Towards Generalized Implementation of Wasserstein Distance in GANs

Minkai Xu*
University of Montreal
minkai.xu@umontreal.ca

Abstract

Wasserstein GANs (WGANs), built upon the Kantorovich-Rubinstein (KR) duality of Wasserstein distance, is one of the most theoretically sound GAN models. However, in practice it does not always outperform other variants of GANs. This is mostly due to the imperfect implementation of the Lipschitz condition required by the KR duality. Extensive work has been done in the community with different implementations of the Lipschitz constraint, which, however, is still hard to satisfy the restriction perfectly in practice. In this paper, we argue that the strong Lipschitz constraint might be unnecessary for optimization. Instead, we take a step back and try to relax the Lipschitz constraint. Theoretically, we first demonstrate a more general dual form of the Wasserstein distance called the Sobolev duality, which relaxes the Lipschitz constraint but still maintains the favorable gradient property of the Wasserstein distance. Moreover, we show that the KR duality is actually a special case of the Sobolev duality. Based on the relaxed duality, we further propose a generalized WGAN training scheme named Sobolev Wasserstein GAN, and empirically demonstrate the improvement over existing methods with extensive experiments.

1 Introduction

Generative adversarial networks (GANs) (Goodfellow et al. 2014) have attracted huge interest in both academia and industry communities due to its effectiveness in a variety of applications. Despite its effectiveness in various tasks, a common challenge for GANs is the training instability (Goodfellow 2016). In literature, many works have been developed to mitigate this problem (Arjovsky and Bottou 2017; Lucic et al. 2017; Heusel et al. 2017a; Mescheder, Nowozin, and Geiger 2017; Mescheder, Geiger, and Nowozin 2018; Yadav et al. 2017).

By far, it is well known that the problem of training instability of the original GANs mainly comes from the ill-behaving distance metric (Arjovsky and Bottou 2017), *i.e.*, the Jensen-Shannon divergence metric, which remains constant when two distributions are disjoint. The Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017) improves this by using the Wasserstein distance, which is able to continuously measure the distance between two distributions. Such a new objective has been shown to be effective in improving training stability.

In practice, since the primal form of the Wasserstein distance is difficult to optimize, the WGAN model (Arjovsky, Chintala, and Bottou 2017) instead proposed to optimize it with the Kantorovich-Rubinstein (KR) duality (Villani 2008). However, though the new WGAN scheme is theoretically more principled, it does not yield better performance in practice compared to other variants of GANs (Lucic et al. 2017). The main obstacle is that the WGAN requires the discriminator (or the critic) to be a Lipschitz function. However, this is very hard to satisfy in practice though a variety of different implementations have been tried such as weight clipping (Arjovsky, Chintala, and Bottou 2017), gradient penalty (GP) (Gulrajani et al. 2017), Lipschitz penalty (LP) (Petzka, Fischer, and Lukovnikov 2018) and spectral normalization (SN) (Miyato et al. 2018). As a result, WGAN is still unable to always achieve very compelling results.

In this paper, we argue that the strong Lipschitz condition might be unnecessary in the inner optimization loop for WGAN's critic. Intuitively, a looser constraint on the critic, which results in a larger function space, can simplify the practical constrained optimization problem of the restricted critic, and the better-trained critic would further benefit the training of the generator. Therefore, instead of developing new methods to impose the Lipschitz constraint, in this paper we propose to relax this constraint. In other words, we move our attention from "how to better implement the Lipschitz constraint" to "how to loosen the Lipschitz constraint". More specifically, in this paper we demonstrate a new dual form of the Wasserstein distance where the Lipschitz constraint is relaxed to the Sobolev constraint (Adams and Fournier 2003; Mroueh et al. 2018). We further show that the new duality with the relaxed constraint indeed is a generalization of the KR duality, and it still keeps the gradient property of the Wasserstein distance. Based on this relaxed duality, we propose a generalized WGAN model called Sobolev Wasserstein GAN. To the best of our knowledge, among the restricted GAN models (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Mroueh et al. 2018; Mroueh and Sercu 2017; Bellemare et al. 2017; Adler and Lunz 2018), Sobolev Wasserstein GAN is the most relaxed one that can still avoid the training instability problem.

The main contributions of this paper can be summarized

[*] Work performed while at Shanghai Jiao Tong University.

as follows:

- We demonstrate the Sobolev duality of Wasserstein distance and demonstrate that the new duality is also capable of alleviating the training instability problem in GANs (Section 3.1). We further clarify the relation between Sobolev duality and other previous metrics and highlight that by far Sobolev duality is the most relaxed metric that can still avoid the non-convergence problem in GANs (Section 3.2).
- Based on Sobolev duality, we introduce the Inequality Constraint Augmented Lagrangian Method (Nocedal and Wright 2006) to build the practical Sobolev Wasserstein GAN (SWGAN) training algorithm (Section 4).

We conduct extensive experiments to study the practical performance of SWGAN. We find that generally our proposed model achieves better sample quality and is less sensitive to the hyper-parameters. We also present a theoretical analysis of the minor sub-optimal equilibrium problem common in WGAN family models, and further propose an improved SWGAN with a better convergence.

2 Preliminaries

2.1 Generative Adversarial Networks

Generative adversarial networks (Goodfellow et al. 2014) perform generative modeling via two competing networks. The generator network G learns to map samples from a noise distribution to a target distribution, while the discriminator network D is trained to distinguish between the real data and the generated samples. Then the generator G is trained to output images that can fool the discriminator D. The process is iterated. Formally, the game between the generator G and the discriminator D leads to the minimax objective:

$$\min_G \max_D \Big\{ \mathbb{E}_{x\sim P_r}[\log(D(x))] + \mathbb{E}_{z\sim P_z}[\log(1-D(G(z))] \Big\}, \tag{1}$$

where P_r denotes the distribution of real data and P_z denotes the noise distribution.

This objective function is proven to be equivalent to the Jensen-Shannon divergence (JSD) between the real data distribution P_r and fake data distribution P_g when the discriminator is optimal. Assuming the discriminator is perfectly trained, the optimal discriminator is as follows:

$$D^*(x) = \frac{P_r}{P_r + P_g}. \tag{2}$$

However, recently (Zhou et al. 2019) points out that the gradients provided by the optimal discriminator in vanilla GAN cannot consistently provide meaningful information for the generator's update, which leads to the notorious training problems in GANs such as gradient vanishing (Goodfellow et al. 2014; Arjovsky and Bottou 2017) and mode collapse (Che et al. 2016; Metz et al. 2016; Kodali et al. 2017b; Arora et al. 2017). This view would be clear when checking the gradient of optimal discriminator in Eq. (2): the value of the optimal discriminative function $D^*(x)$ at each point is independent of other points and only reflects the local densities of $P_r(x)$ and $P_g(x)$, thus, when the supports of the two distributions are disjoint, the gradient produced by a well-trained discriminator is uninformative to guide the generator (Zhou et al. 2019).

2.2 Wasserstein Distance

Let P_r and P_g be two data distributions on $\mathbb{R}^n$. The Wasserstein-1 distance between P_r and P_g is defined as

$$W(P_r, P_g) = \inf_{\pi\in\Pi(P_r,P_g)} \mathbb{E}_{(x_i,x_j)\sim\pi}[\|x_i - x_j\|], \tag{3}$$

where the coupling π of P_r and P_g is the probability distribution on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals P_r and P_g, and $\Pi(P_r, P_g)$ denotes the set of all joint distributions π. The Wasserstein distance can be interpreted as the minimum cost of transporting one probability distribution to another. The Kantorovich-Rubinstein (KR) duality (Villani 2008) provides a new way to evaluate the Wasserstein distance between distributions. The duality states that

$$\begin{aligned} W(P_r, P_g) = \sup\nolimits_f \Big\{ \mathbb{E}_{x\sim P_r}[f(x)] - \mathbb{E}_{x\sim P_g}[f(x)] \Big\}, \\ \textit{s.t.}\ \ f(x_i) - f(x_j) \le \|x_i - x_j\|,\ \forall x_i, \forall x_j. \end{aligned} \tag{4}$$

where the supremum is taken over all functions $f\colon \mathcal{X} \to \mathbb{R}$ whose Lipschitz constant is no more than one.

2.3 Wasserstein GAN

The training instability issues of vanilla GAN is considered to be caused by the unfavorable property of distance metric (Arjovsky and Bottou 2017), *i.e.*, the JSD remains constant when the two distributions are disjoint. Accordingly, (Arjovsky, Chintala, and Bottou 2017) proposed Wasserstein distance in the form of KR duality Eq. (4) as an alternative objective.

The Wasserstein distance requires to enforce the Lipschitz condition on the critic network D. It has been observed in previous work that imposing Lipschitz constraint in the critic leads to improved stability and sample quality (Arjovsky, Chintala, and Bottou 2017; Kodali et al. 2017b; Fedus et al. 2017; Farnia and Tse 2018). Besides, some researchers also found that applying Lipschitz continuity condition to the generator can benefit the quality of generated samples (Zhang et al. 2018; Odena et al. 2018). Formally, in WGAN family, with the objective being Wasserstein distance, the optimal critic f^* under Lipschitz constraint holds the following property (Gulrajani et al. 2017):

Proposition 1. *Let π^* be the optimal coupling in Eq.* (3)*, then the optimal function f^* in KR duality Eq.* (4) *satisfies that: let $x_t = tx_i + (1-t)x_j$ with $0 \le t \le 1$, if f^* is differentiable and $\pi^*(x, x) = 0$ for all x, then it holds that $P_{(x_i,x_j)\sim\pi^*}[\nabla f^*(x_t) = \frac{x_i - x_j}{\|x_i - x_j\|}] = 1$.*

This property indicates that for each coupling of generated datapoint x_j and real datapoint x_i in π^*, the gradient at any linear interpolation between x_i and x_j is pointing towards the real datapoint x_i with unit norm. Therefore, guided by the gradients, the generated sample x_j would move toward the real sample x_i. This property provides the explanation, from the gradient perspective, on why WGAN can overcome the training instability issue.

2.4 Sobolev Space

Let $\mathcal{X}$ be a compact space in $\mathbb{R}^n$ and let μ(x) to be a distribution defined on $\mathcal{X}$ as a dominant measure. Functions in the Sobolev space $W^{1,2}(\mathcal{X}, \mu)$ (Adams and Fournier 2003) can be written as:

$$W^{1,2}(\mathcal{X}, \mu) = \Big\{ f : \mathcal{X} \to \mathbb{R}, \textstyle\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 \mu(x) dx < \infty \Big\}. \tag{5}$$

Restrict functions to the Sobolev space $W^{1,2}(\mathcal{X}, \mu)$ vanishing at the boundary and denote this space by $W_0^{1,2}(\mathcal{X}, \mu)$, then the semi-norm in $W_0^{1,2}(\mathcal{X}, \mu)$ can be defined as:

$$\|f\|_{W_0^{1,2}(\mathcal{X}, \mu)} = \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 \mu(x) dx}. \tag{6}$$

Given the notion of semi-norm, we can define the Sobolev unit ball constraint as follows:

$$\begin{aligned} \mathcal{F}_S(\mathcal{X}, \mu) = \Big\{ & f : \mathcal{X} \to \mathbb{R},\ f \in W_0^{1,2}(\mathcal{X}, \mu), \\ & \|f\|_{W_0^{1,2}(\mathcal{X}, \mu)} \leq 1 \Big\}. \end{aligned} \tag{7}$$

Sobolev unit ball is a function class that restricts the square root of integral of squared gradient norms according to the dominant measure $\mu(x)$.

2.5 Sobolev GAN

After WGAN, many works are devoted to improving GAN model by imposing restrictions on the critic function. Typical instances are the GANs based on Integral Probability Metric (IPM) (Mroueh and Sercu 2017; Bellemare et al. 2017). Among them, Sobolev GAN (SGAN) (Mroueh et al. 2018) proposed using Sobolev IPM as the metric for training GANs, which restricts the critic network D in $\mathcal{F}_S(\mathcal{X}, \mu)$:

$$\mathcal{S}_\mu(P_r, P_g) = \sup_{f \in \mathcal{F}_S(\mathcal{X}, \mu)} \Big\{ \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_g}[f(x)] \Big\}. \tag{8}$$

The following choices of measure μ for $\mathcal{F}_S$ are considered, which we will take as our baselines:

- $\mu = \frac{P_r + P_g}{2}$: the mixed distribution of P_r and P_g;
- μ_{gp}: $x = tx_i + (1-t)x_j$, where $x_i \sim P_r$, $x_j \sim P_g$ and $t \sim$ U$[0, 1]$, *i.e.*, the distribution defined by the interpolation lines between P_r and P_g as in (Gulrajani et al. 2017).

Let F_{P_r} and F_{P_g} be the cumulative distribution functions (CDF) of P_r and P_g respectively, and assume that the n partial derivatives of F_{P_r} and F_{P_g} exist and are continuous. Define the differential operator $D^- = (D^{-1}, ..., D^{-n})$ where $D^{-i} = \dfrac{\partial^{n-1}}{\partial x_1 ... \partial x_{i-1} \partial x_{i+1} ... \partial x_n}$, which computes $(n-1)$ high-order partial derivative excluding the i-th dimension. Let $x^{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$. According to (Mroueh et al. 2018), the Sobolev IPM in Eq. (8) has the following equivalent form:

$$\mathcal{S}_\mu(P_r, P_g) = \frac{1}{n} \sqrt{\mathbb{E}_{x \sim \mu} \sum_{i=1}^{n} \left(\frac{D^{-i} F_{P_r}(x) - D^{-i} F_{P_g}(x)}{\mu(x)} \right)^2}. \tag{9}$$

Note that, for each i, $D^{-i} F_P(x)$ is the cumulative distribution of the variable X_i given the other variables $X^{-i} = x^{-i}$ weighted by the density function of X^{-i} at x^{-i}, *i.e.*,

$$D^{-i} F_P(x) = P_{[X^{-i}]}(x^{-i}) F_{P_{[X_i | X^{-i} = x^{-i}]}}(x_i). \tag{10}$$

Thus, the Sobolev IPM can be seen as a comparison of coordinate-wise conditional CDFs. Furthermore, (Mroueh et al. 2018) also proves that the optimal critic f^* in SGAN holds the following property:

$$\nabla_x f^*(x) = \frac{1}{n \mathcal{S}_\mu(P_r, P_g)} \frac{D^- F_{P_g}(x) - D^- F_{P_r}(x)}{\mu(x)}. \tag{11}$$

3 Sobolev Duality of Wasserstein Distance

3.1 Sobolev Duality

Let x_i and x_j be two points in $\mathbb{R}^n$. The linear interpolation between x_i and x_j can be written as $x = tx_i + (1-t)x_j$ with $0 \leq t \leq 1$. Regarding x as a random variable on the line between x_i and x_j, we can then define its probability distribution as $\mu^{x_i, x_j}(x)$, which we will later use as the dominant measure for Sobolev space. Formally, let t be the random variable that follows the uniform distribution U$[0, 1]$. Then $\mu^{x_i, x_j}(x)$ can be written as:

$$\mu^{x_i, x_j}(x) = \begin{cases} \dfrac{1}{\|x_i - x_j\|}, & x = tx_i + (t-1)x_j, \\ 0, & otherwise. \end{cases} \tag{12}$$

With the above defined notation, we propose our new dual form of Wasserstein distance as follows, which we call *Sobolev duality*[1]

$$\begin{aligned} & W(P_r, P_g) = \sup\nolimits_f \Big\{ \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_g}[f(x)] \Big\}, \\ & \textit{s.t. } f \in \mathcal{F}_S(\mathcal{X}, \mu^{x_i, x_j}),\ \forall x_i \sim P_r, \forall x_j \sim P_g, \end{aligned} \tag{13}$$

where $\mathcal{F}_S(\mathcal{X}, \mu^{x_i, x_j})$ denotes the Sobolev unit ball of

$$\|f\|_{W_0^{1,2}(\mathcal{X}, \mu^{x_i, x_j})} = \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 \mu^{x_i, x_j}(x) dx} \leq 1. \tag{14}$$

Note that the support of $\mu^{x_i, x_j}(x)$ is the straight line between x_i and x_j. Thus $\|f\|_{W^{1,2}(\mathcal{X}, \mu^{x_i, x_j})}$ is the square root of the path integral of squared gradient norms from x_i to x_j, *i.e.*, the constraint is restricting the gradient integral on each line between P_r and P_g to be no more than 1.

Corresponding to Proposition 1 of KR duality, we highlight the following property of the Sobolev duality:

[1] We provide the proofs of *Sobolev duality* and Proposition 2 in Appendix.

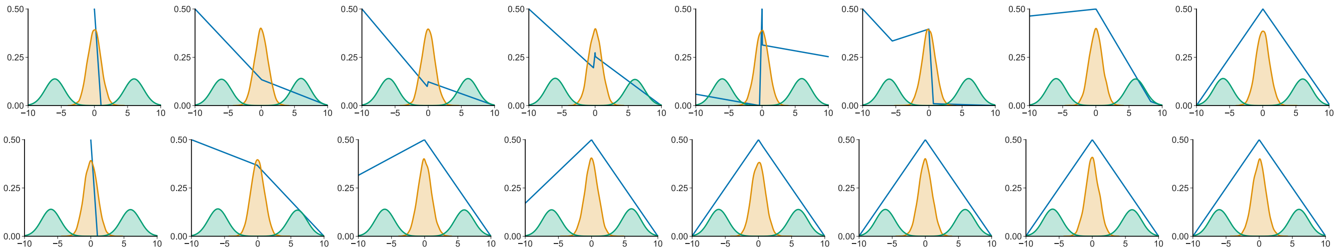


Figure 1: 1D Training comparison. Upper: WGAN. Lower: SWGAN. Orange: real data, sampled from $N(0,1)$. Green: fake data, sampled from $\frac{1}{2}(N(-5,1)+N(5,1))$. Blue: the re-scaled critic, which is normalized to $[0,0.5]$. From left to right, different columns correspond to iteration $0, 30, 60, 90, 120, 180, 240, 300$ respectively. The critic of SWGAN holds a faster and smoother convergence.

Proposition 2. *Let π^* be the optimal coupling in Eq.* (3), *then the optimal function f^* in Sobolev duality Eq.* (13) *satisfies that: let $x_t = tx_i + (1-t)x_j$ with $0 \leq t \leq 1$, if f^* is differentiable and $\pi^*(x,x) = 0$ for all x, then it holds that* $P_{(x_i,x_j)\sim\pi^*}[\nabla f^*(x_t) = \frac{x_i - x_j}{\|x_i - x_j\|}] = 1$.

That is, with Sobolev duality, the gradient direction for every fake datum is the same as WGAN. Hence, enforcing the Sobolev duality constraint on discriminator can be an effective alternative of the Lipschitz condition to guarantee a stable training for GAN model.

3.2 Relation to Other Metrics

Relation to KR Duality in Eq. (4). As indicated by Proposition 2, the optimal critic f^* of Sobolev duality actually holds the same gradient property as KR duality in Proposition 1. However, as clarified below, the constraint in Sobolev duality is indeed looser than KR duality, which would potentially benefit the optimization.

In the classic KR duality , f is restricted under Lipschitz condition, *i.e.*, the gradient norms of *all points* in the metric space are enforced to no more than 1. By contrast, in our Sobolev duality, we restrict *the integral* of squared gradient norms *over each line* between P_r and P_g. This implies that Lipschitz continuity is a sufficient condition of the constraint in Sobolev duality. In summary, Sobolev duality is a generalization of KR duality where the constraint is relaxed, while still keeps the same property of training stability.

Relation to Sobolev IPM in Eq. (8). We now clarify the difference between Sobolev IPM in Eq. (8) and Sobolev duality of Wasserstein distance in Eq. (13). In the former metric, when implementing μ_{gp} (defined in Section 2.5), the *total integral* of squared gradient norms on all interpolation lines between P_r and P_g is enforced to no more than 1; while in the latter metric, the integral *over each interpolation line* between P_r and P_g is restricted. Therefore, Sobolev duality enforces stronger constraint than Sobolev IPM.

However, we should also note that the stronger constraint is necessary to ensure the favorable gradient property in Proposition 2. By contrast, as shown in Eq. (11), Sobolev IPM measures coordinate-wise conditional CDF, which cannot always provide gradients as good as the optimal transport plan in Wasserstein distance. A toy example is provided in Appendix to show the case that Sobolev IPM is sometimes insufficiently constrained to ensure the convergence.

4 Sobolev Wasserstein GAN

Now we define the GAN model with Sobolev duality, which we name as Sobolev Wasserstein GAN (SWGAN). Formally, SWGAN can be written as:

$$\begin{aligned}\min_G \max_D \mathcal{L}_S(D_w, G_\theta) \\ = \mathbb{E}_{x\sim P_r} D_w(x) - \mathbb{E}_{z\sim P_z} D_w(G_\theta(z)),\end{aligned} \tag{15}$$

with the constraint that

$$\mathbb{E}_{x\sim\mu^{x_i,x_j}} \|\nabla_x D_w(x)\|^2 \leq 1, \forall x_i \sim P_r, \forall x_j \sim P_g, \tag{16}$$

where μ^{x_i,x_j} is the interpolation distribution on lines between pairs of points x_i and x_j as defined in Eq. (12).

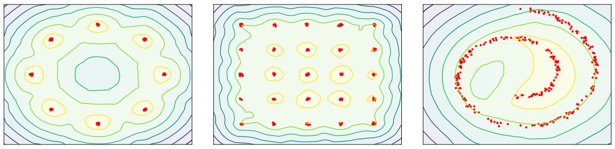
Let Ω_{ij} denote $1 - \mathbb{E}_{x\sim\mu^{x_i,x_j}} \|\nabla_x D_w(x)\|^2$, then the constraint is to restrict Ω_{ij} to be greater than or equal to 0 for all the pairs of (x_i, x_j). Inspired by (Mroueh et al. 2018), we define the following Augmented Lagrangian inequality regularization (Nocedal and Wright 2006) corresponding to SWGAN ball constraints:

$$\begin{aligned}\mathcal{L}_{al}^{(ij)}(w,\theta,\alpha) &= \alpha(\Omega_{ij} - s_{ij}) - \frac{\rho}{2}(\Omega_{ij} - s_{ij})^2, \\ \mathcal{L}_{al}(w,\theta,\alpha) &= \mathbb{E}_{x_i\sim P_r}\mathbb{E}_{x_j\sim P_g}\mathcal{L}_{al}^{(ij)}(w,\theta,\alpha).\end{aligned} \tag{17}$$

where α is the Lagrange multiplier, ρ is the quadratic penalty weight and s_{ij} represents the slack variables. Practically, s_{ij} is directly substituted by its optimal solution:

$$s_{ij}^* = \max\left\{\Omega_{ij} - \frac{\alpha}{\rho}, 0\right\}. \tag{18}$$

As in (Arjovsky, Chintala, and Bottou 2017) and (Mroueh et al. 2018), the regularization term in Eq. (17) is added to the loss only when training the critic. To be more specific, the training process is: given the generator parameters θ, we train the discriminator by maximizing $\mathcal{L}_S + \mathcal{L}_{al}$; then given the discriminator parameters w, we train the generator via minimizing $\mathcal{L}_S$. We leave the detailed training procedure in Appendix.



(a) 8 Gaussians

(b) 25 Gaussians

(c) Swiss Roll

Figure 2: Level sets of SWGANs critic. Yellow corresponds to high values and purple to low. The training samples are indicated in red and the generated distribution is fixed at the real data plus Gaussian noise.

5 Experiments

We tested SWGAN on both synthetic density modeling and real-world image generation task.[2]

5.1 Synthetic Density Modeling

1D Distribution Modeling. Displaying the level sets is a standard qualitative approach to evaluate the learned critic function for two-dimensional data sets (Gulrajani et al. 2017; Kodali et al. 2017a; Petzka, Fischer, and Lukovnikov 2018). Here, we consider both real data distribution P_r and generated fake distribution P_g are fixed simple one-dimensional Gaussian distributions. Our goal is to investigate whether the critics of both WGAN and SWGAN can be efficiently optimized to provide the favorable gradients presented in Proposition 1 and 2. We observed that while both critics can be trained to the theoretical optimum, the latter one can always enjoy a faster convergence compared with the former one. An empirical example is visualized in Fig. 1. As shown here, with the same initial states and hyper-parameters, the critic of SWGAN holds a faster and smoother convergence towards the optimal state. This meaningful observation verifies our conjecture that a larger function space of the critic would benefit the training.

Level Sets of the Critic. In this section we give another 2D level sets visualization. As analyzed in Section 3.1, Sobolev constraint in SWGAN is a generalization of Lipschitz constraint. Therefore, theoretically SWGAN critic should also be capable of modeling more challenging real and fake data and providing meaningful gradients. To demonstrate this, we train SWGAN critics to optimality on several toy distributions. The value surfaces of the critics are plotted in Figure 2, which shows good fitness of the distribution following our theoretical analysis.

5.2 Real-world Image Generation

Experimental Setup **Controlled variables.** To make the comparisons more convincing, we also include extended versions of the existing GAN models to control the contrastive variables. The controlled variables includes:

- Sampling size. In SWGAN-AL we need to sample m points on each interpolation line between P_r and P_g, while in WGAN-GP (Gulrajani et al. 2017) and SGAN (Mroueh et al. 2018) only one point is sampled. To yield a more fair comparison, we perform additional experiments of WGAN and SGAN with the sampling size equal to m.

Generator $G(z)$

Operation	Kernel size	Resample	Output Dims
Noise	N/A	N/A	128
Linear	N/A	N/A	$128 \times 4 \times 4$
Residual block	$[3 \times 3] \times 2$	Up	$128 \times 8 \times 8$
Residual block	$[3 \times 3] \times 2$	Up	$128 \times 16 \times 16$
Residual block	$[3 \times 3] \times 2$	Up	$128 \times 32 \times 32$
Conv, tanh	3×3	N/A	$3 \times 32 \times 32$

Critic $D(x)$

Operation	Kernel size	Resample	Output Dims
Residual block	$[3 \times 3] \times 2$	Down	$128 \times 16 \times 16$
Residual block	$[3 \times 3] \times 2$	Down	$128 \times 8 \times 8$
Residual block	$[3 \times 3] \times 2$	N/A	$128 \times 8 \times 8$
Residual block	$[3 \times 3] \times 2$	N/A	$128 \times 8 \times 8$
ReLU, mean pool	N/A	N/A	128
Linear	N/A	N/A	1

Table 1: ResNet architecture.

- Optimization method. In our baseline WGAN-GP (Gulrajani et al. 2017), the restriction is imposed by Penalty Method (PM). By contrast, SGAN and SWGAN-AL use Augmented Lagrangian Method (ALM). ALM is a more advanced algorithm than PM for strictly imposing the constraint. To see the practical difference, we add experiment settings of SWGAN with penalty regularization term (named SWGAN-GP). Formally, the penalty can be written as:

$$\mathcal{L}_{gp}(w, \theta) = -\lambda\, \mathbb{E}_{x_i \sim P_r} \mathbb{E}_{x_j \sim P_g} \Omega_{ij}^2(D_w, G_\theta), \qquad (19)$$

  where λ is the gradient penalty coefficient. $\mathcal{L}_{gp}$ is the alternative term of the ALM penalty $\mathcal{L}_{al}$ in Eq. (17) for the training of SWGAN-GP.

Baselines. For comparison, we also evaluated the WGAN-GP (Gulrajani et al. 2017) and Sobolev GAN (SGAN) (Mroueh et al. 2018) with different sampling sizes and penalty methods. The choice of baselines is due to their close relation to SWGAN as analyzed in Section 3.2. We omit other previous methods since as a representative of state-of-the-art GAN model, WGAN-GP has been shown to rival or outperform a number of former methods, such as the original GAN (Goodfellow et al. 2014), Energy-based generative adversarial network (Zhao, Mathieu, and LeCun 2016), the original WGAN with weight clipping (Arjovsky, Chintala, and Bottou 2017), Least Squares GAN (Mao et al. 2017), Boundary equilibrium GAN (Berthelot, Schumm, and Metz 2017) and GAN with denoising feature matching (Warde-Farley and Bengio 2016).

Evaluation metrics. Since GAN lacks the capacity to perform reliable likelihood estimations (Theis, Oord, and Bethge 2015), we instead concentrate on evaluating the quality of generated images. We choose to compare the maximal Frechet Inception Distances (FID) (Heusel et al.

[2] Code is available at https://github.com/MinkaiXu/SobolevWassersteinGAN.

GANs	CIFAR-10		Tiny-ImageNet	
	IS	FID	IS	FID
WGAN-GP*	7.85±.07	18.21±.12	8.17±.03	18.70±.05
WGAN-GP with $m = 8$	7.88±.09	18.08±.22	8.17±.04	18.69±.10
WGAN-AL	7.79±.09	17.86±.16	8.26±.03	18.70±.06
WGAN-AL with $m = 8$	7.89±.09	17.52±.27	8.31±.02	18.61±.09
SGAN*, $\mu = \frac{P_r+P_g}{2}$	7.81±.11	17.89±.27	8.30±.04	18.90±.04
SGAN*, $\mu = \mu_{GP}$	7.83±.10	18.03±.24	8.31±.03	18.90±.08
SGAN, $\mu = \mu_{GP}$ with $m = 8$	7.86±.09	17.74±.24	8.33±.03	18.75±.07
SWGAN-GP	**7.98±.08**	17.50±.19	8.38±.03	18.50±.03
SWGAN-AL	7.93±.09	**16.75±.24**	**8.41±.03**	**18.32±.05**

* denotes the vanilla version of our baselines.

Table 2: Performance of GANs on CIFAR-10 and Tiny-ImageNet.

2017b) and Inception Scores (Salimans et al. 2016) reached during training iterations, both computed from 50K samples. A high image quality corresponds to high Inception and low FID scores. Specifally, IS is defined as $\exp(\mathbb{E}_x \operatorname{KL}(p(y|x)||p(y)))$, where $p(y|x)$ is the distribution of label y conditioned on generated data x, and $p(y)$ is the marginal distribution. IS combines both the confidence of the class predictions for each synthetic images (quality) and the integral of the marginal probability of the predicted classes (diversity). The classification probabilities were estimated by the Inception model (Salimans et al. 2016), a classifier pre-trained upon the ImageNet dataset (Deng et al. 2009). However, in practice we note that IS is hard to detect the mode collapse problems. FID use the same Inception model to capture computer-vision-specific features of a collection of real and generated images, and then calculate the Frechet distance (also called $Wasserstein$-2 distance) (Aronov et al. 2006) between two activation distributions. The intuition of IS is that high-quality images should lead to high confidence in classification, while FID is aiming to measure the computer-vision-specific similarity of generated images to real ones through Frechet distance (*a.k.a.*, *Wasserstein*-2 distance) (Aronov et al. 2006).

Data. We test different GANs on CIFAR-10 (Krizhevsky, Hinton et al. 2009) and Tiny-ImageNet (Deng et al. 2009) , which are standard datasets widely used in GANs literatures. Both datasets consist of tens of thousands of real-world color images with class labels.

Network architecture. For all experimental settings, we follow WGAN-GP (Gulrajani et al. 2017) and adopt the same Residual Network (ResNet) (He et al. 2016) structures and hyperparameters. Specifically, the generator and critic are residual networks. (Gulrajani et al. 2017) use pre-activation residual blocks with two 3×3 convolutional layers each and ReLU nonlinearity. Batch normalization is used in the generator but not the critic. Some residual blocks perform downsampling (in the critic) using mean pooling after the second convolution, or nearest-neighbor upsampling (in the generator) before the second convolution. Formally, we present our ResNet architecture in Table 1. Further architectural details can be found in our open-source model.

Other implementation details. For SWGAN metaparameter, we choose 8 as the sample size m. Adam optimizer (Kingma and Ba 2014) is set with learning rate decaying from $2 \cdot 10^{-4}$ to 0 over 100K iterations with $\beta_1 = 0, \beta_2 = 0.9$. We used 5 critic updates per generator update, and the batch size used was 64.

Results We here also introduce WGAN with Augmented Lagrangian Method (WGAN-AL) for further comparison, which is similar to SGAN (Mroueh et al. 2018). Scores in terms of FID and IS on CIFAR-10 and Tiny-ImageNet are reported in Table 2. Some representative samples from the resulting generator of SWGAN are provided in Fig. 3 and Fig. 4. Some representative samples from the resulting generator of SWGAN are provided in Appendix. In experiments, we note that IS is remarkably unstable during training and among different initializations, while FID is fairly stable.

From Table 2, we can see that SWGANs generally work

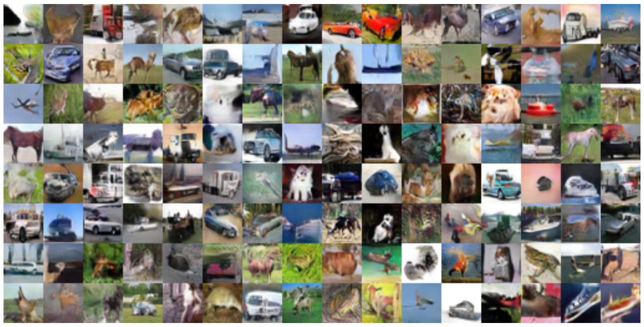


Figure 3: Generated CIFAR-10 samples.

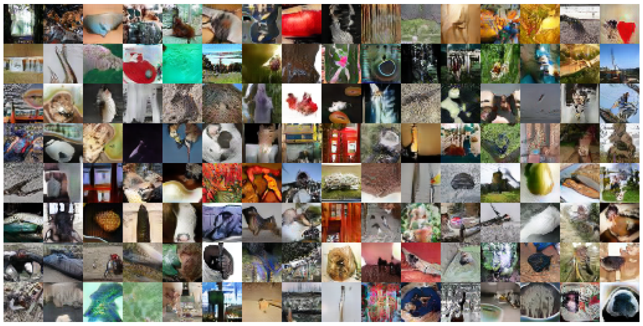


Figure 4: Generated Tiny-ImageNet samples.

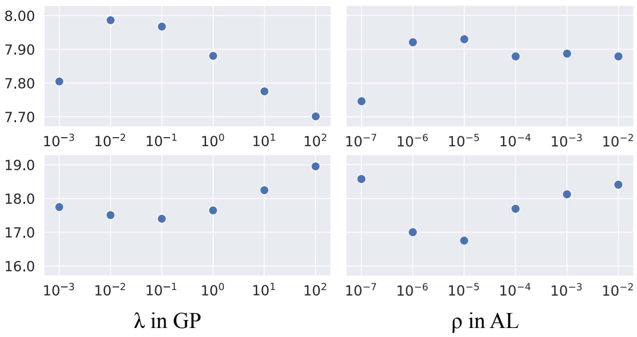


Figure 5: The comparison of SWGANs with different regularization terms and parameters. Top: Inception scores. Bottom: Frechet Inception Distances.

better than the baseline models. The experimental results also show that WGAN and SGAN tend to have slightly better performance when using ALM or sample more interpolation points. However, compared with SWGAN-AL and SWGAN-GP, the performances in these cases are still not competitive enough. This indicates that the larger sampling size and ALM optimization algorithm are not the key elements for the better performance of SWGAN, *i.e.*, these results evidence that *it is the relaxed constraint in Sobolev duality that leads to the improvement*, which is in accordance with our motivation that a looser constraint would simplify the constrained optimization problem and lead to a stronger GAN model.

We further test SWGAN with different regularization terms and parameters on CIFAR-10. The scores are shown in Figure 5. As shown in Figure 5, generally ALM is a better choice when considering FID, while GP is better for IS. A meaningful observation is that SWGAN is not sensitive to different values of penalty weights ρ and λ. By contrast, a previous large scale study reported that the performance of WGAN-GP holds strong dependence on the penalty weight λ (see Figure 8 and 9 in (Lucic et al. 2017)). This phenomenon demonstrates a more smooth and stable convergence and well-behaved critic performance throughout the whole training process of SWGAN.

6 Conclusion

In this paper, we proposed a new dual form of Wasserstein distance with the Lipschitz constraint relaxed and demonstrate that it is still capable of eliminating the training instability issues. This new dual form leads to a generalized WGAN model. We built Sobolev Wasserstein GAN based on the proposed duality and provided empirical evidence that our GAN model outperforms the previous approaches, which either impose the strong Lipschitz penalty or cannot theoretically guarantee the convergence.

This work was motivated by the intuition that with a less restricted function space, the critic would be easier to be trained to optimum, thus benefiting the training of GANs. To the best of our knowledge, Sobolev Wasserstein GAN is the GAN model with the most relaxed restriction that can still avoid the training instability problem. In the future, we hope that practitioners can take a step back and investigate whether we can further relax the constraint imposed on the function space of the critic and what the minimal requirement for the convergence guarantee could be.

Acknowledgments

We would like to thank Zhiming Zhou, Guansong Lu, Weinan Zhang and Jian Tang for the extremely helpful discussions and comments. The Shanghai Jiao Tong University (SJTU) team is supported by "New Generation of AI 2030" Major Project (2018AAA0100900) and National Natural Science Foundation of China (62076161, 61772333, 61632017).

References

Adams, R. A.; and Fournier, J. J. 2003. *Sobolev spaces*, volume 140. Elsevier.

Adler, J.; and Lunz, S. 2018. Banach wasserstein gan. In *Advances in Neural Information Processing Systems*, 6754–6763.

Arjovsky, M.; and Bottou, L. 2017. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223.

Aronov, B.; Har-Peled, S.; Knauer, C.; Wang, Y.; and Wenk, C. 2006. Fréchet distance for curves, revisited. In *European Symposium on Algorithms*, 52–63. Springer.

Arora, S.; Ge, R.; Liang, Y.; Ma, T.; and Zhang, Y. 2017. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573* .

Bellemare, M. G.; Danihelka, I.; Dabney, W.; Mohamed, S.; Lakshminarayanan, B.; Hoyer, S.; and Munos, R. 2017. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743* .

Berthelot, D.; Schumm, T.; and Metz, L. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* .

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2016. Mode Regularized Generative Adversarial Networks. *arXiv preprint arXiv:1612.02136* .

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Farnia, F.; and Tse, D. 2018. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems*, 5248–5258.

Fedus, W.; Rosca, M.; Lakshminarayanan, B.; Dai, A. M.; Mohamed, S.; and Goodfellow, I. 2017. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446* .

Goodfellow, I. 2016. Advances in neural information processing systems 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160* .

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 5767–5777.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017a. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Klambauer, G.; and Hochreiter, S. 2017b. GANs trained by a two time-scale update rule converge to a Nash equilibrium. *arXiv preprint arXiv:1706.08500* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017a. How to train your DRAGAN. *arXiv preprint arXiv:1705.07215* 2(4).

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017b. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215* .

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2017. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337* .

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which Training Methods for GANs do actually Converge? In *International Conference on Machine Learning*.

Mescheder, L.; Nowozin, S.; and Geiger, A. 2017. The numerics of gans. In *Advances in neural information processing systems*.

Metz, L.; Poole, B.; Pfau, D.; and Sohl-Dickstein, J. 2016. Unrolled Generative Adversarial Networks. *arXiv preprint arXiv:1611.02163* .

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=B1QRgziT-.

Mroueh, Y.; Li, C.-L.; Sercu, T.; Raj, A.; and Cheng, Y. 2018. Sobolev GAN. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SJA7xfb0b.

Mroueh, Y.; and Sercu, T. 2017. Fisher gan. In *Advances in neural information processing systems*.

Nocedal, J.; and Wright, S. 2006. *Numerical optimization*. Springer Science & Business Media.

Odena, A.; Buckman, J.; Olsson, C.; Brown, T. B.; Olah, C.; Raffel, C.; and Goodfellow, I. 2018. Is generator conditioning causally related to gan performance? *arXiv preprint arXiv:1802.08768* .

Petzka, H.; Fischer, A.; and Lukovnikov, D. 2018. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=B1hYRMbCW.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*.

Theis, L.; Oord, A. v. d.; and Bethge, M. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844* .

Villani, C. 2008. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg. ISBN 9783540710509. URL https://books.google.com/books?id=hV8o5R7_5tkC.

Warde-Farley, D.; and Bengio, Y. 2016. Improving generative adversarial networks with denoising feature matching. In *International Conference on Learning Representations*.

Yadav, A.; Shah, S.; Xu, Z.; Jacobs, D.; and Goldstein, T. 2017. Stabilizing Adversarial Nets With Prediction Methods. *arXiv preprint arXiv:1705.07364* .

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* .

Zhao, J.; Mathieu, M.; and LeCun, Y. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* .

Zhou, Z.; Liang, J.; Song, Y.; Yu, L.; Wang, H.; Zhang, W.; Yu, Y.; and Zhang, Z. 2019. Lipschitz Generative Adversarial Nets. In *International Conference on Machine Learning*.