

Lipschitz Generative Adversarial Nets

Zhiming Zhou¹ Jiadong Liang² Yuxuan Song¹ Lantao Yu³ Hongwei Wang³ Weinan Zhang¹ Yong Yu¹
Zhihua Zhang²

Abstract

In this paper we show that generative adversarial networks (GANs) without restriction on the discriminative function space commonly suffer from the problem that the gradient produced by the discriminator is uninformative to guide the generator. By contrast, Wasserstein GAN (WGAN), where the discriminative function is restricted to 1-Lipschitz, does not suffer from such a gradient uninformativeness problem. We further show in the paper that the model with a compact dual form of Wasserstein distance, where the Lipschitz condition is relaxed, may also theoretically suffer from this issue. This implies the importance of Lipschitz condition and motivates us to study the general formulation of GANs with Lipschitz constraint, which leads to a new family of GANs that we call Lipschitz GANs (LGANs). We show that LGANs guarantee the existence and uniqueness of the optimal discriminative function as well as the existence of a unique Nash equilibrium. We prove that LGANs are generally capable of eliminating the gradient uninformativeness problem. According to our empirical analysis, LGANs are more stable and generate consistently higher quality samples compared with WGAN.

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 are popular and widely used, but they are notoriously hard to train (Goodfellow, 2016). The underlying obstacles, though have been heavily studied (Arjovsky & Bottou, 2017; Lucic et al., 2017; Heusel et al., 2017; Mescheder et al., 2017;

¹Shanghai Jiao Tong University ²Peking University ³Stanford University. Correspondence to: Zhiming Zhou <heyohai@apex.sjtu.edu.cn>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

2018; Yadav et al., 2017), are still not fully understood.

The objective of GAN is usually defined as a distance metric between the real distribution \mathcal{P}_r and the generative distribution \mathcal{P}_g , which implies that $\mathcal{P}_r = \mathcal{P}_g$ is the unique global optimum. The nonconvergence of traditional GANs has been considered as a result of ill-behaving distance metric (Arjovsky & Bottou, 2017), i.e., 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, which can properly measure the distance between two distributions no matter whether their supports are disjoint.

In this paper, we conduct a further study on the convergence of GANs from the perspective of the informativeness of the gradient of the *optimal discriminative function* f^* . We show that for GANs that have no restriction on the discriminative function space, e.g., the vanilla GAN and its most variants, $f^*(x)$ is only related to the densities of the local point x and does not reflect any information about other points in the distributions. We demonstrate that under these circumstances, the gradient of the optimal discriminative function with respect to its input, on which the generator updates generated samples, usually tells nothing about the real distribution. We refer to this phenomenon as the *gradient uninformativeness*, which is substantially different from the gradient vanishing and is a fundamental cause of nonconvergence of GANs.

According to the analysis of Gulrajani et al. (2017), Wasserstein GAN can avoid the gradient uninformativeness problem. Meanwhile, we show in the paper that the Lipschitz constraint in the Kantorovich-Rubinstein dual of the Wasserstein distance can be relaxed, leading to a new equivalent dual; and with the new dual form, the gradient may also not reflect any information about how to refine \mathcal{P}_g towards \mathcal{P}_r . It suggests that Lipschitz condition would be a vital element for resolving the gradient uninformativeness problem.

Motivated by the above analysis, we investigate the general formulation of GANs with Lipschitz constraint. We show that under a mild condition, penalizing Lipschitz constant guarantees the existence and uniqueness of the optimal discriminative function as well as the existence of the unique Nash equilibrium between f^* and \mathcal{P}_g where $\mathcal{P}_r = \mathcal{P}_g$. It leads to a new family of GANs that we call Lipschitz GANs

Table 1: Comparison of different objectives in GANs.

	ϕ	φ	\mathcal{F}	$f^*(x)$	Gradient Vanishing	Gradient Uninformative	$f^*(x)$ Uniqueness
Vanilla GAN	$-\log(\sigma(-x))$	$-\log(\sigma(x))$	$\{f: \mathbb{R}^n \rightarrow \mathbb{R}\}$	$\log \frac{\mathcal{P}_r(x)}{\mathcal{P}_g(x)}$	Yes	Yes	Yes
Least-Squares GAN	$(x - \alpha)^2$	$(x - \beta)^2$	$\{f: \mathbb{R}^n \rightarrow \mathbb{R}\}$	$\frac{\alpha \cdot \mathcal{P}_g(x) + \beta \cdot \mathcal{P}_r(x)}{\mathcal{P}_r(x) + \mathcal{P}_g(x)}$	No	Yes	Yes
μ -Fisher GAN	x	$-x$	$\{f: \mathbb{R}^n \rightarrow \mathbb{R}, \mathbb{E}_{x \sim \mu} f(x) ^2 \leq 1\}$	$\frac{1}{\mathcal{F}_\mu(\mathcal{P}_r, \mathcal{P}_g)} \frac{\mathcal{P}_r(x) - \mathcal{P}_g(x)}{\mu(x)}$	No	Yes	Yes
Wasserstein GAN	x	$-x$	$\{f: \mathbb{R}^n \rightarrow \mathbb{R}, k(f) \leq 1\}$	N/A	No	No	No
Lipschitz GAN	any ϕ and φ satisfying Eq. (11)		$\{f: \mathbb{R}^n \rightarrow \mathbb{R}\}; k(f)$ is penalized	N/A	No	No	Yes

(LGANs). We show that LGANs are generally capable of eliminating the gradient uninformativeness in the manner that with the optimal discriminative function, the gradient for each generated sample, if nonzero, will point towards some real sample. This process continues until the Nash equilibrium $\mathcal{P}_r = \mathcal{P}_g$ is reached.

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 LGANs and their theoretical analysis. We conduct the empirical analysis in Section 5. Finally, we discuss related work in Section 6 and conclude the paper in Section 7.

2. Preliminaries

In this section we first give some notions and then present a general formulation for generative adversarial networks.

2.1. Notation and Notions

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

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

In this paper and in most existing GANs, the metrics d_X and d_Y are by default Euclidean distance which we also denote by $\|\cdot\|$. The smallest constant k is called the (best) Lipschitz constant of f , denoted by $k(f)$.

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

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

where $\Pi(\mathcal{P}_r, \mathcal{P}_g)$ denotes the set of all probability measures with marginals \mathcal{P}_r and \mathcal{P}_g . 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}_r and \mathcal{S}_g denote the supports of \mathcal{P}_r and \mathcal{P}_g , respectively. We say two distributions are disjoint if their supports are disjoint.

The Kantorovich-Rubinstein (KR) duality (Villani, 2008) provides a way of more efficiently computing of Wasserstein

distance. The duality states that

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

The constraint in Eq. (3) implies that f is Lipschitz continuous with $k(f) \leq 1$. Interestingly, we have a more compact dual form of the Wasserstein distance. That is,

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

The proof for this dual form is given in Appendix A.5. We see that this new dual relaxes the Lipschitz continuity condition of the dual form in Eq. (3).

2.2. Generative Adversarial Networks (GANs)

Typically, GANs can be formulated as

$$\begin{aligned} \min_{f \in \mathcal{F}} J_D &\triangleq \mathbb{E}_{z \sim \mathcal{P}_z} [\phi(f(g(z)))] + \mathbb{E}_{x \sim \mathcal{P}_r} [\varphi(f(x))], \\ \min_{g \in \mathcal{G}} J_G &\triangleq \mathbb{E}_{z \sim \mathcal{P}_z} [\psi(f(g(z)))] \end{aligned} \quad (5)$$

where \mathcal{P}_z is the source distribution of the generator in \mathbb{R}^m and \mathcal{P}_r is the target (real) distribution in \mathbb{R}^n . The generative function $g: \mathbb{R}^m \rightarrow \mathbb{R}^n$ learns to output samples that share the same dimension as samples in \mathcal{P}_r , while the discriminative function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\phi, \varphi, \psi: \mathbb{R} \rightarrow \mathbb{R}$ are loss metrics. We denote the implicit distribution of the generated samples by \mathcal{P}_g .

We list the choices of \mathcal{F} , ϕ and φ in some representative GAN models in Table 1. In these GANs, the gradient that the generator receives from the discriminator with respect to (w.r.t.) a generated sample $x \in \mathcal{S}_g$ is

$$\nabla_x J_G(x) \triangleq \nabla_x \psi(f(x)) = \nabla_{f(x)} \psi(f(x)) \cdot \nabla_x f(x), \quad (6)$$

where the first term $\nabla_{f(x)} \psi(f(x))$ is a step-related 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 .

We use f^* to denote the optimal discriminative function, i.e., $f^* \triangleq \arg \min_{f \in \mathcal{F}} J_D$. For further notation, we let $\dot{J}_D(x) \triangleq \mathcal{P}_g(x)\phi(f(x)) + \mathcal{P}_r(x)\varphi(f(x))$. It has $J_D = \int \dot{J}_D(x) dx$.

2.3. The Gradient Vanishing

The gradient vanishing problem has been typically thought as a key factor for causing the nonconvergence of GANs, i.e., the gradient becomes zero when the discriminator is perfectly trained.

[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 GAN ([Mao et al., 2016](#)), which aims at addressing the gradient vanishing problem, also focused on $\nabla_{f(x)}\psi(f(x))$.

[Arjovsky & Bottou \(2017\)](#) provided a new perspective for understanding the gradient vanishing. They argued that \mathcal{S}_r and \mathcal{S}_g are usually disjoint and the gradient vanishing stems from the ill-behaving of traditional distance metrics, i.e., the distance between \mathcal{P}_r and \mathcal{P}_g remains constant when they are disjoint. The Wasserstein distance was thus used ([Arjovsky et al., 2017](#)) as an alternative metric, which can properly measure the distance between two distributions no matter whether they are disjoint.

3. The Gradient Uninformativeness

In this paper we pay our main attention on the gradient direction of the optimal discriminative function, i.e., $\nabla_x f^*(x)$, along which the generated sample x is updated. We show that for many distance metrics, 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 factor of resulting in nonconvergence and instability in the training of traditional GANs.

The gradient uninformativeness is substantially different from the gradient vanishing. The gradient vanishing is about the scalar term $\nabla_{f(x)}\psi(f(x))$ in $\nabla_x J_G(x)$ 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 defined by $\nabla_x f^*(x)$. The two issues are orthogonal, though they sometimes exist simultaneously. See Table 1 for a summary of issues for representative GANs.

Next, we discuss the gradient uninformativeness in the taxonomy of restrictions on the discriminative function space \mathcal{F} . We will show that for unrestricted GANs, gradient uninformativeness commonly exists; for restricted GANs, such an issue might still exist; and with Lipschitz condition, it generally does not exist.

3.1. Unrestricted GANs

For many GAN models, there is no restriction on \mathcal{F} . Typical cases include f -divergence based GANs, such as the vanilla GAN ([Goodfellow et al., 2014](#)), Least-Squares GAN ([Mao et al., 2016](#)) and f -GAN ([Nowozin et al., 2016](#)).

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

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

Hence, 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 x receives from the optimal discriminative function, does not reflect any information about \mathcal{P}_r .

Typical situation is that \mathcal{S}_r and \mathcal{S}_g are disjoint, which is common in practice according to ([Arjovsky & Bottou, 2017](#)). To further distinguish the gradient uninformativeness from the gradient vanishing, we consider an ideal case: \mathcal{S}_r and \mathcal{S}_g are totally overlapped and both consist of n discrete points, but their probability masses over these points are different. In this case, $\nabla_x f^*(x)$ for each generated sample is still uninformative, but the gradient does not vanish.

3.2. Restricted GANs: Fisher GAN as an Instance

Some GANs impose restrictions on \mathcal{F} . Typical instances are the Integral Probability Metric (IPM) based GANs ([Mroueh & Sercu, 2017](#); [Mroueh et al., 2017](#); [Bellemare et al., 2017](#)) and the Wasserstein GAN ([Arjovsky et al., 2017](#)). We next show that GANs with restriction on \mathcal{F} might also suffer from the gradient uninformativeness.

The optimal discriminative function of μ -Fisher IPM $\mathcal{F}_\mu(\mathcal{P}_r, \mathcal{P}_g)$, the generalized objective of the Fisher GAN ([Mroueh et al., 2017](#)), has the following form:

$$f^*(x) = \frac{1}{\mathcal{F}_\mu(\mathcal{P}_r, \mathcal{P}_g)} \frac{\mathcal{P}_r(x) - \mathcal{P}_g(x)}{\mu(x)}, \quad (7)$$

where μ is a distribution whose support covers \mathcal{S}_r and \mathcal{S}_g , and $\frac{1}{\mathcal{F}_\mu(\mathcal{P}_r, \mathcal{P}_g)}$ 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.

3.3. The Wasserstein GAN

As shown by ([Gulrajani et al. \(2017\)](#)), the gradient of the optimal discriminative function in the KR dual form of the Wasserstein distance has the following property:

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

$$\mathbb{P}_{(x,y) \sim \pi^*} \left[\nabla_{x_t} f^*(x_t) = \frac{y - x}{\|y - x\|} \right] = 1. \quad (8)$$

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 , i.e., the gradient at any x_t is pointing towards the real sample y ; (ii) these (x, y) pairs match the optimal coupling π^* in the optimal transport perspective. It implies that WGAN is able to overcome the gradient uninformative ness as well as the gradient vanishing.

Our concern turns to the reason why WGAN can avoid gradient uninformative ness. To address this question, we alternatively apply the compact dual of the Wasserstein distance in Eq. (4) and study the optimal discriminative function.

Since there is generally no closed-form solution for f^* in Eq. (4), we take an illustrative example, but the conclusion is general. Let $Z \sim U[0, 1]$ be a uniform variable on interval $[0, 1]$, \mathcal{P}_r be the distribution of $(1, Z)$ in \mathbb{R}^2 , and \mathcal{P}_g be the distribution of $(0, Z)$ in \mathbb{R}^2 . According to Eq. (4), we have an optimal f^* as follows

$$f^*(x) = \begin{cases} 1, & \forall x \in \mathcal{S}_r; \\ 0, & \forall x \in \mathcal{S}_g. \end{cases} \quad (9)$$

Though having the constraint “ $f(x) - f(y) \leq d(x, y)$, $\forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g$,” the Wasserstein distance in this dual form also only defines the values of $f^*(x)$ on \mathcal{S}_r and \mathcal{S}_g . For each generated sample x which is isolated or at the boundary (there does not exist $\epsilon > 0$ such that it holds $y \in \mathcal{S}_r \cup \mathcal{S}_g$ for all y with $0 < \|y - x\| < \epsilon$), the gradient of $f^*(x)$ is theoretically undefined and thus cannot provide useful information about \mathcal{P}_r . We can consider the more extreme case where \mathcal{S}_g are isolated points to make it clearer.

These examples imply that Lipschitz condition would be critical for resolving the gradient uninformative ness problem. Motivated by this, we study the general formulation of GANs with Lipschitz constraint, which leads to a family of more general GANs that we call Lipschitz GANs. We will see that in Lipschitz GANs, the similarity measure between \mathcal{P}_r and \mathcal{P}_g might not be some Wasserstein distance, but they still perform very well.

4. Lipschitz GANs

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

In this paper, we investigate the general formulation of GANs with Lipschitz constraint, 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 Generative Adversarial

Nets (LGANs) as:

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

In this work, we further assume that the loss functions ϕ and φ satisfy the following conditions:

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

The assumptions for the losses ϕ and φ are very mild. Note that in WGAN $\phi(x) = \varphi(-x) = x$ is used, which satisfies Eq. (11). There are many other instances, such as $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$, $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + 1}$ and $\phi(x) = \varphi(-x) = \exp(x)$. Meanwhile, there also exist losses 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 & Ye, 2017; Miyato et al., 2018).

To devise a loss in LGANs, it is practical to let ϕ be an increasing function with non-decreasing derivative and set $\phi(x) = \varphi(-x)$. Moreover, the linear combinations of such losses still satisfy Eq. (11). Figure 13 illustrates some of these loss metrics.

Note that $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$ is the objective of vanilla GAN. As we have shown, the vanilla GAN suffers from the gradient uninformative ness problem. However, as we will show next, when imposing the Lipschitz regularization, the resulting model as a specific case of LGANs behaves very well.

4.1. Theoretical Analysis

We now present the theoretical analysis of LGANs. First, we consider the existence and uniqueness of the optimal discriminative function.

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

Note that although WGAN does not satisfy the condition in Theorem 1, its solution 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. The following theorems can be regarded as a generalization of Proposition 1 to LGANs.

Theorem 2. *Assume $\phi'(x) > 0$, $\varphi'(x) < 0$, and the optimal discriminator f^* exists and is smooth. We have*

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

- (c) If $\mathcal{S}_r = \mathcal{S}_g$ and $\mathcal{P}_r \neq \mathcal{P}_g$, then there exists (x, y) pair with both points in $\mathcal{S}_r \cup \mathcal{S}_g$ and $y \neq x$ 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^* under the objective $J_D + \lambda \cdot k(f)^2$, where it holds that $\mathcal{P}_r = \mathcal{P}_g$ and $k(f^*) = 0$.

The proof is given in Appendix A.2. This theorem states the basic properties of LGANs, including the existence of unique Nash equilibrium where $\mathcal{P}_r = \mathcal{P}_g$ and 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 uninformativity problem.

It is worth noticing that the penalty $k(f)$ is in fact necessary for Property-(c) and Property-(d). The reason is due to the existence of the case that $\nabla_{f^*(x)} \mathring{J}_D(x) = 0$ for $\mathcal{P}_r(x) \neq \mathcal{P}_g(x)$. Minimizing $k(f)$ guarantees that the only Nash equilibrium is achieved when $\mathcal{P}_r = \mathcal{P}_g$. In WGAN, 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 $\mathcal{P}_r = \mathcal{P}_g$ where any function subject to 1-Lipschitz constraint is an optimal f^* in WGAN. It implies that minimizing $k(f)$ also benefits WGAN.

4.2. Refining the Bounding Relationship

From Theorem 2, we know that for any point x , as long as $\mathring{J}_D(x)$ does not hold a zero gradient with respect to $f^*(x)$, $f^*(x)$ must be bounded by another point y such that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$. We further clarify that when there is a bounding relationship, it must involve both real sample(s) and fake sample(s). 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 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 only it is not the overlapping case, all fake samples hold $\nabla_{f^*(x)} \mathring{J}_D(x) > 0$, while all real samples hold $\nabla_{f^*(y)} \mathring{J}_D(y) < 0$.

Note that there might exist a dozen real and fake samples that bound each other. Under the Lipschitz continuity condition, the bounding relationship on the value surface of f^* is

the basic building block that connects \mathcal{P}_r and \mathcal{P}_g , and each fake sample with $\nabla_{f^*(x)} \mathring{J}_D(x) \neq 0$ lies in at least one of these bounded relationships. Next we will further interpret the implication of bounding relationship and show that it guarantees meaningful $\nabla_x f^*(x)$ for all involved points.

4.3. The Implication of Bounding Relationship

Recall that the Proposition 1 states that $\nabla_{x_t} f^*(x_t) = \frac{y-x}{\|y-x\|}$. We next show that it is actually a direct consequence of bounding relationship between x and y . 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 on the value surface of f . Any point in this line holds the maximum gradient slope k , and the gradient direction at any point in this line is pointing towards the $x \rightarrow y$ direction. The proof is provided in Appendix A.4.

Combining Theorems 2 and 3, we can conclude that when \mathcal{S}_r and \mathcal{S}_g are disjoint, the gradient $\nabla_x f^*(x)$ for each generated sample $x \in \mathcal{S}_g$ points towards some real sample $y \in \mathcal{S}_r$, which guarantees that $\nabla_x f^*(x)$ -based updating would pull \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 $\nabla_x f^*(x)$ must point towards some real sample $y \in \mathcal{S}_r$. And in the fully overlapped case, according to Property-(c), unless $\mathcal{P}_r = \mathcal{P}_g$, there must exist at least one pair of (x, y) in strict bounding relationship and $\nabla_x f^*(x)$ pulls x towards y . Finally, Property-(d) guarantees that the only Nash equilibrium is $\mathcal{P}_r = \mathcal{P}_g$ where $\nabla_x f^*(x) = 0$ for all generated samples.

5. Empirical Analysis

In this section, we empirically study the gradient uninformativity problem and the performance of various objectives of Lipschitz GANs. The anonymous code is provided in the supplemental material.

5.1. Gradient Uninformativeness in Practice

According to our analysis, $\nabla_x f^*(x)$ for most traditional GANs is uninformative. Here we investigate the practical behaviors of the gradient uninformativity. Note that the behaviors of GANs without restriction on \mathcal{F} are essentially identical. We choose the Least-Squares GAN whose f^* is relatively simple as the representative and study it with a set

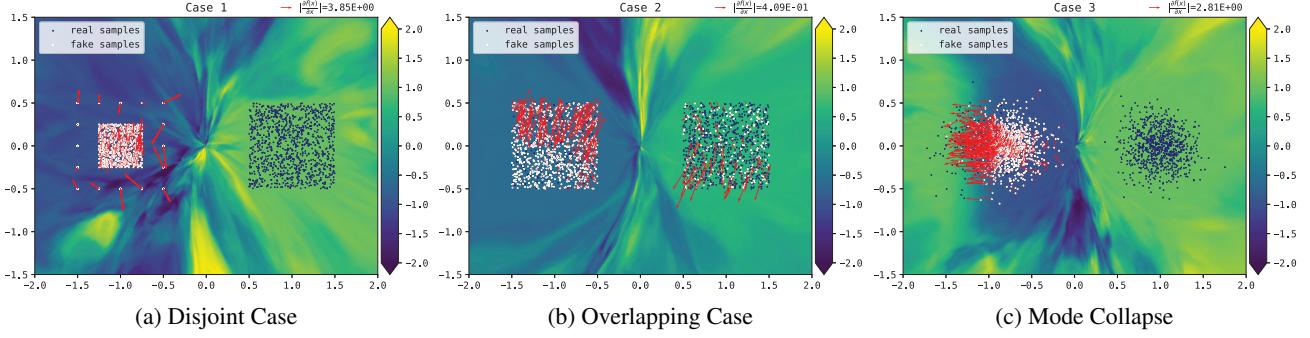


Figure 1: Practical behaviors of gradient uninformativeness: noisy gradient. Local greedy gradient leads to mode collapse.

of synthetic experiments which benefits the visualization.

The results are shown in Figure 1. We find that the gradient is very random, which we believe is the typical practical behavior of the gradient uninformativeness. Given the non-deterministic property of $f^*(x)$ for points out of $\mathcal{S}_r \cup \mathcal{S}_g$, $\nabla_x f^*(x)$ is highly sensitive to the hyper-parameters. We actually conduct the same experiments with a set of different hyper-parameters. The rest is provided in Appendix B.

In Section 3, we discussed the gradient uninformativeness under the circumstances that the fake sample is not surrounded by real samples. Actually, the problem of $\nabla_x f^*(x)$ in traditional GANs is more general, which can also be regarded as the gradient uninformativeness. For example, in the case of Figure 1b where the real and fake samples are both evenly distributed in the two regions with different densities, $f^*(x)$ is constant in each region and undefined outside. It theoretically has zero $\nabla_x f^*(x)$ for inner points and undefined $\nabla_x f^*(x)$ for boundary points. They in practice also behave as noisy gradient. We note that in the totally overlapping and continuous case, $\nabla_x f^*(x)$ is also ill-behaving, which seems to be an intrinsic cause of mode collapse, as illustrated in Figure 1c where \mathcal{P}_r and \mathcal{P}_g are both devised to be Gaussian(s).

5.2. Verifying $\nabla_x f^*(x)$ of LGANs

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 2. 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 2 and 3, respectively. In Figure 2, 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 non-trivial. 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 . This result visually demonstrates that the gradient of a generated sample is towards a real sample. Note that the final results of Figure 3 keep almost identical when varying the loss metric ϕ and φ in the family of LGANs.

5.3. Stabilized Discriminative Functions

The Wasserstein distance is a very special case that has solution under Lipschitz constraint. It is the only case where both ϕ and φ have constant derivative. As a result, f^* under the Wasserstein distance 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 WGAN; Karras et al. (2017) and Adler & 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 5.

5.4. Max Gradient Penalty (MaxGP)

LGANs impose penalty on the Lipschitz constant of the discriminative function. There are works that investigate different implementations of Lipschitz continuity in GANs, such as gradient penalty (GP) (Gulrajani et al., 2017), Lipschitz penalty (LP) (Petzka et al., 2017) and spectral normalization (SN) (Miyato et al., 2018). However, the existing regularization methods do not directly penalize the Lipschitz constant. According to (Adler & Lunz, 2018), Lipschitz constant $k(f)$

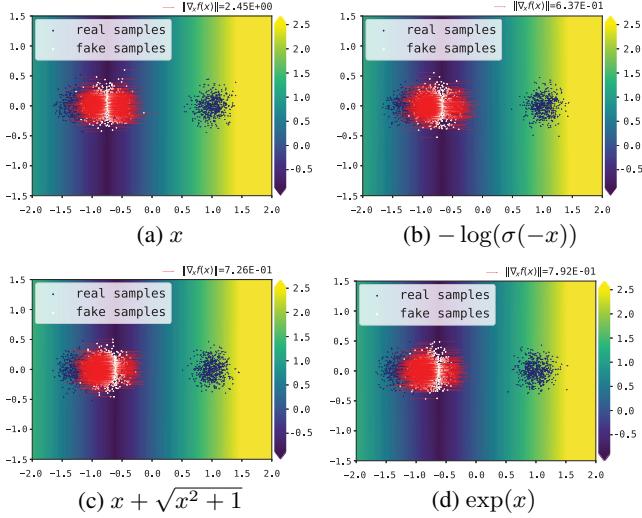


Figure 2: $\nabla_x f^*(x)$ in LGANs point towards real samples.

is equivalent to the maximum scale of $\|\nabla_x f(x)\|$. Both GP and LP penalize all gradients whose scales are larger than the given target Lipschitz constant k_0 . SN directly restricts the Lipschitz constant via normalizing the network weights by their largest eigenvalues. However, it is currently unclear how to effectively penalize the Lipschitz constant with SN.

To directly penalize Lipschitz constant, we approximate $k(f)$ in Eq. (10) with the maximum sampled gradient scale:

$$k(f) \simeq \max_x \|\nabla_x f(x)\|. \quad (12)$$

Practically, we follow (Gulrajani et al., 2017) and sample x as random interpolation of real and fake samples. We provide more details of this algorithm (MaxGP) in Appendix C.

According to our experiments, MaxGP in practice is usually comparable with GP and LP. However, in some of our synthetic experiments, we find that MaxGP is able to achieve the optimal discriminative function while GP and LP fail, e.g., the problem of solving f^* in Figure 3. Also, in some real data experiments, we find the training with GP or LP diverges and it is able to converge if we switch to MaxGP, e.g., the training with metric $\phi(x) = \varphi(-x) = \exp(x)$.

5.5. Benchmark with Unsupervised Image Generation

To quantitatively compare the performance of different objectives under Lipschitz constraint, 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 $\phi(x) = \varphi(-x) = (x + \alpha)^2$. To make the comparison simple, we fix $\psi(x)$ in the objective of generator as $-x$. We set $\alpha = 1.0$ in the experiment.

The strict monotonicity assumption of ϕ and φ is critical

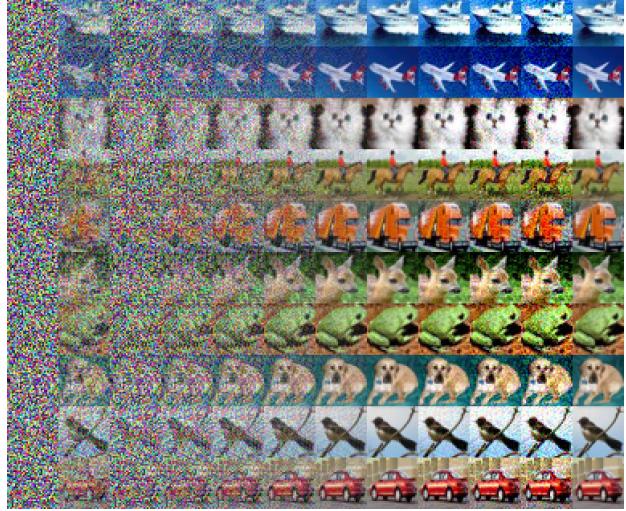


Figure 3: $\nabla_x f^*(x)$ gradation with CIFAR-10.

in Theorem 2 to theoretically guarantee the existences of bounding relationships for *arbitrary datas*. But if we further assume S_r and S_g 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 : for the hinge loss, it would mean $2\alpha < k(f) \cdot \|y - x\|$ for all $y \in S_r$ and $x \in S_g$.

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 WGAN-GP in our implementation achieves IS 7.71 ± 0.03 and FID 18.86 ± 0.13 on CIFAR-10. We use MaxGP for all experiments 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 is remarkably unstable during training and among different initializations. By contrast, FID is fairly stable.

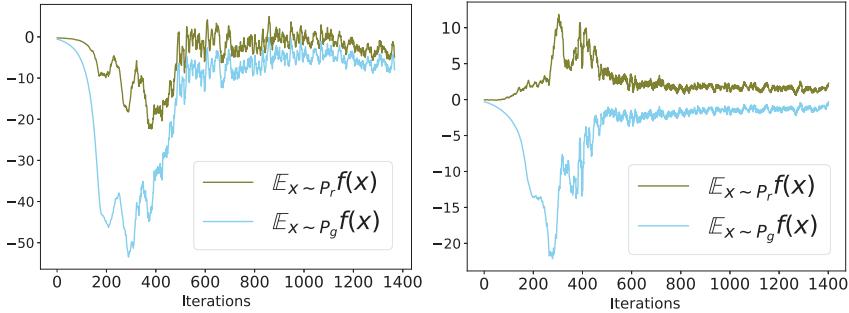
From Table 2, we can see that LGANs generally work better than WGAN. Different LGANs have relatively similar final results, while the objectives $\phi(x) = \varphi(-x) = \exp(x)$ and $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + 1}$ achieve the best performances. The hinge loss and quadratic loss with a suitable λ turn out to also work pretty good. We plot the training curves in terms of FID in Figures 4 and 6. Due to page limitation, we leave more results and details in Appendix D.

6. Related Work

WGAN (Arjovsky et al., 2017) based on the KR dual does not suffer from the gradient uninformativeness problem. We have shown that the Lipschitz constraint in the KR dual of the Wasserstein distance can be relaxed. With the new dual form, the resulting model suffers from the gradient

Table 2: Quantitative comparisons with unsupervised image generation.

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
$\exp(x)$	8.03 ± 0.03	15.64 ± 0.07	8.67 ± 0.04	14.90 ± 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	8.82 ± 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


 Figure 5: $f^*(x)$ in LGANs is more stable. Left: WGAN. Right: LGANs.

uninformativeness problem.

We have shown that Lipschitz constraint is able to ensure the convergence for a family of GAN objectives, which is not limited to the Wasserstein distance. For example, Lipschitz continuity is also introduced to the vanilla GAN (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 vanilla GAN 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 & Tse, 2018) also provide some analysis on how f -divergence behaviors when combined with Lipschitz. However, their analysis is limited to the symmetric f -divergence.

Fedus et al. (2017) also argued that divergence is not the primary guide of the training of GANs. However, they thought that the vanilla GAN with a non-saturating generator objective somehow works. According to our analysis, given the optimal f^* , the vanilla GAN has no guarantee on its convergence. Unterthiner et al. (2017) provided some arguments on the unreliability of $\nabla_x f^*(x)$ in traditional GANs, which motivates their proposal of Coulomb GAN. However, the arguments there are not thorough. By contrast, we identify the gradient uninformativeness problem and link it to the restrictions on \mathcal{F} . Moreover, we have accordingly proposed a new solution, i.e., the Lipschitz GANs.

Some work studies the suboptimal convergence of GANs (Mescheder et al., 2017; 2018; Arora et al., 2017; Liu et al., 2017; Farnia & Tse, 2018), which is another important direction for theoretically understanding GANs. Despite the fact

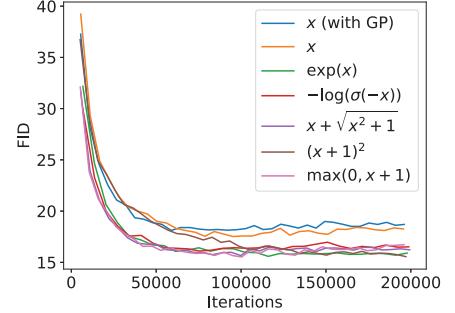


Figure 4: Training curves on CIFAR.

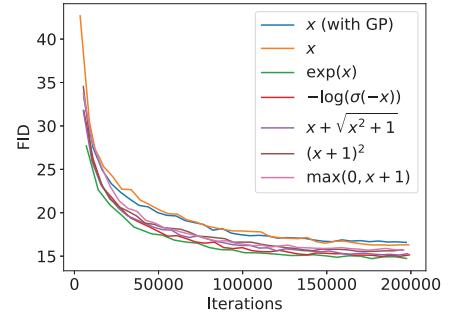


Figure 6: Training curves on Tiny.

that the behaviors of suboptimal can be different, we think the optimal should well-behave in the first place, e.g., informative gradient and stable Nash equilibrium. 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.

7. Conclusion

In this paper we 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 unrestricted GANs and also appears in restricted 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 show 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) and the helpful discussions with Dachao Lin. Jiadong Liang and Zhihua Zhang have been supported by Beijing Municipal Commission of Science and Technology under Grant No. 181100008918005, and by Beijing Academy of Artificial Intelligence (BAAI).

References

- Adler, J. and Lunz, S. Banach Wasserstein GAN. *arXiv preprint arXiv:1806.06621*, 2018.
- Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.
- Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.
- Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.
- Farnia, F. and Tse, D. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems 31*. 2018.
- Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: GANs do not need to decrease divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.
- Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*, 2017.
- Lim, J. H. and Ye, J. C. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.
- Liu, S., Bousquet, O., and Chaudhuri, K. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pp. 5545–5553, 2017.
- Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are GANs created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. *arXiv preprint ArXiv:1611.04076*, 2016.
- Mescheder, L., Nowozin, S., and Geiger, A. The numerics of GANs. In *Advances in Neural Information Processing Systems*, pp. 1825–1835, 2017.
- Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, pp. 3478–3487, 2018.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- Mroueh, Y. and Sercu, T. Fisher GAN. In *Advances in Neural Information Processing Systems*, pp. 2510–2520, 2017.
- Mroueh, Y., Li, C., Sercu, T., Raj, A., and Cheng, Y. Sobolev GAN. *arXiv preprint arXiv:1711.04894*, 2017.
- Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.
- Odena, A., Buckman, J., Olsson, C., Brown, T. B., Olah, C., Raffel, C., and Goodfellow, I. Is generator conditioning causally related to GAN performance? *arXiv preprint arXiv:1802.08768*, 2018.
- Petzka, H., Fischer, A., and Lukovnicov, D. On the regularization of Wasserstein GANs. *arXiv preprint arXiv:1709.08894*, 2017.
- Qi, G.-J. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2226–2234, 2016.

Unterthiner, T., Nessler, B., Klambauer, G., Heusel, M., Ramsauer, H., and Hochreiter, S. Coulomb GANs: Provably optimal nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819*, 2017.

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

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

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

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

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 $\|f\|_{Lip} \leq k$. 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

$$\begin{aligned} \mathfrak{G}(f) &= \mathbb{E}_{X \sim \mathcal{P}_g} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(f(Y)) \\ &\geq \mathbb{E}_{X \sim \mathcal{P}_g} (\phi'(a_0)(f(x) - a_0) + \phi(a_0)) + \mathbb{E}_{Y \sim \mathcal{P}_r} (\varphi'(a_0)(f(x) - a_0) + \varphi(a_0)) \\ &= \phi'(a_0) \mathbb{E}_{X \sim \mathcal{P}_g} f(x) + \varphi'(a_0) \mathbb{E}_{Y \sim \mathcal{P}_r} f(Y) + C \\ &= (\phi'(a_0) + \varphi'(a_0)) \mathbb{E}_{X \sim \mathcal{P}_g} f(X) + \varphi'(a_0) (\mathbb{E}_{Y \sim \mathcal{P}_r} f(Y) - \mathbb{E}_{X \sim \mathcal{P}_g} f(X)) + C \\ &= k \varphi'(a_0) (\mathbb{E}_{Y \sim \mathcal{P}_r} \frac{1}{k} f(Y) - \mathbb{E}_{X \sim \mathcal{P}_g} \frac{1}{k} f(X)) + C \\ &\geq -k \varphi'(a_0) W_1(\mathcal{P}_r, \mathcal{P}_g) + C. \end{aligned} \tag{13}$$

Therefore, we get the lower bound. \square

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

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

Proof. Since ϕ, φ are convex functions, we have

$$\begin{aligned} \mathfrak{G}(f) &= \mathbb{E}_{X \sim \mathcal{P}_g} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(f(Y)) \\ &\geq \mathbb{E}_{X \sim \mathcal{P}_g} (\phi'(a_1)(f(x) - a_1) + \phi(a_1)) + \mathbb{E}_{Y \sim \mathcal{P}_r} (\varphi'(a_1)(f(x) - a_1) + \varphi(a_1)) \\ &= \phi'(a_1) \mathbb{E}_{X \sim \mathcal{P}_g} f(x) + \varphi'(a_1) \mathbb{E}_{Y \sim \mathcal{P}_r} f(Y) + C_1 \\ &= (\phi'(a_1) + \varphi'(a_1)) \mathbb{E}_{X \sim \mathcal{P}_g} f(X) + \varphi'(a_1) (\mathbb{E}_{Y \sim \mathcal{P}_r} f(Y) - \mathbb{E}_{X \sim \mathcal{P}_g} f(X)) + C_1 \\ &= (\phi'(a_1) + \varphi'(a_1)) \mathbb{E}_{X \sim \mathcal{P}_g} f(X) + k \varphi'(a_1) (\mathbb{E}_{Y \sim \mathcal{P}_r} \frac{1}{k} f(Y) - \mathbb{E}_{X \sim \mathcal{P}_g} \frac{1}{k} f(X)) + C_1 \\ &\geq (\phi'(a_1) + \varphi'(a_1)) \mathbb{E}_{X \sim \mathcal{P}_g} f(X) - k \varphi'(a_1) W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1 \\ &\geq (\phi'(a_1) + \varphi'(a_1)) f(0) - k(\phi'(a_1) + \varphi'(a_1)) \mathbb{E}_{X \sim \mathcal{P}_g} \|X\| - k \varphi' W_1(\mathcal{P}_r, \mathcal{P}_g) + C_1. \end{aligned} \tag{14}$$

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

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 $\lim_{n \rightarrow \infty} \mathfrak{G}(f_n) = \inf(\mathfrak{G}(f))$. Suppose that $\{r_i\}_{i=1}^\infty$ is the sequence of all rational points in $\text{dom}(f)$. 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 $x \in \text{dom}(f)$, 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

$$\begin{aligned} \lim_{m,l \rightarrow \infty} |f_{mm}(x) - f_{ll}(x)| &\leq \lim_{m,l \rightarrow \infty} (|f_{mm}(x) - f_{mm}(r)| + |f_{mm}(r) - f_{ll}(r)| + |f_{ll}(r) - f_{ll}(x)|) \\ &\leq \lim_{m,l \rightarrow \infty} \left(\frac{\epsilon}{10} + \frac{\epsilon}{10} + |f_{mm}(r) - f_{ll}(r)| \right) = \frac{\epsilon}{5} \end{aligned} \quad (15)$$

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

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

$$\phi(g_n(x)) \geq \phi(g_n(0) - k\|x\|) \geq \phi(-C' - k\|x\|) \geq \phi'(a_0)(-C' - k\|x\| - a_0) + \phi(a_0) = -k\phi'(a_0)\|x\| + C'' \quad (16)$$

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

By Fatou's Lemma,

$$\begin{aligned} \mathbb{E}_{X \sim \mathcal{P}_g} (\phi(g(X)) + k\phi'(a_0)\|X\| - C'') &= \mathbb{E}_{X \sim \mathcal{P}_g} \lim_{n \rightarrow \infty} (\phi(g_n(X)) + k\phi'(a_0)\|X\| - C'') \\ &\leq \lim_{n \rightarrow \infty} \mathbb{E}_{X \sim \mathcal{P}_g} (\phi(g_n(X)) + k\phi'(a_0)\|X\| - C'') \\ &= \lim_{n \rightarrow \infty} \mathbb{E}_{X \sim \mathcal{P}_g} \phi(g_n(X)) + \mathbb{E}_{X \sim \mathcal{P}_g} (k\phi'(a_0)\|X\| - C'') \end{aligned} \quad (17)$$

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

$$\begin{aligned} \mathfrak{G}(g) &= \mathbb{E}_{X \sim \mathcal{P}_g} \phi(g(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(g(Y)) \leq \lim_{n \rightarrow \infty} \mathbb{E}_{X \sim \mathcal{P}_g} \phi(g_n(X)) + \lim_{n \rightarrow \infty} \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(g_n(Y)) \\ &\leq \lim_{n \rightarrow \infty} (\mathbb{E}_{X \sim \mathcal{P}_g} \phi(g_n(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(g_n(Y))) = \inf_{\|f\|_{Lip} \leq k} \mathfrak{G}(f) \end{aligned} \quad (18)$$

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

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 $\|f\|_{Lip} \leq k$, 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 $\lim_{n \rightarrow \infty} \mathfrak{T}(f_n) = \inf(\mathfrak{T}(f))$. Without loss of generality, we assume that $f_n(0) = 0, \forall n \in \mathbb{N}^+$. Because $\|f_n\|_{Lip} \leq k$, 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 $\mathfrak{T}(f^*) = \inf_{\|f\|_{Lip} \leq k} \mathfrak{T}(f)$. \square

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:

- $\phi' \geq 0, \varphi' \leq 0$;
- 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 $\|f\|_{Lip} \leq k$, 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. We have proved most conditions in previous lemmas. And we only have to consider the condition that for any $x \in \mathbb{R}$, $\phi'(x) + \varphi'(x) \geq 0$ (or $\phi'(x) + \varphi'(x) \leq 0$) 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 $\phi'(x) + \varphi'(x) \geq 0$ for all x and there exists a_1 such that $\phi'(a_1) + \varphi'(a_1) > 0$. Then we know $\forall x \leq a_0$, $\phi'(x) + \varphi'(x) = 0$, which leads to $\forall x \leq a_0$, $\phi'(x) = -\varphi'(x)$. Thus, for any $x \leq a_0$, $0 \leq \phi''(x) = -\varphi''(x) \leq 0$, which means $\forall x \leq a_0$, $\phi(x) = -\varphi(x) = tx$, $t \geq 0$. Similar to the previous lemmas, we can get a series of functions $\{f_n\}_{n=1}^\infty$ such that $\lim_{n \rightarrow \infty} \mathfrak{G}(f_n) = \inf(\mathfrak{G}(f))$. 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 $f_n(0) \leq C$ with Lemma 2. On the other hand, when $C > k \cdot \text{diam}(\mathcal{S}_r \cup \mathcal{S}_g) + a_0$, 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 $\mathfrak{G}(f^*) = \inf_{\|f\|_{Lip} \leq k} \mathfrak{G}(f)$. \square

Lemma 6 (Theorem 1 Part I). *Under the same assumption of Lemma 5, we have $\mathfrak{F}(f) = \mathbb{E}_{X \sim \mathcal{P}_g} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(f(Y)) + \lambda \|f\|_{Lip}^\alpha$ 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 $\mathfrak{F}(f) = \mathfrak{G}(f) + \lambda \|f\|_{Lip}^\alpha \geq -\|f\|_{Lip} \varphi'(a_0) W_1(\mathcal{P}_r, \mathcal{P}_g) + \lambda \|f\|_{Lip}^\alpha$. 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 $\lim_{n \rightarrow \infty} \mathfrak{F}(f_n) = \inf_{f \in \text{dom } \mathfrak{F}} \mathfrak{F}(f)$. 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 $\lim_{n \rightarrow \infty} g_n = g$, then by Fatou's Lemma, we have $\mathfrak{G}(g) \leq \underline{\lim}_{n \rightarrow \infty} \mathfrak{G}(g_n)$.

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

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 $\mathfrak{F}(f) = \mathbb{E}_{X \sim \mathcal{P}_g} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi(f(Y)) + \lambda \|f\|_{Lip}^\alpha$ 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

$$\begin{aligned} \left\| \frac{f_1 + f_2}{2} \right\|_{Lip} &= \sup_{x,y} \frac{\frac{f_1(x)+f_2(x)}{2} - \frac{f_1(y)+f_2(y)}{2}}{\|x-y\|} \\ &\leq \sup_{x,y} \frac{1}{2} \frac{|f_1(x)-f_1(y)| + |f_2(x)-f_2(y)|}{\|x-y\|} \\ &\leq \frac{1}{2} \left(\sup_{x,y} \frac{|f_1(x)-f_1(y)|}{\|x-y\|} + \sup_{x,y} \frac{|f_2(x)-f_2(y)|}{\|x-y\|} \right) \\ &= \frac{1}{2} (\|f_1\|_{Lip} + \|f_2\|_{Lip}). \end{aligned} \tag{19}$$

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

$$\begin{aligned} \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha &\leq \lambda \left(\frac{1}{2} (\|f_1\|_{Lip} + \|f_2\|_{Lip}) \right)^\alpha \\ &\leq \lambda \frac{1}{2} (\|f_1\|_{Lip}^\alpha + \|f_2\|_{Lip}^\alpha). \end{aligned} \quad (20)$$

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

$$\begin{aligned} \mathfrak{G}\left(\frac{f_1 + f_2}{2}\right) &= \mathbb{E}_{X \sim \mathcal{P}_g} \phi\left(\frac{f_1 + f_2}{2}\right) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi\left(\frac{f_1 + f_2}{2}\right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\ &< \mathbb{E}_{X \sim \mathcal{P}_g} \left(\frac{\phi(f_1) + \phi(f_2)}{2} \right) + \mathbb{E}_{Y \sim \mathcal{P}_r} \varphi\left(\frac{f_1 + f_2}{2}\right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\ &\leq \mathbb{E}_{X \sim \mathcal{P}_g} \left(\frac{\phi(f_1) + \phi(f_2)}{2} \right) + \mathbb{E}_{Y \sim \mathcal{P}_r} \left(\frac{\varphi(f_1) + \varphi(f_2)}{2} \right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\ &\leq \mathbb{E}_{X \sim \mathcal{P}_g} \left(\frac{\phi(f_1) + \phi(f_2)}{2} \right) + \mathbb{E}_{Y \sim \mathcal{P}_r} \left(\frac{\varphi(f_1) + \varphi(f_2)}{2} \right) + \lambda \frac{1}{2} (\|f_1\|_{Lip}^\alpha + \|f_2\|_{Lip}^\alpha) \\ &= \frac{1}{2} (\mathfrak{G}(f_1) + \mathfrak{G}(f_2)) = \inf \mathfrak{G}(f) \end{aligned} \quad (21)$$

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

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

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

Proof.

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

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

$\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} = 0$ 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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} = 0$ 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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} \neq 0$, then it is obvious that $\frac{\partial J_D^*}{\partial k(f^*)} \neq 0$.

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

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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} = 0$, 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 $\mathcal{P}_r = \mathcal{P}_g$. \square

Proof of Theorem 2.

(a): Let k be the Lipschitz constant of f^* . Consider x with $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} \neq 0$. 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 $\frac{|f^*(z) - f^*(w)|}{\|z - w\|} = \frac{\int_w^z f'^*(t) dt}{\|z - w\|}$. Let $\epsilon \rightarrow 0$, we have $t \rightarrow x$. Then $|f'(t)| \rightarrow |f'(x)|$. Let $\delta \rightarrow 0$, we have $(k - \delta) \rightarrow k$. 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. $\lim_{n \rightarrow \infty} \frac{|f(y) - f(y_n)|}{\|y - y_n\|} = k(y)$. 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 $\epsilon_2 \rightarrow 0$ and $\delta_2 \rightarrow 0$. 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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} \neq 0$, without loss of generality, we can assume $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} > 0$. Then, for all $y \in B(x, \epsilon_2)$, we have $\frac{\partial \hat{J}_D(y)}{\partial f^*(y)} > 0$, as long as ϵ_2 is small enough.

Now we change the value of $f^*(y)$ for $y \in B(x, \epsilon_2)$. Let $g(y) = \begin{cases} f^*(y) - \frac{\epsilon_2}{N}(1 - \frac{\|x-y\|}{\epsilon_2}), & y \in B(x, \epsilon_2); \\ f^*(y) & \text{otherwise.} \end{cases}$. Because

$\frac{\partial \hat{J}_D(y)}{\partial f^*(y)} > 0$, $\forall y \in B(x, \epsilon_2)$, 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 $\frac{|g(y) - g(z)|}{\|y - z\|} \leq \frac{|f^*(y) - f^*(z)| + |\frac{\epsilon_2}{N}(1 - \frac{\|x-y\|}{\epsilon_2}) - \frac{\epsilon_2}{N}(1 - \frac{\|x-z\|}{\epsilon_2})|}{\|y - z\|} = \frac{|f^*(y) - f^*(z)| + \frac{\epsilon_2}{N}(\frac{\|x-y\| - \|x-z\|}{\|y-z\|})}{\|y - z\|} \leq \frac{|f^*(y) - f^*(z)|}{\|y - z\|} + \frac{1}{N} \frac{\|y-z\|}{\|y-z\|} = \frac{|f^*(y) - f^*(z)|}{\|y - z\|} + \frac{1}{N} \leq k - \delta_2 + \frac{1}{N} < k$ (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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} = \mathcal{P}_g(x) \frac{\partial \phi(f^*(x))}{\partial f^*(x)} + \mathcal{P}_r(x) \frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = \mathcal{P}_g(x) \frac{\partial \phi(f^*(x))}{\partial f^*(x)} > 0$, 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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} \neq 0$. 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 $\frac{\partial \hat{J}_D(x)}{\partial f(x)} \frac{\partial f(x)}{\partial x} = 0$. 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 $\frac{\partial \hat{J}_D(\hat{x})}{\partial f(\hat{x})} \neq 0$. 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 $\|\frac{\partial f(\hat{x})}{\partial \hat{x}}\| = k(f) \neq 0$. This is contradictory to that $\frac{\partial \hat{J}_D(\hat{x})}{\partial f(\hat{x})} \frac{\partial f(\hat{x})}{\partial \hat{x}} = 0$. 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. \square

Remark 1. For the Wasserstein distance, $\nabla_{f^*(x)} \hat{J}_D(x) = 0$ 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. \square

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

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

Proof. Consider x with $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} > 0$. 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 $\frac{\partial \hat{J}_D(y)}{\partial f^*(y)} \geq 0$. 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 as 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 $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} < 0$ such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$. Given the fact $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} > 0$ and $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} < 0$, 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. \square

Lemma 12. For any x , if $\frac{\partial \hat{J}_D(x)}{\partial f(x)} > 0$, then $\mathcal{P}_g(x) > 0$. For any y , if $\frac{\partial \hat{J}_D(y)}{\partial f(y)} < 0$, then $\mathcal{P}_r(y) > 0$.

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

Proof of Theorem 3.

For any $x \in \mathcal{S}_g$, if $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} > 0$, according to Lemma 11, there exists a y with $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} < 0$ 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. \square

Remark 2. $\frac{\partial \hat{J}_D(x)}{\partial f^*(x)} < 0$ 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 a $x' \in \mathcal{S}_g$ that bounds this point.

A.4. Proof of Theorem 4

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.

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 $\|\cdot\|_p$ and $f(y) - f(x) = k\|y - x\|_p$, then $f(x_t) = f(x) + t \cdot k\|y - x\|_p$

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

$$\begin{aligned} f(y) - f(x) &= f(y) - f(x_t) + f(x_t) - f(x) \\ &\leq f(y) - f(x_t) + k\|x_t - x\|_p = f(y) - f(x_t) + t \cdot k\|y - x\|_p \\ &\leq k\|y - x_t\|_p + t \cdot k\|y - x\|_p = k \cdot (1-t)\|y - x\|_p + t \cdot k\|y - x\|_p \\ &= k\|y - x\|_p. \end{aligned} \tag{22}$$

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

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 to k .

Proof.

$$\begin{aligned}\frac{\partial f(x_t)}{\partial v} &= \lim_{h \rightarrow 0} \frac{f(x_t + hv) - f(x_t)}{h} = \lim_{h \rightarrow 0} \frac{f(x_t + h\frac{y-x}{\|y-x\|_2}) - f(x_t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x_t + \frac{h}{\|y-x\|_2}) - f(x_t)}{h} = \lim_{h \rightarrow 0} \frac{\frac{h}{\|y-x\|_2} \cdot k\|y - x\|_2}{h} = k.\end{aligned}\quad \square$$

Proof of Theorem 4. Assume $p = 2$. According to (Adler & Lunz, 2018), if $f(x)$ is k -Lipschitz with respect to $\|\cdot\|_2$ and $f(x)$ is differentiable at x_t , then $\|\nabla f(x_t)\|_2 \leq k$. Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. We have

$$k^2 = k \frac{\partial f(x_t)}{\partial v} = k \langle v, \nabla f(x_t) \rangle = \langle kv, \nabla f(x_t) \rangle \leq \|kv\|_2 \|\nabla f(x_t)\|_2 = k^2. \quad (23)$$

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

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

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

The Wasserstein distance is given as follows

$$W_1(\mathcal{P}_r, \mathcal{P}_g) = \inf_{\pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g)} \mathbb{E}_{(x,y) \sim \pi} [d(x, y)], \quad (24)$$

where $\Pi(\mathcal{P}_r, \mathcal{P}_g)$ denotes the set of all probability measures with marginals \mathcal{P}_r and \mathcal{P}_g on the first and second factors, respectively. The Kantorovich-Rubinstein (KR) dual (Villani, 2008) is written as

$$\begin{aligned}W_{KR}(\mathcal{P}_r, \mathcal{P}_g) &= \sup_f \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g} [f(x)], \\ &\text{s.t. } f(x) - f(y) \leq d(x, y), \forall x, \forall y.\end{aligned} \quad (25)$$

We will prove that Wasserstein distance in its dual form can also be written as

$$\begin{aligned}W_{LL}(\mathcal{P}_r, \mathcal{P}_g) &= \sup_f \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g} [f(x)], \\ &\text{s.t. } f(x) - f(y) \leq d(x, y), \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g,\end{aligned} \quad (26)$$

which relaxes the constraint in the KR dual form of Wasserstein distance.

Theorem 5. Given $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$, we have $W_{KR}(\mathcal{P}_r, \mathcal{P}_g) = W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = W_1(\mathcal{P}_r, \mathcal{P}_g)$.

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

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

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:

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

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

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

B. The Practical Behaviors of Gradient Uninformativeness

To study the practical behaviors of gradient uninformativeness, we conducted a set of experiments with various hyper-parameter settings. We use the Least-Squares GAN in this experiments as an representative of traditional GANs. The value surface and the gradient of generated samples under various situations are plotted as follows.

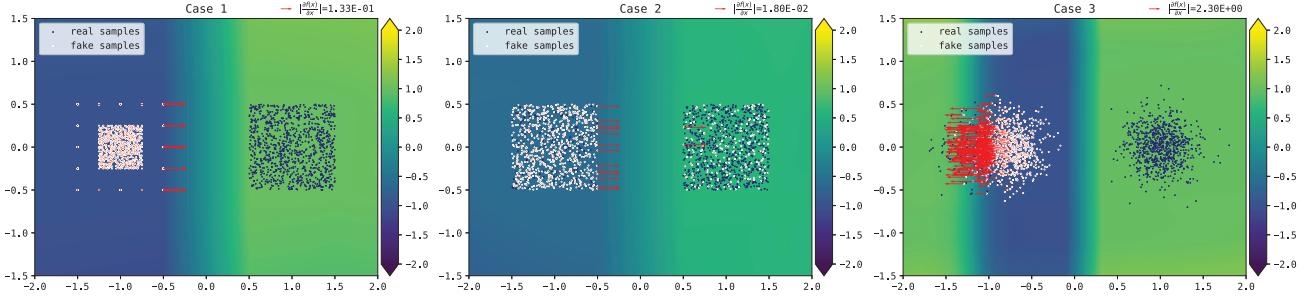


Figure 7: ADAM with lr=1e-2, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=1.

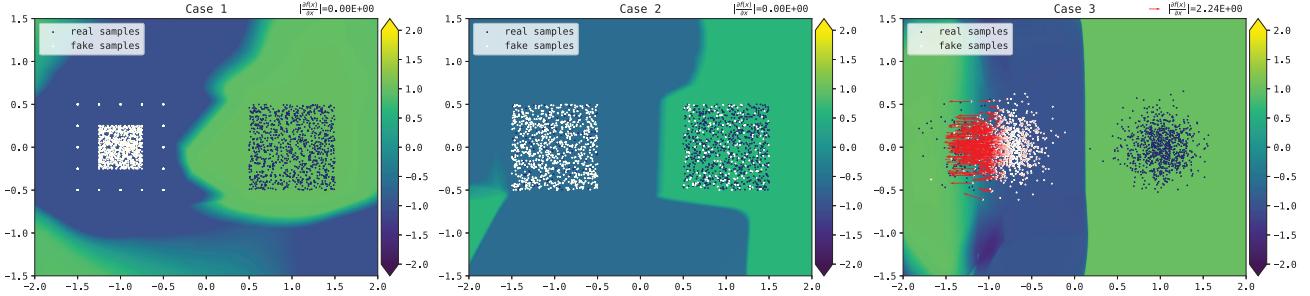


Figure 8: ADAM with lr=1e-2, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=4.

These experiments shown that the practical f highly depend on the hyper-parameter setting. Given limited capacity, the neural network try to learn the best f . When the neural network is capable of learning approximately the optimal f^* , how the actual f approaches f^* and how the points whose gradients are theoretically undefined behave highly depends the optimization details and the characteristics of the network.

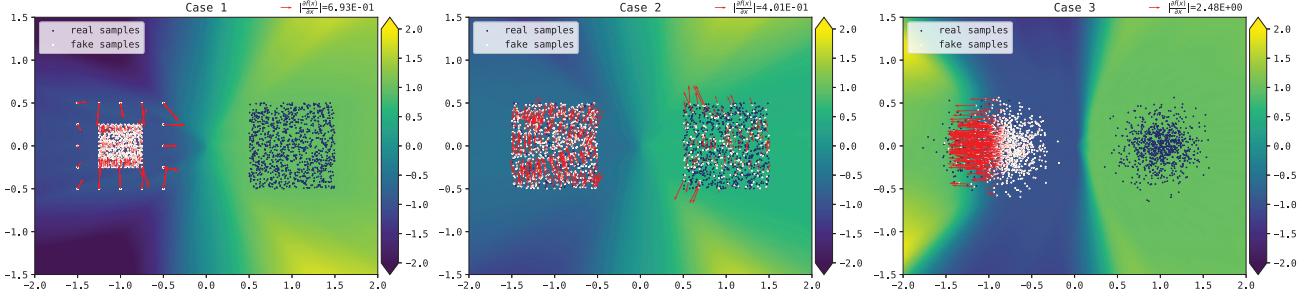


Figure 9: ADAM with lr=1e-5, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=4.

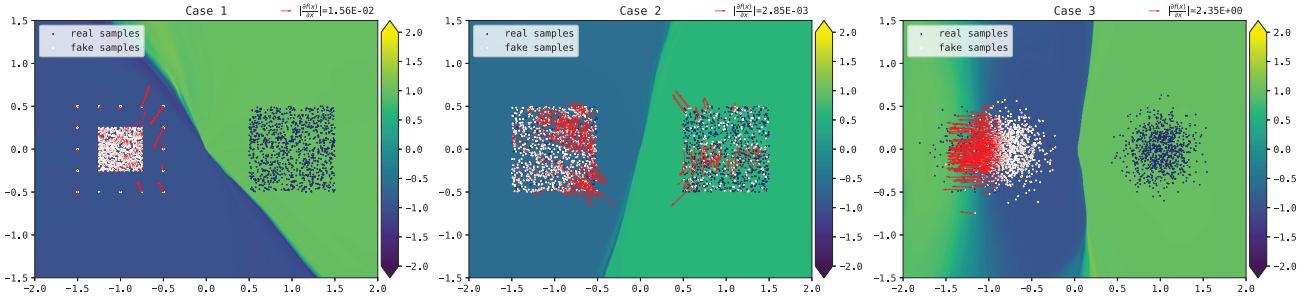


Figure 10: SGD with lr=1e-3. MLP with SELU activations, #hidden units=128, #layers=64.

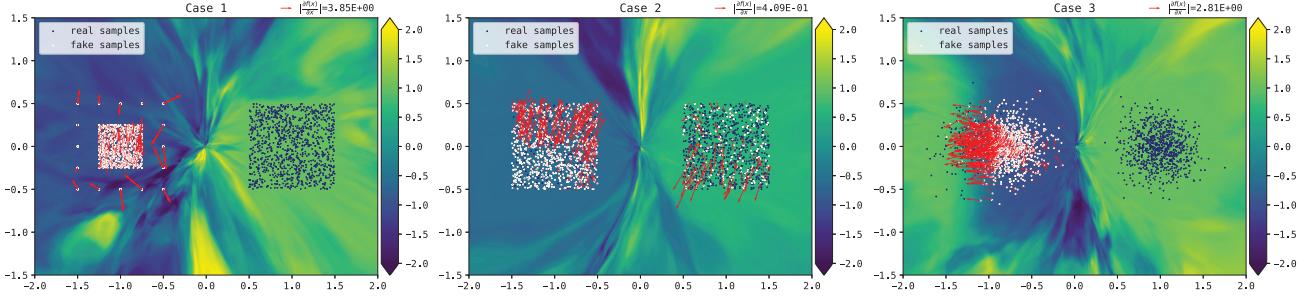


Figure 11: SGD with lr=1e-4. MLP with SELU activations, #hidden units=128, #layers=64.

C. On the Implementation of Lipschitz continuity for GANs

Typical techniques for enforcing k -Lipschitz includes: spectral normalization (Miyato et al., 2018), gradient penalty (Gulrajani et al., 2017), and Lipschitz penalty (Petzka et al., 2017). Before moving into the detailed discussion of these methods, we would like to provide several important notes in the first place.

Firstly, enforcing k -Lipschitz in the blending-region of \mathcal{P}_r and \mathcal{P}_g is actually sufficient.

Define $B(\mathcal{S}_r, \mathcal{S}_g) = \{\hat{x} = x \cdot t + y \cdot (1-t) \mid x \in \mathcal{S}_r \text{ and } y \in \mathcal{S}_g \text{ and } t \in [0, 1]\}$. It is clear that f is 1-Lipschitz in $B(\mathcal{S}_r, \mathcal{S}_g)$ implies $f(x) - f(y) \leq d(x, y), \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g$. Thus, it is a sufficient constraint for Wasserstein distance in Eq. (4). In fact, $f(x)$ is k -Lipschitz in $B(\mathcal{P}_r, \mathcal{P}_g)$ is also a sufficient condition for all properties described in Lipschitz GANs.

Secondly, enforcing k -Lipschitz with regularization would provide a dynamic Lipschitz constant k .

Lemma 15. *With Wasserstein GAN objective, we have $\min_{f \in \mathcal{F}_{k,Lip}} J_D(f) = k \cdot \min_{f \in \mathcal{F}_{1,Lip}} J_D(f)$.*

Assuming we can directly control the Lipschitz constant $k(f)$ of f , the total loss of the discriminator becomes $J(k) \triangleq \min_{f \in \mathcal{F}_{k,Lip}} J_D(f) + \lambda \cdot (k - k_0)^2$. With Lemma 15, let $\alpha = -\min_{f \in \mathcal{F}_{1,Lip}} J_D(f)$, then $J(k) = -k \cdot \alpha + \lambda \cdot (k - k_0)^2$, and $J(k)$ achieves its minimum when $k = \frac{\alpha}{2\lambda} + k_0$. When α goes to zero, i.e., \mathcal{P}_g converges to \mathcal{P}_r , the optimal k decreases. And when $\mathcal{P}_r = \mathcal{P}_g$, we have $\alpha = 0$ and the optimal $k = k_0$. The similar analysis applies to Lipschitz GANs.

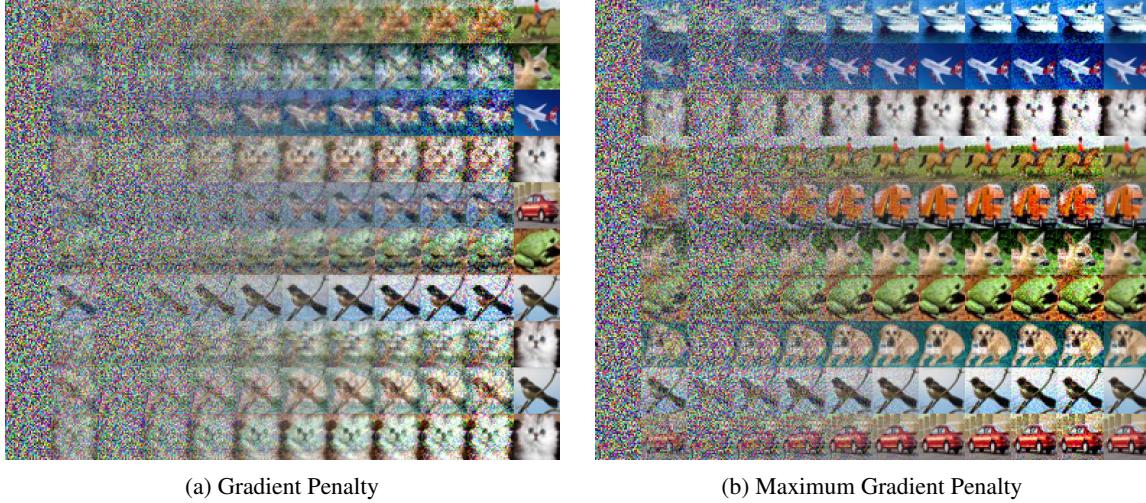


Figure 12: Comparison between gradient penalty and maximum gradient penalty, with \mathcal{P}_r and \mathcal{P}_g consist of ten real and noise images, respectively. The leftmost in each row is a $x \in \mathcal{S}_g$ and the second is its gradient $\nabla_x f^*(x)$. The interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing ϵ , which will pass through a real sample, and the rightmost is the nearest $y \in \mathcal{S}_r$.

C.1. Existing Methods

For practical methods, though spectral normalization (Miyato et al., 2018) recently demonstrates their excellent results in training GANs, spectral normalization is an absolute constraint for Lipschitz over the entire space, i.e., constricting the maximum gradient of the entire space, which is unnecessary. On the other side, we also notice both penalty methods proposed in (Gulrajani et al., 2017) and (Petzka et al., 2017) are not exact implementation of the Lipschitz continuity condition, because it does not directly penalty the maximum gradient, but penalties all gradients towards the given target Lipschitz constant or penalties all these greater than one towards the given target.

We also empirically found that the existing methods including spectral normalization (Miyato et al., 2018), gradient penalty (Gulrajani et al., 2017), and Lipschitz penalty (Petzka et al., 2017) all fail to converge to the optimal $f^*(x)$ in some of our synthetic experiments.

C.2. The New Method

Note that this practical method of imposing Lipschitz continuity is not the key contribution of this work. We leave the more rigorous study on this topic as our further work. We introduce it for the necessity for understanding our paper and reproducing of experiments.

Combining the idea of spectral normalization and gradient penalty, we developed a new way of implementing the regularization of Lipschitz continuity in our experiments. Spectral normalization is actually constraining the maximum gradient over the entire space. And as we argued previously, enforcing Lipschitz continuity in the blending region is sufficient. Therefore, we propose to restricting the maximum gradient over the blending region:

$$J_{\text{maxgp}} = \lambda \max_{x \sim B(\mathcal{S}_r, \mathcal{S}_g)} [\|\nabla_x f(x)\|^2] \quad (27)$$

In practice, we sample x from $B(\mathcal{S}_r, \mathcal{S}_g)$ as in (Gulrajani et al., 2017; Petzka et al., 2017) using training batches of real and fake samples.

We compare the practical result of (centralized) gradient penalty $\mathbb{E}_{x \sim B} [\|\nabla_x f(x)\|^2]$ and the proposed maximum gradient penalty in Figure 12. Before switching to maximum gradient penalty, we struggled for a long time and cannot achieve a high quality result as shown in Figure 12b. The other forms of gradient penalty (Gulrajani et al., 2017; Petzka et al., 2017) perform similar as $\mathbb{E}_{x \sim B} [\|\nabla_x f(x)\|^2]$.

To improve the stability and reduce the bias introduced via batch sampling, one can further keep track x with the maximum $\|\nabla_x f(x)\|$. A practical and light weight method is to maintain a list S_{max} that has the currently highest (top-k) $\|\nabla_x f(x)\|_2$

(initialized with random samples), use the S_{\max} as part of the batch that estimates J_{maxgp} , and update the S_{\max} after each batch updating of the discriminator. According to our experiments, it is usually does not improve the training significantly.

D. Extended Discussions and More Details

D.1. Various ϕ and φ That Satisfies Eq. (11)

For Lipschitz GANs, ϕ and φ are required to satisfy Eq. (11). Eq. (11) is actually quite general and there exists many other instances, e.g., $\phi(x) = \varphi(-x) = x$, $\phi(x) = \varphi(-x) = -\log(\sigma(-x))$, $\phi(x) = \varphi(-x) = x + \sqrt{x^2 + \alpha}$ with $\alpha > 0$, $\phi(x) = \varphi(-x) = \exp(x)$, etc. We plot these instances of ϕ and φ in Figure 13.

To devise a loss satisfies Eq. (11), it is practical to let ϕ be an increasing function with non-decreasing derivative and set $\phi(x) = \varphi(-x)$. 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).

D.2. Experiment Details

In our experiments with real datas (CIFAR-10, Tiny Imagenet and Oxford 102), we follow the network architecture and hyper-parameters in (Gulrajani et al., 2017). The network architectures are detailed in Table 3. We use Adam optimizer with beta1=0.0, beta2=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. We use MaxGP for all our experiments of LGANs and search the best penalty weight λ in [0.01, 0.1, 1.0, 10.0]. Please check more details in our codes. For all experiments in Table 2, we only change ϕ and φ and the dataset, and all other components are fixed.

We plot the IS training curve of LGANs in Figure 14 and 15. We provide the visual results of LGANs in Figure 16, 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 18.

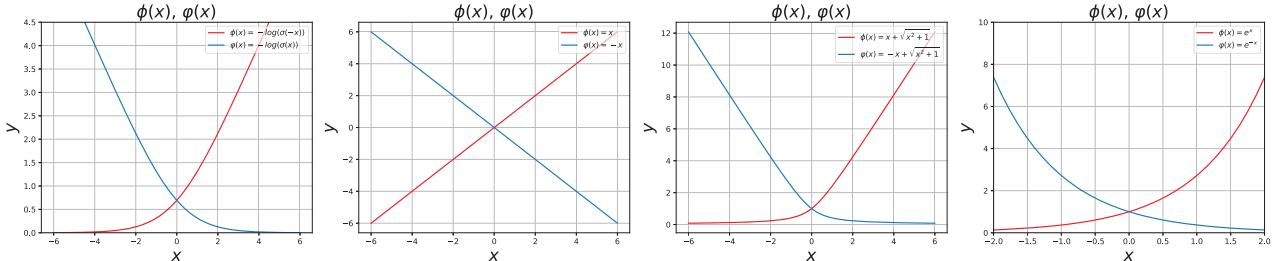


Figure 13: Various ϕ and φ that satisfies Eq. (11).

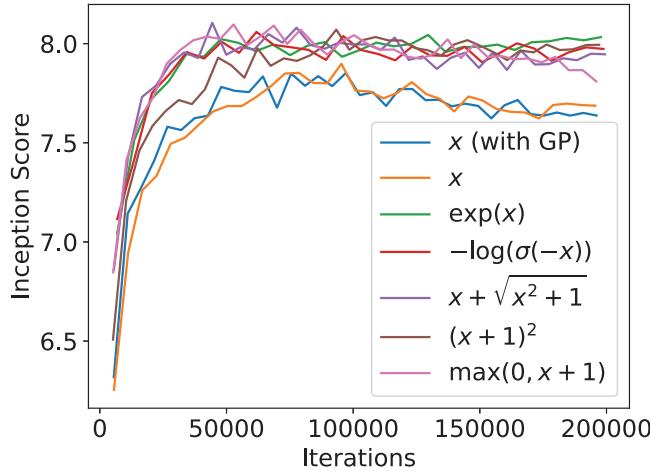


Figure 14: IS training curves on CIFAR-10.

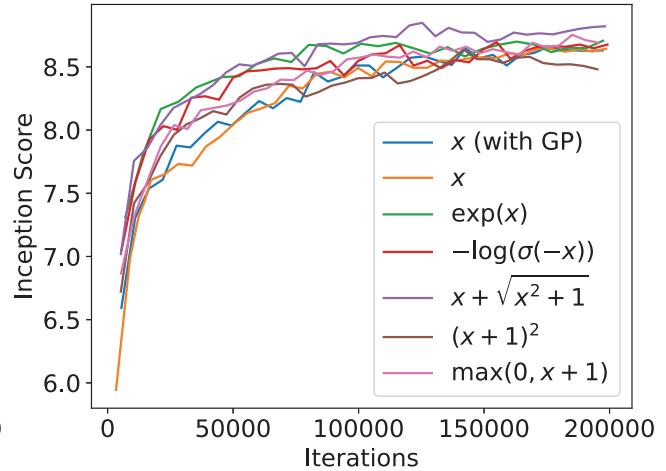


Figure 15: IS training curves on Tiny ImageNet.



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



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



Figure 18: Random samples of LGANs with different loss metrics on Oxford 102.

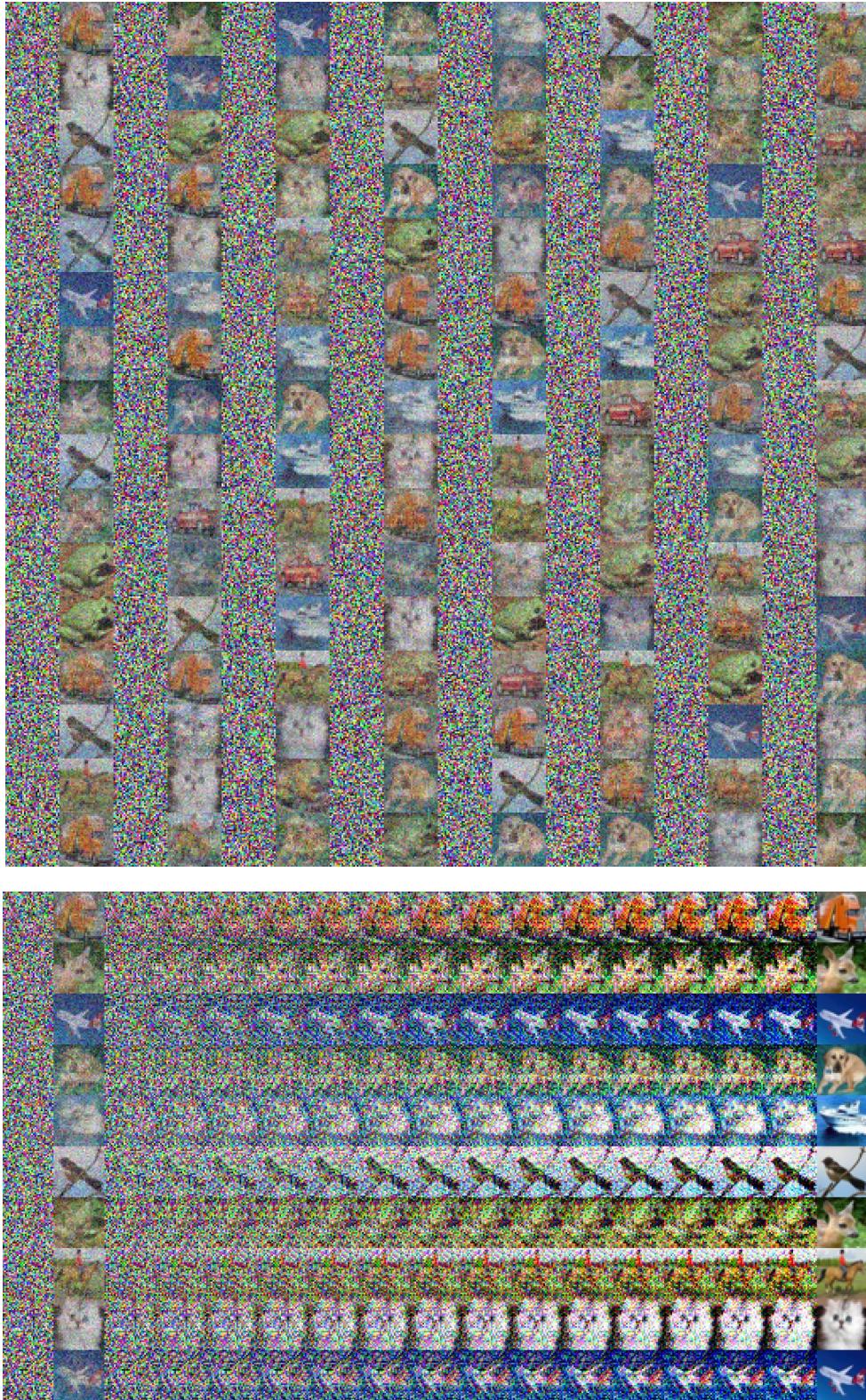


Figure 19: The gradient of LGANs with real world data, where \mathcal{P}_r consists of ten images and \mathcal{P}_g is Gaussian noise. Up: Each odd column are $x \in \mathcal{S}_g$ and the nearby column are their gradient $\nabla_x f^*(x)$. Down: the leftmost in each row is $x \in \mathcal{S}_g$, the 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 $y \in \mathcal{S}_r$.

Generator:				Discriminator:			
Operation	Kernel	Resample	Output Dims	Operation	Kernel	Resample	Output Dims
Noise	N/A	N/A	128	Residual Block	$3 \times 3 \times 2$	Down	$128 \times 16 \times 16$
Linear	N/A	N/A	$128 \times 4 \times 4$	Residual Block	$3 \times 3 \times 2$	Down	$128 \times 8 \times 8$
Residual block	3×3	UP	$128 \times 8 \times 8$	Residual Block	$3 \times 3 \times 2$	N/A	$128 \times 8 \times 8$
Residual block	3×3	UP	$128 \times 16 \times 16$	Residual Block	$3 \times 3 \times 2$	N/A	$128 \times 8 \times 8$
Residual block	3×3	UP	$128 \times 32 \times 32$	ReLU,mean pool	N/A	N/A	128
Conv & Tanh	3×3	N/A	$3 \times 32 \times 32$	Linear	N/A	N/A	1

Table 3: The network architectures.