Relaxed Wasserstein, with Applications to GANs and Distributionally Robust Optimization

Xin Guo [*]   Johnny Hong [†]   Tianyi Lin [‡]   Nan Yang [§]

September 9, 2018

Abstract

Comparing probability distributions is an integral part of many modern data-driven applications, such as generative adversarial networks (GANs) and distributionally robust optimization (DRO). We propose a novel class of statistical divergences called *Relaxed Wasserstein* (RW) divergence, which generalizes Wasserstein divergence and is parametrized by the class of strictly convex and differentiable functions. We establish for RW divergence several probabilistic properties, many of which are crucial for the success of Wasserstein divergence. In addition, we derive theoretical results showing that the underlying convex function in RW plays an important role in variance stabilization, shedding light on the choice of appropriate convex function. We develop a version of GANs based on RW divergence and demonstrate via empirical experiments that RW-based GANs (RWGANs) lead to superior performance in image generation problems compared to existing approaches. In particular, we find that in our experiments RWGANs are fastest in generating meaningful images compared to other GANs. We also illustrate the use of RW divergence in constructing ambiguity sets for DRO problems, and the robust portfolio problem under mean-variance framework.

1 Introduction

Statistical divergences play an important role in many data-driven applications. One striking example is generative adversarial networks (GANs, Goodfellow et al. (2014)), which have been used for high resolution image generation (Denton et al., 2015; Radford et al., 2015), image inpainting (Yeh et al., 2016), image super-resolution (Ledig et al., 2016), visual manipulation (Zhu et al., 2016), text-to-image synthesis (Reed et al., 2016), video generation (Vondrick et al., 2016), semantic segmentation (Luc et al., 2016), and abstract reasoning diagram generation (Ghosh et al., 2017). A recurring theme to improve

---

[*]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. Email: xinguo@berkeley.edu.

[†]Department of Statistics, University of California, Berkeley, USA. Email: jcyhong@berkeley.edu.

[‡]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. Email: darren_lin@berkeley.edu.

[§]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. Email: nanyang@berkeley.edu.

GANs training is the choice of statistical divergences. The first proposed class of statistical divergences is based on the Jensen-Shannon (JS) divergence, which is essentially the symmetric version of the Kullback-Leibler (KL) divergence. It is shown in (Arjovsky et al., 2017) that JS divergence is undesirable with unstable training, suggesting Wasserstein-L^1 divergence as an alternative. The resulting Wasserstein GANs (WGANs) outperform the original GANs in several aspects. The Wasserstein-L^1 divergence is continuous, differentiable and has a duality representation, allowing a very stable gradient flow in the process of training. Besides stability, the Wasserstein-L^1 divergence also avoids the issue of mode collapse and further provides meaningful learning curves that can be used for debugging and for hyperparameter searching. With additional weight clipping (Arjovsky et al., 2017) and gradient penalty (WGANs-GP, Gulrajani et al. (2017)), the volatility of the gradient is somehow controlled. The Wasserstein-L^2 divergence, also known as Mallows distance, on the other hand, has been widely used in topics such as statistical testing (Munk and Czado, 1998; De Wet, 2002), machine learning (Zhou et al., 2005), optimal transport (Villani, 2008) and stochastic games (Lasry and Lions, 2007).

In this paper, we propose a novel class of statistical divergence called *Relaxed Wasserstein* (RW) divergence. RW divergence is Wasserstein divergence parametrized by the class of strictly convex and differentiable functions, which contain different curvature information. In this paper, we first show that RW divergence is dominated by the total variation (TV) distance and squared Wasserstein-L^2 divergence (Theorem 3.1). In parallel to the Wasserstein-L^2 divergence, we obtain its nonasymptotic moment estimate (Theorem 3.2) and its concentration inequality (Theorem 3.3). By comparing with Wasserstein divergence, we show RW is a reasonable divergence.

For application purposes, we establish an important lemma (Lemma 3.4) which states that RW divergence can be a distorted Wasserstein-L^2 divergence with some residual terms independent of the coupling. This decomposition immediately leads to the continuity and differentiability of RW divergence (Theorem 3.5). From a practical perspective, especially in light of stochastic gradient descent for GANs, these properties ensure the plausibility of a gradient descent procedure. Using the decomposition lemma again, we establish the duality representation of RW divergence (Theorem 3.6), which gives rise to an explicit formula for the gradient evaluation and an asymmetric clipping procedure (Corollary 3.6.1). Our numerical experiments show that this asymmetric clipping is useful for controlling the volatility of the gradient.

An important question in using RW divergence is the choice of the underlying convex function. We establish the connection between Fisher information and the Hessian of a convex function (Proposition 1) and derive the asymptotic distribution of Bregman divergences (Theorem 3.7). These theoretical results shed light from a variance stabilizing perspective on why Wasserstein-L^2 might not be a desirable choice of statistical divergence and how to select the convex function for RW.

We illustrate the use of RW divergence in GANs. In particular, we introduce Relaxed Wasserstein GANs (RWGANs) and compare RWGANs with several state-of-the-art GANs in image generation. Our numerical results show that despite the fastest rate of training, WGANs-GP fail to converge; RWGANs are robust and converge faster than WGANs, suggesting that RWGANs strike a balance between WGANs and WGANs-GP and RWGANs might be more desirable for large-scale computations. Furthermore, we observe in our experiments RWGANs are fastest in generating meaningful images compared to other GANs. As a byproduct, our experiment provides some evidences that an appropriate weight clipping has the potential

to be competitive with gradient penalty in WGANs.

As another application of RW divergence, we discuss how to use RW divergence to construct ambiguity sets for distrbutionally robust optimization (DRO) problems, in which statistical divergences are of critical importance. Solutions to traditional optimization problems are usually sensitive to the model parameters, which is a major drawback. Robust optimization solves this issue by formulating problems under appropriate uncertainty sets for the model parameters and/or for the solutions against a certain measure of robustness. For instance, tractable uncertainty sets can be formulated in terms of chance constraints and expectation constraints under a given distribution $\mathbb{P}$ (Jiang and Guan, 2012). However, in most data-driven research where the distribution $\mathbb{P}$ itself is usually unknown, the concept of ambiguity set is introduced (Bayraksan and Love, 2015). The key idea of DRO is as follows: instead of optimizing under one particular distribution and under a deterministic set, it formulates optimization problems with a set of possible distributions, under the concept of ambiguity sets. The ambiguity set contains distributions that are not far away from the nominal distribution, measured by the divergence function. In this regard, various choices of divergence functions have been discussed in the literature, for example, KL and f-divergences (Namkoong and Duchi, 2016; Van Parys et al., 2017) and Wasserstein distances (Esfahani and Kuhn, 2015; Shafieezadeh-Abadeh et al., 2015; Wozabal, 2012; Gao and Kleywegt, 2016; Blanchet et al., 2018). In this paper, we show that using RW divergences as a divergence function in DRO problems leads to simple forms of asymptotically valid ambiguity sets. We finally discuss the construction of robust portfolios under mean-variance framework.

The rest of the paper is organized as follows. Section 2 provides the preliminaries and notations that will be used throughout the paper. Section 3 describes the RW divergence and discusses its theoretical properties. In particular, we discuss choices of the convex function parametrizing the RW divergence in Section 3.3. Section 4.1 discusses the implementation of RWGANs and presents two numerical studies on real data examples. Section 4.2 explores the application of RW divergence in constructing ambiguity sets for robust optimization problems and robust portfolio constructions. Section 5 concludes our paper.

2 Background

In this section, we review the definitions and properties of Bregman divergence and Wasserstein divergence.

2.1 Notations

Throughout the paper, the following notations are used unless otherwise stated.

If $x \in \mathbb{R}^d$ denotes a vector in Euclidean space and X represents a matrix, then $x^\top$ denotes the transpose of this vector x, $\|x\|_q$ denotes that $q-$norm of x, and $\log(x)$ denotes the component-wise logarithm of this vector x. $X \succeq 0$ or $\succ 0$ means that X is positive semi-definite or positive definite, respectively. $\mathcal{X} \subset \mathbb{R}^d$ denotes a set where the diameter of $\mathcal{X}$ is defined as

$$\text{diam}(\mathcal{X}) = \max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|_2 \,,$$

and $1_{\mathcal{X}}$ denotes an indicator function of the set $\mathcal{X}$. If $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions, $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions defined on $\mathcal{X}$, then $\Pi(\mathbb{P}, \mathbb{Q})$ denotes the set of all couplings of $\mathbb{P}$ and $\mathbb{Q}$, i.e., the set of all joint distributions over $\mathcal{X} \times \mathcal{X}$ with marginal distributions being $\mathbb{P}$ and $\mathbb{Q}$. We use ϕ for a strictly convex and twice-differentiable function with an L-Lipschitz continuous gradient, i.e.,

$$0 \prec \nabla^2 \phi(x) \preceq LI_d,$$

where $x \in \text{dom}(\phi)$, i.e., the domain of ϕ, and I_d is an identity matrix in $\mathbb{R}^{d \times d}$. For the statistical learning setup, we define $\mathbb{P}_r$ as an unknown true probability distribution, $\mathbb{P}_n$ as the empirical distribution based on n observations from $\mathbb{P}_r$, and $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^d\}$ as a parametric family of probability distributions.

2.2 Wasserstein Divergence

Definition 2.1. *The Wasserstein divergence of order p between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$W_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} [c(x, y)]^p \ \pi(dx, dy) \right)^{1/p}, \tag{1}$$

where $p \geq 1$. $c(\cdot, \cdot) \geq 0$ is a metric supported on $\mathcal{X} \times \mathcal{X}$. An important special case is the Wasserstein-L^q divergence of order p as follows,

$$W_p^{L^q}(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_q^p \ \pi(dx, dy) \right)^{1/p}. \tag{2}$$

$q = 2$ and $\mathcal{X} = \mathbb{R}^d$ in (2) corresponds to the squared Wasserstein-L^2 divergence of order 2:

$$W_2^{L^2}(\mathbb{P}, \mathbb{Q}) \;\; = \;\; \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2^2 \ \pi(dx, dy) \right)^{1/2} \tag{3}$$

Remark 1. *Given $\mathbb{P}$ and $\mathbb{Q}$, we have the following two properties of the Wasserstein divergence of order p,*

1. *$W_p(\mathbb{P}, \mathbb{Q}) \geq 0$ and the equality holds if and only if $\mathbb{P} = \mathbb{Q}$ almost everywhere.*

2. *$W_p(\mathbb{P}, \mathbb{Q})$ is a metric since $W_p(\mathbb{P}, \mathbb{Q}) = W_p(\mathbb{Q}, \mathbb{P})$ and*

$$W_p(\mathbb{P}, \mathbb{Q}) \leq W_p(\mathbb{P}, \mathbb{S}) + W_p(\mathbb{S}, \mathbb{Q}),$$

where $\mathbb{S}$ is another probability distribution.

2.3 Bregman Divergence

Definition 2.2 ((Jones and Byrne, 1990)). *Given any strictly convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence is defined as*

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle, \tag{4}$$

for any $x, y \in \mathbb{R}^d$.

In particular, we have

- L^2 divergence: $D_\phi(x,y) = \|x-y\|_2^2$ if $\phi(x) = \|x\|_2^2$,

- Itakura-Saito divergence: $D_\phi(x,y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$ if $\phi(x) = -\log x$,

- KL divergence: $D_\phi(x,y) = x^\top \log(\frac{x}{y})$ if $\phi(x) = x^\top \log(x)$, and

- Mahalanobis divergence: $D_\phi(x,y) = (x-y)^\top A(x-y)$ if $\phi(x) = x^\top Ax$ and $A \succeq 0$.

Remark 2. *1. $D_\phi(x,y) \geq 0$, due to the convexity of ϕ and the equality holds if and only if $x = y$.*

2. $D_\phi(x,y)$ is not a metric: it is not symmetric and it violates the triangle inequality.

3. Bregman divergences are asymptotically equivalent to f-divergences (in particular, χ^2-divergence) under some conditions (Pardo and Vajda, 2003), and are the unique class of divergences where the conditional expectation is the optimal predictor (Banerjee et al., 2005a).

4. In statistical learning, the Bregman divergence is extensively exploited for K-means clusterings (Banerjee et al., 2005b).

In addition, the following lemma will be useful for our analysis.

Lemma 2.1. *Assume that $\phi : \mathcal{X} \to \mathbb{R}$ is a strictly convex and twice-differentiable function with an L-Lipschitz continuous gradient,*

$$D_\phi(x,y) \leq \frac{L}{2}\|x-y\|_2^2$$

for any $x,y \in \mathcal{X} \subset \mathbb{R}^d$.

Proof. This is clear,

$$
\begin{aligned}
D_\phi(x,y) &= \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y \rangle \\
&= \int_0^1 \langle \nabla\phi(tx + (1-t)y), x-y \rangle dt - \langle \nabla\phi(y), x-y \rangle \\
&= \int_0^1 \langle \nabla\phi(tx + (1-t)y) - \nabla\phi(y), x-y \rangle dt \\
&\leq \left(\int_0^1 t\, dt \right) L\|x-y\|_2^2 = \frac{L}{2}\|x-y\|_2^2.
\end{aligned}
$$

where the second equality comes from the mean value theorem and the inequality comes from the fact that ϕ is a twice-differentiable function with an L-Lipschitz continuous gradient. $\square$

3 Relaxed Wasserstein Divergence

We now propose a new class of statistical divergence called Relaxed Wasserstein (RW) divergence, parametrized by Wasserstein divergence and Bregman divergence. The term *relaxed* refers to the fact that RW divergence relaxes the symmetry of cost function $c(x,y)$ in Equation (1) and extends to a broader class of asymmetric divergences.

Definition 3.1. *The Relaxed Wasserstein divergence between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y) \, \pi(dx, dy),$$

where D_ϕ is the Bregman divergence with a strictly convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$.

Remark 3. *1. $W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \geq 0$ and the equality holds if and only if $\mathbb{P} = \mathbb{Q}$ almost everywhere.*

2. $W_{D_\phi}(\mathbb{P}, \mathbb{Q})$ is not a metric since $D_\phi(x, y)$ is asymmetric.

3. $W_{D_\phi}(\mathbb{P}, \mathbb{Q})$ includes two important special cases, $W_2^{L^2}$ and W_{KL}. More specifically, $W_{D_\phi} = W_2^{L^2}$ when $\phi(x) = \|x\|_2^2$, and $W_{D_\phi} = W_{KL}$ when $\phi(x) = -x^\top \log(x)$.

3.1 Probabilistic Properties

In this section, we establish several probabilistic properties of RW divergence. Recall that the Wasserstein divergence is controlled by weighted Total Variation (TV) distance (Theorem 6.15 (Villani, 2008) for more details). In parallel, we show that the RW divergence is dominated by the weighted TV distance and the squared Wasserstein-L^2 divergence.

Definition 3.2. *The Total Variation distance between the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$TV(\mathbb{P}, \mathbb{Q}) := \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|, \tag{5}$$

where A is a Borel set.

Theorem 3.1. *Assume that $\phi : \mathcal{X} \to \mathbb{R}$ is a strictly convex and twice-differentiable function with an L-Lipschitz continuous gradient, then*

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \leq L \left[\text{diam}(\mathcal{X})\right]^2 \cdot TV(\mathbb{P}, \mathbb{Q}), \tag{6}$$

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{2}L \cdot \left[W_2^{L^2}(\mathbb{P}, \mathbb{Q})\right]^2, \tag{7}$$

where $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$.

Proof. For the inequality (6), define π as the transfer plan that keeps all the mass shared by $\mathbb{P}$ and $\mathbb{Q}$ fixed and distributes the rest uniformly, i.e.,

$$\pi(dx, dy) = (\mathbb{P} \wedge \mathbb{Q})(dx)\delta_{\{y=x\}} + \frac{1}{a}(\mathbb{P} - \mathbb{Q})_+(dx) \cdot (\mathbb{P} - \mathbb{Q})_-(dy),$$

where $\mathbb{P} \wedge \mathbb{Q} = \mathbb{P} - (\mathbb{P} - \mathbb{Q})_+$ and $a = (\mathbb{P} - \mathbb{Q})_+ [\mathcal{X}] = (\mathbb{P} - \mathbb{Q})_- [\mathcal{X}]$. Then

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) &\leq \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y)\ \pi\,(dx, dy) \\
&= \frac{1}{a} \int_{\mathcal{X} \times \mathcal{X}} [\phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle]\ (\mathbb{P} - \mathbb{Q})_+ (dx) \cdot (\mathbb{P} - \mathbb{Q})_- (dy) \\
&= \frac{1}{a} \int_{\mathcal{X} \times \mathcal{X}} \left[\int_0^1 \langle \nabla \phi(tx + (1-t)y) - \nabla \phi(y), x - y \rangle\ dt \right] (\mathbb{P} - \mathbb{Q})_+ (dx) \cdot (\mathbb{P} - \mathbb{Q})_- (dy) \\
&\leq \frac{1}{a} \int_{\mathcal{X} \times \mathcal{X}} \left[\left(\int_0^1 t\,dt \right) L \|x - y\|_2^2 \right] (\mathbb{P} - \mathbb{Q})_+ (dx)\, (\mathbb{P} - \mathbb{Q})_- (dy) \\
&\leq \frac{L}{2a} \int_{\mathcal{X} \times \mathcal{X}} \left[\|x - y\|_2^2 \right] (\mathbb{P} - \mathbb{Q})_+ (dx)\, (\mathbb{P} - \mathbb{Q})_- (dy) \\
&\leq \frac{L}{a} \int_{\mathcal{X} \times \mathcal{X}} \left[\|x - x_0\|_2^2 + \|x_0 - y\|_2^2 \right] (\mathbb{P} - \mathbb{Q})_+ (dx)\, (\mathbb{P} - \mathbb{Q})_- (dy) \\
&\leq L \left[\int_{\mathcal{X}} \|x - x_0\|_2^2\ (\mathbb{P} - \mathbb{Q})_+ (dx) + \int_{\mathcal{X}} \|y - x_0\|_2^2\ (\mathbb{P} - \mathbb{Q})_- (dy) \right] \\
&= L \int_{\mathcal{X}} \|x - x_0\|_2^2\, |\mathbb{P} - \mathbb{Q}|\,(dx) = L\,[\mathrm{diam}(\mathcal{X})]^2 \cdot |\mathbb{P}(\mathcal{X}) - \mathbb{Q}(\mathcal{X})| \\
&\leq L\,[\mathrm{diam}(\mathcal{X})]^2 \cdot TV(\mathbb{P}, \mathbb{Q}),
\end{aligned}
$$

where the first inequality comes from Definition 3.1, the first equality from Definition 2.2 and the definition of the specific π, the second inequality is by Lemma 2.1, the fourth inequality by the triangle inequality, and the last inequality by Definition 3.2.

For the inequality (7), we have

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) &= \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} D_\phi(x, y)\ \pi(dx, dy) \\
&\leq \frac{1}{2} L \cdot \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2^2\ \pi(dx, dy) \\
&= \frac{1}{2} \cdot \left[W_2^{L^2}(\mathbb{P}, \mathbb{Q}) \right]^2,
\end{aligned}
$$

where the inequality holds thanks to Lemma 2.1 and the fact that $\pi(dx, dy) \geq 0$ for any coupling $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$. $\qquad\square$

Next, we establish another key probabilistic property of RW divergence, i.e., the nonasymptotic moment estimates and the concentration inequality. To begin, define two statistics

$$
M_q(\mathbb{P}_r) = \int_{\mathcal{X}} \|x\|_2^q\ \mathbb{P}_r(dx), \quad \text{and} \quad \mathcal{E}_{\alpha, \gamma}(\mathbb{P}_r) = \int_{\mathcal{X}} \exp\left(\gamma \|x\|_2^\alpha \right)\ \mathbb{P}_r(dx).
$$

Theorem 3.2 (Nonasymptotic Moment Estimate). *Assume that $M_q(\mathbb{P}_r) < +\infty$ for some $q > 2$, then there exists a constant $C(q, d) > 0$ such that, for $n \geq 1$,*

$$
\mathbb{E}\left[W_{D_\phi}(\mathbb{P}_n, \mathbb{P}_r) \right] \leq \frac{C(q, d) L M_q^{\frac{2}{q}}(\mathbb{P}_r)}{2} \cdot
\begin{cases}
n^{-\frac{1}{2}} + n^{-\frac{q-2}{q}}, & 1 \leq d \leq 3,\ q \neq 4, \\
n^{-\frac{1}{2}} \log(1 + n) + n^{-\frac{q-2}{q}}, & d = 4,\ q \neq 4, \\
n^{-\frac{2}{d}} + n^{-\frac{q-2}{q}}, & d \geq 5,\ q \neq d/(d-2).
\end{cases}
$$

Theorem 3.3 (Concentration Inequality). *Assume one of the following three conditions holds,*

$$\text{Either} \quad \exists\ \alpha > 2,\ \exists\ \gamma > 0, such\ that\ \mathcal{E}_{\alpha,\gamma}(\mathbb{P}_r) < \infty, \tag{8}$$

$$or \quad \exists\ \alpha \in (0,2),\ \exists\ \gamma > 0, such\ that\ \mathcal{E}_{\alpha,\gamma}(\mathbb{P}_r) < \infty, \tag{9}$$

$$or \quad \exists\ q > 4, such\ that\ M_q(\mathbb{P}_r) < \infty. \tag{10}$$

Then for $n \geq 1$ and $\epsilon > 0$,

$$\mathbb{P}_r\left(W_{D_\phi}(\mathbb{P}_n, \mathbb{P}_r) \geq \epsilon\right) \leq a(n,\epsilon)1_{\{\epsilon \leq \frac{L}{2}\}} + b(n,\epsilon),$$

where

$$a(n,\epsilon) = C_1 \begin{cases} \exp\left(-\frac{4cn\epsilon^2}{L^2}\right), & 1 \leq d \leq 3, \\ \exp\left(-\frac{4cn\epsilon^2}{L^2}\log^2\left(2 + \frac{L}{2\epsilon}\right)\right), & d = 4, \\ \exp\left(-cn\left(\frac{2\epsilon}{L}\right)^{\frac{d}{2}}\right), & d \geq 5, \end{cases}$$

and

$$b(n,\epsilon) = C_2 \begin{cases} \exp\left(-cn\left(\frac{2\epsilon}{L}\right)^{\frac{\alpha}{2}}\right) \cdot 1_{\{\epsilon > \frac{L}{2}\}}, & under\ condition\ (8), \\ \exp\left(-c\left(\frac{2n\epsilon}{L}\right)^{\frac{\alpha-\epsilon}{2}}\right) \cdot 1_{\{\epsilon \leq \frac{L}{2}\}} + \exp\left(-c\left(\frac{2n\epsilon}{L}\right)^{\frac{\alpha}{2}}\right) \cdot 1_{\{\epsilon > \frac{L}{2}\}}, & 0 < \epsilon < \alpha, under\ condition\ (9), \\ n\left(\frac{2n\epsilon}{L}\right)^{-\frac{q-\epsilon}{2}}, & 0 < \epsilon < q, under\ condition\ (10). \end{cases}$$

where c, C_1 and C_2 are constants depending on q and d.

Theorem 3.2 and Theorem 3.3 show that the importance of Lipchitz constant L of the underlying function ϕ in the statistical behaviour of RW divergence. The proof follows from Theorem 1 and Theorem 2 presented in Fournier and Guillin (2015) and Theorem 3.1 in this paper.

3.2   Continuity, Differentiability and Duality Representation

In this section, we establish the continuity, differentiability and duality representation of RW divergence, demonstrating that RW divergence is a reasonable choice for the GANs. We first present a simple yet important lemma.

Lemma 3.4 (Decomposition of RW divergence). *The RW divergence can be decomposed in terms of the distorted squared Wasserstein-L_2 divergence of order 2 with several additional residual terms independent of the choice of coupling π, i.e.,*

$$W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\left[W_2^{L^2}\left(\mathbb{P}, \mathbb{Q} \circ (\nabla\phi)^{-1}\right)\right]^2$$
$$+ \int_{\mathcal{X}}\left[\phi(x) - \frac{1}{2}\|x\|_2^2\right]\mathbb{P}(dx) + \int_{\mathcal{X}}\left[\langle\nabla\phi(x), x\rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2\right]\mathbb{Q}(dx).$$
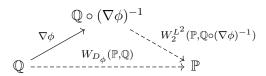
See Figure 1.

Figure 1: The decomposition of W_{D_ϕ} where the solid arrow denotes transformation and the dashed arrows denote the divergences between probability distributions.

Proof. First, we need to prove that the inverse of ϕ is well-defined. Since $\nabla^2\phi(x) \succ 0$, $\forall x \in \mathcal{X}$, the gradient mapping $\nabla\phi : \mathcal{X} \to \mathbb{R}^d$ has a positive-definite Jacobian matrix at each point. Applying the mean value theorem yields that ϕ is injective so the inverse of $\nabla\phi$ exists and is bijective. Denote it as

$$(\nabla\phi)^{-1} : \nabla\phi(\mathcal{X}) \to \mathcal{X},$$

then

$$\mathbb{Q} \circ (\nabla\phi)^{-1} : \mathbb{R}^d \to \mathbb{R}$$

is also a probability distribution. Thus

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P},\mathbb{Q}) &= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{X}} [\phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle] \; \pi(dx, dy) \\
&= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{X}} \left[\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|\nabla\phi(y)\|_2^2 - \langle\nabla\phi(y), x\rangle\right] \; \pi(dx, dy) \\
&\quad + \int_{\mathcal{X}\times\mathcal{X}} \left[\phi(x) - \frac{1}{2}\|x\|_2^2\right] \; \pi(dx, dy) + \int_{\mathcal{X}\times\mathcal{X}} \left[\langle\nabla\phi(y), y\rangle - \phi(y) - \frac{1}{2}\|\nabla\phi(y)\|^2\right] \; \pi(dx, dy) \\
&= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{X}} \left[\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|\nabla\phi(y)\|_2^2 - \langle\nabla\phi(y), x\rangle\right] \; \pi(dx, dy) \\
&\quad + \int_{\mathcal{X}} \left[\phi(x) - \frac{1}{2}\|x\|_2^2\right] \; \mathbb{P}(dx) + \int_{\mathcal{X}} \left[\langle\nabla\phi(x), x\rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2\right] \; \mathbb{Q}(dx).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\left[W_2^{L^2}\left(\mathbb{P}, \mathbb{Q}\circ(\nabla\phi)^{-1}\right)\right]^2 &= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q}\circ(\nabla\phi)^{-1})} \int_{\mathcal{X}\times\mathbb{R}^d} \|x - y\|_2^2 \; \pi(dx, dy) \\
&= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathbb{R}^d} \|x - \nabla\phi(y)\|_2^2 \; \pi(dx, dy) \\
&= \inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{X}} \left[\|x\|_2^2 + \|\nabla\phi(y)\|_2^2 - 2\langle\nabla\phi(y), x\rangle\right] \; \pi(dx, dy).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\left[W_2^{L^2}\left(\mathbb{P}, \mathbb{Q}\circ(\nabla\phi)^{-1}\right)\right]^2 \\
&\quad + \int_{\mathcal{X}} \left[\phi(x) - \frac{1}{2}\|x\|_2^2\right] \; \mathbb{P}(dx) + \int_{\mathcal{X}} \left[\langle\nabla\phi(x), x\rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2\right] \; \mathbb{Q}(dx).
\end{aligned}
$$

$\square$

Now we are ready to present our main results on the continuity and differentiability of the parametrized RW divergence in the generative modeling.

Definition 3.3 (Generative modeling). *The procedure of generative modeling is to approximate an unknown probability distribution $\mathbb{P}_r$ by constructing a class of suitable parametric probability distributions $\mathbb{P}_\theta$. More specifically, define a latent variable $Z \in \mathcal{Z}$ with a fixed probability distribution $\mathbb{P}_Z$ and a sequence of parametric functions $g_\theta : \mathcal{Z} \to \mathcal{X}$. Then $\mathbb{P}_\theta$ is defined as the probability distribution of $g_\theta(Z)$.*

Theorem 3.5 (Continuity and Differentiability of RW divergence). *1. $W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous in θ if g_θ is continuous in θ.*

2. $W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ is differentiable almost everywhere if g_θ is locally Lipschitz with a constant $L(\theta, z)$ such that $\mathbb{E}\left[L(\theta, Z)^2\right] < \infty$, i.e., for each given (θ_0, z_0), there exists a neighborhood $\mathcal{N}$ such that

$$\|g_\theta(z) - g_{\theta_0}(z_0)\|_2 \leq L(\theta_0, z_0)\left(\|\theta - \theta_0\|_2 + \|z - z_0\|_2\right).$$

for any $(\theta, z) \in \mathcal{N}$.

Proof. It follows from Lemma 3.4 that $W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta) = T_1 + T_2$, where

$$T_1 = \frac{1}{2}\left[W_2^{L^2}\left(\mathbb{P}_r, \mathbb{P}_\theta \circ (\nabla\phi)^{-1}\right)\right]^2,$$

$$T_2 = \int_{\mathcal{X}}\left[\phi(x) - \frac{1}{2}\|x\|_2^2\right]\mathbb{P}_r(dx) + \int_{\mathcal{X}}\left[\langle\nabla\phi(x), x\rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2\right]\mathbb{P}_\theta(dx).$$

We observe that T_2 is continuous and differentiable with respect to θ since ϕ is a twice differentiable function. Furthermore, since $(\nabla\phi)^{-1}$ is also continuous and differentiable, it suffices to show that $W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous in θ if g_θ is continuous in θ, and differentiable almost everywhere if g_θ is locally Lipschitz with a constant $L(\theta, z)$ such that $\mathbb{E}\left[L(\theta, Z)^2\right] < \infty$ for any θ.

Given two vectors $\theta_0, \theta \in \mathbb{R}^d$, we define π as a joint distribution of $(g_\theta(Z), g_{\theta_0}(Z))$ where $Z \sim \mathbb{P}_Z$, then

$$W_2^{L^2}(\mathbb{P}_\theta, \mathbb{P}_{\theta_0}) \leq \left(\int_{\mathcal{X}\times\mathcal{X}}\|x - y\|_2^2 \ \pi(dx, dy)\right)^{1/2}$$

$$= \left(\int_{\mathcal{Z}}\|g_\theta(z) - g_{\theta_0}(z)\|_2^2 \ \mathbb{P}_Z(dz)\right)^{1/2},$$

where $\|g_\theta(z) - g_{\theta_0}(z)\|_2^2 \to 0$, $\forall z \in \mathcal{Z}$, since g_θ is continuous in θ. Furthermore, $\|g_{\theta_1}(z) - g_{\theta_2}(z)\|_2^2$ is uniformly bounded on $\mathcal{Z}$ since $g_\theta(x) \in \mathcal{X}$ and $\mathcal{X}$ is a compact set. Therefore, applying the bounded convergence theorem yields

$$\left|W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_\theta) - W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_{\theta_0})\right| \leq W_2^{L^2}(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})$$

$$\leq \left(\int_{\mathcal{Z}}\|g_\theta(z) - g_{\theta_0}(z)\|_2^2 \ \mathbb{P}_Z(dz)\right)^{1/2} \to 0, \quad \text{as } \theta \to \theta_0.$$

where the first inequality comes from the triangle inequality.

Given a pair (θ_0, z_0), the local Lipschitz continuity of g_θ implies that there exists a neighborhood $\mathcal{N}$ such that $\|g_\theta(z) - g_{\theta_0}(z_0)\|_2 \leq L(\theta_0, z_0)\left(\|\theta - \theta_0\|_2 + \|z - z_0\|_2\right)$ for any $(\theta, z) \in \mathcal{N}$. Then

$$\int_{\mathcal{Z}}\|g_\theta(z_0) - g_{\theta_0}(z_0)\|_2^2 \ \mathbb{P}_Z(dz_0) \leq \int_{\mathcal{Z}}[L(\theta_0, z_0)]^2 \cdot \|\theta - \theta_0\|_2^2 \ \mathbb{P}_Z(dz_0)$$

$$= \|\theta - \theta_0\|_2^2 \cdot \mathbb{E}\left[L(\theta_0, Z)^2\right].$$

Therefore,

$$
\begin{aligned}
\left| W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_\theta) - W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_{\theta_0}) \right| &\leq W_2^{L^2}(\mathbb{P}_\theta, \mathbb{P}_{\theta_0}) \\
&\leq \left(\int_{\mathcal{Z}} \|g_\theta(z_0) - g_{\theta_0}(z_0)\|_2^2 \ \mathbb{P}_Z(dz_0) \right)^{1/2} \\
&\leq \|\theta - \theta_0\|_2 \cdot \mathbb{E}\left[L(\theta, Z)^2 \right]^{1/2},
\end{aligned}
$$

which implies that $W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_\theta)$ is locally Lipschitz. Applying the Rademacher's theorem (Evans and Gariepy, 2015) yields that $W_2^{L^2}(\mathbb{P}_r, \mathbb{P}_\theta)$ is differentiable with respect to θ almost everywhere. $\qquad \square$

Next is the the duality representation of RW divergence.

Theorem 3.6 (Duality Representation of RW divergence). *Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ such that*

$$
\int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P} + \mathbb{Q})(dx) < +\infty,
$$

then there exists a Lipschitz continuous function $f : \mathcal{X} \to \mathbb{R}$ such that the RW divergence has the following duality representation

$$
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \phi(x) \ (\mathbb{P} - \mathbb{Q})(dx) + \int_{\mathcal{X}} \langle \nabla\phi(x), x \rangle \ \mathbb{Q}(dx) - \left(\int_{\mathcal{X}} f(x) \ \mathbb{P}(dx) + \int_{\mathcal{X}} f^*(\nabla\phi(x)) \ \mathbb{Q}(dx) \right),
$$

where f^* is the conjugate of f, such that $f^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - f(x)$.

Proof. First,

$$
\left[W_2^{L^2}(\mathbb{P}, \mathbb{Q}) \right]^2 = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2^2 \ \pi(dx, dy) \tag{11}
$$

$$
= \int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P} + \mathbb{Q})(dx) - \sup_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} 2x^\top y \ \pi(dx, dy), \tag{12}
$$

then it follows from Proposition 3.1 (Brenier, 1991) that there exists a Lipschitz continuous function $f : \mathcal{X} \to \mathbb{R}$ such that the squared Wasserstein-L_2 divergence of order 2 has a duality representation:

$$
\begin{aligned}
\left[W_2^{L^2}(\mathbb{P}, \mathbb{Q}) \right]^2 &= \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2^2 \ \pi(dx, dy) \\
&= \int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P} + \mathbb{Q})(dx) - 2 \left(\int_{\mathcal{X}} f(x) \ \mathbb{P}(dx) + \int_{\mathcal{X}} f^*(x) \ \mathbb{Q}(dx) \right),
\end{aligned}
$$

where $f^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - f(x)$. By Lemma 3.4,

$$
\begin{aligned}
W_{D_\phi}(\mathbb{P}, \mathbb{Q}) &= \frac{1}{2} \left[W_2^{L^2}\left(\mathbb{P}, \mathbb{Q} \circ (\nabla\phi)^{-1}\right) \right]^2 \\
&\quad + \int_{\mathcal{X}} \left[\phi(x) - \frac{1}{2}\|x\|_2^2 \right] \mathbb{P}(dx) + \int_{\mathcal{X}} \left[\langle \nabla\phi(x), x \rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2 \right] \mathbb{Q}(dx), \\
&= \frac{1}{2} \left(\int_{\mathcal{X}} \|x\|_2^2 \ \mathbb{P}(dx) + \int_{\mathcal{X}} \|\nabla\phi(x)\|_2^2 \ \mathbb{Q}(dx) \right) - \left(\int_{\mathcal{X}} f(x) \ \mathbb{P}(dx) + \int_{\mathcal{X}} f^*(\nabla\phi(x)) \ \mathbb{Q}(dx) \right) \\
&\quad + \int_{\mathcal{X}} \left[\phi(x) - \frac{1}{2}\|x\|_2^2 \right] \mathbb{P}(dx) + \int_{\mathcal{X}} \left[\langle \nabla\phi(x), x \rangle - \phi(x) - \frac{1}{2}\|\nabla\phi(x)\|^2 \right] \mathbb{Q}(dx) \\
&= \int_{\mathcal{X}} \phi(x) \ (\mathbb{P} - \mathbb{Q})(dx) + \int_{\mathcal{X}} \langle \nabla\phi(x), x \rangle \ \mathbb{Q}(dx) - \left(\int_{\mathcal{X}} f(x) \ \mathbb{P}(dx) + \int_{\mathcal{X}} f^*(\nabla\phi(x)) \ \mathbb{Q}(dx) \right).
\end{aligned}
$$

11

$\square$

Finally, we show that Theorem 3.6 allows for an explicit formula for the gradient evaluation in the generative modeling (Definition 3.3), providing the theoretical guarantee for the RWGANs training.

Corollary 3.6.1 (Gradient Evaluation). *Under the setting of generative modeling, we assume that g_θ is locally Lipschitz with a constant $L(\theta, z)$ such that $\mathbb{E}\left[L(\theta, Z)^2\right] < \infty$ and*

$$\int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P}_r + \mathbb{P}_\theta)(dx) < +\infty.$$

Then there exists a Lipschitz continuous solution $f : \mathcal{X} \to \mathbb{R}$ such that the gradient of the RW divergence has an explicit form of

$$\nabla_\theta \left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)\right] = \mathbb{E}_Z \left[[\nabla_\theta g_\theta(Z)]^\top \nabla^2 \phi(g_\theta(Z)) g_\theta(Z)\right] + \mathbb{E}_Z \left[\nabla_\theta f (\nabla \phi(g_\theta(Z)))\right].$$

Proof. Since g_θ is a locally Lipschitz and $\int_{\mathcal{X}} \|x\|_2^2 \ (\mathbb{P}_r + \mathbb{P}_\theta)(dx) < +\infty$, it follows from Theorem 3.5 and Theorem 3.6 that $W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)$ is differentiable almost everywhere and there exists a Lipschitz continuous function $\tilde{f} : \mathcal{X} \to \mathbb{R}$ such that the RW divergence has a duality representation as

$$W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta) = \int_{\mathcal{X}} \phi(x) \ (\mathbb{P}_r - \mathbb{P}_\theta)(dx) + \int_{\mathcal{X}} \langle \nabla \phi(x), x \rangle \, \mathbb{P}_\theta(dx) - \left(\int_{\mathcal{X}} \tilde{f}(x) \, \mathbb{P}_r(dx) + \int_{\mathcal{X}} \tilde{f}^* (\nabla \phi(x)) \ \mathbb{P}_\theta(dx)\right).$$

By the envelope theorem (Milgrom and Segal, 2002), we obtain that

$$
\begin{aligned}
\nabla_\theta \left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)\right] &= \nabla_\theta \left[-\int_{\mathcal{X}} \phi(x) \, \mathbb{P}_\theta(dx) + \int_{\mathcal{X}} \langle \nabla \phi(x), x \rangle \, \mathbb{P}_\theta(dx) - \int_{\mathcal{X}} \tilde{f}^* (\nabla \phi(x)) \ \mathbb{P}_\theta(dx)\right] \\
&= \nabla_\theta \left[-\int_{\mathcal{Z}} \phi(g_\theta(z)) \, \mathbb{P}_Z(dz) + \int_{\mathcal{Z}} \langle \nabla \phi(g_\theta(z)), g_\theta(z) \rangle \, \mathbb{P}_Z(dz) - \int_{\mathcal{Z}} \tilde{f}^* (\nabla \phi(g_\theta(z))) \ \mathbb{P}_Z(dz)\right] \\
&= -\int_{\mathcal{Z}} [\nabla_\theta g_\theta(z)]^\top \nabla \phi(g_\theta(z)) \, \mathbb{P}_Z(dz) + \int_{\mathcal{Z}} [\nabla_\theta g_\theta(z)]^\top \nabla \phi(g_\theta(z)) \, \mathbb{P}_Z(dz) \\
&\quad + \int_{\mathcal{Z}} [\nabla_\theta g_\theta(z)]^\top \nabla^2 \phi(g_\theta(z)) g_\theta(z) \, \mathbb{P}_Z(dz) - \int_{\mathcal{Z}} \nabla_\theta \tilde{f}^* (\nabla \phi(g_\theta(z))) \ \mathbb{P}_Z(dz) \\
&= \int_{\mathcal{Z}} [\nabla_\theta g_\theta(z)]^\top \nabla^2 \phi(g_\theta(z)) g_\theta(z) \, \mathbb{P}_Z(dz) - \int_{\mathcal{Z}} \nabla_\theta \tilde{f}^* (\nabla \phi(g_\theta(z))) \ \mathbb{P}_Z(dz)
\end{aligned}
$$

Letting $f = -\tilde{f}^*$,

$$\nabla_\theta \left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta)\right] = \mathbb{E}_Z \left[[\nabla_\theta g_\theta(Z)]^\top \nabla^2 \phi(g_\theta(Z)) g_\theta(Z)\right] + \mathbb{E}_Z \left[\nabla_\theta f (\nabla \phi(g_\theta(Z)))\right],$$

where f is Lipschitz continuous. $\square$

3.3 Choices of ϕ

While RW divergence provides the flexibility of choosing the underlying Bregman divergence, in practice one would like to have a principled way of determining this choice. The following theoretical results shed light on how to choose an appropriate convex function ϕ in Bregman divergence D_ϕ.

Our first result connects Fisher information of a distribution of an exponential family and the Hessian of ϕ.

Proposition 1. *Suppose $X \sim \mathbb{P}_\theta$ belongs to a regular exponential family. Let $\mu = \mathbb{E}(X)$, ψ be the cumulant function, and ϕ be the convex conjugate of ψ. Let*

$$(\mathcal{I})_{ij} = -\mathbb{E}\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \tag{13}$$

be the Fisher information matrix of the underlying true distribution, with L being the likelihood function. Assume that ψ is three-time differentiable. Then

$$\mathcal{I}(\mu) = \mathbb{E}\left[\nabla_\mu^2 D_\phi(x, \mu)\right] = \nabla^2 \phi(\mu). \tag{14}$$

Proof. Equation (13) follows directly from the representation $p_\theta(x) = \exp(-D_\phi(x, \mu) - g_\phi(x))$. Equation (14) follows from a straightforward calculation,

$$\mathbb{E}\left[\nabla_\mu^2 D_\phi(x, \mu)\right] = \mathbb{E}\left[\nabla_\mu^2 [\phi(x) - \phi(\mu) - \nabla\phi(\mu)^T(x - \mu)]\right]$$
$$= \mathbb{E}\left[\nabla_\mu[-\nabla^2\phi(\mu)(x - \mu)]\right] = \mathbb{E}[\nabla_\mu^2\phi(\mu)] = \phi''(\mu).$$

$\square$

The next result shows that asymptotically, Bregman divergence between the true parameters and the corresponding maximum likelihood estimator of the parameters will converge in distribution to a finite weighted sum of independent χ^2 distributed random variables.

Theorem 3.7. *Suppose there exists a family of probability distributions $\mathbb{P}_\theta$ parametrized by $\theta \in \Theta \subset \mathbb{R}^m$. Given i.i.d data $\{X_i\}_{i=1}^n$, and $\hat{\theta}_n$ the maximum likelihood estimator of θ, then*

$$\lim_{n \to \infty} nD_\phi(\theta, \hat{\theta}_n) \xrightarrow{d} \frac{1}{2}\sum_{i=1}^r \beta_i Z_i^2,$$

where Z_i's are independent standard Gaussian random variables, β_i's are the non-zero eigenvalues of the matrix $H\Sigma$, and $r = rank(\Sigma^T H \Sigma)$, with H the Hessian of ϕ at θ and Σ the inverse Fisher information matrix.

Proof. First, write the Taylor expansion of ϕ around $\hat{\theta}_n$,

$$\phi(\theta) = \phi(\hat{\theta}_n) + \langle \theta - \hat{\theta}_n, \nabla\phi(\hat{\theta}_n)\rangle + \frac{1}{2}(\theta - \hat{\theta}_n)^T H(\hat{\theta}_n)(\theta - \hat{\theta}_n) + o(\|\theta - \hat{\theta}_n\|_2^2),$$

where $H(\hat{\theta})$ is the Hessian of $\phi(x)$ at $x = \hat{\theta}$. Notice that by the properties of maximum likelihood estimators, as $n \to \infty$, $\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{d} N(0, \mathcal{I}^{-1}) \overset{d}{=} N(0, \Sigma)$. Also, both $H(\hat{\theta}_n) \to H(\theta)$ and $n \cdot o(\|\theta - \hat{\theta}_n\|_2^2) \to 0$ in probability. Therefore by the Slutsky's theorem,

$$nD_\phi(\theta, \hat{\theta}_n) = n(\phi(\theta) - \phi(\hat{\theta}_n) - \langle\theta - \hat{\theta}_n, \nabla\phi(\hat{\theta}_n)\rangle)$$
$$= \frac{1}{2}\sqrt{n}(\theta - \hat{\theta}_n)^T H \sqrt{n}(\theta - \hat{\theta}_n) + n \cdot o(\|\theta - \hat{\theta}_n\|_2^2)$$
$$\xrightarrow{d} \frac{1}{2}X^T H X,$$

where $X \overset{d}{=} N(0, \Sigma)$. Let $S \in \mathbb{R}^{d \times s}$ be a square root of Σ. Since Σ and H are positive semidefinite, by spectral theorem, we can write $S^T H S = R^T \Lambda R$, where $\Lambda = diag(\beta_1, \ldots, \beta_r)$, which is the diagonal matrix of non-zero eigenvalues of $S^T H S$, and also the diagonal matrix of non-zero eigenvalues of $H\Sigma$, $r = rank(\Sigma H \Sigma)$, and R is the matrix of corresponding orthonormal eigenvectors. Then

$$X^T H X \overset{d}{=} (SY)^T H SY \overset{d}{=} Y^T R^T \Lambda R Y \overset{d}{=} Z^T \Lambda Z = \sum_{i=1}^{r} \beta_i Z_i^2,$$

where Z_i are independent standard Gaussian random variables. Therefore, we have the quadratic form of Gaussian variables $\sqrt{n}(\theta - \hat{\theta}_n)^T H \sqrt{n}(\theta - \hat{\theta}_n) \overset{d}{=} \sum_{i=1}^{r} \beta_i Z_i^2$. $\qquad \square$

To see how Theorem 3.7 and Proposition 1 shed light on the choice of ϕ, let us first consider the squared loss $D_\phi(x, y) = ||x - y||^2$. In this case, the corresponding ϕ is $||x||_2^2$ and the Hessian is $H = 2I$, so the weights in the weighted sum of χ_1^2 random variables in Theorem 3.7 are determined purely from the eigenvalues of the inverse Fisher information matrix Σ, which is the negative inverse of Hessian of the likelihood function. In other words, the convergence behavior of $nD_\phi(\theta, \hat{\theta}_n)$ is purely determined from the curvature of the likelihood surface at θ. If the likelihood surface is close to being flat at θ in certain directions, some of the eigenvalues of Σ will be undesirably large, resulting in a large asymptotic variance for $nD_\phi(\theta, \hat{\theta})$. This suggests that the squared loss function and hence Wasserstein-L^2 might not be a suitable choice as a divergence measure when the underlying likelihood function is likely to be flat at the true parameter θ.

Moreover, in light of Theorem 3.7, H can be used as a tool to stabilize the asymptotic variations of $nD_\phi(\theta, \hat{\theta}_n)$. A potential choice of H is Σ^{-1}, the Fisher information matrix of the likelihood function. Then $H\Sigma = I$, so all the associated eigenvalues β_i's are ones and the resulting asymptotic variance is always $r/2$, independent of the curvature of the underlying likelihood surface. To ensure that $H = \Sigma^{-1}$, if the underlying likelihood function is from an exponential family, ϕ can be simply chosen to be the associated Bregman divergence by Proposition 1. Note that with this choice $nD_\phi(\theta, \hat{\theta}_n)$ is equivalent to the classical likelihood ratio statistic. In a more general setting, if a reasonable estimate of the Fisher information matrix at θ is available, say $\hat{\Sigma}^{-1}$, a reasonable choice of Bregman divergence is the Mahalanobis distance $D_\phi(x, y) = (x - y)^T \hat{\Sigma}^{-1}(x - y)$, provided that the objective is to stabilize the asymptotic variance of $nD_\phi(\theta, \hat{\theta}_n)$. Indeed, the corresponding ϕ is $\phi(x) = x^T \hat{\Sigma}^{-1} x$ and the Hessian is $\hat{\Sigma}^{-1}$, so the matrix $H\Sigma$ in Theorem 3.7 is close to being the identity matrix.

4 Applications

4.1 RWGANs

In this section, we will present numerical evaluation on image generations to demonstrate the effectiveness and efficiency of using RW in GANs. We will first review the basics of GANs (Section 4.1.1) and the computation of the gradient of RW in training GANs (RWGANs) (Section 4.1.2). We then describe our experiment framework and settings (Section 4.1.3). Finally we report the experimental results under RWGANs versus other nine well-established variants of GANs (Section 4.1.4).

4.1.1 GANs and Jensen-Shannon divergence.

The goal of the GANs is to estimate a probability distribution $\mathbb{P}_r$ hidden in the data. As defined in Definition 3.3, one can define a random variable Z with a fixed distribution $\mathbb{P}_Z$ and pass it through a parametric function $g_\theta : \mathcal{Z} \to \mathcal{X}$ to construct a probability distribution $\mathbb{P}_\theta$. In practice, the parametric function g_θ is implemented using a neural network called *Generator G*. Meanwhile, another neural network *Discriminator D* will assign a score between 0 to 1 to the generated samples, either from the empirical distribution $\mathbb{P}_r$ or the approximate distribution $\mathbb{P}_\theta = g_\theta(Z)$. A higher score from the discriminator D would indicate that the sample is more likely to be from the empirical distribution. A GAN is trained by optimizing G and D iteratively until D can no longer distinguish between samples from $\mathbb{P}_r$ or $\mathbb{P}_\theta$. In this light, one can learn the probability distribution $\mathbb{P}_r$ by adapting θ and fitting the data with $\mathbb{P}_\theta$. This approximation is done by finding a solution f that optimizes a given cost function between $\mathbb{P}_r$ and $\mathbb{P}_\theta$.

Mathematically, training of GANs with an optimal discriminator is minimizing the Jensen-Shannon divergence between $\mathbb{P}_r$ and $\mathbb{P}_\theta$. Indeed, recall that GANs is a min-max game of

$$\min_G \max_D \left\{ \mathbb{E}_{x \sim \mathbb{P}_r}[\log D(x)] + \mathbb{E}_{z \sim \mathbb{P}(z)}[\log(1 - D(G(z)))] \right\}. \tag{15}$$

If we fix G and optimize for D, the optimal discriminator would be $D_G^*(x) = \frac{p_r(x)}{p_r(x)+p_g(x)}$, where p_r and p_g are density functions of $\mathbb{P}_r$ and $\mathbb{P}_\theta = g_\theta(Z)$ respectively. Plugging this back to Equation (15), we have

$$\min_G \left\{ \mathbb{E}_{x \sim \mathbb{P}_r}[\log \frac{p_r(x)}{p_r(x) + p_g(x)}] + \mathbb{E}_{x \sim \mathbb{P}_\theta(Z)}[\log \frac{p_g(x)}{p_r(x) + p_g(x)}] \right\} \tag{16}$$

$$= -\log 4 + 2JS(\mathbb{P}_r, \mathbb{P}_\theta), \tag{17}$$

where the last term is the Jensen-Shannon (JS) divergence.

In (Arjovsky et al., 2017), the JS divergence is replaced with Wasserstein distance. In the following section, we replace the JS divergence with RW divergence, and show that it would result in better performance.

4.1.2 Gradient descent and smoothness of RW divergence

In the training of GANs, Descent methods are typically used to minimize (17). Similarly, differentiability is needed for RW divergence in the RWGANs approach. As in WGANs, despite the theoretical explicit formulas derived in the duality representation and the gradient evaluation (Theorem 3.6 and Corollary 3.6.1), it is infeasible to directly compute such an f in practice. Nevertheless, since the Wasserstein divergence is parametrized by any strictly convex function in RWGANs, we obtain a great deal of flexibility in the choice of loss functions. For example, one can choose an appropriate ϕ such that

$$\nabla_\theta \left[W_{D_\phi}(\mathbb{P}_r, \mathbb{P}_\theta) \right] \approx \mathbb{E}_Z \left[\nabla_\theta f \left(\nabla \phi(g_\theta(z)) \right) \right].$$

For instance, one can try the KL divergence where $\nabla^2 \phi(x) = \text{diag}(1/x)$, observing that

$$\mathbb{E}_Z \left[[\nabla_\theta g_\theta(Z)]^\top \nabla^2 \phi(g_\theta(Z)) g_\theta(Z) \right] = \mathbb{E}_Z \left[[\nabla_\theta g_\theta(Z)]^\top \vec{1} \right] \leq C,$$

where C is a constant depending on the Lipschitz constant of g_θ. This implies that this term is controlled by θ during the process of training. The numerical results in section 4.1.4 confirm the effectiveness of our heuristic.

4.1.3 Experimental framework and settings

Experimental framework. The similarity between our experimental framework and the one in WGANs (Arjovsky et al., 2017) is: we apply back-propagation to train the generator and discriminator networks, and update the parameters once in the generative model and n_{critic} times in the discriminator network.

The differences between ours and the WGANs are: 1) we use $\nabla\phi$ to do the asymmetric clipping instead of the symmetric clipping. Note that the asymmetric clipping guarantees the Lipschitz continuity of f and $\nabla\phi(w) \in [-c, c]$; 2) we use a scaling parameter S to stabilize the asymmetric clipping. This is critical for the experiment since it reduces the variance of the gradient updates; 3) we adopt RMSProp (Tieleman and Hinton, 2012) instead of ADAM (Kingma and Ba, 2014), which allows a choice of a larger step-size and avoids the non-stationary problem (Mnih et al., 2016).

The details are described in Algorithm 1, where the boxed equation highlights the asymmetric clipping procedure, one of the key algorithmic differences between WGANs and RWGANs.

Algorithm 1 RWGANs. The default values $\alpha = 0.0005$, $c = 0.005$, $S = 0.01$, $m = 64$, $n_{critic} = 5$.

Require: α: the learning rate; c: the clipping parameter; m: the batch size; n_{critic}, the number of iterations of the critic per generator iteration; $N_{\max}$, the maximum number of one forward pass and one backward pass of all the training examples.

Require: w_0, initial critic parameters; θ_0: initial generator's parameters.

  for $N = 1, 2, \ldots, N_{\max}$ **do**
    for $t = 0, \ldots, n_{critic}$ **do**
      Sample a batch of real data $\{x_i\}_{i=1}^m$ from $\mathbb{P}_r$.
      Sample a batch of prior samples $\{z_i\}_{i=1}^m$ from $p(z)$.
      $g_w \leftarrow \frac{1}{m}\sum_{i=1}^m [\nabla_w f_w(x_i) - \nabla_w f_w(g_\theta(z_i))]$.
      $w \leftarrow w + \alpha \cdot \mathrm{RMSProp}(w, g_w)$.
      $\boxed{w \leftarrow \mathrm{clip}\left(w, -S \cdot (\nabla\phi)^{-1}(-c), S \cdot (\nabla\phi)^{-1}(c)\right).}$
    end for
    Sample a batch of prior samples $\{z_i\}_{i=1}^m$ from $p(z)$.
    $g_\theta \leftarrow -\frac{1}{m}\sum_{i=1}^m \nabla_\theta f_w(\nabla\phi(g_\theta(z_i)))$.
    $\theta \leftarrow \theta - \alpha \cdot \mathrm{RMSProp}(\theta, g_\theta)$.
  end for

Experimental settings. In order to test RWGANs, we adopt nine baseline methods as discussed in the introduction. They are RWGANs, WGANs (Arjovsky et al., 2017), WGANs-GP (Gulrajani et al., 2017), CGANs (Mirza and Osindero, 2014), InfoGANs (Chen et al., 2016), GANs (Goodfellow et al., 2014), LSGANs (Mao et al., 2016), DRAGANs (Kodali et al., 2017), BEGANs (Berthelot et al., 2017), EBGANs (Zhao et al., 2016), and ACGANs (Odena et al., 2017). The implementation of all these approaches is based on publicly available online information. In addition, we use the following four standard and well-known datasets in our experiment.

1. MNIST is a dataset of handwritten digits. It has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

2. Fashion-MNIST is an alternative dataset of Zalando's article images to MNIST. It consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 gray-scale image, associated with a label from 10 classes.

4.1.4 Experimental Results.

We start our experiment by training models using the ten different GANs procedures on MNIST and Fashion-MNIST. The architecture is DCGAN (Radford et al., 2015) and the maximum number of epochs is 100.

Figure 2 shows the training curves of the negative critic loss of all candidate approaches. The figure indicates that RWGANs and WGANs are stable with the smallest variances, where RWGANs has a slight higher variance partly due to the use of a larger step-size and asymmetric clipping. This slightly higher variance, nevertheless, speeds up the rate of training. Indeed, as illustrated in Figure 3. RWGANs is the fastest to generate meaningful images. Note that CGANs and InfoGANs seem faster but the images they have generated are not meaningful, as they fall into undesirable local optima in the optimization procedure.

4.2 Robust Optimization and Robust Portfolio Construction

In this section, we show the potential applications of RW to the robust optimization problems. Consider the following setting in robust optimization and machine learning:

$$\min_{\theta} \max_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(X, Y; \theta)]$$

which minimizes the loss function ℓ with available features X (and potentially labels Y in the supervised learning setting) over the parameter θ. The data follows probability distribution $\mathbb{P}$, which is allowed to vary inside an ambiguity set $\mathcal{P}$. In the data driven setting where iid samples $\{X_i\}_{i=1}^n$ are drawn from an underlying probability distribution $\mathbb{P}$, we consider ambiguity sets defined as the Relaxed-Wasserstein ball centered at the empirical distribution subject to certain constraints: $\mathcal{P} = \{\mathbb{P} : W_{D_\phi}(\mathbb{P}, \mathbb{P}_n) \leq \delta, \mathbb{E}_{\mathbb{P}}[h(X, \theta)] = 0\}$.

Now let X be a random variable in $\mathbb{R}^m$, with i.i.d copies $X_1, \ldots, X_n$, $\theta \in \mathbb{R}^l$ be the model parameter of interest, and $h(\cdot, \cdot)$ be the optimality condition of the parameter θ to be calibrated. Then one can easily extend Proposition 1 in Blanchet et al. (2016) to RW divergence, with little modification of the proof.

Proposition 2. *Let $h(\cdot, \theta) : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$ be Borel measurable and integrable, and $\Omega = \{(u, x) \in \mathbb{R}^m \times \mathbb{R}^m : D_\phi(u, x) < \infty\}$ be Borel measurable and non-empty. Further, suppose that 0 lies in the interior of the convex hull of $\{h(u, \theta) : u \in \mathbb{R}^m\}$. Define the Robust Wasserstein Profile (RWP) function as*

$$R_n(\theta) = \inf\{W_{D_\phi}(\mathbb{P}, \mathbb{P}_n) : \mathbb{E}_{\mathbb{P}}[h(X, \theta)] = 0\}.$$

Then

$$R_n(\theta) = \sup_{\lambda \in \mathbb{R}^r} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}^m} \{\lambda^T h(u, \theta) - D_\phi(u, X_i)\} \right\}.$$

If the limiting distribution of $R_n(\theta)$ is known, then inverting the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of the limiting distribution would give a $1 - \alpha$ confidence region for θ. The proposition above can be easily extended to Relaxed Wasserstein and yield meaningful results when the cost function of the optimal transport cost $c(u, x)$ is chosen to be Bregman divergence, i.e., $D_\phi(u, x)$.

Here we present two examples of special choices of ϕ. For simplicity, let $h(x, \theta) = x - \theta$. Choose $c(u, x) = D_\phi(u, x)$ for any strictly convex ϕ. Proposition 2 then implies

$$R_n(\theta) = \sup_{\lambda \in \mathbb{R}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \{\lambda(u - \theta) - D_\phi(u, X_i)\} \right\}$$

$$= \sup_{\lambda \in \mathbb{R}} \left\{ \lambda\theta - \frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \{\lambda u - \phi(u) + \phi(X_i) + \phi'(X_i)(u - X_i)\} \right\}.$$

We know

$$\sup_u \{\lambda u - \phi(u) + \phi'(X_i)u\} = \psi(\lambda + \phi'(W_i)),$$

where ψ is the convex conjugate of ϕ. Then

$$R_n(\theta) = \sup_{\lambda \in \mathbb{R}} \left\{ \lambda\theta - \frac{1}{n} \sum_{i=1}^{n} \{\phi(X_i) - \phi'(X_i)X_i + \psi(\lambda + \phi'(X_i))\} \right\} \tag{18}$$

Example 1. *If $\phi(x) = x^2$, which corresponds to the L^2 distance, then Equation (18) is reduced to*

$$R_n(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \theta) \right)^2.$$

Applying the Central Limit Theorem, $R_n(\theta)$ converges to a χ^2 distribution.

Example 2. *Take $\phi(x) = x \log x - x$, which corresponds to the KL divergence, then $\phi'(x) = \log x$, $\psi(x) = \psi'(x) = e^x$. The first order condition for λ in Equation (18) gives us*

$$\theta = \frac{1}{n} \sum_{i=1}^{n} \psi'(\lambda + \phi'(X_i)).$$

Solve for λ and we get

$$\lambda = \log \theta - \log(\frac{1}{n} \sum_{i=1}^{n} X_i).$$

Then

$$R_n(\theta) = \theta \log \theta - \theta - \theta \log \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) + \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which by the Central Limit Theorem and the continuous mapping theorem, it converges to a normal distribution plus the logarithm of a normal distribution. The parameters of the limiting distribution can be easily estimated using data. Moreover, to construct a $1 - \alpha$ confidence region for θ, one only needs to find the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of the limiting distribution.

Now we can specialize to the mean-variance portfolio construction problem, which is to construct a portfolio with a required expected return such that the risk, measured by the variance of the portfolio, is minimized. A robust version of this mean variance portfolio problem is to consider all possible distributions for the returns of assets in the ambiguity set, and then to optimize under the worst case scenario.

Mathematically, let $\pi \in \mathbb{R}^d$ be the vector of the weights of d assets in the portfolio, $Var_{\mathbb{P}}(R) \in \mathbb{R}^{d \times d}$ be the variance of the vector of returns R under a probability measure $\mathbb{P}$, $\mathcal{U}(\mathbb{P}_n) = \{\mathbb{P} : W_{D_\phi}(\mathbb{P}, \mathbb{P}_n) \leq \delta\}$ be the ambiguity set, which is a RW ball centered at the empirical distribution $\mathbb{P}_n$, and $\mathcal{F}_{\delta,\alpha}(n) = \{\pi : \pi^T 1 = 1, \min_{\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)}[E_{\mathbb{P}}(\pi^T R)] \geq \alpha\}$ be the feasible region of π such that the portfolio has minimum return of α in the worst case. If $\phi(x) = x^2$, then according to Theorem 1 in Blanchet et al. (2018), the following duality result holds:

$$\min_{\pi \in \mathcal{F}_{\delta,\alpha}(n)} \max_{\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)} \pi^T Var_{\mathbb{P}}(R)\pi = \min_{\pi \in \mathcal{F}_{\delta,\alpha}(n)} \left(\sqrt{\pi^T Var_{\mathbb{P}_n}(R)\pi} + \sqrt{\delta}\|\pi\|_2 \right)^2. \tag{19}$$

That is to say, the distributionally robust optimization problem of minimizing portfolio variance is equivalent to minimizing the standard deviation of the portfolio under the empirical distribution, plus an L^2 penalization term.

5   Conclusion

We propose a novel class of statistical divergence called RW divergence and establish several important theoretical properties. Numerical experiments, with RW parametrized by the KL divergence in image generation, show that RWGANs is a promising trade-off between WGANs and WGANs-GP, achieving both the robustness and efficiency during the learning process. The asymmetric clipping in RWGANs is a viable alternative to the gradient penalty and the symmetric clipping in WGANs, avoiding the low-quality samples and the failure of convergence. We also discuss a potential application of RW divergence in the context of robust optimization and explain how it can be used to construct ambiguity sets.

The flexible framework of RW divergences raises a natural question on whether one can select ϕ according to the data and the structure of the problem. This question is partially addressed by Proposition 1 and Theorem 3.7: with the objective of variance stabilization, a reasonable choice of the Bregman divergence is the Mahalanobis distance with the corresponding covariance matrix being the estimated Fisher information matrix.

While we highlight only the applications of RW to GANs and robust optimization, we believe that the theoretical results of RW divergence can be a valuable addition to the rich theory for optimal transport, where regularities of Wasserstein-based cost functions have been extensively studied (Caffarelli, 1991, 1992;

Chen and Figalli, 2017; Villani, 2008). With the extension of Bregman divergence to the functional space (Frigyik et al., 2008), the application of RW divergence in martingale optimal transport is also promising.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. *ICML*, pages 214–223, 2017.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705–1749, 2005b.

G. Bayraksan and D. K. Love. Data-driven stochastic programming using phi-divergences. *Tutorials in Operations Research*, 2015.

D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks. *ArXiv Preprint: 1703.10717*, 2017.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *ArXiv Preprint: 1610.05627*, 2016.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *ArXiv Preprint: 1802.04885*, 2018.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

L. Caffarelli. Some regularity properties of solutions of Monge Ampère equation. *Communications on Pure and Applied Mathematics*, 44(8-9):965–969, 1991.

L. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

S. Chen and A. Figalli. Partial $W^{2,p}$ regularity for optimal transport maps. *Journal of Functional Analysis*, 272(11):4588–4605, 2017.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *NIPS*, pages 2172–2180, 2016.

T De Wet. Goodnes-of-fit tests for location and scale families based on a weighted l 2-wasserstein distance measure. *Test*, 11(1):89–107, 2002.

E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. *NIPS*, pages 1486–1494, 2015.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *arXiv.org*, 2015.

L. Evans and R. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 2015.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional bregman divergence. *IEEE Int. Symp. Inf. Theory*, pages 1681–1685, 2008.

R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv.org*, April 2016.

A. Ghosh, V. Kulharia, A. Mukerjee, V. Namboodiri, and M. Bansal. Contextual RNN-GANs for abstract reasoning diagram generation. *AAAI*, pages 1382–1388, 2017.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, pages 2672–2680, 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *ArXiv Preprint: 1704.00028*, 2017.

R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, pages 1–37, 2012.

L. K. Jones and C. L. Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Transactions on Information Theory*, 1990.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ArXiv Preprint: 1412.6980*, 2014.

N. Kodali, J. Abernethy, J. Hays, and Z. Kira. How to train your DRAGAN. *ArXiv Preprint: 1705.07215*, 2017.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1): 229–260, 2007.

C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR, abs/1609.04802*, 2016.

P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *In NIPS Workshop on Adversarial Training*, 2016.

X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. *ArXiv Preprint: 1611.04076*, 2016.

P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *ArXiv Preprint: 1411.1784*, 2014.

V. Mnih, A. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *ICML*, pages 1928–1937, 2016.

Axel Munk and Claudia Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241, 1998.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. *ICML*, pages 2642–2651, 2017.

M. C. Pardo and I. Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, 49(7):1860–1868, 2003.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Arxiv Preprint: 1511.06434*, 2015.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. *ICML*, pages 1060–1069, 2016.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.

Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118*, 2017.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. *NIPS*, pages 613–621, 2016.

D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 2012.

R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *ArXiv Preprint: 1607.07539*, 2016.

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *ArXiv Preprint: 1609.03126*, 2016.

Ding Zhou, Jia Li, and Hongyuan Zha. A new mallows distance based metric for comparing clusterings. In *Proceedings of the 22nd international conference on Machine learning*, pages 1028–1035. ACM, 2005.

J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, pages 597–613, 2016.

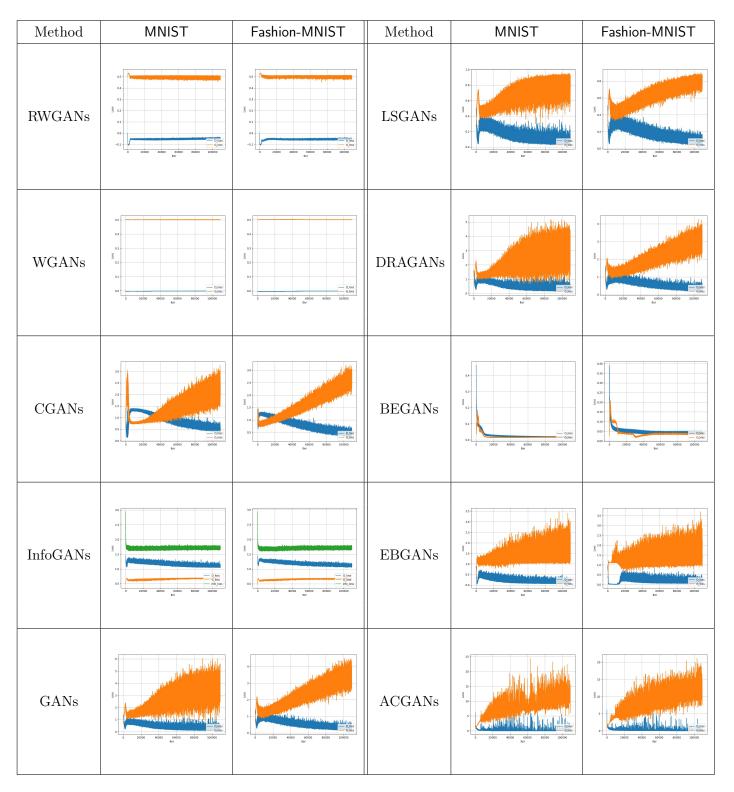| Method	MNIST	Fashion-MNIST	Method	MNIST	Fashion-MNIST
RWGANs			LSGANs		
WGANs			DRAGANs		
CGANs			BEGANs		
InfoGANs			EBGANs		
GANs			ACGANs		

Figure 2: Training curves of the negative critic loss at different stages of training on MNIST and Fashion-MNIST. G_{loss} and D_{loss} refer to the loss in generative and discriminative nets, which is plotted in orange and blue lines, respectively.

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
RWGANs				
WGANs				
CGANs				
InfoGANs				
GANs				

Method	$N = 1$	$N = 10$	$N = 25$	$N = 100$
LSGANs				
DRAGANs				
BEGANs				
EBGANs				
ACGANs				

Figure 3: Sample qualities at different stages of training on MNIST.