,\ W_2 = 3.02,\ W^{(\lambda)} = 2.93$



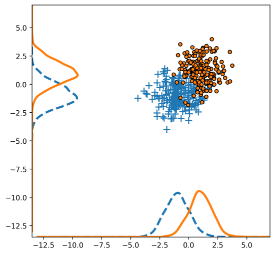
$W_1 = 5.32,\ W_2 = 6.62$ and $W^{(\lambda)} = 3.31$

where $\lambda = 3$.

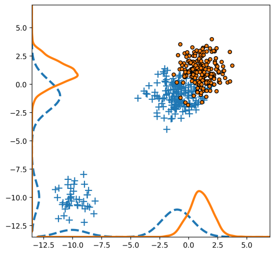
The ability to lift the ground distance d^p is one of the perks of W_p and it makes it a suitable tool in statistics and ML. Interestingly, this desirable feature becomes a negative aspect as far as robustness is concerned.

Example (Robustness issues and a preview of the solution)

Given two measure μ (target) and ν (original), OT embeds the distributions geometry: when the underlying distribution is contaminated by outliers, **the marginal constraints force OT** to transport outlying values, inducing an undesirable extra cost, which entails large changes in W_p.



$W_1 = 3.00,\ W_2 = 3.02,\ W^{(\lambda)} = 2.93$



$W_1 = 5.32,\ W_2 = 6.62$ and $W^{(\lambda)} = 3.31$

where $\lambda = 3$. Let's meet $W^{(\lambda)}$...

Robust Optimal Transport (ROBOT)

Robust OT (ROBOT) problem is defined in Mukherjee et al. (2021):

$$\underbrace{\min_{\gamma,s} \int \boxed{c(x,y)}\, \gamma(x,y)\mathrm{d}x\mathrm{d}y}_{\text{standard OT}} + \underbrace{\lambda \|s\|_{\mathrm{TV}}}_{\text{penalization}}$$

Robust OT (ROBOT) problem is defined in Mukherjee et al. (2021):

$$\begin{aligned}
&\underbrace{\min_{\gamma,s} \int \boxed{c(x,y)}\, \gamma(x,y)\mathrm{d}x\mathrm{d}y}_{\text{standard OT}} + \underbrace{\lambda \|s\|_{\mathrm{TV}}}_{\text{penalization}} \\
\text{s.t.} \quad & \int \gamma(x,y)\mathrm{d}y = \mu(x) + s(x) \geq 0 \\
& \int s(x)\mathrm{d}x = 0 \\
& \int \gamma(x,y)\mathrm{d}x = \nu(y),
\end{aligned} \tag{4}$$

where $\lambda > 0$ is a **regularization parameter**, which controls for the role of s. The latter introduces a **modification of the measure μ:**

Robust OT (ROBOT) problem is defined in Mukherjee et al. (2021):

$$\begin{aligned} \underbrace{\min_{\gamma,s} \int \boxed{c(x,y)}\,\gamma(x,y)\mathrm{d}x\mathrm{d}y}_{\text{standard OT}} + \underbrace{\lambda \|s\|_{\mathrm{TV}}}_{\text{penalization}} \\ \text{s.t.} \quad & \int \gamma(x,y)\mathrm{d}y = \mu(x) + s(x) \geq 0 \\ & \int s(x)\mathrm{d}x = 0 \\ & \int \gamma(x,y)\mathrm{d}x = \nu(y), \end{aligned} \tag{4}$$

where $\lambda > 0$ is a **regularization parameter**, which controls for the role of s. The latter introduces a **modification of the measure μ: The outlier is eliminated from the sample (no mass conservation, unbalanced OT).**

Robust OT (ROBOT) problem is defined in Mukherjee et al. (2021):

$$\begin{aligned} &\underbrace{\min_{\gamma,s} \int \boxed{c(x,y)}\, \gamma(x,y)\mathrm{d}x\mathrm{d}y}_{\text{standard OT}} + \underbrace{\lambda\|s\|_{\mathrm{TV}}}_{\text{penalization}} \\ \text{s.t.} \quad & \int \gamma(x,y)\mathrm{d}y = \mu(x) + s(x) \geq 0 \\ & \int s(x)\mathrm{d}x = 0 \\ & \int \gamma(x,y)\mathrm{d}x = \nu(y), \end{aligned} \tag{4}$$

where $\lambda > 0$ is a **regularization parameter**, which controls for the role of s. The latter introduces a **modification of the measure μ: The outlier is eliminated from the sample (no mass conservation, unbalanced OT).**

Remark

Mukherjee et al. (2021) prove that solving (4) is equivalent to

$$\inf \left\{ \int \boxed{c_\lambda(x,y)})\mathrm{d}\gamma(x,y) : \gamma \in \Gamma(\mu,\nu) \right\}, \tag{5}$$

which is similar to the original OT problem, but the cost function $c(x,y) = d(x,y)$ is replaced by $\boxed{c_\lambda = \min\{c, 2\lambda\}}$ that is bounded from above by 2λ.

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

$$W^{(\lambda)}(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int c_\lambda(x, y) \, \mathrm{d}\gamma(x, y) \right\} \tag{6}$$

is the Robust Wasserstein distance and it 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

- *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

- *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

- *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,*

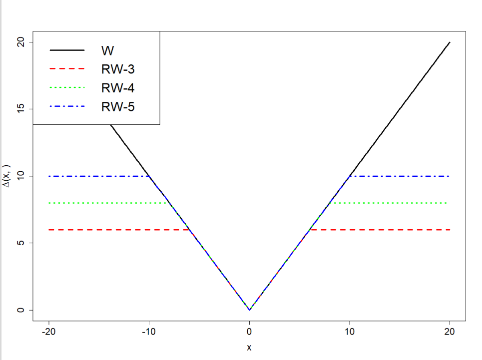
$$\hat{\theta}_n^\lambda = \arg\min W^{(\lambda)}(\hat{\mu}_n, \mu_\theta)$$

converges to

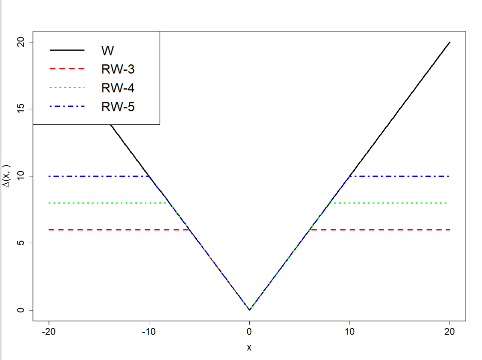
$$\theta_\star = \arg\min W^{(\lambda)}(\mu_\star, \mu_\theta)$$

- *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.*

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



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).

Using **synthetic data**, we illustrate the performance of MERWE considering the problem of estimation of a (location) parameter in the univariate setting. Specifically, we study the following settings:

- Finite moments (sum of log-normal r.v.s, with and w/o ε of contamination), for different sample sizes
- Infinite moments of different order (symmetric α-stable r.v.s with different values of α, with and w/o ε of contamination)

In all cases, we compare the MERWE (based on $W^{(\lambda)}$) to MEWE (based on the extant W_1). Our goal is two-fold: (1) illustrate the robustness of MERWE; (2) illustrate that the MERWE works even when the underlying generative model has infinite moments.

Finite moments:

SETTINGS	$n = 100$				$n = 200$				$n = 1000$			
	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	0.009	0.041	0.092	0.002	0.011	0.036	0.085	0.001	0.007
$\varepsilon = 0.1, \eta = 4$	0.035	0.089	0.001	0.012	0.029	0.096	0.001	0.015	0.013	0.098	≈ 0	0.017
$\varepsilon = 0.2, \eta = 1$	**0.071**	**0.157**	**0.007**	**0.028**	0.086	0.177	0.008	0.033	**0.081**	**0.172**	**0.006**	**0.030**
$\varepsilon = 0.2, \eta = 4$	0.046	0.204	0.003	0.045	0.034	0.202	0.001	0.042	0.017	0.194	≈ 0	0.038
$\varepsilon = 0$	**0.036**	**0.034**	**0.001**	**0.001**	0.022	0.021	≈ 0	≈ 0	**0.012**	**0.010**	≈ 0	≈ 0

Finite moments:

SETTINGS	$n = 100$				$n = 200$				$n = 1000$			
	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	0.009	0.041	0.092	0.002	0.011	0.036	0.085	0.001	0.007
$\varepsilon = 0.1, \eta = 4$	0.035	0.089	0.001	0.012	0.029	0.096	0.001	0.015	0.013	0.098	≈ 0	0.017
$\varepsilon = \mathbf{0.2}, \eta = \mathbf{1}$	**0.071**	**0.157**	**0.007**	**0.028**	0.086	0.177	0.008	0.033	**0.081**	**0.172**	**0.006**	**0.030**
$\varepsilon = 0.2, \eta = 4$	0.046	0.204	0.003	0.045	0.034	0.202	0.001	0.042	0.017	0.194	≈ 0	0.038
$\varepsilon = \mathbf{0}$	**0.036**	**0.034**	**0.001**	**0.001**	0.022	0.021	≈ 0	≈ 0	**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, $n = 200$*

Finite moments:

SETTINGS	$n=100$				$n=200$				$n=1000$			
	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	0.009	0.041	0.092	0.002	0.011	0.036	0.085	0.001	0.007
$\varepsilon=0.1, \eta=4$	0.035	0.089	0.001	0.012	0.029	0.096	0.001	0.015	0.013	0.098	≈ 0	0.017
$\boldsymbol{\varepsilon=0.2, \eta=1}$	**0.071**	**0.157**	**0.007**	**0.028**	0.086	0.177	0.008	0.033	**0.081**	**0.172**	**0.006**	**0.030**
$\varepsilon=0.2, \eta=4$	0.046	0.204	0.003	0.045	0.034	0.202	0.001	0.042	0.017	0.194	≈ 0	0.038
$\boldsymbol{\varepsilon=0}$	**0.036**	**0.034**	**0.001**	**0.001**	0.022	0.021	≈ 0	≈ 0	**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.*

Infinite moments:

SETTINGS	Cauchy				Stable ($\alpha = 0.5$)				Stable ($\alpha = 1.1$)			
	BIAS		MSE		BIAS		MSE		BIAS		MSE	
	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE	MERWE	MEWE
$\varepsilon = 0.1, \eta = 1$	0.084	1.531	0.010	3.627	0.087	3.178	0.011	13.730	0.089	0.658	0.011	1.029
$\varepsilon = 0.1, \eta = 4$	0.205	1.529	0.047	3.656	0.163	3.173	0.034	13.706	0.206	0.745	0.047	1.050
$\varepsilon = 0.2, \eta = 1$	**0.180**	**1.502**	**0.037**	**3.601**	0.170	3.155	0.036	12.838	**0.181**	**0.675**	**0.037**	**0.941**
$\varepsilon = 0.2, \eta = 4$	0.459	1.820	0.223	4.690	0.383	3.140	0.165	12.713	0.484	1.072	0.244	1.801
$\varepsilon = 0$	**0.045**	**1.550**	**0.003**	**3.740**	0.044	3.118	0.003	12.600	**0.041**	**0.612**	**0.002**	**0.893**

Remark

The MEWE has larger bias and MSE than the ones yielded by the MERWE. This aspect is particularly evident for the distributions with undefined first moment, namely the Cauchy distribution. If we increase α to 1.1, the absence of the second moment still entails a worse performance of MEWE wrt to the MERWE.

We propose RWGAN-1 and RWGAN-2, which are two RWGAN deep learning models: both approaches are based on dual version of ROBOT. 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:

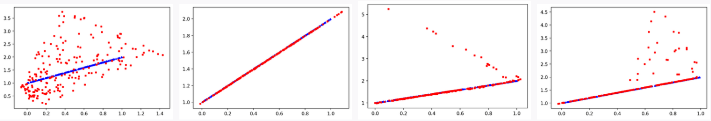
$$\begin{gathered} X_{i_1}^{(n)} \sim \mathrm{U}(0,1), X_{i_2}^{(n)} = X_{i_1}^{(n)} + 1, \\ X_i^{(n)} = (X_{i_1}^{(n)}, X_{i_2}^{(n)}), i = 1, 2, \ldots, n_1, \\ X_i^{(n)} = (X_{i_1}^{(n)}, X_{i_2}^{(n)} + \eta), i = n_1 + 1, n_1 + 2, \ldots, n, \end{gathered} \tag{7}$$

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.

WGAN	RWGAN-1	RWGAN-2	RWGAN-B

WGAN | RWGAN-1 | RWGAN-2 | RWGAN-B

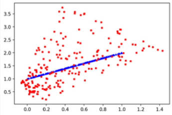
10%

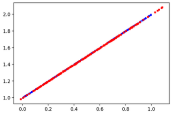


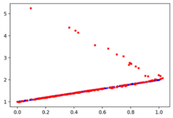
(a) $W^1 = 0.5864$ (b) $W^1 = 0.0514$ (c) $W^1 = 0.1560$ (d) $W^1 = 0.1771$

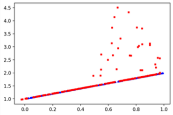
	WGAN	RWGAN-1	RWGAN-2	RWGAN-B

10%

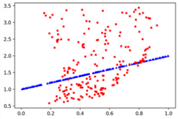

(a) $W^1 = 0.5864$

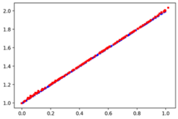

(b) $W^1 = 0.0514$

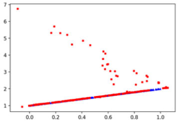

(c) $W^1 = 0.1560$

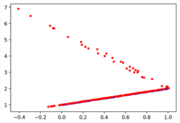

(d) $W^1 = 0.1771$

20%


(i) $W^1 = 0.5646$

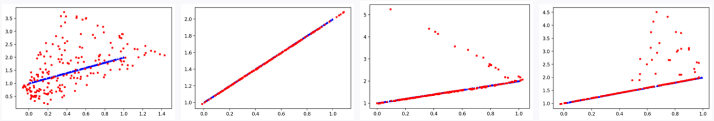

(j) $W^1 = 0.0470$


(k) $W^1 = 0.2938$


(l) $W^1 = 0.3229$

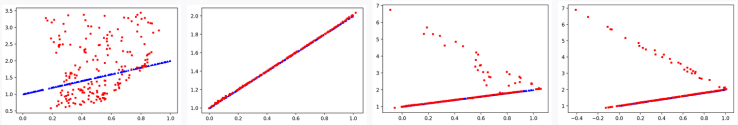
	WGAN	RWGAN-1	RWGAN-2	RWGAN-B

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-2 and RWGAN-B are able to generate data roughly consistent with the uncontaminated distribution, but they still produce some abnormal points when the proportion and size of outliers increase. RWGAN-1 performs better than its competitors and generates data that agree with the uncontaminated distribution, even when the proportion and size of outliers are large.

Take home message

In the paper:

- 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 $\lambda > 0$

In the paper:

- 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

In the paper:

- 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

In the paper:

- 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

In the paper:

- 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 **machine learning problems** (generative adversarial networks and domain adaptation)

In the paper:

- 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 **machine learning problems** (generative adversarial networks and domain adaptation)
- We illustrate the applicability of ROBOT for RWGAN using the **Fashion-MNIST dataset**