Understanding the Effectiveness of Lipschitz Regularization in the Discriminator of Generative Adversarial Nets

Zhiming Zhou HEYOHAIZHOU@GMAIL.COM
Department of Computer Science
Shanghai University of Finance and Economics
100 Wudong Road, Yangpu District, Shanghai 200433, China

Jiadong Liang JDLIANG@PKU.EDU.CN
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Haidian District, Beijing 100871, China

Yong Yu YYU@APEX.SJTU.EDU.CN
Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dongchuan Road, Minhang District, Shanghai 200240, China

Zhihua Zhang ZHZHANG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Haidian District, Beijing 100871, China

Abstract

In this paper, we study the convergence of GANs (generative adversarial nets) from the perspective of optimal discriminative function. We show that GANs without regularization in the discriminative function space commonly suffer from an issue that the gradients from the discriminator can be uninformative to guide the generator, while Lipschitz regularization in the discriminative function space can generally resolve the gradient uninformativeness issue and guarantee GANs' convergence. We provide a thorough theoretical study upon Lipschitz regularization and Lipschitz regularized GANs, which deepens the current understanding of Lipschitz regularization for GANs and dissects how Lipschitz regularization resolve the gradient uninformativeness issue. We expand on the theoretical and practical implications of the gradient uninformativeness issue, elaborating on how it causes training instability and mode collapse, and how common practices against these issues take effect. Based on these analyses, we identify and clarify the key elements of the convergence of GANs.

Keywords: Lipschitz regularization, generative adversarial nets, optimal discriminative function, gradient uninformativeness, convergence, training instability, mode collapse

1. Introduction

Generative Adversarial Nets (GANs, Goodfellow et al., 2014), as one of the most promising generative models, have been successfully applied in various related tasks. However, GANs are also well-known for their difficulties in training (Goodfellow, 2016). The common issues include training instability, mode collapse, low sample quality, etc. The underlying obstacles, though have been heavily studied (Arjovsky and Bottou, 2017; Mescheder et al., 2017, 2018; Metz et al., 2017; Unterthiner et al., 2018; Sun et al., 2020), are still not fully understood.

The objective of GANs is usually defined as a distance metric between the target (real) distribution P_t and the generation (fake) distribution P_g formed by generated samples, which implies that $P_g = P_t$ is the unique global optimum. The training issues of conventional GANs have been considered to stem from the illness of the distance metric (Arjovsky and Bottou, 2017), e.g., the Jensen–Shannon divergence (Goodfellow et al., 2014) between P_g and P_t keeps constant when their supports are disjoint. Arjovsky et al. (WGANs, 2017) accordingly suggested using the Wasserstein distance as the replacement, which can properly measure the distance between two distributions no matter whether their supports are disjoint.

To estimate the Wasserstein distance in its dual form for ease of computation, it is required to introduce a Lipschitz regularization in the discriminative function space of WGANs. Observing its immediate effectiveness in improving the training stability and sample quality, Lipschitz regularization has rapidly become popular in various GANs models (Gulrajani et al., 2017; Fedus et al., 2018; Miyato et al., 2018; Karras et al., 2018; Zhang et al., 2019). However, the underlying working mechanism of Lipschitz regularization is still obscure.

In this paper, we propose to study the convergence of GANs from the perspective of the gradients of the optimal discriminative function with respect to generated samples. By inspecting the optimal discriminative function and its gradient with respect to generated samples, i.e., inspecting the gradient flow at the connecting point between the generator and the discriminator, the understanding of GANs' training can be much more clear. The reason is that: (i) it takes the G-D structure of GANs (i.e., the generator and discriminator structure) into account; (ii) inspecting the samples instead of the probabilities fits the sample-based nature of GANs. In this view, GANs can be understood to work as follows: G models the samples to be updated and updates them according to D, while D, whose behavior is theoretically defined by the optimal discriminative function and practically affected by the training details, tells the generator how these samples should be updated via its gradient with respect to these samples.

We demonstrate that the theoretical convergence of GANs heavily depends on the regularization in the discriminative function space, i.e., whether there is a regularization in the discriminator and what regularization it is. We show that if there is no regularization in the discriminative function space, the GANs provably suffers from a gradient uninformativeness issue (which means in many cases that have been proved to commonly exist, the gradient that the generator receives from the optimal discriminator does not tell any information of the target distribution), and hence generally does not guarantee its convergence. We also demonstrate that not an arbitrary regularization in the discriminative function space can resolve this gradient uninformativeness issue.

However, interestingly, we find that Lipschitz regularization can generally resolve the gradient uninformativeness issue and guarantee GANs' convergence. We provide a provably valid construction of these Lipschitz regularized GANs, which turns out to be very mild and cover most popular choices of GANs' objective, and hence explains how Lipschitz regularization works when combined with these GANs.

We provide detailed analysis upon why Lipschitz regularization can generally resolve the gradient uninformativeness issue, showing that Lipschitz regularization makes the gradients of the optimal discriminative function with respect to the generated samples point directly towards real samples (i.e., samples in target distributions). We demonstrate the existence and uniqueness of the optimal discriminative function for GANs under Lipschitz regularization, prove that there is only a single Nash equilibrium between the optimal generative function and the optimal discriminative function, and show that otherwise the GANs will always move samples from locations where there is too much to locations where there is too little.

We identify the new family of GANs with Lipschitz regularization in the discriminative function space as Lipschitz regularized GANs and shorten as LGANs. In trying to attain the optimal discriminative function of LGANs, with which we can then verify its theoretical properties, we find various underlying issues in the current implementations of Lipschitz regularization. We hence provide a rigorous study upon the implementation of Lipschitz regularization, and thereby propose two revised versions with theoretical justification. On the basis of that, we construct several instances of LGANs, empirically verify their theoretical properties, and show their consistently superior performance over WGANs.

With the aforementioned line well elaborated, we further clarify the theoretical and practical implications of the gradient uninformativeness issue. We stress that gradient uninformativeness implies theoretically undefined sample updating behavior, while at the same time, we show that the common practices on the design of network architectures and the choice of hyperparameters typically lead to simple and smooth discriminative functions, which tend to make the theoretically undefined gradient point towards the target distribution. We suggest viewing these practical choices (or tricks) as forming implicit regularizations in the discriminative function space, hence mitigated the impact of the gradient uninformativeness issue. We conjecture that this could be the reason why these GANs without theoretical convergence guarantee, though being unstable and hard to use, could somehow work in practice with well-tuned training schemes and hyperparameters.

TODO
TODO
TODO
TODO
TODO

TODO
TODO
TODO
TODO
TODO

The remainder of this paper is organized as follows. In Section 2, we provide some preliminaries that will be used in this paper. In Section 3, we define and formalize the gradient uninformativeness issue. In Section 4, we present the Lipschitz regularized GANs and their theoretical properties. In Section 5, we study the implementation of Lipschitz regularization. In Section 6, we provide the empirical analysis of the Lipschitz regularized GANs. In Section 7, we study the theoretical and practical behaviors of the gradient uninformativeness issue. In Section 8, we study the formation of mode collapse and its relationship to gradient uninformativeness. In Section 9, we summarize and clarify the keys to convergence guarantee of GANs. Finally, we discuss related work in Section 10 and conclude the paper in Section 11.

2. Preliminaries

In this section, we first introduce some notions that will be used in this paper, including the Lipschitz continuity and the Wasserstein distance. We then present a generalized formulation for GANs whose discriminator takes a single sample as input, and introduce the key research object of this paper, i.e., the gradients that the generator receives from the discriminator with respect to generated samples. Related notations are also given therein.

2.1 Lipschitz Continuity and Lipschitz Constant

Given two metric spaces (X, d_X) and (Y, d_Y), a function $f\colon X \to Y$ is said to be k_0-Lipschitz continuous, if there exists a constant $k_0 \geq 0$ such that

$$d_Y(f(x_1), f(x_2)) \leq k_0 \cdot d_X(x_1, x_2), \forall\, x_1, x_2 \in X. \tag{1}$$

The smallest constant k_0 is called the Lipschitz constant of f, which we denote by $\kappa(f)$.

In this paper, the metrics d_X and d_Y are by default the Euclidean distance, i.e., the ℓ_2-norm of the displacement of the two parameters[1], which we denote by $\|\cdot\|$.

2.2 The Wasserstein Distance and its Duals

The first-order Wasserstein distance W_1 between two probability distributions is defined as

$$W_1(P_g, P_t) = \inf_{\pi \in \Pi(P_g, P_t)} \mathbb{E}_{(x,y)\sim\pi}\left[d(x, y)\right], \tag{2}$$

where $\Pi(P_g, P_t)$ denotes the set of all probability measures with marginals P_g and P_t. It can be interpreted as the minimum transport cost from the distribution P_g to the distribution P_t. Hence, π is usually called the transport plan. We use π^* to denote the optimal transport plan, and use S_g and S_t to denote the supports of P_g and P_t respectively.

The Kantorovich-Rubinstein (KR) duality (Villani, 2008) provides a more efficient way of computing the Wasserstein distance. The duality states that

$$\begin{aligned} W_1(P_g, P_t) = \sup_f\ & \mathbb{E}_{x\sim P_g}\left[f(x)\right] - \mathbb{E}_{x\sim P_t}\left[f(x)\right], \\ s.t.\ \ & f(x) - f(y) \leq d(x, y),\ \forall x, \forall y. \end{aligned} \tag{3}$$

1. When switching to other norms, the property of the gradients will get changed. Different norms will induce gradients with different properties. See Appendix A.5 for some basic arguments on this.

The constraint in Eq. (3) implies that f is Lipschitz continuous with $\kappa(f) \leq 1$.

Interestingly, we find a more compact dual form of the Wasserstein distance. That is,

$$\begin{aligned} W_1(P_g, P_t) &= \sup_f \mathbb{E}_{x\sim P_g}[f(x)] - \mathbb{E}_{x\sim P_t}[f(x)], \\ s.t.\ \ & f(x) - f(y) \leq d(x,y),\ \forall x \in S_g, \forall y \in S_t. \end{aligned} \tag{4}$$

This new dual form relaxes the Lipschitz continuity condition from over the entire space to respectively over their supports. The proof for this dual form is given in Appendix A.1.

As shown in WGANs-GP (Gulrajani et al., 2017), the gradient of the optimal discriminative function in the KR dual form of the Wasserstein distance has the following property:

Proposition 1 *Let π^* be the optimal transport plan in Eq. (2) and $x_t = t \cdot x + (1-t) \cdot y$ with $0 \leq t \leq 1$. If the optimal discriminative function f^* in Eq. (3) is differentiable and $\pi^*(x, x) = 0$ for all x, then it holds that*

$$\mathbb{P}_{(x,y)\sim\pi^*}\left[\nabla_{x_t} f^*(x_t) = \frac{y-x}{\|y-x\|}\right] = 1. \tag{5}$$

This proposition implies: (i) for each generated sample x, there exists a real sample y such that $\nabla_{x_t} f^*(x_t) = \frac{y-x}{\|y-x\|}$ for all linear interpolations x_t between x and y, which means the gradient at any linear interpolation x_t is pointing towards the real sample y with a unit length; (ii) these (x, y) pairs match the optimal transport plan π^*, i.e., the directions of $\nabla_x f^*(x)$ indicate the optimal transport; (iii) WGANs does not suffer from the gradient vanishing, and the gradients are not uninformative.

2.3 Generalized Formulation of Generative Adversarial Nets

Typically, GANs, whose discriminator takes a single sample as input, can be formulated as

$$\begin{aligned} &\min_{f\in\mathcal{F}}\ J_D(f) = \mathbb{E}_{z\sim P_s}[\phi(f(g(z)))] + \mathbb{E}_{x\sim P_t}[\varphi(f(x))], \\ &\min_{g\in\mathcal{G}}\ J_G(g) = \mathbb{E}_{z\sim P_s}[\psi(f(g(z)))], \end{aligned} \tag{6}$$

where P_s is the source distribution of the generator in $\mathbb{R}^m$ (e.g., Gaussian noise) and P_t is the target (real) distribution in $\mathbb{R}^n$. The generative function $g\colon \mathbb{R}^m \to \mathbb{R}^n$ learns to output samples that share the same dimension and characteristics as samples in P_t, while the discriminative function $f\colon \mathbb{R}^n \to \mathbb{R}$ learns to output a real-valued score indicating the authenticity of a given sample in the target space. ϕ, φ, $\psi\colon \mathbb{R} \to \mathbb{R}$ are the loss metrics[2]. $\mathcal{F}$ and $\mathcal{G}$ denote the discriminative and generative function spaces, respectively.

We denote the implicit distribution of the generated samples by P_g, and use f^* to denote the optimal discriminative function, i.e., $f^* = \arg\min_{f\in\mathcal{F}} J_D(f)$.

We list the choices of ϕ, φ, $\mathcal{F}$, and the corresponding f^* of several representative GANs in Table 1. Without loss of generality, the basic common pattern of these GANs is that the

2. By *loss metric*, we refer to the $\mathbb{R} \to \mathbb{R}$ function ϕ, φ, ψ, or their combination. When referring to the entire J_D and/or J_G, with/without with the regularization(s) in $\mathcal{F}$ and/or $\mathcal{G}$, we will use the word *objective*.

	ϕ	φ	$\mathcal{F}$	$f^*(x)$
Original GANs	$-\log(\sigma(-x))$	$-\log(\sigma(x))$	Free	$\log \frac{P_t(x)}{P_g(x)}$
Least-Squares GANs	$(x-a)^2$	$(x-b)^2$	Free	$\frac{a\cdot P_g(x)+b\cdot P_t(x)}{P_g(x)+P_t(x)}$
μ-Fisher GANs	x	$-x$	$\mathbb{E}_{x\sim\mu}\lvert f(x)\rvert^2 \le 1$	$\frac{P_g(x)-P_t(x)}{\mathcal{F}_\mu(P_g,P_t)\cdot\mu(x)}$
Wasserstein GANs	x	$-x$	$\kappa(f)\le 1$	N/A
Lipschitz Regularized GANs	any ϕ and φ satisfying Eq. (12)		$\kappa(f)$ regularized	N/A

Table 1: The comparison of different objectives of GANs.

discriminator forces $f(x)$ to be small for generated samples, while forces $f(x)$ to be large for real samples. Typically, ψ is chosen to be the negative of ϕ, forming a minimax formulation. We introduce a free ψ to make the framework more general. Different choices of ϕ, φ, ψ, $\mathcal{F}$, and $\mathcal{G}$, if valid, lead to different distance metrics and hence, in a sense, different GANs.

For subsequent needs, we let

$$\begin{aligned} \mathring{J}_D(f,x) &= P_g(x)\phi(f(x)) + P_t(x)\varphi(f(x)), \\ \dot{J}_G(x) &= \psi(f(x)). \end{aligned} \tag{7}$$

It has $J_D(f) = \int \mathring{J}_D(f,x)\,\mathrm{d}x$ and $J_G(g) = \int P_g(x)\dot{J}_G(x)\,\mathrm{d}x$.

In these GANs, the gradient that the generator receives from the discriminator with respect to a generated sample $x \in S_g$ is

$$\nabla_x \dot{J}_G(x) = \nabla_x \psi(f(x)) = \nabla_{f(x)}\psi(f(x)) \cdot \nabla_x f(x), \tag{8}$$

where the first term $\nabla_{f(x)}\psi(f(x))$ is a scalar, and the second term $\nabla_x f(x)$ is a vector with the same dimension as x which indicates the direction that the generator should follow for optimizing the generated sample x.

2.4 The Gradient Vanishing

The gradient vanishing issue has been considered as a critical phenomenon that indicates the existence of training issues in GANs. For original GANs, when the discriminator is trained to optimum, $\nabla_{f(x)}\psi(f(x))$ tends to become zero. Goodfellow et al. (2014) suggested bypassing it by using a revised loss metric ψ for the generator. Actually, only the scalar $\nabla_{f(x)}\psi(f(x))$ is changed. The Least-Squares GANs (Mao et al., 2017), which aims at addressing the gradient vanishing issue, also focused on $\nabla_{f(x)}\psi(f(x))$. We can actually show that the Least-Squares GANs, as well as the revised original GANs, may still suffer from vanishing gradient due to the zeroness of $\nabla_x f(x)$. For example, in a region where $\frac{P_t(x)}{P_g(x)}$ is constant (see Figure 14b).

Arjovsky and Bottou (2017) provided a new perspective for understanding the gradient vanishing, concerning the overall properties of distance metrics. They argued that S_g and S_t are usually disjoint, and the gradient vanishing stems from the illness of conventional distance metrics when S_g and S_t are disjoint, e.g., the JS divergence between P_g and P_t would remain constant. The Wasserstein distance was thus proposed by Arjovsky et al. (2017) as a replacement to conventional distance metrics, which can properly measure the distance between two distributions no matter whether their supports are disjoint or not.

3. The Gradient Uninformativeness

In this paper, we will pay our main attention to the gradient direction, which is more interesting and more important than the gradient scale. We will consider the optimal discriminative function $f^*(x)$ and its gradient $\nabla_x f^*(x)$. $\nabla_x f^*(x)$ means along which the generator will be told by the well-optimized discriminator to update the generated sample x.

We show that for many distance metrics and hence many GANs, such a gradient may fail to bring any useful information about P_t. Consequently, P_g is not guaranteed to converge to P_t. We name this phenomenon as the *gradient uninformativeness* and argue that it is a fundamental cause of nonconvergence and instability in the training of conventional GANs.

The gradient uninformativeness is substantially different from the gradient vanishing. The gradient vanishing is about the scalar term $\nabla_{f(x)}\psi(f(x))$ or the overall scale of $\nabla_x \dot{J}_G(x)$, while the gradient uninformativeness is about the direction of $\nabla_x \dot{J}_G(x)$, which is entirely defined by $\nabla_x f^*(x)$. The two issues are orthogonal, though they sometimes exist simultaneously.

Next, we discuss the gradient uninformativeness in the taxonomy of the regularization in the discriminative function space. We will show that: (i) for unregularized GANs, the gradient uninformativeness commonly exists; (ii) for GANs with regularization, such an issue may still exist; (iii) with Lipschitz regularization, the issue generally does not exist.

3.1 Unregularized GANs: The Issue Commonly Exists

For many GANs, there is no regularization in the discriminative function space. Typical instances include f-divergence based GANs, such as the original GANs (Goodfellow et al., 2014) and Least-Squares GANs (Mao et al., 2017).

In these GANs, the value of the optimal discriminative function at each point is independent of other points and only reflects the local densities $P_g(x)$ and $P_t(x)$:

$$f^*(x) = \underset{f(x)\in\mathbb{R}}{\arg\min}\ P_g(x)\phi(f(x)) + P_t(x)\varphi(f(x)),\ \forall x. \tag{9}$$

Given that $f^*(x)$ only reflects the local densities $P_g(x)$ and $P_t(x)$, for each generated sample x that is not surrounded by real samples (formally, there exists $\epsilon > 0$ such that for all y with $0 < \|y - x\| < \epsilon$, it holds that $y \notin S_t$), $f^*(x)$ in the surroundings of x would contain no information about P_t.

Thus, we can claim that $\nabla_x f^*(x)$, the gradient that the generator receives from the optimal discriminative function for updating this sample, does not reflect any information about P_t. Hence, there is no guarantee upon whether the generator can update the sample towards getting closer to the target distribution, nor the overall convergence.

The typical situation is that S_g and S_t are disjoint, which is common in practice according to Arjovsky and Bottou (2017). That is, the gradient uninformativeness commonly exists in unregularized GANs.

To further distinguish the gradient uninformativeness from the gradient vanishing, we consider an ideal case: S_g and S_t are totally overlapped, both consisting of n discrete points, while their probability masses over these points are different. Checking Eq. (9) for this case,

we can see that $\nabla_x f^*(x)$ for each generated sample is still uninformative, because there is no real sample around. But the gradient does not vanish and is actually undefined[3].

3.2 Regularized GANs: The Issue May Still Exist

There also exists GANs that impose regularization in the discriminative function space. Typical instances include the Integral Probability Metric (IPM) based GANs (Mroueh and Sercu, 2017; Mroueh et al., 2018; Bellemare et al., 2017) and the Wasserstein GANs (Arjovsky et al., 2017). We next show that GANs with regularization in the discriminative function space might also suffer from the gradient uninformativeness.

The optimal discriminative function of the μ-Fisher IPM $\mathcal{F}_\mu(P_g, P_t)$, i.e., the f^* of the generalized objective of the Fisher GANs (Mroueh et al., 2018), has the following form:

$$f^*(x) = \frac{1}{\mathcal{F}_\mu(P_g, P_t)} \frac{P_g(x) - P_t(x)}{\mu(x)}, \tag{10}$$

where $\mathcal{F}_\mu(P_g, P_t)$ is the measured distance between the two distributions which can be regarded as a constant, and μ is a distribution whose support covers S_g and S_t.

It can be observed that $f^*(x)$ of the μ-Fisher IPM, in its core, also only reflects the local densities, and its gradient does not reflect any useful information about P_t at other locations. Similar to the above, we can conclude that for each generated sample x that is not surrounded by real samples, $\nabla_x f^*(x)$ is uninformative to guide the generator.

4. Lipschitz Regularized GANs

Lipschitz regularization has recently become popular in GANs. It has been constantly observed that introducing Lipschitz continuity as a regularization in the discriminator leads to improved stability and sample quality (Arjovsky et al., 2017; Gulrajani et al., 2017; Fedus et al., 2018; Miyato et al., 2018; Karras et al., 2018; Zhang et al., 2019).

In this section, we investigate the generalized formulation of GANs with Lipschitz regularization, where the Lipschitz constant of the discriminative function is penalized, to theoretically analyze the properties of such GANs.

In particular, we define the Lipschitz regularized Generative Adversarial Nets (LGANs) as:

$$\begin{aligned} &\min_{f\in\mathcal{F}} \mathbb{E}_{z\sim P_s}[\phi(f(g(z)))] + \mathbb{E}_{x\sim P_t}[\varphi(f(x))] + \frac{\rho}{2}\cdot\kappa(f)^\alpha, \\ &\min_{g\in\mathcal{G}} \mathbb{E}_{z\sim P_s}[\psi(f(g(z)))]. \end{aligned} \tag{11}$$

We will show that, if $\rho > 0$ and $\alpha > 1$, then once the following condition holds, the proposed LGANs is a well-defined objective of GANs, which generally resolves the gradient

3. See Sections 7 and 8 for a deeper understanding of this issue.

uninformativeness issue and has theoretical guarantee on its convergence:

$$\begin{cases} \phi \text{ is convex}, \\ \varphi \text{ is convex}, \\ \exists\, a,\ \phi'(a) = -\varphi'(a) \neq 0. \end{cases} \tag{12}$$

This condition is a generalized version of the one given in Zhou et al. (2019). If we force $\phi(x) = \varphi(-x)$, which is typical in practice, the condition can be simplified to

$$\begin{cases} \phi \text{ is convex}, \\ \phi'(0) \neq 0. \end{cases} \tag{13}$$

This condition is very straightforward and mild. Requiring ϕ to be increasing means that the discriminator should force small $f(x)$ for generated samples. Requiring φ to be decreasing means that the discriminator should force large $f(x)$ for real samples. The other constraints are included because, otherwise, this problem is not guaranteed to have a solution.

Note that in WGANs, the loss metric $\phi(x) = \varphi(-x) = x$ is used, which satisfies Eq. (12). There are many other instances satisfying Eq. (12), e.g., $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$, $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + 1}$, and $\phi(x) = \varphi(-x) = \exp(x)$.

Note that it is trivial to find more ϕ and φ by rescaling and offsetting along the axes. The linear combination of two or more ϕ or φ also keep satisfying Eq. (12). We illustrate some of these loss metrics in Figure 1.

There also exist loss metrics used in GANs that do not satisfy Eq. (12), e.g., the quadratic loss (Mao et al., 2017) and the hinge loss (Zhao et al., 2017; Lim and Ye, 2017; Miyato et al., 2018). Nevertheless, we will discuss and study them in the experiments (see Section 6).

Note that $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$ corresponds to the loss metric of the original GANs. As we have shown, the original GANs suffers from the gradient uninformativeness issue. However, as we will next show, when imposing the Lipschitz regularization, the resulting model, as a specific instance of LGANs, behaves very well.

4.1 Theoretical Analysis of Lipschitz Regularized GANs

We now present the theoretical analysis of LGANs. The intention of the analysis is two folds. The first is to verify that the formulation is a valid one. The second is to elaborate how the gradient uninformativeness issue is resolved. All proofs are deferred to Appendix A.

For the first fold, we will demonstrate the existence and uniqueness of the optimal discriminative function for GANs under Lipschitz regularization, and prove that there is only a single Nash equilibrium between the optimal generative function and the optimal discriminative function where $P_g = P_t$, and show that, otherwise, the GANs will always move samples from locations where there is relatively too much to locations where there is relatively too little.

For the second fold, we will, all along the way, provide detailed analysis upon why Lipschitz regularization can generally resolve the gradient uninformativeness issue, and at last show

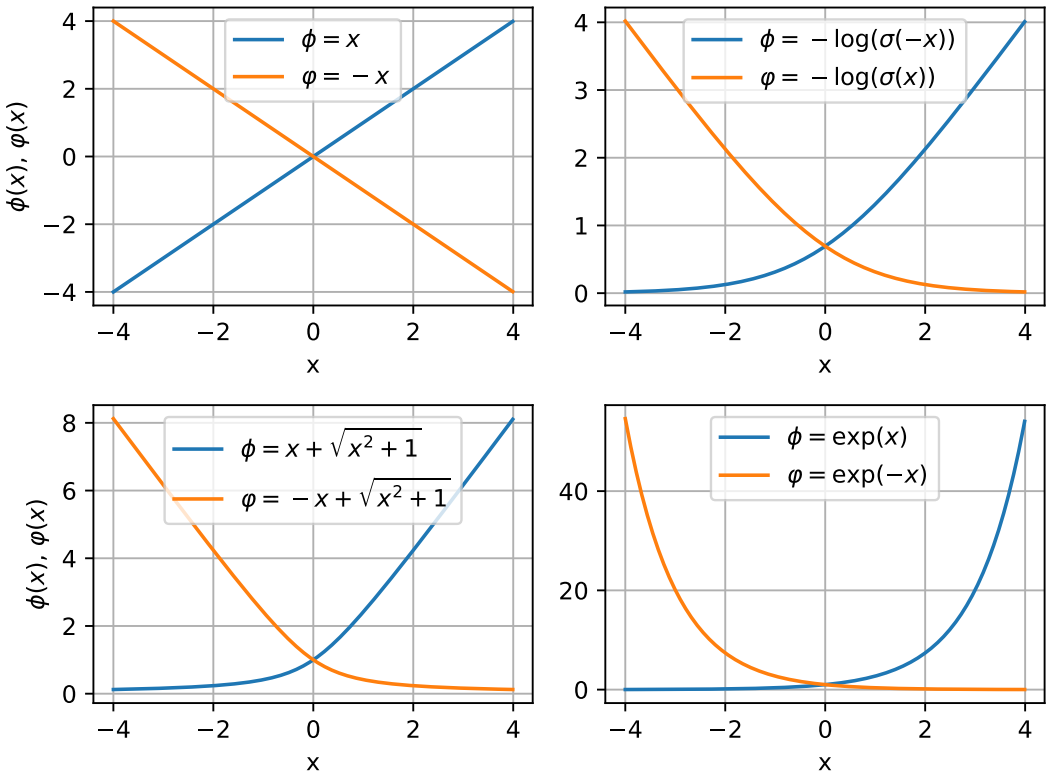


Figure 1: Exemplifying ϕ and φ that satisfies Eq. (12).

that Lipschitz regularization makes the gradients of the optimal discriminative function with respect to generated samples point directly towards real samples (i.e., samples in target distributions), hence, in a sense, being remarkably informative.

4.1.1 Existence and Uniqueness of the Optimal Discriminative Function

First, we consider the existence and uniqueness of the optimal discriminative function.

Theorem 2 *Under Assumption* (12)*, the optimal discriminative function f^* of Eq.* (11) *exists. Moreover, if ϕ or φ is strictly convex, it is unique.*

Note that for LGANs with WGANs' loss metric, i.e., $\phi(x) = \varphi(-x) = x$, which does not satisfy the condition that ϕ or φ is strictly convex, the solution of Eq. (11), i.e., the optimal discriminative function f^*, still exists but is not unique. Specifically, if f^* is an optimal solution, then $f^* + d$ for any $d \in \mathbb{R}$ is also an optimal solution. And this is actually the only special case. For all other instances that satisfy Eq. (12), ϕ or φ is strictly convex.

4.1.2 Unique Nash Equilibrium and the Existence of Bounding Relationships

The following theorems can be regarded as a generalization of Proposition 1 of WGANs to LGANs, with more detailed analysis of bounding relationships and equilibrium.

Theorem 3 *Under Assumption* (12)*, we have the optimal discriminative function f^* exists. If we further assume f^* is smooth, we have:*

(a) For all $x \in S_g \cup S_t$, if it holds that $\nabla_{f^*(x)} \mathring{J}_D(f^*, x) \neq 0$, then there exists $y \in S_g \cup S_t$ with $y \neq x$ such that $|f^*(y) - f^*(x)| = \kappa(f^*) \cdot \|y - x\|$;

(b) For all $x \in S_g \cup S_t - S_g \cap S_t$, there exists $y \in S_g \cup S_t$ with $y \neq x$ such that $|f^*(y) - f^*(x)| = \kappa(f^*) \cdot \|y - x\|$;

(c) If $S_g = S_t$ and $P_g \neq P_t$, then there exists (x, y) pair in $S_g \cup S_t$ with $y \neq x$ such that $|f^*(y) - f^*(x)| = \kappa(f^*) \cdot \|y - x\|$ and $\nabla_{f^*(x)} \mathring{J}_D(f^*, x) \neq 0$;

(d) There is a unique Nash equilibrium between g^* and f^*, where $P_g = P_t$ and $\kappa(f^*) = 0$.

This theorem states the basic properties of LGANs, including: (i) the existence of unique Nash equilibrium where $P_g = P_t$ and $\kappa(f^*) = 0$; (ii) the existence of *bounding relationships* in the optimal discriminative function, i.e., $\exists y \neq x$ such that $|f^*(y) - f^*(x)| = \kappa(f^*) \cdot \|y - x\|$. The former ensures that the objective is a well-defined distance metric, and the latter, as we will next show, eliminates the gradient uninformativeness issue.

It is worth noticing that penalizing $\kappa(f)$, instead of simply restricting the maximum of $\kappa(f)$ as in WGANs, is indeed necessary for Theorem 3-(c) and Theorem 3-(d). It is due to the existence of cases where $\nabla_{f^*(x)} \mathring{J}_D(f^*, x) = 0$ for x with $P_g(x) \neq P_t(x)$ when the loss metric is not $\phi(x) = \varphi(-x) = x$, i.e., when the loss metric is strictly convex.

Minimizing $\kappa(f)$, in any case, guarantees that the Nash equilibrium is achieved only when $P_g = P_t$. With WGANs' loss metric, minimizing $\kappa(f)$ is not necessary. However, if $\kappa(f)$ is

not minimized towards zero, $\nabla_x f^*(x)$ is not guaranteed to be zero when $P_g = P_t$ (Mescheder et al., 2018). This implies that minimizing $\kappa(f)$ also benefits WGANs.

4.1.3 The Properties of Bounding Relationship

From Theorem 3, we know, as long as P_g still has not fully converged to P_t, there must exist point x with $f^*(x)$ being bounded by another point y, such that $|f^*(y)-f^*(x)| = \kappa(f^*)\cdot\|y-x\|$.

We here further clarify that, when there is a bounding relationship, it must involve both real sample(s) and fake sample(s). And surely, because the discriminator forces large values for real samples and small values for generated samples, in a bounding relationship, under a fully optimized discriminative function, the values of real samples should always be larger.

In particular, we have:

Theorem 4 *Under the conditions in Theorem 3, we have*

1) *For any $x \in S_g$, if $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$, then there must exist some $y \in S_t$ with $y \neq x$ such that $f^*(y) - f^*(x) = \kappa(f^*)\cdot\|y - x\|$ and $\nabla_{f^*(y)}\mathring{J}_D(f^*, y) < 0$;*
2) *For any $y \in S_t$, if $\nabla_{f^*(y)}\mathring{J}_D(f^*, y) < 0$, then there must exist some $x \in S_g$ with $y \neq x$ such that $f^*(y) - f^*(x) = \kappa(f^*)\cdot\|y - x\|$ and $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$.*

The intuition behind the above theorem is that samples from the same distribution, e.g., the generated samples, will not bound each other to violate the optimality of $\mathring{J}_D(f^*, x)$. So, when there is a strict bounding relationship, i.e., if it involves points that hold $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) \neq 0$, it must involve both real and generated samples.

It is worth noting that, if it is the disjoint case, all fake samples hold $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$, while all real samples hold $\nabla_{f^*(y)}\mathring{J}_D(f^*, y) < 0$. And in any case, $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$ implies $x \in S_g$, though x might at the same time belong to S_t. Nevertheless, $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$ means f^* can assign a lower value to x to better optimize the objective if it is not bounded by some other sample. Similarly, $\nabla_{f^*(y)}\mathring{J}_D(f^*, y) < 0$ implies $y \in S_t$ and f^* can assign a higher value to x.

Furthermore, $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) > 0$ and $\nabla_{f^*(y)}\mathring{J}_D(f^*, y) < 0$, implies, $\frac{P_t(x)}{P_g(x)} < \frac{P_t(y)}{P_g(y)}$. That is, in a sense, the bounding relationship is bridging a location x where generated samples are relatively too much to a location y where generated samples are relatively too little.

Note that there might exist a chain of bounding relationships that involves a dozen of fake samples and real samples, and every two bound each other. It can be shown that these points all lie in the same line in the value surface of f^*, i.e., in the space of $(x, f^*(x))$, and bound each other.

The bounding line in the value surface of f^* is the basic building block that connects P_g and P_t, and each generated sample with $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) \neq 0$ lies in at least one of these bounding lines. We will next show that these bounding lines that connect P_g and P_t provide reliable channels for guiding the samples in P_g to move to P_t.

4.1.4 The Implication of Bounding Relationship

Recall that Proposition 1 states $\nabla_{x_t} f^*(x_t) = \frac{y-x}{\|y-x\|}$. We next show that this is actually a direct consequence of the bounding relationship between x and y, i.e., a bounding relationship guarantees meaningful $\nabla_x f^*(x)$ for all involved points, making it point towards real samples. We formally state it as follows:

Theorem 5 *Assume function f is Lipschitz continuous and differentiable. Then for all x and y that satisfy $y \neq x$ and $f(y) - f(x) = \kappa(f) \cdot \|y - x\|$, we have $\nabla_{x_t} f(x_t) = \kappa(f) \cdot \frac{y-x}{\|y-x\|}$ for all $x_t = t \cdot x + (1-t) \cdot y$ with $0 \leq t \leq 1$.*

In other words, if two points x and y bound each other in terms of $f(y) - f(x) = \kappa(f) \cdot \|y - x\|$, there is a straight line between x and y in the value surface of f. At any point in this straight line, the gradient holds the maximum norm $\kappa(f)$ and the gradient direction is pointing towards the $x \to y$ direction.

Note that the requirement of differentiability of f^* is not critical. It does not hold only when one sample is bounded by multiple bounding lines, i.e., being bounded in different directions, hence forming multiple sub-gradients and being non-differentiable.

Note that the type of the norm is critical for this theorem. We assumed ℓ_2-norm in this theorem and the entire paper, which is the typical choice in practice. When switching to other norms, e.g., ℓ_1-norm, the gradient may have different properties. See Appendix A.5 for more details.

4.1.5 Summing Up

Combining Theorems 2, 3, 4 and 5, we can conclude that, as long as $\rho > 0$ and $\alpha > 1$, and the loss metrics ϕ and φ satisfy the condition Eq. (12), and f^* is smooth and differentiable:

- According to Theorem 3-(a), when S_g and S_t are disjoint, the gradient of the optimal (or practically well-trained) discriminative function for each generated sample $x \in S_g$ points towards a certain real sample $y \in S_t$, which guarantees that $\nabla_x f^*(x)$-based generator update would tend to move P_g towards P_t at every step.
- In fact, Theorem 3 provides further guarantee on the convergence. Theorem 3-(b) implies that, for any generated sample $x \in S_g$ that does not lie in S_t, its gradient under optimal discriminative function must point towards a certain real sample $y \in S_t$.
- In the fully overlapped case, according to Theorem 3-(c), unless $P_g = P_t$, there must exist at least one pair of (x, y) in strict bounding relationship and $\nabla_x f^*(x)$ pulls x towards y. We can further claim that it is moving sample from location where there is too much (with relatively low $\frac{P_t(x)}{P_g(x)}$) to location where there is too little (with relatively high $\frac{P_t(x)}{P_g(x)}$).
- Lastly, Theorem 3-(d) guarantees that the Nash equilibrium between the optimal generative function and the optimal discriminative function is reached if and only if $P_g = P_t$. At the Nash equilibrium, it holds $\kappa(f^*) = 0$, and consequently $\nabla_x f^*(x) = 0$ for all generated samples, which means the training will fully stop.

5. Max-Gradient Penalty and Augmented Lagrangian

Arjovsky et al. (2017) initially (as far as the authors know) introduced the requirement for Lipschitz regularization in GANs. After that, researchers (Kodali et al., 2017; Fedus et al., 2018; Miyato et al., 2018) empirically found that Lipschitz regularization is also useful when combined with other GANs objectives, e.g., the original GANs' objective.

Recently, such a phenomenon was also theoretically explained (Farnia and Tse, 2018; Zhou et al., 2019), i.e., combining Lipschitz regularization with common objectives of GANs yields variant distance metrics that are able to provide a continuous measure between the real and fake distributions, similar to the Wasserstein distance.

As it stands, Lipschitz regularization is a promising technique for improving the training of GANs with theoretical guarantee. However, accurate and effective implementation of Lipschitz regularization remains challenging.

5.1 Existing Lipschitz Regularization Implementations

A fair number of works are devoted to investigating the implementation of Lipschitz regularization. The initial attempt in Arjovsky et al. (2017) achieves the Lipschitz regularization via **Weight Clipping** (WC), i.e., restricting the maximum value of all weighting parameters in the neural network. However, it was later shown to lead to suboptimal solutions (Gulrajani et al., 2017; Petzka et al., 2018).

Accordingly, new methods named **Gradient Penalty** (GP) and **Lipschitz Penalty** (LP) were proposed. The two methods share the same spirit and achieve Lipschitz continuity via penalizing the sampled gradient norms towards a given target value. The target value is typically 1, however, not necessary (Karras et al., 2018; Adler and Lunz, 2018). They are based on the fact that the Lipschitz constant of a function is equivalent to its max gradient norm (Adler and Lunz, 2018), i.e., the maximum of the norm of its gradients.

Formally, the two methods employ the following regularization terms, respectively:

$$R_{gp} = \frac{\rho}{2} \cdot \mathbb{E}_{x\sim P_{\hat{x}}}[(\|\nabla_x f(x)\| - k_0)^\alpha], \tag{14}$$

$$R_{lp} = \frac{\rho}{2} \cdot \mathbb{E}_{x\sim P_{\hat{x}}}[(\max\{0, \|\nabla_x f(x)\| - k_0\})^\alpha], \tag{15}$$

where α is typically 2 and $P_{\hat{x}}$ denotes the sampling distribution determined by the sample strategy, which is typically random linear interpolation between the real and fake samples.

Petzka et al. (2018) argued that the gradient penalty is less reasonable, because k_0-Lipschitz does not necessarily imply that the gradient norm at every sample point is k_0. Hence, their proposed Lipschitz penalty was to only penalize gradients whose norm is larger than k_0.

Apart from those already mentioned, Miyato et al. (2018) provided a new direction for enforcing the Lipschitz continuity, named **Spectral Normalization** (SN) (Yoshida and Miyato, 2017), which is based on another fact that the Lipschitz constant of a linear function $h(x) = Wx$ is equivalent to the maximum singular value of the weight matrix W.

Given that the singular value of a weight matrix is (easily) obtained (e.g., by the power iteration), they proposed to divide the weight of each linear layer of a neural network by its

maximum singular value, i.e.,

$$\bar{W}_{SN} = W/\sigma(W), \tag{16}$$

where $\sigma(W)$ denotes the maximum singular value of W.

As a result of the spectral normalization, the Lipschitz constant of every linear layer is fixed as 1. Then, if the nonlinearity parts, e.g., activation functions, are also Lipschitz continuous, which is true for common choices like $ReLU$ and $tanh$, the resulting model will have an upper bound on the Lipschitz constant.

Remark 6 *It is worth noting that the spectral normalization results in a hard global restriction on the Lipschitz constant, while the gradient penalty and the Lipschitz penalty are soft partial regularizations, where the locality depends on the choice of $P_{\hat{x}}$.*

5.2 Analysis on Lipschitz Regularization Implementations

Before moving into the detailed discussion of these methods, we would like to provide several important notes in the first place, e.g., the sufficiency of partial Lipschitz continuity.

5.2.1 Partial Lipschitz Regularization is Sufficient

The most common choice of $P_{\hat{x}}$ in the gradient penalty and the Lipschitz penalty is the distribution formed by random linear interpolations between the real and fake samples. Currently, why such a choice is valid is still not clear, and people tend to believe that it is only a deleterious practical trick (Miyato et al., 2018).

Here, we provide a theoretical justification for such a choice. We will first demonstrate with the Wasserstein distance, and then arguably extend it to LGANs.

To get such a conclusion, we need to delve more deep into the KR duality Eq. (3) and our newly developed compact dual form Eq. (4). For KR duality, x and y are required to sample from the entire sample space, which is hence equivalent to Lipschitz regularization. However, with the compact dual form, we know that x and y are actually only necessarily required to sample from S_g and S_t, respectively.

It is worth noting that given the constraints in the compact dual form, any other constraints in the KR duality do not affect the final result of $W_1(P_g, P_t)$. And more importantly, any f^* in the compact dual form corresponds to (at least) one f^* in the KR duality with the value of f^* on S_g and S_t unchanged. Thus, any f^* in the compact dual form Eq. (4) also holds the following key property of the Wasserstein distance (Villani, 2008):

Theorem 7 *Let π^* be the optimal transport plan in the primal form of the Wasserstein distance Eq. (2) and f^* be the optimal discriminative function in the compact dual form Eq. (4). It holds that*

$$P_{(x,y)\sim\pi^*}[f^*(x) - f^*(y) = d(x, y)] = 1. \tag{17}$$

Note that the Proposition 1 is based on Eq. (17) and the 1-Lipschitz continuity of f^*. And we can further notice that f^* being partially Lipschitz continuous is sufficient for the proof.

Formally, we have:

Theorem 8 *Let $S_{\hat{x}} = \{\hat{x} = t \cdot x + (1-t) \cdot y \mid x \in S_g,\ y \in S_t,\ t \in [0,1]\}$ denote the support of the linear interpolations between P_g and P_t. Enforcing partial 1-Lipschitz over $S_{\hat{x}}$ is sufficient to maintain the property of Eq.* (5) *for the Wasserstein distance, i.e.,*

$$\mathbb{P}_{(x,y)\sim\pi^*}\left[\nabla_{x_t} f^*(x_t) = \frac{y-x}{\|y-x\|}\right] = 1. \tag{18}$$

It means the Wasserstein distance and its desired gradient property for GANs keep unchanged when the constraints outside the blending region $S_{\hat{x}}$ are dropped. That is, 1-Lipschitz in the blending region has covered all the necessities of the constraints.

From our analysis on how Lipschitz regularization works, we can tell these constraints outside $S_{\hat{x}}$ are also unnecessary for the forming of bounding relationships. Hence, we can similarly conclude that Lipschitz regularization over $S_{\hat{x}}$ is also sufficient for LGANs.

Note that Theorem 8 also indicates that for GANs, restricting the global Lipschitz constant, e.g., using the spectral normalization, might be unnecessarily too strong. Given that partial Lipschitzness is sufficient, henceforward, by Lipschitz constant or Lipschitz continuity, we by default mean that over the region $S_{\hat{x}}$.

5.2.2 Superfluous Constraints in Current Partial Lipschitz Implementations

We next show that, although imposing partial Lipschitz regularization is sufficient, the current implementations of partial Lipschitz regularization, i.e., the gradient penalty and the Lipschitz penalty, contain superfluous constraints and are hence biased.

The gradient penalty and the Lipschitz penalty impose the Lipschitz continuity via penalty method. Penalty method is a soft regularization where the constraint is usually slightly drifted. During the training, a penalty based method would provide a dynamic Lipschitz constant, which is usually much larger than the target Lipschitz constant, depending on the weight of the regularization.

Same as above, we demonstrate the analysis by the Wasserstein distance[4]. Let $W_1(P_g, P_t, k) = \sup_{\kappa(f)\le k} \mathbb{E}_{x\sim P_g}[f(x)] - \mathbb{E}_{x\sim P_t}[f(x)]$. It holds that $W_1(P_g, P_t, k) = k \cdot W_1(P_g, P_t)$. Because $W_1(P_g, P_t) = \sup_{\kappa(f)\le k} \mathbb{E}_{x\sim P_g}[f(x)/k] - \mathbb{E}_{x\sim P_t}[f(x)/k]$.

Assume we can directly optimize the Lipschitz constant k and consider the following objective:

$$J_{wgans\text{-}gp}(k) = -W_1(P_g, P_t, k) + \frac{\rho}{2} \cdot (k - k_0)^\alpha. \tag{19}$$

Note that in contrast with the *sup* in the dual form of Wasserstein distance, *we write all objectives in the forms of minimization problems.* Hence, it becomes $-W_1(P_g, P_t, k)$. Given that P_g and P_t is fixed, $W_1(P_g, P_t)$ is a constant and we denote it as c. Then,

$$J_{wgans\text{-}gp}(k) = -c \cdot k + \frac{\rho}{2} \cdot (k - k_0)^\alpha. \tag{20}$$

4. To extend the $J_{wgans\text{-}xx}$ parts of analysis to LGANs, we just need to write J_D as a function of k and replace $W_1(P_g, P_t, k)$ with $J_D(k)$. The closed-form solution is hard to attain, but because $J_D(k)$ is an increasing function of k with decreasing derivative, similar qualitative analysis is easily attainable.

When $\alpha > 1$, its *minimum* is reached when $k^* = (\frac{2c}{\rho\alpha})^{\frac{1}{\alpha-1}} + k_0$. Particularly, when $\alpha = 2$, $k^* = \frac{c}{\rho} + k_0$. Note that replacing $(k - k_0)^\alpha$ with $(\max\{0, k - k_0\})^\alpha$, i.e., analogizing switching from the gradient penalty to the Lipschitz penalty, will result in the same optimal k^*.

From the above, we can see that[5], when ρ is small or the distance between P_g and P_t is large (i.e., if c is large), the resulting Lipschitz constant can be much larger than k_0.

Under these circumstances, both the gradient penalty and the Lipschitz penalty introduce superfluous constraints. Saying $k_0 = 1$ and the current Lipschitz constant is 100, sampled points whose gradient is larger than 1 but smaller than 100, are penalized, unintentionally.

We will see in the experiments that these superfluous constraints alter the optimal discriminative function and damage the property of the gradient received by the generator.

Petzka et al. (2018) noted that Lipschitz penalty has a connection to regularized Wasserstein distance. However, regularized Wasserstein distance also alters the property of the optimal discriminative function and leads to blurry π^* (Seguy et al., 2018). That is, their results are not contradictory to our results.

5.3 The Proposed Lipschitz Regularization Implementations

Now we present our proposed methods towards more efficient (i.e., partial instead of global) and unbiased (without superfluous constraints) implementation of Lipschitz regularization.

5.3.1 MaxGP: Max Gradient Regularization with Penalty Method

Given that the partial Lipschitz continuity over the support of the linear interpolations between the real and fake distributions, i.e., $S_{\hat{x}}$, is sufficient for all desired properties in GANs, we would consider only restricting the Lipschitz constant in such a region.

Similar to gradient penalty, we can regularize the Lipschitz constant via penalty method. But, to avoid the superfluous constraints, we need to only penalize the max gradient norm in $S_{\hat{x}}$, which is equivalent to the Lipschitz constant in the region of $S_{\hat{x}}$.

The resulting regularization is as follows:

$$R_{maxgp} = \frac{\rho}{2}([\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\|] - k_0)^\alpha. \tag{21}$$

In analogy to Lipschitz penalty, we can also extend the penalty term with $\max\{0, \cdot\}$. However, when only regularizing the max gradient norm, it is less necessary. Because it will only take effect when the discriminator is underfitting.

Practically, we can directly use the max gradient norm estimated within a mini-batch, e.g., follow Gulrajani et al. (2017) and sample x as random linear interpolations of the real and fake samples in parallel mini-batches:

$$\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\| \approxeq \max_{x \in B_{mini}} \|\nabla_x f(x)\|. \tag{22}$$

5. From another perspective, as c goes to zero, i.e., as P_g converges to P_t, the optimal Lipschitz constant k^* decreases. And finally, when $P_g = P_t$, we have $c = 0$ and the optimal Lipschitz constant k^* equals k_0.

where $B_{mini} = \{t \cdot x + (1-t) \cdot y \mid x \in B_{fake}, y \in B_{real}, t \in [0,1]\}$ with B_{fake} and B_{real} indicating n randomly sampled fake samples and real samples, respectively.

We can also consider further introducing more accurate estimation of $\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\|$ to improve the stability and reduce the error introduced via batch sampling.

A practical and lightweight method for a more accurate estimation of $\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\|$ is to maintain a buffer B_{maxbuf} that stores these x that achieve the current historical top-k $\|\nabla_x f(x)\|$, which can be initialized with random samples. During training, use the samples buffered in B_{maxbuf} as part of (or as additional) the batch that estimates the current max gradient norm, and update B_{maxbuf} thereupon.

We have studied these two in experiments. According to our experiments, the historical buffer is usually unnecessary and directly using the max gradient norm in a mini-batch seems good enough, though we do not exclude the possible benefits of historical buffer or other more accurate estimations of max gradient norm or Lipschitz constant.

We conjecture that, when the training goes to the relatively later stage, the surface of f is relatively smooth and many locations hold the max gradient norm. Hence, typically, a mini-batch estimation could be accurate enough for successful training.

5.3.2 MaxAL: Max Gradient Regularization with Augmented Lagrangian

With the penalty method, the constraint usually cannot be strictly satisfied. The resulting Lipschitz constant, as discussed around Eq. (19), is usually drifted.

In the circumstances of GANs, strictly imposing a given Lipschitz constant might benefit the control of variables in contrast experiments, e.g., when comparing different networks and objectives. Also, if one would like to strictly evaluate the Wasserstein distance, a strict restriction of 1-Lipschitz would be favorable. Otherwise, it needs to estimate the Lipschitz constant and divide the loss by the estimated Lipschitz constant.

In the situation where people would like the constraint to be strictly imposed, the augmented Lagrangian is a classic alternative to the penalty method for strict constraint satisfaction. It extends the penalty method by introducing an extra Lagrange multiplier term. The regularization terms derived from the augmented Lagrangian can be written as follows:

$$R_{maxal} = \lambda_{al} \cdot (\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\| - k_0) + \frac{\rho}{2} \cdot (\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\| - k_0)^{\alpha}, \tag{23}$$

where λ_{al} is the Lagrange multiplier and $\alpha > 1$.

Given that the augmented Lagrangian is a simple extension to the penalty method and there exists potential benefits, we investigate its theoretical properties and practical performance in imposing Lipschitz regularization.

5.3.3 First Order Optimality Analysis: Part-I, MaxAL Properties

Some interesting properties of the augmented Lagrangian can be easily derived with its first order optimality analysis. For clarity, we denote $\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\|$ as k. Still, we

demonstrate with the Wasserstein distance for simplicity. Let the overall objective be:

$$J_{wgans\text{-}al}(k) = -W_1(P_g, P_t, k) + \lambda_{al} \cdot (k - k_0) + \frac{\rho}{2} \cdot (k - k_0)^{\alpha}. \tag{24}$$

Similar as previous, because $W_1(P_g, P_t, k) = k \cdot W_1(P_g, P_t)$ and $W_1(P_g, P_t)$ is a constant, we denote $W_1(P_g, P_t)$ as c and then $W_1(P_g, P_t, k)$ equals $c \cdot k$.

Then, what is optimizing is

$$J_{wgans\text{-}al}(k) = -c \cdot k + \lambda_{al} \cdot (k - k_0) + \frac{\rho}{2} \cdot (k - k_0)^{\alpha}. \tag{25}$$

Then, the first order optimality conditions can be written down as follows:

$$\begin{aligned} &\nabla_{\lambda_{al}} J_{wgans\text{-}al} = k - k_0 = 0, \\ &\nabla_k J_{wgans\text{-}al} = -c + \lambda_{al} + \frac{\rho}{2} \cdot \alpha \cdot (k - k_0)^{\alpha - 1} = 0. \end{aligned} \tag{26}$$

Thereby, when the augmented Lagrangian converges, $k = k_0$ and $\lambda_{al} = W_1(P_g, P_t)$.

5.3.4 First Order Optimality Analysis: Part-II, How to Optimize MaxAL

We now explain the suggested optimization schema for MaxAL. To move on, we need to introduce the Lagrange multiplier method and its first order optimality analysis.

The Lagrange multiplier method is also a classical method for constrained optimization. It introduces a Lagrange multiplier term into the original optimization problem, i.e.,

$$R_{maxlm} = \lambda_{lm} \cdot (\max_{x \sim P_{\hat{x}}} \|\nabla_x f(x)\| - k_0), \tag{27}$$

where λ_{lm} is the Lagrange multiplier.

The so-called augmented Lagrangian can also be viewed as an extension of the Lagrange multiplier method, where the penalty term is regarded as the augmentation.

Consider the optimization problem of

$$J_{wgans\text{-}lm}(k) = -W_1(P_g, P_t, k) + \lambda_{lm} \cdot (k - k_0). \tag{28}$$

The first order optimality condition of the Lagrange multiplier method can be written as:

$$\begin{aligned} &\nabla_{\lambda_{lm}} J_{wgans\text{-}lm} = k - k_0 = 0, \\ &\nabla_k J_{wgans\text{-}lm} = -c + \lambda_{lm} = 0. \end{aligned} \tag{29}$$

When the Lagrange multiplier method converges, it also holds $k = k_0$ and $\lambda_{lm} = W_1(P_g, P_t)$.

The superiority of the augmented Lagrangian over the Lagrange multiplier method can be understood as: with the driven force of the penalty term, it is much easier for the augmented Lagrangian to reach the first order optimality.

From the first order optimality conditions, we can see that c, which is fixed and being the real target, when converged, is equal to $\lambda_{al} + \frac{\rho}{2} \cdot \alpha \cdot (k - k_0)^{\alpha - 1}$ in the augmented Lagrangian,

meanwhile, is equal to λ_{lm} in the Lagrange multiplier method, which means the following should hold:

$$\lambda_{lm} = \lambda_{al} + \frac{\rho}{2} \cdot \alpha \cdot (k - k_0)^{\alpha - 1}. \tag{30}$$

Recall that the augmented Lagrangian can be understood as the Lagrange multiplier method with extra penalty term, which means λ_{al} shall play the role of λ_{lm}. Hence, the commonly suggested update rule for λ_{al} in the augmented Lagrangian is the following:

$$\lambda_{al}^{t+1} = \lambda_{al}^{t} + \frac{\rho}{2} \cdot \alpha \cdot (k - k_0)^{\alpha - 1}, \tag{31}$$

where t indicates the iteration or timestamp.

Thus, to upgrade from the penalty method to the augmented Lagrangian, one only needs to introduce the Lagrange multiplier term and add an extra update step for λ_{al} according to Eq. (31) after each iteration of the discriminator.

5.4 Empirical Analysis of Lipschitz Regularization Implementations

In this section, we empirically study the proposed Lipschitz regularization implementations, showing its superiority over existing methods.

Specifically, we study and compare the practical behaviors of spectral normalization (SN), gradient penalty (GP), max gradient regularization with penalty method (MaxGP, namely max gradient penalty), and max gradient regularization with augmented Lagrangian (MaxAL). In our experiments, we find the Lipschitz penalty shares a very similar performance with the gradient penalty, so we omit the Lipschitz penalty for clarity.

We use multilayer perceptrons for all toy experiments and use a Resnet architecture (He et al., 2016) that is similar to the one used in Gulrajani et al. (2017) for all other real data experiments. We use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0$ and $\beta_2 = 0.9$.

Frechet Inception Distance (FID) (Heusel et al., 2017) is used to quantitatively evaluate the resulting models. Another well-known metric is Inception Score (Salimans et al., 2016). However, it is not well explained, and the resulting score is highly unstable (Zhou et al., 2018; Borji, 2019). Hence, such results are not included here.

For experiments in this section, we by default use WGANs' objective because its sample gradients, at optimum, correspond to the optimal transport plan, which help us check the optimality of these results. Following the common choice, we set $k_0 = 1$ and $\alpha = 2$.

The code for reproducing these results is provided at `https://github.com/ZhimingZhou/MaxGP-MaxAL-for-reproduce`.

5.4.1 Two Dimensional Toy Data

To intuitively study the property of different methods, we first test their performances with simple two-dimensional data. In this experiment, we randomly sample two data points in two-dimensional space as P_g and another two points as P_t. We fix these two distributions and train a discriminator with different implementations of Lipschitz regularization.

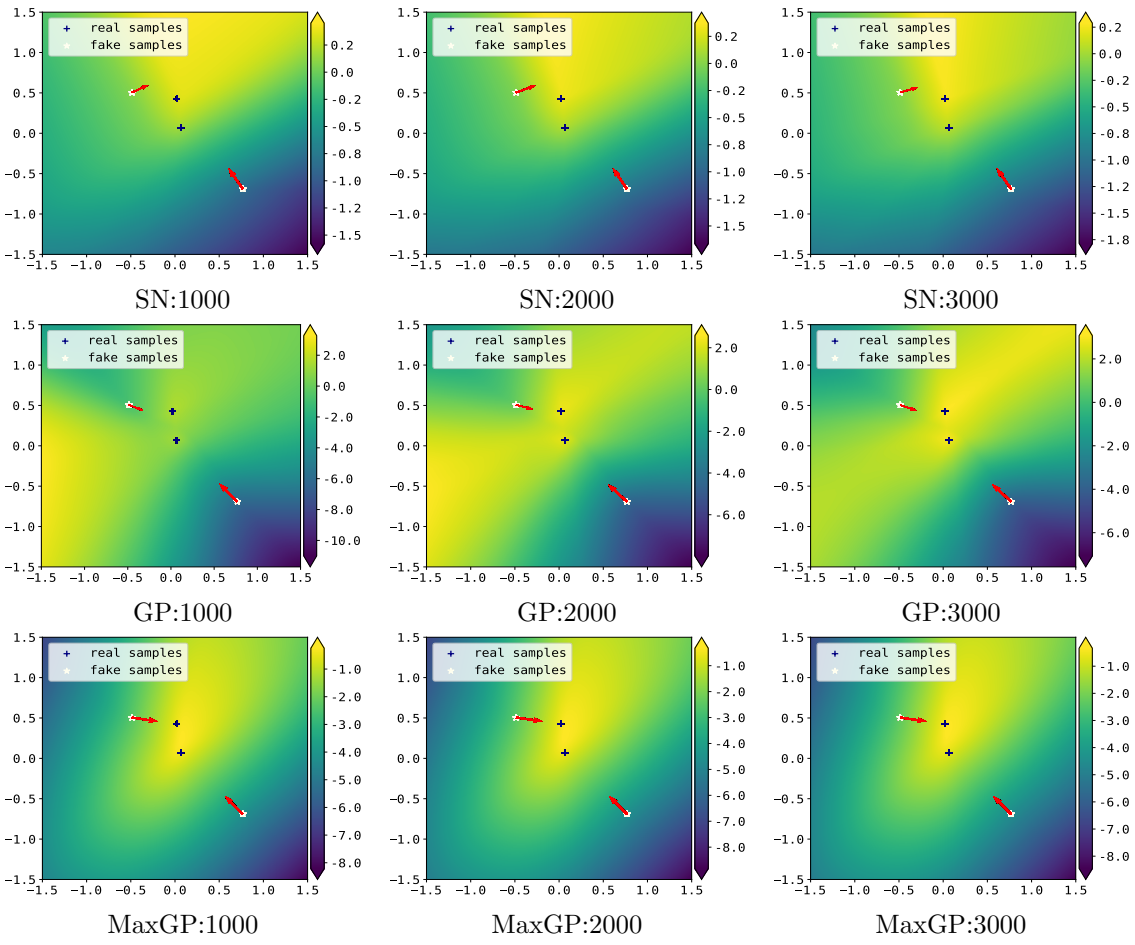


Figure 2: With P_g and P_t both being two random sampled points in two-dimensional space, we train the discriminator of WGANs using SN, GP and MaxGP, respectively. The numbers after the name of the methods are the corresponding iteration numbers. The arrows in the figures indicate the gradient scales and directions. From the results, we notice that: (i) SN in this case fails to achieve the optimal discriminative function; (ii) the discriminator trained with GP is oscillatory; (iii) MaxGP stably converges to the optimal. Note that MaxGP, as well as SN, converged with 1000 iterations.

We check whether these methods are able to achieve the optimal discriminative function, by verifying their gradients with respect to the generated samples, which should follow the Proposition 1 and point towards their target real samples that minimize the transport cost.

Our first interesting observation is that SN in some cases fails to achieve the optimal discriminative function. As shown in Figure 2, SN quickly converges to a suboptimal solution and sticks there. We have tried to avoid all possible external causes that we can think of to this problem[6], but this issue keeps existing. It suggests the existence of internal deficiency of the current SN-based Lipschitz regularization system. Hence, we believe the current SN-based Lipschitz regularization implementation, though computationally cheaper, is less reliable. We leave further investigation of this issue as future work.

In Figure 2, we can also notice that GP leads to oscillation in the value surface of the discriminative function. It seems there is no stable optimum for the discriminative function under GP, evidencing that the superfluous constraints do affect the optimal discriminative function. By contrast, we see that MaxGP quickly converges to the optimal discriminative function (according to the Figure 2, the convergence is achieved within 1000 iterations) and stably holds at the optimal state (keep almost unchanged). Note that the gradients of the generated samples are pointing towards the real samples in an optimal transport manner.

5.4.2 Toy Real World Data

We further compare these Lipschitz implementations with real world data. We still want to check whether these methods converge to the optimal discriminative function.

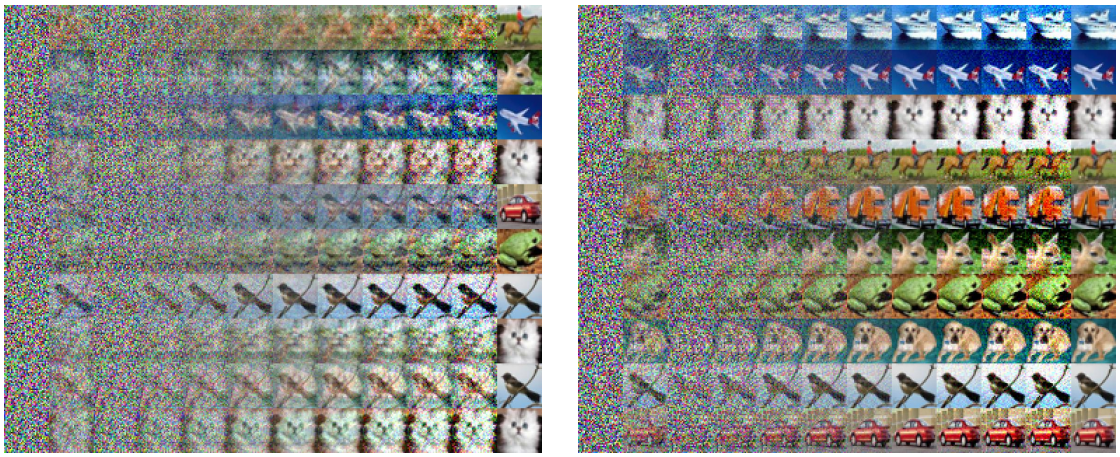
However, we found practically the optimal discriminative function is almost unachievable when trained with an entire real world dataset, even it is a small dataset like CIFAR-10 or MNIST. Hence, in this experiment, we use a small subset of the real world dataset instead. Specifically, we select ten representative CIFAR-10 images as P_t and use ten random noise images as P_g. Same as above, we fix the two distributions, train the discriminator till optimal, and check the gradients of the resulting discriminative functions.

For the high dimensional case, visualizing the gradient direction is nontrivial. Hence, we plot the gradients and corresponding increments in Figure 3, where the leftmost in each row is a sample x in S_g and the second is its gradient $\nabla_x f(x)$. The interiors are $x + \epsilon \cdot \nabla_x f(x)$ with increasing ϵ, and the rightmost in each row is the real sample y in S_t that is closest to any point of that half-line.

From the results, MaxGP is also able to achieve the optimal discriminative function in the high dimensional case. We see that the gradients of the ten noise images in P_g are pointing towards the ten real images in P_t, respectively.

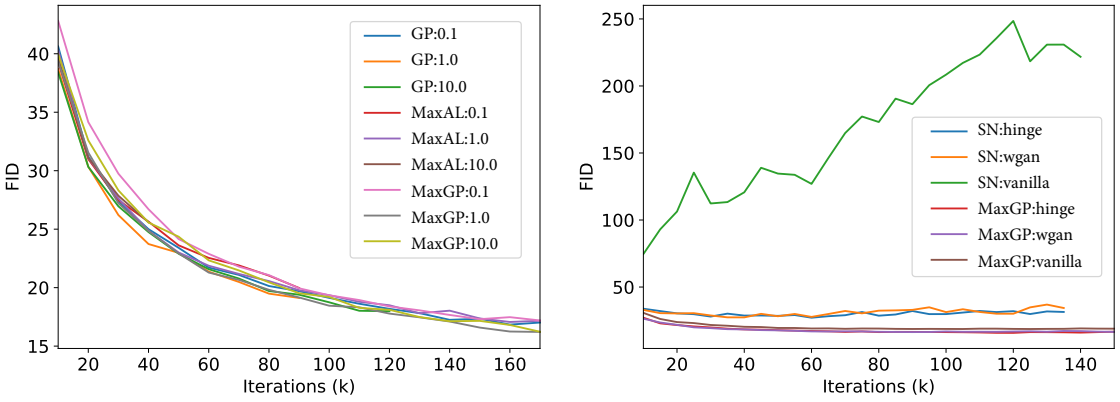
However, the resulting gradients of GP do not clearly point towards real samples. The gradient tends to be a blending of several images in the target domain, and it also appears a sort of mode collapse (multiple cats and birds). This experiment once again verifies that these superfluous constraints unintentionally introduced by GP are harmful.

6. For example, we have tried increasing the network capacity, increasing the number of the power iteration, and training the discriminator for a very long time with a decreasing learning rate.



(a) GP (b) MaxGP

Figure 3: With P_g and P_t being ten fixed noise and real images, respectively, we train the discriminator of WGANs using GP and MaxGP towards optimum. The leftmost in each row is a sample x in S_g and the second is the gradient $\nabla_x f(x)$. The interiors are $x + \epsilon \cdot \nabla_x f(x)$ with increasing ϵ. The rightmost is the nearest real sample y in S_t. GP fails to achieve the optimal discriminative function, and the samples tend to collapse to a subset. By contrast, with MaxGP, the gradients of P_g samples perfectly follow the optimal transport.



(a) (b)

Figure 4: The quantitative comparison of different implementations of Lipschitz regularization. Unsupervised CIFAR-10 generation with WGANs' loss metric in terms of FID training curve. The number after the name of the method is the regularization weight ρ and the string after the method name indicates the loss metric it used. GP, MaxGP and MaxAL achieve very similar results, and they are not very sensitive to the regularization weight ρ. The training of SN, when using WGANs' loss metric, diverges. When switched to the hinge loss or original GANs' loss metric, the final results of SN are still clearly worse than MaxGP.

5.4.3 Sample Quality on CIFAR-10

We now test the practical difference when training a complete GANs model, using these methods to impose Lipschitz regularization.

In this experiment, we not only train the model with WGANs' loss metric, but also with the hinge loss (Miyato et al., 2018) and original GANs' loss metric (Goodfellow et al., 2014), which have been found also work well under Lipschitz regularization (Fedus et al., 2018).

The results in terms of the training curve of FID are plotted in Figure 4. In Figure 4a, we compare GP, MaxGP and MaxAL with different regularization weights under WGANs' loss metric. The visual results are also provided in Figure 6. We see that the training progresses, final FIDs, and visual results are quite similar to each other.

As we found in the previous experiment, the optimal discriminative function is already very hard to achieve even if P_g and P_t both only consist of ten images. In fact, when trained with an entire dataset, the gradients $\nabla_x f(x)$ of the generated samples are highly blurry.

Hence, we believe that the reason why these methods do not show obvious differences in these real world applications lies in the optimization level. That is, the attainable/attained discriminative function could be too far from the optimal, so no matter whether it is biased or not, the final result appears similar.

We tend to believe the issue lies in the discriminative function space. That is, the current convolutional neural networks, though being powerful/suitable for other tasks like classification, are not expressive enough or not favorable for expressing the Lipschitz regularized optimal discriminative function.

In this experiment, we initially use WGANs' loss metric for all methods. However, we found that with the Resnet architecture (Gulrajani et al., 2017), SN constantly fails to converge. We note that in Miyato et al. (2018), when using Resnet architecture, the model with SN is trained using a hinge loss.

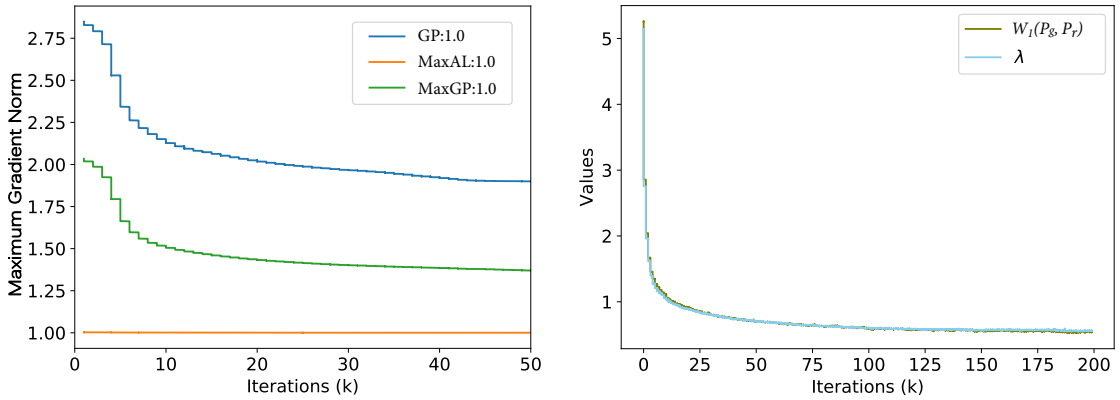
We therefore also test SN with the hinge loss, and in addition, the original GAN's loss metric. The results are plotted in Figure 4b. We also include the results of MaxGP with these loss metrics for comparison. As we can see, the results of MaxGP are generally better than SN.

Lastly, we inspect the properties of MaxAL. As shown in Figure 5a, MaxAL is able to quickly drive the Lipschitz constant to the given target and keep the Lipschitz constant fairly stable during the training. By contrast, the Lipschitz constants under GP and MaxGP keep changing, decreasing as P_g getting closer to P_t (echoing our analysis around Eq. (20)).

Another interesting fact about MaxAL is that, when trained with WGANs' loss metric, the optimal λ is equivalent to $W_1(P_g, P_t)$. We verify this fact by plotting these two terms during training together. As shown in Figure 5b, the two lines are basically overlapped.

5.5 Summary on Lipschitz Regularization Implementations

We have demonstrated that regularizing the Lipschitz constant over the support of the interpolations of real and fake samples is sufficient to gain the desired gradient properties



(a) (b)

Figure 5: The favorable properties of MaxAL. The number after the name of the method is the regularization weight ρ. With MaxAL, the Lipschitz constant quickly converges to the given target 1. By contrast, the Lipschitz constants achieved by GP and MaxGP are dynamic, which decreases as P_g converges to P_t. In addition, when trained with WGANs' loss metric, the value of the λ is equivalent to the Wasserstein distance.

induced by Lipschitz continuity. It provides theoretical guarantee on the validity of the sampling strategy of gradient-penalty based methods.

The analysis also suggests that global restriction on the Lipschitz constant is unnecessary. Combined with the fact that we found the current method for global Lipschitz regularization, i.e., the spectral normalization, somehow fails or shows significantly inferior performances, we presently tend to champion these gradient-penalty based partial Lipschitz regularization.

On the other hand, we have observed that the current implementations of partial Lipschitz regularization, i.e., the gradient penalty and the Lipschitz penalty, introduce superfluous constraints to the optimization problem, which evidently alter the optimal discriminative function and impair the favorable gradient properties, leading to instability during training.

We have accordingly proposed a revision to the gradient penalty and demonstrated that the proposed method, MaxGP, is able to achieve the optimal discriminative function in an unbiased manner. In addition, we have suggested augmented Lagrangian as a simple yet good alternative to the penalty method, resulting in the proposed MaxAL, which is able to strictly impose a given Lipschitz constant.

(a) GP (b) SN

(c) MaxGP (d) MaxAL

Figure 6: The visual comparison of different implementations of Lipschitz regularization. Unsupervised CIFAR-10 generation with WGANs' loss metric. Since the training with SN diverges when using WGANs' loss metric, here we plot its result with hinge loss, instead.

6. Empirical Analysis of Lipschitz Regularized GANs

Prescribed the theoretically sound and practically well-behaving implementations of Lipschitz regularization, we now verify the theoretical properties of LGANs and benchmark various instances of LGANs. We will show LGANs' consistently superior performances over WGANs.

In the following experiments, we use MaxGP for Lipschitz regularization, and whenever necessary we search the best regularization weight $\rho/2$ in $[0.01, 0.1, 1.0, 10.0]$. To adopt MaxGP for LGANs, we just need to set $k_0 = 0$. We follow the common choice and set $\alpha = 2$.

In fact, MaxAL can also be used, if a strict target Lipschitz constant is somehow preferred. Note that if the target Lipschitz constant k_0 is small enough and P_g and P_t are not close enough, all required bounding relationships can be established. One can further consider decreasing k_0 in MaxAL during training as needed. Penalizing the Lipschitz constant is to ensure the establishment of effective bounding relationships when $P_g \neq P_t$. Nevertheless, the following experiments are based on MaxGP, which is a more natural choice for LGANs.

The code for reproducing these results is provided at `https://github.com/ZhimingZhou/LGANs-for-reproduce`.

6.1 Verifying $\nabla_x f^*(x)$ in LGANs Point Towards Real Samples

One important theoretical benefit of LGANs is that $\nabla_x f^*(x)$ for each generated sample is guaranteed to point towards a certain real sample. We here verify the direction of $\nabla_x f^*(x)$ with a set of ϕ and φ that satisfy Eq. (12).

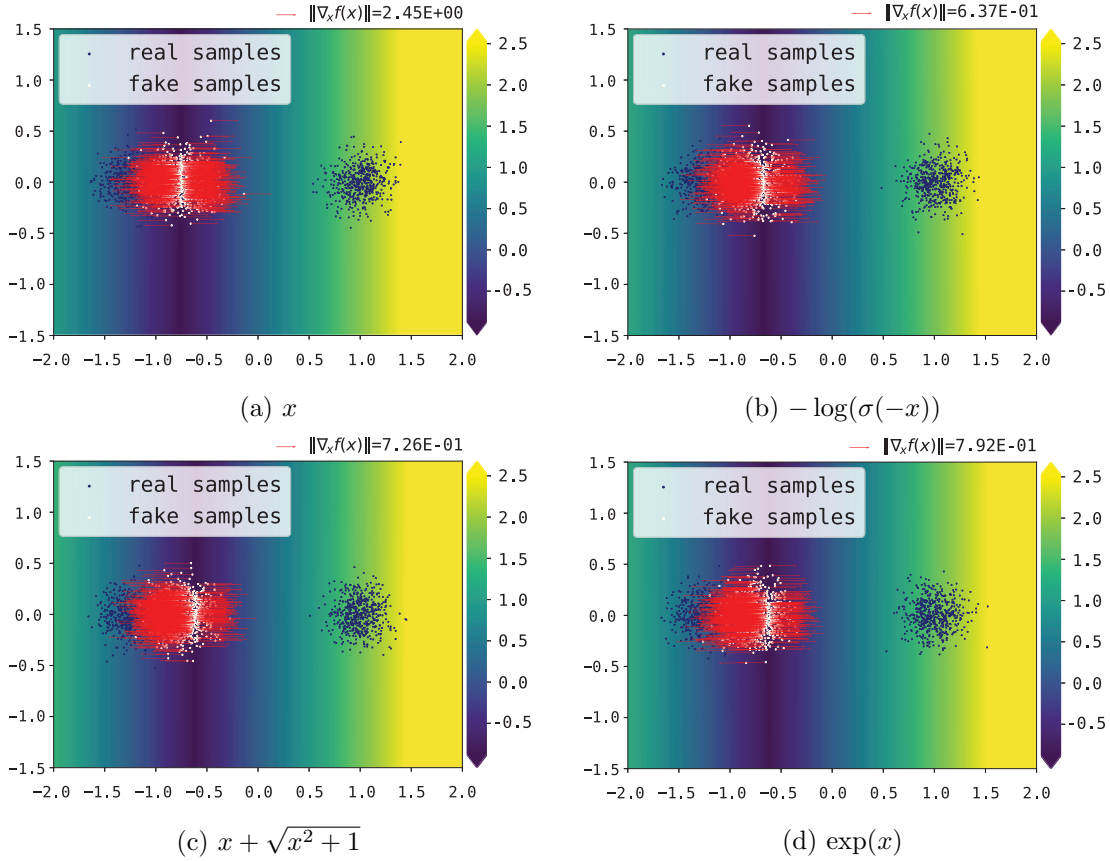
The tested loss metrics include: (a) $\phi(x) = \varphi(-x) = x$; (b) $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$; (c) $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + 1}$; (d) $\phi(x) = \varphi(-x) = \exp(x)$. They are tested in two scenarios: two-dimensional toy data and real world high dimensional data.

In the two-dimensional case, P_t consists of two distant Gaussians and P_g is fixed as one Gaussian which is close to one of the two real Gaussians, as illustrated in Figure 7. For the latter case, we use the CIFAR-10 training set. To make the solving of f^* feasible, we use ten CIFAR-10 images as P_t and ten fixed noise images as P_g. Note that we fix P_g on purpose because, to verify the direction of $\nabla_x f^*(x)$, learning P_g is not necessary.

The results are shown in Figures 7 and 8, respectively. In Figure 7, we can easily see that $\nabla_x f^*(x)$ of each generated sample is pointing towards a certain real sample.

Note that the gradients in LGANs do not always follow the optimal transport, which though is intuitively good, turns out to not necessarily imply a better performance. According to our experiments, LGANs consistently outperform WGANs, and the strict convexity of LGANs seems to be more important. It might be related to the fact that these strictly convex loss metrics, with annealing gradient scales at infinity, weaken the benefit of further discriminating these well-identified and enable the discriminator to pay more attention to these ill-identified.

Not following the optimal transport, LGANs may temporarily tend to gather samples. However, this is not really a problem as it seems to be, because LGANs will later redistribute



(a) x (b) $-\log(\sigma(-x))$

(c) $x + \sqrt{x^2+1}$ (d) $\exp(x)$

Figure 7: Verifying $\nabla_x f^*(x)$ in LGANs point towards real samples. P_t consists of two distant Gaussians and P_g is fixed as one Gaussian that is close to one of the two real Gaussians. From these results, we can see that $\nabla_x f^*(x)$ of each generated sample is pointing towards a certain real sample. LGANs with different loss metrics induce different gradients. We can see that the gradients in LGANs do not always follow the optimal transport, which though is intuitively good, turns out to not necessarily imply a better performance. According to our experiments, LGANs consistently outperform WGANs, and the strict convexity of LGANs seems to be more important. LGANs may temporarily tend to gather samples. However, this is not really a problem as it seems to be, because LGANs will later redistribute these exceedingly gathered samples, as part of continually moving samples from where there is relatively too much to where there is relatively too little until final convergence.

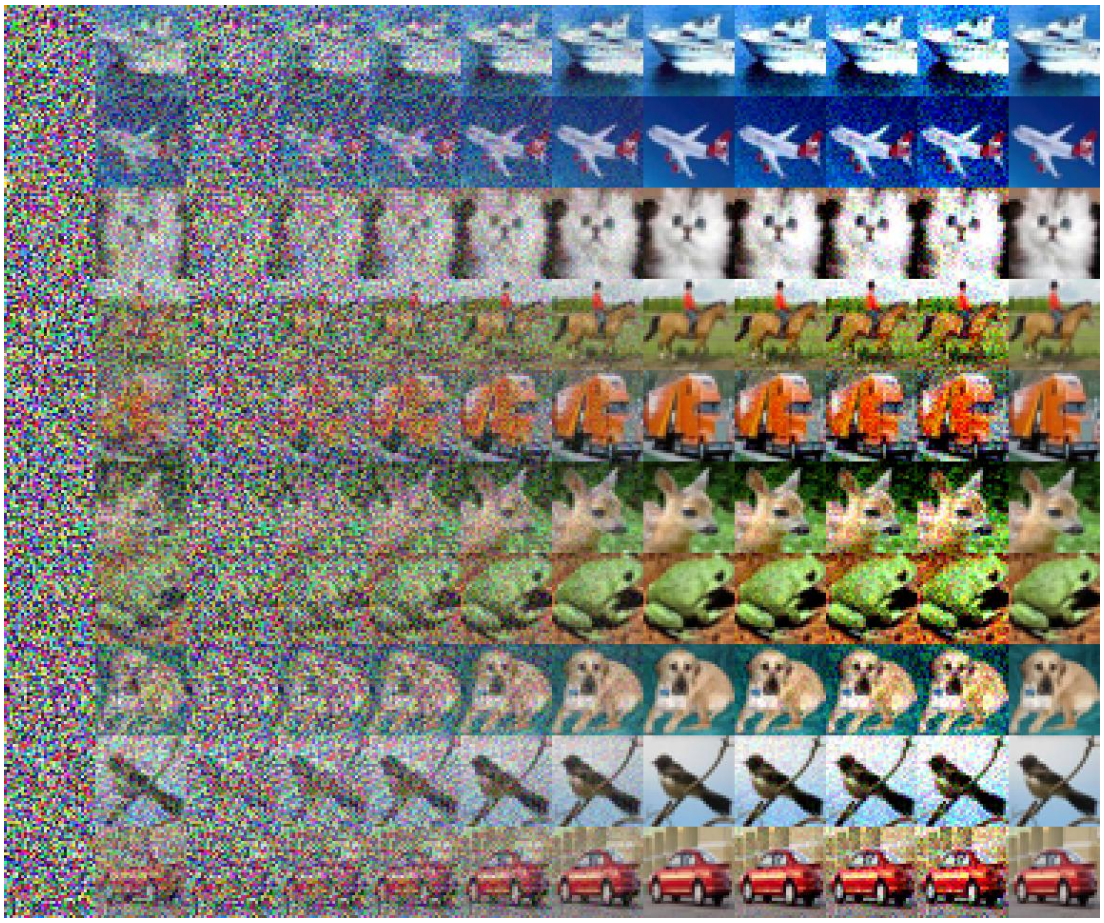


Figure 8: Verifying $\nabla_x f^*(x)$ in LGANs point towards real samples. With P_g and P_t being ten fixed noise and real images, respectively, we train the discriminator of LGANs towards optimum. The leftmost in each row is a sample x in S_g and the second is the gradient $\nabla_x f(x)$. The interiors are $x + \epsilon \cdot \nabla_x f(x)$ with increasing ϵ. The rightmost is the real sample y in S_t that is nearest to this half-line. This result visually demonstrates that $\nabla_x f^*(x)$ of generated samples are pointing towards real samples. Note that according to our experiments, the final results of this experiment keep almost identical when varying the loss metrics ϕ and φ in the family of LGANs.

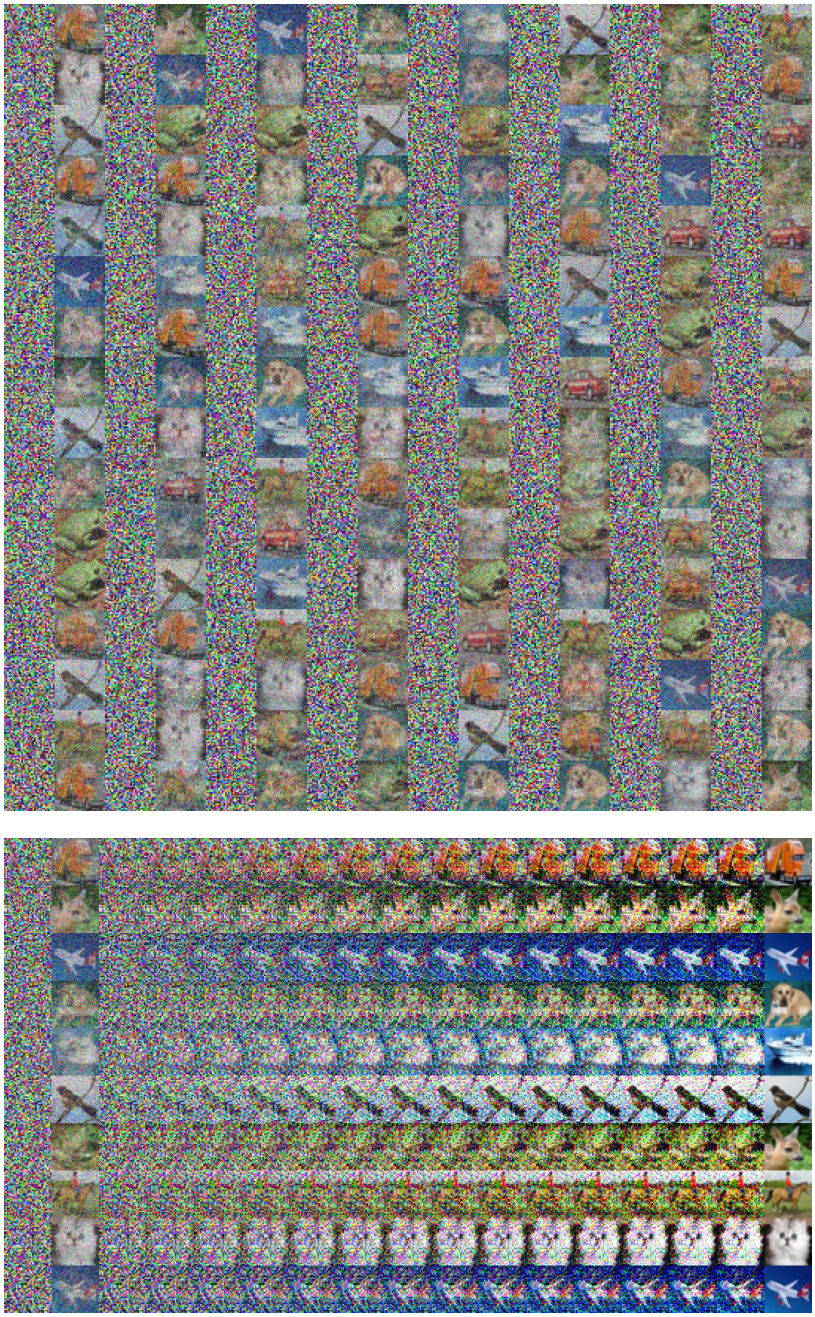


Figure 9: Verifying $\nabla_x f^*(x)$ in LGANs points towards real samples. P_t consists of ten images and P_g is a fixed Gaussian noise distribution. Up: Each odd column are $x \in S_g$ and the nearby column are their gradients $\nabla_x f^*(x)$. Down: the leftmost in each row is $x \in S_g$, the second are their gradients $\nabla_x f^*(x)$, the interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing ϵ, and the rightmost is the nearest $y \in S_t$.

these exceedingly gathered samples, as part of continually moving samples from where there is relatively too much to where there is relatively too little until final convergence.

For the high dimensional case, visualizing the gradient direction is nontrivial. Hence, we plot the gradient and corresponding increments. In Figure 8, the leftmost in each row is a sample x in S_g and the second is its gradient $\nabla_x f(x)$. The interiors are $x + \epsilon \cdot \nabla_x f(x)$ with increasing ϵ and the rightmost is the real sample y in S_t that is closest to any point of this half-line. This result visually demonstrates that in LGANs, $\nabla_x f^*(x)$ of generated samples are pointing towards real samples.

Note that the final results of Figure 8 keep almost identical when varying the loss metrics ϕ and φ in the family of LGANs, which we believe is reasonable. According to our analysis, when S_g and S_t are disjoint, LGANs behave just like WGANs in the sense that every sample in S_g must get bounded by a real sample. It seems unlikely, at least in our case, that multiple samples in S_g are bounded by the same sample in S_t.

We provide an extra experiment in Figure 9 for verifying $\nabla_x f^*(x)$ in LGANs under a more complex setting, where P_g is a fixed Gaussian distribution. We can see that even in this case, the gradient of each generated sample is basically pointing towards one of the real samples, though not as clear as in Figure 8.

6.2 The Benefit of Uniqueness of f^* in LGANs: Stabilized f.

The loss metric that corresponds to the Wasserstein distance is a very special case that satisfies Eq. (12). It is the only case where both ϕ and φ have constant derivatives, i.e., both are not strictly convex (otherwise, the uniqueness holds).

As a result, f^* under the Wasserstein distance loss metric has a free offset, i.e., given any optimal discriminative function f^*, $f^* + d$ with any $d \in \mathbb{R}$ is also optimal. In practice, it behaves as oscillations of f during the training.

The oscillations seem to harm the practical performance of WGANs. Karras et al. (2018) and Adler and Lunz (2018) introduced regularizations to the discriminative function to prevent f from drifting during the training. By contrast, any other instance of LGANs does not have this issue. We illustrate the practical difference in Figure 10.

Note that upon this oscillation effect, WGANs and LGANs with Wasserstein distance loss metric are essentially the same. The difference lies in the resulting value of Lipschitz constant: WGANs forces it being or towards one, while LGANs with Wasserstein distance loss metric penalize it to make it as small as possible.

Nonetheless, the substantial change happens when P_g converges to P_t. At that time, the training of LGANs will fundamentally stop with wholly zero gradients as $\kappa(f) = 0$. But WGANs, with $\kappa(f)$ requested to be one, will keep fluctuating (Mescheder et al., 2018).

6.3 Benchmark with Unsupervised Image Generation

To quantitatively compare the performance of different loss metrics under Lipschitz regularization, we test them with unsupervised image generation tasks.

In this part of experiments, we also include the hinge loss $\phi(x) = \varphi(-x) = \max(0, x+a)$ and quadratic loss (Mao et al., 2017), which do not fit the assumption of strict monotonicity. For the quadratic loss, we set $\phi(x) = \varphi(-x) = (x+a)^2$. We set $a = 1.0$ in all the experiments.

The strict monotonicity assumption of ϕ and φ is critical in Theorem 3 to theoretically guarantee the existence of bounding relationships for *arbitrary data*. But if we further assume that the supports of S_g and S_r are limited, it is possible that there exists a suitable ρ, which results in a proper scale of $\kappa(f)$ (see the arguments around Eq. (19)), such that all the real and fake samples lie in a strict monotone region of φ and ϕ. Then, the hinge loss, and even the quadratic loss, may also work well. For the hinge loss, it would imply $\kappa(f) \cdot \|y - x\| < 2a$ for all $x \in S_g$ and for all $y \in S_t$.

Our tentative experiments show that the choice of ψ in J_G does not play a central role. Even so, in this experiment, we choose to fix the loss metric as $\psi(x) = -x$. The thought behind our current choice is that: if we choose to use the minimax formulation $\psi(x) = -\phi(x)$, though we can get the minimax explanation of what the generator is minimizing, it will have some intuitively strange property: when ϕ is strictly convex, these well-identified samples (with low $f(x)$ values, which somehow indicate larger distances to the target distribution) will get small gradient weights $\nabla_{f(x)}\psi(f(x))$ (likely, the scale of $\nabla_x f(x)$ is around $\kappa(f)$). On the other hand, setting $\psi(x) = -x$ would update samples with evenly distributed weights. Note that $\psi(x) = \phi(-x)$ can be another choice. We believe the choice of ψ is an interesting research topic. However, stagnating at the intuition level, we leave it to future work.

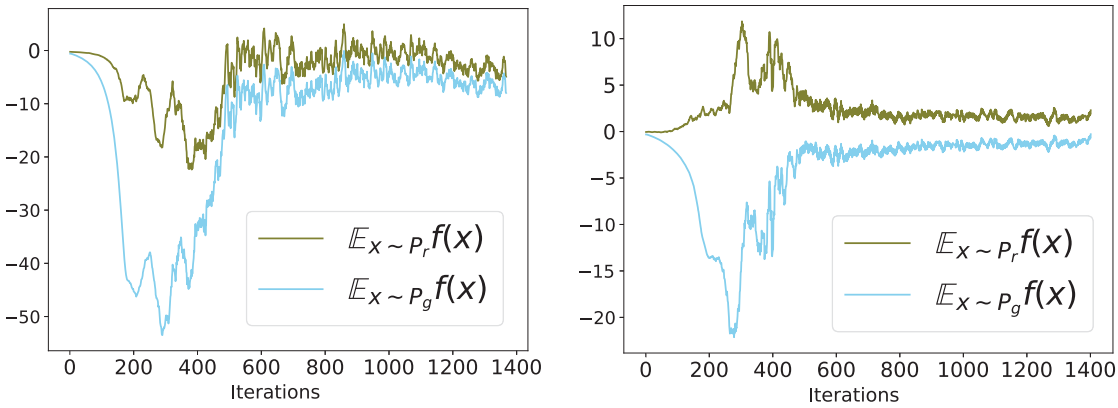
The results in terms of Inception Score (IS, Salimans et al., 2016) and Frechet Inception Distance (FID, Heusel et al., 2017) are presented in Table 2. For these experiments, we adopt the network structures and hyperparameter setting from Gulrajani et al. (2017). We use $200,000$ iterations for better convergence and use $500k$ samples to evaluate IS and FID for preferable stability. We note that IS, though popular, is not well explained (Zhou et al., 2018; Borji, 2019). And it is highly unstable during the training. By contrast, FID is fairly stable. Still, we include IS for better reference. We plot the training curves in terms of FID in Figure 11. The training curves in terms of IS are deferred to Figure 15 in Appendix.

From Figure 11 and Table 2, we can clearly tell that LGANs instances with strictly convex loss metrics consistently and clearly outperform WGANs, while LGANs with WGANs' loss metric shares a similar performance to WGANs. Different instances of LGANs have relatively similar final results, while the loss metrics $\phi(x) = \varphi(-x) = x + \sqrt{x^2+1}$ and $\phi(x) = \varphi(-x) = \exp(x)$ that have relatively stronger convexity perform relatively better.

This is probably because LGANs with strictly convex loss metric has annealing gradient scales at infinity, which weakens the benefit of further discriminating these well-identified and enables the discriminator to pay more attention to these ill-identified.

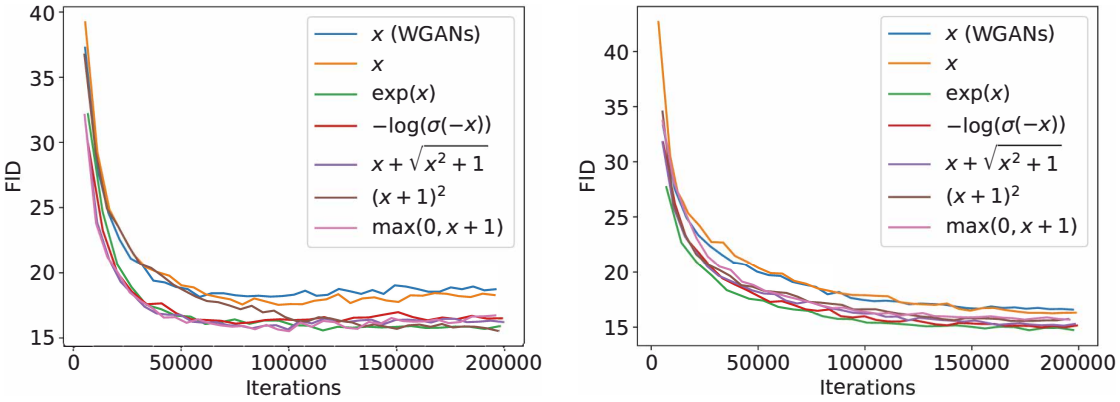
The hinge loss and quadratic loss turn out to also work pretty good.

We provide the visual results of LGANs with different loss metrics in Figures 12 and 13.



(a) WGANs. (b) LGANs.

Figure 10: The benefit of the uniqueness of f^* in LGANs. f is more stable during training.

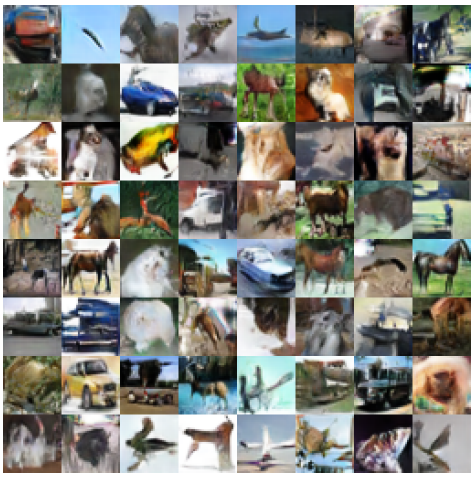


(a) CIFAR-10. (b) Tiny-ImageNet.

Figure 11: Training curves in terms of FID. WGANs and a set of instances of LGANs.

Loss Metrics	CIFAR-10		Tiny-ImageNet	
	IS	FID	IS	FID
x	7.68 ± 0.03	18.35 ± 0.12	8.66 ± 0.04	16.47 ± 0.04
$-\log(\sigma(-x))$	7.95 ± 0.04	16.47 ± 0.11	8.70 ± 0.04	15.05 ± 0.07
$x+\sqrt{x^2+1}$	7.97 ± 0.03	16.03 ± 0.09	$\mathbf{8.82 \pm 0.03}$	15.11 ± 0.06
$\exp(x)$	$\mathbf{8.03 \pm 0.03}$	$\mathbf{15.64 \pm 0.07}$	8.67 ± 0.04	$\mathbf{14.90 \pm 0.07}$
$(x+1)^2$	7.97 ± 0.04	15.90 ± 0.09	8.53 ± 0.04	15.72 ± 0.11
$\max(0, x+1)$	7.91 ± 0.04	16.52 ± 0.12	8.63 ± 0.04	15.75 ± 0.06

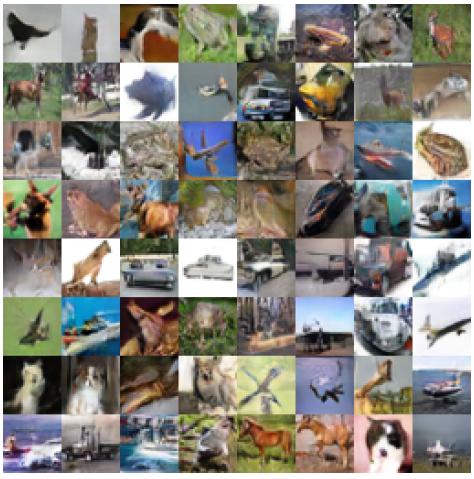
Table 2: The quantitative comparisons. WGANs' loss metric and other instances of LGANs.



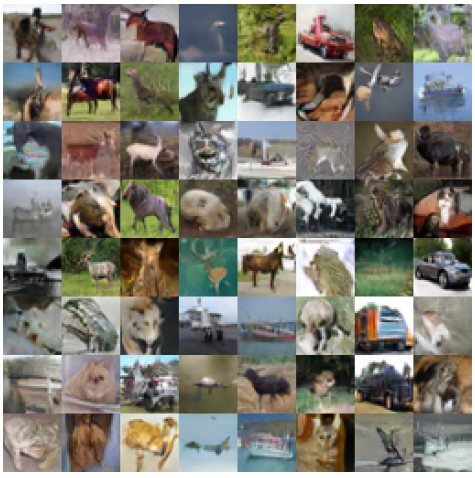
(a) x

(b) $\exp(x)$

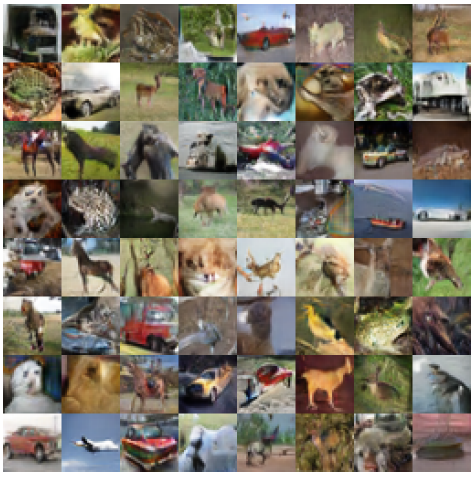
(c) $-\log(\sigma(-x))$



(d) $x + \sqrt{x^2 + 1}$

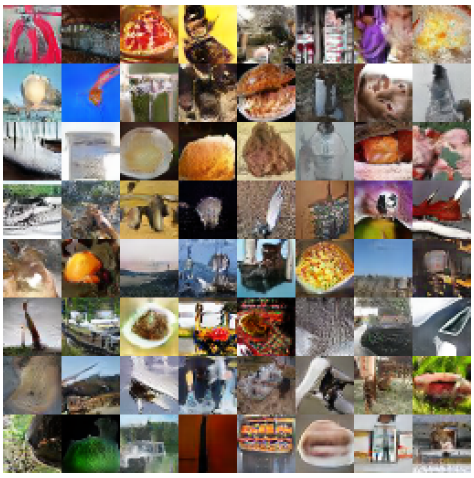


(e) $(x + 1.0)^2$

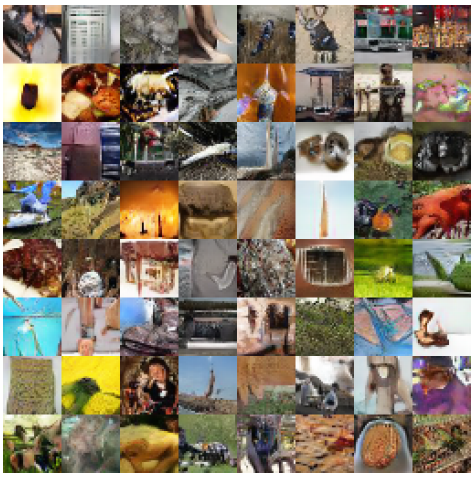


(f) $\max(0, x + 1.0)$

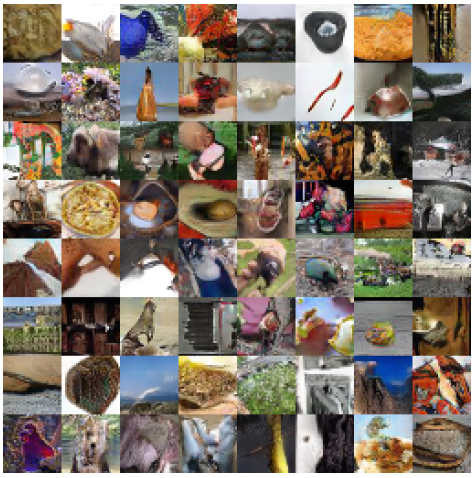
Figure 12: Random samples of LGANs with different loss metrics on CIFAR-10.



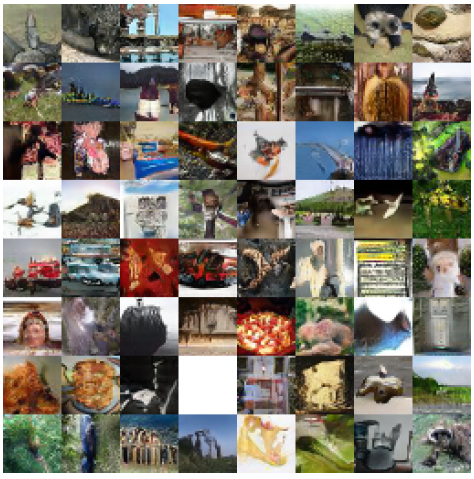
(a) x



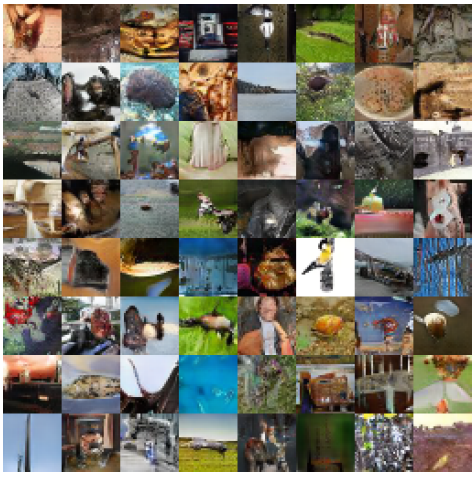
(b) $\exp(x)$



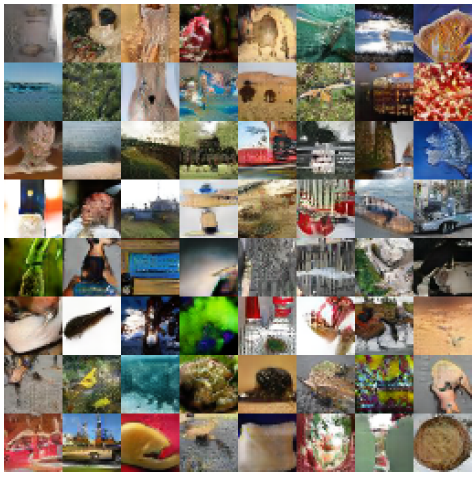
(c) $-\log(\sigma(-x))$



(d) $x + \sqrt{x^2 + 1}$



(e) $(x + 1.0)^2$



(f) $\max(0, x + 1.0)$

Figure 13: Random samples of LGANs with different loss metrics on Tiny-ImageNet.

7. Concretizing the Gradient Uninformativeness

We showed in Section 3 that the gradient uninformativeness is a fundamental cause of training instability of GANs. However, there, only formal arguments are provided to argue for its existence and implication. In this section, we study the theoretical and practical behaviors of the gradient uninformativeness, and demonstrate that practical training schemes could introduce implicit regularization and hence make unregularized GANs possibly work.[7]

7.1 Gaps and Region-Wise Gradient Uninformativeness

To provide a clear intuition of the theoretical behavior of gradient uninformativeness, we introduce the concept of gaps and region-wise gradient uninformativeness.

For unregularized GANs, the optimal discriminative function is undefined on points that is not covered by the supports of the real and fake distributions (i.e., not covered by $S_t \cup S_g$). We name these points with undefined optimal discriminative function value as *gaps*.

If a generated sample is isolated, which means there are gaps around the sample, it is clear that gradient uninformativeness means undefined gradient and hence undefined sample update behavior. However, if a generated sample suffers from the gradient uninformativeness (i.e., no real samples around) but is not isolated (i.e., there are other generated samples around), it might be difficult to describe its uninformative sample update behavior on itself. To overcome this, *we propose to consider a region that is surrounded by gaps as an entirety.*

If such a region is merely formed by generated samples, it is trivial to draw the conclusion that the gradient of the entire region is uninformative and how its support should be updated is undefined.

We know that for an isolated generated sample, no matter whether there is a real sample underneath, it suffers from the gradient uninformativeness. Similarly, for a region that is surrounded by gaps, the existence of real samples inside will not change the fact that the entire region has no information of outside samples, which prevents samples from effectively moving from one region to another (though inside the region, some alignments of real and fake distributions might occur). That is, if consider a region as a whole, its entire gradient is uninformative to other regions/samples, just like an isolated point.

Note that aligning or mapping a source dataset with finite samples to a target dataset, i.e., P_s has finite samples, is the typical case of the existence of isolated samples. When P_s is continuous distribution, a region that is merely formed by generated samples is the typical case. Some unbounded continuous distributions whose tails are light, e.g., Gaussian, practically can also be approximately viewed as a region.

Remark 9 *In fact, for unregularized GANs, even if $P_g = P_t$, as long as there are isolated samples or regions (given $P_g = P_t$, it means these are the attributes of P_t), their gradients will always be uninformative, i.e., there is no theoretical convergence. By contrast, if $P_g = P_t$, the optimal discriminative function of LGANs in any case has substantial zero gradient for all generated samples.*

7. To deliver the key massage without complexity, we restrict our discussion to unregularized GANs.

Remark 10 *For the problem caused by gaps with undefined optimal discriminative function values, a natural solution might require the optimal discriminative function to be at least continuously defined such that it bridges every generate sample to the target distribution. Note that in LGANs, each sample is connected to the target distribution by at least one bounding line. As an example of other possible directions, Unterthiner et al. (2018) define the discriminative function as the potential field of charged particles which also bridges all generated samples to the target distribution.*

7.2 Implicit Regularization in Practical Training

To study the practical behaviors of gradient uninformativeness and understanding how GANs that theoretically suffer from the gradient uninformativeness issue work in practice, we conduct a set of experiments with various hyper-parameter settings and visualize the learned discriminative function and the gradients of generated samples in Figure 14.

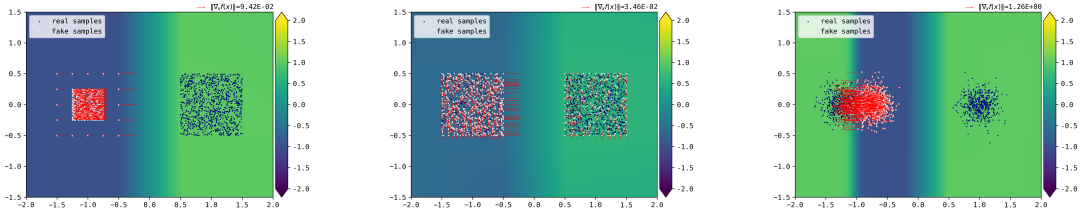
We choose the Least-Squares GANs whose optimal discriminative function is relatively simple as the representative of unregularized GANs in this experiment, which benefits the visualization. In Figure 14, the white points are simulated generated samples, while the blue points are the real samples. The red arrows indicate the gradients of the generated samples, which is beautified by setting the upper and lower bounds of the length. The top-right note gives the gradient scale of unit length. The background color indicates the values of $f(x)$, i.e., shows the value surface of the discriminative function. For Adam, we set $\beta_1 = 0.0$ and $\beta_2 = 0.9$. Please view it in digital mode and zoom in to see the details.

From the figures, we see that even in the showed simple two-dimensional cases, under some hyper-parameter settings, the resulting value surface can be very complex, where the uninformative and theoretically undefined gradients appear clearly chaotic and unfavorable behaviors. On the other hand, it can be noticed that some typical settings (e.g., simple neural network architecture with limited capacity, relatively large learning rate, Adam optimizer) tend to form a relatively simple and smooth value surface, e.g., monotonically increasing from S_g to S_t, making the theoretically uninformative/undefined gradients much more meaningful.

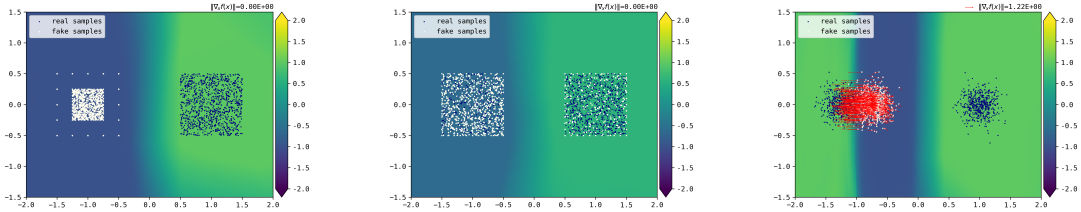
These results show that the practical discriminative function, especially these values on the gaps, highly depends on the hyper-parameter setting. Note that the values on the gaps determine the gradients of the boundary samples of regions and hence determine the updating/moving direction of regions. That is, one can find these settings that implicitly regularize the discriminative function towards being favorable (basically, the gaps are filled such that the increasing direction is towards the target distribution) to mitigate the impact of the gradient uninformativeness issue and make unregularized GANs more likely to work.[8]

Nevertheless, without any explicit theoretical guarantee on the convergence, unregularized GANs are practically hard to use, being unstable and sensitive to hyper-parameters. By contrast, GANs with convergence guarantee (e.g., LGANs) can be much more easy to use. One just need to enlarge the capacity of the discriminator and train it to the best.

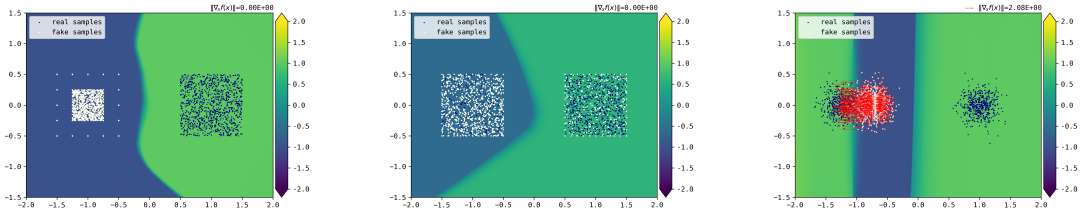
8. Another important empirical technique is to delicately balance the generator and the discriminator. This can be understood as trying to avoid the optimal discriminative function that may overstretch the surface.



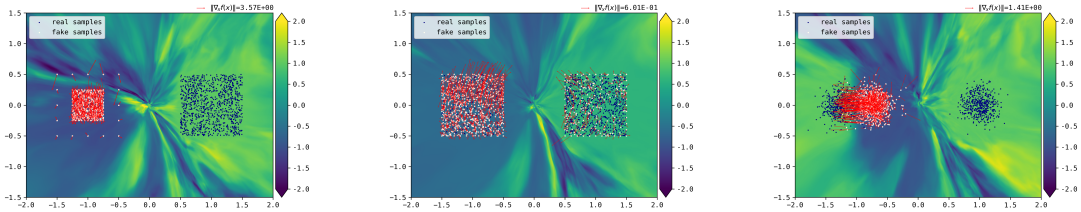
(a) Adam with LR=1e-2. MLP with ReLU. #hidden units=128, #layers=1.



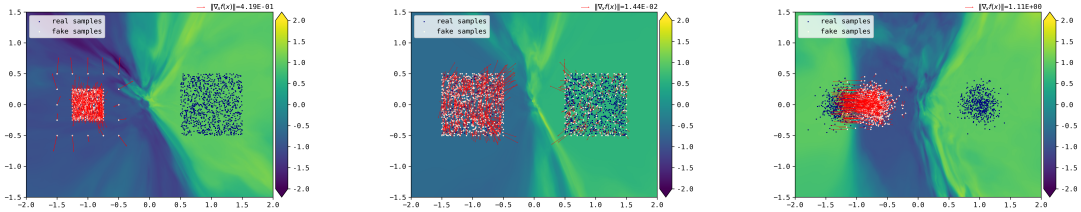
(b) Adam with LR=1e-2. MLP with ReLU. #hidden units=128, #layers=4.



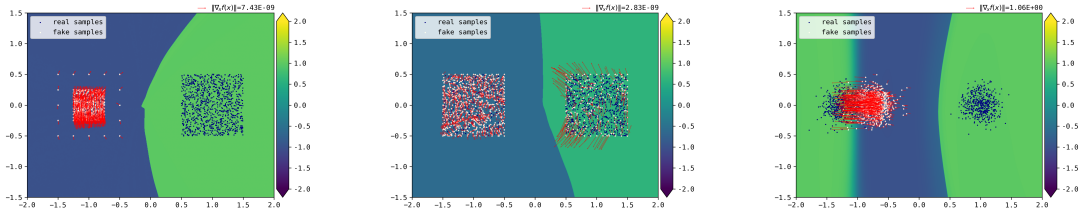
(c) Adam with LR=1e-2. MLP with ReLU. #hidden units=128, #layers=16.



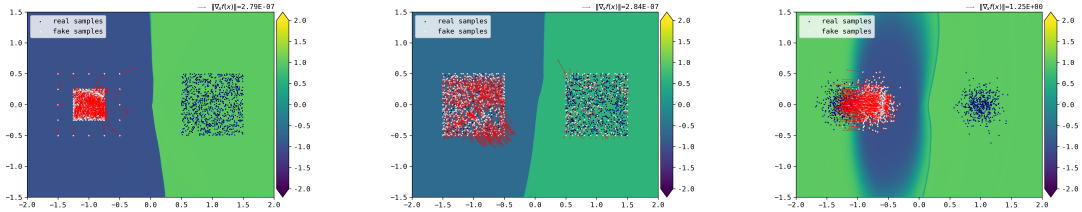
(d) SGD with LR=1e-3. MLP with SELU. #hidden units=128, #layers=64.



(e) SGD with LR=1e-2. MLP with SELU. #hidden units=128, #layers=64.



(f) Adam with LR=1e-3. MLP with SELU. #hidden units=128, #layers=64.



(g) Adam with LR=1e-4. MLP with SELU. #hidden units=128, #layers=64.

Figure 14: The practical discriminative function highly depends on the hyper-parameter setting. Some particular settings (e.g., simple or low-capacity architecture, Adam, large learning rate) tend to lead to simple and smooth value surfaces, where gradients of generated samples generally point towards the target distribution, while some other settings (e.g., complex network, SGD, small learning rate) tend to lead to complex, unfavorable value surfaces. These favorable and unfavorable settings are basically consistent with the empirical success and failure of unregularized GANs in common practice.

Specifically: (a) A simple architecture with insufficient capacity leads to a simple surface. This could be the reason why the DCGAN architecture (Radford et al., 2015) can be very useful for stabilizing the training of GANs. (b) As the network capacity increases, under a proper setting and with a proper learning rate, we may get the optimal discriminative function. (c) However, if the network gets too powerful, since it can approximate the optimal discriminative function possibly in many ways, the perfect optimal discriminative function gets hardly attainable and the value surface starts to get complex. This could be the reason why a too powerful discriminator will lead to training failure of unregularized GANs. (d) As an extreme case, we find the value surface of a deep network learned with SGD and small LR can be very complex. (e) Switching to a larger LR can significantly reduce the complexity of the value surface. (f) Switching to Adam can further lead to a simpler surface. (g) The value surface of Adam keeps very simple, even if we use a quite small learning rate.

8. An Intrinsic Cause of Mode Collapse

Mode collapse is one of the most common issues in GANs' training, which refers to the phenomenon that the generator only learns to produce/imitate parts of the target distribution, while missing some others.

A good deal of literature has tried to study the source of mode collapse (Che et al., 2017; Metz et al., 2017; Kodali et al., 2017; Arora et al., 2017) and/or measure the degree of mode collapse (Odena et al., 2017; Arora and Zhang, 2017; Zhong et al., 2019).

The most recognized cause of mode collapse is that: if the generator is much stronger than the discriminator, particularly, if fix the discriminator and train the generator till optimum, it may learn to only produce the sample(s) in the local or global maximum(s) of the current discriminative function.

The argument above is true for most GANs, even for LGANs. However, from the perspective of optimal discriminative function, there actually exists a much more fundamental cause of mode collapse, i.e., (region-wise) gradient uninformativeness. At a high level, we can say that mode collapse is a manifestation of the nonconvergence caused by gradient uninformativeness, which can be easily explained by the region-wise behavior of gradient uninformativeness.

8.1 Lack of Internal Force to Escape From Mode Collapse

Firstly, we argue that if there are no gaps in P_t, i.e., if P_t is continuously distributed and covers the entire space, theoretically, i.e, if continuously keep the optimality of the generator and discriminator during the training, unregularized GANs can achieve global convergence, no matter what is the initial state of P_g and how P_t is distributed.

The insight behind is that $f^*(x)$-based sample update will always move samples from location where $f^*(x)$ is relatively large to nearby location where $f^*(x)$ is relatively small. Therefore, the global convergence is achieved if and only if f^* gets constant, which implies $P_g = P_t$.

More formally,

Remark 11 *For unregularized GANs, if P_t is continuously distributed and covers the entire space, the Nash equilibrium between the optimal generative function (g^*) and the optimal discriminative function (f^*) is reached if and only if f^* gets constant.*

Remark 12 *For unregularized GANs, constant f^* implies constant $\frac{P_g(x)}{P_t(x)}$ and hence $P_g = P_t$.*

That is, if P_t is continuously distributed and covers the entire space, the generator (ideally) can escape from the mode collapse state. Note that, if P_t covers the entire space, and generator initially only produce noises, it can be considered that the generator is mode-collapsed to these noises, and theoretically, it can later disperse to the entire space.

However, we can also easily get the conclusion that: for unregularized GANs, if there are isolated samples/regions in P_t that does not connect to the samples/regions that P_g currently covers, there is no internal force to guide P_g to cover these isolated samples/regions in P_t, because of the existence of gaps and the correlated uninformative gradients.

Given that the supports of P_g and P_t are typically disjoint, we consider that there lacks of (sufficient) internal force for the generator of unregularized GANs to escape from the mode collapse state. By contrast, LGANs in any case build bounding lines between S_g and S_t. Following the bounding lines, the generator can move samples from one mode to another, and hence can escape from the mode collapse state. In other words, the bounding relationship in LGANs, which is due to the existence of Lipschitz regularization, forms a solid internal force to escape from the mode collapse state.

8.2 The Formation of Mode Collapse

The initial state of the generator is typically producing noises. Being different from P_t, it needs to disperse or update the noise samples towards P_t and thereby cover the modes.

Given that when S_g and S_t are disjoint, the gradients are uninformative, in the extreme case, the generator may fail to find any mode of P_t. This actually also happens in practice from time to time, but people tend to regard it as ordinary training failure.

By tuning the hyper-parameters or changing the network architectures, people find training schema that has better discriminative function values on the gaps, e.g., monotonically increasing from these noises to one mode of P_t. Following the surface of the discriminative function, the generator moves generated samples to the mode. If unfortunately, the values of the gaps around this mode, after the previous process, get smaller than these values inside the mode, these samples will be trapped in this mode, forming a static mode collapse state.

When generated samples collapsed to a mode, the values of the optimal discriminative function on that mode will typically get smaller, and the values of the optimal discriminative function on modes that are missed by the generator are relatively high[9]. If the hyper-parameters and network architectures are tuned well, e.g., prefer monotonic surface, it is likely that the values of the gaps outside the mode are monotonically increasing from this mode towards another mode. This will lead to the typical phenomenon of mode collapse during training, i.e., generated samples are cycling among different modes.

More complicated scenarios, e.g., covering multiple modes while missing some others, can be simulated similarly. The key massage is that: with undefined behaviors of uninformative gradients, there is no theoretical convergence guarantee; however, with partially well-tuned hyper-parameters and network architectures, which make the values of discriminative function on gaps less informative, we may get P_g somehow partially moved to P_t, appeared as *mode collapse*. These simulations also illustrated how the tuning of unregularized GANs can be sensitive and hard to control.

9. The Essence of Convergence Guarantee

In Section 3, we show that unregularized GANs, suffering from the gradient uninformativeness issue, generally does not guarantee its convergence. It implies the necessity of regularization in the discriminative function space. Later in Section 7 and 8, by digging

9. According to Eq. (9). Assuming $\phi'(x) > 0$ and $\varphi'(x) < 0$. For unregularized GANs, if $\phi'(x) > 0$, then $f^*(x)$ is negatively correlated with $P_g(x)$; if $\varphi'(x) < 0$, then $f^*(x)$ is positively correlated with $P_t(x)$.

into the implications of gradient uninformativeness, we further show that the more fundamental cause of nonconvergence can be traced back to the gaps (with undefined optimal discriminative function values) that separating S_g and S_t, and the underlying mechanism of how regularization resolves this issue seems to boil down to the elimination of the gaps, e.g., making S_g and S_t connected by the bounding lines in the optimal discriminative function.

Apart from these results presented, our analysis also indicates a set of useful amendments to the common understandings of the convergence of GANs, which we will illustrate in the following subsections.

9.1 Adversarial Training is Not the Key to Convergence

As the name generative adversarial nets suggested, adversarial is a key conceptional difference between GANs and other generative models, and it is also typically regarded as the key factor to the better quality of samples.

However, it is worth stressing that adversarial training, i.e., the process of discriminating generated samples by assigning low scores while requiring the generated samples to have high activations, does not necessarily guarantee the convergence. In unregularized GANs, for example, due to the existence of gaps, the direction of increasing activation for boundary samples can be arbitrary. In open-ended gap regions, chasing a high activation along the value surface of the gaps can lead a sample arbitrarily far away.

Be that as it may, we have to also recognize that adversarial is indeed a key factor to the high quality of GANs samples, in the sense that: once the sample successfully gets a high activation, though there is no guarantee on the success of this process, it means a high quality in terms of the indistinguishability by the discriminator (Zhou et al., 2018).[10]

The previous analysis is mainly based on unregularized GANs. For regularized GANs, things can be more complex. For a single sample, a higher activation does not always imply better sample quality. However, if consider the entirety . indistinguishability by discriminator. Adversarial training does not guarantee the effective decreasing in terms of the distance metric.

9.2 GANs is Essentially a Sample-based Framework

As we argued in the above, perceiving the minimax between the generator and discriminator as an adversarial training is, though interesting, not helpful for understanding the convergence of GANs. As what Goodfellow et al. (2014) also did, to argue the convergence of the proposed GANs model, it is conventional to show that the max part of the minimax is estimating a distance metric between P_g and P_t. Thereby, we can view GANs as the generator is minimizing a distance metric between P_g and P_t estimated by the discriminator.

10. More strictly, for unregularized GANs, it means the sample is currently at a location where the density ratio $\frac{P_t(x)}{P_g(x)}$ estimated by the discriminator is high, which means it is close to the real samples (i.e., $P_t(x)$ is high), and generated samples are not seriously mode-collapsed to this location (i.e., $P_g(x)$ is not high).

However, we have to highlight that GANs is essentially a sample-based framework, where the probability density/mass of P_g is not directly adjustable, and hence convergence analysis that assumes directly adjustable probability can be unreliable.

First of all, the discriminator is clearly sample-based, which takes samples as input to estimate the distance, instead of the probabilities. As a result, the gradient to the generator is passed through the sample space, i.e., the discriminator tells the generator how to minimize the distance metric by telling how each sample should be updated.

For the generator, if the latent distribution is fixed, which is the typical case, the generator can only adjust the generative function, i.e., how the latent vectors is mapped. In typical optimization schemes, the change of the map is further restricted to be continuous, and hence it cannot directly increase the amount of sample in one location and decrease the amount of sample in another location, i.e., it cannot directly adjust the probability density/mass of P_g. Even if the latent distribution is learnable (Kingma and Welling, 2014; Gurumurthy et al., 2017), whose change may then directly affect the probability density/mass of P_g, due to the sample-based natural of the discriminator, the gradient to the parameters of the latent distribution is passed through the sample space, which means the generator is still essentially trying to match the adjustment of the sample distribution suggested by the discriminator. Nothing directly upon the adjustment of density is specified nor executed.

9.3 Conventional Divergences are Problematic Even in Fully Overlapped Case

already mentioned locality. but more importantly,

Following the arguments in Arjovsky et al. (2017), people tend to believe that the key to successful training of GANs is a good distance metric. It is true that the distance metric is critical and the training issue when S_g and S_t are disjoint can be explained by the illness of conventional distance metric. But we find that when S_g and S_t are fully overlapped, the gradient uninformativeness still exists, which can not be well explained by the distance metric.

9.4 $\nabla_x f^*(x)$ of Wasserstein Distance may also be Uninformative

The Relaxed Dual Form of Wasserstein Distance also Suffers from the Gradient Uninformativeness

Does a good divergence necessarily guarantee convergence?

9.5 The Sample-based Nature of GANs

https://dl.acm.org/doi/pdf/10.1145/3422622

[receive info from sample-based gradient from discriminator.]

if p is optimized.

but p is actually not optimized, the gradient cannot flow to p.

G-D structure. how you parameterize the model. Sample based.

It is common to show that the defined objective of GANs is a distance or divergence between P_g and P_t to argue for its validity. However, it is hardly showed that the distance metric can be successfully optimized. The existence of gradient uninformativeness suggests that it is not a trivial problem.

In this section, we investigate the source of gradient uninformativeness through the lens of the envelope theorem (Milgrom and Segal, 2002) and at the same time answer the question of what are the necessary properties for a distance metric to be $\nabla_x f^*(x)$-optimizable.

9.6 Preliminary: The Envelope Theorem

The envelope theorem is a classic result about the differentiation properties of an optimization problem. For ease of reference, we restate it in the context of GANs as follows:

Let the parameter of the generator be θ and the parameter of discriminator be ϑ. Let $J_D(\theta, \vartheta) = \mathbb{E}_{z\sim P_s}[\phi(f_\vartheta(g_\theta(z)))] + \mathbb{E}_{x\sim P_t}[\varphi(f_\vartheta(x))]$. GANs optimize θ with respect to

$$J(\theta) = \arg\min_{\vartheta} J_D(\theta, \vartheta) \quad s.t. \quad s(\theta, \vartheta) \leq 0. \tag{32}$$

where $s(\theta, \vartheta) \leq 0$ is the constraint that corresponds to the regularization in the discriminative function space, and $J(\theta)$ is a distance metric between P_{g_θ} and P_t.

The corresponding Lagrangian dual problem is given by

$$L(\theta, \vartheta, \lambda) = J_D(\theta, \vartheta) + \lambda \cdot s(\theta, \vartheta), \tag{33}$$

where λ is the Lagrange multiplier.

Let ϑ^* and λ^* together be the solution that minimizes the objective function $L(\vartheta, \lambda; \theta)$. According to the envelope theorem, if J and L are *continuously differentiable*, we have that

$$\nabla_\theta J(\theta) = \nabla_\theta L(\theta, \vartheta, \lambda)\Big|_{\vartheta=\vartheta^*, \lambda=\lambda^*} = \nabla_\theta L(\theta; \vartheta^*, \lambda^*) = \nabla_\theta J_D(\theta; \vartheta^*) + \lambda^* \cdot \nabla_\theta s(\theta; \vartheta^*). \tag{34}$$

9.7 Continuity and Differentiability with Respect to Sample

In Section 3, we show that unregularized GANs commonly suffer from a gradient uninformativeness issue. It is typically considered that the training difficulty of unregularized GANs stems from the fact that its implied distance metric cannot work properly (e.g., being constant) when S_g and S_t are disjoint.

Here, by showing that when the S_g and S_t are totally overlapped, unregularized GANs still suffer from the gradient uninformativeness, we argue that there exists a more fundamental cause of the training instability.

Given that $J(\theta)$ is a distance metric between P_g and P_t, $\nabla_\theta J(\theta)$ tells how the gradient of such a distance metric is delivered to the parameter of the generators, i.e., θ.

9.8 Gradient Information Embedded in the Discriminative Function

9.8.1 Unregularized GANs

As an illustrative example, we first consider the following setting: let P_g be a distribution on two points a and $1+a$ in $\mathbb{R}$ with probability of p and $1-p$, respectively. And the target distribution P_t is evenly distributed on points 0 and 1. Here a and p are the learnable parameters of the generator, and a currently equals 0, which means P_g and P_t are totally overlapped.

In this setting, we allow the generator to directly change the probability distribution indicated by p and also the location of samples indicated by a. Note that

$$J_D(a,p,\vartheta) = p\cdot\phi(f_\vartheta(a)) + (1-p)\cdot\phi(f_\vartheta(1+a)) + 0.5\cdot\varphi(f_\vartheta(0)) + 0.5\cdot\varphi(f_\vartheta(1)).$$

For unregularized GANs, we know that: Theoretically, $f_{\vartheta^*}(x)$ is only defined on the two or four points 0 and 1, a and $1-a$, depending on the value of a. In any case, $\nabla_x f_{\vartheta^*}(x)$ is undefined for all points. And finite value of $f_{\vartheta^*}(x)$ requires $a=0$. If $a\neq 0$, then $|f_{\vartheta^*}(x)|=\infty$ for all these four points.

Now let's consider the gradient of J, applying the envelope theorem:

$$\begin{aligned}
&\nabla_a J(a,p) = \nabla_a J_D(a,p;\vartheta^*) = p\nabla_{f_{\vartheta^*}(a)}\phi(f_{\vartheta^*}(a))\nabla_a f_{\vartheta^*}(a) + (1-p)\nabla_{f_{\vartheta^*}(1+a)}\phi(f_{\vartheta^*}(1+a))\nabla_{(1+a)} f_{\vartheta^*}(1+a);\\
&\nabla_p J(a,p) = \nabla_p J_D(a,p;\vartheta^*) = \phi(f_{\vartheta^*}(a)) - \phi(f_{\vartheta^*}(1+a)).
\end{aligned} \tag{35}$$

Because there is no constraint or regularization, we ignore the term $\lambda^*\cdot\nabla_p s(a,p;\vartheta^*)$.

The formulation of $\nabla_p J(a,p)$ means that: if $a=0$, which leads to well-defined finite-valued $f_{\vartheta^*}(a)$ and $f_{\vartheta^*}(1+a)$, then p has a well-defined gradient; if $a\neq 0$, then the gradient of p is exceptional. On the other side, as evidenced by $\nabla_a J(a,p)$, its gradient for a is always undefined, because $\nabla_x f_{\vartheta^*}(x)$ is always undefined.

We understand the above analysis as:

- The undefined gradient with respect to a stems from the fact that $J(a,p)$ as a function of a is actually not continuously differentiable (by contrast, the Wasserstein distance would be continuously differentiable), i.e., the envelope theorem is actually inapplicable to the setting.
- The well-defined gradient with respect to the density or probability p is interesting, and it reveals that a fundamental limitation of the GANs framework, i.e., it is sample-based because the discriminator takes a sample as input. If GANs is somehow density-based or has direct access to the probability parameters, it might be applicable in more cases.
- In this prototype, p is an explicit parameter in the objective function, hence it might be optimized. However, in practice, the p is usually implicitly given by the different amounts of samples in different locations. Hence, the gradient from J, in practice, also can not pass to p. Because it needs to pass via $\nabla_x f_{\vartheta^*}(x)\cdot\nabla_p x$ and $\nabla_x f_{\vartheta^*}(x)$ is not well-defined.

As a summary, for unregularized GANs, it is common that the overall objective J (the one that is already fully optimized over the discriminator or f, playing the role as a distance

metric between the real and fake distributions to guide the optimization of the generator) is not differentiable with respect to the location of samples, which leads to the undefined gradients. And unfortunately, in the current sample-based GANs formulation where the discriminator takes a sample as input, the gradient must be passed via $\nabla_x f_{\vartheta^*}(x)$. The above two combined together makes unregularized GANs sometimes theoretically not optimizable.

So, given the fact the current GANs formulation is sample-based and the gradient must be passed via $\nabla_x f_{\vartheta^*}(x)$, we maybe should switch to sample-based distance metrics, e.g., optimal transport based metric like Wasserstein distance or these implicitly implied by LGANs.

For the fully overlapped case, J should also have well-defined gradients for the parameters that change the location of samples. However, the underlying objective of J is convex with respect to P_g does not imply the model of J is convex with respect to the generator's parameter θ. And as a matter of fact, we have already known that the gradients from J with respect to samples in unregularized GANs only reflect the local information and tend to lead to model collapse (see Section 8). So, clearly, well-defined gradients or optimizable is actually not a sufficient condition for convergence. The key may lie in the (big) gap between sample-based optimization and density-based distance metric.

9.8.2 Wasserstein Distance with Compact Dual

Arjovsky et al. (2017) has already provided the envelope theorem based analysis for the KR duality of Wasserstein distance. Here, we will analyze our newfound compact dual of Wasserstein distance to gain a deeper understanding on the essence of convergence of GANs.

For Wasserstein distance with the compact dual, to make the analysis even simple, we consider the following case: let P_g be a delta distribution at θ in $\mathbb{R}$ with $\theta < 1$, while P_t is a delta distribution at 1. Then, $J_D(\theta, \vartheta) = f_\vartheta(1) - f_\vartheta(\theta)$ and the constraint is $f_\vartheta(1) - f_\vartheta(\theta) \leq (1 - \theta)$.

Note that, $f_{\vartheta^*}(\theta)$ is only necessarily defined on θ and 1, and $\nabla_x f_{\vartheta^*}(\theta)$ is also undefined for all sample points. Due to the free offset property of Wasserstein distance, we further assume $f(1) = 1$ without loss of generality. Then the problem is simplified as: $J_D(\theta, \vartheta) = 1 - f_\vartheta(\theta)$ with the constraint $f_\vartheta(\theta) - \theta \leq 0$.

The Lagrangian dual problem is given by

$$L(\theta, \vartheta, \lambda) = 1 - f_\vartheta(\theta) + \lambda \cdot (f_\vartheta(\theta) - \theta)). \tag{36}$$

From the envelope theorem, we have

$$\begin{aligned} \nabla_\theta J(\theta) = \nabla_\theta L(\theta; \vartheta^*, \lambda^*) &= -\nabla_\theta f_{\vartheta^*}(\theta) + \lambda^* \cdot \nabla_\theta (f_{\vartheta^*}(\theta) - \theta) \\ &= -\nabla_\theta f_{\vartheta^*}(\theta) + (\lambda^* \cdot \nabla_\theta f_{\vartheta^*}(\theta) - \lambda^*) = (\lambda^* - 1) \cdot \nabla_\theta f_{\vartheta^*}(\theta) - \lambda^*. \end{aligned} \tag{37}$$

By first order optimality condition of the optimal ϑ^* and λ^*, we have:

$$\begin{aligned} \nabla_{\vartheta^*} L &= (\lambda^* - 1) \nabla_{\vartheta^*} f_{\vartheta^*}(\theta) = 0, \\ \nabla_{\lambda^*} L &= f_{\vartheta^*}(\theta) - \theta = 0. \end{aligned} \tag{38}$$

We can notice that $\lambda^* = 1$ is one of its solutions. Applying it to Eq. (37), we get $\nabla_\theta J(\theta) = 1$, which is reasonable and true, and more importantly, we notice that the sample gradient $\nabla_\theta f_{\vartheta^*}(\theta)$, though is still undefined, it is eliminated by the gradient from the constraint.

In summary, for Wasserstein distance with compact dual, because the parameter of the generator is also in the constraint(s). When applying the envelope theorem, it is necessary to consider the gradient from the constraint(s). And it seems the undefined gradient $\nabla_\theta f_{\vartheta^*}(\theta)$ will be somehow eliminated. And the actual gradient, which really takes effect, may come from the remaining part of the gradient from the constraint(s). See, by first order optimality condition Eq (38), it holds $f_{\vartheta^*}(\theta) = \theta$.

9.8.3 With Lipschitz Condition or Lipschitz Regularization

WGANs with Wasserstein distance in KR duality does not involve the parameters of the generator in the constraint of the optimization problem (has the Lipschitz condition). LGANs penalizes the Lipschitz constant, which also does not involve the parameters of the generator in the constraints. So, as long as J is continuously differentiable, the envelope theorem is applicable and we have

$$\begin{aligned}\nabla_\theta J(\theta; \vartheta^*) &= \nabla_\theta J_D(\theta; \vartheta^*) \\ &= \nabla_\theta \mathbb{E}_{z\sim P_s}[\phi(f_{\vartheta^*}(g_\theta(z)))] + \mathbb{E}_{x\sim P_t}[\psi(f_{\vartheta^*}(x))] \\ &= \nabla_\theta \mathbb{E}_{z\sim P_s}[\phi(f_{\vartheta^*}(g_\theta(z)))].\end{aligned} \tag{39}$$

With the Lipschitz condition or penalizing the Lipschitz constant, the objective is intuitively continuously differentiable with respect to P_g. If the generative function is continuous and locally Lipschitz with respect to its parameter θ, then the overall objective J should be continuously differentiable with respect to the generator's parameter θ.

In fact, we have shown in the paper that Lipschitz continuity with respect to Euclidean distance results in excellent gradient properties in terms of $\nabla_{g_\theta(z)} f_{\vartheta^*}(g_\theta(z))$. So, if the generator is continuously differentiable with respect to θ, i.e., if $\nabla_\theta g_\theta(z)$ is well-defined, then $\nabla_\theta \phi(f_{\vartheta^*}(g_\theta(z)))$ and hence Eq. (39) is well-defined, and is expected to well behave.

9.8.4 Sample-Based Distribution Estimation

In unregularized GANs, if $S_g \cup S_t$ does not cover the whole input space, $f^*(x)$ would be undefined outside $S_g \cup S_t$. As a result, the gradient for samples, which are isolated or at the boundary, can be problematic. This also leads to a more serious problem: it prevents samples in one region from adapting to other regions and consequently prevents P_g from converging to P_t.

From the above envelope theorem based analysis, one could notice that the sample-based distribution estimation (i.e., implicit density models, which GANs belong to) is quite different from explicit density estimation (where the distribution is directly parameterized).

When directly parameterizing the distribution (which is usually intractable), the density of sample points can be directly optimized, while in sample-based distribution estimation, to increase or decrease the density of a certain point, it requires modifying samples from being the support of one probability to another.

This is why cases with totally-overlapped distributions also suffer from the faulty gradient. Such a conclusion also reminds us that we need to be cautious when understanding or

proving GANs at the distribution level, because with the discriminator taking a sample as input, GANs is actually sample-based. Note that the generator can actually be density based, if applicable, e.g., when modeling some easily parameterizable distribution.

10. Related Work

We show that Lipschitz regularization is able to ensure the convergence for a family of GANs objectives, which is not limited to the Wasserstein distance. For example, Lipschitz regularization is also imposed to the original GANs (Miyato et al., 2018; Kodali et al., 2017; Fedus et al., 2018), achieving improvements in the quality of generated samples. As a matter of fact, the original GANs' loss metric $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$ is a special case of our LGANs. Thus, our analysis explains why and how Lipschitz regularization works under the original GANs' objective.

Farnia and Tse (2018) also provided some analysis on how the f-divergence would behave when combined with Lipschitz continuity condition, resulting in a new well-behaving distance metric. However, their analysis is limited to the symmetric f-divergence.

Fedus et al. (2018) argued that divergence is not the primary guide of the training of GANs. However, they thought that the original GANs with a non-saturating generator loss metric somehow work. According to our analysis, the original GANs and more generally all unregularized GANs have no guarantee on the convergence, no matter what generator loss metric is used. We have also provided a reasonable explanation on how these unregularized GANs that do not guarantee the convergence work in practice.

Unterthiner et al. (2018) provided some arguments on the unreliability/issues of $\nabla_x f^*(x)$ in the original GANs, which motivates their proposal of Coulomb GANs. However, the arguments in their paper are not thorough. By contrast, we provide a systematic and thorough study over the gradient issues in unregularized GANs. We accordingly propose a new solution, i.e., the Lipschitz regularized GANs, which is a strong rival to the Coulomb GANs, with superior efficiency and sample quality.

Some prior work studies the suboptimal convergence of GANs (Mescheder et al., 2017, 2018; Arora et al., 2017; Liu et al., 2017; Farnia and Tse, 2018; Zhang et al., 2018), which is another important direction for theoretically understanding GANs. Despite the fact that the behavior of suboptimal can be different from the optimal, we think it should, first of all, well-behave under the optimum condition.

Researchers found that applying Lipschitz continuity condition to the generator also benefits the quality of generated samples (Zhang et al., 2019; Odena et al., 2018). Qi (2020) studied the Lipschitz condition from the perspective of loss-sensitive with a Lipschitz data density assumption. However, these are actually different branches and not necessarily related. Their discussions or comparisons are hence out of the scope of this paper.

There exist generative models that do not use $\nabla_x f^*(x)$ as the primal guide of the generator's update. For example, Sanjabi et al. (2018) updated the generator according to the optimal transport plan. However, the sample quality of this branch of works is currently limited.

There are also GANs where the discriminator's input is not a single sample. For example, Li et al. (2017) required a batch of samples, simulating the distribution. And Jolicoeur Martineau (2018) required simultaneously inputting one real sample and one fake sample. Our analysis does not directly apply to their models, but the similar spirit, i.e., analyzing whether the gradient flow between the generator and the discriminator are effective, given the optimal discriminator, can be used to analyze their models.

11. Conclusion

In this paper, we have studied the training instability issue of GANs from the perspective of the optimal discriminative function. By defining the concept of the gradient uninformativeness issue, we have shown that unregularized GANs, where there is no regularization in the discriminative function space, commonly do not guarantee the convergence, which can be a fundamental cause of the training instability.

We have developed the Lipschitz regularized GANs (LGANs) as a general solution to the gradient uninformativeness issue and shown the various favorable theoretical properties. Verifying these theoretical properties raises the requirement of strict Lipschitz regularization implementation. Accordingly, we have explored the existing Lipschitz regularization implementations and found their underlying issues, and hence naturally proposed the max gradient norm penalty and its augmented Lagrangian version as alternatives. We have also verified the theoretical properties of LGANs, showing their consistently superior performance over WGANs.

To provide a more comprehensive understanding of the training instability issue of GANs, we have further studied the gradient issues of unregularized GANs and these gradient issues' practical behaviors, from which we learn the mechanisms of how unregularized GANs work in practice and found a fundamental cause of mode collapse, i.e., the locality of $f^*(x)$ and the gradients.

Finally, to reveal the essence of convergence guarantee of GANs, we have discussed the gradient flow between the generator and the discriminator with the help of the envelope theorem, and found that the key issue might lie in the sample-based nature of the current GANs framework because the discriminator takes a sample as input and the information interchange between the generator and the discriminator must be passed via $\nabla_x f^*(x)$.

Acknowledgements

This work is sponsored by APEX-YITU Joint Research Program. The authors have been supported by National Natural Science Foundation of China (61702327, 61772333, 61632017), Shanghai Sailing Program (17YF1428200), Beijing Municipal Commission of Science and Technology (181100008918005), and Beijing Academy of Artificial Intelligence (BAAI). Zhiming Zhou would like to thank Dachao Lin for a lot of helpful discussions on the central theorems and proofs of LGANs. Zhiming Zhou would like to thank Yuxuan Song and Lantao Yu for their fruitful discussions on the initial idea of LGANs. Zhiming Zhou would like to thank Hongwei Wang for his helps on the writing.

Appendix A. Proofs

In this section, we provide proofs and correlated lemmas to theorems in the main text. Interested reader may find these lemmas and proof details useful.

A.1 Proof of the Compact Dual Form of Wasserstein Distance

We here provide a proof for the proposed compacted dual form of Wasserstein distance, i.e., Eq. (4). We will use $W_1(P_g, P_t)$ to denote the primal form of Wasserstein distance, while use $W_{KR}(P_g, P_t)$ to denote its Kantorovich-Rubinstein (KR) duality and use $W_{KRC}(P_g, P_t)$ to denote the compact dual form.

Theorem 13 *Given $W_1(P_g, P_t) = W_{KR}(P_g, P_t)$, we have $W_1(P_g, P_t) = W_{KRC}(P_g, P_t)$.*

Proof

(i) For any f that satisfies "$f(x) - f(y) \leq d(x, y),\ \forall x, \forall y$", it must satisfy "$f(x) - f(y) \leq d(x, y),\ \forall x \in S_t, \forall y \in S_g$". Thus, $W_{KR}(P_g, P_t) \leq W_{KRC}(P_g, P_t)$.

(ii)
- Let $F_{KRC} = \{f |\, f(x) - f(y) \leq d(x, y),\ \forall x \in S_g, \forall y \in S_t\}$.
- $\forall \pi \in \Pi(P_g, P_t)$, we have the following:

$$\begin{aligned} W_{KRC}(P_g, P_t) &= \sup_{f \in F_{KRC}} \mathbb{E}_{x \sim P_g}[f(x)] - \mathbb{E}_{x \sim P_t}[f(x)] \\ &= \sup_{f \in F_{KRC}} \mathbb{E}_{(x,y) \sim \pi}[f(x) - f(y)] \\ &\leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)]. \end{aligned} \tag{40}$$

- That is, $W_{KRC}(P_g, P_t) \leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)], \forall \pi \in \Pi(P_g, P_t)$.
- Thereby, $W_{KRC}(P_g, P_t) \leq \inf_{\pi \in \Pi(P_g, P_t)} \mathbb{E}_{(x,y) \sim \pi}[d(x, y)] = W_1(P_g, P_t)$.

(iii) Combining (i) and (ii), we have $W_{KR}(P_g, P_t) \leq W_{KRC}(P_g, P_t) \leq W_1(P_g, P_t)$. Given $W_{KR}(P_g, P_t) = W_1(P_g, P_t)$, we have $W_{KR}(P_g, P_t) = W_{KRC}(P_g, P_t) = W_1(P_g, P_t)$.

■

A.2 Proof of Theorem 2

In this proof, we assume P_g and P_t are in the Wasserstein space of order 1, which implies $\mathbb{E}_{x \sim P_g}\|x\| < +\infty$ and $\mathbb{E}_{y \sim P_t}\|y\| < +\infty$, such that $W_1(P_g, P_t)$ is well-defined and has finite value (Villani, 2008, Definition 6.4).

Lemma 14 *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$.*

- *If there exists $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$. Then $\mathfrak{G}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))]$, with f subject to $\kappa(f) \le k_0$, is lower bounded.*

Proof

Given these conditions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))] \\
&\ge \mathbb{E}_{x\sim P_g}[\phi'(a_0)(f(x) - a_0) + \phi(a_0)] + \mathbb{E}_{y\sim P_t}[\varphi'(a_0)(f(x) - a_0) + \varphi(a_0)] \\
&= \phi'(a_0)\mathbb{E}_{x\sim P_g}[f(x)] + \varphi'(a_0)\mathbb{E}_{y\sim P_t}[f(y)] + C_0 \\
&= (\phi'(a_0) + \varphi'(a_0))\mathbb{E}_{x\sim P_g}[f(x)] + \varphi'(a_0)(\mathbb{E}_{y\sim P_t}[f(y)] - \mathbb{E}_{x\sim P_g}[f(x)]) + C_0 \\
&= \kappa(f)\varphi'(a_0)(\mathbb{E}_{y\sim P_t}[\tfrac{f(y)}{\kappa(f)}] - \mathbb{E}_{x\sim P_g}[\tfrac{f(y)}{\kappa(f)}]) + C_0 \qquad (41) \\
&\ge -\kappa(f)|\varphi'(a_0)|W_1(P_g, P_t) + C_0 \\
&\ge -k_0|\varphi'(a_0)|W_1(P_g, P_t) + C_0.
\end{aligned}$$

Therefore, $\mathfrak{G}(f)$ is lower bounded. ■

Lemma 15 *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$.*

- *If there exists $a_1 \in \mathbb{R}$ such that $\phi'(a_1) + \varphi'(a_1) > 0$. Then $\mathfrak{G}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))]$, with $\kappa(f) \le k_0$, is finite implies $f(x) < +\infty$ for all x.*
- *If there exists $a_2 \in \mathbb{R}$ such that $\phi'(a_2) + \varphi'(a_2) < 0$. Then $\mathfrak{G}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))]$, with $\kappa(f) \le k_0$, is finite implies $f(x) > -\infty$ for all x.*

Proof

Given these conditions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))] \\
&\ge \mathbb{E}_{x\sim P_g}[\phi'(a_1)(f(x) - a_1) + \phi(a_1)] + \mathbb{E}_{y\sim P_t}(\varphi'(a_1)[f(x) - a_1) + \varphi(a_1)] \\
&= \phi'(a_1)\mathbb{E}_{x\sim P_g}[f(x)] + \varphi'(a_1)\mathbb{E}_{y\sim P_t}[f(y)] + C_1 \\
&= (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{x\sim P_g}[f(x)] + \varphi'(a_1)(\mathbb{E}_{y\sim P_t}[f(y)] - \mathbb{E}_{x\sim P_g}[f(x)]) + C_1 \\
&= (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{x\sim P_g}[f(x)] + \kappa(f)\varphi'(a_1)(\mathbb{E}_{y\sim P_t}[\tfrac{f(y)}{\kappa(f)}] - \mathbb{E}_{x\sim P_g}[\tfrac{f(x)}{\kappa(f)}]) + C_1 \qquad (42) \\
&\ge (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{x\sim P_g}[f(x)] - \kappa(f)|\varphi'(a_1)|W_1(P_g, P_t) + C_1 \\
&\ge (\phi'(a_1) + \varphi'(a_1))(f(0) - k_0\mathbb{E}_{x\sim P_g}\|x\|) - k_0|\varphi'(a_1)|W_1(P_g, P_t) + C_1.
\end{aligned}$$

If $f(0) \to +\infty$, then $\mathfrak{G}(f) \to +\infty$. Hence, $f(0) < +\infty$. Because $\kappa(f) \le k_0$, we have $f(x) < +\infty$ for all x. The other case can be proved symmetrically. ■

Lemma 16 *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$.*

- *If there exist $a_0, a_1, a_2 \in \mathbb{R}$ such that*

$$\begin{aligned}\phi'(a_0) + \varphi'(a_0) &= 0,\\ \phi'(a_1) + \varphi'(a_1) &> 0,\\ \phi'(a_2) + \varphi'(a_2) &< 0.\end{aligned}$$

Then $\mathfrak{G}(f) = \mathbb{E}_{x\sim P_t}[\phi(f(x))] + \mathbb{E}_{y\sim P_g}[\varphi(f(y))]$, with f subject to $\kappa(f) \le k_0$, has global minima.

Proof

According to Lemma 14, $\mathfrak{G}(f)$ has a lower bound, i.e., $inf(\mathfrak{G}(f)) > -\infty$. Thus, we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{G}(f_n) = inf(\mathfrak{G}(f))$.

Let $\{r_i\}_{i=1}^{\infty}$ be the sequence of all rational points in $dom(f)$. According to Lemma 15, for any $x \in \mathbb{R}$, $\{f_n(x) \,|\, n \in \mathbb{R}\}$ is bounded. By Bolzano-Weierstrass theorem, there is a subsequence $\{f_{1,n}\}_{n=1}^{\infty} \subseteq \{f_n\}_{n=1}^{\infty}$ such that $\{f_{1,n}(r_1)\}_{n=1}^{\infty}$ converges, and there is a subsequence $\{f_{2,n}\}_{n=1}^{\infty} \subseteq \{f_{1,n}\}_{n=1}^{\infty}$ such that $\{f_{2,n}(r_2)\}_{n=1}^{\infty}$ converges. For r_i, there is a subsequence $\{f_{i,n}\}_{n=1}^{\infty} \subseteq \{f_{i-1,n}\}_{n=1}^{\infty}$ such that $\{f_{i,n}(r_i)\}_{n=1}^{\infty}$ converges. Then the sequence $\{f_{m,n}\}_{n=1}^{\infty}$ converges at $\{r_i\}_{i=1}^{m}$.

For all $x \in dom(f)$, $\forall \epsilon > 0$, $\exists r \in \{r_i\}_{n=1}^{\infty}$ such that $\|x - r\| \le \frac{\epsilon}{10k_0}$. So,

$$\begin{aligned}&\lim_{m,l,n\to\infty} |f_{m,n}(x) - f_{l,n}(x)|\\ \le &\lim_{m,l,n\to\infty} (|f_{m,n}(x) - f_{m,n}(r)| + |f_{m,n}(r) - f_{l,n}(r)| + |f_{l,n}(r) - f_{l,n}(x)|) \qquad (43)\\ \le &\lim_{m,l,n\to\infty} (\frac{\epsilon}{10} + \frac{\epsilon}{10} + |f_{m,n}(r) - f_{l,n}(r)|) = \frac{\epsilon}{5}.\end{aligned}$$

Let $\epsilon \to 0$, then $\lim_{m,l,n\to\infty} |f_{m,n}(x) - f_{l,n}(x)| = 0$. So, we can claim that $\{\{f_{m,n}\}_{n=1}^{\infty}\}_{m=1}^{\infty}$ converges at x. Let $g_m = \lim_{n\to\infty} f_{m,n}$. Assume $\{g_m\}_{m=1}^{\infty}$ converges to g.

According to Lemma 15, $\forall m \in \mathbb{N}, \exists C'$ such that $|g_m(0)| \le C'$. So,

$$\begin{aligned}\phi(g_m(x)) &\ge \phi'(a_0)(g_m(x) - a_0) + \phi(a_0)\\ &\ge \phi'(a_0)(g_m(0) - k_0\|x\| - a_0) + \phi(a_0)\\ &= \phi'(a_0)g_m(0) - k_0\phi'(a_0)\|x\| - a_0\phi'(a_0) + \phi(a_0) \qquad (44)\\ &\ge -|\phi'(a_0)|C' - k_0\phi'(a_0)\|x\| - a_0\phi'(a_0) + \phi(a_0)\\ &= -k_0\phi'(a_0)\|x\| + C''.\end{aligned}$$

That is, $\phi(g_m(x)) + k_0\phi'(a_0)\|x\| - C'' \ge 0$.

By Fatou's Lemma,

$$\begin{aligned}
&\mathbb{E}_{x\sim P_g}[\phi(g(x)) + k_0\phi'(a_0)\|x\| - C''] \\
=&\mathbb{E}_{x\sim P_g} \varliminf_{m\to\infty} [\phi(g_m(x)) + k_0\phi'(a_0)\|x\| - C''] \\
\leq& \varliminf_{m\to\infty} \mathbb{E}_{x\sim P_g}[\phi(g_m(x)) + k_0\phi'(a_0)\|x\| - C''] \\
=& \varliminf_{m\to\infty} \mathbb{E}_{x\sim P_g}[\phi(g_m(x))] + \mathbb{E}_{x\sim P_g}[k_0\phi'(a_0)\|x\|] - C''.
\end{aligned} \tag{45}$$

It means $\mathbb{E}_{x\sim P_g}[\phi(g(x))] \leq \varliminf_{m\to\infty} \mathbb{E}_{x\sim P_g}[\phi(g_m(x))]$. Similarly, we have $\mathbb{E}_{y\sim P_t}[\varphi(g(y))] \leq \varliminf_{m\to\infty} \mathbb{E}_{y\sim P_t}[\varphi(g_m(y))]$.

Combining these inequalities, we have

$$\begin{aligned}
\mathfrak{G}(g) &= \mathbb{E}_{x\sim P_g}[\phi(g(x))] + \mathbb{E}_{y\sim P_t}[\varphi(g(y))] \\
&\leq \varliminf_{m\to\infty} \mathbb{E}_{x\sim P_g}[\phi(g_m(x))] + \varliminf_{m\to\infty} \mathbb{E}_{y\sim P_t}[\varphi(g_m(y))] \\
&\leq \varliminf_{m\to\infty} (\mathbb{E}_{x\sim P_g}[\phi(g_m(x))] + \mathbb{E}_{y\sim P_t}[\varphi(g_m(y))]) \\
&= \inf_{\kappa(f)\leq k_0} \mathfrak{G}(f).
\end{aligned} \tag{46}$$

Note that for any $x, y \in dom(g)$, $|g(x) - g(y)| \leq \lim_{m\to\infty}(|g(x) - g_m(x)| + |g_m(x) - g_m(y)| + |g_m(y) - g(y)|) \leq k_0\|x - y\|$, i.e., $\kappa(g) \leq k_0$. That is, g achieves the global minima. ■

Lemma 17 *$\mathfrak{T}(f) = \mathbb{E}_{x\sim P_g}[f(x)] - \mathbb{E}_{y\sim P_t}[f(y)]$, with f subject to $\kappa(f) \leq k_0$, has global minima.*

Proof

It is easy to find that for any $C \in \mathbb{R}$, $\mathfrak{T}(f + C) = \mathfrak{T}(f)$. Similar to the previous lemma, we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{T}(f_n) = inf(\mathfrak{T}(f))$. Without loss of generality, we assume that $f_n(0) = 0, \forall n \in \mathbb{N}^+$. Because $\kappa(f_n) \leq k_0$, we can claim that for any $x \in \mathbb{R}$, $\{f_n(x) \,|\, n \in \mathbb{R}\}$ is bounded. Then we can imitate the method used in Lemma 16 to find the optimal function f^* such that $\mathfrak{T}(f^*) = \inf_{\kappa(f)\leq k_0} \mathfrak{T}(f)$. ■

Lemma 18 *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$.*

- *If there is $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$. Then, if we further assume $S_g \cup S_t$ are bounded, $\mathfrak{G}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))]$, with f subject to $\kappa(f) \leq k_0$, has global minima.*

Proof

We have proved most conditions in previous Lemmas, and we only have to consider the condition that:

- for any $x \in \mathbb{R}$, $\phi'(x) + \varphi'(x) \geq 0$ and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$;
- for any $x \in \mathbb{R}$, $\phi'(x) + \varphi'(x) \leq 0$ and there exists a_2 such that $\phi'(a_2) + \varphi'(a_2) < 0$.

Without loss of generality, we assume that $\phi'(x)+\varphi'(x) \geq 0$ for all x and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$. We know that $\forall x \leq a_0,\ \phi'(x) + \varphi'(x) = 0$, that is, $\phi'(x) = -\varphi'(x)$. Thus, for any $x \leq a_0$, $0 \leq \phi''(x) = -\varphi''(x) \leq 0$, which means $\forall x \leq a_0,\ \phi(x) = -\varphi(x) = tx,\ t \geq 0$.

Similar to the previous Lemmas, we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{G}(f_n) = inf(\mathfrak{G}(f))$. Actually we can assume that for all $n \in \mathbb{N}^+$, there is $f_n(0) \in [-C, C]$, where C is a constant: (i) it is not difficult to find $f_n(0) \leq C$ with Lemma 15; (ii) on the other hand, if $f(0) < -k_0 \cdot \max_{x\in S_t \cup S_g} \|x\| + a_0$, we have $f(x) < a_0$ for all $x \in S_t \cup S_g$, then $\mathfrak{G}(f) = \mathfrak{G}(f - f(0) - k_0 \cdot \max_{x\in S_g\cup S_t} \|x\| + a_0)$. That is, we can have $f(0) >= -k_0 \cdot \max_{x\in S_g\cup S_t} \|x\| + a_0$.

Because $\kappa(f_n) \leq k_0$, we have that for any $x \in \mathbb{R}$, $\{f_n(x) \,|\, n \in \mathbb{R}\}$ is bounded. So we can imitate the method used in Lemma 16 and find the optimal function f^* such that $\mathfrak{G}(f^*) = \inf_{\kappa(f)\leq k_0} \mathfrak{G}(f)$. ■

Lemma 19 (Theorem 2, Existence) *Under the same assumption of Lemma 18, we have* $\mathfrak{F}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))] + \frac{\rho}{2} \cdot \kappa(f)^\alpha$ *with* $\rho > 0$ *and* $\alpha > 1$ *has global minima.*

Proof

When $\kappa(f) = \infty$, it is trivial that $\mathfrak{F}(f) = \infty$. When $\kappa(f) < \infty$, combining Lemma 14, we have $\mathfrak{F}(f) = \mathfrak{G}(f) + \frac{\rho}{2} \cdot \kappa(f)^\alpha \geq -\kappa(f)|\varphi'(a_0)|W_1(P_t, P_g) + \frac{\rho}{2} \cdot \kappa(f)^\alpha + C_0$. When $\rho > 0$ and $\alpha > 1$, the right term is a convex function about $\kappa(f)$, it has a lower bound. So we can find a sequence $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mathfrak{F}(f_n) = \inf_{f\in dom}\, \mathfrak{F}(f)$. It is no doubt that there exists a constant C such that $\kappa(f_n) \leq C$ for all f_n. Then, similar as in previous Lemmas, we can show that, for any point x, $\{f_n(x)\}$ is bounded. So we can imitate the method used in Lemma 16 to find the sequence $\{g_n\}$ such that $\{g_n\} \subseteq \{f_n\}$ and $\{g_n\}_{n=1}^{\infty}$ converge at every point x. Let $\lim_{n\to\infty} g_n = g$. As shown in Lemma 16, we have $\mathfrak{G}(g) \leq \underline{\lim}_{n\to\infty} \mathfrak{G}(g_n)$.

If $\kappa(g) > \underline{\lim}_{n\to\infty} \kappa(g_n)$, then there exist x, y such that $\frac{|g(x)-g(y)|}{\|x-y\|} \geq \underline{\lim}_{n\to\infty} \kappa(g_n) + \epsilon \geq \underline{\lim}_{n\to\infty} \frac{|g_n(x)-g_n(y)|}{\|x-y\|} + \epsilon$. That is, $|g(x) - g(y)| \geq \underline{\lim}_{n\to\infty} |g_n(x) - g_n(y)| + \epsilon\|x - y\| = |g(x) - g(y)| + \epsilon\|x - y\| > |g(x) - g(y)|$. The contradiction implies $\kappa(g) \leq \underline{\lim}_{n\to\infty} \kappa(g_n)$.

Then we have $\mathfrak{F}(g) = \mathfrak{G}(g)+\frac{\rho}{2}\cdot\kappa(g)^\alpha \leq \underline{\lim}_{n\to\infty} \mathfrak{G}(g_n)+\underline{\lim}_{n\to\infty} \frac{\rho}{2}\cdot\kappa(g_n)^\alpha \leq \underline{\lim}_{n\to\infty}(\mathfrak{G}(g_n)+ \frac{\rho}{2} \cdot \kappa(g_n)^\alpha) = \inf \mathfrak{F}(f)$. Thus, the global minima exists. ■

Lemma 20 (Theorem 2, Uniqueness) *Let* ϕ *and* φ *be two convex functions, whose domains are both* $\mathbb{R}$. *If there is* $a_0 \in \mathbb{R}$ *such that* $\phi'(a_0) + \varphi'(a_0) = 0$. *Then if* ϕ *or* φ *is strictly convex, then the minimizer of* $\mathfrak{F}(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{y\sim P_t}[\varphi(f(y))] + \frac{\rho}{2} \cdot \kappa(f)^\alpha$ *with* $\rho > 0$ *and* $\alpha > 1$ *is unique (on the support of* $S_g \cup S_t$*).*

Proof

Without loss of generality, we assume that ϕ is strictly convex. By the strict convexity of ϕ, we have $\forall x, y \in \mathbb{R},\ \phi(\frac{x+y}{2}) < \frac{1}{2}(\phi(x) + \phi(y))$.

Assume f_1 and f_2 are two different minimizers of $\mathfrak{F}(f)$.

First, we have

$$\begin{aligned}\kappa\Big(\frac{f_1+f_2}{2}\Big) &= \sup_{x,y} \frac{\frac{f_1(x)+f_2(x)}{2} - \frac{f_1(y)+f_2(y)}{2}}{\|x-y\|} \\ &\le \sup_{x,y} \frac{1}{2}\frac{|f_1(x)-f_1(y)| + |f_2(x)-f_2(y)|}{\|x-y\|} \\ &\le \frac{1}{2}\Big(\sup_{x,y}\frac{|f_1(x)-f_1(y)|}{\|x-y\|} + \sup_{x,y}\frac{|f_2(x)-f_2(y)|}{\|x-y\|}\Big) \\ &= \Big(\frac{\kappa(f_1)+\kappa(f_2)}{2}\Big).\end{aligned} \tag{47}$$

Given $\rho > 0$ and $\alpha > 1$, we have

$$\begin{aligned}\frac{\rho}{2}k\Big(\frac{f_1+f_2}{2}\Big)^\alpha &\le \frac{\rho}{2}\Big(\frac{\kappa(f_1)+\kappa(f_2)}{2}\Big)^\alpha \\ &\le \frac{\rho}{2}\Big(\frac{\kappa(f_1)^\alpha+\kappa(f_2)^\alpha}{2}\Big).\end{aligned} \tag{48}$$

Then we have

$$\begin{aligned}\mathfrak{F}\Big(\frac{f_1+f_2}{2}\Big) &= \mathbb{E}_{x\sim P_g}\phi\Big(\frac{f_1(x)+f_2(x)}{2}\Big) + \mathbb{E}_{y\sim P_t}\varphi\Big(\frac{f_1(y)+f_2(y)}{2}\Big) + \frac{\rho}{2}k\Big(\frac{f_1+f_2}{2}\Big)^\alpha \\ &< \mathbb{E}_{x\sim P_g}\Big(\frac{\phi(f_1(x))+\phi(f_2(x))}{2}\Big) + \mathbb{E}_{y\sim P_t}\varphi\Big(\frac{f_1(y)+f_2(y)}{2}\Big) + \frac{\rho}{2}k\Big(\frac{f_1+f_2}{2}\Big)^\alpha \\ &\le \mathbb{E}_{x\sim P_g}\Big(\frac{\phi(f_1(x))+\phi(f_2(x))}{2}\Big) + \mathbb{E}_{y\sim P_t}\Big(\frac{\varphi(f_1(y))+\varphi(f_2(y))}{2}\Big) + \frac{\rho}{2}k\Big(\frac{f_1+f_2}{2}\Big)^\alpha \\ &\le \mathbb{E}_{x\sim P_g}\Big(\frac{\phi(f_1(x))+\phi(f_2(x))}{2}\Big) + \mathbb{E}_{y\sim P_t}\Big(\frac{\varphi(f_1(y))+\varphi(f_2(y))}{2}\Big) + \frac{\rho}{2}\Big(\frac{\kappa(f_1)^\alpha+\kappa(f_2)^\alpha}{2}\Big) \\ &= \frac{1}{2}(\mathfrak{F}(f_1)+\mathfrak{F}(f_2)) = \inf\mathfrak{F}(f).\end{aligned} \tag{49}$$

We get a contradiction $\mathfrak{F}(\frac{f_1+f_2}{2}) < \inf\mathfrak{F}(f)$, which implies that the minimizer of $\mathfrak{F}(f)$ is unique. ■

A.3 Proof of Theorem 3

Let $J_D(f) = \mathbb{E}_{x\sim P_g}[\phi(f(x))] + \mathbb{E}_{x\sim P_t}[\varphi(f(x))]$. Let $\mathring{J}_D(f,x) = P_g(x)\phi(f(x)) + P_t(x)\varphi(f(x))$. It has $J_D(f) = \int \mathring{J}_D(f,x)\,\mathrm{d}x$. Let $J(f) = J_D(f) + \frac{\rho}{2}\cdot\kappa(f)^\alpha$. It has $J(f) = J(\bar{f},\kappa(f)) = J_D(\bar{f},\kappa(f)) + \frac{\rho}{2}\cdot\kappa(f)^\alpha$, where $\bar{f} = f/\kappa(f)$ and $J_D(\bar{f},\kappa(f)) = \mathbb{E}_{x\sim P_g}[\phi(\bar{f}(x)\cdot\kappa(f))] +$

$\mathbb{E}_{x\sim P_t}[\varphi(\bar{f}(x)\cdot\kappa(f))]$. Let $J_{D^*}(\kappa(f)) = \min_{\bar{f}} \mathbb{E}_{x\sim P_g}[\phi(\bar{f}(x)\cdot\kappa(f))] + \mathbb{E}_{x\sim P_t}[\varphi(\bar{f}(x)\cdot\kappa(f))]$. It has $\min_f J(f) = \min_{\kappa(f)} J_{D^*}(\kappa(f)) + \frac{\rho}{2}\cdot\kappa(f)^\alpha$. Let $J_*(\kappa(f)) = J_{D^*}(\kappa(f)) + \frac{\rho}{2}\cdot\kappa(f)^\alpha$. Let $f^* = \arg\min_f J(f)$. It has $\kappa(f^*) = \arg\min_{\kappa(f)} J_*(\kappa(f))$.

Lemma 21 *It holds $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$ for all x, if and only if, $\kappa(f^*) = 0$.*

Proof

(i) If $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$ holds for all x, then $\kappa(f^*) = 0$.

For the optimal f^*, it holds that $\nabla_{\kappa(f^*)}J_* = \nabla_{\kappa(f^*)}J_{D^*} + \frac{\rho}{2}\cdot\alpha\kappa(f^*)^{\alpha-1} = 0$.

$\nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$ for all x implies $\nabla_{\kappa(f^*)}J_{D^*} = 0$. Thus, we conclude that $\kappa(f^*) = 0$.

(ii) If $\kappa(f^*) = 0$, then $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$ holds for all x.

For the optimal f^*, it holds that $\nabla_{\kappa(f^*)}J = \nabla_{\kappa(f^*)}J_{D^*} + \frac{\rho}{2}\cdot\alpha\kappa(f^*)^{\alpha-1} = 0$.

So, $\kappa(f^*) = 0$ implies $\nabla_{\kappa(f^*)}J_{D^*} = 0$. Besides, $\kappa(f^*) = 0$ also implies $\forall x, y, f^*(x) = f^*(y)$. If there exists some point x such that $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) \neq 0$, it is obvious that $\nabla_{\kappa(f^*)}J_{D^*} \neq 0$. It is contradictory to $\nabla_{\kappa(f^*)}J_{D^*} = 0$. Thus, we have $\forall x, \nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$. ■

Lemma 22 *If $\forall x, y, f^*(x) = f^*(y)$, then $P_t = P_g$.*

Proof

$\forall x, y, f^*(x) = f^*(y)$ implies $\kappa(f^*) = 0$. According to Lemma 21, for all x it holds $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) = 0$, i.e., $P_g(x)\nabla_{f^*(x)}\phi(f^*(x)) + P_t(x)\nabla_{f^*(x)}\varphi(f^*(x)) = 0$. Thus, $\frac{P_g(x)}{P_t(x)} = -\frac{\nabla_{f^*(x)}\varphi(f^*(x))}{\nabla_{f^*(x)}\phi(f^*(x))}$. That is, $\frac{P_g(x)}{P_t(x)}$ has a constant value, which implies $P_t = P_g$. ■

Proof [**Theorem 3**]

(a): Consider x with $\nabla_{f^*(x)}\mathring{J}_D(f^*, x) \neq 0$. Let $\kappa(f, x) = \sup_y \frac{|f(y)-f(x)|}{\|y-x\|}$.

(i) If $\forall\delta$ s.t. $\forall\epsilon$ there exist $z, w \in B(x, \epsilon)$ such that $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} \geq \kappa(f^*) - \delta$, which means there exists t such that $f^{*\prime}(t) \geq \kappa(f^*) - \delta$, because $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} = \frac{\int_w^z f^{*\prime}(t)dt}{\|z-w\|}$. Let $\epsilon \to 0$, we have $t \to x$. Assume f^* is smooth, then $f^{*\prime}(t) \to f^{*\prime}(x)$. Let $\delta \to 0$, we have $f^{*\prime}(t) \to \kappa(f^*)$. we have that $f^{*\prime}(x) = \kappa(f^*)$.

(ii) Assume that $\exists\delta$ s.t. $\exists\epsilon$ and for all $z, w \in B(x, \epsilon)$, $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} < \kappa(f^*) - \delta$. Consider the following condition, for all δ_2 and $\epsilon_2 \in (0, \epsilon/2)$, $\exists y \in B(x, \epsilon_2)$, such that $\kappa(y) > \kappa(f^*) - \delta_2$. Then there exists a sequence of $\{y_n\}_{n=1}^\infty$ s.t. $\lim_{n\to\infty}\frac{|f(y)-f(y_n)|}{\|y-y_n\|} = k(y)$. Then there exists a y' such that $\frac{|f(y)-f(y')|}{\|y-y'\|} \geq k - \delta_2$. According to the assumption, we have $\|y - y'\| \geq \frac{\epsilon}{2}$. Then $k(x) \geq \frac{|f^*(x)-f^*(y)|}{\|x-y\|} \geq \frac{|f^*(y)-f^*(y')|-|f^*(x)-f^*(y)|}{\|x-y\|+\|y-y'\|} \geq \frac{|f^*(y)-f^*(y')|-k\|x-y\|}{\|x-y\|+\|y-y'\|} \geq (k -$

$\delta_2)\frac{\|y-y'\|}{\|x-y\|+\|y-y'\|} - k\frac{\|x-y\|}{\|x-y\|+\|y-y'\|} \geq (1-\frac{\epsilon_2}{\epsilon_2+\|y-y'\|})(k-\delta_2) - k\frac{\epsilon_2}{\|y-y'\|} \geq (1-\frac{\epsilon_2}{\epsilon_2+\|y-y'\|})(k-\delta_2) - k\frac{\epsilon_2}{\|y-y'\|}$. Let $\epsilon_2 \to 0$ and $\delta_2 \to 0$. We get $k(x) = k$, which means there exists a y such that $|f^*(y) - f^*(x)| = k\|y-x\|$.

(iii) Now we can assume $\exists\delta_2$ s.t. $\exists\epsilon_2$ and for all $y \in B(x,\epsilon_2)$, such that $k(y) \leq k - \delta_2$. If $\nabla_{f^*(x)}\mathring{J}_D(f^*,x) \neq 0$, without loss of generality, we can assume $\nabla_{f^*(x)}\mathring{J}_D(f^*,x) > 0$. Then, for all $y \in B(x,\epsilon_2)$, we have $\nabla_{f^*(y)}\mathring{J}_D(y) > 0$, as long as ϵ_2 is small enough. Now we change the value of $f^*(y)$ for $y \in B(x,\epsilon_2)$. Let $g(y) = \begin{cases} f^*(y) - \frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2}), & y \in B(x,\epsilon_2); \\ f^*(y) \qquad \text{otherwise.} \end{cases}$. Because $\nabla_{f^*(y)}\mathring{J}_D(y) > 0$, $\forall y \in B(x,\epsilon_2)$, when N is sufficiently large, it is not difficult to show $J_D(g) < J_D(f^*)$. We next verify that $\|g\|_{Lip} \leq k$. For any y, z, if $y, z \notin B(x,\epsilon_2)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} < k$. If $y \in B(x,\epsilon_2)$, $z \notin B(x,\epsilon_2)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} \leq \frac{|(f^*(y)-f^*(z)|+\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2}))}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2})}{\epsilon_2-\|x-y\|} = \frac{|(f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N} \leq k(y) + \frac{1}{N} \leq k - \delta_2 + \frac{1}{N} < k$ (when $N \gg \frac{1}{\delta_2}$). If $y, z \in B(x,\epsilon)$, then $\frac{|g(y)-g(z)|}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|+|\frac{\epsilon_2}{N}(1-\frac{\|x-y\|}{\epsilon_2})-\frac{\epsilon_2}{N}(1-\frac{\|x-z\|}{\epsilon_2})|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{\frac{\epsilon_2}{N}(\frac{\|x-y\|-\|x-z\|}{\epsilon_2})|}{\|y-z\|} \leq \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N}\frac{\|y-z\|}{\|y-z\|} = \frac{|f^*(y)-f^*(z)|}{\|y-z\|} + \frac{1}{N} \leq k - \delta_2 + \frac{1}{N} < k$ (when $N \gg \frac{1}{\delta_2}$). So, we have $\|g\|_{Lip} \leq k$. But we have $J_D(g) < J_D(f^*)$. The contradiction tells us that there must exist a y such that $|f^*(y) - f^*(x)| = k\|y-x\|$.

(b): For $x \in S_t \cup S_g - S_t \cap S_g$, assuming $P_g(x) \neq 0$ and $P_t(x) = 0$, we have $\nabla_{f^*(x)}\mathring{J}_D(f^*,x) = P_g(x)\nabla_{f^*(x)}\phi(f^*(x)) + P_t(x)\nabla_{f^*(x)}\varphi(f^*(x)) = P_g(x)\nabla_{f^*(x)}\phi(f^*(x)) > 0$, because $P_g(x) > 0$ and $\nabla_{f^*(x)}\phi(f^*(x)) > 0$. Then according to (a), there must exist a y such that $|f^*(y) - f^*(x)| = \kappa(f^*)\cdot\|y-x\|$. The other situation can be proved in the same way.

(c): According to Lemma 22, in the situation that $P_t \neq P_g$, for the optimal f^*, there must exist at least one pair of points x and y such that $y \neq x$ and $f^*(x) \neq f^*(y)$. It also implies that $\kappa(f^*) > 0$. Then according to Lemma 21, there exists a point x such that $\nabla_{f^*(x)}\mathring{J}_D(f^*,x) \neq 0$. According to (a), there exists y with $y \neq x$ satisfying that $|f^*(y) - f^*(x)| = \kappa(f^*)\cdot\|y-x\|$.

(d): In Nash equilibrium state, it holds that, for any $x \in S_t \cup S_g$, $\nabla_{\kappa(f)}J = \nabla_{\kappa(f)}J_{D^*} + 2\frac{\rho}{2}\cdot\kappa(f) = 0$ and $\nabla_{f(x)}\mathring{J}_D(f,x)\nabla_x f(x) = 0$. We claim that in the Nash equilibrium state, the Lipschitz constant $\kappa(f)$ must be 0. If $\kappa(f) \neq 0$, according to Lemma 21, there must exist a point $\hat{x}$ such that $\nabla_{f(\hat{x})}\mathring{J}_D(f,\hat{x}) \neq 0$. And according to (a), it must hold that $\exists\hat{y}$ fitting $|f(\hat{y}) - f(\hat{x})| = \kappa(f)\cdot\|\hat{x}-\hat{y}\|$. According to Theorem 5, we have $\left\|\nabla_{\hat{x}}f(\hat{x})\right\| = \kappa(f) \neq 0$. This is contradictory to that $\nabla_{f(\hat{x})}\mathring{J}_D(f,\hat{x})\nabla_{\hat{x}}f(\hat{x}) = 0$. Thus $\kappa(f) = 0$. That is, $\forall x \in S_t \cup S_g$, $\nabla_x f(x) = 0$, which means $\forall x, y, f(x) = f(y)$. According to Lemma 22, $\forall x, y, f(x) = f(y)$ implies $P_t = P_g$. Thus $P_t = P_g$ is the only Nash equilibrium in our system. ■

Remark 23 *For the Wasserstein distance, $\nabla_{f^*(x)}\mathring{J}_D(f^*,x) = 0$ if and only if $P_t(x) = P_g(x)$. For the Wasserstein distance, penalizing the Lipschitz constant also benefits: at the convergence state, it will hold $\nabla_x f^*(x) = 0$ for all x.*

A.4 Proof of Theorem 4

Lemma 24 *Let k be the Lipschitz constant of f. If $f(a)-f(b)=k\|a-b\|$ and $f(b)-f(c)=k\|b-c\|$, then $f(a)-f(c)=k\|a-c\|$ and $(a,f(a)),(b,f(b)),(c,f(c))$ lies in the same line.*

Proof $f(a)-f(c)=f(a)-f(b)+f(b)-f(c)=k\|a-b\|+k\|b-c\|\geq k\|a-c\|$. Because the Lipschitz constant of f is k, we have $f(a)-f(c)\leq k\|a-c\|$. Thus $f(a)-f(c)=k\|a-c\|$. Because the triangle equality holds, we have a,b,c is in the same line. Furthermore, because $f(a)-f(b)=k\|a-b\|$, $f(b)-f(c)=k\|b-c\|$ and $f(a)-f(c)=k\|a-c\|$, we have $(a,f(a)),(b,f(b)),(c,f(c))$ lies in the same line. ■

Lemma 25

- *For any x with $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)>0$, there exists a y with $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)<0$ such that $f^*(y)-f^*(x)=\kappa(f^*)\|y-x\|$;*
- *For any y with $\nabla_{f^*(y)}\mathring{J}_D(f^*,y)<0$, there exists a x with $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)>0$ such that $f^*(y)-f^*(x)=\kappa(f^*)\|y-x\|$.*

Proof Consider x with $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)>0$. According to Theorem 3, there exists y such that $|f^*(y)-f^*(x)|=\kappa(f^*)\|y-x\|$. Assume that for every y that holds $|f^*(y)-f^*(x)|=\kappa(f^*)\|y-x\|$, it has $\nabla_{f^*(y)}\mathring{J}_D(f^*,y)\geq 0$. Consider the set $S(x)=\{y\mid f^*(y)-f^*(x)=\kappa(f^*)\|y-x\|\}$. Note that, according to Lemma 24, any z that holds $f^*(z)-f^*(y)=\kappa(f^*)\|z-y\|$ for any $y\in S(x)$ will also be in $S(x)$. Similar to the proof of (a) in Theorem 3, we can decrease the value of $f^*(y)$ for all $y\in S(x)$ to construct a better f. By contradiction, we have that there must exist a y with $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)<0$ such that $|f^*(y)-f^*(x)|=\kappa(f^*)\|y-x\|$. Given the fact $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)>0$ and $\nabla_{f^*(x)}\mathring{J}_D(f^*,x)<0$, we can conclude that $f^*(y)>f^*(x)$ and $f^*(y)-f^*(x)=\kappa(f^*)\|y-x\|$. Otherwise, if $f^*(x)-f^*(y)=\kappa(f^*)\|y-x\|$, then we can construct a better f by decreasing $f^*(x)$ and increasing $f^*(y)$ which does not break the k-Lipschitz constraint. The other case can be proved similarly. ■

Lemma 26 *For any x, if $\nabla_{f(x)}\mathring{J}_D(f,x)>0$, then $P_g(x)>0$. For any y, if $\nabla_{f(y)}\mathring{J}_D(f,y)<0$, then $P_t(y)>0$.*

Proof $\nabla_{f(x)}\mathring{J}_D(f,x)=P_g(x)\nabla_{f(x)}\phi(f(x))+P_t(x)\nabla_{f(x)}\varphi(f(x))$. And we know $\phi'(x)>0$ and $\varphi'(x)<0$. Naturally, $\nabla_{f(x)}\mathring{J}_D(f,x)>0$ implies $P_g(x)>0$. Similarly, $\nabla_{f(y)}\mathring{J}_D(f,y)<0$ implies $P_t(y)>0$. ■

Proof [**Theorem 4**]

For any $x \in S_g$, if $\nabla_{f^*(x)} \mathring{J}_D(f^*, x) > 0$, according to Lemma 25, there exists a y with $\nabla_{f^*(x)} \mathring{J}_D(f^*, x) < 0$ such that $f^*(y) - f^*(x) = \kappa(f^*)\|y - x\|$. According to Lemma 26, we have $P_t(y) > 0$. That is, there is a $y \in S_t$ such that $f^*(y) - f^*(x) = \kappa(f^*)\|y - x\|$. We can prove the other case symmetrically. ■

Remark 27 *$\nabla_{f^*(x)} \mathring{J}_D(f^*, x) < 0$ for some $x \in S_g$ means x is at the overlapping region of S_t and S_g. It can be regarded as a $y \in S_t$, and one can apply the other rule which guarantees that there exists an $x' \in S_g$ that bounds this point.*

A.5 Proof of Theorem 5 and the Necessity of Euclidean Distance

In this section, we will prove Theorem 5, i.e., Lipschitz continuity with l_2-norm (Euclidean Distance) can guarantee that the gradient is directly pointing towards some sample, and at the same time, demonstrate that the other norms do not have this property.

Let (x, y) be such that $y \neq x$, and we define $x_t = x + t \cdot (y - x)$ with $t \in [0, 1]$.

Lemma 28 *If $f(x)$ is k-Lipschitz with respect to $\|\cdot\|_p$ and $f(y) - f(x) = k\|y - x\|_p$, then $f(x_t) = f(x) + t \cdot k\|y - x\|_p$*

Proof As we know $f(x)$ is k-Lipschitz, with the property of norms, we have

$$\begin{aligned} f(y) - f(x) &= f(y) - f(x_t) + f(x_t) - f(x) \\ &\leq f(y) - f(x_t) + k\|x_t - x\|_p = f(y) - f(x_t) + t \cdot k\|y - x\|_p \\ &\leq k\|y - x_t\|_p + t \cdot k\|y - x\|_p = k \cdot (1 - t)\|y - x\|_p + t \cdot k\|y - x\|_p \\ &= k\|y - x\|_p. \end{aligned} \tag{50}$$

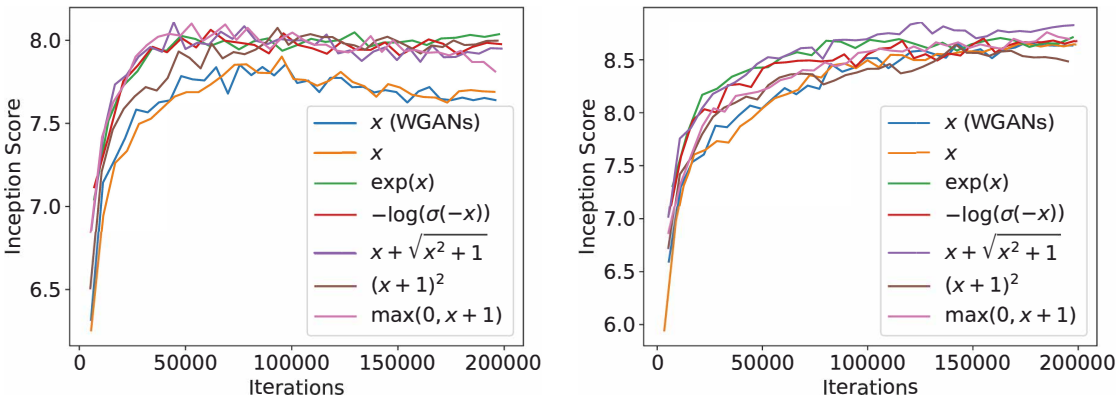
$f(y) - f(x) = k\|y - x\|_p$ implies all the inequalities are equalities. Therefore, $f(x_t) = f(x) + t \cdot k\|y - x\|_p$. ■

Lemma 29 *Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. If $f(x_t) = f(x) + t \cdot k\|y - x\|_2$, then $\nabla_v f(x_t)$ equals k.*

Proof

$$\begin{aligned} \nabla_v f(x_t) &= \lim_{h \to 0} \frac{f(x_t + hv) - f(x_t)}{h} = \lim_{h \to 0} \frac{f(x_t + h\frac{y-x}{\|y-x\|_2}) - f(x_t)}{h} \\ &= \lim_{h \to 0} \frac{f(x_{t+\frac{h}{\|y-x\|_2}}) - f(x_t)}{h} = \lim_{h \to 0} \frac{\frac{h}{\|y-x\|_2} \cdot k\|y - x\|_2}{h} = k. \end{aligned}$$

■



(a) CIFAR-10. (b) Tiny-ImageNet.

Figure 15: Training curves in terms of IS. WGANs and a set of instances of LGANs.

Proof [**Theorem 5**] Assume $p = 2$. According to (Adler and Lunz, 2018), if $f(x)$ is k-Lipschitz with respect to $\|.\|_2$ and $f(x)$ is differentiable at x_t, then $\|\nabla f(x_t)\|_2 \leq k$. Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. We have

$$k^2 = k\nabla_v f(x_t) = k < v, \nabla f(x_t) >=< kv, \nabla f(x_t) > \leq \|kv\|_2 \|\nabla f(x_t)\|_2 = k^2. \tag{51}$$

Because the equality holds only when $\nabla f(x_t) = kv = k\frac{y-x}{\|y-x\|_2}$, we have that $\nabla f(x_t) = k\frac{y-x}{\|y-x\|_2}$. ■

Above proof utilizes the property that $\|\nabla f(x_t)\|_2 \leq k$, which is derived from that $f(x)$ is k-Lipschitz with respect to $\|.\|_2$. However, other norms do not satisfy this property. Specifically, according to the theory in (Adler and Lunz, 2018): if a differentiable function f is k-Lipschitz with respect to norm $\|.\|_p$, then the Lipschitz continuity actually implies a bound on the dual norm of gradients, i.e., $\|\nabla f\|_q \leq k$. Here $\|.\|_q$ is the dual norm of $\|.\|_p$, which satisfies $\frac{1}{p} + \frac{1}{q} = 1$. As we could notice, a norm is equal to its dual norm if and only if $p = 2$. Switching to l_p-norm with $p \neq 2$, it is actually bounding the l_q-norm of the gradients. However, bounding the l_q-norm of the gradients does not guarantee the gradient direction at fake samples point towards real samples. A counter-example is provided as follows.

Consider a function $g(x, y) = x + y$ on $\mathbb{R}^2$. We have for all (x_1, y_1), (x_2, y_2), there is $g(x_1, y_1) - g(x_2, y_2) = (x_1 - x_2) + (y_1 - y_2) \leq |x_1 - x_2| + |y_1 - y_2| = \|(x_1, y_1) - (x_2, y_2)\|_1$, which means g is a 1-Lipschitz function with respect to l_1-norm. According to the above, the dual norm of ∇g is bounded, with $\|\nabla g\|_\infty \leq 1$; one could also verify that ∇g is equal to $(1, 1)$ at every point in $\mathbb{R}^2$ with $\|\nabla g\|_\infty = 1$. However, selecting two points $A = (0, 0)$ and $B = (2, 1)$, we have $g(A) - g(B) = \|A - B\|_1$, but we can notice that $\nabla g(A) = (1, 1)$, which is **not directly** pointing towards B.

Note that different norms will induce different gradients with different properties (Adler and Lunz, 2018). We here expect the gradient directly points towards a real sample.

References

Jonas Adler and Sebastian Lunz. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6754–6763, 2018.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

Sanjeev Arora and Yi Zhang. Do GANs actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets. In *International Conference on Machine Learning (ICML)*, pages 224–232, 2017.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding (CVIU)*, 179:41–65, 2019.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Farzan Farnia and David Tse. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5248–5258, 2018.

William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *International Conference on Learning Representations (ICLR)*, 2018.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

Alexia Jolicoeur Martineau. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*, 2017.

Chengtao Li, David Alvarez-Melis, Keyulu Xu, Stefanie Jegelka, and Suvrit Sra. Distributional adversarial networks. *arXiv preprint arXiv:1706.09549*, 2017.

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5545–5553, 2017.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1825–1835, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning (ICML)*, pages 3478–3487, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Youssef Mroueh and Tom Sercu. Fisher GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2513–2523, 2017.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *International Conference on Learning Representations (ICLR)*, 2018.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning (ICML)*, pages 2642–2651, 2017.

Augustus Odena, Jacob Buckman, Catherine Olsson, Tom Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning (ICML)*, pages 3849–3858, 2018.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Guo-Jun Qi. Loss-sensitive generative adversarial networks on Lipschitz densities. *International Journal of Computer Vision (IJCV)*, 128(5):1118–1140, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2234–2242, 2016.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7091–7101, 2018.

Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations (ICLR)*, 2018.

Ruoyu Sun, Tiantian Fang, and Alexander Schwing. Towards a better global loss landscape of gans. *Advances in Neural Information Processing Systems*, 33, 2020.

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably optimal nash equilibria via potential fields. In *International Conference on Learning Representations (ICLR)*, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 7354–7363, 2019.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *International Conference on Learning Representations (ICLR)*, 2017.

Peilin Zhong, Yuchen Mo, Chang Xiao, Pengyu Chen, and Changxi Zheng. Rethinking generative mode coverage: A pointwise guaranteed approach. *Advances in Neural Information Processing Systems*, 32:2088–2099, 2019.

Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Weinan Zhang, Jun Wang, and Yong Yu. Activation maximization generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhiming Zhou, Jiadong Liang, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Yong Yu, and Zhihua Zhang. Lipschitz generative adversarial nets. In *International Conference on Machine Learning (ICML)*, pages 7584–7593, 2019.