Lipschitz Generative Adversarial Nets

Zhiming Zhou 1  Jiadong Liang 2  Yuxuan Song 1  Lantao Yu 3  Hongwei Wang 3  Weinan Zhang 1  Yong Yu 1  Zhihua Zhang 2

Abstract

In this paper, we study the convergence of generative adversarial networks (GANs) from the perspective of the informativeness of the gradient of the optimal discriminative function. We show that 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 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; 2018; Yadav et al., 2017), are still not fully understood.

The objective of GAN is usually defined as or proved equivalent to a distance metric between the real distribution $P_r$ and the generative distribution $P_g$, which implies that $P_r = P_g$ 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 $P_r$ and $P_g$ 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 $P_g$ towards $P_r$. 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 $P_g$ where $P_r = P_g$. It
leads to a new family of GANs that we call Lipschitz GANs (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 \(f : X \to Y\) is said to be Lipschitz continuous if there exists a constant \(k \geq 0\) such that

\[ d_Y(f(x_1), f(x_2)) \leq k \cdot d_X(x_1, x_2), \forall x_1, x_2 \in X. \]  

(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

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

(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 \(\pi^*\) 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

\[ W_1(\mathcal{P}_r, \mathcal{P}_g) = \sup_f \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{y \sim \mathcal{P}_g} [f(y)], \]  

\[ \text{s.t. } f(x) - f(y) \leq d(x, y), \forall x, \forall y. \]  

(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,

\[ W_1(\mathcal{P}_r, \mathcal{P}_g) = \sup_{\phi, \varphi} \mathbb{E}_{x \sim \mathcal{P}_r} [\phi(f(x))] - \mathbb{E}_{y \sim \mathcal{P}_g} [\varphi(f(x))], \]  

\[ \text{s.t. } f(x) - f(y) \leq d(x, y), \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g. \]  

(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

\[ \min_{f \in \mathcal{F}} J_D \triangleq \mathbb{E}_{z \sim \mathcal{P}_z} [\phi(f(g(z)))] + \mathbb{E}_{x \sim \mathcal{P}_r} [\varphi(f(x))], \]  

\[ \min_{g \in \mathcal{G}} J_G \triangleq \mathbb{E}_{z \sim \mathcal{P}_z} [\psi(f(g(z)))]\]  

(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 \(g : \mathbb{R}^m \to \mathbb{R}^n\) learns to output samples that share the same dimension as samples in \(\mathcal{P}_r\), while the discriminative function \(f : \mathbb{R}^n \to \mathbb{R}\) learns to output a score indicating the authenticity of a given sample. Here \(\mathcal{F}\) and \(\mathcal{G}\) denote discriminative and generative function spaces, respectively; and \(\phi, \varphi, \psi : \mathbb{R} \to \mathbb{R}\) are loss metrics. We denote the implicit distribution of the generated samples by \(\mathcal{P}_g\).

We list the choices of \(\mathcal{F}, \phi\) and \(\varphi\) 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

\[ \nabla_x J_G(x) \triangleq \nabla_x \psi(f(x)) = \nabla f(x) \psi(f(x)) \cdot \nabla f(x), \]  

(6)

where the first term \(\nabla f(x) \psi(f(x))\) is a step-related scalar, and the second term \(\nabla f(x)\) 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 \(J_D(x) \triangleq \mathcal{P}_g(x)\phi(f(x)) + \mathcal{P}_r(x)\varphi(f(x)).\) It has \(J_D = \int J_D(x) dx\).
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 $S_r$ and $S_g$ are usually disjoint and the gradient vanishing stems from the ill-behaving of traditional distance metrics, i.e., the distance between $P_r$ and $P_g$ 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 Un informativeness

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 $P_r$. Consequently, $P_g$ is not guaranteed to converge to $P_r$. 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 $P_r(x)$ and $P_g(x)$:

$$f^*(x) = \arg\min_{f(x)\in \mathbb{R}} P_g(x)\varphi(f(x)) + P_r(x)\psi(f(x)), \forall x.$$  

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 $y \notin S_r$), $f^*(x)$ in the surrounding of $x$ would contain no information about $P_r$. Thus $\nabla_x f^*(x)$, the gradient that $x$ receives from the optimal discriminative function, does not reflect any information about $P_r$.

Typical situation is that $S_r$ and $S_g$ 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: $S_r$ and $S_g$ 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 $\mu$-Fisher IPM $F_\mu(P_r, P_g)$, the generalized objective of the Fisher GAN (Mroueh et al., 2017), has the following form:

$$f^*(x) = \frac{1}{F_\mu(P_r, P_g)} \frac{P_r(x) - P_g(x)}{\mu(x)},$$  

(7)

where $\mu$ is a distribution whose support covers $S_r$ and $S_g$, and $\frac{1}{\sqrt{P_r P_g}}$ is a constant. It can be observed that $\mu$-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 $\pi^*$ 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

$$P_{(x, y) \sim \pi^*} \left[\nabla x f^*(x_t) = \frac{y - x}{\|y - x\|}\right] = 1.$$  

(8)
This proposition indicates: (i) for each generated sample \( x \), there exists a real sample \( y \) such that \( \nabla_x f^*(x) = \frac{y - x}{\|y - x\|} \) for all linear interpolations \( x_t \) between \( x \) and \( y \), i.e., the gradient at any \( x \) is pointing towards the real sample \( y \); (ii) these \((x, y)\) pairs match the optimal coupling \( \pi^* \) in the optimal transport perspective. It implies that WGAN is able to overcome the gradient un informativeness as well as the gradient vanishing.

Our concern turns to the reason why WGAN can avoid gradient un informativeness. 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

\[
    f^*(x) = \begin{cases} 
        1, & \forall x \in \mathcal{S}_r; \\
        0, & \forall x \in \mathcal{S}_g.
    \end{cases}
\] (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 \( \mathcal{P}_r \). More extremely, we can consider the case where \( \mathcal{S}_g \) are isolated points.

These examples imply that Lipschitz condition would be critical for resolving the gradient un informativeness 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:

\[
    \min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathcal{P}_r}[\phi(f(g(z)))] + \mathbb{E}_{x \sim \mathcal{P}_r}[\psi(f(x))] + \lambda \cdot k(f)^2,
\]

\[
    \min_{g \in \mathcal{G}} \mathbb{E}_{g \sim \mathcal{P}_r}[\psi(f(g(z)))).
\] (10)

In this work, we further assume that the loss functions \( \phi \) and \( \psi \) satisfy the following conditions:

\[
    \begin{align*}
    & \phi'(x) > 0, \phi''(x) < 0, \\
    & \phi''(x) > 0, \phi'''(x) \geq 0, \\
    & \exists a, \phi'(a) + \phi'(a) = 0.
    \end{align*}
\] (11)

The assumptions for the losses \( \phi \) and \( \psi \) 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 \( \phi \) be 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 15 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 un informativeness 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 \( \phi \) or \( \varphi \) 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 \( \phi'(x) > 0, \phi'(x) < 0 \), 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 \( \nabla f^*(x) \gamma \mathcal{P}_d(x) \neq 0 \), 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 \( S_r = S_g \) and \( P_r \neq P_g \), then there exists \((x, y)\) pair with both points in \( S_r \cup S_g \) and \( y \neq x \) such that \( |f(y) - f(x)| = k(f^*) \cdot \|y - x\| \) and \( \nabla f(x) J_D(x) \neq 0 \);

(d) There is a unique Nash equilibrium under the objective \( J_D \)+ \( \lambda \cdot k(f^2) \), where it holds that \( P_r = P_g \) 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 \( P_r = P_g \) 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, the latter, as we will show next, eliminates the gradient uninformativeness 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 \( \nabla f(x) J_D(x) = 0 \) for \( P_r(x) \neq P_g(x) \). Minimizing \( k(f) \) guarantees that the only Nash equilibrium is achieved when \( P_r = P_g \). In WGAN, minimizing \( k(f) \) is not necessary. However, if \( k(f) \) is not minimized towards zero, \( \nabla f^*(x) \) is not guaranteed to be zero at the convergence state \( P_r = P_g \) 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 \( J_D(x) \) 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 \( x \in S_r \), if \( \nabla f(x) J_D(x) > 0 \), then there must exist some \( y \in S_r \), with \( y \neq x \) such that \( f^*(y) - f^*(x) = k(f^*) \cdot \|y - x\| \) and \( \nabla f(y) J_D(y) < 0 \);

2) For any \( y \in S_r \), if \( \nabla f(y) J_D(y) < 0 \), then there must exist some \( x \in S_g \) with \( y \neq x \) such that \( f^*(y) - f^*(x) = k(f^*) \cdot \|y - x\| \) and \( \nabla f(x) J_D(x) > 0 \).

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 \( J_D(x) \). So, when there is strict bounding relationship (i.e., it involves points that hold \( \nabla f(x) J_D(x) \neq 0 \)), it must involve both real and fake samples. It is worth noticing that except the overlapping case, all fake samples hold \( \nabla f(x) J_D(x) > 0 \), while all real samples hold \( \nabla f(y) J_D(y) < 0 \).

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 \( P_r \) and \( P_g \), and each fake sample with \( \nabla f(x) J_D(x) \neq 0 \) 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 \( \nabla f^*(x) \) for all involved points.

4.3. The Implication of Bounding Relationship

Recall that the Proposition 1 states that \( \nabla x f(x) = \frac{y - x}{\|y - x\|} \). 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 \( \nabla x f(x) = k \cdot \frac{y - x}{\|y - x\|} \) for all \( x = tx + (1 - t)y \) 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 \( S_r \) and \( S_g \) are disjoint, the gradient \( \nabla x f^*(x) \) for each generated sample \( x \in S_g \) points towards some real sample \( y \in S_r \), which guarantees that \( \nabla x f^*(x) \)-based updating would pull \( P_g \) towards \( P_r \) at every step.

In fact, Theorem 2 provides further guarantee on the convergence. Property-(b) implies that for any generated sample \( x \in S_g \) that does not lie in \( S_r \), its gradient \( \nabla x f^*(x) \) must point towards some real sample \( y \in S_r \). And in the fully overlapped case, according to Property-(c), unless \( P_r = P_g \), 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 \( P_r = P_g \) where \( \nabla x f^*(x) = 0 \) for all generated samples.

5. Empirical Analysis

In this section, we empirically study the gradient uninformativeness 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 uninformativeness. Note that the behaviors of GANs without restriction on \( F \) are essentially identical. We choose the Least-Squares GAN whose \( f^* \) is relatively simple as the representative and study it with a set
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 \( S_r \cup S_g \), \( \nabla_x f^*(x) \) 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.6.

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). See more details in Appendix B.4.

### 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 \( \phi \) and \( \varphi \) 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 \( \epsilon \) 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 \( \phi \) and \( \varphi \) 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 \( \phi \) and \( \varphi \) 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) (Paszke 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 \( h(f) \)
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:

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

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 $\phi$ and $\varphi$ is critical in Theorem 2 to theoretically guarantee the existences of bounding relationships for arbitrary datas. But if we further assume $S_r$ and $S_y$ are limited, it is possible that there exists a suitable $\lambda$ such that all real and fake samples lie in a strict monotone region of $\phi$ and $\varphi$: for the hinge loss, it would mean $2\alpha < k(f) \cdot \|y - x\|$ for all $y \in S_r$ and $x \in S_y$.

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 $\lambda$ 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 $\lambda$ 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 un informativeness 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
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 an 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 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) has provided a thorough study of GANs with Lipschitz density assumption on data distribution.

### 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 \( P_r \). That is, in a sense, there is no guarantee that \( P_g \) will converge to \( P_r \). Typical case is that \( P_r \) and \( P_g \) 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 \( P_r = P_g \) where \( k(f^*) = 0 \). Our experiments show LGANs lead to more stable discriminative functions and achieve higher sample qualities.
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. and Zhang, Y. Do GANs actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 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.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

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.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

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

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., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 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.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. 2018.

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.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. Solving approximate Wasserstein GANs to stationarity. *arXiv preprint arXiv: 1802.08249*, 2018.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

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 $X \sim P_r, Y \sim P_r$. Assume $\mathbb{E}_{X \sim P_r} ||X|| < \infty$ and $\mathbb{E}_{Y \sim P_r} ||Y|| < \infty$. Let $\mathcal{G}(f) = \mathbb{E}_{X \sim P_r} \phi(f(X)) + \mathbb{E}_{Y \sim P_r} \varphi(f(Y))$. Let $\|f\|_{\text{Lip}}$ denote the Lipschitz constant of $f$. Let $S_r$ and $S_g$ denote the supports of $P_r$ and $P_g$, respectively. Let $W_1(P_r, P_g)$ denote the 1-st Wasserstein distance between $P_r$ and $P_g$.

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

**Proof.** Given that $\phi, \varphi$ are convex functions, we have

$$
\mathcal{G}(f) = \mathbb{E}_{X \sim P_r} \phi(f(X)) + \mathbb{E}_{Y \sim P_r} \varphi(f(Y)) \\
\geq \mathbb{E}_{X \sim P_r} (\phi'(a_0)(f(x) - a_0) + \phi(a_0)) + \mathbb{E}_{Y \sim P_r} (\varphi'(a_0)(f(x) - a_0) + \varphi(a_0)) \\
= \phi'(a_0)\mathbb{E}_{X \sim P_r} f(x) + \varphi'(a_0)\mathbb{E}_{Y \sim P_r} f(Y) + C \\
= (\phi'(a_0) + \varphi'(a_0))\mathbb{E}_{X \sim P_r} f(X) + (\varphi'(a_0))\mathbb{E}_{Y \sim P_r} f(Y) - \mathbb{E}_{X \sim P_r} f(X) + C \\
= k\varphi'(a_0)(\mathbb{E}_{Y \sim P_r} 1_k f(Y) - \mathbb{E}_{X \sim P_r} 1_k f(X)) + C \\
\geq -k\varphi'(a_0)W_1(P_r, P_g) + C.
$$

Therefore, we get the lower bound. \hfill \Box

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

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

**Proof.** Since $\phi, \varphi$ are convex functions, we have

$$
\mathcal{G}(f) = \mathbb{E}_{X \sim P_r} \phi(f(X)) + \mathbb{E}_{Y \sim P_r} \varphi(f(Y)) \\
\geq \mathbb{E}_{X \sim P_r} (\phi'(a_1)(f(x) - a_1) + \phi(a_1)) + \mathbb{E}_{Y \sim P_r} (\varphi'(a_1)(f(x) - a_1) + \varphi(a_1)) \\
= \phi'(a_1)\mathbb{E}_{X \sim P_r} f(x) + \varphi'(a_1)\mathbb{E}_{Y \sim P_r} f(Y) + C_1 \\
= (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{X \sim P_r} f(X) + \varphi'(a_1)\mathbb{E}_{Y \sim P_r} f(Y) - \mathbb{E}_{X \sim P_r} f(X) + C_1 \\
= (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{X \sim P_r} f(X) + k\varphi'(a_1)(\mathbb{E}_{Y \sim P_r} 1_k f(Y) - \mathbb{E}_{X \sim P_r} 1_k f(X)) + C_1 \\
\geq (\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{X \sim P_r} f(X) - k\varphi'(a_1)W_1(P_r, P_g) + C_1 \\
\geq (\phi'(a_1) + \varphi'(a_1))f(0) - k(\phi'(a_1) + \varphi'(a_1))\mathbb{E}_{X \sim P_r} ||X|| - k\varphi'W_1(P_r, P_g) + C_1.
$$

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

**Lemma 3.** Let $\phi$ and $\varphi$ be two convex functions, whose domains are both $\mathbb{R}$. If $\phi$ and $\varphi$ 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 $\mathcal{G}(f) = \mathbb{E}_{X \sim P_r} \phi(f(X)) + \mathbb{E}_{Y \sim P_g} \varphi(f(Y))$, where $f$ is subject to $\|f\|_{\text{Lip}} \leq k$, has global minima.

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

- $\|f^*\|_{\text{Lip}} \leq k$;
- $\forall f$ s.t. $\|f\|_{\text{Lip}} \leq k$, we have $\mathcal{G}(f^*) \leq \mathcal{G}(f)$.
We denote \( \phi \). That is, \( \phi \) in Lemma 3 and find the optimal function \( f \) such that \( \| \{ f_{m} \} \| \leq C \). Furthermore, for all \( \{ f_{m} \} \), we claim that \( \{ f_{m} \} \) converges at \( x \). Actually, \( \epsilon > 0 \), find \( r \in \{ r_{i} \} \) such that \( \| x - r \| \leq \frac{\epsilon}{10} \), we have

\[
\lim_{m,l \to \infty} |f_{mm}(x) - f_{ll}(x)| \leq \lim_{m,l \to \infty} \left( |f_{mm}(x) - f_{mm}(r)| + |f_{mm}(r) - f_{ll}(r)| + |f_{ll}(r) - f_{ll}(x)| \right)
\]

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

We denote \( \{ f_{mn} \} \) as \( \{ g_{n} \} \) and \( \{ g_{n} \} \) 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 \( \phi' \geq 0 \), \( \phi' \leq 0 \), we have

\[
\phi(g_{n}(x)) \geq \phi(g_{n}(0) - k\|x\|) \geq \phi(-C' - k\|x\|) \geq \phi'(0)(-C' - k\|x\| - a_{0}) + \phi(a_{0}) = -k\phi'(0)\|x\| + C''
\]

That is, \( \phi(g_{n}(x)) + k\phi'(0)\|x\| - C'' \geq 0 \).

By Fatou’s Lemma,

\[
\mathbb{E}_{X \sim P_{Y}}(\phi(g(X)) + k\phi'(0)\|X\| - C'') = \mathbb{E}_{X \sim P_{Y}} \lim_{n \to \infty} (\phi(g_{n}(X)) + k\phi'(0)\|X\| - C'')
\]

\[
\leq \lim_{n \to \infty} \mathbb{E}_{X \sim P_{Y}} (\phi(g_{n}(X)) + k\phi'(0)\|X\| - C'')
\]

\[
= \lim_{n \to \infty} \mathbb{E}_{X \sim P_{Y}} \phi(g_{n}(X)) + \mathbb{E}_{X \sim P_{Y}} (k\phi'(0)\|X\| - C'')
\]

It means \( \mathbb{E}_{X \sim P_{Y}} \phi(g(X)) \leq \lim_{n \to \infty} \mathbb{E}_{X \sim P_{Y}} \phi(g_{n}(X)) \). Similarly, we have \( \mathbb{E}_{Y \sim P_{Y}} \varphi(g(Y)) \leq \lim_{n \to \infty} \mathbb{E}_{Y \sim P_{Y}} \varphi(g_{n}(Y)) \). Combining the two inequalities, we have

\[
\mathbb{E} \leq \lim_{n \to \infty} \mathbb{E}_{X \sim P_{Y}} \phi(g_{n}(X)) + \mathbb{E}_{Y \sim P_{Y}} \varphi(g_{n}(Y))
\]

\[
\leq \lim_{n \to \infty} \mathbb{E}_{X \sim P_{Y}} \phi(g_{n}(X)) + \mathbb{E}_{Y \sim P_{Y}} \varphi(g_{n}(Y)) = \inf_{\|f\|_{Lip} \leq k} \mathbb{E} \phi(g)
\]

Note that for any \( x, y \in \text{dom}(g) \), \( |g(x) - g(y)| \leq \lim_{n \to \infty} (|g(x) - g_{n}(x)| + |g_{n}(x) - g_{n}(y)| + |g_{n}(y) - g(y)|) \leq k\|x - y\| \). That is, \( \|g\|_{Lip} \leq k \), \( \mathbb{E} \phi(g) = \inf_{\|f\|_{Lip} \leq k} \mathbb{E} \phi(f) \).

**Lemma 4** (Wasserstein distance). \( \mathbb{E} = \mathbb{E}_{X \sim P_{Y}} f(X) - \mathbb{E}_{Y \sim P_{Y}} f(Y) \), where \( f \) is subject to \( \|f\|_{Lip} \leq k \), has global minima.

**Proof.** It is easy to find that for any \( C \in \mathbb{R} \), \( \mathbb{E} + C = \mathbb{E} \). Similar to the previous lemma, we can get a series of functions \( \{ f_{n} \} \) such that \( \lim_{n \to \infty} \mathbb{E} = \inf \mathbb{E} \). 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 \( x \in \mathbb{R} \), \( \{ f_{n}(x) \} \) is bounded. Then we can imitate the method used in Lemma 3 and find the optimal function \( f^{*} \) such that \( \mathbb{E}(f^{*}) = \inf_{\|f\|_{Lip} \leq k} \mathbb{E}(f) \).

**Lemma 5.** Let \( \phi \) and \( \varphi \) be two convex functions, whose domains are both \( \mathbb{R} \). If we further suppose that the support sets \( S_{r} \) and \( S_{\gamma} \) are bounded. Then if \( \phi \) and \( \varphi \) satisfy the following properties:

- \( \phi' \geq 0 \), \( \varphi' \leq 0 \);
- \( \text{There is } a_{0} \in \mathbb{R} \text{ such that } \phi'(a_{0}) + \varphi'(a_{0}) = 0 \).
Lipschitz Generative Adversarial Nets

We have \( \mathcal{G}(f) = \mathbb{E}_{X \sim \mathcal{P}_x} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_y} \varphi(f(Y)) \), where \( f \) is subject to \( \|f\|_{\text{Lip}} \leq k \), has global minima.

That is, \( \exists f^* \), s.t.

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

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 \( \lim_{n \to \infty} \mathcal{G}(f_n) = \inf_{f \in \mathcal{D}} \mathcal{G}(f) \). 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 \( C > k \), \( \text{diam}(\mathcal{S}_2) + a_0 \), 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, \( \mathcal{G}(f) = \mathcal{G}(f - f(0)) \). This is the reason we can assume \( f_n(0) \in [-C, C] \). Because \( \|f_n\|_{\text{Lip}} \leq k \), we can assert that for any \( x \in \mathbb{R} \), \( \{f_n(x) \mid n \in \mathbb{R}\} \) is bounded. So we can imitate the method used in Lemma 3 and find the optimal function \( f^* \) such that \( \mathcal{G}(f^*) = \inf_{\|f\|_{\text{Lip}} \leq k} \mathcal{G}(f) \).

Lemma 6 (Theorem 1 Part I). Under the same assumption of Lemma 5, we have \( \mathcal{F}(f) = \mathbb{E}_{X \sim \mathcal{P}_x} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_y} \varphi(f(Y)) + \lambda \|f\|_{\text{Lip}}^2 \) with \( \lambda > 0 \) and \( \alpha > 1 \) has global minima.

Proof. When \( \|f\|_{\text{Lip}} = \infty \), it is trivial that \( \mathcal{F}(f) = \infty \). And when \( \|f\|_{\text{Lip}} < \infty \), combining Lemma 1, we have \( \mathcal{F}(f) = \mathcal{G}(f) + \lambda \|f\|_{\text{Lip}}^2 \geq -\|f\|_{\text{Lip}} \phi'(a_0) W_1(\mathcal{P}_R, \mathcal{P}_G) + \lambda \|f\|_{\text{Lip}}^2 \). When \( \lambda > 0 \) and \( \alpha > 1 \), the right term is a convex function about \( \|f\|_{\text{Lip}} \); it has a lower bound. So we can find a sequence \( \{f_n\}_{n=1}^\infty \) such that \( \lim_{n \to \infty} \mathcal{F}(f_n) = \inf_{f \in \mathcal{D}} \mathcal{F}(f) \). It is no doubt that there exists a constant \( C \) such that \( \|f_n\|_{\text{Lip}} \leq C \) for all \( f_n \). Then it is not difficult to show for any point \( x \), \( \{f_n(x)\} \) is bounded. So we can imitate the method used in main theorem to find the sequence \( \{g_n\} \) such that \( \{g_n\} \subseteq \{f_n\} \) and \( \{g_n\}_{n=1}^\infty \) converge at every point \( x \). Suppose \( \lim_{n \to \infty} g_n = g \), then by Fatou’s Lemma, we have \( \mathcal{G}(g) \leq \lim_{n \to \infty} \mathcal{G}(g_n) \).

Next, we prove that \( \|g\|_{\text{Lip}} \leq \lim_{n \to \infty} \|g_n\|_{\text{Lip}} \). If the claim holds, then \( \mathcal{G}(g) = \mathcal{G}(g) + \lambda \|g\|_{\text{Lip}}^2 \geq \lim_{n \to \infty} \mathcal{G}(g_n) + \lim_{n \to \infty} \lambda \|g_n\|_{\text{Lip}}^2 \geq \lim_{n \to \infty} \mathcal{G}(g_n) + \lambda \|g_n\|_{\text{Lip}}^2 = \inf_{f \in \mathcal{D}} \mathcal{F}(f) \). Thus, the global minima exists. In fact, if \( \|g\|_{\text{Lip}} > \lim_{n \to \infty} \|g_n\|_{\text{Lip}} \), then there exist \( x, y \) such that \( |g(x) - g(y)| \geq \lim_{n \to \infty} |g_n(x) - g_n(y)| + \epsilon \geq \lim_{n \to \infty} |g_n(x) - g_n(y)| + \epsilon \|x - y\| = |g(x) - g(y)| + \epsilon \|x - y\| > |g(x) - g(y)| \). The contradiction tells us that \( \|g\|_{\text{Lip}} \leq \lim_{n \to \infty} \|g_n\|_{\text{Lip}} \).

Lemma 7 (Theorem 1 Part II). Let \( \phi \) and \( \varphi \) be two convex functions, whose domains are both \( \mathbb{R} \). If \( \phi \) or \( \varphi \) is strictly convex, then the minimizer of \( \mathcal{G}(f) = \mathbb{E}_{X \sim \mathcal{P}_x} \phi(f(X)) + \mathbb{E}_{Y \sim \mathcal{P}_y} \varphi(f(Y)) + \lambda \|f\|_{\text{Lip}}^2 \) 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 \( \phi \) is strictly convex. By the strict convexity of \( \phi \), we have \( \forall x, y \in \mathbb{R}, \ \phi\left(\frac{x+y}{2}\right) < \frac{1}{2}(\phi(x) + \phi(y)) \). Assume \( f_1 \) and \( f_2 \) are two different minimizers of \( \mathcal{G}(f) \).

First, we have

\[
\left\| f_1 + f_2 \right\|_{\text{Lip}} \leq \sup_{x, y} \frac{|f_1(x) + f_2(x) - f_1(y) - f_2(y)|}{\|x - y\|} \leq \sup_{x, y} \frac{1}{2} \left( \frac{|f_1(x) - f_1(y)|}{\|x - y\|} + \frac{|f_2(x) - f_2(y)|}{\|x - y\|} \right) 
\]
And given \( \lambda > 0 \) and \( \alpha > 1 \), we further have

\[
\lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \leq \lambda \left( \frac{1}{2} \left( \|f_1\|_{Lip} + \|f_2\|_{Lip} \right) \right)^\alpha \\
\leq \frac{1}{2} \left( \|f_1\|_{Lip}^\alpha + \|f_2\|_{Lip}^\alpha \right).
\]

(20)

Let \( \hat{s}(f_1) = \hat{s}(f_2) = \inf \hat{s}(f) \). Then we have

\[
\Theta \left( \frac{f_1 + f_2}{2} \right) = E_{X \sim \mathcal{P}_x} \phi \left( \frac{f_1 + f_2}{2} \right) + E_{Y \sim \mathcal{P}_y} \varphi \left( \frac{f_1 + f_2}{2} \right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\
\leq E_{X \sim \mathcal{P}_x} \left( \phi(f_1) + \phi(f_2) \frac{2}{2} \right) + E_{Y \sim \mathcal{P}_y} \varphi \left( \frac{f_1 + f_2}{2} \right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\
\leq E_{X \sim \mathcal{P}_x} \left( \phi(f_1) + \phi(f_2) \frac{2}{2} \right) + E_{Y \sim \mathcal{P}_y} \varphi \left( \frac{f_1 + f_2}{2} \right) + \lambda \left\| \frac{f_1 + f_2}{2} \right\|_{Lip}^\alpha \\
= \frac{1}{2} \left( \Theta(f_1) + \Theta(f_2) \right) = \inf \Theta(f)
\]

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

### A.2. Proof of Theorem 2

Let \( J_D = E_{x \sim \mathcal{P}_x} [\phi(f(x))] + E_{x \sim \mathcal{P}_x} [\varphi(f(x))] \). Let \( \hat{J}_D(x) = \mathcal{P}_y(x) \phi(f(x)) + \mathcal{P}_r(x) \varphi(f(x)) \). Clearly, \( J_D = \int_{\mathbb{R}} \hat{J}_D(x) dx \).

Let \( J^*_D(k) = \min_{f \in \mathcal{F}_{Lip}} J_D = \min_{f \in \mathcal{F}_{Lip}} \mathbb{E}_{x \sim \mathcal{P}_x} [\phi(k \cdot f(x) + b)] + \mathbb{E}_{x \sim \mathcal{P}_y} [\varphi(k \cdot f(x) + b)] \).

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

**Proof.**

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

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

\( \frac{\partial J_D(x)}{\partial f(x)} = 0 \) 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 \( \frac{\partial J_D(x)}{\partial f(x)} = 0 \) 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 \( \forall x, f^*(x) = f^*(y) \).

Given \( \forall x, y, f^*(x) = f^*(y) \), if there exists some point \( x \) such that \( \frac{\partial J_D(x)}{\partial f(x)} \neq 0 \), then it is obvious that \( \frac{\partial J_D}{\partial k(f^*)} \neq 0 \).

It is contradictory to \( \frac{\partial J_D}{\partial f(x)} = 0 \). Thus we have \( \forall x, \frac{\partial J_D(x)}{\partial f(x)} = 0 \). \( \square \)

**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 \( \frac{\partial J_D(x)}{\partial f^*(x)} = 0 \), i.e., \( \mathcal{P}_r(x) \frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = 0 \).

Thus, \( \mathcal{P}_r(x) \frac{\partial \varphi(f^*(x))}{\partial f^*(x)} = 0 \). That is, \( \mathcal{P}_r(x) \frac{\partial \varphi(f^*(x))}{\partial f^*(x)} \) 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 $\frac{\partial J_{D}(x)}{\partial f^*(x)} \neq 0$. Define $k(y) = \sup_y \frac{|f(y) - f^*(x)|}{\|y - x\|}$.

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

(ii) Assume $\exists \delta > 0$ and $\forall z, w \in B(x, \epsilon)$ such that $\frac{|f^*(z) - f^*(x)|}{\|z - w\|} < k - \delta$. Consider the following condition, for all $\delta_2$ and $\epsilon_2 \in (0, \epsilon/2), \exists y \in B(x, \epsilon_2)$, such that $\|y - x\| = k - \delta_2$. Then there exists a sequence of $\{y_n\}_{n=1}^{\infty}$ s.t. $\lim_{n \to \infty} |f(y_n) - f(x)| = \epsilon \cdot \epsilon_2$. We get $\frac{|f^*(y_n) - f^*(x)|}{\|y_n - x\|} = \epsilon \cdot \epsilon_2 \geq (1 - \epsilon_2)\|x - y\|$. Let $\epsilon_2 \to 0$ and $\delta_2 \to 0$. We get $k(x) = k$, which means there exists a $y$ such that $|f^*(y) - f^*(x)| = k\|y - x\|$.

(iii) Now we can assume $\exists \delta_2 > 0$ and $\forall y \in B(x, \epsilon_2)$, such that $k(y) \leq k - \delta_2$. If $\frac{\partial J_{D}(x)}{\partial f^*(x)} \neq 0$, without loss of generality, we can assume $\frac{\partial J_{D}(x)}{\partial f^*(x)} > 0$. Then, for all $y \in B(x, \epsilon_2)$, we have $\frac{\partial J_{D}(y)}{\partial f^*(y)} > 0$, as long as $\epsilon_2$ is small enough.

Now we change the value of $f^*(y)$ for $y \in B(x, \epsilon_2)$. Let $g(y) = \left\{ \begin{array}{ll} f^*(y) & \frac{\partial J_{D}(y)}{\partial f^*(y)} < 0, \\ f^*(x) & \text{otherwise.} \end{array} \right.$ Because $\frac{\partial J_{D}(y)}{\partial f^*(y)} > 0, \forall y \in B(x, \epsilon_2)$, 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 $y, z, \in B(x, \epsilon_2)$, then $\frac{|g(y) - g(z)|}{\|y - z\|} \leq |f^*(y) - f^*(z)| = \epsilon \cdot \epsilon_2$. If $y \in B(x, \epsilon_2)$, then $\frac{|g(y) - g(z)|}{\|y - z\|} \leq |f^*(y) - f^*(z)| + \frac{N}{\epsilon_2} \leq \epsilon \cdot \epsilon_2 + \frac{N}{\epsilon_2} \leq \epsilon \cdot \epsilon_2 + \frac{N}{\epsilon_2} \leq k - \delta_2 + \frac{N}{\epsilon_2} < k$ (when $N > \frac{1}{\delta_2}$).

(b) For $x \in S_r \cup S_g - S_r \cap S_g$, assuming $\mathcal{P}_g(x) \neq 0$ and $\mathcal{P}_r(x) = 0$, we have $\frac{\partial J_{D}(x)}{\partial f^*(x)} = \mathcal{P}_g(x) \frac{\partial f^*(x)}{\partial f^*(x)} + \mathcal{P}_r(x) \frac{\partial f^*(x)}{\partial f^*(x)} = \mathcal{P}_g(x) \frac{\partial f^*(x)}{\partial f^*(x)} > 0$, because $\mathcal{P}_g(x) > 0$ and $\frac{\partial f^*(x)}{\partial f^*(x)} > 0$. Then according to (a), there must exist a $x$ 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 $f^*(y) \neq f^*(x)$. It also implies that $k(f^*) > 0$. Then according to Lemma 8, there exists a point $x$ such that $\frac{\partial J_{D}(x)}{\partial f^*(x)} \neq 0$. According to (a), there exists $y$ with $y \neq x$ satisfying that $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$. But we have $|f^*(y) - f^*(x)| = k(f^*) \cdot \|y - x\|$.

(d) In Nash equilibrium state, it holds that, for any $x \in S_r \cup S_g$, $\frac{\partial J_{D}(x)}{\partial f^*(x)} = \frac{\partial J_{D}(x)}{\partial f^*(x)} + 2\lambda \cdot k(f) = 0$ and $\frac{\partial J_{D}(x)}{\partial f^*(x)} + 2\lambda \cdot k(f) = 0$. 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 $\frac{\partial J_{D}(x)}{\partial f^*(x)} \neq 0$. And according to (a), it must hold that $\exists y$ fitting $|f(\hat{y}) - f(\hat{x})| = k(f) \cdot \|\hat{x} - \hat{y}\|$. According to Theorem 4, we have $\left\| \frac{\partial f(x)}{\partial x} \right\| = k(f) \neq 0$. This is contradictory to that $\frac{\partial J_{D}(x)}{\partial f^*(x)} = 0$. Thus $k(f) = 0$. That is, $\forall x \in S_r \cup S_g$, $\frac{\partial f(x)}{\partial x} = 0$, which means $\forall x, y, f(x) = f(y)$. According to Lemma 9, $\forall x, y, f(x) = f(y)$ implies $\mathcal{P}_r = \mathcal{P}_g$. Thus $\mathcal{P}_r = \mathcal{P}_g$ is the only Nash equilibrium in our system.

Remark 1. For the Wasserstein distance, $\nabla_{f^*(x)} J_{D}(x) = 0$ if and only if $\mathcal{P}_r(x) = \mathcal{P}_g(x)$. For the Wasserstein distance, penalizing the Lipschitz constant also benefits: at the convergence state, it will hold $\frac{\partial f(x)}{\partial x} = 0$ for all $x$. 
A.3. Proof of Theorem 3

Lemma 10. Let $k$ be the Lipschitz constant of $f$. If $f(a) - f(b) = k\|a - b\|$ and $f(b) - f(c) = k\|b - c\|$, then $f(a) - f(c) = k\|a - c\|$ and $(a, f(a)), (b, f(b)), (c, f(c))$ lies in the same line.

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

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

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

Proof. Consider $x$ with $\frac{\partial J_D(x)}{\partial f(x)} > 0$. 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 $\frac{\partial J_D(y)}{\partial f(y)} \geq 0$. Consider the set $S(x) = \{ y | f^*(y) - f^*(x) = k(f^*)\|y - x\| \}$. Note that, according to Lemma 10, any $z$ that holds $f^*(z) - f^*(y) = k(f^*)\|z - y\|$ for any $y \in S(x)$ will also be in $S(x)$. Similar to the proof of (a) in Theorem 2, we can decrease the value of $f^*(y)$ for all $y \in S(x)$ to construct a better $f$. By contradiction, we have that there must exist a $y$ with $\frac{\partial J_D(x)}{\partial f(y)} < 0$ such that $|f^*(y) - f^*(x)| = k(f^*)\|y - x\|$. Given the fact $\frac{\partial J_D(x)}{\partial f(x)} > 0$ and $\frac{\partial J_D(x)}{\partial f(y)} < 0$, we can conclude that $f^*(y) > f^*(x)$ and $f^*(y) - f^*(x) = k(f^*)\|y - x\|$. Otherwise, if $f^*(y) - f^*(x) = k(f^*)\|y - x\|$, then we can construct a better $f$ by decreasing $f^*(y)$ and increasing $f^*(x)$ which does not break the $\kappa$-Lipschitz constraint. The other case can be proved similarly.

Lemma 12. For any $x$, if $\frac{\partial J_D(x)}{\partial f(x)} > 0$, then $P_g(x) > 0$. For any $y$, if $\frac{\partial J_D(y)}{\partial f(y)} < 0$, then $P_r(y) > 0$.

Proof. $\frac{\partial J_D(x)}{\partial f(x)} = P_g(x) \frac{\partial \phi(f(x))}{\partial f(x)} + P_r(x) \frac{\partial \varphi(f(x))}{\partial f(x)}$. And we know $\phi'(x) > 0$ and $\varphi'(x) < 0$. Naturally, $\frac{\partial J_D(x)}{\partial f(x)} \geq 0$ implies $P_g(x) > 0$. Similarly, $\frac{\partial J_D(y)}{\partial f(y)} < 0$ implies $P_r(y) > 0$.

Proof of Theorem 3.

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

Remark 2. $\frac{\partial J_D(x)}{\partial f(x)} < 0$ for some $x \in S_g$ means $x$ is at the overlapping region of $S_r$ and $S_g$. It can be regarded as a $y \in S_r$, and one can apply the other rule which guarantees that there exists a $x' \in S_g$ that bounds this point.

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

In this section, we delve deep into the relationship between gradient properties and different norms in Lipschitz continuity condition. We will prove Theorem 4, i.e., Lipschitz continuity with $l_2$-norm (Euclidean Distance) can guarantee that the gradient is directly pointing towards some sample, and at the same time, demonstrate that the other norms do not have this property.

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

Lemma 13. If $f(x)$ is $k$-Lipschitz with respect to $\| \cdot \|_p$ 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

$$f(y) - f(x) = f(y) - f(x_t) + f(x_t) - f(x) \leq f(y) - f(x_t) + k\|x_t - x\|_p \leq k\|y - x_t\|_p + t \cdot k\|y - x\|_p = k \cdot (1 - t)\|y - x\|_p + t \cdot k\|y - x\|_p \leq k\|y - x\|_p.$$  \hfill (22)

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

**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.**

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

**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

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

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

Above proof utilizes the property that $\|\nabla f(x_t)\|_2 \leq k$, which is derived from that $f(x)$ is $k$-Lipschitz with respect to $\|\cdot\|_2$. However, other norms do not satisfy this property. Specifically, according to the theory in (Adler & Lunz, 2018): if a differentiable function $f$ is $k$-Lipschitz with respect to norm $\|\cdot\|_p$, then the Lipschitz continuity actually implies a bound on the dual norm of gradients, i.e., $\|\nabla f\|_q \leq k$. Here $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$, which satisfies $\frac{1}{p} + \frac{1}{q} = 1$.

As we could notice, a norm is equal to its dual norm if and only if $p = 2$. Switching to $l_p$-norm with $p \neq 2$, it is actually bounding the $l_q$-norm of the gradients. However, bounding the $l_q$-norm of the gradients does not guarantee the gradient direction at fake samples point towards real samples. A counter-example is provided as follows.

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

Note that different norm will induce different gradient with different property (Adler & Lunz, 2018). We here expect the gradient directly points towards a real sample.

**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

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

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

$$W_{KR}(\mathcal{P}_r, \mathcal{P}_y) = \sup_f \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathcal{P}_y} [f(x)],$$

s.t. $f(x) - f(y) \leq d(x, y), \forall x, \forall y$. \hfill (25)
We will prove that Wasserstein distance in its dual form can also be written as
\[
W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = \sup_f \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g} [f(x)],
\]
\[
s.t. f(x) - f(y) \leq d(x, y), \ \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g,
\]
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 \( F_{LL} = \{ f | f(x) - f(y) \leq d(x, y), \ \forall x \in \mathcal{S}_r, \forall y \in \mathcal{S}_g \} \).

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

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

\[
\forall \pi \in \Pi(\mathcal{P}_r, \mathcal{P}_g), \text{ we have the following:}
\]
\[
W_{LL}(\mathcal{P}_r, \mathcal{P}_g) = \sup_{f \in F_{LL}} \mathbb{E}_{x \sim \mathcal{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathcal{P}_g} [f(x)]
\]
\[
= \sup_{f \in F_{LL}} \mathbb{E}_{(x, y) \sim \pi} [f(x) - f(y)]
\]
\[
= \sup_{f \in F_{LL}} \mathbb{E}_{(x, y) \sim \pi} [(f(x) - f(y))I_A] + \mathbb{E}_{(x, y) \sim \pi} [(f(x) - f(y))I_{A^c}]
\]
\[
\leq \mathbb{E}_{(x, y) \sim \pi} [|y - x| I_A]
\]
\[
\leq \mathbb{E}_{(x, y) \sim \pi} [d(x, y)].
\]

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

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

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

**B. Systematical Study on Gradient Issues of Traditional GANs: Gradient Uninformativeness+**

Recall the gradient that the generator receives from the discriminator with respect to a generated sample \( x \in \mathcal{S}_g \) is
\[
\nabla_x J_G(x) \triangleq \nabla_x \psi(f(x)) = \nabla_{f(x)} \psi(f(x)) \cdot \nabla_x f(x), \tag{27}
\]
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 \( x \).

In the following, we will show that \( \nabla_{f(x)} \psi(f(x)) \) may lead to Type-I Gradient Vanishing, and \( \nabla_x f(x) \) is involved with both Type-II Gradient Vanishing and Unwarranted Gradient Direction (the general problem of gradient direction, including gradient uninformativeness).

**B.1. Type-I Gradient Vanishing**

The well-known gradient vanishing problem (Goodfellow et al., 2014; Arjovsky & Bottou, 2017) mainly refers to the problem that in the vanilla GAN, when the discriminator perfectly distinguishes the fake samples, the generator receives a vanishing gradient, i.e., the gradient scale approaches zero, which blocks its learning process.
Gradient vanishing can be explained in the divergence level (Arjovsky et al., 2017), i.e., when \( S_r \) and \( S_g \) are totally disjoint, the JS divergence which the discriminator in the vanilla GAN estimates, is locally constant with respect to \( P_g \).

It should be noted that, in terms of Eq. (27), the gradient vanishing problem in the vanilla GAN mainly stems from the vanishing scale in the scalar term \( \nabla f(x) \psi(f(x)) \), which we refer to as the Type-I Gradient Vanishing. We will show in the next section that the vector term \( \nabla_x f(x) \) may also be zero, which leads to another type of gradient vanishing that we call the Type-II Gradient Vanishing. Interestingly, the Least-Squares GAN (Mao et al., 2016) which adopts the quadratic loss as the objective function, avoids the Type-I Gradient Vanishing but might suffers from the Type-II Gradient Vanishing.

The occurrence of Type-I Gradient Vanishing, i.e., \( \nabla f(x) \psi(f(x)) = 0 \), has two necessary conditions: (i) the existence of an extreme point, i.e., \( s = \{ x | \nabla_x \psi(x) = 0 \} \neq \emptyset \); (ii) the accessibility of extreme point, i.e., \( \{ x \in S_g | f(x) \in s \} \neq \emptyset \). In the vanilla GAN, by switching to an alternative generator objective function \( \psi(x) = -\log \sigma(x) \), it avoids the accessibility of extreme point and hence solves the Type-I gradient vanishing problem. Wasserstein GAN (Arjovsky et al., 2017), with \( \psi(x) = x \), avoids the existence of extreme point and thus avoids the Type-I Gradient Vanishing. The Least-Squares GAN (Mao et al., 2016), with \( \phi(x) = (x - \alpha)^2, \psi(x) = (x - \gamma)^2 \) and \( \alpha \neq \gamma \), avoids the Type-I Gradient Vanishing via avoiding the accessibility of extreme point.

B.2. Type-II Gradient Vanishing and Unwarranted Gradient Direction

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

Generally, if a sample \( x \) is the local optimum of \( f^* \), then it suffers a substantive zero-gradient, which we refer to as the Type-II Gradient Vanishing. We broadly name gradients that do not guarantee convergence as Unwarranted Gradient. To highlight the importance of gradient direction, we refer to this problem as Unwarranted Gradient Direction problem, which includes (i) uninformative gradient; (ii) theoretically undefined gradient; (iii) non-convergent Type-II Gradient Vanishing; (iv) local-greedy gradient, etc. As an important subcase of Unwarranted Gradient Direction problem, we also separately name the problem caused by uninformative gradient as the gradient un informativeness problem.

- Uninformative gradient: if the optimal discriminative function only reflects the local densities, when a generated sample is not surrounded by real samples, its gradient tells nothing about \( P_r \). Typical situation is that \( P_r \) and \( P_g \) are disjoint, which is common in practice (Arjovsky & Bottou, 2017). This issue is independent of gradient vanishing: consider the
ideal case where $\mathcal{P}_r$ and $\mathcal{P}_g$ are totally overlapped and both consist of $n$ discrete points, but their probability mass over these points are different; it suffers from the gradient uninformativeness, but not the gradient vanishing.

- Theoretically undefined gradient: for generated sample $x$, if the optimal discriminative function is not fully defined in the surrounding of $x$, e.g., it is isolated and at the boundary. It suffers from a theoretically undefined gradient.

- Non-convergent Type-II Gradient Vanishing: for generated samples that theoretically has a Type-II Gradient Vanishing, despite the practical existence of Type-II Gradient Vanishing (e.g., Figure 7a), it more commonly suffers the non-convergent version of Type-II Gradient Vanishing, which behaves as noisy gradient (e.g., fake samples in the central of the left region in Figure 7b).

- Local-greedy gradient: when the optimal discriminative function only reflects the local densities, even if the gradient is well-defined and non-zero, the gradient update based on $\nabla_x f(x)$ is local greedy, which turns out to be an intrinsic cause of mode collapse. See Section B.4 for the details.

Samples that suffer from the uninformative gradient might at the same time suffer from theoretically undefined gradient: the uninformative gradient happens when the generated sample is not surrounded by reals samples; if it is further not fully surrounded by any kind of (real or fake) samples, it also suffers the theoretically undefined gradient. Meanwhile, theoretically undefined gradient is not a sub-problem uninformative gradient: for samples at the boundary, it suffers theoretically undefined gradient, but its gradient might be uninformative if some real samples are in the surrounding.

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

### B.3. Sample-Based Distribution Estimation

In unrestricted GANs, if $S_r \cup S_g$ does not cover the whole input space, $f^*(x)$ would be undefined outside $S_r \cup S_g$. As a result, the gradient for samples which are isolated or at the boundary can be problematic, regardless of whether $\mathcal{P}_r$ and $\mathcal{P}_g$ are overlapped or not. This also leads to a more serious problem: it prevents samples in one region from adapting to other regions, and consequently prevents $\mathcal{P}_g$ from converging to $\mathcal{P}_r$.

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

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

This is why cases with totally-overlapped distributions also suffer from the Unwarranted Gradient Direction.

#### B.3.1. Sample-Based Estimation Issues+: The Target Point of Generator in Traditional GANs

Sample-based analysis also explains a weird phenomenon about the optimal target point of generator in traditional GANs. Taking the Least-Squares GAN as an example: note that the generator objective of the Least-Squares GAN is $(x - \gamma)^2$ and the $\gamma$ derived from the Pearson $X^2$ divergence is $\gamma = \frac{\alpha + \beta}{2}$ (e.g. $\alpha = -1$, $\beta = 1$, $\gamma = 0$), but, in practice, $\gamma = \beta$ usually works better than $\gamma = \frac{\alpha + \beta}{2}$.

When $\mathcal{P}_g(x) = \mathcal{P}_r(x)$, we have $f^*(x) = \frac{\alpha + \beta}{2}$. And $\alpha \leq f^*(x) \leq \frac{\alpha + \beta}{2}$ means $\mathcal{P}_g(x) \geq \mathcal{P}_r(x)$. If the target $\gamma$ equals to $\frac{\alpha + \beta}{2}$, samples from points where $\mathcal{P}_g(x) \geq \mathcal{P}_r(x)$ cannot adapt to locations where $\mathcal{P}_g(x) \leq \mathcal{P}_r(x)$, i.e., where $\frac{\alpha + \beta}{2} \leq f^*(x) \leq \beta$. As a result, $\mathcal{P}_g$ would never converge to $\mathcal{P}_r$, and $\gamma$ actually needs to be the same as $\beta$ to avoid this issue. However, in the mean time, $\gamma = \beta$ would lead to non-zero gradient scale in terms of $\nabla f(x) \psi(f(x))$ for each sample\(^2\), even when $\mathcal{P}_g$ converges to $\mathcal{P}_r$.

In sample-based optimization, if $f^*(x)$ is a monotonically increasing function of $\frac{\mathcal{P}_r(x)}{\mathcal{P}_g(x)}$, this issue is generally inevitable, including vanilla GAN. Generally, $\gamma$ has to equal to the maximum possible value of $f^*(x)$ to avoid the above-paragraph

\(^2\)The $\nabla f(x)$ is undefined for sample that is isolated or at the boundary.
The typical result is that when fake samples get close to a mode of \( P_r \), \( \nabla_x f^*(x) \) move them towards the mode.

mentioned problem; however, because this value is not equal to the value of \( f^*(x) \) when \( P_r(x) = P_g(x) \), it results into non-zero gradient scale at convergence.

**B.4. The Cause of Mode Collapse**

Previously, we mainly discuss the problem of \( \nabla_x f^*(x) \) in the case where \( P_r \) and \( P_g \) are disjoint or discrete. In this section, we extend our discussion to the overlapping and continuous cases.

In the disjoint or discrete cases, we argue that in traditional GANs (typically the unrestricted GANs), \( f^*(x) \) on \( P_g \) does not reflect any information about the location of other points in \( P_r \), which will lead to an unfeasible \( \nabla_x f^*(x) \) and thus non-convergence. In the overlapping and continuous case, things are actually different, \( f^*(x) \) around each point is also defined, and its gradient \( \nabla_x f^*(x) \) now reflects the local variation of \( f^*(x) \).

For most traditional GANs, \( f^*(x) \) mainly reflects the local information about the density \( P_r(x) \) and \( P_g(x) \). However, it is worth noting that \( f^*(x) \) is usually an increasing function with respect to \( P_r(x) \) while a decreasing function with respect to \( P_g(x) \). For instance, \( f^*(x) \) in the vanilla GAN is \( \log P_r(x)/P_g(x) \).

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

Mode collapse is a notorious problem in GANs’ training, which refers to the phenomenon that the generator only learns to produce part of \( P_r \). Many literature try to study the source of mode collapse (Che et al., 2016; Metz et al., 2016; Kodali et al., 2017; Arora et al., 2017) and measure the degree of mode collapse (Odena et al., 2016; Arora & Zhang, 2017).

The most recognized cause of mode collapse is that, if the generator is much stronger than the discriminator, it may learn to only produce the sample(s) in the local or global maximum of \( f(x) \) for the current discriminator. This argument is true for most of GAN models. However, from our perspective on \( f^*(x) \) and its gradient, there actually exists a much more fundamental cause of mode collapse, i.e., the locality of \( f^*(x) \) in traditional GANs and the locality of gradient operator \( \nabla \).

In traditional GANs, \( f^*(x) \) is a function of local densities \( P_r(x) \) and \( P_g(x) \), which is local, and the gradient operator \( \nabla \) is also a local operator. As the result, \( \nabla_x f^*(x) \) only reflects its local variations and cannot capture the statistic of \( P_r \) and \( P_g \) that is far from itself. If \( f^*(x) \) in the surrounding area of \( x \) is well-defined, \( \nabla_x f^*(x) \) will move \( x \) towards the nearby location where the value of \( f^*(x) \) is higher. It does not take the global statistics into account.

The typical result is that when fake samples get close to a mode of \( P_r \), they move towards the mode and get stuck there (due to the locality). Assume \( P_r \) consists of two Gaussian distributions (A and B) that are distant from each other, while the current \( P_g \) is uniformly distributed over its support and close to real Gaussian A. In this case, \( \nabla_x f(x) \) of all fake samples will point towards the center of Gaussian A. If \( P_g \) is a Gaussian with the same standard deviation as Gaussian A, \( \nabla_x f(x) \) in the vanilla GAN and Least-Squares GAN shows almost identical behaviors, which is illustrated in Figure 8. In Fisher GAN, if \( \mu(x) \) is uniform, the case is even worse: a large amount of points that are relatively far from Gaussian A will move away from A (but the direction does not necessarily point towards B, though it does in our 1-D case).

In the overlapping and continuous case, though \( \nabla_x f^*(x) \) indeed carries information about \( P_r \). But \( \nabla_x f^*(x) \)-based updating turns out to be a local greedy strategy, which is still unfavorable.
B.5. The Envelope Theorem Perspective

Here we explain the gradient issues from the perspective of the envelope theorem. The envelope theorem (Milgrom & Segal, 2002) is a result about the differentiability properties of the objective function of a parameterized optimization problem.

Let the parameter of discriminator be \( \vartheta \) and the parameter of generator be \( \theta \). \( J_D(\vartheta, \theta) = \mathbb{E}_{x \sim p_x}[\phi(f(g(z; \theta); \vartheta))] + \mathbb{E}_{z \sim p_z}[\varphi(f(x; \vartheta))] \). Consider the problem

\[
J(\theta) = \arg \min_{\vartheta} J_D(\vartheta, \theta) \quad \text{s.t.} \quad s(\vartheta, \theta) \leq 0.
\]

The Lagrangian dual problem is given by

\[
L(\vartheta, \lambda, \theta) = J_D(\vartheta, \theta) + \lambda \cdot s(\vartheta, \theta),
\]

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

\[
\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial L(\vartheta, \lambda, \theta)}{\partial \theta} \bigg|_{\vartheta = \lambda^*, \lambda = \lambda^*} = \frac{\partial J_D(\vartheta^*, \theta)}{\partial \theta} + \lambda^* \frac{\partial s(\vartheta^*, \theta)}{\partial \theta}.
\]

### B.5.1. UNRESTRICTED GANs

As an illustrative sample, we first consider the following setting: let \( P_g \) be a distribution on two points \( a \) and \( 1 + a \) in \( \mathbb{R} \) with probability of \( p \) and \( 1 - p \), respectively, and the real distribution is evenly distributed on points 0 and 1. Here \( a \) and \( p \) are the learnable parameters of the generator, and \( a \) currently equals to 0 which means \( P_g \) and \( P_r \) are totally overlapped. In this setting, we allow the generator to directly change the probability distribution indicated by \( p \) and also the location of samples indicated by \( a \). Note that \( J_D(\vartheta, a, p) = p\phi(f(a; \vartheta)) + (1 - p)\phi(f(1 + a; \vartheta)) + 0.5\varphi(f(0; \vartheta)) + 0.5\varphi(f(1; \vartheta)) \).

For unrestricted GANs, we know that theoretically, \( f(x; \vartheta^*) \) is only defined on the two or four points 0 and 1, \( a \) and \( 1 - a \). In any case, \( \frac{\partial f(x; \vartheta^*)}{\partial x} \) is undefined for all points. Also if \( a \neq 0 \), we have \( |f(x; \vartheta^*)| = \infty \) for all these defined points.

Now let’s consider the gradient of \( J \), applying the envelope theorem:

\[
\frac{\partial J(a, p)}{\partial p} = \frac{\partial J_D(\vartheta^*, a, p)}{\partial p} = \phi(f(a; \vartheta^*)) - \phi(f(1 + a; \vartheta^*));
\]

\[
\frac{\partial J(a, p)}{\partial a} = \frac{\partial J_D(\vartheta^*, a, p)}{\partial a} = p \frac{\partial \phi(f(a; \vartheta^*))}{\partial f(a; \vartheta^*)} \frac{\partial f(a; \vartheta^*)}{\partial a} + (1 - p) \frac{\partial \phi(f(1 + a; \vartheta^*))}{\partial f(1 + a; \vartheta^*)} \frac{\partial f(1 + a; \vartheta^*)}{\partial (1 + a)}.
\]

Because there is no restriction, we ignore the term \( \lambda^* \frac{\partial s(\vartheta^*, a, p)}{\partial p} \).

The Eq. (31) means: if \( a = 0 \), \( p \) has a well-defined gradient; if \( a \neq 0 \), the gradient of \( p \) is exceptional and zero. However, according to Eq. (32), its gradient for \( a \) is always undefined because \( \frac{\partial f(x; \vartheta^*)}{\partial x} \) is undefined. Actually, the undefined gradient with respect to \( a \) stems from the fact that \( J(a, p) \) as a function of \( a \) is actually not continuously differentiable, i.e., envelope theorem is inapplicable for Eq. (32).

Generally, for unrestricted GANs, it is common that \( J \) is non-differentiable with respect to the location of samples, which leads to the undefined gradients. Unfortunately, in the common practice of GANs, the parameterization of generator is more like the parameter \( a \) here, which updates the distribution via changing the location of samples. For the fully overlapped case, \( J \) should also have well-defined gradient for the parameters that change the location of samples. However, the underlying objective of \( J \) is convex with respect to \( P_g \) does not imply the model of \( J \) is convex with respect to \( \theta \). We currently believe that the mode collapse is due to the non-convexity that stems from the parameterization.

### B.5.2. WGAN with Compact Dual

For WGAN with the compact dual, to make the analysis even simple, we consider the following case: let \( P_g \) be a delta distribution at \( \theta \) in \( \mathbb{R} \) with \( \theta < 1 \), while \( P_r \) is a delta distribution at 1. \( J_D(\vartheta, \theta) = f(1; \vartheta) - f(\theta; \vartheta) \) and the restriction is that \( f(1; \vartheta) - f(\theta; \vartheta) - (1 - \theta) \leq 0 \). We know that for the optimal \( f(x, \vartheta^*) \), it has \( f(\theta; \vartheta^*) = f(1; \vartheta^*) - 1 + \theta \). Due to the free offset property of WGAN, we further assume \( f(1) = 1 \) without loss of generality. Then the problem is simplified as: \( J_D(\vartheta, \theta) = 1 - f(\theta; \vartheta) \) with the restriction \( f(\theta; \vartheta) = \theta \).
The Lagrangian dual problem is given by

\[ L(\vartheta, \lambda, \theta) = 1 - f(\vartheta; \theta) + \lambda(f(\theta; \vartheta) - \theta). \] (33)

From the envelope theorem, we have

\[
\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial J_D(\vartheta^*, \theta)}{\partial \theta} = -\frac{\partial f(\theta; \vartheta^*)}{\partial \theta} + \lambda^* \frac{\partial (f(\theta; \vartheta^*) - \theta)}{\partial \theta} 
= (\lambda^* - 1) \frac{\partial f(\theta; \vartheta^*)}{\partial \theta} - \lambda^*. \] (34)

For the optimal \( \lambda^* \) and \( \vartheta^* \), we have \( \frac{\partial L}{\partial \vartheta} = (\lambda^* - 1) \frac{\partial f(\theta; \vartheta^*)}{\partial \theta} = 0 \). Let’s set \( \lambda^* = 1 \) for one of the solutions and apply it to Eq. 34. We get \( \frac{\partial J(\theta)}{\partial \theta} = 1 \), which is true. And we notice that the undefined gradient \( \frac{\partial f(\theta; \vartheta^*)}{\partial \theta} \) is eliminated by the gradient from the constraint.

In summary, for WGAN with compact dual, because the parameter of the generator is also in the constraint(s). When applying the envelope theorem, it is necessary to consider the gradient from the constraint(s). And it turns out that the undefined gradient \( \frac{\partial f(\theta; \vartheta^*)}{\partial \theta} \) will be somehow eliminated and the actual gradient comes from the remaining part of the gradient from the constraint(s).

### B.5.3. FOR LIPSCHITZ CONDITION AND LIPSCHITZ GANS

The Lipschitz condition (WGAN in KR) does not involve the parameters of generator in the constraint of the optimization problem. Lipschitz GANs enforce Lipschitz as a penalty which does not involve constraints. So as long as \( J \) is continuously differentiable, the envelope theorem is applicable and we have

\[
\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial E_{z \sim P_z}[\phi(f(g(z; \theta); \vartheta^*))]}{\partial \theta}. \] (35)

With the Lipschitz condition or penalty on Lipschitz constant, the objective is intuitively continuously differentiable with respect to \( \theta \). If the generative function is continuous and locally-Lipschitz with respect to its parameter \( \theta \), then the objective will be continuously differentiable with respect to \( \theta \). We have shown in the paper that Lipschitz constraint with respect to Euclidean distance results in excellent gradient properties in terms of \( \frac{\partial f(x; \vartheta^*)}{\partial x} \), i.e., the right side of Eq. (35) well behaves, if the generator is continuously differentiable with respect to \( \theta \).

### B.6. The Practical behaviors of Gradient Issues

To study the practical behaviors of various gradient issues and understanding how GANs that theoretically have gradient issues work in practice. 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 9: ADAM with lr=1e-2, beta1=0.0, beta2=0.9. MLP with RELU activations, #hidden units=1024, #layers=1.

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 \( f \) which might lead to a simple value surface. When the neural network is capable of learning the (approximate) optimal \( f^* \), how the actual \( f \) approaches \( f^* \) and how the theoretically undefined points behaviors...
highly depends the optimization details and the characteristics of the network. For points whose gradient are theoretically undefined, they practical behaviors highly depend on the detailed setting, which means it is hard to control and is controlled by hyper-parameters tuning.

These experiments also shed some light on how hyper-parameters influence GANs training. First, the value surface of traditional GANs is highly dependent on the network and training parameters. Second, some particular settings lead to
simple (or other favorable) value surface. According to the above experiments: (i) a low-capacity network tends to learn a simple surface; (ii) ADAM, compared with SGD, tends to learn a simpler surface; (iii) large learning rate tends to learn a simpler surface than small learning rate; (iv) piece-wise linear activation tends to result in simpler value surface, compared with highly nonlinear activation function. Hyper-parameter setting that leads to simple value surface is more likely to have successful training, which are basically consistent with the empirical success and failure of traditional GANs in common practice.

B.6.1. EXPLANATION ON THE EMPIRICAL SUCCESS OF TRADITIONAL GANs

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

As shown in Appendix B.6, hyper-parameters are important in influencing the value surface of \( f^\ast \). Some typical settings (e.g., simplified neural network architecture, relu or leaky relu activation, relatively high learning rate, Adam optimizer, etc.) tend to form a relatively simple/smooth value surface (e.g., monotonically increasing from \( P_g \) to \( P_r \)), making \( \nabla_x f^\ast(x) \) much more meaningful.

That is, one can find these settings, where \( \nabla_x f^\ast(x) \) or \( \nabla_x f(x) \) is more favourable, to enable traditional GANs to work. In opposite, we have tried highly-nonlinear activation such as swish (Ramachandran et al., 2018) in the discriminator. It turns out traditional GANs are very likely to fail. By contrast, our proposed Lipschitz GANs are compatible with highly-nonlinear activation. Another important empirical technique is to delicately balance the generator and the discriminator or limit the capacity of the discriminator. This is to avoid the fatal optimal \( f^\ast \).

Nevertheless, without theoretically guarantee on its convergence, traditional GANs are practically hard to use, sensitive to hyper-parameters and easily broken.

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 \( P_r \) and \( P_g \) is actually sufficient.

Define \( B(S_r, S_g) = \{ \hat{x} = x \cdot t + y \cdot (1-t) \mid x \in S_r \text{ and } y \in S_g \text{ and } t \in [0,1] \} \). It is clear that \( f \) is 1-Lipschitz in \( B(S_r, S_g) \) implies \( f(x) - f(y) \leq d(x,y), \forall x \in S_r, \forall y \in S_g \). Thus, it is a sufficient constraint for Wasserstein distance in Eq. (4). In fact, \( f(x) \) is \( k \)-Lipschitz in \( B(P_r, P_g) \) 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 \( \min_{f \in \mathcal{F}_{1-Lip}} J_D(f) = k \cdot \min_{f \in \mathcal{F}_{1-Lip}} J_D(f) \).

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

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 \( f^\ast(x) \) in some of our
C.2. The New Method

Notice. 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:

$$J_{\text{maxgp}} = \lambda \max_{x \sim B(S_r, S_g)} \left[ \left\| \nabla_x f(x) \right\|^2 \right]$$

In practice, we sample $x$ from $B(S_r, S_g)$ 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 $\mathbb{E}_{x \sim B} [\left\| \nabla_x f(x) \right\|^2]$ and the proposed maximum gradient penalty in Figure 14. Before switching to maximum gradient penalty, we struggled for a long time and cannot achieve a high quality result as shown in Figure 14b. The other forms of gradient penalty (Gulrajani et al., 2017; Petzka et al., 2017) perform similar as $\mathbb{E}_{x \sim B} [\left\| \nabla_x f(x) \right\|^2]$.

To improve the stability and reduce the bias introduced via batch sampling, one can further keep track $x$ with the maximum $\left\| \nabla_x f(x) \right\|$. A practical and light weight method is to maintain a list $S_{\text{max}}$ that has the currently highest (top-k) $\left\| \nabla_x f(x) \right\|_2$ (initialized with random samples), use the $S_{\text{max}}$ as part of the batch that estimates $J_{\text{maxgp}}$, and update the $S_{\text{max}}$ 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. Non-differentiable Optimal Discriminative Function

If $f^*$ is $k$-Lipschitz and $f^*(y) - f^*(x) = k \cdot d(x, y)$, we say that $(x, y)$ are coupled. When a sample $x$ is coupled with more than one $y$ and these $y$ lie in different directions of $x$, $f^*$ is non-differentiable at $x$ and it will has sub-gradient along each direction. When $f^*$ is non-differentiable, as we noticed in the experiments, it usually behaviors as that the gradient

![Figure 14: Comparison between gradient penalty and maximum gradient penalty, with $P_r$ and $P_g$ consist of ten real and noise images, respectively. The leftmost in each row is a $x \in S_g$ and the second is its gradient $\nabla_x f^*(x)$. The interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing $\epsilon$, which will pass through a real sample, and the rightmost is the nearest $y \in S_r$.](image)
We plot the IS training curve of LGANs in Figures 16 and 17. We provide the visual results of LGANs in Figures 18, 19.

The non-differentiable problem also traces back to the Monge problem (Villani, 2008), where the generated sample is mapping to a single target. It theoretically discusses under which condition each sample has a single target, which by nature solves the non-differentiable problem. However, for the Monge problem to be solvable, it requires \( d(x, y) \) to be a strictly convex and super-linear (Villani, 2008). Unfortunately, the Euclidean distance, which is necessary to ensure the gradient direction from fake sample directly points toward real sample, does not fit this condition. So we currently does not figure out a practical solution to take advantage of Monge problem related theories.

Nonetheless, even if \( f^* \) is non-differentiable, the gradient is also usually somehow pointing towards the real samples. And the empirical founding is that when the \( P_g \) get close to \( P_r \), the non-differentiable problem diminishes.

### D.2. Connection between Lipschitz GAN and Optimal Transport

The first order Wasserstein distance, also named as the Earth Mover’s distance, is a special form of optimal transport (Villani, 2008), which measures the minimal cost of moving the source distribution to the target distribution, and the optimal coupling \( \pi(x, y) \) describes the transport plan, i.e., how much density we should move from \( x \) to \( y \).

 Naturally, updating the generator according to the optimal coupling would pull \( P_g \) towards \( P_r \). However, updating the generator according to the optimal coupling \( \pi(x, y) \) is a totally different mechanism for training GANs. In typical GANs, we update the generator following \( \nabla_x f^*(x) \). An interesting fact is that: assuming \( f^*(x) \) is differentiable, with Wasserstein distance objective, when updating the generator according to \( \nabla_x f^*(x) \), it follows the optimal coupling \( \pi \), if we use the KR dual of Wasserstein distance and \( d(x, y) \) represents the Euclidean distance (Proposition 1).

In general optimal transport, \( d(x, y) \) is not required to be a distance and can be any cost function. To the best knowledge of the authors, it is hard to access the coupling information from \( \nabla_x f^*(x) \) if \( d(x, y) \) is arbitrary. However, given the optimal coupling \( \pi \), directly updating each sample towards its target is also practicable. An instance of this line of work is (Sanjabi et al., 2018), where the objective (Seguy et al., 2017) of generator is \( \mathbb{E}_{x \sim P_g} [\mathbb{E}_{y \sim \pi(x)} [d(x, y)]] \).

In summary, we think training GANs with optimal mapping and with Lipschitz continuity are two mechanisms with different underlying principles that have some connections.

### D.3. Various \( \phi \) and \( \varphi \) That Satisfies Eq. (11)

For Lipschitz GANs, \( \phi \) and \( \varphi \) 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 \( \phi \) and \( \varphi \) in Figure 15.

To devise a loss satisfies Eq. (11), it is practical to let \( \phi \) 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 \( \phi \) and \( \varphi \) within a function class, and linear combination of two or more \( \phi \) or \( \varphi \) from different function classes also keep satisfying Eq. (11).

### D.4. Experiment Details

In our experiments with real data (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 \( \lambda \) 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 \( \phi \) and \( \varphi \) and the dataset, and all other components are fixed.

We plot the IS training curve of LGANs in Figure 16 and 17. We provide the visual results of LGANs in Figure 18, Figure 19.
for CIFAR-10 and Tiny Imagenet, respectively. As an extra experiment, we also provide the visual results of LGANs on Oxford 102 in Figure 20.

![Figure 15: Various $\phi$ and $\psi$ that satisfies Eq. (11).](image)

![Figure 16: IS training curves on CIFAR-10.](image)

![Figure 17: IS training curves on Tiny ImageNet.](image)

| Generator: | Operation | Kernel | Resample | Output Dims |
|------------|-----------|--------|----------|-------------|
| Noise      | N/A       | N/A    |          | 128         |
| Linear     | N/A       | N/A    |          | 128×4×4     |
| Residual block | 3×3 | UP     | 128×8×8  |
| Residual block | 3×3 | UP     | 128×16×16 |
| Residual block | 3×3 | UP     | 128×32×32 |
| Conv & Tanh | 3×3 | N/A    | 3×32×32  |

| Discriminator: | Operation | Kernel | Resample | Output Dims |
|----------------|-----------|--------|----------|-------------|
| Residual Block | 3×3×2     | Down   | 128×16×16 |
| Residual Block | 3×3×2     | Down   | 128×8×8   |
| Residual Block | 3×3×2     | N/A    | 128×8×8   |
| Residual Block | 3×3×2     | N/A    | 128×8×8   |
| ReLU, mean pool | N/A     | N/A    | 128      |
| Linear         | N/A       | N/A    | 1         |

Table 3: The network architectures.
Figure 18: Random samples of LGANs with different loss metrics on CIFAR-10.
Figure 19: Random samples of LGANs with different loss metrics on Tiny Imagenet.
Figure 20: Random samples of LGANs with different loss metrics on Oxford 102.
Figure 21: 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 $x \in S_g$ and the nearby column are their gradient $\nabla_x f^*(x)$. Down: the leftmost in each row is $x \in S_g$, the second are their gradients $\nabla_x f^*(x)$, the interiors are $x + \epsilon \cdot \nabla_x f^*(x)$ with increasing $\epsilon$, and the rightmost is the nearest $y \in S_r$. 