A Convex Duality Framework for GANs

Farzan Farnia[*]
farnia@stanford.edu

David Tse[*]
dntse@stanford.edu

Abstract

Generative adversarial network (GAN) is a minimax game between a generator mimicking the true model and a discriminator distinguishing the samples produced by the generator from the real training samples. Given an unconstrained discriminator able to approximate any function, this game reduces to finding the generative model minimizing a divergence measure, e.g. the Jensen-Shannon (JS) divergence, to the data distribution. However, in practice the discriminator is constrained to be in a smaller class $\mathcal{F}$ such as neural nets. Then, a natural question is how the divergence minimization interpretation changes as we constrain $\mathcal{F}$. In this work, we address this question by developing a convex duality framework for analyzing GANs. For a convex set $\mathcal{F}$, this duality framework interprets the original GAN formulation as finding the generative model with minimum JS-divergence to the distributions penalized to match the moments of the data distribution, with the moments specified by the discriminators in $\mathcal{F}$. We show that this interpretation more generally holds for f-GAN and Wasserstein GAN. As a byproduct, we apply the duality framework to a hybrid of f-divergence and Wasserstein distance. Unlike the f-divergence, we prove that the proposed hybrid divergence changes continuously with the generative model, which suggests regularizing the discriminator's Lipschitz constant in f-GAN and vanilla GAN. We numerically evaluate the power of the suggested regularization schemes for improving GAN's training performance.

1 Introduction

Learning a probability model from data samples is a fundamental task in unsupervised learning. The recently developed generative adversarial network (GAN) [1] leverages the power of deep neural networks to successfully address this task across various domains [2]. In contrast to traditional methods of parameter fitting like maximum likelihood estimation, the GAN approach views the problem as a *game* between a *generator* G whose goal is to generate fake samples that are close to the real data training samples and a *discriminator* D whose goal is to distinguish between the real and fake samples. The generator creates the fake samples by mapping from random noise input.

The following minimax problem is the original GAN problem, also called *vanilla GAN*, introduced in [1]

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}\big[\log D(\mathbf{X})\big] + \mathbb{E}\big[\log\big(1 - D(G(\mathbf{Z}))\big)\big]. \tag{1}$$

Here $\mathbf{Z}$ denotes the generator's noise input, $\mathbf{X}$ represents the random vector for the real data distributed as $P_{\mathbf{X}}$, and $\mathcal{G}$ and $\mathcal{F}$ respectively represent the generator and discriminator function sets. Implementing this minimax game using deep neural network classes $\mathcal{G}$ and $\mathcal{F}$ has lead to the state-of-the-art generative model for many different tasks.

To shed light on the probabilistic meaning of vanilla GAN, [1] shows that given an unconstrained discriminator D, i.e. if $\mathcal{F}$ contains all possible functions, the minimax problem (1) will reduce to

$$\min_{G\in\mathcal{G}} \mathrm{JSD}(P_{\mathbf{X}}, P_{G(\mathbf{Z})}), \tag{2}$$

[*]Department of Electrical Engineering, Stanford University.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

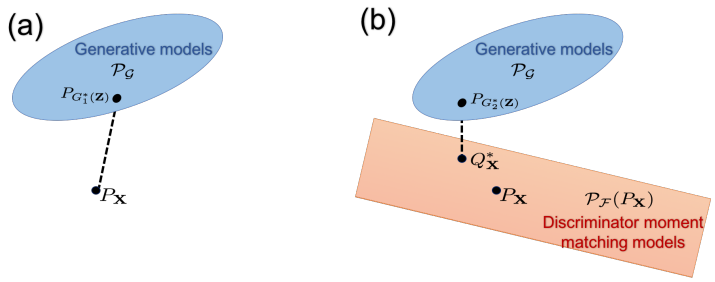

Figure 1: (a) Divergence minimization in (2) between $P_\mathbf{X}$ and generative models $\mathcal{P}_\mathcal{G}$ for unconstrained $\mathcal{F}$, (b) Divergence minimization in (3) between generative models $\mathcal{P}_\mathcal{G}$ and discriminator moment matching models $\mathcal{P}_\mathcal{F}(P_\mathbf{X})$.

where JSD denotes the Jensen-Shannon (JS) divergence. The optimization problem (2) can be interpreted as finding the closest generative model to the data distribution $P_\mathbf{X}$ (Figure 1a), where distance is measured using the JS-divergence. Various GAN formulations were later proposed by changing the divergence measure in (2): f-GAN [3] generalizes vanilla GAN by minimizing a general f-divergence; Wasserstein GAN (WGAN) [4] considers the first-order Wasserstein (the earth-mover's) distance; MMD-GAN [5, 6, 7] considers the maximum mean discrepancy; Energy-based GAN [8] minimizes the total variation distance as discussed in [4]; Quadratic GAN [9] finds the distribution minimizing the second-order Wasserstein distance.

However, GANs trained in practice differ from this minimum divergence formulation, since their discriminator is not optimized over an unconstrained set and is constrained to smaller classes such as neural nets. As shown in [10], constraining the discriminator is in fact necessary to guarantee good generalization properties for GAN's learned model. Then, how does the minimum divergence interpretation (2) change as we constrain $\mathcal{F}$? A standard approach used in [10, 11] is to view the maximum discriminator objective as an $\mathcal{F}$-based distance between distributions. For unconstrained $\mathcal{F}$, the $\mathcal{F}$-based distance reduces to the original divergence measure, e.g. the JS-divergence in vanilla GAN.

While $\mathcal{F}$-based distances have been shown to be useful for analyzing GAN's generalization properties [10], their connection to the original divergence measure remains unclear for a constrained $\mathcal{F}$. Then, what is the interpretation of GAN minimax game with a constrained discriminator? In this work, we address this question by interpreting the dual problem to the discriminator optimization. To analyze the dual problem, we develop a convex duality framework for general divergence minimization problems. We apply the duality framework to the f-divergence and optimal transport cost families, providing interpretation for f-GAN, including vanilla GAN minimizing JS-divergence, and Wasserstein GAN.

Specifically, we generalize the interpretation for unconstrained $\mathcal{F}$ in (2) to any linear space discriminator set $\mathcal{F}$. For this class of discriminator sets, we interpret vanilla GAN as the following JS-divergence minimization between two sets of probability distributions, the set of generative models and the set of discriminator moment-matching distributions (Figure 1b),

$$\min_{G\in\mathcal{G}} \min_{Q\in\mathcal{P}_\mathcal{F}(P_\mathbf{X})} \mathrm{JSD}(P_{G(\mathbf{Z})}, Q). \tag{3}$$

Here $\mathcal{P}_\mathcal{F}(P_\mathbf{X})$ contains any distribution Q satisfying the moment matching constraint $\mathbb{E}_Q[D(\mathbf{X})] = \mathbb{E}_P[D(\mathbf{X})]$ for all discriminator D's in $\mathcal{F}$. More generally, we show that a similar interpretation applies to GANs trained over any convex discriminator set $\mathcal{F}$. We further discuss the application of our duality framework to neural net discriminators with bounded Lipschitz constant. While a set of neural network functions is not necessarily convex, we prove any convex combination of Lipschitz-bounded neural nets can be approximated by uniformly combining boundedly-many neural nets. This result applied to our duality framework suggests considering a uniform mixture of multiple neural nets as the discriminator.

As a byproduct, we apply the duality framework to the minimum sum hybrid of f-divergence and the first-order Wasserstein (W_1) distance, e.g. the following hybrid of JS-divergence and W_1 distance:

$$d_{\mathrm{JSD},W_1}(P_1, P_2) := \min_Q \; W_1(P_1, Q) + \mathrm{JSD}(Q, P_2). \tag{4}$$

We prove that this hybrid divergence enjoys a continuous behavior in distribution P_1. Therefore, the hybrid divergence provides a remedy for the discontinuous behavior of the JS-divergence when optimizing the generator parameters in vanilla GAN. [4] observes this issue with the JS-divergence in vanilla GAN and proposes to instead minimize the continuously-changing W_1 distance in WGAN. However, as empirically demonstrated in [12] vanilla GAN with Lipschitz-bounded discriminator remains the state-of-the-art method for training deep generative models in several benchmark tasks. Here, we leverage our duality framework to prove that the hybrid d_{JSD,W_1}, which possesses the same continuity property as in W_1 distance, is in fact the divergence measure minimized in vanilla GAN with 1-Lipschitz discriminator. Our analysis hence provides an explanation for why regularizing the discriminator's Lipschitz constant via gradient penalty [13] or spectral normalization [12] improves the training performance in vanilla GAN. We then extend our focus to the hybrid of f-divergence and the second-order Wasserstein (W_2) distance. In this case, we derive the f-GAN (e.g. vanilla GAN) problem with its discriminator being adversarially trained using Wasserstein risk minimization [14]. We numerically evaluate the power of these families of hybrid divergences in training vanilla GAN.

2 Divergence Measures

2.1 Jensen-Shannon divergence

The Jensen-Shannon divergence is defined in terms of the KL-divergence (denoted by KL) as

$$\mathrm{JSD}(P,Q) := \frac{1}{2}\,\mathrm{KL}(P\|M) + \frac{1}{2}\,\mathrm{KL}(Q\|M)$$

where $M = \frac{P+Q}{2}$ is the mid-distribution between P and Q. Unlike the KL-divergence, the JS-divergence is symmetric $\mathrm{JSD}(P,Q) = \mathrm{JSD}(Q,P)$ and bounded $0 \le \mathrm{JSD}(P,Q) \le 1$.

2.2 f-divergence

The f-divergence family [15] generalizes the KL and JS divergence measures. Given a convex lower semicontinuous function f with $f(1) = 0$, the f-divergence d_f is defined as

$$d_f(P,Q) := \mathbb{E}_P\big[f\big(\frac{q(\mathbf{X})}{p(\mathbf{X})}\big)\big] = \int p(\mathbf{x}) f\big(\frac{q(\mathbf{x})}{p(\mathbf{x})}\big)\,\mathrm{d}\mathbf{x}. \tag{5}$$

Here $\mathbb{E}_P$ denotes expectation over distribution P and $p,\ q$ denote the density functions for distributions $P,\ Q$, respectively. The KL-divergence and the JS-divergence are members of the f-divergence family, corresponding to respectively $f_{\mathrm{KL}}(t) = t\log t$ and $f_{\mathrm{JSD}}(t) = \frac{t}{2}\log t - \frac{t+1}{2}\log\frac{t+1}{2}$.

2.3 Optimal transport cost, Wasserstein distance

The optimal transport cost for cost function $c(\mathbf{x},\mathbf{x}')$, which we denote by OT_c, is defined as

$$\mathrm{OT}_c(P,Q) := \inf_{M\in\Pi(P,Q)} \mathbb{E}\big[c(\mathbf{X},\mathbf{X}')\big], \tag{6}$$

where $\Pi(P,Q)$ contains all couplings with marginals P, Q. The Kantorovich duality [16] shows that for a non-negative lower semi-continuous cost c,

$$\mathrm{OT}_c(P,Q) = \max_{D\text{ c-concave}} \mathbb{E}_P\big[D(\mathbf{X})\big] - \mathbb{E}_Q\big[D^c(\mathbf{X})\big], \tag{7}$$

where we use D^c to denote D's c-transform defined as $D^c(\mathbf{x}) := \sup_{\mathbf{x}'}\ D(\mathbf{x}') - c(\mathbf{x},\mathbf{x}')$ and call D c-concave if D is the c-transform of a valid function. Considering the norm-based cost $c_q(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^q$ with $q \ge 1$, the qth order Wasserstein distance W_q is defined based on the c_q optimal transport cost as

$$W_q(P,Q) := OT_{c_q}(P,Q)^{1/q} = \inf_{M\in\Pi(P,Q)} \mathbb{E}\big[\,\|\mathbf{X}-\mathbf{X}'\|^q\,\big]^{1/q}. \tag{8}$$

An important special case is the first-order Wasserstein (W_1) distance corresponding to the difference norm cost $c_1(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|$. Given cost function c_1, a function D is c-concave if and only if D is 1-Lipschitz, and the c-transform $D^c = D$ for any 1-Lipschitz D. Therefore, the Kantorovich duality (7) implies that

$$W_1(P,Q) = \max_{D\text{ 1-Lipschitz}} \mathbb{E}_P\big[D(\mathbf{X})\big] - \mathbb{E}_Q\big[D(\mathbf{X})\big]. \tag{9}$$

Another notable special case is the second-order Wasserstein (W_2) distance, corresponding to the difference norm-squared cost $c_2(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^2$.

3 Divergence minimization in GANs: a convex duality framework

In this section, we develop a convex duality framework for analyzing divergence minimization problems conditioned to moment-matching constraints. Our framework generalizes the duality framework developed in [17] for the f-divergence family.

For a general divergence measure $d(P, Q)$, we define d's conjugate over distribution P, which we denote by d^*_P, as the following mapping from real-valued functions of $\mathbf{X}$ to real numbers

$$d^*_P(D) := \sup_Q \mathbb{E}_Q[D(\mathbf{X})] - d(P, Q). \tag{10}$$

Here the supremum is over all distributions on $\mathbf{X}$ with support set $\mathcal{X}$. We later show the following theorem, which is based on the above definition, recovers various well-known GAN formulations, when applied to divergence measures discussed in Section 2.

Theorem 1. *Suppose divergence $d(P, Q)$ is non-negative, lower semicontinuous and convex in distribution Q. Consider a convex set of continuous functions $\mathcal{F}$ and assume support set $\mathcal{X}$ is compact. Then,*

$$\begin{aligned} &\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}_{P_\mathbf{X}}[D(\mathbf{X})] - d^*_{P_{G(\mathbf{Z})}}(D) \\ &= \min_{G\in\mathcal{G}} \min_{Q} \big\{ d(P_{G(\mathbf{Z})}, Q) + \max_{D\in\mathcal{F}} \{ \mathbb{E}_{P_\mathbf{X}}[D(\mathbf{X})] - \mathbb{E}_Q[D(\mathbf{X})] \} \big\}. \end{aligned} \tag{11}$$

Proof. We defer the proof to the Appendix. □

Theorem 1 interprets (11)'s LHS minimax problem as searching for the closest generative model to the distributions penalized to share the same moments specified by $\mathcal{F}$ with $P_\mathbf{X}$. The following corollary of Theorem 1 shows if we further assume that $\mathcal{F}$ is a linear space, then the penalty term penalizing moment mismatches can be moved to the constraints. This reduction reveals a divergence minimization problem between generative models and the following set $\mathcal{P}_\mathcal{F}(P)$ which we call the set of discriminator moment matching distributions,

$$\mathcal{P}_\mathcal{F}(P) := \big\{ Q : \forall D \in \mathcal{F},\ \mathbb{E}_Q[D(\mathbf{X})] = \mathbb{E}_P[D(\mathbf{X})] \big\}. \tag{12}$$

Corollary 1. *In Theorem 1 suppose $\mathcal{F}$ is further a linear space, i.e. for any $D_1, D_2 \in \mathcal{F}$ and $\lambda \in \mathbb{R}$ we have $D_1 + \lambda D_2 \in \mathcal{F}$. Then,*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}_{P_\mathbf{X}}[D(\mathbf{X})] - d^*_{P_{G(\mathbf{Z})}}(D) = \min_{G\in\mathcal{G}} \min_{Q\in\mathcal{P}_\mathcal{F}(P_\mathbf{X})} d(P_{G(\mathbf{Z})}, Q). \tag{13}$$

In next section, we apply this duality framework to divergence measures discussed in Section 2 and show how to derive various GAN problems through the developed framework.

4 Duality framework applied to different divergence measures

4.1 f-divergence: f-GAN and vanilla GAN

Theorem 2 shows the application of Theorem 1 to f-divergences. We use f^* to denote f's convex-conjugate [18], defined as $f^*(u) := \sup_t ut - f(t)$. Note that Theorem 2 applies to any f-divergence d_f with non-decreasing convex-conjugate f^*, which holds for all f-divergence examples discussed in [3] with the only exception of Pearson χ^2-divergence.

Theorem 2. *Consider f-divergence d_f where the corresponding f has a non-decreasing convex-conjugate f^*. In addition to Theorem 1's assumptions, suppose $\mathcal{F}$ is closed to adding constant functions, i.e. $D + \lambda \in \mathcal{F}$ if $D \in \mathcal{F}$, $\lambda \in \mathbb{R}$. Then, the minimax problem in the LHS of* (11) *and* (13)*, will reduce to*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}[D(\mathbf{X})] - \mathbb{E}\big[f^*\big(D(G(\mathbf{Z}))\big)\big]. \tag{14}$$

Proof. We defer the proof to the Appendix. □

The minimax problem (14) is in fact the f-GAN problem [3]. Theorem 2 hence reveals that f-GAN searches for the generative model minimizing f-divergence to the distributions matching moments specified by $\mathcal{F}$ to the moments of true distribution.

Example 1. *Consider the JS-divergence, i.e. f-divergence corresponding to* $f_{\mathrm{JSD}}(t) = \frac{t}{2}\log t - \frac{t+1}{2}\log\frac{t+1}{2}$. *Then,* (14) *up to additive and multiplicative constants reduces to*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}[D(\mathbf{X})] + \mathbb{E}\big[\log\big(1-\exp(D(G(\mathbf{Z}))\big)\big]. \tag{15}$$

Moreover, if for function set $\tilde{\mathcal{F}}$ *the corresponding* $\mathcal{F} = \{D:\ D(\mathbf{x}) = -\log(1+\exp(\tilde{D}(\mathbf{x}))),\ \tilde{D}\in\tilde{\mathcal{F}}\}$ *is a convex set, then* (15) *will reduce to the following minimax game which is the vanilla GAN problem* (1) *with sigmoid activation applied to the discriminator output,*

$$\min_{G\in\mathcal{G}} \max_{\tilde{D}\in\tilde{\mathcal{F}}} \mathbb{E}\big[\log\frac{1}{1+\exp(\tilde{D}(\mathbf{X}))}\big] + \mathbb{E}\big[\log\frac{\exp(\tilde{D}(\mathbf{X}))}{1+\exp(\tilde{D}(\mathbf{X}))}\big]. \tag{16}$$

4.2 Optimal Transport Cost: Wasserstein GAN

Theorem 3. *Let divergence* d *be optimal transport cost* OT_c *where* c *is a non-negative lower semicontinuous cost function. Then, the minimax problem in the LHS of* (11) *and* (13) *reduces to*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}[D(\mathbf{X})] - \mathbb{E}\big[D^c(G(\mathbf{Z}))\big]. \tag{17}$$

Proof. We defer the proof to the Appendix. □

Therefore the minimax game between G and D in (17) can be viewed as minimizing the optimal transport cost between generative models and the distributions matching moments over $\mathcal{F}$ with $P_{\mathbf{X}}$'s moments. The following example applies this result to the first-order Wasserstein distance and recovers the WGAN problem [4] with a constrained 1-Lipschitz discriminator.

Example 2. *Let the optimal transport cost in* (17) *be the* W_1 *distance, and suppose* $\mathcal{F}$ *is a convex subset of 1-Lipschitz functions. Then, the minimax problem* (17) *will reduce to*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}[D(\mathbf{X})] - \mathbb{E}\big[D(G(\mathbf{Z}))\big]. \tag{18}$$

Therefore, the moment-matching interpretation also holds for WGAN: for a convex set $\mathcal{F}$ of 1-Lipschitz functions WGAN finds the generative model with minimum W_1 distance to the distributions penalized to share the same moments over $\mathcal{F}$ with the data distribution. We discuss two more examples in the Appendix: 1) for the indicator cost $c_I(\mathbf{x},\mathbf{x}') = \mathbb{I}(\mathbf{x}\neq\mathbf{x}')$ corresponding to the total variation distance we draw the connection to the energy-based GAN [8], 2) for the second-order cost $c_2(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^2$ we recover [9]'s quadratic GAN formulation under the LQG setting assumptions, i.e. linear generator, quadratic discriminator and Gaussian input data.

5 Duality framework applied to neural net discriminators

We applied the duality framework to analyze GAN problems with convex discriminator sets. However, a neural net set $\mathcal{F}_{nn} = \{f_{\mathbf{w}}:\ \mathbf{w}\in\mathcal{W}\}$, where $f_{\mathbf{w}}$ denotes a neural net function with fixed architecture and weights $\mathbf{w}$ in feasible set $\mathcal{W}$, does not generally satisfy this convexity assumption. Note that a linear combination of several neural net functions in $\mathcal{F}_{nn}$ may not remain in $\mathcal{F}_{nn}$.

Therefore, we apply the duality framework to $\mathcal{F}_{nn}$'s convex hull, which we denote by $\mathrm{conv}(\mathcal{F}_{nn})$, containing any convex combination of neural net functions in $\mathcal{F}_{nn}$. However, a convex combination of infinitely-many neural nets from $\mathcal{F}_{nn}$ is characterized by infinitely-many parameters, which makes optimizing the discriminator over $\mathrm{conv}(\mathcal{F}_{nn})$ computationally intractable. In the following theorem, we show that although a function in $\mathrm{conv}(\mathcal{F}_{nn})$ is a combination of infinitely-many neural nets, that function can be approximated by uniformly combining boundedly-many neural nets in $\mathcal{F}_{nn}$.

Theorem 4. *Suppose any function* $f_{\mathbf{w}}\in\mathcal{F}_{nn}$ *is L-Lipschitz and bounded as* $|f_{\mathbf{w}}(\mathbf{x})|\leq M$. *Also, assume that the* k*-dimensional random input* $\mathbf{X}$ *is norm-bounded as* $\|\mathbf{X}\|_2\leq R$. *Then, any function in* $\mathrm{conv}(\mathcal{F}_{nn})$ *can be uniformly approximated over the ball* $\|\mathbf{x}\|_2\leq R$ *within* ϵ*-error by a uniform combination* $\hat{f}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_{\mathbf{w}_i}(\mathbf{x})$ *of* $m = \mathcal{O}(\frac{M^2 k\log(LR/\epsilon)}{\epsilon^2})$ *functions* $(f_{\mathbf{w}_i})_{i=1}^m\in\mathcal{F}_{nn}$.

Proof. We defer the proof to the Appendix. □

The above theorem suggests using a uniform combination of multiple discriminator nets to find a better approximation of the solution to the divergence minimization problem in Theorem 1 solved over $\mathrm{conv}(\mathcal{F}_{nn})$. Note that this approach is different from MIX-GAN [10] proposed for achieving equilibrium in GAN minimiax game. While our approach considers a uniform combination of multiple neural nets as the discriminator, MIX-GAN considers a randomized combination of the minimax game over multiple neural net discriminators and generators.

6 Minimum-sum hybrid of f-divergence and Wasserstein distance: GAN with Lipschitz or adversarially-trained discriminator

Here we apply the convex duality framework to a novel class of divergence measures. For each f-divergence d_f we define divergence d_{f,W_1}, which is the minimum sum hybrid of d_f and W_1 divergences, as follows

$$d_{f,W_1}(P_1, P_2) := \inf_Q \; W_1(P_1, Q) + d_f(Q, P_2). \tag{19}$$

The above infimum is taken over all distributions on random $\mathbf{X}$, searching for distribution Q minimizing the sum of the Wasserstein distance between P_1 and Q and the f-divergence from Q to P_2. Note that the hybrid of JS-divergence and W_1-distance defined earlier in (4) is a special case of the above definition. While f-divergence in f-GAN does not change continuously with the generator parameters, the following theorem shows that similar to the continuous behavior of W_1-distance shown in [19, 4] the proposed hybrid divergence changes continuously with the generative model. We defer the proofs of this section's results to the Appendix.

Theorem 5. *Suppose $G_{\boldsymbol{\theta}} \in \mathcal{G}$ is continuously changing with parameters $\boldsymbol{\theta}$. Then, for any Q and $\mathbf{Z}$, $d_{f,W_1}(P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}, Q)$ will behave continuously as a function of $\boldsymbol{\theta}$. Moreover, if $G_{\boldsymbol{\theta}}$ is assumed to be locally Lipschitz, then $d_{f,W_1}(P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}, Q)$ will be differentiable w.r.t. $\boldsymbol{\theta}$ almost everywhere.*

Our next result reveals the minimax problem dual to minimizing this hybrid divergence with symmetric f-divergence component. We note that this symmetricity condition is met by the JS-divergence and the squared Hellinger divergence among the f-divergence examples discussed in [3].

Theorem 6. *Consider d_{f,W_1} with a symmetric f-divergence d_f, i.e. $d_f(P, Q) = d_f(Q, P)$, satisfying the assumptions in Theorem 2. If the composition $f^* \circ D$ is 1-Lipschitz for all $D \in \mathcal{F}$, the minimax problem in Theorem 1 for the hybrid d_{f,W_1} reduces to the f-GAN problem, i.e.*

$$\min_{G \in \mathcal{G}} \max_{D \in \mathcal{F}} \; \mathbb{E}[D(\mathbf{X})] - \mathbb{E}\big[\, f^*\big(D(G(\mathbf{Z}))\big)\,\big]. \tag{20}$$

The above theorem reveals that when the Lipschitz constant of discriminator D in f-GAN is properly regularized, then solving the f-GAN problem over the regularized discriminator also minimizes the continuous divergence measure d_{f,W_1}. As a special case, in the vanilla GAN problem (16) we only need to constrain discriminator $\tilde{D}$ to be 1-Lipschitz, which can be done via the gradient penalty [13] or spectral normalization of $\tilde{D}$'s weight matrices [12], and then we minimize the continuously-behaving d_{JSD,W_1}. This result is also consistent with [12]'s empirical observations that regularizing the Lipschitz constant of the discriminator improves the training performance in vanilla GAN.

Our discussion has so far focused on the mixture of f-divergence and the first order Wasserstein distance, which suggests training f-GAN over Lipschitz-bounded discriminators. As a second solution, we prove that the desired continuity property can also be achieved through the following hybrid using the second-order Wasserstein (W_2) distance-squared:

$$d_{f,W_2}(P_1, P_2) := \inf_Q \; W_2^2(P_1, Q) + d_f(Q, P_2). \tag{21}$$

Theorem 7. *Suppose $G_{\boldsymbol{\theta}} \in \mathcal{G}$ continuously changes with parameters $\boldsymbol{\theta} \in \mathbb{R}^k$. Then, for any distribution Q and random vector $\mathbf{Z}$, $d_{f,W_2}(P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}, Q)$ will be continuous in $\boldsymbol{\theta}$. Also, if we further assume $G_{\boldsymbol{\theta}}$ is bounded and locally-Lipschitz w.r.t. $\boldsymbol{\theta}$, then the hybrid divergence $d_{f,W_2}(P_{G_{\boldsymbol{\theta}}(\mathbf{Z})}, Q)$ is almost everywhere differentiable w.r.t. $\boldsymbol{\theta}$.*

The following result shows that minimizing d_{f,W_2} reduces to f-GAN problem where the discriminator is being adversarially trained.

Theorem 8. *Assume d_f and $\mathcal{F}$ satisfy the assumptions in Theorem 6. Then, the minimax problem in Theorem 1 corresponding to the hybrid d_{f,W_2} divergence reduces to*

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{F}} \mathbb{E}[D(\mathbf{X})] + \mathbb{E}\big[\min_{\mathbf{u}} -f^*\big(D\big(G(\mathbf{Z})+\mathbf{u}\big)\big) + \|\mathbf{u}\|^2\big]. \tag{22}$$

The above result reduces minimizing the hybrid d_{f,W_2} divergence to an f-GAN minimax game with a new third player. Here the third player assists the generator by perturbing the generated fake samples in order to make them harder to be distinguished from the real samples by the discriminator. The cost for perturbing a fake sample $G(\mathbf{Z})$ to $G(\mathbf{Z})+\mathbf{u}$ will be $\|\mathbf{u}\|^2$, which constrains the power of the third player who can be interpreted as an adversary to the discriminator. To implement the game between these three players, we can adversarially learn the discriminator while we are training GAN, using the Wasserstein risk minimization (WRM) adversarial learning scheme discussed in [14].

7 Numerical Experiments

To evaluate our theoretical results, we used the CelebA [20] and LSUN-bedroom [21] datasets. Furthermore, in the Appendix we include the results of our experiments over the MNIST [22] dataset. We considered vanilla GAN [1] with the minimax formulation in (16) and DCGAN [23] convolutional architecture for discriminator and generator. We used the code provided by [13] and trained DCGAN via Adam optimizer [24] for 200,000 generator iterations. We applied 5 discriminator updates for each generator update.

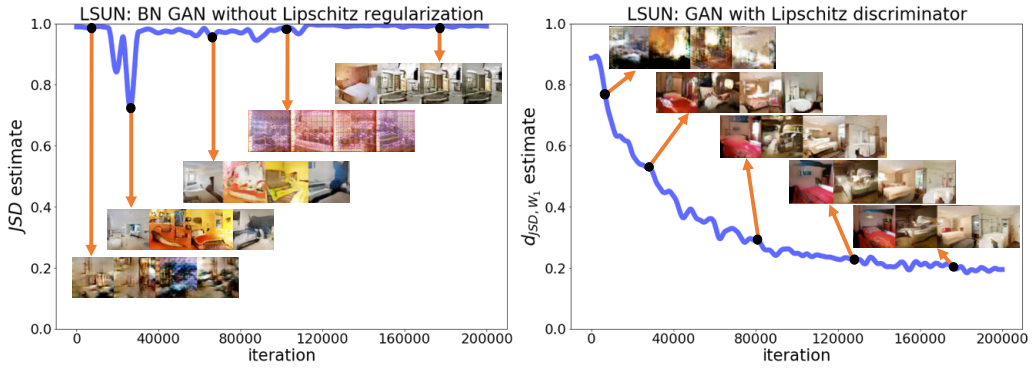

Figure 2: Divergence estimate in DCGAN trained over LSUN samples, (left) JS-divergence in standard DCGAN regularized with batch normalization, (right) hybrid d_{JSD,W_1} in DCGAN with 1-Lipschitz discriminator regularized via spectral normalization.

Figure 2 shows how the discriminator loss evaluated over 2000 validation samples, which is an estimate of the divergence measure, changes as we train the DCGAN over LSUN samples. Using standard DCGAN regularizied by only batch normalization (BN) [25], we observed (Figure 2-left) that the JS-divergence estimate always remains close to its maximum value 1 and also poorly correlates with the visual quality of generated samples. In this experiment, the GAN training failed and led to mode collapse starting at about the 110,000th iteration. On the other hand, after replacing BN with spectral normalization (SN) [12] to ensure the discriminator's Lipschitzness, the discriminator loss decreased in a desired monotonic fashion (Figure 2-right). This observation is consistent with Theorems 5 and 6 showing that the discriminator loss becomes an estimate for the hybrid d_{JSD,W_1} divergence changing continuously with the generator parameters. Also, the samples generated by the Lipschitz-regularized DCGAN looked qualitatively better and correlated well with the estimate of d_{JSD,W_1} divergence.

Figure 3 shows the results of similar experiments over the CelebA dataset. Again, we observed (Figure 3-top left) that the JS-divergence estimate remains close to 1 while training DCGAN with BN. However, after applying two different Lipschitz regularization methods, SN and the gradient penalty (GP) [13] in Figures 3-top right and bottom left, we observed that the hybrid d_{JSD,W_1} changed nicely and monotonically, and correlated properly with the sharpness of samples generated. Figure 3-bottom right shows that a similar desired behavior can also be achieved using the second-order hybrid d_{JSD,W_2} divergence. In this case, we trained the DCGAN discriminator via the WRM adversarial learning scheme [14].

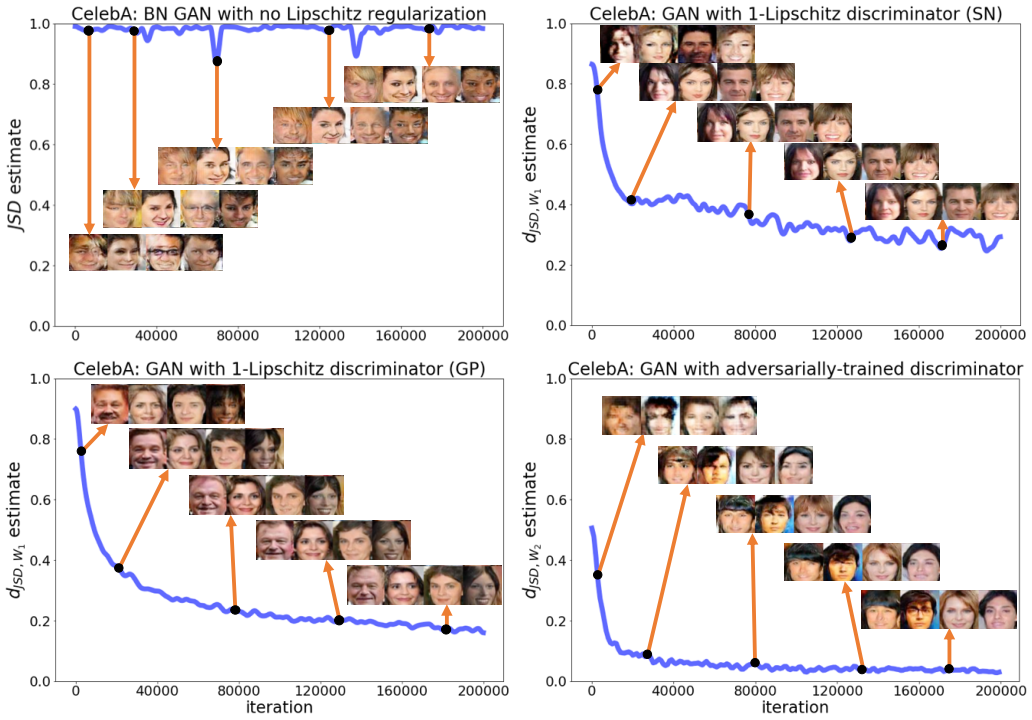

Figure 3: Divergence estimate in DCGAN trained over CelebA samples, (top-left) JS-divergence in DCGAN regularized with batch normalization, (top-right) hybrid d_{JSD,W_1} in DCGAN with spectrally-normalized discriminator, (bottom-left) hybrid d_{JSD,W_1} in DCGAN with 1-Lipschitz discriminator regularized via the gradient penalty, (bottom-right) hybrid d_{JSD,W_2} in DCGAN with discriminator being adversarially-trained using WRM.

8 Related Work

Theoretical studies of GAN have focused on three different aspects: approximation, generalization, and optimization. On the approximation properties of GAN, [11] studies GAN's approximation power using a moment-matching approach. The authors view the maximized discriminator objective as an $\mathcal{F}$-based adversarial divergence, showing that the adversarial divergence between two distributions takes its minimum value if and only if the two distributions share the same moments over $\mathcal{F}$. Our convex duality framework interprets their result and further draws the connection to the original divergence measure. [26] studies the f-GAN problem through an information geometric approach based on the Bregman divergence and its connection to f-divergence.

Analyzing GAN's generalization performance is another problem of interest in several recent works. [10] proves generalization guarantees for GANs in terms of $\mathcal{F}$-based distance measures. [27] uses an elegant approach based on the Birthday Paradox to empirically study the generalizibility of GAN's learned models. [28] develops a quantitative approach for examining diversity and generalization in GAN's learned distribution. [29] studies approximation-generalization trade-offs in GAN by analyzing the discriminative power of $\mathcal{F}$-based distances. Regarding optimization properties of GAN, [30, 31] propose duality-based methods for improving the optimization performance in training deep generative models. [32] suggests applying noise convolution with input data for boosting the training performance in f-GAN. Moreover, several other works including [33, 34, 35, 9, 36] explore the optimization and stability properties of training GANs. Finally, we note that the same convex analysis approach used in this paper has further provided a powerful theoretical framework to analyze various supervised and unsupervised learning problems [37, 38, 39, 40, 41].

Acknowledgments: We are grateful for support under a Stanford Graduate Fellowship, the National Science Foundation grant under CCF-1563098, and the Center for Science of Information (CSoI), an NSF Science and Technology Center under grant agreement CCF-0939370.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *International Conference on Machine Learning*, 2017.

[5] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[6] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

[7] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2200–2210, 2017.

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

[9] Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

[10] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

[11] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5551–5559, 2017.

[12] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 2018.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[14] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[15] Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

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

[17] Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pages 139–153, 2006.

[18] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[19] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[21] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[22] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun.com/exdb/mnist/*, 1998.

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

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

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[26] Richard Nock, Zac Cranko, Aditya K Menon, Lizhen Qu, and Robert C Williamson. f-gans in an information geometric nutshell. In *Advances in Neural Information Processing Systems*, pages 456–464, 2017.

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

[28] Shibani Santurkar, Ludwig Schmidt, and Aleksander Madry. A classification-based perspective on gan distributions. *arXiv preprint arXiv:1711.00970*, 2017.

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

[30] Xu Chen, Jiang Wang, and Hao Ge. Training generative adversarial networks via primal-dual subgradient methods: a Lagrangian perspective on GAN. In *International Conference on Learning Representations*, 2018.

[31] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, 2018.

[32] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems*, pages 2015–2025, 2017.

[33] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5591–5600, 2017.

[34] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1823–1833, 2017.

[35] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[36] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. Solving approximate wasserstein gans to stationarity. *arXiv preprint arXiv:1802.08249*, 2018.

[37] Miroslav Dudík, Steven J Phillips, and Robert E Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8(Jun):1217–1260, 2007.

[38] Meisam Razaviyayn, Farzan Farnia, and David Tse. Discrete rényi classifiers. In *Advances in Neural Information Processing Systems*, pages 3276–3284, 2015.

[39] Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, pages 4240–4248, 2016.

[40] Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian Ziebart. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems*, pages 559–567, 2016.

[41] Rizal Fathony, Mohammad Ali Bashiri, and Brian Ziebart. Adversarial surrogate losses for ordinal regression. In *Advances in Neural Information Processing Systems*, pages 563–573, 2017.