When OT meets MoM: Robust estimation of Wasserstein Distance

Guillaume Staerman[1] **Pierre Laforgue**[2] **Pavlo Mozharovskyi**[1] **Florence d'Alché-Buc**[1]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris
[2] DSRC & Dept. of Computer Science, Università degli Studi di Milano, Italy

Abstract

Originated from Optimal Transport, the Wasserstein distance has gained importance in Machine Learning due to its appealing geometrical properties and the increasing availability of efficient approximations. It owes its recent ubiquity in generative modelling and variational inference to its ability to cope with distributions having non overlapping support. In this work, we consider the problem of estimating the Wasserstein distance between two probability distributions when observations are polluted by outliers. To that end, we investigate how to leverage a Medians of Means (MoM) approach to provide robust estimates. Exploiting the dual Kantorovitch formulation of the Wasserstein distance, we introduce and discuss novel MoM-based robust estimators whose consistency is studied under a data contamination model and for which convergence rates are provided. Beyond computational issues, the choice of the partition size, *i.e.,* the unique parameter of theses robust estimators, is investigated in numerical experiments. Furthermore, these MoM estimators make Wasserstein Generative Adversarial Network (WGAN) robust to outliers, as witnessed by an empirical study on two benchmarks CIFAR10 and Fashion MNIST.

1 Introduction

Computing distances between probability distributions has become a central question in numerous modern Machine Learning applications, ranging from generative modeling to clustering. Optimal Transport (OT) [1, 2] offers an appealing and insightful tool to solve this problem, building upon the Wasserstein distance. Given two probability distributions, the latter is defined in terms of the solution to the Monge-Kantorovich optimal mass transportation problem. Interestingly, it relies on a ground distance between points to build a distance between probability distributions [3]. For that reason, the Wasserstein distance stands out from the divergences usually exploited in generative modeling, like the f-divergences [4, 5], by its ability to take into account the underlying geometry of the space, capturing the difference between probability distributions even when they have non-overlapping supports. This appealing property has been successfully exploited in Generative Adversarial Networks (GANs) [6, 7, 8], as well as in Variational Autoencoders (VAEs) [9], where the Wasserstein distance can advantageously replace an f-divergence as the loss function. Many other applications [10, 11, 12] rely on the entropic-regularized approximations introduced by [13], which has considerably alleviated the inherent computational complexity of the Wasserstein distance in the discrete case, by drawing on the Sinkhorn-Knopp algorithm. A common feature to almost all these works is that the Wasserstein distance is estimated from finite samples. While this problem has long been theoretically studied under the i.i.d. assumption [14, 15, 16], it has never been tackled through the lens of robustness to outliers, a crucial issue in Reliable Machine Learning. Indeed, data is nowadays collected at a large scale in unmastered acquisition conditions, and through a large variety of devices and platforms. The resulting datasets often present undesirable influential observations, whether they are errors or rare observations. The presence of corrupted data may heavily damage the quality of estimators, calling for dedicated methods such as JS/TV-GANs [17] in the particular case of robust shift-parameter estimation, Robust Divergences in variational inference [18], or more general tools from robust statistics [19].

The aim of this work is to propose outliers-robust estimators of the Wasserstein distance, and illustrate their application in generative modeling. To that end, we explore how to combine a Median-of-Means approach

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

with Optimal Transport. The Median-of-Means (MoM) is a robust mean estimator firstly introduced in complexity theory during the 1980s [20, 21, 22]. Following the seminal deviation study by [23], MoM has lately witnessed a surge of interest, mainly due to its attractive sub-gaussian behavior, under the sole assumption that the underlying distribution has finite variance [24]. Originally devoted to scalar random variables, MoM has notably been extended to random vectors [25, 26, 27] and U-statistics [28, 29]. As a natural alternative to the empirical mean, MoM has become the cornerstone of several robust learning procedures in heavy-tailed situations, including bandits [30] and MoM-tournaments [31]. A more recent line of work now focuses on MoM's ability to deal with outliers. Aside from concentration results in a contaminated context [32, 33], it has yielded promising applications in robust mean embedding [34], and the more general MoM-minimization framework [35].

In this paper, we introduce and study outliers-robust estimators of the Wasserstein distance based on the MoM methodology. Our contribution is threefold:

- Focusing on the Kantorovich-Rubinstein duality [36], we present three novel MoM-based estimators, leveraging in particular Medians of U-statistics (MoU). In the realistic setting of contaminated data, we show their strong consistency, and provide non-asymptotic bounds as well.
- We propose a dedicated algorithm to compute these three estimators in practice. Applied on a parametric family of Lipschitz functions, *e.g.* neural networks with clipped weights, it performs a MoM/MoU gradient descent algorithm. A sensitivity analysis of the unique parameter of these estimators is also provided throught numerical experiments on toy datasets.
- We robustify WGANs (w.r.t. outliers) using a MoM-based estimator as loss function. We show the benefits of this approach through convincing numerical results on two contaminated well known benchmarks: CIFAR10 and Fashion MNIST.

2 Background and preliminaries

Before introducing the problem to be addressed, we recall some key notions about the Wasserstein distance and the Medians-of-Means estimator. Let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of $\mathbb{R}^d$, for some $d \in \mathbb{N}^*$. We denote by $\mathcal{M}_+^1(\mathcal{X})$ the space of all probability measures on $\mathcal{X}$, and consider two distributions μ and ν from $\mathcal{M}_+^1(\mathcal{X})$ and $\mathcal{M}_+^1(\mathcal{Y})$. For any $K \in \mathbb{N}^*$, the median of $\{z_1, \dots, z_K\} \in \mathbb{R}^K$ is denoted by $\underset{1\leq k\leq K}{\text{med}}\{z_k\}$.

2.1 Wasserstein Distance

Given $p \in [1, \infty)$, the Wasserstein distance of order p between two arbitrary measures μ and ν is defined through the resolution of the Monge-Kantorovitch mass transportation problem [1, 3]:

$$\mathcal{W}_p(\mu,\nu) = \min_{\pi \,\in\, \mathcal{U}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^p d\pi(x\times y) \right)^{1/p}, \tag{1}$$

where $\mathcal{U}(\mu,\nu) = \{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) : \int \pi(x,y)dy = \mu(x); \int \pi(x,y)dx = \nu(y)\}$ is the set of joint probability distributions with marginals μ and ν. In the remainder of this paper, we focus on the Wasserstein of order 1, $\mathcal{W}_1$, omitting the subscript 1 for notation simplicity. By the dual Kantorovich-Rubinstein formulation [36], with $\mathcal{B}_L$ the unit ball of the Lipschitz functions space, a useful rewriting of the 1-Wasserstein distance is:

$$\mathcal{W}(\mu,\nu) = \sup_{\Phi\in\mathcal{B}_L} \mathbb{E}_\mu\left[\Phi(X)\right] - \mathbb{E}_\nu\left[\Phi(Y)\right]. \tag{2}$$

Of particular interest is the problem of estimating the Wasserstein distance between μ and ν given a finite number of observations. The usual assumption is to rely upon two samples $\mathbf{X} = \{X_1, \dots, X_n\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_m\}$, composed of i.i.d. realizations drawn respectively from μ and ν. The corresponding empirical distributions denoted by $\hat{\mu}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, and $\hat{\nu}_m = (1/m)\sum_{j=1}^m \delta_{Y_j}$. The natural questions are then: *how to compute the estimator $\mathcal{W}(\hat{\mu}_n, \hat{\nu}_m)$, and does it converge towards $\mathcal{W}(\mu,\nu)$?* In the dual formulation (2), computing $\mathcal{W}(\hat{\mu}_n, \hat{\nu}_m)$ is equivalent to replace the expectations with empirical means. The unit ball of Lipschitz functions can be replaced with a parameterized family of Lipschitz functions, more amenable for learning when $\mathcal{W}(\hat{\mu}_n, \hat{\nu}_m)$ is used as a loss function, see *e.g.* Wasserstein GANs [7]. From the theoretical side, a substantial number of works have studied the convergence of $\mathcal{W}(\hat{\mu}_n, \hat{\nu}_m)$ under the i.i.d. setting described above. Statistical rates of convergence of the original OT problem are known to be slow rates with respect to the dimension d of the input space, *i.e.* they are of order $O(n^{-1/d})$ [14, 15, 16, 37, 38].

2.2 Median-of-Means

Given an i.i.d. sample $\mathbf{X} = \{X_1, \dots, X_n\}$ drawn from μ, the Median-of-Means (MoM) is an estimator of $\mathbb{E}_\mu[X]$ built as follows. First, choose $K_{\mathbf{X}} \leq n$, and partition $\mathbf{X}$ into $K_{\mathbf{X}}$ disjoint blocks $\mathcal{B}_1^{\mathbf{X}}, \dots, \mathcal{B}_{K_{\mathbf{X}}}^{\mathbf{X}}$ of size $B_{\mathbf{X}} = n/K_{\mathbf{X}}$. If n cannot be divided by $K_{\mathbf{X}}$, some observations may be removed. Then, empirical means are computed on each of the $K_{\mathbf{X}}$ blocks. The estimator returned is finally the median of the empirical means thus computed. For a function $\Phi\colon \mathcal{X} \to \mathbb{R}$, the MoM

estimator of $\mathbb{E}_\mu[\Phi(X)]$ is then formally given by:

$$\mathrm{MoM}_{\mathbf{X}}[\Phi] = \underset{1\leq k\leq K_{\mathbf{X}}}{\mathrm{med}} \Big\{ \frac{1}{B_{\mathbf{X}}} \sum_{i\in\mathcal{B}_k^{\mathbf{X}}} \Phi(X_i) \Big\}. \qquad (3)$$

This estimator provides an attractive alternative to the sample mean $\overline{\Phi}_{\mathbf{X}} = (1/n)\sum_{i=1}^n \Phi(X_i)$ for robust learning. Indeed, it has been shown to (i) exhibit a sub-Gaussian behavior under only a finite variance assumption, making it particularly suited to heavy-tailed distributions, and (ii) be non-sensitive to outliers. MoM also nicely adapts to multisample U-statistics of arbitrary degrees [39]. Indeed, assume that one is interested in estimating $\mathbb{E}_{\mu\otimes\nu}[h(X,Y)]$, for some kernel $h\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. Given the samples $\mathbf{X} = \{X_1,\dots,X_n\}$ and $\mathbf{Y} = \{Y_1,\dots,Y_m\}$, a natural idea then consists in partitioning both $\mathbf{X}$ and $\mathbf{Y}$ into $\mathcal{B}_1^{\mathbf{X}},\dots,\mathcal{B}_{K_{\mathbf{X}}}^{\mathbf{X}}$ and $\mathcal{B}_1^{\mathbf{Y}},\dots,\mathcal{B}_{K_{\mathbf{Y}}}^{\mathbf{Y}}$ respectively, with $K_{\mathbf{Y}}\leq m$, and $B_{\mathbf{Y}} = m/K_{\mathbf{Y}}$. One may then compute U-statistics on each pair of blocks (k,l) for $k\leq K_{\mathbf{X}}$ and $l\leq K_{\mathbf{Y}}$, and return the median of the $K_{\mathbf{X}}\times K_{\mathbf{Y}}$ U-statistics. However, this construction introduces dependence between the base estimators, making the theoretical study more difficult. An alternative then consists in choosing $K_{\mathbf{X}} = K_{\mathbf{Y}} = K$, and considering only the diagonal blocks (see Figure 1c). These two estimators are referred to as (diagonal) Median-of-U-statistics (MoU), and using $\mathcal{B}_{k,l}^{\mathbf{XY}}$ to denote the block of tuples (X_i,Y_j) such that $X_i\in\mathcal{B}_k^{\mathbf{X}}$ and $Y_j\in\mathcal{B}_l^{\mathbf{Y}}$, they are formally given by:

$$\mathrm{MoU}_{\mathbf{XY}}[h] = \underset{\substack{1\leq k\leq K_{\mathbf{X}}\\ 1\leq l\leq K_{\mathbf{Y}}}}{\mathrm{med}} \Big\{ \frac{1}{B_{\mathbf{X}}B_{\mathbf{Y}}} \sum_{(i,j)\in\mathcal{B}_{k,l}^{\mathbf{XY}}} h(X_i,Y_j) \Big\},$$

$$\mathrm{MoU}^{\mathrm{diag}}_{\mathbf{XY}}[h] = \underset{1\leq k\leq K}{\mathrm{med}} \Big\{ \frac{1}{B_{\mathbf{X}}B_{\mathbf{Y}}} \sum_{(i,j)\in\mathcal{B}_{k,k}^{\mathbf{XY}}} h(X_i,Y_j) \Big\}.$$

3 When Wasserstein meets MoM

In this section, we investigate how MoM estimators can be leveraged to define and analyze new estimators of $\mathcal{W}(\mu,\nu)$ that exhibit strong theoretical guarantees in presence of outliers. In order to assess robustness, we place ourselves in the realistic $\mathcal{O}\cup\mathcal{I}$ framework, see *e.g.* [19, 40], devoted to data contamination. In this setting, the i.i.d. assumption is relaxed, and the following assumption is instead adopted.

Assumption 1 *Sample* $\mathbf{X}$ *is polluted with* $n_{\mathcal{O}} < n/2$ *(possibly adversarial) outliers. The remaining* $n - n_{\mathcal{O}}$ *points are informative data, or* inliers*, independently distributed according to* μ*. A similar assumption is made on* $\mathbf{Y}$*, which is supposed to contain* $m_{\mathcal{O}} < m/2$ *arbitrary outliers, and* $m - m_{\mathcal{O}}$ *inliers drawn from* ν*. Inliers are assumed to lie in a compact set* $\mathcal{K}\subset\mathbb{R}^d$*. In contrast, no assumption is made on the outliers, that may not be bounded. The proportions of outliers in samples* $\mathbf{X}$ *and* $\mathbf{Y}$ *are denoted by* $\tau_{\mathbf{X}} = n_{\mathcal{O}}/n$ *and* $\tau_{\mathbf{Y}} = m_{\mathcal{O}}/m$ *respectively.*

3.1 MoM and MoU-based estimators

Starting from the expression of the dual expression (2), we observe that it can be considered with a two-fold perspective. The first one consists in considering the Wasserstein distance as the supremum of the difference between two expected values. The second one, obtained by linearity of the expectation, rather regards $\mathcal{W}(\mu,\nu)$ as the supremum of single expected values, but taken with respect to the tuple (X,Y), and associated to the kernel: $h_\phi\colon (X,Y)\mapsto\phi(X)-\phi(Y)$.

Although quite elementary at first sight, this two-fold perspective gains complexity when applied to the empirical distributions $\hat{\mu}_n$ and $\hat{\nu}_m$. Indeed, following the first perspective, the natural estimator obtained is the supremum of the differences between two empirical averages, while the second one leads to the supremum of 2-samples U-statistics of degrees $(1,1)$ and kernels h_ϕ. So far, both points of view are strictly equivalent by linearity of the expectation and the empirical mean. However, this equivalence breaks down as soon as non-linearities are introduced, through MoM-like estimators for instance. We therefore introduce three distinct estimators of $\mathcal{W}(\mu,\nu)$, that differ upon which estimator of Section 2.2 is used.

Definition 2 *We define the Median-of-Means and the Median-of-U-statistics estimators of the* 1*-Wasserstein distance as follows:*

$$\mathcal{W}_{\mathrm{MoM}}(\hat{\mu}_n,\hat{\nu}_m) = \sup_{\phi\in\mathcal{B}_L} \{\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathrm{MoM}_{\mathbf{Y}}[\phi]\},$$

$$\mathcal{W}_{\mathrm{MoU}}(\hat{\mu}_n,\hat{\nu}_m) = \sup_{\phi\in\mathcal{B}_L} \{\mathrm{MoU}_{\mathbf{XY}}[h_\phi]\},$$

$$\mathcal{W}_{\mathrm{MoU-diag}}(\hat{\mu}_n,\hat{\nu}_m) = \sup_{\phi\in\mathcal{B}_L} \{\mathrm{MoU}^{\mathrm{diag}}_{\mathbf{XY}}[h_\phi]\}.$$

While $\mathcal{W}_{\mathrm{MoM}}$ relies on the difference between individual median blocks, $\mathcal{W}_{\mathrm{MoU\text{-}diag}}$ considers the median over all possible combinations of blocks between $\mathbf{X}$ and $\mathbf{Y}$. As an intermediate step, $\mathcal{W}_{\mathrm{MoU\text{-}diag}}$ looks after diagonal blocks only. The latter formulation is used in [34] to derive robust mean embedding and Maximum Mean Discrepancy estimators. The theoretical analysis is made simpler by the independence between the blocks, but the estimator suffers from an increased variance due to the important loss of information, see Figure 1c and [28]. It should be noticed however that $\mathcal{W}_{\mathrm{MoU\text{-}diag}}$ enjoys a much lower computational cost in practice.

One elegant way to combine both benefits, *i.e.* small loss of information and low computational cost, is to

consider randomized blocks [29]. Instead of partitioning the dataset, this method builds blocks by sampling them independently through simple Sampling Without Rejection (SWoR). One consequence is the possibility for the randomized blocks to overlap (see $\mathscr{B}_1^{\mathbf{X}}$, $\mathscr{B}_2^{\mathbf{X}}$, $\mathscr{B}_3^{\mathbf{X}}$ in Figure 1a), making the estimator's concentration analysis more difficult. Nevertheless, guarantees similar to that of MoM have been established (up to constants), and the extension to 2-sample U-statistics built on randomized blocks allows for a better exploration of the grid than through MoU^{diag}, see Figure 1e. However, despite the possibility to reach every part of the grid, the exploration scheme illustrated in Figure 1e have a fixed structure (*e.g.* always 3 cells per column, 4 cells per row). The *totally free* alternative, as depicted in Figure 1f, consists in sampling directly from the pairs of observations, which generates incomplete U-statistics. If no theoretical guarantees have been established for this extension due to the complex replication setting between blocks, it still benefits from good empirical results [29], consistent with the grid covering it allows.

Another important question to be addressed is: *how to handle the non-differentiability introduced by the median operator?* Indeed, the Wasserstein distance often acts as a loss function, *e.g.* in generative modeling (VAEs, GANs), and optimizing through a MoM/MoU-based criterion then becomes crucial. One answer is to uses a MoM-gradient descent algorithm [35]. It consists in performing a mini-batch gradient step based on the median block. In order to avoid local minima, authors propose shuffle the partition at each step of the descent, leading to the minimization of an expected MoM loss (w.r.t. the shuffling) that is more stable. Notice that this method goes beyond random partitions, and easily adapts to the randomized extensions discussed above.

3.2 Theoretical guarantees

We now establish the statistical guarantees satisfied by the estimators introduced in Definition 2 under Assumption 1. First notice that if $\hat{\mu}_{\text{MoM}}$ denotes with a language abuse the *measure* such that for all application $\phi\colon \mathbb{R}^d \to \mathbb{R}$ it holds $\mathbb{E}_{\hat{\mu}_{\text{MoM}}}[\phi] = \text{MoM}_{\mathbf{X}}[\phi]$, it is direct to see that $\mathcal{W}_{\text{MoM}}(\hat{\mu}_n, \hat{\nu}_m) = \mathcal{W}(\hat{\mu}_{\text{MoM}}, \hat{\nu}_{\text{MoM}})$. Then, it holds $\mathcal{W}_{\text{MoM}}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{W}(\mu, \nu) \leq \mathcal{W}(\mu, \hat{\mu}_{\text{MoM}}) + \mathcal{W}(\hat{\nu}_{\text{MoM}}, \nu)$, and one may only focus on the theoretical guarantees of the right-hand side terms. Before stating our main results, we need an additional assumption on the numbers of outliers $n_{\mathcal{O}}$ and $m_{\mathcal{O}}$, which are assumed to grow sub-linearly with respect to n and m.

Assumption 3 *There exist $C_{\mathcal{O}} \geq 1$ and $0 \leq \alpha_{\mathcal{O}} < 1$ such that $n_{\mathcal{O}} \leq C_{\mathcal{O}}^2\, n^{\alpha_{\mathcal{O}}}$ and $m_{\mathcal{O}} \leq C_{\mathcal{O}}^2\, m^{\alpha_{\mathcal{O}}}$.*

We start by an asymptotic result establishing the strong consistency of estimators in Definition 2. It highlights the different outlier configurations allowed through conditions on the proportions of outliers $\tau_{\mathbf{X}}$ and $\tau_{\mathbf{Y}}$.

Proposition 4 *Suppose that samples $\mathbf{X}$ and $\mathbf{Y}$ satisfy Assumptions 1 and 3. Then, choosing $K_{\mathbf{X}} = \lceil \sqrt{2\tau_{\mathbf{X}}}\, n \rceil$, it holds:*

$$\mathcal{W}(\hat{\mu}_{\text{MoM}}, \mu) \xrightarrow{a.s} 0.$$

If moreover $\tilde{\tau} := \tau_{\mathbf{X}} + \tau_{\mathbf{Y}} - \tau_{\mathbf{X}}\tau_{\mathbf{Y}} < 1/2$, then choosing $K_{\mathbf{X}} = \lceil \sqrt{2\tilde{\tau}}\, n \rceil$ and $K_{\mathbf{Y}} = \lceil \sqrt{2\tilde{\tau}}\, m \rceil$, it holds:

$$\left| \mathcal{W}_{\text{MoU}}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{W}(\mu, \nu) \right| \xrightarrow{a.s} 0.$$

If finally $\tau_{\mathbf{X}} + \tau_{\mathbf{Y}} < 1/2$ and $n = m$, then choosing $K_{\mathbf{X}} = K_{\mathbf{Y}} = \lceil \sqrt{2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})}\, n \rceil$, it holds:

$$\left| \mathcal{W}_{\text{MoU}-\text{diag}}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{W}(\mu, \nu) \right| \xrightarrow{a.s} 0.$$

The key argument in the proof of Proposition 4 consists in converting the convergence of the different estimators into the convergences of blocks containing no outliers. Numbers of blocks $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$ are chosen such that (i) such blocks are always in majority, and (ii) their sizes $n/K_{\mathbf{X}}$ and $m/K_{\mathbf{Y}}$ go to infinity as n and m go to infinity. Any other choice of $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$ that satisfies this two conditions also ensures convergence. If the outliers proportions are unknown, building $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$ from upper bounds of $\tau_{\mathbf{X}}$ and $\tau_{\mathbf{Y}}$ thus does not impact Proposition 4. The conditions on $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$ also constrain the proportions of outliers admitted, as illustrated in Figure 1d. The assumption $n = m$ for $\mathcal{W}_{\text{MoU}-\text{diag}}$ is necessary to be able to build a majority of sane blocks. Our next proposition now investigates the nonasymptotic behavior of the proposed estimators.

Proposition 5 *Suppose that samples $\mathbf{X}$ and $\mathbf{Y}$ satisfy Assumption 1, and define $\Gamma\colon \tau \mapsto \sqrt{1+\sqrt{2\tau}}/\sqrt{1-2\tau}$. Then, for all $\delta \in]0, \exp(-4n\sqrt{2\tau_{\mathbf{X}}})]$, choosing $K_{\mathbf{X}} = \lceil \sqrt{2\tau_{\mathbf{X}}}\, n \rceil$, it holds with probability at least $1-\delta$:*

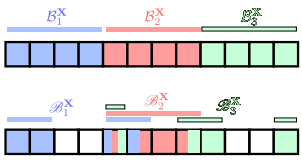
$$\mathcal{W}(\hat{\mu}_{\text{MoM}}, \mu) \leq \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}} + C_2(\tau_{\mathbf{X}}) \sqrt{\frac{\log(1/\delta)}{n}},$$

with $C_1(\tau) = 2 + C_L C_2(\tau)$, $C_2(\tau) = 4\ \text{diam}(\mathcal{K})\ \Gamma(\tau)$, and C_L a universal constant depending only on $\mathcal{B}_L$.

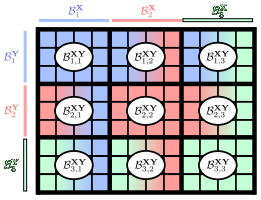
If furthermore $\tau_{\mathbf{X}} + \tau_{\mathbf{Y}} < 1/2$ and $n = m$, then for all $\delta \in]0, \exp(-4n\sqrt{2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})})]$, choosing $K_{\mathbf{X}} = K_{\mathbf{Y}} = \lceil \sqrt{2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})}\, n \rceil$, it holds with probability at least $1-\delta$:

$$\left| \mathcal{W}_{\text{MoU}-\text{diag}}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{W}(\mu, \nu) \right| \leq \frac{2C_1(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})}{n^{1/(d+2)}} + 2C_2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}}) \sqrt{\frac{\log(1/\delta)}{n}}.$$

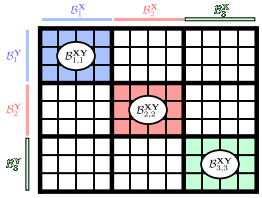
The proof derives from concentration results established in [33], combined with a generic chaining argument. It



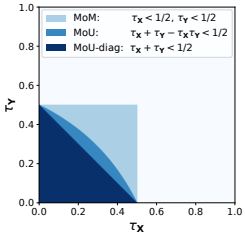
(a) 1D standard and randomized MoM



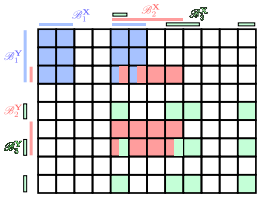
(b) $\text{MoU}_{\mathbf{XY}}$



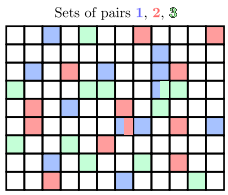
(c) $\text{MoU}^{\text{diag}}_{\mathbf{XY}}$



(d) Admitted proportion of outliers



(e) MoU based on randomized blocks



(f) MoU based on randomized pairs

Figure 1: Sampling strategies to build MoM and MoU, as well as admitted proportion of outliers.

should be noticed that constant $C_2(\tau_{\mathbf{X}})$ explodes as $\tau_{\mathbf{X}}$ goes to $1/2$, which is expected: the more outliers, the more difficult it is to estimate $\mathcal{W}(\mu,\nu)$. We also stress that the dependence in $1/\sqrt{1-2\tau_{\mathbf{X}}}$ is better than the $1/(1-2\tau_{\mathbf{X}})^{3/2}$ term exhibited in [34]. Integrating the deviation probabilities of Proposition 5 and using Assumption 3, we finally obtain our main theorem, that provides a nonasymptotic control on the expected value of our estimators deviations from $\mathcal{W}(\mu,\nu)$.

Theorem 6 *Suppose that samples* $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy Assumptions 1 and 3, and recall the notation used in Proposition 5. Let* $\beta \in [0,1]$*, then for all* n *such that* $n^{\frac{1}{d+2}+\frac{1-\beta}{2}} \geq C_1(\tau_{\mathbf{X}})/(2C_2(\tau_{\mathbf{X}})(2\tau_{\mathbf{X}})^{\frac{1}{4}})$*, it holds:*

$$\mathbb{E}\left[\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu)\right] \;\leq\; \frac{\kappa_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}} + \frac{\kappa_2(\tau_{\mathbf{X}})}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\kappa_3(\tau_{\mathbf{X}})}{n^{\beta/2}},$$

with $\kappa_1(\tau) = C_1(\tau)$*,* $\kappa_2(\tau) = 2C_{\mathcal{O}}C_2(\tau)(2/\tau)^{1/4}$*, and* $\kappa_3(\tau) = \sqrt{\pi}C_2(\tau)/2$*.*

Of course, the above bound only makes sense if $\beta > \alpha_{\mathcal{O}}$*. In particular, if* $\alpha_{\mathcal{O}} \leq d/(d+2)$*, setting* $\beta = 1$ *gives that for all* n *such that* $n^{\frac{1}{d+2}} \geq C_1(\tau_{\mathbf{X}})/(2C_2(\tau_{\mathbf{X}})(2\tau_{\mathbf{X}})^{\frac{1}{4}})$*, with the notation* $\kappa = \kappa_1 + \kappa_2 + \kappa_3$*, it holds:*

$$\mathbb{E}\left[\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu)\right] \;\leq\; \kappa(\tau_{\mathbf{X}})\; n^{-1/(d+2)}.$$

If furthermore $\tau_{\mathbf{X}}+\tau_{\mathbf{Y}} < 1/2$ *and* $n = m$*, then for all* n *s.t.* $n^{\frac{1}{d+2}} \geq C_1(\tau_{\mathbf{X}}+\tau_{\mathbf{Y}})/(2C_2(\tau_{\mathbf{X}}+\tau_{\mathbf{Y}})(2(\tau_{\mathbf{X}}+\tau_{\mathbf{Y}}))^{\frac{1}{4}})$*, with the notation* $\kappa' = 2\kappa_1 + 2\sqrt{2}\kappa_2 + 2\kappa_3$*, it holds:*

$$\mathbb{E}\left|\mathcal{W}_{\text{MoU}-\text{diag}}(\hat{\mu}_n,\hat{\nu}_m) - \mathcal{W}(\mu,\nu)\right| \leq\; \kappa'(\tau_{\mathbf{X}}+\tau_{\mathbf{Y}})\; n^{-1/(d+2)}.$$

Theorem 6 highlights that the estimators proposed in Definition 2 remarkably resist to the presence of outliers in the training datasets. The price to pay is a slightly slower rate of order $O(n^{-1/(d+2)})$, that becomes equivalent in high dimension – the usual setting of Optimal Transport – to the standard $O(n^{-1/d})$ rate. Interestingly, the dependence in the outliers growing rate $\alpha_{\mathcal{O}}$ is made explicit, and is in line with expectations (see below). Unfortunately, the dependency between the blocks makes the nonasymptotic analysis harder for $\mathcal{W}_{\text{MoU}}$ and the computationally cheap randomized extensions discussed in Section 3.1. This theoretical challenge is left for future work. We stress that there is no *median-of-means miracle*. If the number of blocks allows to cancel the outliers impact, the statistical performance then scales with the block size, *i.e.* as $1/\sqrt{B_{\mathbf{X}}} = \sqrt{K_{\mathbf{X}}/n}$. Since $K_{\mathbf{X}}$ is roughly $2n_{\mathcal{O}}$, this means a $\sqrt{n_{\mathcal{O}}/n}$ rate. So if one allows $n_{\mathcal{O}}$ to grow proportionally to n, the bound becomes vacuous. To get guarantees improving with n, we thus need $n_{\mathcal{O}}$ to scale as $n^{\alpha_{\mathcal{O}}}$, for some $\alpha_{\mathcal{O}} < 1$, and the resulting rate is $n^{(1-\alpha_{\mathcal{O}})/2}$, as found in Theorem 6. We finally point out that the condition on n ensures $\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu) \geq C_1(\tau_{\mathbf{X}})/n^{1/(d+2)} - C_2(\tau_{\mathbf{X}})\sqrt{\log(1/\delta)/n^{\beta}}$, as the right hand side is negative while $\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu)$ is positive by construction. A less stringent condition might be derived, using *e.g.* the nature of the functions in $\mathcal{B}_L$.

Remark 7 *The unique property of the Wasserstein distance we used, compared to other Integral Probability Metrics (IPMs) [41], is the way to bound the entropy of the unit ball of Lipschitz functions. The present analysis can thus be extended in a direct fashion to any other IPM that has finite entropy.*

4 MoM-based estimators in practice

In this section, we first propose a novel algorithm to approximate the MoM/MoU-based estimators using

neural networks and provide an empirical study of its behaviour on two toy datasets. Then, we show how to robustify Wasserstein-GANs and present MoMWGAN, a MoM-based variant of GAN, which is evaluated on two well-known image benchmarks.

4.1 Approximation algorithm

As show in Section 3, MoM/MoU-based estimation of the Wasserstein distance offers a robust alternative to the classical empirical estimator of $\mathcal{W}$. Indeed, the empirical estimator of $\mathcal{W}$ would not converge towards the target in the $\mathcal{O} \cup \mathcal{I}$ framework. The proposed estimators are consistent and have convergence rates of order $O(n^{-1/(d+2)})$ with the $\mathcal{O} \cup \mathcal{I}$ framework.These convergence rates are similar, when d is not too small, to those of the empirical estimator of $\mathcal{W}$ in a non-contaminated setting. Nevertheless, the question of computing these estimators raises two major difficulties: (i) the optimization over the unit ball of Lipschitz functions is intractable, which is a difficulty common to the approximation of the standard Wassertein Distance, and (ii) the non-differentiability of the median-based loss. The first issue is well known of the practioners of the Wasserstein distance who usually prefer to rely on its primal definition with an entropy-based regularization [13]. However, learning algorithms devoted to Wasserstein GANs overcome this by weight clipping [7] or gradient penalization [8] to impose to the GAN a Lipchitz constraint. Similarly we propose here to limit Φ to be a neural network with similar constraints on weights to ensure its M-Lipschitzianity. This enables to approximate the Wasserstein distance up to a (unknown) multiplicative coefficient M.To overpass (ii), one can adopt MoM/MoU gradient descent. Exploited in the context of robust classification [35], using MoM/MoU gradient descent has been proved to be equivalent to minimize the expectation over the sampling strategy of blocks of $\mathcal{W}_{\text{MoM}}, \mathcal{W}_{\text{MoU-diag}}$ and $\mathcal{W}_{\text{MoU}}$. Combining these techniques, we design novel algorithms to compute approximations of the proposed estimators: $\widetilde{\mathcal{W}}_{\text{MoM}}$ (see Algorithm 1), $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$ (see the Supplementary Material).

4.2 Empirical study

We denote I_2 the identity matrix of dimension 2 and $\mathbf{v}$, the vector $(v, v)^\top$ with $v \in \mathbb{R}$.

Toy datasets. Two simulated datasets in 2D space with different kinds of anomalies are used in the experiments. The random vectors X_1 and X_2 are chosen to be distributed according a mixture of a standard Gaussian distribution and an "anomaly" distribution, respectively $\mathcal{A}_1$ and $\mathcal{A}_2$ defined as follows. $\mathcal{A}_1$ is the uniform distribution $\mathcal{U}[-\mathbf{50}, \mathbf{50}]$ that mimics *isolated outliers* while $\mathcal{A}_2$ is the standard Cauchy distribution shifted by 25, defined to mimic *aggregate outliers* (see *e.g.* [42]). The random vector Y is Gaussian with $Y \sim \mathcal{N}(\mathbf{5}, I_2)$, Datasets $\mathcal{D}_1 = (\mathbf{X}_1, \mathbf{Y})$ and $\mathcal{D}_2 = (\mathbf{X}_2, \mathbf{Y})$ contain 500 independent and identical copies of (X_1, Y), (X_2, Y) respectively, with the same proportion of outliers $\tau_{\mathbf{X}} = \tau_{\mathbf{Y}}$.

Algorithm 1 Approximation of $\mathcal{W}_{\text{MoM}}(\mathbf{X}, \mathbf{Y})$.

Initialization: η, the learning rate. c, the clipping parameter. w_0, the initial weigths. $K_{\mathbf{X}}, K_{\mathbf{Y}}$ the number of blocks for $\mathbf{X}$ and $\mathbf{Y}$.

1: **for** $t = 0, \dots, n_{\text{iter}}$ **do**
2: Sample $K_{\mathbf{X}}$ disjoint blocks $\mathcal{B}_1^{\mathbf{X}}, \dots, \mathcal{B}_{K_{\mathbf{X}}}^{\mathbf{X}}$ and $K_{\mathbf{Y}}$ disjoint blocks $\mathcal{B}_1^{\mathbf{Y}}, \dots, \mathcal{B}_{K_{\mathbf{Y}}}^{\mathbf{Y}}$ from a sampling scheme
3: Find both median blocks $\mathcal{B}_{med}^{\mathbf{X}}$ and $\mathcal{B}_{med}^{\mathbf{Y}}$
4:
$$G_w \leftarrow \lfloor K_{\mathbf{X}}/n \rfloor \sum_{i \in \mathcal{B}_{med}^{\mathbf{X}}} \nabla_w \phi_w(X_i) - \lfloor K_{\mathbf{Y}}/m \rfloor \sum_{j \in \mathcal{B}_{med}^{\mathbf{Y}}} \nabla_w \phi_w(Y_j))$$
5: 7.1 $w \leftarrow w + \eta \times \text{RMSProp}(w, G_w)$
6: 7.2 $w \leftarrow \text{clip}(w, -c, c)$
7: **end for**
8: **Output**: w, $\widetilde{\mathcal{W}}_{\text{MoM}}$, ϕ_w.

Evaluation metrics. The Lipchitz constant M being unknown and highly depending of the clipping parameter choice, it wouldn't be appropriate to compare the true 1-Wasserstein value, equal to $\sqrt{50}$, with $\widetilde{\mathcal{W}}_{\text{MoM}}, \widetilde{\mathcal{W}}_{\text{MoU-diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$. Therefore, we propose to compare $\widetilde{\mathcal{W}}_{\text{MoM}}$, $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$ to $\widetilde{\mathcal{W}}$, the 1-Wasserstein distance approximated by Algorithm 1, when MoM is not used, *e.g.* $K_{\mathbf{X}} = K_{\mathbf{Y}} = 1$, by measuring the absolute error between them.

Influence of $K_{\mathbf{X}}, K_{\mathbf{Y}}$. The numbers of blocks, $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$, are crucial parameters for computation. They define the trade-off between the robustness of the estimator and computational burden. However the theory does not give enough insights about their value: the necessary assumption for the consistency is only that they should be greater than $2\tau_X n$ (see Section 3.2). An empirical study of the influence of their values on the behavior of the approximations of $\mathcal{W}_{\text{MoM}}, \mathcal{W}_{\text{MoU-diag}}$ and $\mathcal{W}_{\text{MoU}}$ is therefore much useful. For sake of simplicity, we set $K_{\mathbf{X}} = K_{\mathbf{Y}}$ in the subsequent experiments.

In a first experiment, we explore the ability of algorithm 1 and variants described in the supplements to override outliers according to the values of $K_{\mathbf{X}}$ and with different rates of outliers $\tau_{\mathbf{X}}$. The approximations $\widetilde{\mathcal{W}}_{\text{MoM}}, \widetilde{\mathcal{W}}_{\text{MoU-diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$ are computed using a simple multi-layer perceptrons with one hidden layer and MoM gradient descent over several $\tau_{\mathbf{X}}$ and $K_{\mathbf{X}}$ on both

datasets. The experiment is repeated 20 times with various seeds. Mean results are displayed. Figure 2 represents absolute deviations between the 1-Wasserstein distance approximated with a MLP when $\tau_{\mathbf{X}} = 0$ and $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$ with various anomalies settings and different values of K_X. The reader is invited to refer to Section B of the supplements to see similar results for $\widetilde{\mathcal{W}}_{\text{MoU}}$ and $\widetilde{\mathcal{W}}_{\text{MoM}}$). Shaded areas, in Figure 2 represent 25%-75% quantiles over the 20 repetitions. On both datasets, we observe that the approximation algorithm succeeds to provide an estimation of $\mathcal{W}_{\text{MoU-diag}}$, able to override outliers with different $\tau_{\mathbf{X}}$ while $K_{\mathbf{X}}$ is high enough. From Section 3.2, we know that $K_{\mathbf{X}}$ needs to be higher than $2\tau_{\mathbf{X}} n$ to have theoretical guarantees. Experiments show that in practice, this condition is not necessary in every situation. For example, when $\tau_{\mathbf{X}} = 0.1$ (*i.e.* 50 anomalies) in Figure 2 (left), only 70 blocks are needed to override outliers. The reason is that hypothesis makes things work in the worst case, *i.e.*, when each outlier is isolated in one block which lead to have $\tau_{\mathbf{X}} n$ contaminated blocks. This is rarely the case in practice, several blocks can be contaminated by many outliers and this is why fewer blocks are needed.

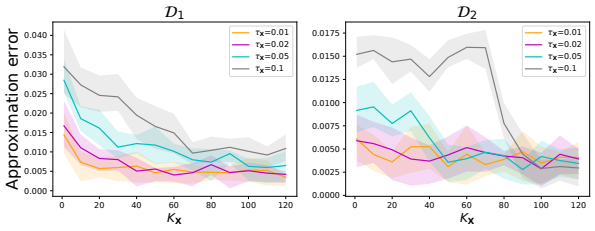


Figure 2: $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$ over $K_{\mathbf{X}}$ for different anomalies proportion $\tau_{\mathbf{X}}$ on $\mathcal{D}_1$ (left) and $\mathcal{D}_2$ (right).

In a second experiment illustrated by Figure 3, we study the convergence of the approximation algorithm with and without anomalies for different values of $K_{\mathbf{X}}$ on $\mathcal{D}_1$. To get a fair comparison between the different settings of the algorithm, we compare the predicted values across the "learning" epoch. Here during one epoch, the algorithm has made a gradient pass over the whole dataset, which means that one epoch corresponds one iteration of the approximation algorithm if $K_{\mathbf{X}} = 1$ (no MoM estimation), and to $K_{\mathbf{X}}$ iterations, in the other cases. In both cases (with or without anomalies), the higher $K_{\mathbf{X}}$ is, the faster the approximation algorithm converges. Without surprise, the MoM approach benefits from the same properties than a mini-batch approach. When there is no anomalies, the distance values reached after convergence are close to the "true" value (obtained with the plain estimator when $K_{\mathbf{X}} = 1$), especially when $K_{\mathbf{X}}$ is lower. This means that the MoM-based algorithm can be used routinely instead of the plain estimator. With 5% of anomalies, one can see that distance values reached after convergence get closer to the target as $K_{\mathbf{X}}$ grows.

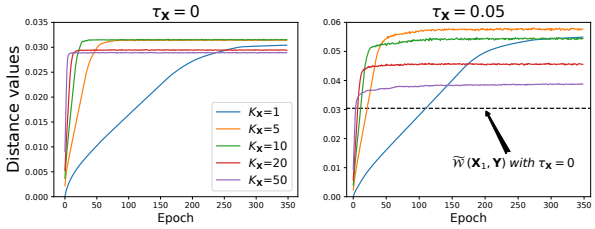


Figure 3: Convergence of $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$ without anomalies (left) and with 5% anomalies (right) for different $K_{\mathbf{X}}$.

4.3 Application to robust Wassertein GANs

In this part, we introduce a robust modification of WGANs, named MoMWGAN, using one of the three proposed estimators in Section 3. The behaviour of likelihood-free generative modeling such as Generative Adversarial Networks in the presence of outliers, *i.e.*, with heavy-tails distributions or contaminated data, has been poorly investigated up to very recently. At our knowledge, the unique reference is [17]. In particular, Gao et al. [17] have studied theoretically and empirically the robustness of f-GAN in the special case of mean estimation for elliptical distributions. In contrast, we illustrate here the theoretical results of section 3 by applying a MoM approach to robustify Wasserstein-GAN and show on two real-world image benchmarks how this new variant of GAN behaves when learned with contaminated data.

Reminder on GAN: Let us briefly recall the GAN principle. A GAN learns a function $g_\theta : \mathcal{Z} \to \mathcal{X}$ such that samples generate by $g_\theta(z) \sim P_\theta$, taking as input a sample z (from some reference measure ξ, often Gaussian) in a latent space $\mathcal{Z}$, are close to those of the true distribution P_r of data. Wasserstein GANs [7, 8] use the 1-Wasserstein Distance under its Kantorovich-Rubinstein dual formula as the loss function. Instead of maximizing over the unit ball of Lipschitz functions, one uses a parametric family of M-Lipschitz functions under the form of neural net with clipped weights w [7]. Following up the theoretical analysis of Section 3, we introduce a MoM-based WGAN (MoMWGAN) model, combining the $\mathcal{W}_{\text{MoM}}$ estimator studied in 3 and WGAN's framework. Following the weight clipping approach, MoMWGAN boils down to the problem:

$$\min_{\theta} \max_{w} \left\{ \text{MoM}_{\mathbf{X}}[f_w] - \frac{1}{m} \sum_{j=1}^{m} f_w(g_\theta(Z_j)),\ k \leq K_{\mathbf{X}} \right\}$$

Note that the MoM procedure is chosen to be only applied on the observed contaminated sample. It is

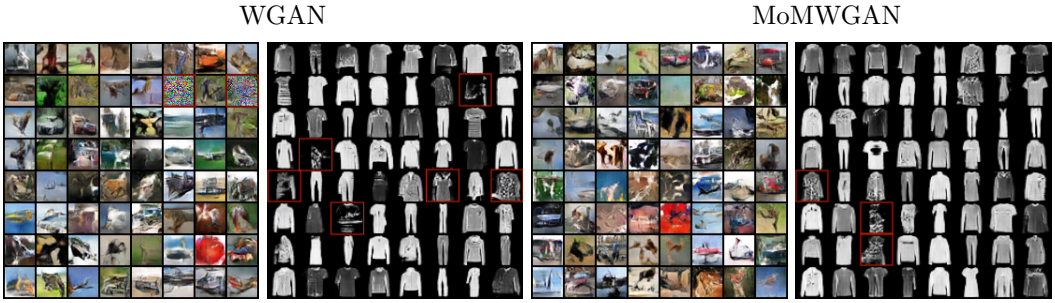


Figure 4: Generated samples from trained WGAN and MoMWGAN on CIFAR10 and Fashion MNIST datasets.

not clear in which way the sample drawn from the currently learned density is polluted and thus defining the number of blocks would be an issue. Optimization in WGAN is usually performed by taking mini-batches to reduce the computational load. In the same spirit, we apply MoM inside contaminated mini-batches as described in Algorithm 2. To get the outliers-robust property observed in the numerical experiments, we pay the price of finding the median block at each step by evaluating the loss which significantly increases the computational complexity.

Algorithm 2 MoMWGAN

Initialization: η, the learning rate. c, the clipping parameter. b, the batch size. n_c, the number of critic iterations per generator iteration, $K_{\mathbf{X}}$ the number of blocks. w_0, θ_0 the initial critic/generator's parameters.

while θ has not converged **do**
  for $t = 0, \dots, n_c$ **do**
    Sample $\{X_i\}_{i=1}^b \sim P_r$ to get $\mathbf{X}_t$ and sample $\{z_i\}_{i=1}^b \sim \xi$ to get $\mathbf{Z}_t$
    Updating w with step 2-6 of Algorithm 1 with $\mathbf{X} = \mathbf{X}_t$ and $\mathbf{Y} = g_\theta(\mathbf{Z}_t)$
  end for
  Sample $\{Z_j\}_{j=1}^b \sim \xi$
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{b} \sum_{j=1}^b f_w(g_\theta(Z_j))$
  $\theta \leftarrow \theta - \eta \times \mathrm{RMSProp}(\theta, g_\theta)$
end while

Numerical experiments To test the robustness of MoMWGAN we contaminated two well-known image datasets, CIFAR10 and Fashion MNIST, with two anomalies settings. *Noise* based-anomalies are added to CIFAR10, *i.e.*, images with random intensity pixels drawn from a uniform law. For Fashion MNIST, the five first classes are considered as "informative data" while the sixth (Sandal) contains the anomalies. In both settings, WGAN and MoMWGAN are trained on the training samples contaminated in a uniform fashion with a proportion of 1.5% of outliers in both datasets. Both models use standard parameters of WGAN. $K_{\mathbf{X}} = 4$ blocks have been used by MoMWGAN in both experiments. To assess performance of the resulting GANs, we generated 50000 generated images using each model (WGAN and MoMGAN) and measured the Fréchet Inception Distance (FID) [43] between the generated examples in both cases and the (real) test sample. Table 1 shows that MoMWGAN improves upon WGAN in terms of outliers-robustness. Furthermore, some generated images are represented in Figure 4. One can see that outliers do not affect MoMWGAN generated samples while WGAN reproduce noise on contaminated CIFAR10 dataset. For Fashion MNIST, one may see that fewer images are degraded with MoMWGAN generator.

	WGAN	MoMWGAN
Polluted CIFAR10	57	55.9
Polluted Fashion MNIST	13.8	13.2

Table 1: FID on polluted datasets.

5 Conclusion and perspectives

In this paper, we have introduced three robust estimators of the Wasserstein distance based on MoM methodology. We have shown asymptotic and non-asymptotic results in the context of polluted data, *i.e.* the $\mathcal{O} \cup \mathcal{I}$ framework. Surpassing computational issues, we have designed an algorithm to compute, in a efficient way, these estimators. Numerical experiments have highlighted the behavior of these estimators over their unique tuning parameter. Finally, we proposed to robustify WGANs using one of the introduced estimators and have shown its benefits on convincing numerical results. The theoretically well-founded MoM approaches to robustify the Wasserstein distance open the door to numerous applications beyond WGAN, including variational generative modeling. The promising MoMGAN deserves more attention and future work will concern the analysis of the estimator it provides.

Acknowlegments

The authors thank Pierre Colombo for his helpful remarks. This work has been funded by BPI France in the context of the PSPC Project Expresso (2017-2021).

References

[1] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, New York, 2003.

[2] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhauser, 2015.

[3] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[4] I. Csiszàr. Eine informationstheoretische ungleichung und ihre anwendung auf den bewis der ergodizität von markhoffschen kette. *Magyer Tud. Akad. Mat. Kutato Int. Koezl*, 8:85–108, 1963.

[5] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f -divergences. *Ann. Statist.*, 37(2):876–904, 04 2009.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS 2014)*, 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017.

[9] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[10] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[11] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Mach. Learn.*, 107(12):1923–1945, 2018.

[12] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, 2018.

[13] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, 2013.

[14] R. M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 02 1969.

[15] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics and Probability Letters*, 76: 1298–1302, 07 2006.

[16] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019.

[17] Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets, 2018.

[18] Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018).*, 2018.

[19] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics (Second Edition)*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[20] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Ltd, 1983.

[21] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[22] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

[23] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

[24] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[25] Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21 (4):2308–2335, 2015.

[26] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[27] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

[28] Emilien Joly and Gábor Lugosi. Robust estimation of u-statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016.

[29] Pierre Laforgue, Stephan Clémençon, and Patrice Bertail. On medians of (randomized) pairwise means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

[30] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[31] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019.

[32] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[33] P. Laforgue, G. Staerman, and S. Clémençon. How robust is the median-of-means? concentration bounds in presence of outliers. arxiv.org/abs/2006.05240, 2020.

[34] Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, and Guillaume Lecué. Monk – outlier-robust mean embedding estimation by median-of-means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

[35] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

[36] Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

[37] Emmanuel Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. *Electron. J. Probab.*, 16(83):2296–2333, 2011.

[38] Nicolas Fournier and Arnaud Guillin. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[39] A. J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

[40] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.*, 48(2):906–931, 04 2020.

[41] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599, 2012.

[42] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009. ISSN 0360-0300.

[43] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pages 6626–6637. 2017.