Lipschitz Regularized Generative Adversarial Nets

A Systematic Study on the Convergence of Generative Adversarial Nets

Zhiming Zhou HEYOHAIZHOU@GMAIL.COM
Department of Computer Science, School of Information Management and Engineering
Shanghai University of Finance and Economics
No.100 Wudong Road, Yangpu District, Shanghai, China

Zhihua Zhang ZHZHANG@MATH.PKU.EDU.CN
Department of Probability and Statistics, School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing 100871, China

Abstract

In this paper, we show that Generative Adversarial Networks (GANs) without regularization in the discriminative function space commonly suffer from a problem that the gradients from the discriminator can be uninformative to guide the generator. We find that this gradient uninformativeness issue is nontrivial, not any simple regularization in the discriminative function space can resolve the gradient uninformativeness issue.

However, interestingly, Lipschitz regularization in the discriminative function space can generally resolve the gradient uninformativeness issue and guarantee the GANs's convergence. We provide a sufficient condition for the construction of these GANs, which turns out to cover most popular choices of GANs objective and hence explain how Lipschitz regularization works when combined with these GANs.

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

The above leads to the Lipschitz regularized GANs, a GANs family. We construct several instances of this GANs family and show their consistent superior performance over WGANs.

In trying to attain the optimal discriminative function and then verifying the theoretical properties of LGANs, we realize various underlying issues in the current implementations of Lipschitz regularization and hence propose improved versions with theoretical guarantee.

Then, to make the understanding of the convergence issues of GANs even more comprehensive, we provide a systematic study of the gradient issues of traditional GANs and how these issues behave in practice, and thereby learn how traditional GANs work in practice.

Finally, we study the gradient flow between the generator and discriminator in various GANs with the help of envelope theorem, trying to figure out the real key to the convergence of GANs and realize that it is worth noticing GANs is essentially a sample-based framework.

Keywords: Lipschitz, Regularization, GANs, Training, Instability, Convergence, Issue.

1. Introduction

Generative Adversarial Networks (GANs, Goodfellow et al. (2014)), as one of the most successful generative models, have shown promising results in various challenging tasks. GANs is popular and widely used, but it is notoriously hard to train (Goodfellow, 2016). The underlying obstacles, though have been heavily studied (Arjovsky and Bottou, 2017; Lucic et al., 2017; Heusel et al., 2017; Mescheder et al., 2017, 2018; Yadav et al., 2017), are still not fully understood.

The objective of GANs is usually defined as a distance metric between the target distribution, or in GANs more commonly named as the real distribution, $\mathcal{P}_r$ and the distribution $\mathcal{P}_g$ formed by generated samples, which implies that $\mathcal{P}_r = \mathcal{P}_g$ is the unique global optimum. The training instability issues of traditional GANs has been considered stems from the illness of the distance metric (Arjovsky and Bottou, 2017), e.g., the distance between $\mathcal{P}_r$ and $\mathcal{P}_g$ keeps constant when their supports are disjoint. Arjovsky et al. (2017) accordingly suggested using the Wasserstein distance as the distance metric, which can properly measure the distance between two distributions no matter whether their supports are disjoint.

In this paper, we propose to study the convergence of GANs from the perspective of the optimal discriminative function f^*. By inspecting f^* and its gradient with respect to samples, the understanding of GANs's training can be much more clear. The reason is that we now take the G-D (i.e., generator and discriminator) structure of GANs into account, and we are inspecting the samples (instead of the densities or probabilities) and the gradients that the generator receives from the discriminator (i.e., from f^*) with respect to the samples to be updated. That is to inspect the connecting point of the generator and the discriminator. In this sense, GANs works as follows. G models the samples to be updated and updates them accordingly, while D or f^* tells the generator how these samples should be updated via its gradient with respect to these samples.

We demonstrate that the 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 find that if there is no regularization in the discriminative function space, the GANs generally does not guarantee its convergence, provably suffering a gradient uninformativeness issue, which means the gradient that the generator receives from the discriminator does not tell any information of the real distribution. We also find that this gradient uninformativeness issue is nontrivial, not any single simple regularization in the discriminative function space can resolve the gradient uninformativeness issue.

However, interestingly, we realize that Lipschitz regularization can generally resolve the gradient uninformativeness issue and guarantee the GANs's convergence. We provide a sufficient condition for the construction of these GANs, which we believe is very close to the necessary condition for GANs where the discriminator has a single input sample. It turns out to cover most popular choices of GANs objective and hence explain how Lipschitz regularization works when combined with these GANs.

We provide detailed analysis upon why Lipschitz regularization can generally resolve the gradient uninformativeness issue and show 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. We prove the existence and uniqueness of the optimal discriminative function for GANs under Lipschitz regularization. We prove that there is only a single Nash equilibrium between the generator and the optimal discriminative function, and show that otherwise, the GANs will always move samples from locations where has too much to locations where has too less.

The above leads to the Lipschitz regularized GANs, a GANs family. We construct several instances of this GANs family, and empirically verify their theoretical properties, and then show their consistent superior performance over WGANs.

In trying to attain the optimal discriminative function and hence we can verify its theoretical properties, we realize various underlying issues in the current implementations of Lipschitz regularization. We hence provide a serious study on the implementation of Lipschitz regularization, and thereby propose two advanced versions with theoretical justification.

Then, to gain a deeper understanding of the convergence issues of GANs, we feel it is required to understand how traditional GANs, especially these unregularized ones, work in practice. We hence provide a systematic study of the gradient issues of traditional GANs and investigate how these gradient issues behave in practice, with which we learn how these theoretically not guaranteed GANs work in practice.

In a nutshell, tuning supplies the lack of theory, and common practices (or tricks) leads to smooth discriminative function or avoids the fatal optimal discriminative function, which favors the training and make the training more likely to success, but still being unstable, sensitive to hyper-parameters and network architecture, and hard to use.

Finally, we study the gradient flow between the generator and discriminator in various GANs, with the help of the envelope theorem, a classic result about the differentiation properties of an optimization problem, to see what is the indispensable element for convergence guarantee.

With the study of unregularized GANs and the compact dual form of Wasserstein distance under envelope theorem, we realize that GANs is essentially a sample-based framework, and the information interchange between the generator and discriminator must flow through gradient of the optimal discriminative function with respect to the samples.

And given the previous analysis of issues of traditional GANs and the effect of Lipschitz regularization in GANs, we believe that f-divergence or these distance metric induced by unregularized GANs is not a favorable distance metric for GANs (given its sample-based nature), and optimal transport based distance metric or these distance metric induced by Lipschitz regularization is more compatible with the current GANs framework.

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 study the gradient uninformativeness issue in detail. In Section 4, we present Lipschitz regularized GANs and their theoretical analysis. In Section 5, We study the implementation of Lipschitz regularization. In Section 6, we provide the empirical analysis of Lipschitz regularized GANs. In Section 7, we study the gradient issues of traditional GANs and learn how they work in practice. In Section 8, we study the essence of convergence via the lens of envelope theorem. Finally, we discuss related work in Section 9 and conclude the paper in Section 10.

2. Preliminaries

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

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-)Lipschitz continuous, if there exists a constant $k \geq 0$ such that

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

In this paper and most existing GANs papers, like the WGANs series, the metrics d_X and d_Y are by default Euclidean distance or norm[1], which we denote by $\|\cdot\|$.

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

2.2 Wasserstein Distance and its Dual Forms, and an Important Proposition

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

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

where $\Pi(\mathcal{P}_g, \mathcal{P}_r)$ denotes the set of all probability measures with marginals $\mathcal{P}_g$ and $\mathcal{P}_r$. It can be interpreted as the minimum cost of transporting the distribution $\mathcal{P}_g$ to the distribution $\mathcal{P}_r$. We use π^* to denote the optimal transport plan, and let $\mathcal{S}_g$ and $\mathcal{S}_r$ denote the supports of $\mathcal{P}_g$ and $\mathcal{P}_r$, respectively.

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

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

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

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

$$\begin{aligned} W_1(\mathcal{P}_g, \mathcal{P}_r) = \sup_f\, & \mathbb{E}_{x\sim\mathcal{P}_g}\left[f(x)\right] - \mathbb{E}_{x\sim\mathcal{P}_r}\left[f(x)\right], \\ & s.t.\, f(x) - f(y) \leq d(x, y),\ \forall x \in \mathcal{S}_g, \forall y \in \mathcal{S}_r. \end{aligned} \tag{4}$$

This new dual form relaxes the Lipschitz continuity condition from over the entire space to over their supports, respectively. The proof for this dual form is given in Appendix A.5.

As shown by 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:

1. When switching to other norms, the property of the gradients will get changed. Different norms will induce different gradients with different properties. See Appendix A.4 for some basic arguments on this.

Table 1: The comparison of different objectives of GANs.

	ϕ	φ	$\mathcal{F}$	$f^*(x)$
Original GANs	$-\log(\sigma(-x))$	$-\log(\sigma(x))$	Free	$\log \frac{\mathcal{P}_r(x)}{\mathcal{P}_g(x)}$
Least-Squares GANs	$(x-\alpha)^2$	$(x-\beta)^2$	Free	$\frac{\alpha\cdot\mathcal{P}_g(x)+\beta\cdot\mathcal{P}_r(x)}{\mathcal{P}_r(x)+\mathcal{P}_g(x)}$
μ-Fisher GANs	x	$-x$	$\mathbb{E}_{x\sim\mu}\lvert f(x)\rvert^2 \leq 1$	$\frac{\mathcal{P}_r(x)-\mathcal{P}_g(x)}{\mathcal{F}_\mu(\mathcal{P}_r,\mathcal{P}_g)\cdot\mu(x)}$
Wasserstein GANs	x	$-x$	$k(f)\leq 1$	*N/A*
Lipschitz Regularized GANs	any ϕ and φ satisfying Eq. (11)		$k(f)$ regularized	*N/A*

Proposition 1. *Let π^* be the optimal transport plan in Eq. (2) and $x_t = tx + (1-t)y$ 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*

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

This proposition indicates: (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 also means the gradient at any interpolation x_t is pointing towards the real sample y; (i) these (x,y) pairs match the optimal coupling π^*, i.e., the direction of $\nabla_{x_t} f^*(x_t)$ indicates the optimal transport; (iii) it also implies that WGANs does not suffer from the gradient vanishing.

2.3 Generalized Formulation of Generative Adversarial Nets

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

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

where $\mathcal{P}_z$ is the source distribution of the generator in $\mathbb{R}^m$ and $\mathcal{P}_r$ is the real (target) 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 as samples in $\mathcal{P}_r$, while the discriminative function $f\colon \mathbb{R}^n \to \mathbb{R}$ learns to output a score indicating the authenticity of a given sample.

Here $\mathcal{F}$ and $\mathcal{G}$ denote discriminative and generative function spaces, respectively. And ϕ, φ, $\psi\colon \mathbb{R}\to\mathbb{R}$ are loss metrics. We denote the implicit distribution of the generated samples by $\mathcal{P}_g$. We use f^* to denote the optimal discriminative function, i.e., $f^* \triangleq \arg\min_{f\in\mathcal{F}} J_D$. For subsequent needs, we let $\mathring{J}_D(x) \triangleq \mathcal{P}_g(x)\phi(f(x)) + \mathcal{P}_r(x)\varphi(f(x))$. It has $J_D = \int \mathring{J}_D(x)dx$.

We list the choices of ϕ, φ and $\mathcal{F}$ of some representative GANs in Table 1. Without loss of generality, the basic common pattern of these GANs is that the discriminator forces $f(x)$ to be small for generated samples, while forces $f(x)$ to be large for real samples. Different choices of ϕ, φ, and $\mathcal{F}$, if valid, lead to different distance metrics and hence, in a sense, different GANs. 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.

2.4 The Gradient that the Generator Receives and Gradient Vanish

In these GANs, the gradient that the generator receives from the discriminator with respect to a generated sample $x \in \mathcal{S}_g$ is

$$\nabla_x J_G(x) \triangleq \nabla_x \psi(f(x)) = \nabla_{f(x)} \psi(f(x)) \cdot \nabla_x f(x), \tag{7}$$

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.

The gradient vanishing problem has been considered as a key phenomenon that indicates the existence of training issues in GANs. For original GANs, when the discriminator is trained to optimum, $\nabla_{f(x)} \psi(f(x))$ becomes zero. Goodfellow et al. (2014) addressed this problem by using an alternative objective for the generator. Actually, only the scalar $\nabla_{f(x)} \psi(f(x))$ is changed. The Least-Squares GANs (Mao et al., 2016), which aims at addressing the gradient vanishing problem, also focused on $\nabla_{f(x)} \psi(f(x))$. And we can actually show that Least-Squares GANs may still suffer from gradient vanishing, due to the zeroness of $\nabla_x f(x)$.

Arjovsky and Bottou (2017) provided a new perspective for understanding the gradient vanishing. They argued that $\mathcal{S}_g$ and $\mathcal{S}_r$ are usually disjoint and the gradient vanishing stems from the illness of traditional distance metrics, i.e., the distance between $\mathcal{P}_g$ and $\mathcal{P}_r$ remains constant when they are disjoint. The Wasserstein distance was thus proposed by Arjovsky et al. (2017) as an alternative distance metric, 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 turns out 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 $\mathcal{P}_r$. Consequently, $\mathcal{P}_g$ is not guaranteed to converge to $\mathcal{P}_r$. We name this phenomenon as the *gradient uninformativeness* and argue that it is a fundamental cause of nonconvergence and instability in the training of traditional GANs[2].

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 $\nabla_x J_G(x)$, while the gradient uninformativeness is about the direction of $\nabla_x J_G(x)$, 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 regularization in the discriminative function space $\mathcal{F}$. We will show that: (i) for unregularized GANs, 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.

2. By traditional GANs, we refer to original GANs, Least-Squares GANs and so on. We will later give a precise definition of traditional GANs in Section 7, where we study how these GANs work in practice.

3.1 Unregularized GANs: Gradient Uninformativeness Commonly Exists

For many GANs, there is no regularization in the discriminative function space $\mathcal{F}$. Typical instances include f-divergence based GANs, such as the original GANs (Goodfellow et al., 2014), Least-Squares GANs (Mao et al., 2016).

In these GANs, the value of the optimal discriminative function at each point, i.e., $f^*(x)$, is independent of other points and only reflects the local densities $\mathcal{P}_g(x)$ and $\mathcal{P}_r(x)$:

$$f^*(x) = \arg\min_{f(x)\in\mathbb{R}} \ \mathcal{P}_g(x)\phi(f(x)) + \mathcal{P}_r(x)\varphi(f(x)), \ \forall x. \tag{8}$$

Given that $f^*(x)$ only reflects the local densities $\mathcal{P}_g(x)$ and $\mathcal{P}_r(x)$, for each generated sample x, which is not surrounded by real samples (there exists $\epsilon > 0$ such that for all y with $0 < \|y - x\| < \epsilon$, it holds that $y \notin \mathcal{S}_r$), $f^*(x)$ in the surrounding of x would contain no information about $\mathcal{P}_r$.

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

Typical situation is that $\mathcal{S}_g$ and $\mathcal{S}_r$ are disjoint, which is common in practice according to Arjovsky and Bottou (2017). That is, gradient uninformativeness commonly exists in unregularized GANs.

To further distinguish the gradient uninformativeness from the gradient vanishing, we consider an ideal case: $\mathcal{S}_g$ and $\mathcal{S}_r$ are totally overlapped and both consist of n discrete points, but their probability masses over these points are different. Check Eq. (8) in this case and you will find 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 being undefined[3].

3.2 Regularized GANs: Gradient Uninformativeness May Still Exist

There also exists some GANs that impose regularization in the discriminative function space $\mathcal{F}$. Typical instances are the *integral probability metric (IPM)* based GANs (Mroueh and Sercu, 2017; Mroueh et al., 2017; 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 μ-Fisher IPM $\mathcal{F}_\mu(\mathcal{P}_g, \mathcal{P}_r)$, i.e., the generalized objective of the Fisher GANs (Mroueh et al., 2017), has the following form:

$$f^*(x) = \frac{1}{\mathcal{F}_\mu(\mathcal{P}_g, \mathcal{P}_r)} \frac{\mathcal{P}_g(x) - \mathcal{P}_r(x)}{\mu(x)}, \tag{9}$$

where μ is a distribution whose support covers $\mathcal{S}_g$ and $\mathcal{S}_r$. $\frac{1}{\mathcal{F}_\mu(\mathcal{P}_g,\mathcal{P}_r)}$ is a constant. It can be observed that μ-Fisher IPM also defines $f^*(x)$ at each point according to the local densities and does not reflect information of other locations. Similar as above, we can conclude that for each generated sample that is not surrounded by real samples, $\nabla_x f^*(x)$ is uninformative.

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

4. Lipschitz Regularized GANs

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

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

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

$$\begin{aligned}&\min_{f\in\mathcal{F}} \mathbb{E}_{z\sim\mathcal{P}_z}[\phi(f(g(z)))] + \mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(f(x))] + \frac{\rho}{2}\cdot k(f)^2,\\ &\min_{g\in\mathcal{G}} \mathbb{E}_{z\sim\mathcal{P}_z}[\psi(f(g(z)))].\end{aligned} \tag{10}$$

We will show that, if $\rho > 0$ and once the following condition holds, the above defined LGANs can generally resolve the gradient uninformativeness issue and guarantee the convergence.

$$\begin{cases}\phi'(x) > 0, \phi''(x) \geq 0,\\ \varphi'(x) < 0, \varphi''(x) \geq 0,\\ \exists\, a,\ \phi'(a) + \varphi'(a) = 0.\end{cases} \tag{11}$$

This condition for the loss metrics ϕ and φ is a sufficient condition for desired properties, and it is actually very mild and should be very close to the necessary condition.

Requiring ϕ to be increasing means that the discriminator will need to force small $f(x)$ for generated samples. Requiring φ to be decreasing, meaning that the discriminator will need to force large $f(x)$ for real samples. The other constraints are included because, otherwise, this problem is not guaranteed to have a solution.

To devise a loss in LGANs, it is practical to let ϕ be an increasing function with non-decreasing derivatives, and then simply set $\phi(x) = \varphi(-x)$ would be sufficient.

Note that in WGANs, loss metrics $\phi(x) = \varphi(-x) = x$ is used, which satisfies Eq. (11). There are many other instances satisfy Eq. (11), such as $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$, $\phi(x) = \varphi(-x) = x + \sqrt{x^2+1}$ and $\phi(x) = \varphi(-x) = \exp(x)$.

Note that rescaling and offsetting along the axes are trivial operation to found more ϕ and φ within a function class, and linear combination of two or more ϕ or φ from different function classes also keep satisfying Eq. (11). We illustrate some of these loss metrics in Figure 1.

Meanwhile, there also exists loss metrics used in GANs that do not satisfy Eq. (11), e.g., the quadratic loss (Mao et al., 2016) and the hinge loss (Zhao et al., 2016; Lim and Ye, 2017; Miyato et al., 2018). Nevertheless, we will study them in experiments (See Section 6).

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

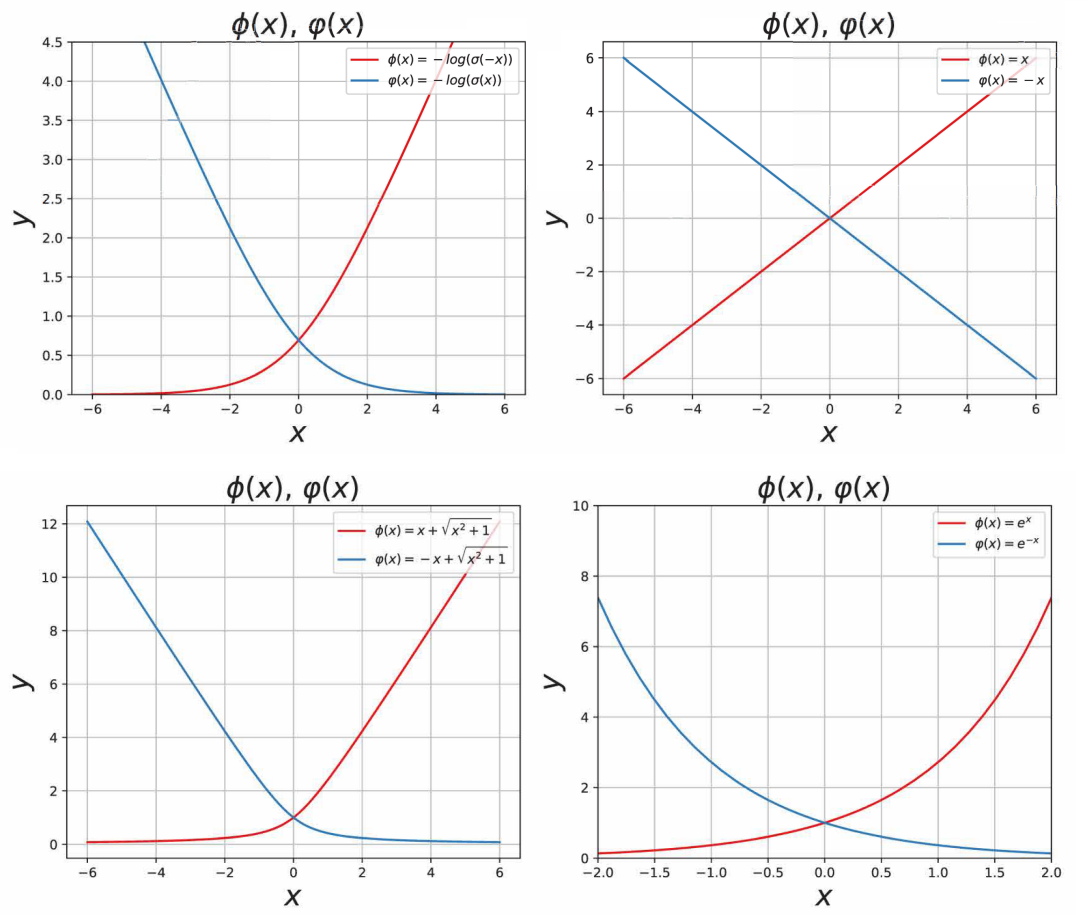


Figure 1: Various ϕ and φ that satisfies Eq. (11).

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 understand how the gradient uninformativeness issue is resolved. All proofs are deferred to Appendix A.

For the first fold, we will prove the existence and uniqueness of the optimal discriminative function for GANs under Lipschitz regularization. And then we will prove that there is only a single Nash equilibrium between the generator and the optimal discriminative function, and show that, otherwise, the GANs will always move samples from locations where has too much to locations where has too less. With all these as a whole, we believe LGANs is a valid GANs formulation.

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 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 extremely 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 1. *Under Assumption* (11)*, the optimal discriminative function f^* of Eq.* (10) *exists. And, if ϕ or φ is strictly convex, it is unique.*

Note that LGANs with the WGANs loss metrics, i.e., $\phi(x) = \varphi(-x) = x$, which does not satisfy the condition that ϕ or φ is strictly convex, the solution of Eq. (10), i.e., the optimal discriminative function f^*, still exists but is not unique. Specifically, if f^* is an optimal solution, then $f^* + \alpha$ for any $\alpha \in \mathbb{R}$ is also an optimal solution. And this is actually the only special case. For all other instances that satisfy the condition, 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-GP to LGANs, with more detailed analysis of bounding relationships and equilibrium.

Theorem 2. *Under Assumption* (11)*, we have the optimal discriminative function f^* exists. If we further assume f^* is smooth, we have:*

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

(b) For all $x \in \mathcal{S}_g \cup \mathcal{S}_r - \mathcal{S}_g \cap \mathcal{S}_r$, there exists $y \in \mathcal{S}_g \cup \mathcal{S}_r$ with $y \neq x$ such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$;

(c) If $\mathcal{S}_g = \mathcal{S}_r$ and $\mathcal{P}_g \neq \mathcal{P}_r$, then there exists (x, y) pair in $\mathcal{S}_g \cup \mathcal{S}_r$ with $y \neq x$ such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$ and $\nabla_{f^*(x)} \mathring{J}_D(x) \neq 0$;

(d) There is a unique Nash equilibrium between $\mathcal{P}_g$ and f^*, where $\mathcal{P}_g = \mathcal{P}_r$ and $k(f^*) = 0$.

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

It is worth noticing that the penalizing $k(f)$, instead of simply restricting the maximum of $k(f)$ as in WGANs, is, in fact, necessary for Property-(c) and Property-(d). It is due to the existence of cases, where $\nabla_{f^*(x)} \mathring{J}_D(x) = 0$ for $\mathcal{P}_g(x) \neq \mathcal{P}_r(x)$, when the loss metrics are not $\phi(x) = \varphi(-x) = x$, i.e., when the loss metric is strictly convex.

Minimizing $k(f)$, in any case, guarantees that the only Nash equilibrium is achieved when $\mathcal{P}_g = \mathcal{P}_r$. With the WGANs loss metrics, minimizing $k(f)$ is not necessary. However, if $k(f)$ is not minimized towards zero, $\nabla_x f^*(x)$ is not guaranteed to be zero at the convergence state $\mathcal{P}_g = \mathcal{P}_r$ (Mescheder et al., 2018). This implies that minimizing $k(f)$ also benefits WGANs.

4.1.3 The Refined Properties of Bounding Relationship

From Theorem 2, we know, as long as $\mathcal{P}_g$ still has not fully converged to $\mathcal{P}_r$, there must exist point x with $f^*(x)$ being bounded by another point y, such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$.

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 high values for real samples and low values for generated samples. In a bounding relationship under fully optimized discriminative function, the values of real samples should always be higher. More formally, we have:

Theorem 3. *Under the conditions in Theorem 2, we have*

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

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

It is worth noticing that, if it is the disjoint case, all fake samples hold $\nabla_{f^*(x)} \mathring{J}_D(x) > 0$, while all real samples hold $\nabla_{f^*(y)} \mathring{J}_D(y) < 0$. And for a generated sample $x \in \mathcal{S}_g$, $\nabla_{f^*(x)} \mathring{J}_D(x) > 0$ means $f^*(x)$ can assign a lower value to x so as to better optimize the objective, if it is not bounded by another sample. See lemmas in the proof for more details.

Note that there might exist a chain of bounding relationships that involves a dozen of fake samples and real samples. 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 to each other. The bounding line in the value surface of f^* is the basic building block that connects $\mathcal{P}_g$ and $\mathcal{P}_r$, and each fake sample with $\nabla_{f^*(x)} \mathring{J}_D(x) \neq 0$ lies in at least one of these bounded lines.

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 bounding relationship between x and y, i.e., 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 4. *Assume function f is differentiable and its Lipschitz constant is k, then for all x and y, which satisfy $y \neq x$ and $f(y) - f(x) = k \cdot \|y - x\|$, we have $\nabla_{x_t} f(x_t) = k \cdot \frac{y-x}{\|y-x\|}$ for all $x_t = tx + (1-t)y$ with $0 \leq t \leq 1$.*

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

With a deep inspection of the proof of Proposition 1 in Gulrajani et al. (2017) and Theorems 4, we believe the differentiable requirement 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.

4.1.5 Summing Up

Combining Theorems 1, 2, 3 and 4, we can conclude that, as long as $\rho > 0$ and the loss metrics ϕ and φ satisfy the condition Eq. (11) and f^* is smooth and differentiable:

- According to Property-(a), when $\mathcal{S}_g$ and $\mathcal{S}_r$ are disjoint, the gradient of the optimal (or practically well-trained) discriminative function for each generated sample $x \in \mathcal{S}_g$, i.e., $\nabla_x f^*(x)$, points towards some real sample $y \in \mathcal{S}_r$, which guarantees that $\nabla_x f^*(x)$-based generator update would tend to move $\mathcal{P}_g$ towards $\mathcal{P}_r$ at every step.
- In fact, Theorem 2 provides further guarantee on the convergence. Property-(b) implies that, for any generated sample $x \in \mathcal{S}_g$ that does not lie in $\mathcal{S}_r$, its gradient under optimal discriminative function, i.e., $\nabla_x f^*(x)$, must point towards some real sample $y \in \mathcal{S}_r$.
- And in the fully overlapped case, according to Property-(c), unless $\mathcal{P}_g = \mathcal{P}_r$, there must exist at least one pair of (x, y) in strict bounding relationship and $\nabla_x f^*(x)$ pulls x towards y, and we can actually also claim that it is moving sample from location where it has too much to where it has too less. See lemmas in the proofs for more details.
- Finally, Property-(d) guarantees that the only Nash equilibrium between the generator and discriminator is reached when $\mathcal{P}_g = \mathcal{P}_r$, and it holds that $k(f) = 0$ and hence $\nabla_x f^*(x) = 0$ for all generated samples, which means the training will fully stop.

As we have already seen, the original GANs suffers from the gradient vanishing and the gradient uninformativeness. However, when imposing the Lipschitz regularization in the discriminative function space, the resulting model, as a special case of LGANs, behaves very well. In the experiments, we will see that it even outperforms WGANs empirically. This further shows the powerfulness of Lipschitz regularization in discriminative function space.

5. Max Gradient Norm Penalty and Augmented Lagrangian

WGANs (Arjovsky et al., 2017) first introduces the requirement for Lipschitz regularization in GANs. After that, researchers (Kodali et al., 2017; Fedus et al., 2017; Miyato et al., 2018) also empirically found that Lipschitz regularization is also useful, when combined with other GANs objectives, e.g., the original GANs objective.

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

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

5.1 Existing Lipschitz Regularization Implementations

Quite a few recent 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 parameters (also named weights) in the neural network. However, it was later shown to lead to suboptimal solutions (Gulrajani et al., 2017; Petzka et al., 2017).

And corresponding alternative methods were thus proposed for imposing the Lipschitz regularization, named **gradient penalty** (GP) and **Lipschitz penalty** (LP), respectively. The two methods share the same spirit and achieve Lipschitz continuity via penalizing the sample gradients towards a given target value. The target value is usually 1, however, not necessary (Karras et al., 2017; 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 introduce the following regularization terms, respectively:

$$L_{gp} = \frac{\rho}{2} \cdot \mathbb{E}_{x \sim \mathcal{P}_{\hat{x}}}[(\|\nabla_x f(x)\| - k_0)^2], \tag{12}$$

$$L_{lp} = \frac{\rho}{2} \cdot \mathbb{E}_{x \sim \mathcal{P}_{\hat{x}}}[(\max\{0, \|\nabla_x f(x)\| - k_0\})^2], \tag{13}$$

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

Petzka et al. (2017) 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. It is also the reason why they proposed 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.

Given the singular value of a weight matrix is (easily) attainable, they proposed to divide the weight of each linear layer of a neural network by its maximum singular value, i.e.,

$$\bar{W}_{SN} = W/\sigma(W), \tag{14}$$

where $\sigma(W)$ denotes the maximum singular value of W. As a result, the Lipschitz constant of every linear layer is fixed as 1. Then if the nonlinearity parts, i.e., 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.

It is worth noting that the spectral normalization results in a *hard global* restriction on the Lipschitz constant, while gradient penalty and Lipschitz penalty are *soft local* regularizations.

5.2 Analysis on Lipschitz Regularization Implementations and Motivations

Before moving into the detailed discussion of these methods, we would like to provide several important notes in the first place.

5.2.1 Local Lipschitz Regularization is Sufficient

The most common choice of $\mathcal{P}_{\hat{x}}$ in gradient penalty and 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 provide arguments to reasonably extend it to LGANs.

To get such 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 $\mathcal{S}_g$ and $\mathcal{S}_r$, respectively.

It is worthy noticing that given the constraints in the compact dual form, any other constraints in the KR duality does not affect the final result of $W_1(\mathcal{P}_g, \mathcal{P}_r)$. 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 $\mathcal{S}_g$ and $\mathcal{S}_r$ unchanged. Thus, any f^* in the compact dual form Eq. (4) also holds the following key property of Wasserstein distance (Villani, 2008):

Theorem 5. *Let π^* be the optimal transport plan in the primal form of Wasserstein distance Eq.* (2) *and f^* be the optimal discriminative function in the compact dual form Eq.* (4)*. It holds that*

$$P_{(x,y)\sim\pi^*}\left[f^*(x) - f^*(y) = d(x,y)\right] = 1. \tag{15}$$

Note that the Proposition 1 is based on Eq. (15) and the 1-Lipschitz continuity of f^*.

Let $S_{\hat{x}} \triangleq \{\hat{x} = x \cdot t + y \cdot (1-t) \mid x \in \mathcal{S}_g,\, y \in \mathcal{S}_r,\, t \in [0,1]\}$ denotes the support of the linear interpolations between the real and fake distributions. And we can further notice that f^* being local Lipschitz continuity over $S_{\hat{x}}$ is sufficient for proving Proposition 1.

Formally, we have:

Theorem 6. *Enforcing the local Lipschitz regularization over $S_{\hat{x}}$ is sufficient to maintain the property of Proposition 1 for Wasserstein distance.*

It means the Wasserstein distance and its desired gradient property for GANs keep unchanged, when you drop constraints outside the blending region $S_{\hat{x}}$. Because the blending region has covered the necessity and very parts.

One can imagine the similar holds for Lipschitz regularized GANs. And with our analysis upon how Lipschitz regularization works, we believe these extra constraints, i.e., these outside $S_{\hat{x}}$, are also unnecessary and are indifferent to the forming of bounding relationships.

Note that Theorem 6 also indicates that, for training GANs, restricting the global Lipschitz constant, e.g., spectral normalization, might be unnecessarily too strong. And henceforward, by Lipschitz constant or Lipschitz continuity, we mostly mean that over the local region $S_{\hat{x}}$.

5.2.2 Superfluous Constraints in Current Local Lipschitz Implementations

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

Gradient penalty and Lipschitz penalty impose the Lipschitz continuity via penalty method. Penalty method is a soft regularization, where the constraint is usually slightly drifted.

And 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, i.e., ρ. Likewise, we use the Wasserstein distance as a demonstration.

Let $W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) \triangleq \sup_{k(f)\leq\tilde{k}} \mathbb{E}_{x\sim\mathcal{P}_g}[f(x)] - \mathbb{E}_{x\sim\mathcal{P}_r}[f(x)]$. It holds that $W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) = \tilde{k} \cdot W_1(\mathcal{P}_g, \mathcal{P}_r)$. Because $W_1(\mathcal{P}_g, \mathcal{P}_r) = \sup_{k(f)\leq\tilde{k}} \mathbb{E}_{x\sim\mathcal{P}_g}[f(x)/\tilde{k}] - \mathbb{E}_{x\sim\mathcal{P}_r}[f(x)/\tilde{k}]$.

Assume we can directly optimize the Lipschitz constant $\tilde{k}$ and consider the following objective:

$$J_{gp}(\tilde{k}) = -W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) + \frac{\rho}{2} \cdot (\tilde{k} - k_0)^2. \tag{16}$$

Given that $\mathcal{P}_g$ and $\mathcal{P}_r$ is fixed, $W_1(\mathcal{P}_g, \mathcal{P}_r)$ is a constant and we denote it as α. Then, $J_{gp}(\tilde{k})$ is quadratic function $-\alpha \cdot \tilde{k} + \frac{\rho}{2} \cdot (\tilde{k} - k_0)^2$, whose optimum is reached when $\tilde{k}^* = \frac{\alpha}{\rho} + k_0$.

Note that replacing $(\tilde{k} - k_0)^2$ with $\max\{0, \tilde{k} - k_0\}^2$, i.e., analogizing switching from gradient penalty to Lipschitz penalty, will result in the same optimal $\tilde{k}^*$.

From the above, we can see that[4], when ρ is small or the distance between $\mathcal{P}_g$ and $\mathcal{P}_r$ is large (i.e., if α is large), the resulting Lipschitz constant can be much larger than k_0.

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

4. From another perspective, as α goes to zero, i.e., as $\mathcal{P}_g$ converges to $\mathcal{P}_r$, the optimal Lipschitz constant $\tilde{k}^*$ decreases. And finally, when $\mathcal{P}_g = \mathcal{P}_r$, we have $\alpha = 0$ and the optimal Lipschitz constant $\tilde{k}^* = k_0$.

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. (2017) 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., 2017). That is, their results are not contradictory to our results here.

5.3 The Proposed Lipschitz Regularization Implementations

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

5.3.1 Max Gradient Norm Regularization with Penalty Method

Given that the local 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 maximum gradient norm in $S_{\hat{x}}$, which is equivalent to the Lipschitz constant in the local region of $S_{\hat{x}}$.

The resulting regularization is as follows:

$$L_{maxgp} = \frac{\rho}{2}([\max_{x\sim\mathcal{P}_{\hat{x}}} \left\|\nabla_x f(x)\right\|] - k_0)^2. \tag{17}$$

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

Practically, we follow Gulrajani et al. (2017) and sample x as random linear interpolations of the real and fake samples in parallel mini-batches. We can either directly use the maximum gradient norm sampled in a mini-batch, or further keep track x with the maximum $\left\|\nabla_x f(x)\right\|$, to improve the stability and reduce the bias introduced via batch sampling.

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

We have studied these two in experiments. According to our experiments, the historical buffer is usually unnecessary and directly using the maximum 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 maximum gradient norm or Lipschitz constant.

We suspect that, when the training goes smoothly, the surface of f is also smooth, in the sense that the Lipschitz constant in different regions are similar. Hence, a mini-batch estimation could be accurate enough for successful training.

5.3.2 Max Gradient Norm Regularization with Augmented Lagrangian

With the penalty method, the constraint is usually not strictly satisfied. The resulting Lipschitz constant, as discussed around Eq. (16), is floating / 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. Because Lipschitz constant may have a huge impact on the performance.

Also, if one would like to strictly evaluate the Wasserstein distance, a strict restriction of the Lipschitz constant being one 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 including an extra Lagrange multiplier term.

The regularization term(s) derived from the augmented Lagrangian can be written as follows:

$$L_{maxal} = \lambda_{al} \cdot (\max_{x \sim \mathcal{P}_{\hat{x}}} \left\| \nabla_x f(x) \right\| - k_0) + \frac{\rho}{2} \cdot (\max_{x \sim \mathcal{P}_{\hat{x}}} \left\| \nabla_x f(x) \right\| - k_0)^2, \tag{18}$$

where λ_{al} is the Lagrange multiplier.

Given that the augmented Lagrangian is a simple extension, and there exists potential benefits, we also investigate the practical performance of augmented Lagrangian in imposing Lipschitz continuity regularization.

5.3.3 First Order Optimality Analysis: Part I, MaxAL Properties

Some interesting properties of the augmented Lagrangian method can be easily derived with its first order optimality analysis. For clarity, we denote $\max_{x \sim \mathcal{P}_{\hat{x}}} \left\| \nabla_x f(x) \right\|$ as $\tilde{k}$. Still, we use the Wasserstein distance for simplicity and demonstration. Let the overall objective be:

$$J_{al}(\tilde{k}) = -W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) + \lambda_{al} \cdot (\tilde{k} - k_0) + \frac{\rho}{2} \cdot (\tilde{k} - k_0)^2. \tag{19}$$

Similar as previous, because $W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) = \tilde{k} \cdot W_1(\mathcal{P}_g, \mathcal{P}_r)$ and $W_1(\mathcal{P}_g, \mathcal{P}_r)$ is a constant, we denote $W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k})$ as $\alpha \cdot \tilde{k}$. Then, what is optimizing is

$$J_{al}(\tilde{k}) = -\alpha \cdot \tilde{k} + \lambda_{al} \cdot (\tilde{k} - k_0) + \frac{\rho}{2} \cdot (\tilde{k} - k_0)^2. \tag{20}$$

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

$$\begin{aligned} \frac{\partial J_{al}}{\partial \lambda_{al}} &= \tilde{k} - k_0 = 0, \\ \frac{\partial J_{al}}{\partial \tilde{k}} &= -\alpha + \lambda_{al} + \rho \cdot (\tilde{k} - k_0) = 0. \end{aligned} \tag{21}$$

Thereby, when the augmented Lagrangian converged, $\tilde{k} = k_0$ and $\lambda_{al} = W_1(\mathcal{P}_g, \mathcal{P}_r)$.

Classic results of augmented Lagrangian also involve the choice of ρ, which is out of the scope of the discussion of this paper and hence is not included.

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

Although the following is the traditional result in optimization, we present it here for ease of reference and 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 into the original optimization problem, i.e.,

$$L_{maxlm} = \lambda_{lm} \cdot (\max_{x \sim \mathcal{P}_{\tilde{x}}} \left\| \nabla_x f(x) \right\| - k_0), \tag{22}$$

where λ is the Lagrange multiplier.

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

Considering the optimization problem of

$$J_{lm}(\tilde{k}) = -W_1(\mathcal{P}_g, \mathcal{P}_r, \tilde{k}) + \lambda_{lm} \cdot (\tilde{k} - k_0). \tag{23}$$

The first order optimality condition of the Lagrangian method can be written down as:

$$\begin{aligned} \frac{\partial J_{lm}}{\partial \lambda_{lm}} &= \tilde{k} - k_0 = 0, \\ \frac{\partial J_{lm}}{\partial \tilde{k}} &= -\alpha + \lambda_{lm} = 0. \end{aligned} \tag{24}$$

That is, when the Lagrangian converges, it also holds $\tilde{k} = k_0$ and $\lambda_{lm} = W_1(\mathcal{P}_g, \mathcal{P}_r)$.

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

Based on the first order optimality conditions of the Lagrange multiplier method and the augmented Lagrangian method, we can see that α, which is fixed and being the real target, is equal to $\lambda_{al} + \rho \cdot (\tilde{k} - k_0)$ in augmented Lagrangian. However, in the true, the unregularized, the original objective, λ_{lm} should be equal to α, which means the following should hold:

$$\lambda_{lm} = \lambda_{al} + \rho \cdot (\tilde{k} - k_0). \tag{25}$$

Plus that the augmented Lagrangian method can be understood as the Lagrangian method with extra penalty term, which means λ_{al} shall play the role of λ_{lm}. Hence, the common suggestion for the update of λ_{al} in the augmented Lagrangian method is using the following update rule: (t indicates the iteration or timestamp)

$$\lambda_{al}^{t+1} = \lambda_{al}^{t} + \rho \cdot (\tilde{k} - k_0). \tag{26}$$

Thus, to upgrade from the penalty method to the augmented Lagrangian method, one need only introduce the Lagrangian multiplier L_{lm} and add an extra update step for λ_{al} according to Eq. (26) 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 gradient penalty, Lipschitz penalty and spectral normalization.

We will study the practical behaviors of various implementations of Lipschitz continuity regularization, including spectral normalization (SN), gradient penalty (GP), max gradient norm regularization with penalty method (or simply termed as max gradient norm penalty) (MaxGP), and max gradient norm regularization with augmented Lagrangian (MaxAL). In our experiments, the Lipschitz penalty shares a very similar performance as the gradient penalty, so we take out the Lipschitz penalty for clarity.

We use multilayer perceptions 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, 2014) with $\beta_1 = 0$ and $\beta_2 = 0.9$.

Frechet Inception Distance (FID) (Heusel et al., 2017) was 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., 2017; Borji, 2018). Hence, we here not include the results of Inception Score.

For this part of experiments, we use the WGANs because its theoretical results correspond to optimal transport, which can be more easily understood and checked.

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 $\mathcal{P}_g$ and another two points as $\mathcal{P}_r$. We fix these two distributions and train a discriminator with different implementations of Lipschitz regularization.

We want to check whether these methods are able to achieve the optimal discriminative function, by verifying the gradients of generated samples, which should follow the Proposition 1 and point towards their target real samples that minimizes 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 currently do not fully understand how such a phenomenon appears. We consider that it might because the global restriction on Lipschitz constant makes the capacity of the discriminator extremely underused such that the optimal discriminative function is not attainable. And we think the probable issues exist in the estimation of maximum singular value, i.e., power iteration, also holds a large portion of the possibility.

We have tried fairly large networks, but it does not help eliminate this problem. We have tried increasing the number of the power iteration that used to acquire the singular value, it does not solve this problem. We have also tried both in-place update of $\bar{W}_{SN}$ and update $\bar{W}_{SN}$ with collection, the problem consistently exists. Training the discriminator for a very

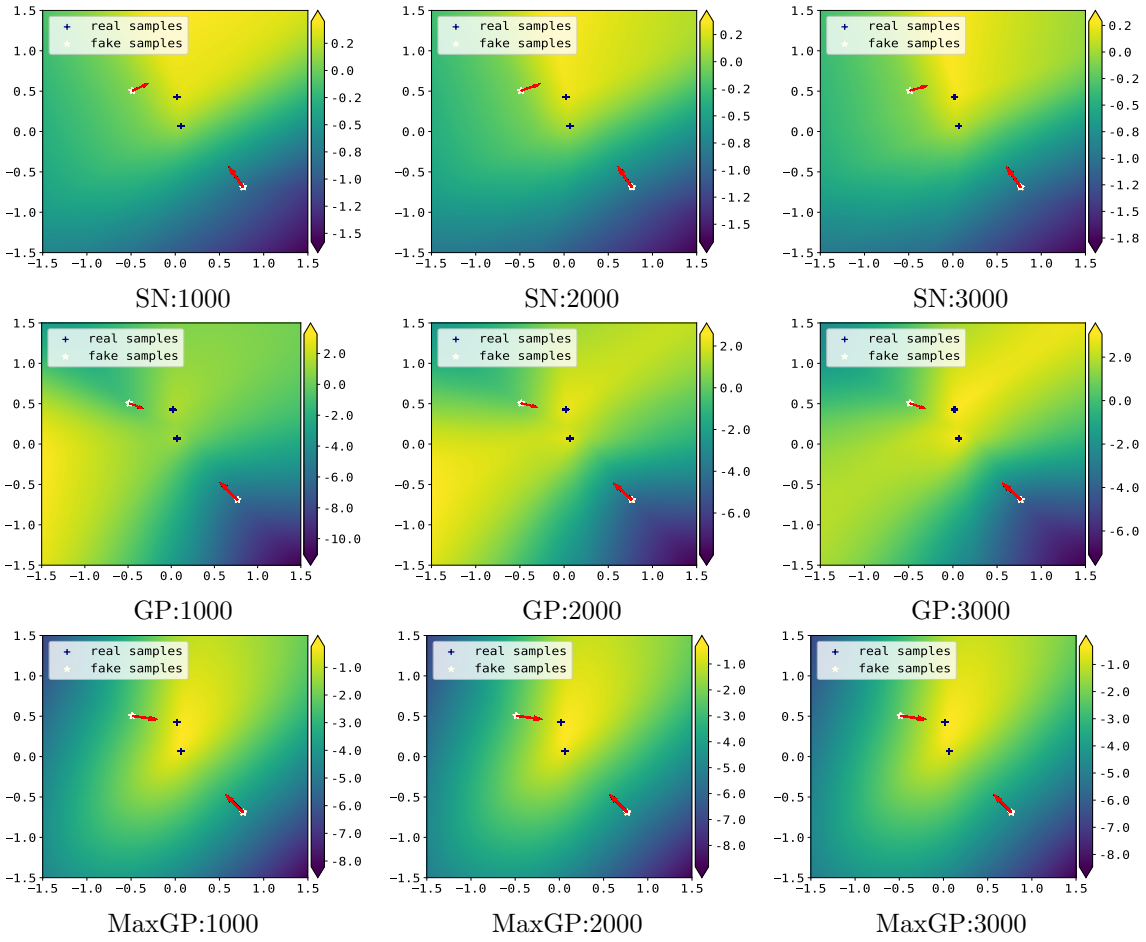
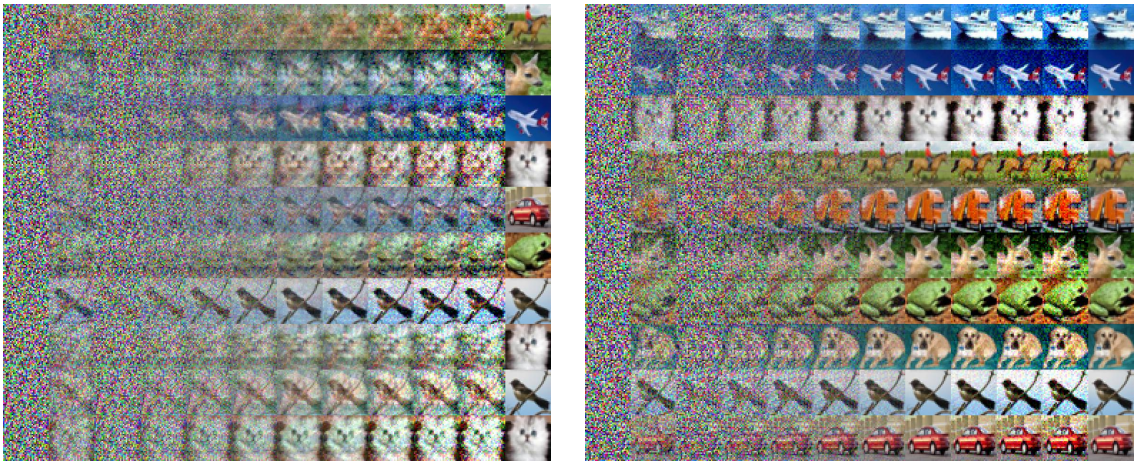


Figure 2: With $\mathcal{P}_g$ and $\mathcal{P}_r$ both being two random sampled points in two dimensional space, we train the discriminator using SN, GP and MaxGP, respectively. The numbers after the name of the methods are the corresponding iteration numbers. The arrows in the figures indicate the gradient scales and directions. From the results, we notice that: (i) SN in this case fails to achieve the optimal discriminator; (ii) the discriminator trained with GP is oscillatory; (iii) MaxGP stably converges to the optimal. Note that the results of SN and MaxGP seem to keep unchanged over iterations. That is because they have already basically converged with 1000 iterations. By contrast, GP keeps oscillatory all the way.

long time with a decreasing learning rate also cannot solve this problem and the final result keeps unchanged. We would leave further investigation as future work.

In Figure 2, we also noticed that GP leads to an oscillatory discriminator, which evidences that the superfluous constraints affect the optimal discriminator, and it turns out to lead to instability. It seems there is no stable optimum for the discriminative function under GP.

By contrast, we see that MaxGP quickly converged to the optimal discriminator (within 1000 iterations) and stably holds at the optimal state (keep almost unchanged), where the gradients of the fake samples point towards the real samples in an optimal transport manner.



(a) GP (b) MaxGP

Figure 3: With $\mathcal{P}_g$ and $\mathcal{P}_r$ being ten fixed noise and real images, respectively, we train discriminator using GP and MaxGP towards optimum. The leftmost in each row is a sample x from $\mathcal{P}_g$ 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 from $\mathcal{P}_r$. As we can see, GP fails to achieve the optimal discriminative function, where the gradients of fake samples do not strictly point towards real samples and tend to collapse to a subset of real samples. By contrast, with MaxGP, the gradients of $\mathcal{P}_g$ samples perfectly follow the optimal transport.

5.4.2 Toy Real World Data

We further compare these methods with real world data. We still want to check whether these methods converge to the optimal discriminative function. However, the real world dataset is too large, and we found practically, the optimal discriminator is almost unachievable. Hence, in this experiment, we use a small subset of the real world dataset, instead. Specifically, we select ten representative CIFAR-10 images as $\mathcal{P}_r$ and use ten random noise as $\mathcal{P}_g$. Then, same as above, we train the discriminator till optimal and check the gradient of the resulting discriminative function of different methods.

For the high dimensional case, visualizing the gradient direction is nontrivial. Hence, we plot the gradient and corresponding increments. In Figure 3, the leftmost in each row is a sample x from $\mathcal{P}_g$ 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 nearest real sample y from $\mathcal{P}_r$, i.e., the real sample that is closest to any point in the gradient directed path.

From the results, MaxGP is also able to achieve the optimal discriminative function in the high dimensional case. We see that the gradient of ten noises in $\mathcal{P}_g$ is pointing towards the ten real images in $\mathcal{P}_r$, 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 inadvertently introduced by GP are harmful.

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 objective, but also with the hinge loss (Miyato et al., 2018) and original GANs objective (Goodfellow et al., 2014; Fedus et al., 2017), which has also found work well under Lipschitz continuity regularization.

The results in terms of training curve of FID are plotted in Figure 5. In Figure 5a, we compare GP, MaxGP and MaxAL with different regularization weights under the objective of WGANs. We see that the training progresses and final results are quite similar to each other. The visual results are also provided in Figure 4.

As we found in the experiments of toy real world data, given $\mathcal{P}_g$ and $\mathcal{P}_r$ both consist of ten images, the optimal discriminator is already very hard to achieve. 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 result is too far from the optimum, so even whether it is biased or not, the final result appears similar.

We have also checked that, with real world dataset, even the relatively small dataset CIFAR-10 or MNIST, the gradient of the generated sample is basically nebulous. Maybe the nebulous gradient somehow points towards $\mathcal{P}_r$ (otherwise, how to explain the progress of the training), but being blurry (averaged?), and definitely not clearly points towards a certain real sample.

That is, in the current hyper-parameter settings, e.g., DCGAN architecture or shallow Resnet, the optimal discriminative function of WGANs is almost impossible to achieve.

It might also be related to the issues of the optimizer. Amam (Kingma and Ba, 2014), the common-used and somewhat powerful optimizer for GANs, is recently shown to not guarantee the convergence (Reddi et al., 2018; Zhou et al., 2018; Zou et al., 2018). We are keeping investigating this phenomenon.

In this experiment, we initially use the WGANs objective for all methods. However, we found that with the Resnet architecture (Gulrajani et al., 2017), SN fails to converge. The same holds with various small modifications of hyper-parameters. We notice 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 GANs objective, which was found to also work well given the Lipschitz regularization.

The results are plotted in Figure 5b. We also include the results of MaxGP with these objectives for comparison. As we can see, the result of MaxGP is generally better than SN.

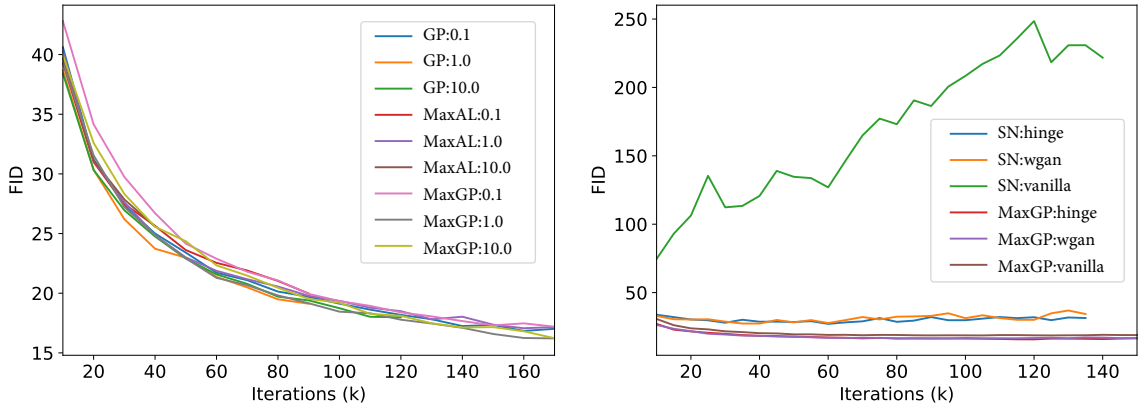
Lastly, we inspect the properties of MaxAL. As shown in Figure 6a, MaxAL is able to quickly restrict the Lipschitz constant to the given target, i.e., 1, and keep the Lipschitz constant fairly stable during the training. By contrast, the Lipschitz constants under GP and MaxGP keep changing, decreasing as $\mathcal{P}_g$ getting closer to $\mathcal{P}_r$.

Another interesting fact about MaxAL is that, when trained with the WGANs objective, the optimal λ is equivalent to $W_1(\mathcal{P}_g, \mathcal{P}_r)$. We verify this fact by plotting these two terms during training together. As shown in Figure 6b, the two lines are basically overlapped.

(a) GP (b) SN

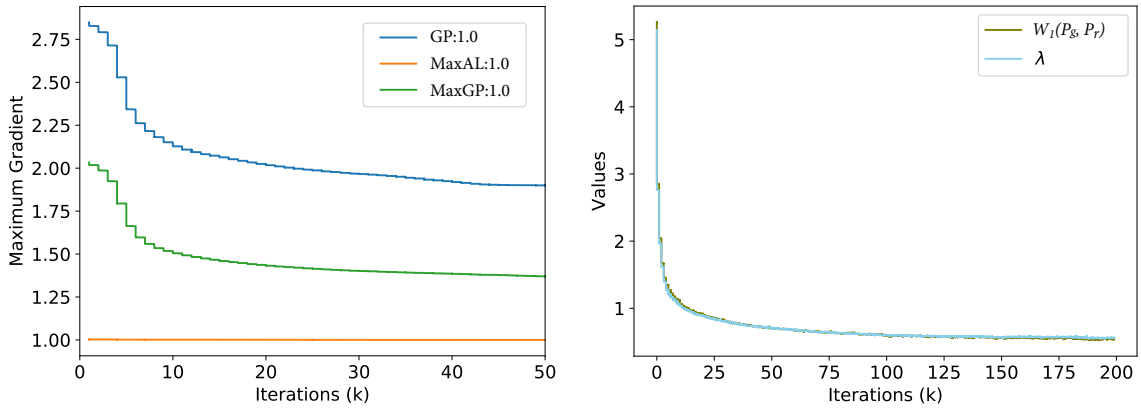
(c) MaxGP (d) MaxAL

Figure 4: The visual comparison of different implementations of Lipschitz regularization under unsupervised CIFAR-10 generation with the WGANs objective. The training with SN diverges, when using WGANs objective. Here we plot its result with hinge loss, instead.



(a) (b)

Figure 5: The quantitative comparison of different implementations of Lipschitz regularization under unsupervised CIFAR-10 generation with the WGANs objective 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 objective 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 diverges, when using WGANs objective. And even when using the hinge loss or original GANs, the final results of SN are still apparently worse than MaxGP.



(a) (b)

Figure 6: The favorable properties of MaxAL. With MaxAL, the Lipschitz constant quickly converges to the given target. By contrast, the Lipschitz constants achieved by GP and MaxGP are dynamic. In addition, the value of λ is equivalent to the Wasserstein distance.

5.5 Summary on Lipschitz Regularization Implementations

Up to now, we demonstrated that restricting the Lipschitz constant over the support of the interpolations of real and fake samples is sufficient to gain the advantageous gradient properties induced by Lipschitz continuity regularization. It provides theoretical guarantee on the validity of these empirical gradient penalty based methods.

In the meantime, it suggests that global restriction on the Lipschitz constant is unnecessary. Combined with the fact that we found the spectral normalization, the method that provides global restriction on the Lipschitz constant somehow fails in many practical scenarios. Although the real issues may actually exist in the estimation of maximum singular value. Given that there is currently no other good alternative for power iteration, we suggest using these methods that regularize local Lipschitz constant in the blending region.

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

We have accordingly proposed revisions to the gradient penalty. Our experiments demonstrated that the proposed MaxGP is able to achieve the optimal discriminative function in an unbiased manner. We also suggested augmented Lagrangian as a simple yet good alternative to the penalty method, which is able to strictly restrict the Lipschitz constant to a given target.

6. Empirical Analysis and Verification of Lipschitz Regularized GANs

With both theoretically sound and practically well-behaving MaxGP, we now verify the theoretical properties of LGANs and benchmark various instances of LGANs, showing its consistent superior performances over WGANs.

To adopt MaxGP for LGANs, we just need to set $k_0 = 0$. In many cases, actually, MaxAL can also be used, if preferred. Because if k_0 is small enough and $\mathcal{P}_g$ and $\mathcal{P}_r$ is not close enough, then all required bounding relationships can be established. Penalizing the Lipschitz constant is to ensure the establishment of effective bounding relationships that moves samples from where has too much to where has too less, when the two distributions are too close. Nevertheless, the following experiments are based on MaxGP.

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 Points Towards Real Sample

One important theoretical benefit of LGANs is that $\nabla_x f^*(x)$ for each generated sample is guaranteed to point towards some real sample. We here verify the gradient direction of $\nabla_x f^*(x)$ with a set of ϕ and φ that satisfy Eq. (11).

The tested objectives 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) $\phi(x) = \varphi(-x) = \exp(x)$. And they are tested in two scenarios: two dimensional toy data and real-world high-dimensional data.

In the two dimensional case, $\mathcal{P}_r$ consists of two Gaussians and $\mathcal{P}_g$ 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 solving f^* feasible, we use ten CIFAR-10 images as $\mathcal{P}_r$ and ten fixed noise images as $\mathcal{P}_g$. Note that we fix $\mathcal{P}_g$ on purpose because to verify the direction of $\nabla_x f^*(x)$, learning $\mathcal{P}_g$ is not necessary.

The results are shown in Figures 7 and 8, respectively. In Figure 7, we can see that the gradient of each generated sample is pointing towards some real sample.

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 from $\mathcal{P}_g$ 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 nearest real sample y from $\mathcal{P}_r$. This result visually demonstrates that the gradient of a generated sample is towards a real sample.

Note that the final results of Figure 8 keep almost identical, when varying the loss metrics ϕ and φ in the LGANs family, which is reasonable. Because when the supports of $\mathcal{P}_g$ and $\mathcal{P}_r$ are disjoint, according to our analysis, LGANs behaves just like WGANs, in the sense that every sample in $\mathcal{S}_g$ must get bounded by a real sample.

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

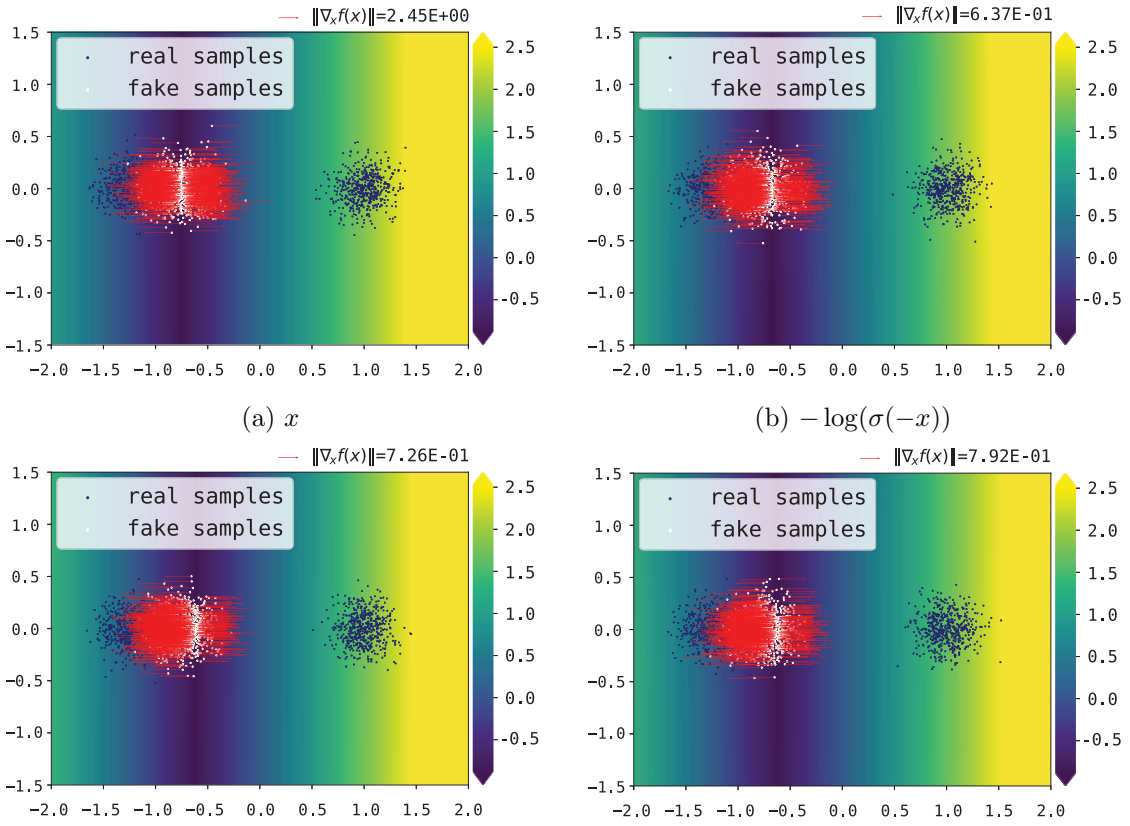
The loss metrics that correspond to the Wasserstein distance is a very special case that has a solution under Lipschitz regularization. It is the only case where both ϕ and φ have constant derivatives, i.e., both are not strictly convex.

As a result, f^* under the Wasserstein distance loss metrics has a free offset, i.e., given some f^*, $f^* + \alpha$ with any $\alpha \in \mathbb{R}$ is also an optimal. In practice, it behaves as oscillations in $f(x)$ during training.

The oscillations affect the practical performance of WGANs. Karras et al. (2017) and Adler and Lunz (2018) introduced regularization to the discriminative function to prevent $f(x)$ drifting during the training. By contrast, any other instance of LGANs does not have this problem. We illustrate the practical difference in Figure 9.

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

The qualitative change happens, when $\mathcal{P}_g$ converges to $\mathcal{P}_r$. At that time, the training LGANs will fundamentally stop with wholly zero gradients passing through G-D, because $k(f) = 0$. But WGANs, requiring $k(f)$ being one, will keep fluctuating.



(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.

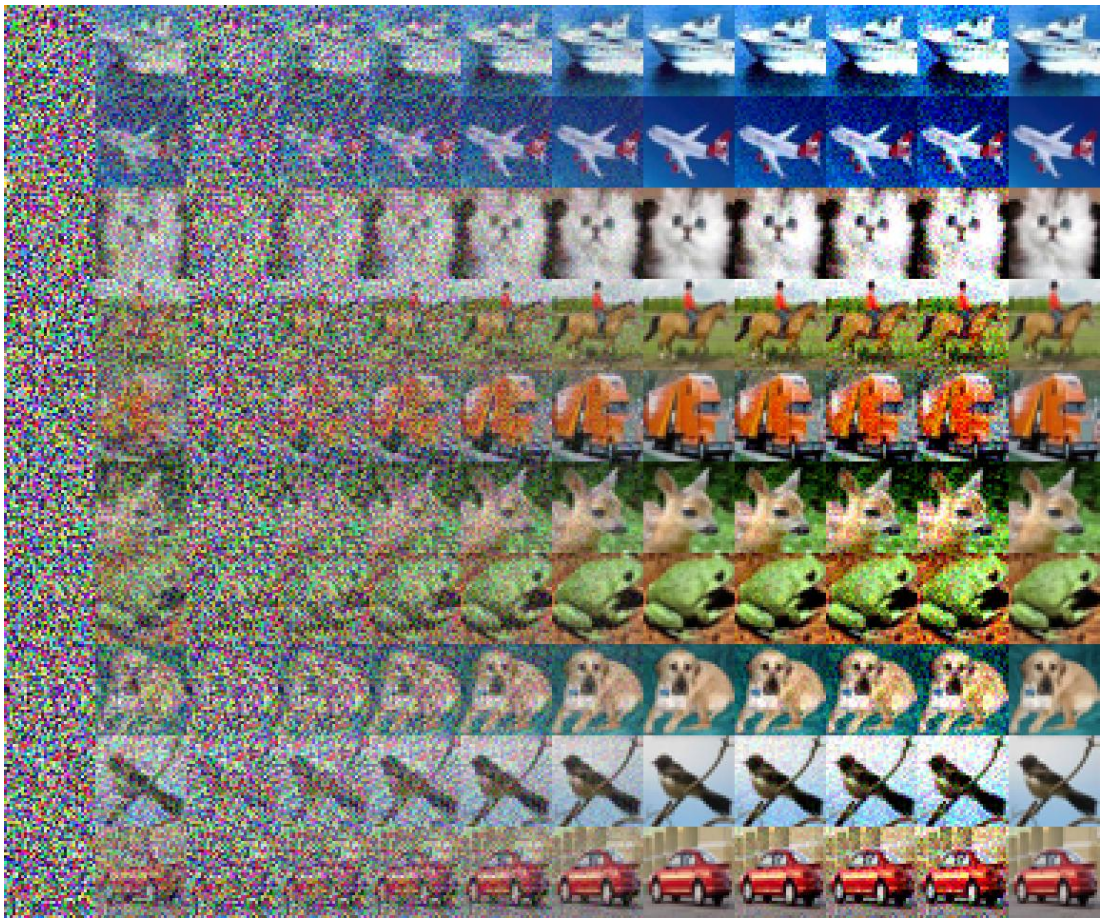
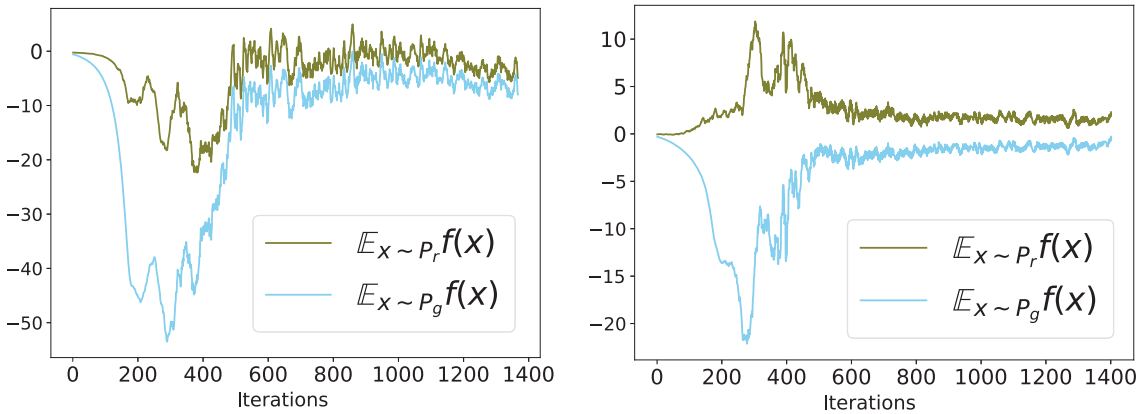


Figure 8: Verifying $\nabla_x f^*(x)$ in LGANs point towards real samples with gradient gradation.



(a) WGANs.

(b) LGANs.

Figure 9: The benefit of the uniqueness of f^* in LGANs. f is more stable during training.

6.3 Benchmark with Unsupervised Image Generation

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

In this part of experiments, we also include the hinge loss $\phi(x) = \varphi(-x) = \max(0, x+\alpha)$ and quadratic loss (Mao et al., 2016), which do not fit the assumption of strict monotonicity. For the quadratic loss, we set $\phi(x) = \varphi(-x) = (x+\alpha)^2$. We set $\alpha = 1.0$ in all the experiments.

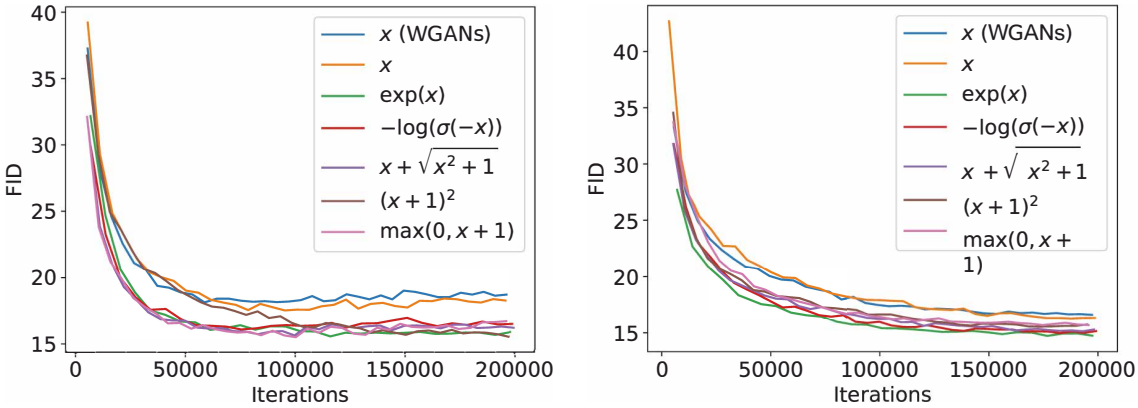
The strict monotonicity assumption of ϕ and φ is critical in Theorem 2 to theoretically guarantee the existence of bounding relationships for *arbitrary datas*. But if we further assume S_g and S_r are limited, it is possible that there exists a suitable λ such that all real and fake samples lie in a strict monotone region of ϕ and φ. Then, the hinge loss and even the quadratic loss may also work well. For hinge loss, it would mean $2\alpha < k(f) \cdot \|y - x\|$ for all $x \in \mathcal{S}_g$ and $y \in \mathcal{S}_r$.

Our tentative experiments show that the choice of $\psi(x)$ does not play a central role. But in this experiment, we still fix the loss metric $\psi(x)$ in the generator's objective as $-x$. The thought behind our current choice of $\psi(x)$ is that: if we choose to use the minimax formulation, though we can get the minimax explanation of what the generator is minimizing, it will have some strange property. That is, when ϕ is strictly convex, then the samples with lower $f(x)$ value will get a smaller gradient scale because it is weighted by $\nabla_{f(x)}\psi(f(x))$; but a lower $f(x)$ value somehow (not strict and not always true) indicates this sample has a larger distance to the real distribution. And on the other hand, setting $\psi(x) = -x$ is also very easy to understand, i.e., updating samples with evenly distributed weights. We believe the choice of $\psi(x)$ is also an interesting research topic and we leave it as 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 all experiments, we adopt the network structures and hyper-parameter setting from (Gulrajani et al., 2017), where WGANs-GP in our implementation achieves IS 7.71 ± 0.03 and FID 18.86 ± 0.13 on CIFAR-10. We use MaxGP and search the best λ in $[0.01, 0.1, 1.0, 10.0]$. 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 being popular, is not well explained (Zhou et al., 2017; Borji, 2018). And it is highly unstable during training. By contrast, FID is fairly stable. We include IS for better reference.

We plot the training curves in terms of FID in Figures 10a and 10b. The training curves in terms of IS are provided in the Appendix. From the Figures and / or Table 2, we can clearly tell that all LGANs instances consistently and remarkably outperform WGANs or LGANs with Wasserstein distance loss metrics. Different instances of LGANs have relatively similar final results, while the objectives $\phi(x) = \varphi(-x) = \exp(x)$ and $\phi(x) = \varphi(-x) = x + \sqrt{x^2+1}$ achieve the best performances.

This is probably because LGANs with strictly convex loss metrics reduces the gradient, i.e., the benefit of further discriminating this sample, of well-identified points towards zero, which enables the discriminator to pay more attention to these ill-identified. And hence LGANs generally work better than WGANs and these instances with relatively strong convexity perform better.



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

Figure 10: Training curves in terms of FID. WGANs and a set of instances of LGANs.

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

Objective	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
$\exp(x)$	$\mathbf{8.03 \pm 0.03}$	$\mathbf{15.64 \pm 0.07}$	8.67 ± 0.04	$\mathbf{14.90 \pm 0.07}$
$-\log(\sigma(-x))$	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
$(x+1)^2$	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

The hinge loss and quadratic loss with a suitable λ turn out to also work pretty good. But with $\alpha = 0.1$, its performance significantly dropped, which is consistent with our analysis.

7. How Traditional GANs Works

To gain a more experienced understanding upon the training issue of unregularized GANs, and at the same time, to understand how unregularized GANs works in practice, we first provide a more systematic discussion and taxonomy of its gradient issues, and then we will study the practical behavior these gradient issues. And we will also explain why mode collapse is common in unregularized GANs.

We realize that the similar analysis apply to some other regularized GANs, whose $f^*(x)$ share the similar properties of unregularized GANs, i.e., only reflect local information and positively correlated with $\mathcal{P}_r(x)$ and negatively correlated with $\mathcal{P}_g(x)$, e.g., Fisher GANs. Hence, we sometimes more generally refer to them as traditional GANs.

In the following, we will show that $\nabla_{f(x)}\varphi(f(x))$ may lead to Type-I gradient vanishing, and $\nabla_x f(x)$ is involved with both Type-II gradient vanishing and faulty gradient direction, which is a generalized concept of gradient uninformativeness.

7.1 Type-I Gradient Vanishing

The well-known gradient vanishing problem (Goodfellow et al., 2014; Arjovsky and Bottou, 2017) mainly refers to the problem in the original GANs. It should be noted that, in terms of Eq. (7), the gradient vanishing problem in the original GANs[5] mainly stems from the vanishing scale in the scalar term $\nabla_{f(x)}\varphi(f(x))$, which we refer to as the Type-I gradient vanishing. We will show in the next section that the vector term $\nabla_x f(x)$ may also be zero, which leads to another type of gradient vanishing that we call the Type-II gradient vanishing.

Interestingly, the Least-Squares GANs (Mao et al., 2016) which adopts the quadratic loss as the loss metric, avoids the Type-I gradient vanishing, but still may suffer from the Type-II gradient vanishing.

The occurrence of Type-I gradient vanishing, i.e., $\nabla_{f(x)}\varphi(f(x)) = 0$, has two necessary conditions: (i) the existence of extreme point, i.e., $s \triangleq \{x \mid \nabla_x\varphi(x) = 0\}$ and $s \neq \emptyset$; (ii) the accessibility of extreme point, i.e., $\{x \in \mathcal{S}_g \mid f(x) \in s\} \neq \emptyset$.

In the original GANs, by switching to an alternative generator objective function $\varphi(x) = -\log\sigma(x)$, it avoids the accessibility of extreme points and hence solves the Type-I gradient vanishing problem.

Wasserstein GANs (Arjovsky et al., 2017), with $\varphi(x) = x$, avoids the existence of extreme points and thus avoids the Type-I gradient vanishing. LGANs avoids the existence of extreme points via penalizing the Lipschitz constant and hence forming bounding relationships.

The Least-Squares GANs (Mao et al., 2016), with $\phi(x) = (x-\alpha)^2$ and $\varphi(x) = (x-\gamma)^2$ and $\alpha \neq \gamma$, avoids the Type-I gradient vanishing via avoiding the accessibility of extreme points.

7.2 Type-II Gradient Vanishing and Faulty Gradient Direction

We here introduce the gradient issues arising from $\nabla_x f(x)$. We will study it from the perspective of the gradients of the optimal discriminative function at sample points, i.e., by analyzing $\nabla_x f^*(x)$.

Generally, if a sample x is at the local optimum of f^*, then it suffers a substantive zero-gradient, which we refer to as the Type-II gradient vanishing.

We broadly name gradients that do not guarantee convergence as faulty gradients. To highlight the importance of gradient direction, we refer to this problem as faulty gradient direction problem, which includes: (i) uninformative gradient; (ii) theoretically undefined gradient; (iii) unconverged Type-II gradient vanishing; (iv) local-greedy gradient.

As an important sub-case of faulty gradient direction problem, we also separately name the problem caused by uninformative gradient as the gradient uninformativeness problem.

The faulty gradient direction problem is orthogonal to gradient vanishing: gradient vanishing is about the scale of the gradient (overall or only the $\nabla_{f(x)}\varphi(f(x))$ part), however, faulty gradient direction is mainly about the direction of the gradient.

5. In the original GANs, the optimal $f^*(x)$ for a fake sample is negative infinite. In practice, the value $f(x)$ of fake samples tends to be different, and thus it does not suffer the Type-II gradient vanishing.

Still one can consider the ideal case, where $\mathcal{P}_g$ and $\mathcal{P}_r$ are totally overlapped and both consist of n discrete points but their probability mass over these points are different, to understand that the two are indeed orthogonal: its gradient direction is meaningless, but the gradient does not necessarily vanish.

- Uninformative gradient: if the optimal discriminative function only reflects the local densities, when a generated sample is not surrounded by reals samples, its gradient tells nothing about $\mathcal{P}_r$. Typical situation is that $\mathcal{P}_g$ and $\mathcal{P}_r$ are disjoint, which is common in practice (Arjovsky and Bottou, 2017).
- Theoretically undefined gradient: for generated sample x, if the optimal discriminative function is not fully defined in the surrounding of x, e.g., it is isolated and at the boundary. It suffers from a theoretically undefined gradient.
- Unconverged Type-II gradient vanishing: for generated samples that theoretically has a Type-II gradient vanishing, despite the practical existence of Type-II gradient vanishing (e.g., Figure 11a), it more commonly suffers the unconverged version of Type-II gradient vanishing, which behaves as noisy gradient (e.g., fake samples in the central of the left region in Figure 11b).
- Local-greedy gradient: when the optimal discriminative function only reflects the local densities, even if the gradient is well-defined and nonzero, the gradient update based on $\nabla_x f(x)$ is local-greedy, which turns out to be an intrinsic cause of mode collapse. See Section 7.5 for more details.

Samples that suffer from the uninformative gradient might at the same time suffer from theoretically undefined gradient: the uninformative gradient happens when the generated sample is not surrounded by reals samples; if it is further not fully surrounded by any kind of (real or fake) samples, it also suffers the theoretically undefined gradient.

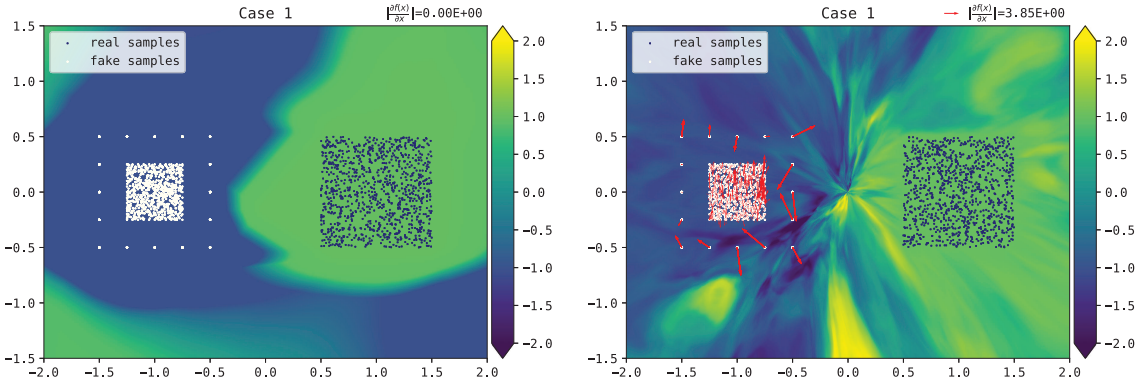
Meanwhile, theoretically undefined gradient is not a sub-problem uninformative gradient: for samples at the boundary, it suffers theoretically undefined gradient, but its gradient might be uninformative, if some real samples are in the surrounding.

If the variation of $f^*(x)$ in a region is too small, due to the precision limitation of the computing device, the practical neural network may not be able to capture the statistical variation. Then it may also behave as the Type-II gradient vanishing.

7.3 The Practical Behaviors of These Gradient Issues

To study the practical behaviors of various gradient issues and understanding how GANs that theoretically has gradient issues works in practice. We conducted a set of experiments with various hyper-parameter settings. We use the Least-Squares GANs as a representative of traditional GANs in this experiment. The value surface and the gradient of generated samples under various situations are plotted as follows.

These experiments showed that the practical f highly depends on the hyper-parameter setting. Given limited capacity, the neural network tries to learn the best f which might lead



(a) Type-II gradient vanishing, converged (b) Type-II gradient vanishing, unconverged

Figure 11: When $\mathcal{S}_g$ and $\mathcal{S}_r$ are disjoint, depending on the hyper-parameters and training states, samples inside $\mathcal{S}_g$ suffer from the Type-II gradient vanishing if converged (Left), or otherwise suffer from faulty gradient direction that stems from unconverged Type-II gradient vanishing (Right). The gradient for a sample point that is isolated or at the boundary is theoretically undefined, and in practice, also behaves as a faulty gradient.

to a simple value surface. When the neural network is capable of learning the (approximate) optimal f^*, how the actual f approaches f^* and how the theoretically undefined points behave highly depends on the optimization details and the characteristics of the network.

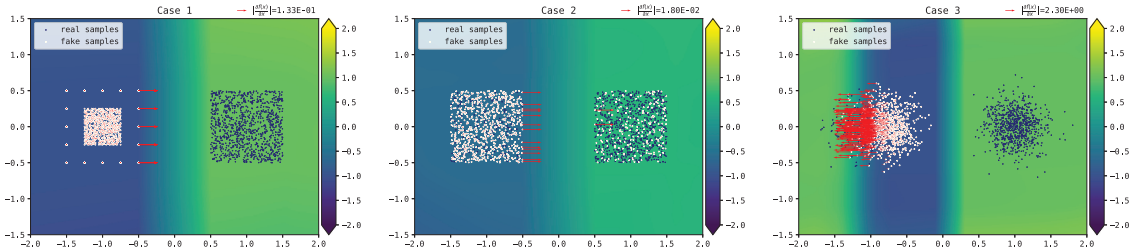
For points whose gradients are theoretically undefined, their practical behaviors highly depend on the detailed setting, which means it is hard to control and is controlled by hyper-parameters tuning.

According to the above experiments: (i) a low-capacity network tends to learn a simple surface; (ii) Adam, compared with SGD, tends to learn a simpler surface; (iii) large learning rate tends to learns a simpler surface than small learning rate; (iv) piecewise linear activation (e.g., ReLU) tends to result in simper value surface, compared with highly nonlinear activation function (e.g., SELU, Klambauer et al. (2017)). For Adam, we set $\beta_1 = 0.0$ and $\beta_2 = 0.9$.

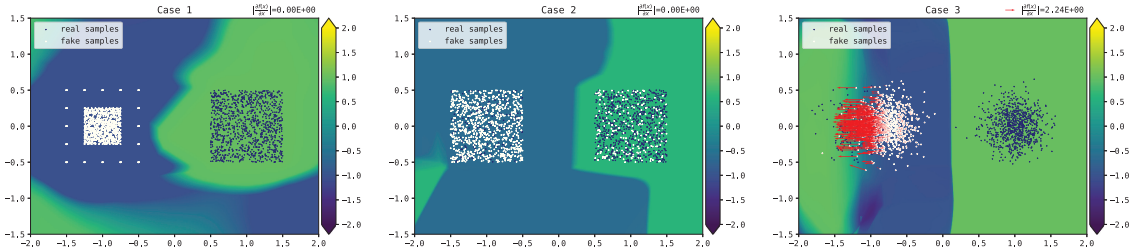
7.4 Explanation on the Empirical Success of Traditional GANs

Although traditional GANs do not have guarantee on its convergence, they have already achieved great success. The thing is that having no guarantee does not mean it cannot converge. It turns out extensive hyper-parameter tuning increases the probability of its convergence.

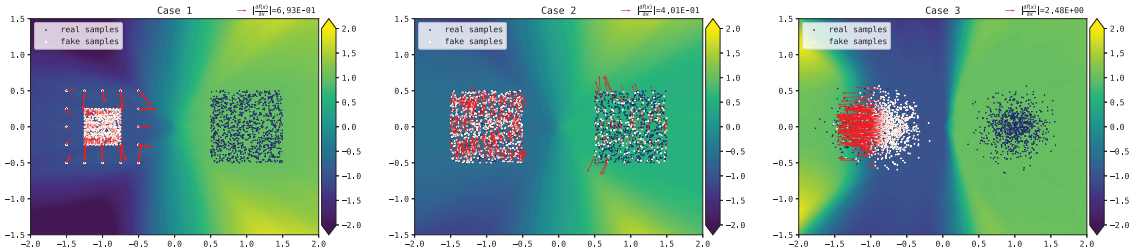
As shown in Figure 12, hyper-parameters, including network architecture, are important in influencing the value surface of f. Some typical settings (e.g., simplified neural network architecture, ReLU or leaky ReLU activation, relatively high learning rate, Adam optimizer, etc.) tend to form a relatively simple / smooth value surface, e.g., monotonically increasing from $\mathcal{S}_g$ to $\mathcal{S}_r$), making the theoretically meaningless $\nabla_x f^*(x)$ much more meaningful. That is, one can find these settings, where $\nabla_x f^*(x)$ or $\nabla_x f(x)$ is more favourable, to enable traditional GANs to work.



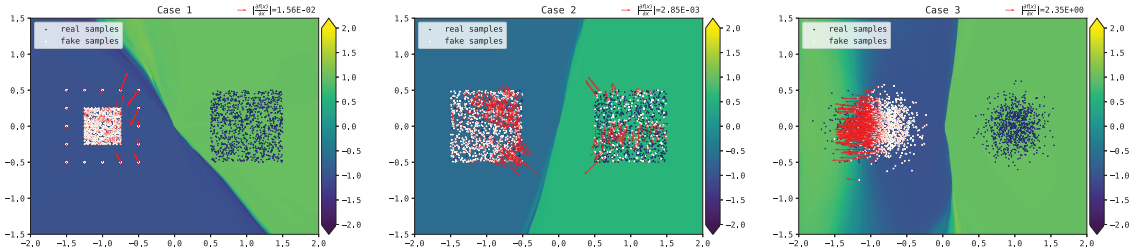
(a) Adam with lr=1e-2, MLP with RELU, #hidden units=1024, #layers=1.



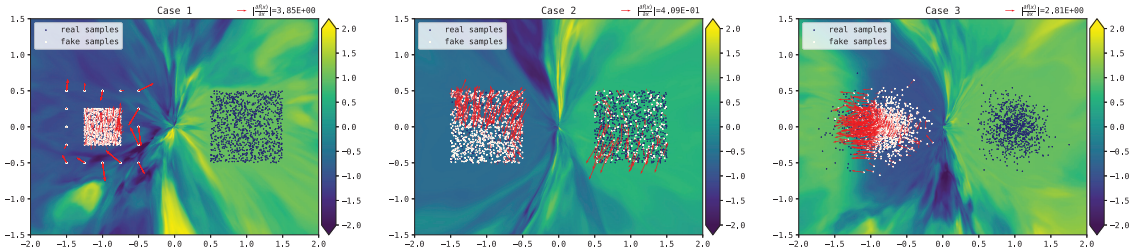
(b) Adam with lr=1e-2, MLP with ReLU, #hidden units=1024, #layers=4.



(c) Adam with lr=1e-5, MLP with ReLU, #hidden units=1024, #layers=4.



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



(e) SGD with lr=1e-4. MLP with SELU, #hidden units=128, #layers=64.

Figure 12: The practical f highly depend on the hyper-parameter setting. Some particular settings lead to simple (or other favorable) value surfaces. Hyper-parameter setting that leads to a simple value surface is more likely to have successful training, which is basically consistent with the empirical success and failure of traditional GANs in common practice.

In contrast, we have tried highly-nonlinear activation such as SWISH (Ramachandran et al., 2018) in the discriminator. It turns out traditional GANs are very likely to fail. By contrast, our proposed Lipschitz GANs are compatible with highly-nonlinear activations.

Another important empirical technique is to delicately balance the generator and the discriminator or limit the capacity of the discriminator. This can be understood as it is trying to avoid the fatal optimal f^* and making the value surface not being overstretched.

Nevertheless, without theoretically guaranteeing its convergence, traditional GANs are practically hard to use, being sensitive to hyper-parameters and easily broken. By contrast, we believe WGANs and LGANs do not have such kind of training issues and can be much more easy to use, especially LGANs, which superior over WGANs with its strictly convex properties: (i) uniqueness of f^*, avoided the possible drifting of f during training; (ii) LGANs lowers the weights for well-distinguished samples in the objective, hence the discriminator can play more attention to these ill-distinguished, which seems to be the key fact that leads to the superior performance of LGANs against WGANs; (iii) LGANs can truly stop the training, when $\mathcal{P}_r$ converged to $\mathcal{P}_r$, with $k(f) = 0$ and entirely zero gradient flow among G and D, while WGANs would not and will keep oscillating (Mescheder et al., 2018).

7.5 The Cause of Mode Collapse

Previously, we mainly discussed the problem of $\nabla_x f^*(x)$ in the cases where $\mathcal{S}_g$ and $\mathcal{P}_r$ are disjoint or being discrete. In this section, we extend our discussion to the overlapping and continuous cases.

In the disjoint or discrete cases, we argue that, in traditional GANs, typically these unregularized GANs, $f^*(x)$ on $\mathcal{P}_g$ does not reflect any information about the location of other points in $\mathcal{P}_r$, which will lead to an unfeasible $\nabla_x f^*(x)$ and thus nonconvergence.

In the overlapping and continuous case, things are actually different, $f^*(x)$ around each point is also defined, and its gradient $\nabla_x f^*(x)$ now reflects the local variation of $f^*(x)$.

For most traditional GANs, $f^*(x)$ mainly reflects the local information about the density $\mathcal{P}_g(x)$ and $\mathcal{P}_r(x)$. However, it is worth noting that $f^*(x)$ is usually an increasing function with respect to $\mathcal{P}_r(x)$, while a decreasing function with respect to $\mathcal{P}_g(x)$. For instance, $f^*(x)$ in the original GANs is $\log \frac{\mathcal{P}_r(x)}{\mathcal{P}_g(x)}$.

Optimizing the generator according to $\nabla_x f^*(x)$ will move the sample x following the direction of increasing $f^*(x)$. Because $f^*(x)$ is positively correlated with $\mathcal{P}_r(x)$ and negatively correlated with $\mathcal{P}_g(x)$, it in a sense means x is becoming more real. However, such a local-greedy property turns out to be a fundamental cause of mode collapse.

Mode collapse is a notorious problem in GANs's training, which refers to the phenomenon that the generator only learns to produce / imitate part(s) of $\mathcal{P}_r$, while missing some others.

A good deal of literature has tried to study the source of mode collapse (Che et al., 2016; Metz et al., 2016; Kodali et al., 2017; Arora et al., 2017) and measure the degree of mode collapse (Odena et al., 2016; Arora and Zhang, 2017).

The most recognized cause of mode collapse is that, if the generator is much stronger than the discriminator, it may learn to only produce the sample(s) in the local or global maximum(s) of $f(x)$ of the current discriminator.

This argument is true for most GANs. However, from our perspective on $f^*(x)$ and its gradient, there actually exists a much more fundamental cause of mode collapse, i.e., the locality of $f^*(x)$ in traditional GANs and the locality of gradient operator ∇.

In traditional GANs, $f^*(x)$ is a function of local densities $\mathcal{P}_r(x)$ and $\mathcal{P}_g(x)$, which is local. And the gradient operator ∇ is also a local operator. As a result, $\nabla_x f^*(x)$ only reflects its local variations and cannot capture the statistic of $\mathcal{P}_r$ and $\mathcal{P}_g$ that is far from itself.

If $f^*(x)$ in the surrounding area of x is well-defined, $\nabla_x f^*(x)$ will move x towards the *nearby* location, where the value of $f^*(x)$ is higher. It does not take the global statistics into account. It will not be aware that there might be some place where samples are missing (i.e., there is a mode missing) and here the samples are too much (i.e., here is a mode collapse).

The typical result is that when fake samples get close to a mode of $\mathcal{P}_r$, they move towards the mode. And then get stuck there, due to the locality. Because here is no internal force to move them out of the mode collapse state. Because they can not tell from $\nabla_x f^*(x)$ its current mode collapse state or the far way mode missing information. They just follow $\nabla_x f^*(x)$ and keep clustered together or vibrate around the mode; or once the entire surface is changing, they follow the entire surface move from one mode to another, or cycling in this manner. This is the practically observed behaviour of mode collapse.

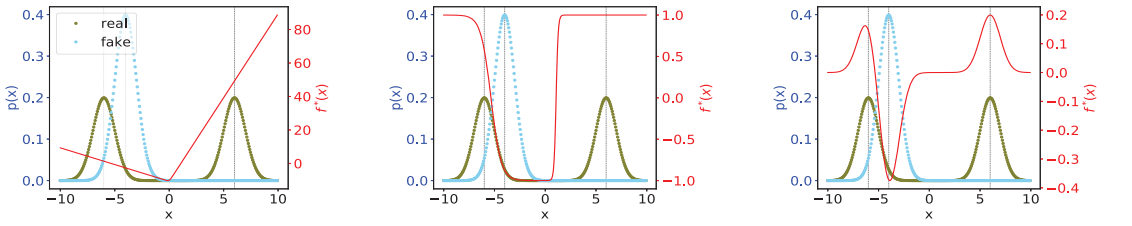
Let's simulate some specific cases for a more intuitive sensation of this phenomenon. Let assume $\mathcal{P}_r$ consists of two Gaussian distributions (A and B) that are distant from each other, while the current $\mathcal{P}_g$ is uniformly distributed over its support and close to real Gaussian A. In this case, $\nabla_x f(x)$ of all fake samples will point towards the center of Gaussian A.

If $\mathcal{P}_g$ is a Gaussian with the same standard deviation as Gaussian A, $\nabla_x f(x)$ in the original GANs and Least-Squares GANs shows almost identical behaviors, which is illustrated in Figure 13. In Fisher GANs, if $\mu(x)$ is uniform, the case is even worse: a large amount of points that are relatively far from Gaussian A will move away from A. Although, in our 1-D case, it is pointing towards B, but this does not necessarily hold in higher dimensions. The third column of Figure 12 simulates the above setting in the two dimensional case, and the samples tend to move towards the nearby mode.

As a summary, in the overlapping and continuous case, though $\nabla_x f^*(x)$ indeed carries information about $\mathcal{P}_r$, $\nabla_x f^*(x)$ based updating turns out to be a local-greedy strategy, which is still unfavorable and is a fundamental cause of mode collapse in traditional GANs, typically these unregularized GANs, whose $f^*(x)$ only reflect the local information and is positively correlated with $\mathcal{P}_r(x)$ and negatively correlated with $\mathcal{P}_g(x)$.

7.6 Adversarial Activation Maximization

Zhou et al. (2017) proposed the adversarial *activation maximization* understanding of the training of unregularized GANs.



(a) Original GAN (b) Least-Squares GAN (c) Fisher GANs with uniform μ

Figure 13: The source of mode collapse. In traditional GANs, $f^*(x)$ is a function of the local densities $\mathcal{P}_g(x)$ and $\mathcal{P}_r(x)$. Given $f^*(x)$ is an increasing function of $\mathcal{P}_r(x)$ and decreasing function of $\mathcal{P}_g(x)$, when fake samples get close to a mode of the $\mathcal{P}_r$, they will follow $\nabla_x f^*(x)$ and move towards the mode. And then, they will keep clustered together or vibrate around the mode; or once the entire surface is changing, they follow the entire surface move from one mode to another, or cycling in this way.

Activation maximization is a traditional technique for visualizing or understanding the neuron(s) in pretrained neural networks. However, the maximized activation of a neuron by itself is not guaranteed to be a valid sample (i.e., not necessarily of high quality), and can be noise, which is also often regarded as a fake sample or adversarial example.

In unregularized GANs, the generator plays the role of doing activation maximization, while the discriminator plays the role of differentiating the fake or adversarial samples, preventing them from getting their desired high activation, and thus ensures the high-activation is achieved by high-quality samples that strongly confuse the discriminator.

With the adversarial training between the generator and discriminator, the adversarial activation maximization process helps solve the problem in vanilla activation maximization, and achieve valid high-activation and the generation of new or realistic samples.

Adversarial activation maximization is interesting, but we have to highlight that it is just another understanding. It does not change the fact that unregularized GANs does not guarantee its convergence. But, indeed, it helps understand how unregularized GANs works in practice and what is its limitations (i.e., the process may fail) from another perspective.

8. The Envelope Theorem Perspective: the Essence of Convergence

Here, we explain the gradient issues from the perspective of the envelope theorem. The envelope theorem (Milgrom and Segal, 2002) is a classic result about the differentiation properties of a (constrained) optimization problem.

8.1 The Envelope Theorem

Let the parameter of discriminator be ϑ and the parameter of generator be θ. $J_D(\vartheta, \theta) = \mathbb{E}_{z\sim\mathcal{P}_z}[\phi(f_\vartheta(g_\theta(z)))] + \mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(f_\vartheta(x))]$. Consider the problem

$$J(\theta) = \arg\min_{\vartheta} J_D(\theta, \vartheta) \quad s.t. \quad s(\theta, \vartheta) \leq 0. \tag{27}$$

The Lagrangian dual problem is given by

$$L(\theta, \vartheta, \lambda) = J_D(\theta, \vartheta) + \lambda \cdot s(\theta, \vartheta), \tag{28}$$

where λ are the Lagrange multipliers.

Let ϑ^* and λ^* together be the solution that minimizes the objective function $L(\theta; \vartheta, \lambda)$.

According to the envelope theorem, if J and L are *continuously differentiable*, we have that

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial L(\theta, \vartheta, \lambda)}{\partial \theta}\Big|_{\vartheta=\vartheta^*, \lambda=\lambda^*} = \frac{\partial L(\theta; \vartheta^*, \lambda^*)}{\partial \theta} = \frac{\partial J_D(\theta; \vartheta^*)}{\partial \theta} + \lambda^* \cdot \frac{\partial s(\theta; \vartheta^*)}{\partial \theta}. \tag{29}$$

8.2 Unregularized GANs

As an illustrative sample, we first consider the following setting: let $\mathcal{P}_g$ be a distribution on two points a and $1 + a$ in $\mathbb{R}$ with probability of p and $1 - p$, respectively. And the real distribution 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 $\mathcal{P}_g$ and $\mathcal{P}_r$ 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 $1 - a$. In any case, $\frac{\partial f_{\vartheta^*}(x)}{\partial x}$ 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:

$$\begin{aligned} \frac{\partial J(a, p)}{\partial a} &= \frac{\partial J_D(a, p; \vartheta^*)}{\partial a} = p \frac{\partial \phi(f_{\vartheta^*}(a))}{\partial f_{\vartheta^*}(a)} \frac{\partial f_{\vartheta^*}(a)}{\partial a} + (1 - p) \frac{\partial \phi(f_{\vartheta^*}(1 + a))}{\partial f_{\vartheta^*}(1 + a)} \frac{\partial f_{\vartheta^*}(1 + a)}{\partial (1 + a)}; \\ \frac{\partial J(a, p)}{\partial p} &= \frac{\partial J_D(a, p; \vartheta^*)}{\partial p} = \phi(f_{\vartheta^*}(a)) - \phi(f_{\vartheta^*}(1 + a)). \end{aligned} \tag{30}$$

Because there is no constraint / regularization, we ignore the term $\lambda^* \cdot \frac{\partial s(a, p; \vartheta^*)}{\partial p}$.

The $\frac{\partial J(a,p)}{\partial p}$ means that: if $a = 0$, which leads to well-defined $f_{\vartheta^*}(x)$, p has a well-defined gradient; if $a \neq 0$, then the gradient of p is exceptional. However, evidenced by $\frac{\partial J(a,p)}{\partial a}$, its gradient for a is always undefined, because $\frac{\partial f_{\vartheta^*}(x)}{\partial x}$ 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, i.e., envelope theorem is actually inapplicable to the setting.
- The well-defined gradient with respect to density / 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, 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 or f, in practice, also can not pass to p. Because it needs to pass via $\frac{\partial f_{\vartheta^*}(x)}{\partial x} \cdot \frac{\partial x}{\partial p}$ and $\frac{\partial f_{\vartheta^*}(x)}{\partial x}$ 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 D or f, playing the role as a distance metric for the real and fake distribution, 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 $\frac{\partial f_{\vartheta^*}(x)}{\partial x}$. The above two combined together makes GANs sometime is theoretically not optimizable.

So, given the fact the current GANs formulation is sample-based and the gradient must be passed via $\frac{\partial f_{\vartheta^*}(x)}{\partial x}$, 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 a well-defined gradient for the parameters that change the location of samples. However, the underlying objective of J is convex with respect to $\mathcal{P}_g$ does not imply the model of J is convex with respect to θ. And as a matter of fact, we have already known that the gradients from the J with respect to samples in unregularized GANs only reflect the local information and tend to lead to model collapse. So, clearly, well-defined gradients or optimizable is not the sufficient condition for convergence. The key should lie in sample-based optimization and maybe because of the big gap between sample-based optimization and density based distance metric.

8.3 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 analyse 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 $\mathcal{P}_g$ be a delta distribution at θ in $\mathbb{R}$ with $\theta < 1$, while $\mathcal{P}_r$ is a delta distribution at 1. $J_D(\vartheta, \theta) = f_\vartheta(1) - f_\vartheta(\theta)$ and the constraint is $f_\vartheta(1) - f_\vartheta(\theta) - (1 - \theta) \leq 0$. We know that for the optimal $f_{\vartheta^*}(x)$, it has $f_{\vartheta^*}(\theta) = f_{\vartheta^*}(1) - 1 + \theta$. Due to the free offset property of Wasserstein distance, without loss of generality, we further assume $f(1) = 1$. Then the problem is simplified as: $J_D(\theta, \vartheta) = 1 - f_\vartheta(\theta)$ with the regularization $f_\vartheta(\theta) - \theta \leq 0$.

Note that, $f_{\vartheta^*}(\theta)$ is only necessarily defined on θ and 1, and $\frac{\partial f_{\vartheta^*}(\theta)}{\partial x}$ is also undefined for all sample points.

The Lagrangian dual problem is given by

$$L(\theta, \vartheta, \lambda) = 1 - f_\vartheta(\theta) + \lambda \cdot (f_\vartheta(\theta) - \theta)). \tag{31}$$

From the envelope theorem, we have

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial L(\theta;\vartheta^*,\lambda^*)}{\partial \theta} = -\frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta} + \lambda^* \cdot \frac{\partial (f_{\vartheta^*}(\theta)-\theta)}{\partial \theta} \\ &= -\frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta} + (\lambda^* \cdot \frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta} - \lambda^*) = (\lambda^*-1)\cdot \frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta} - \lambda^*.\end{aligned} \tag{32}$$

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

$$\begin{aligned}\frac{\partial L}{\partial \vartheta^*} &= (\lambda^*-1)\frac{\partial f_{\vartheta^*}(\theta)}{\partial \vartheta^*} = 0, \\ \frac{\partial L}{\partial \lambda^*} &= f_{\vartheta^*}(\theta) - \theta = 0,\end{aligned} \tag{33}$$

We can notice that $\lambda^* = 1$ is one of its solutions. Applying it to Eq. (32), we get $\frac{\partial J(\theta)}{\partial \theta} = 1$, which is reasonable and true, and more importantly, we notice that the sample gradient $\frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta}$ 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 $\frac{\partial f_{\vartheta^*}(\theta)}{\partial \theta}$ 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 (33), it holds $f_{\vartheta^*}(\theta) = \theta$.

8.4 With Lipschitz Condition or Lipschitz Regularization

WGANs, with Wasserstein distance in KR duality, does not involve the parameters of the generator in the constraint (i.e, the Lipschitz condition) of the optimization problem. 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

$$\begin{aligned}\frac{\partial J(\theta;\vartheta^*)}{\partial \theta} &= \frac{\partial J_D(\theta;\vartheta^*)}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_{z\sim\mathcal{P}_z}[\phi(f_{\vartheta^*}(g_\theta(z)))] + \mathbb{E}_{x\sim\mathcal{P}_r}[\psi(f_{\vartheta^*}(x))]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_{z\sim\mathcal{P}_z}[\phi(f_{\vartheta^*}(g_\theta(z)))]}{\partial \theta}.\end{aligned} \tag{34}$$

With the Lipschitz condition or penalizing the Lipschitz constant, the objective is intuitively continuously differentiable with respect to $\mathcal{P}_g$. If the generative function is continuous and locally Lipschitz with respect to its parameter θ, then the objective should be continuously differentiable with respect to 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 $\frac{\partial f^*_\vartheta(g_\theta(z))}{\partial g_\theta(z)}$. So, if the generator is continuously differentiable with respect to θ, i.e., if $\frac{\partial g_\theta(z)}{\partial \theta}$ is well-defined, then $\frac{\partial \phi(f_{\vartheta^*}(g_\theta(z)))}{\partial \theta}$ and hence Eq. (34) is well-defined and is expected to well behave.

8.5 Sample-Based Distribution Estimation

In unregularized GANs, if $\mathcal{S}_g \cup \mathcal{S}_r$ does not cover the whole input space, $f^*(x)$ would be undefined outside $\mathcal{S}_g \cup \mathcal{S}_r$. 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 $\mathcal{P}_g$ from converging to $\mathcal{P}_r$.

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 any sample point can be directly optimized, while in sample-based distribution estimation, to increase / decrease the density of a certain point, it requires modifying samples from being the support of one probability distribution to another.

This is why cases with totally-overlapped distributions also suffer from the faulty gradient direction. Such a conclusion also reminds us that we need to be cautious when understanding or proving GANs at the distribution level, given GANs is sample-based, with the discriminator requiring a sample as input.

8.5.1 The Choice of Target Point of Generator in Traditional GANs

The notion that GANs is sample-based also explains a weird phenomenon about the optimal target point of the generator in traditional GANs.

Taking the Least-Squares GANs as an example. Note that the generator objective of the Least-Squares GANs is $(x-\gamma)^2$ and the γ derived from the Pearson $\mathcal{X}^2$ divergence is $\frac{\alpha+\beta}{2}$ (e.g. $\alpha=-1$, $\beta=1$, $\gamma=0$). But, in practice, $\gamma=\beta$ usually works better than $\gamma=\frac{\alpha+\beta}{2}$.

When $\mathcal{P}_g(x)=\mathcal{P}_r(x)$, we have $f^*(x)=\frac{\alpha+\beta}{2}$. And $\alpha \le f^*(x) \le \frac{\alpha+\beta}{2}$ means $\mathcal{P}_g(x) \ge \mathcal{P}_r(x)$. If the target γ equals $\frac{\alpha+\beta}{2}$, samples from points where $\mathcal{P}_g(x) \ge \mathcal{P}_r(x)$ cannot adapt to locations where $\mathcal{P}_g(x) \le \mathcal{P}_r(x)$, i.e., where $\frac{\alpha+\beta}{2} \le f^*(x) \le \beta$. As a result, $\mathcal{P}_g$ would never converge to $\mathcal{P}_r$, and γ actually needs to be the same as β to avoid this issue.

However, in the meantime, $\gamma=\beta$ would lead to nonzero gradient scale in terms of $\nabla_{f(x)}\varphi(f(x))$ for each sample[6], even when $\mathcal{P}_g$ converges to $\mathcal{P}_r$.

The arguments above can actually be more general: In sample-based optimization, if $f^*(x)$ is a monotonically increasing function of $\mathcal{P}_r(x)$ and a monotonically decreasing function of $\mathcal{P}_g(x)$, this issue is generally inevitable, including original GANs. Generally speaking, the target point has to equal to the maximum possible value of $f^*(x)$ to avoid the above mentioned problem. However, if doing so, because this value is not equal to the value of $f^*(x)$ when $\mathcal{P}_r(x)=\mathcal{P}_g(x)$, it results into nonzero gradient scale at convergence.

6. The $\nabla_x f(x)$ is undefined for sample that is isolated or at the boundary.

9. Related Work

We have shown 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 introduced to the original GANs (Miyato et al., 2018; Kodali et al., 2017; Fedus et al., 2017), achieving improvements in the quality of generated samples. As a matter of fact, the original GANs objective $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$ is a special case of our LGANs. Thus, our analysis explains why and how it works. Farnia and Tse (2018) also provided some analysis on how f-divergence behaves when combined with Lipschitz. However, their analysis is limited to the symmetric f-divergence.

Fedus et al. (2017) 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 objective somehow works. According to our analysis, given the optimal f^*, the original GANs has no guarantee on its convergence. And we have also provided a reasonable explanation on how these traditional GANs, who does not guarantee its convergence, works in practice.

Unterthiner et al. (2017) provided some arguments on the unreliability of $\nabla_x f^*(x)$ in traditional GANs, which motivates their proposal of Coulomb GANs. However, the arguments there are not thorough. By contrast, we provided a systematic and thorough study over the gradient issues in traditional GANs. And we have also accordingly proposed a new solution, i.e., the Lipschitz GANs, which shall be a strong rival to their proposed Coulomb GANs, with superior efficiency and sample quality.

Some 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., 2017), which is another important direction for theoretically understanding GANs. Despite the fact that the behaviors of suboptimal can be different, we think it should well-behave under the optimum condition in the first place.

Researchers found that applying Lipschitz continuity condition to the generator also benefits the quality of generated samples (Zhang et al., 2018; Odena et al., 2018). And Qi (2017) studied the Lipschitz condition from the perspective of loss-sensitive with a Lipschitz data density assumption. These are actually different branches and not necessarily related. Their discussions are out of the scope of this paper.

There exists generative models that do not use $\nabla_x f^*(x)$ for the primal guide of generator update. For example, Sanjabi et al. (2018) updates the generator according to the optimal transport plan. Currently, 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) requires a batch of sample, simulating the distribution, while Jolicoeur-Martineau (2018) requires simultaneously input one real sample and one fake sample. Our analysis does not directly apply to their models, but the similar spirit, i.e., analysing whether the gradient flaw between G and D is effective, assuming optimal discriminator, can be used to analyse their models.

10. Conclusion

In this paper, we first have studied one fundamental cause of failure in the training of GANs, i.e., the gradient uninformativeness issue. In particular, for generated samples which are not surrounded by real samples, the gradients of the optimal discriminative function $\nabla_x f^*(x)$ tell nothing about $\mathcal{P}_r$. That is, in a sense, there is no guarantee that $\mathcal{P}_g$ will converge to $\mathcal{P}_r$. Typical case is that $\mathcal{P}_r$ and $\mathcal{P}_g$ are disjoint, which is common in practice. The gradient uninformativeness is common for unregularized GANs and also appears in regularized GANs.

To address the nonconvergence problem caused by uninformative $\nabla_x f^*(x)$, we have proposed LGANs and shown that it makes $\nabla_x f^*(x)$ informative in the way that the gradient for each generated sample points towards some real sample. We have also shown that in LGANs, the optimal discriminative function exists and is unique, and the only Nash equilibrium is achieved when $\mathcal{P}_r = \mathcal{P}_g$ where $k(f^*) = 0$. Our experiments showed LGANs lead to more stable discriminative functions and achieve higher sample qualities.

Acknowledgements

This work is sponsored by APEX-YITU Joint Research Program. The authors thank the support of National Natural Science Foundation of China (61702327, 61772333, 61632017), Shanghai Sailing Program (17YF1428200). Zhiming Zhou personally thanks Jiadong Liang and Dachao Lin for a lot of helpful discussions on the central theorems and proofs of LGANs. Zhiming Zhou personally thanks Yuxuan Song and Lantao Yu for their fruitful discussions with me on the initial idea of LGANs. Zhiming Zhou personally thanks Hongwei Wang and Weinan Zhang for their suggestions and helps on the writing and presentation. Zhiming Zhou personally thanks Yong Yu for his unreserved support all these years. Zhihua Zhang has been supported by Beijing Municipal Commission of Science and Technology under Grant No. 181100008918005 and by Beijing Academy of Artificial Intelligence (BAAI).

Appendix A. Proofs

A.1 Proof of Theorem 1

Let X, Y be two random vectors such that $X \sim \mathcal{P}_g, Y \sim \mathcal{P}_r$. Assume $\mathbb{E}_{X\sim\mathcal{P}_g}\|X\| < \infty$ and $\mathbb{E}_{Y\sim\mathcal{P}_r}\|Y\| < \infty$. Let $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y))$. Let $\|f\|_{Lip}$ denote the Lipschitz constant of f. Let $\mathcal{S}_r$ and $\mathcal{S}_g$ denote the supports of $\mathcal{P}_r$ and $\mathcal{P}_g$, respectively. Let $W_1(\mathcal{P}_r, \mathcal{P}_g)$ denote the 1-st Wasserstein distance between $\mathcal{P}_r$ and $\mathcal{P}_g$.

Lemma 1. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. Assume f is subject to $\|f\|_{Lip} \le k$. If there is $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$, then we have a lower bound for $\mathfrak{G}(f)$.*

Proof Given that ϕ, φ are convex functions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) \\
&\ge \mathbb{E}_{X\sim\mathcal{P}_g}(\phi'(a_0)(f(x)-a_0)+\phi(a_0)) + \mathbb{E}_{Y\sim\mathcal{P}_r}(\varphi'(a_0)(f(x)-a_0)+\varphi(a_0)) \\
&= \phi'(a_0)\mathbb{E}_{X\sim\mathcal{P}_g}f(x) + \varphi'(a_0)\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) + C \\
&= (\phi'(a_0)+\varphi'(a_0))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + \varphi'(a_0)(\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}f(X)) + C \\
&= k\varphi'(a_0)(\mathbb{E}_{Y\sim\mathcal{P}_r}\frac{1}{k}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}\frac{1}{k}f(X)) + C \\
&\ge -k\varphi'(a_0)W_1(\mathcal{P}_r, \mathcal{P}_g) + C.
\end{aligned} \tag{35}$$

Therefore, we get the lower bound.

Lemma 2. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. Assume f is subject to $\|f\|_{Lip} \le k$.*

- *If there exists $a_1 \in \mathbb{R}$ such that $\phi'(a_1) + \varphi'(a_1) > 0$, then we have: if $f(0) \to +\infty$, then $\mathfrak{G}(f) \to +\infty$;*
- *If there exists $a_2 \in \mathbb{R}$ such that $\phi'(a_2) + \varphi'(a_2) < 0$, then we have: if $f(0) \to -\infty$, then $\mathfrak{G}(f) \to +\infty$.*

Proof Since ϕ, φ are convex functions, we have

$$\begin{aligned}
\mathfrak{G}(f) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) \\
&\ge \mathbb{E}_{X\sim\mathcal{P}_g}(\phi'(a_1)(f(x)-a_1)+\phi(a_1)) + \mathbb{E}_{Y\sim\mathcal{P}_r}(\varphi'(a_1)(f(x)-a_1)+\varphi(a_1)) \\
&= \phi'(a_1)\mathbb{E}_{X\sim\mathcal{P}_g}f(x) + \varphi'(a_1)\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) + C_1 \\
&= (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + \varphi'(a_1)(\mathbb{E}_{Y\sim\mathcal{P}_r}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}f(X)) + C_1 \\
&= (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) + k\varphi'(a_1)(\mathbb{E}_{Y\sim\mathcal{P}_r}\frac{1}{k}f(Y) - \mathbb{E}_{X\sim\mathcal{P}_g}\frac{1}{k}f(X)) + C_1 \\
&\ge (\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}f(X) - k\varphi'(a_1)W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1 \\
&\ge (\phi'(a_1)+\varphi'(a_1))f(0) - k(\phi'(a_1)+\varphi'(a_1))\mathbb{E}_{X\sim\mathcal{P}_g}\|X\| - k\varphi' W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1.
\end{aligned} \tag{36}$$

Thus, if $f(0) \to +\infty$, then $\mathfrak{G}(f) \to +\infty$. And we can prove the other case symmetrically.

Lemma 3. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. If ϕ and φ satisfy the following properties:*

- $\phi' \geq 0, \varphi' \leq 0$;
- *There exist $a_0, a_1, a_2 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0, \phi'(a_1) + \varphi'(a_1) > 0, \phi'(a_2) + \varphi'(a_2) < 0$.*

Then we have $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_r}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_g}\varphi(f(Y))$, where f is subject to $\|f\|_{Lip} \leq k$, has global minima.

That is, $\exists f^*,\ s.t.$

- $\|f^*\|_{Lip} \leq k$;
- $\forall f\ s.t.\ \|f\|_{Lip} \leq k$, *we have* $\mathfrak{G}(f^*) \leq \mathfrak{G}(f)$.

Proof According to Lemma 1, $\mathfrak{G}(f)$ has a lower bound, which means $inf(\mathfrak{G}(f)) > -\infty$. Thus we can get a series of functions $\{f_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty}\mathfrak{G}(f_n) = inf(\mathfrak{G}(f))$. Suppose that $\{r_i\}_{i=1}^{\infty}$ is the sequence of all rational points in $dom(f)$. Due to Lemma 2, for any $x \in \mathbb{R}$, $\{f_n(x)|n \in \mathbb{R}\}$ is bounded. By Bolzano-Weierstrass theorem, there is a subsequence $\{f_{1n}\} \subseteq \{f_n\}$ such that $\{f_{1n}(r_1)\}_{n=1}^{\infty}$ converges. And there is a subsequence $\{f_{2n}\} \subseteq \{f_{1n}\}$ such that $\{f_{2n}(r_2)\}_{n=1}^{\infty}$ converges. As for r_i, there is a subsequence $\{f_{in}\} \subseteq \{f_{i-1n}\}$ such that $\{f_{in}(r_i)\}_{n=1}^{\infty}$ converges. Then the sequence $\{f_{nn}\}_{n=1}^{\infty}$ will converge at r_i.

Furthermore, for all $x \in dom(f)$, we claim that $\{f_{nn}\}_{n=1}^{\infty}$ converges at x. Actually, $\forall\epsilon > 0$, find $r \in \{r_i\}$ such that $\|x - r\| \leq \frac{\epsilon}{10k}$, we have

$$\begin{aligned}\lim_{m,l\to\infty}|f_{mm}(x) - f_{ll}(x)| &\leq \lim_{m,l\to\infty}(|f_{mm}(x) - f_{mm}(r)| + |f_{mm}(r) - f_{ll}(r)| + |f_{ll}(r) - f_{ll}(x)|) \\ &\leq \lim_{m,l\to\infty}(\frac{\epsilon}{10} + \frac{\epsilon}{10} + |f_{mm}(r) - f_{ll}(r)|) = \frac{\epsilon}{5}\end{aligned} \tag{37}$$

Let $\epsilon \to 0$, then we get $\lim_{m,l\to\infty}|f_{mm}(x) - f_{ll}(x)| = 0$.

We denote $\{f_{nn}\}_{n=1}^{\infty}$ as $\{g_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ converges to g. Due to Lemma 2, we know that $\exists C'$ such that $|g_n(0)| \leq C',\ \forall n \in \mathbb{N}$. Because $\phi' \geq 0, \varphi' \leq 0$, we have

$$\phi(g_n(x)) \geq \phi(g_n(0) - k\|x\|) \geq \phi(-C' - k\|x\|) \geq \phi'(a_0)(-C' - k\|x\| - a_0) + \phi(a_0) = -k\phi'(a_0)\|x\| + C'' \tag{38}$$

That is, $\phi(g_n(x)) + k\phi'(a_0)\|x\| - C'' \geq 0$.

By Fatou's Lemma,

$$\begin{aligned}\mathbb{E}_{X\sim\mathcal{P}_g}(\phi(g(X)) + k\phi'(a_0)\|X\| - C'') &= \mathbb{E}_{X\sim\mathcal{P}_g}\varliminf_{n\to\infty}(\phi(g_n(X)) + k\phi'(a_0)\|X\| - C'') \\ &\leq \varliminf_{n\to\infty}\mathbb{E}_{X\sim\mathcal{P}_g}(\phi(g_n(X)) + k\phi'(a_0)\|X\| - C'') \\ &= \varliminf_{n\to\infty}\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \mathbb{E}_{X\sim\mathcal{P}_g}(k\phi'(a_0)\|X\| - C'')\end{aligned} \tag{39}$$

It means $\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g(X)) \le \underline{\lim}_{n\to\infty}\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X))$. Similarly, we have $\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g(Y)) \le \underline{\lim}_{n\to\infty}\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y))$. Combining the two inequalities, we have

$$\begin{aligned}\mathfrak{G}(g) &= \mathbb{E}_{X\sim\mathcal{P}_g}\phi(g(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g(Y)) \le \varliminf_{n\to\infty}\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \varliminf_{n\to\infty}\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y)) \\ &\le \varliminf_{n\to\infty}\left(\mathbb{E}_{X\sim\mathcal{P}_g}\phi(g_n(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(g_n(Y))\right) = \inf_{\|f\|_{Lip}\le k}\mathfrak{G}(f)\end{aligned} \tag{40}$$

Note that for any $x, y \in dom(g)$, $|g(x)-g(y)| \le \lim_{n\to\infty}(|g(x)-g_n(x)| + |g_n(x)-g_n(y)| + |g_n(y)-g(y)|) \le k\|x-y\|$. That is, $\|g\|_{Lip} \le k$, $\mathfrak{G}(g) = \inf_{\|f\|_{Lip}\le k}\mathfrak{G}(f)$.

Lemma 4 (Wasserstein distance)**.** *$\mathfrak{T}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}f(X) - \mathbb{E}_{Y\sim\mathcal{P}_r}f(Y)$, where f is subject to $\|f\|_{Lip} \le k$, 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 $\|f_n\|_{Lip} \le k$, 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 3 and find the optimal function f^* such that $\mathfrak{T}(f^*) = \inf_{\|f\|_{Lip}\le k}\mathfrak{T}(f)$.

Lemma 5. *Let ϕ and φ be two convex functions, whose domains are both $\mathbb{R}$. If we further suppose that the support sets $\mathcal{S}_r$ and $\mathcal{S}_g$ are bounded. Then if ϕ and φ satisfy the following properties:*

- *$\phi' \ge 0, \varphi' \le 0$;*
- *There is $a_0 \in \mathbb{R}$ such that $\phi'(a_0) + \varphi'(a_0) = 0$.*

We have $\mathfrak{G}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y))$, where f is subject to $\|f\|_{Lip} \le k$, has global minima.

That is, $\exists f^*$, s.t.

- *$\|f^*\|_{Lip} \le k$*
- *$\forall f$ s.t. $\|f\|_{Lip} \le k$, we have $\mathfrak{G}(f^*) \le \mathfrak{G}(f)$.*

Proof We have proved most conditions in previous lemmas. And we only have to consider the condition that for any $x \in \mathbb{R}$, $\phi'(x) + \varphi'(x) \ge 0$ (or $\phi'(x) + \varphi'(x) \le 0$) and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$ (or $\phi'(a_1) + \varphi'(a_1) < 0$).

Without loss of generality, we assume that $\phi'(x) + \varphi'(x) \ge 0$ for all x and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$. Then we know $\forall x \le a_0,\ \phi'(x) + \varphi'(x) = 0$, which leads to $\forall x \le a_0,\ \phi'(x) = -\varphi'(x)$. Thus, for any $x \le a_0$, $0 \le \phi''(x) = -\varphi''(x) \le 0$, which means $\forall x \le a_0,\ \phi(x) = -\varphi(x) = tx,\ t \ge 0$. 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. In fact, it is not difficult to find $f_n(0) \le C$ with Lemma 2. On the other hand, when $C > k \cdot diam(\mathcal{S}_r \cup \mathcal{S}_g) + a_0$, then: if $f(0) < -C$, we have $f(X) < a_0$ for all $X \in \mathcal{S}_r \cup \mathcal{S}_g$. In this case, $\mathfrak{G}(f) = \mathfrak{G}(f - f(0) - C)$.

This is the reason we can assume $f_n(0) \in [-C, C]$. Because $\|f_n\|_{Lip} \leq k$, we can assert 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 3 and find the optimal function f^* such that $\mathfrak{G}(f^*) = \inf_{\|f\|_{Lip} \leq k} \mathfrak{G}(f)$.

Lemma 6 (Theorem 1 Part I)**.** *Under the same assumption of Lemma 5, we have* $\mathfrak{F}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) + \lambda\|f\|^{\alpha}_{Lip}$ *with* $\lambda > 0$ *and* $\alpha > 1$ *has global minima.*

Proof When $\|f\|_{Lip} = \infty$, it is trivial that $\mathfrak{F}(f) = \infty$. And when $\|f\|_{Lip} < \infty$, combining Lemma 1, we have $\mathfrak{F}(f) = \mathfrak{G}(f) + \lambda\|f\|^{\alpha}_{Lip} \geq -\|f\|_{Lip}\varphi'(a_0)W_1(\mathcal{P}_r, \mathcal{P}_g) + \lambda\|f\|^{\alpha}_{Lip}$. When $\lambda > 0$ and $\alpha > 1$, the right term is a convex function about $\|f\|_{Lip}$, 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 $\|f_n\|_{Lip} \leq C$ for all f_n. Then it is not difficult to show for any point x, $\{f_n(x)\}$ is bounded. So we can imitate the method used in main theorem to find the sequence $\{g_n\}$ such that $\{g_n\} \subseteq \{f_n\}$ and $\{g_n\}_{n=1}^{\infty}$ converge at every point x. Suppose $\lim_{n\to\infty} g_n = g$, then by Fatou's Lemma, we have $\mathfrak{G}(g) \leq \underline{\lim}_{n\to\infty}\mathfrak{G}(g_n)$.

Next, We prove that $\|g\|_{Lip} \leq \underline{\lim}_{n\to\infty}\|g_n\|_{Lip}$. If the claim holds, then $\mathfrak{F}(g) = \mathfrak{G}(g) + \lambda\|g\|^{\alpha}_{Lip} \leq \underline{\lim}_{n\to\infty}\mathfrak{G}(g_n) + \underline{\lim}_{n\to\infty}\lambda\|g_n\|^{\alpha}_{Lip} \leq \underline{\lim}_{n\to\infty}(\mathfrak{G}(g_n) + \lambda\|g_n\|^{\alpha}_{Lip}) = \inf\mathfrak{F}(f)$. Thus, the global minima exists. In fact, if $\|g\|_{Lip} > \underline{\lim}_{n\to\infty}\|g_n\|_{Lip}$, then there exist x, y such that $\frac{|g(x)-g(y)|}{\|x-y\|} \geq \underline{\lim}_{n\to\infty}\|g_n\|_{Lip} + \epsilon \geq \underline{\lim}_{n\to\infty}\frac{|g_n(x)-g_n(y)|}{\|x-y\|} + \epsilon$. i.e. $|g(x) - g(y)| \geq \underline{\lim}_{n\to\infty}|g_n(x)-g_n(y)|+\epsilon\|x-y\| = |g(x)-g(y)|+\epsilon\|x-y\| > |g(x)-g(y)|$. The contradiction tells us that $\|g\|_{Lip} \leq \underline{\lim}_{n\to\infty}\|g_n\|_{Lip}$.

Lemma 7 (Theorem 1 Part II)**.** *Let* ϕ *and* φ *be two convex functions, whose domains are both* $\mathbb{R}$*. If* ϕ *or* φ *is strictly convex, then the minimizer of* $\mathfrak{F}(f) = \mathbb{E}_{X\sim\mathcal{P}_g}\phi(f(X)) + \mathbb{E}_{Y\sim\mathcal{P}_r}\varphi(f(Y)) + \lambda\|f\|^{\alpha}_{Lip}$ *with* $\lambda > 0$ *and* $\alpha > 1$ *is unique (in the support of* $\mathcal{S}_r \cup \mathcal{S}_g$*).*

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

$$\begin{aligned}\Big\|\frac{f_1+f_2}{2}\Big\|_{Lip} &= \sup_{x,y}\frac{\frac{f_1(x)+f_2(x)}{2} - \frac{f_1(y)+f_2(y)}{2}}{\|x-y\|}\\ &\leq \sup_{x,y}\frac{1}{2}\frac{|f_1(x)-f_1(y)| + |f_2(x)-f_2(y)|}{\|x-y\|}\\ &\leq \frac{1}{2}\Big(\sup_{x,y}\frac{|f_1(x)-f_1(y)|}{\|x-y\|} + \sup_{x,y}\frac{|f_2(x)-f_2(y)|}{\|x-y\|}\Big)\\ &= \frac{1}{2}(\|f_1\|_{Lip} + \|f_2\|_{Lip}).\end{aligned} \tag{41}$$

And given $\lambda > 0$ and $\alpha > 1$, we further have

$$\begin{aligned}\lambda\Big\|\frac{f_1+f_2}{2}\Big\|^{\alpha}_{Lip} &\leq \lambda\Big(\frac{1}{2}(\|f_1\|_{Lip} + \|f_2\|_{Lip})\Big)^{\alpha}\\ &\leq \lambda\frac{1}{2}(\|f_1\|^{\alpha}_{Lip} + \|f_2\|^{\alpha}_{Lip}).\end{aligned} \tag{42}$$

Let $\mathfrak{F}(f_1)=\mathfrak{F}(f_2)=\inf\mathfrak{F}(f)$. Then we have

$$\begin{aligned}\mathfrak{G}\Big(\frac{f_1+f_2}{2}\Big)&=\mathbb{E}_{X\sim\mathcal{P}_g}\phi\Big(\frac{f_1+f_2}{2}\Big)+\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi\Big(\frac{f_1+f_2}{2}\Big)+\lambda\Big\|\frac{f_1+f_2}{2}\Big\|^\alpha_{Lip}\\&<\mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big)+\mathbb{E}_{Y\sim\mathcal{P}_r}\varphi\Big(\frac{f_1+f_2}{2}\Big)+\lambda\Big\|\frac{f_1+f_2}{2}\Big\|^\alpha_{Lip}\\&\le\mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big)+\mathbb{E}_{Y\sim\mathcal{P}_r}\Big(\frac{\varphi(f_1)+\varphi(f_2)}{2}\Big)+\lambda\Big\|\frac{f_1+f_2}{2}\Big\|^\alpha_{Lip}\\&\le\mathbb{E}_{X\sim\mathcal{P}_g}\Big(\frac{\phi(f_1)+\phi(f_2)}{2}\Big)+\mathbb{E}_{Y\sim\mathcal{P}_r}\Big(\frac{\varphi(f_1)+\varphi(f_2)}{2}\Big)+\lambda\frac{1}{2}(\|f_1\|^\alpha_{Lip}+\|f_2\|^\alpha_{Lip})\\&=\frac{1}{2}(\mathfrak{G}(f_1)+\mathfrak{G}(f_2))=\inf\mathfrak{G}(f)\end{aligned}\tag{43}$$

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

A.2 Proof of Theorem 2

Let $J_D=\mathbb{E}_{x\sim\mathcal{P}_g}[\phi(f(x))]+\mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(f(x))]$. Let $\mathring{J}_D(x)=\mathcal{P}_g(x)\phi(f(x))+\mathcal{P}_r(x)\varphi(f(x))$. Clearly, $J_D=\int_{\mathbb{R}^n}\mathring{J}_D(x)dx$. Let $J^*_D(k)=\min_{f\in\mathcal{F}_{\mathbf{k}\text{-Lip}}}J_D=\min_{f\in\mathcal{F}_{1\text{-Lip}},b}\mathbb{E}_{x\sim\mathcal{P}_g}[\phi(k\cdot f(x)+b)]+\mathbb{E}_{x\sim\mathcal{P}_r}[\varphi(k\cdot f(x)+b)]$.

Let $k(f)$ denote the Lipschitz constant of f. Define $J=J_D+\lambda\cdot k(f)^2$ and $f^*=\arg\min_f[J_D+\lambda\cdot k(f)^2]$.

Lemma 8. *It holds $\frac{\partial\mathring{J}_D(x)}{\partial f^*(x)}=0$ for all x, if and only if, $k(f^*)=0$.*

Proof

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

For the optimal f^*, it holds that $\frac{\partial J}{\partial k(f^*)}=\frac{\partial J^*_D}{\partial k(f^*)}+2\lambda\cdot k(f^*)=0$.

$\frac{\partial\mathring{J}_D(x)}{\partial f^*(x)}=0$ for all x implies $\frac{\partial J^*_D}{\partial k(f^*)}=0$. Thus we conclude that $k(f^*)=0$.

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

For the optimal f^*, it holds that $\frac{\partial J}{\partial k(f^*)}=\frac{\partial J^*_D}{\partial k(f^*)}+2\lambda\cdot k(f^*)=0$.

$k(f^*)=0$ implies $\frac{\partial J^*_D}{\partial k(f^*)}=0$. $k(f^*)=0$ also implies $\forall x,y,f^*(x)=f^*(y)$.

Given $\forall x,y,f^*(x)=f^*(y)$, if there exists some point x such that $\frac{\partial\mathring{J}_D(x)}{\partial f^*(x)}\neq0$, then it is obvious that $\frac{\partial J^*_D}{\partial k(f^*)}\neq0$.

It is contradictory to $\frac{\partial J^*_D}{\partial k(f^*)}=0$. Thus we have $\forall x$, $\frac{\partial\mathring{J}_D(x)}{\partial f^*(x)}=0$.

Lemma 9. *If $\forall x,y,f^*(x)=f^*(y)$, then $\mathcal{P}_r=\mathcal{P}_g$.*

Proof $\forall x, y, f^*(x) = f^*(y)$ implies $k(f^*) = 0$. According to Lemma 8, for all x it holds $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = 0$, i.e., $\mathcal{P}_g(x)\frac{\partial \phi(f^*(x))}{\partial f^*(x)} + \mathcal{P}_r(x)\frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = 0$. Thus, $\frac{\mathcal{P}_g(x)}{\mathcal{P}_r(x)} = -\frac{\frac{\partial \varphi(f^*(x))}{\partial f^*(x)}}{\frac{\partial \phi(f^*(x))}{\partial f^*(x)}}$. That is, $\frac{\mathcal{P}_g(x)}{\mathcal{P}_r(x)}$ has a constant value, which straightforwardly implies $\mathcal{P}_r = \mathcal{P}_g$.

Proof [**Proof of Theorem 2**]

(a): Let k be the Lipschitz constant of f^*. Consider x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$. Define $k(x) = \sup_y \frac{|f(y)-f(x)|}{\|y-x\|}$.

(i) If $\forall \delta$ s.t. $\forall \epsilon$ there exist $z, w \in B(x, \epsilon)$ such that $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} \geq k - \delta$, which means there exists t such that $f'(t) \geq k - \delta$, because $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} = \frac{\int_w^z f^{*\prime}(t)dt}{\|z-w\|}$. Let $\epsilon \to 0$, we have $t \to x$. Then $|f^{*\prime}(t)| \to |f^{*\prime}(x)|$. Let $\delta \to 0$, we have $(k - \delta) \to k$. Assume f^* is smooth, we have that $|f'(x)| = k$, which means there exists a y such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(ii) Assume that $\exists \delta$ s.t. $\exists \epsilon$ and for all $z, w \in B(x, \epsilon)$, $\frac{|f^*(z)-f^*(w)|}{\|z-w\|} < k - \delta$. Consider the following condition, for all δ_2 and $\epsilon_2 \in (0, \epsilon/2)$, $\exists y \in B(x, \epsilon_2)$, such that $k(y) > k - \delta_2$. 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 $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$, without loss of generality, we can assume $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$. Then, for all $y \in B(x, \epsilon_2)$, we have $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} > 0$, 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 $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} > 0$, $\forall y \in B(x, \epsilon_2)$, 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 exists a y such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(b): For $x \in \mathcal{S}_r \cup \mathcal{S}_g - \mathcal{S}_r \cap \mathcal{S}_g$, assuming $\mathcal{P}_g(x) \neq 0$ and $\mathcal{P}_r(x) = 0$, we have $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f^*(x))}{\partial f^*(x)} + \mathcal{P}_r(x)\frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f^*(x))}{\partial f^*(x)} > 0$, because $\mathcal{P}_g(x) > 0$ and $\frac{\partial \phi(f^*(x))}{\partial f^*(x)} > 0$.

Then according to (a), there must exist a y such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$. The other situation can be proved in the same way.

(c): According to Lemma 9, in the situation that $\mathcal{P}_r \neq \mathcal{P}_g$, 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 $k(f^*) > 0$. Then according to Lemma 8, there exists a point x such that $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} \neq 0$. According to (a), there exists y with $y \neq x$ satisfying that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$.

(d): In Nash equilibrium state, it holds that, for any $x \in \mathcal{S}_r \cup \mathcal{S}_g$, $\frac{\partial J}{\partial k(f)} = \frac{\partial J_D^*}{\partial k(f)} + 2\lambda \cdot k(f) = 0$ and $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} \frac{\partial f(x)}{\partial x} = 0$. We claim that in the Nash equilibrium state, the Lipschitz constant $k(f)$ must be 0. If $k(f) \neq 0$, according to Lemma 8, there must exist a point $\hat{x}$ such that $\frac{\partial \mathring{J}_D(\hat{x})}{\partial f(\hat{x})} \neq 0$. And according to (a), it must hold that $\exists \hat{y}$ fitting $|f(\hat{y}) - f(\hat{x})| = k(f) \cdot \|\hat{x} - \hat{y}\|$. According to Theorem 4, we have $\left\| \frac{\partial f(\hat{x})}{\partial \hat{x}} \right\| = k(f) \neq 0$. This is contradictory to that $\frac{\partial \mathring{J}_D(\hat{x})}{\partial f(\hat{x})} \frac{\partial f(\hat{x})}{\partial \hat{x}} = 0$. Thus $k(f) = 0$. That is, $\forall x \in \mathcal{S}_r \cup \mathcal{S}_g$, $\frac{\partial f(x)}{\partial x} = 0$, which means $\forall x, y, f(x) = f(y)$. According to Lemma 9, $\forall x, y, f(x) = f(y)$ implies $\mathcal{P}_r = \mathcal{P}_g$. Thus $\mathcal{P}_r = \mathcal{P}_g$ is the only Nash equilibrium in our system.

Remark 1. *For the Wasserstein distance, $\nabla_{f^*(x)} \mathring{J}_D(x) = 0$ if and only if $\mathcal{P}_r(x) = \mathcal{P}_g(x)$. For the Wasserstein distance, penalizing the Lipschitz constant also benefits: at the convergence state, it will hold $\frac{\partial f^*(x)}{\partial x} = 0$ for all x.*

A.3 Proof of Theorem 3

Lemma 10. *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 11. *For any x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$, there exists a y with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$.*

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

Proof Consider x with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$. According to Theorem 2, there exists y such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$. Assume that for every y that holds $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$, it has $\frac{\partial \mathring{J}_D(y)}{\partial f^*(y)} \geq 0$. Consider the set $S(x) = \{y \mid f^*(y) - f^*(x) = k(f^*)\|y - x\|\}$. Note that, according to Lemma 10, any z that holds $f^*(z) - f^*(y) = k(f^*)\|z - y\|$ for any $y \in S(x)$ will also be in $S(x)$. Similar to the proof of (a) in Theorem 2, 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 $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$.

Given the fact $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$ and $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$, we can conclude that $f^*(y) > f^*(x)$ and $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. Otherwise, if $f^*(x) - f^*(y) = k(f^*)\|y - x\|$, 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 12. *For any x, if $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} > 0$, then $\mathcal{P}_g(x) > 0$. For any y, if $\frac{\partial \mathring{J}_D(y)}{\partial f(y)} < 0$, then $\mathcal{P}_r(y) > 0$.*

Proof $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} = \mathcal{P}_g(x)\frac{\partial \phi(f(x))}{\partial f(x)} + \mathcal{P}_r(x)\frac{\partial \varphi(f(x))}{\partial f(x)}$. And we know $\phi'(x) > 0$ and $\varphi'(x) < 0$. Naturally, $\frac{\partial \mathring{J}_D(x)}{\partial f(x)} > 0$ implies $\mathcal{P}_g(x) > 0$. Similarly, $\frac{\partial \mathring{J}_D(y)}{\partial f(y)} < 0$ implies $\mathcal{P}_r(y) > 0$.

Proof [**Proof of Theorem 3**]

For any $x \in \mathcal{S}_g$, if $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} > 0$, according to Lemma 11, there exists a y with $\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. According to Lemma 12, we have $\mathcal{P}_r(y) > 0$. That is, there is a $y \in \mathcal{S}_r$ such that $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. We can prove the other case symmetrically.

Remark 2. *$\frac{\partial \mathring{J}_D(x)}{\partial f^*(x)} < 0$ for some $x \in \mathcal{S}_g$ means x is at the overlapping region of $\mathcal{S}_r$ and $\mathcal{S}_g$. It can be regarded as a $y \in \mathcal{S}_r$, and one can apply the other rule which guarantees that there exists an $x' \in \mathcal{S}_g$ that bounds this point.*

A.4 Proof of Theorem 4 and the Necessity of Euclidean Distance

In this section, we will prove Theorem 4, 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 13. *If $f(x)$ is k-Lipschitz with respect to $\|.\|_p$ 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

$$\begin{aligned} f(y) - f(x) &= f(y) - f(x_t) + f(x_t) - f(x) \\ &\leq f(y) - f(x_t) + k\|x_t - x\|_p = f(y) - f(x_t) + t \cdot k\|y - x\|_p \\ &\leq k\|y - x_t\|_p + t \cdot k\|y - x\|_p = k \cdot (1 - t)\|y - x\|_p + t \cdot k\|y - x\|_p \\ &= k\|y - x\|_p. \end{aligned} \tag{44}$$

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

Lemma 14. *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 $\frac{\partial f(x_t)}{\partial v}$ equals k.*

Proof

$$\frac{\partial f(x_t)}{\partial v} = \lim_{h\to 0}\frac{f(x_t+hv)-f(x_t)}{h} = \lim_{h\to 0}\frac{f(x_t+h\frac{y-x}{\|y-x\|_2})-f(x_t)}{h}$$
$$= \lim_{h\to 0}\frac{f(x_{t+\frac{h}{\|y-x\|_2}})-f(x_t)}{h} = \lim_{h\to 0}\frac{\frac{h}{\|y-x\|_2}\cdot k\|y-x\|_2}{h} = k.$$

Proof [**Proof of Theorem 4**] 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 $\|\nabla f(x_t)\|_2 \le k$. Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. We have

$$k^2 = k\frac{\partial f(x_t)}{\partial v} = k\langle v, \nabla f(x_t)\rangle = \langle kv, \nabla f(x_t)\rangle \le \|kv\|_2\|\nabla f(x_t)\|_2 = k^2. \tag{45}$$

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 $\|\nabla f(x_t)\|_2 \le k$, 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., $\|\nabla f\|_q \le k$. 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 $g(x_1,y_1)-g(x_2,y_2) = (x_1-x_2)+(y_1-y_2) \le |x_1-x_2|+|y_1-y_2| = \|(x_1,y_1)-(x_2,y_2)\|_1$, 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 $\|\nabla g\|_\infty \le 1$; 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.

A.5 Proof of the Compact Dual Form of Wasserstein Distance

We here provide a proof for our new dual form of Wasserstein distance, i.e., Eq. (4).

Theorem 7. *Given $W_{KR}(\mathcal{P}_g,\mathcal{P}_r) = W_1(\mathcal{P}_g,\mathcal{P}_r)$, we have $W_{KR}(\mathcal{P}_g,\mathcal{P}_r) = W_{LL}(\mathcal{P}_g,\mathcal{P}_r) = W_1(\mathcal{P}_g,\mathcal{P}_r)$.*

Proof

(i) For any f that satisfies "$f(x) - f(y) \leq d(x, y),\ \forall x, \forall y$", it must satisfy "$f(x) - f(y) \leq d(x, y),\ \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g$".

Thus, $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) \leq W_{LL}(\mathcal{P}_r, \mathcal{P}_g)$.

(ii) Let $F_{LL} = \{f |\, f(x) - f(y) \leq d(x, y),\ \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g\}$.

Let $A = \{(x, y) \,|\, x \in \mathcal{S}_r, y \in \mathcal{S}_g\}$ and $I_A = \begin{cases} 1, & (x, y) \in A; \\ 0, & otherwise \end{cases}$.

Let A^c denote the complementary set of A and define I_{A^c} accordingly.

$\forall \pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g)$, we have the following:

$$\begin{aligned} W_{LL}(\mathcal{P}_r, \mathcal{P}_g) &= \sup_{f \in F_{LL}} \mathbb{E}_{x \sim \mathcal{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g}[f(x)] \\ &= \sup_{f \in F_{LL}} \mathbb{E}_{(x,y) \sim \pi}[f(x) - f(y)] \\ &= \sup_{f \in F_{LL}} \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y))I_A] + \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y))I_{A^c}] \\ &= \sup_{f \in F_{LL}} \mathbb{E}_{(x,y) \sim \pi}[(f(x) - f(y))I_A] \\ &\leq \mathbb{E}_{(x,y) \sim \pi}[\|y - x\| I_A] \\ &\leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)]. \end{aligned}$$

$W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq \mathbb{E}_{(x,y) \sim \pi}[d(x, y)], \forall \pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g)$

$\Rightarrow W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq \inf_{\pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g)} \mathbb{E}_{(x,y) \sim \pi}[d(x, y)] = W_1(\mathcal{P}_r, \mathcal{P}_g)$.

(iii) Combining (i) and (ii), we have $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) \leq W_{LL}(\mathcal{P}_r, \mathcal{P}_g) \leq W_1(\mathcal{P}_r, \mathcal{P}_g)$.

Given $I(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$, we have $I(\mathcal{P}_r, \mathcal{P}_g) = W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$.

Appendix B. Experiment Details and More Results

We use multilayer perception 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 (high dimensional, images) experiments. See the code for the detailed architectures.

We use Adam optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.9$, and the learning rate is 0.0002, which linear decays to zero in $200,000$ iterations. We use 5 discriminator updates per generator update. For all experiments that involve regularization and hence have a hyper-parameter ρ, we search the best regularization weight $\frac{\rho}{2}$ in $[0.01, 0.1, 1.0, 10.0]$.

For all experiments that are aimed for a contrast test, we only change the necessary part(s), saying the objectives or loss metrics or the implementations of Lipschitz regularization, and keep the rest identical or as the same as possible. Please check more details in our codes.

We provide the visual results of LGANs in Figure 16 and Figure 17 for CIFAR-10 and Tiny-ImageNet, respectively. As an extra experiment, we also provide the visual results of LGANs on Oxford 102 in Figure 15. We provide an extra experiment for verifying $\nabla_x f^*(x)$ in LGANs under a more complex setting, where $\mathcal{P}_g$ is a Gaussian distribution, in Figure 14. We plot the training curve of LGANs in terms of Inception Score in Figure 18.

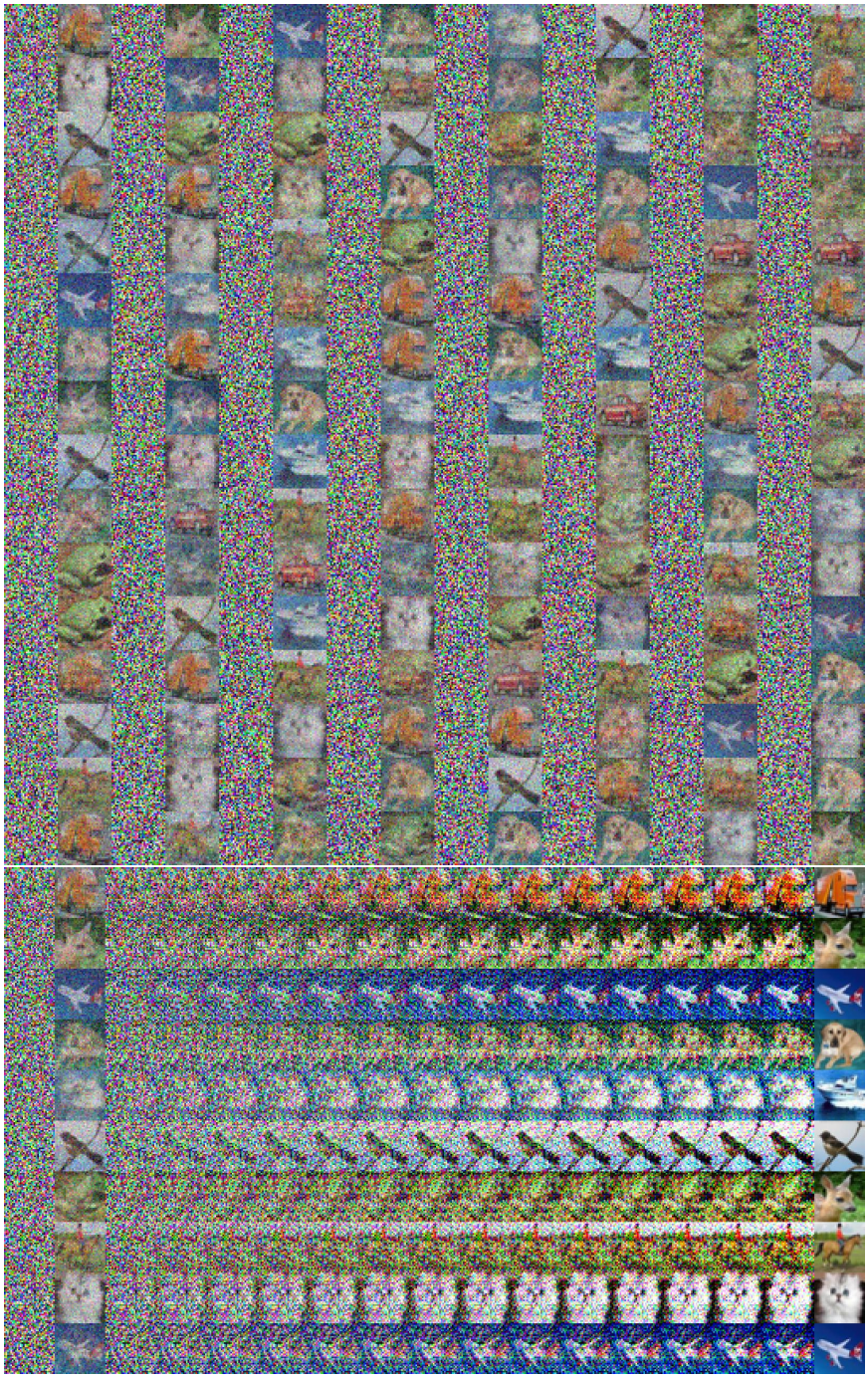
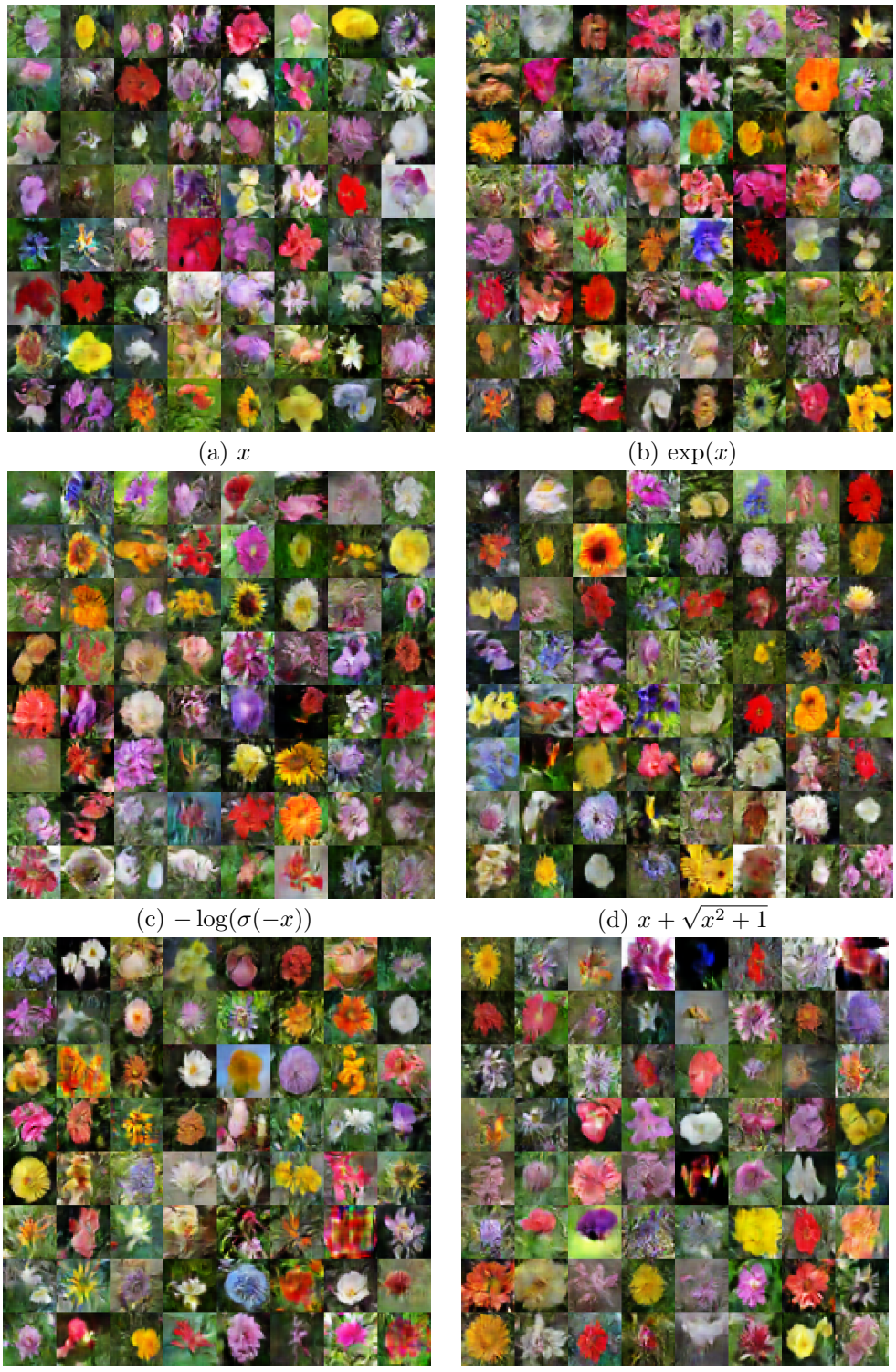


Figure 14: Verifying $\nabla_x f^*(x)$ in LGANs points towards real samples. Here, $\mathcal{P}_r$ consists of ten images and $\mathcal{P}_g$ is Gaussian noise. Up: Each odd column are $x \in \mathcal{S}_g$ and the nearby column are their gradients $\nabla_x f^*(x)$. Down: the leftmost in each row is $x \in \mathcal{S}_g$, 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 $y \in \mathcal{S}_r$.



(a) x (b) $\exp(x)$

(c) $-\log(\sigma(-x))$ (d) $x + \sqrt{x^2 + 1}$

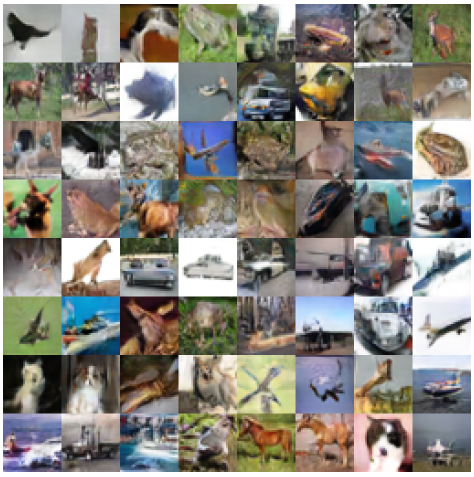
(e) $(x + 1.0)^2$ (f) $\max(0, x + 1.0)$

Figure 15: Random samples of LGANs with different loss metrics on Oxford 102.

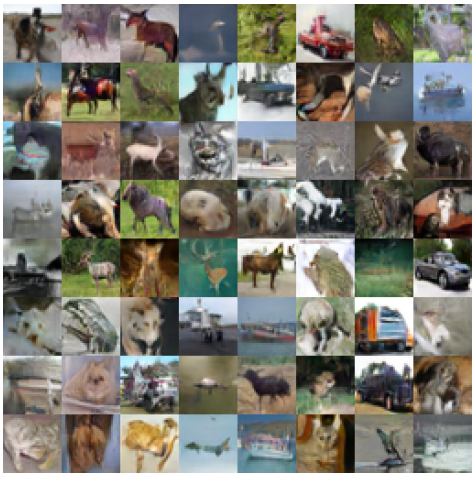
(a) x

(b) $\exp(x)$

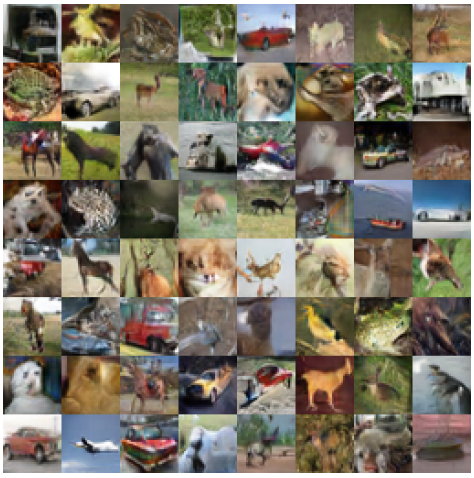
(c) $-\log(\sigma(-x))$



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

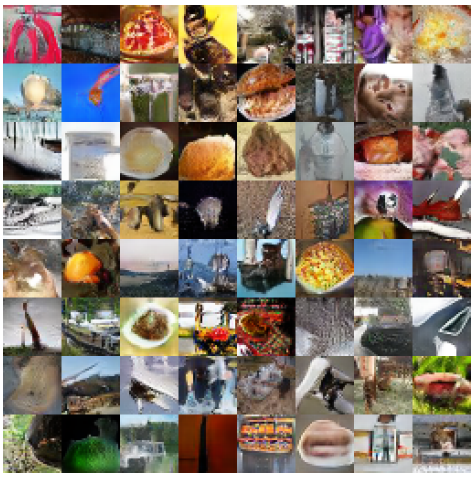


(e) $(x + 1.0)^2$

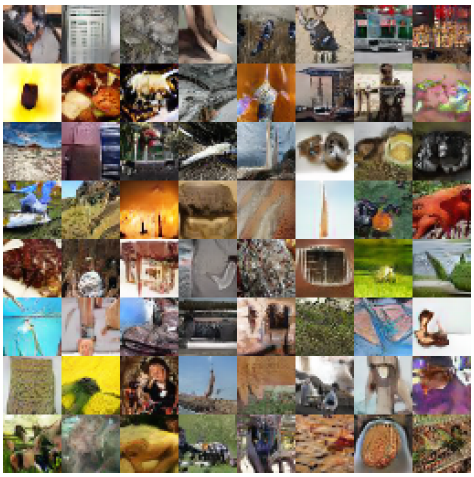


(f) $\max(0, x + 1.0)$

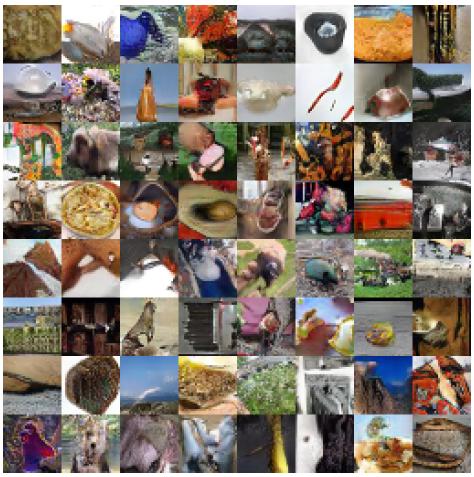
Figure 16: Random samples of LGANs with different loss metrics on CIFAR-10.



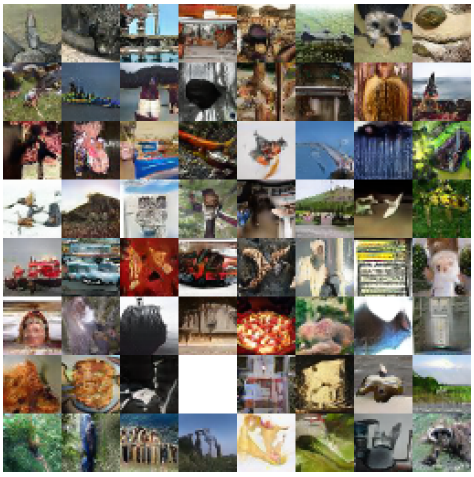
(a) x



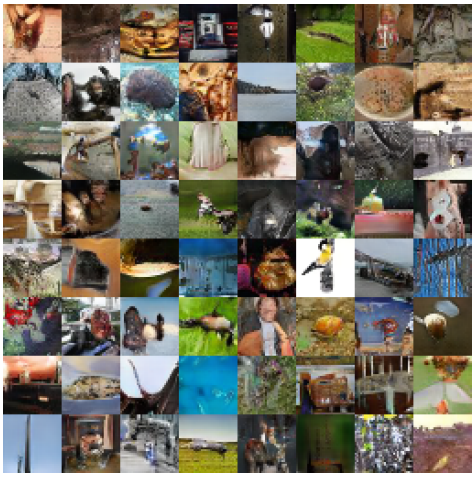
(b) $\exp(x)$



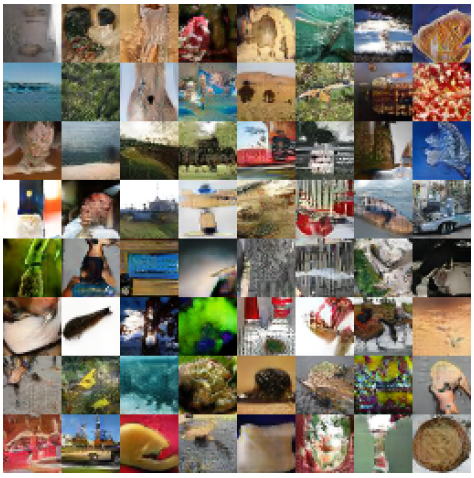
(c) $-\log(\sigma(-x))$



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

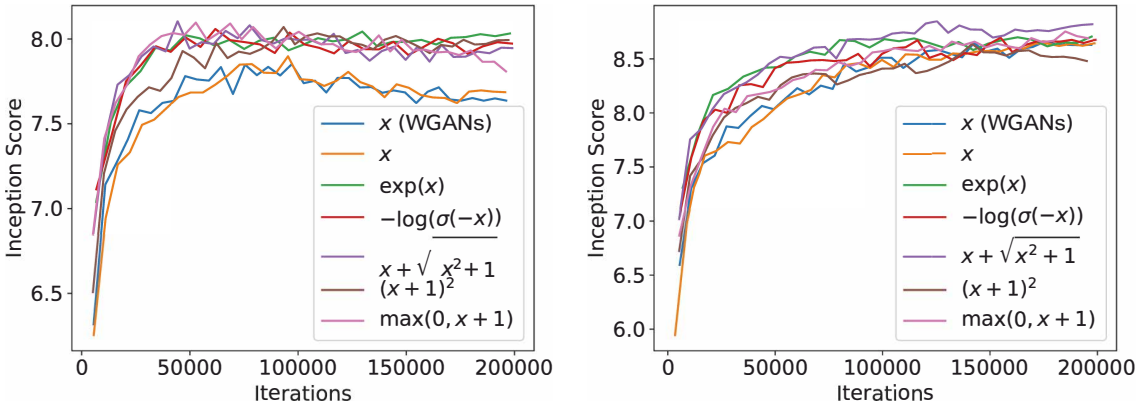


(e) $(x + 1.0)^2$



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

Figure 17: Random samples of LGANs with different loss metrics on Tiny-ImageNet.



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

Figure 18: Training curves in terms of IS. WGANs and a set of instances of LGANs.

References

Jonas Adler and Sebastian Lunz. Banach Wasserstein GAN. *arXiv preprint arXiv:1806.06621*, 2018.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 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 (GANs). *arXiv preprint arXiv:1703.00573*, 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. *arXiv preprint arXiv:1802.03446*, 2018.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Farzan Farnia and David Tse. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems 31*. 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 divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

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*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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*, pages 6626–6637, 2017.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980, 2017.

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*, pages 5545–5553, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *arXiv preprint ArXiv:1611.04076*, 2016.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. In *Advances in Neural Information Processing Systems*, 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*, pages 3478–3487, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

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. *arXiv preprint arXiv:1802.05957*, 2018.

Youssef Mroueh and Tom Sercu. Fisher GAN. In *Advances in Neural Information Processing Systems*, pages 2510–2520, 2017.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. *arXiv preprint arXiv:1711.04894*, 2017.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016.

Augustus Odena, Jacob Buckman, Catherine Olsson, Tom B Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to GAN performance? *arXiv preprint arXiv:1802.08768*, 2018.

Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of Wasserstein GANs. *arXiv preprint arXiv:1709.08894*, 2017.

Guo-Jun Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. 2018.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint*, 2018.

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*, pages 2226–2234, 2016.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee. Solving approximate Wasserstein GANs to stationarity. *arXiv preprint arXiv: 1802.08249*, 2018.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably optimal nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819*, 2017.

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

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

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. *arXiv preprint arXiv:1805.08318*, 2018.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Zhiming Zhou, Shu Rong, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Activation maximization generative adversarial nets. *arXiv preprint arXiv:1703.02000*, 2017.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. *arXiv preprint arXiv:1810.00143*, 2018.

Zhiming Zhou, Jiadong Liang, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Yong Yu, and Zhihua Zhang. Lipschitz generative adversarial nets. *arXiv preprint arXiv:1902.05687*, 2019.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. *arXiv preprint arXiv:1811.09358*, 2018.