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 score, 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 convolutional neural nets. Then, a natural question is how the divergence minimization interpretation will change as we constrain \( \mathcal{F} \). In this work, we address this question by developing a convex duality framework for analyzing GAN minimax problems. For a convex set \( \mathcal{F} \), this duality framework interprets the original vanilla GAN problem as finding the generative model with the 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. We further apply the convex duality framework to explain why regularizing the discriminator’s Lipschitz constant, e.g. via spectral normalization or gradient penalty, can greatly improve the training performance in a general f-GAN problem including the vanilla GAN formulation. We prove that Lipschitz regularization can be interpreted as convolving the original divergence score with the first-order Wasserstein distance, which results in a continuously-behaving target divergence measure. We numerically explore the power of Lipschitz regularization for improving the continuity behavior and training performance in GAN problems.

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]

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

(1)

Here \( Z \) denotes the generator’s noise input, \( X \) represents the random vector for the real data distributed as \( P_X \), 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.

∗Department of Electrical Engineering, Stanford University.

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

$$\min_{G \in \mathcal{G}} \text{JSD}(P_X, P_G(Z)),$$

(2)

where $\text{JSD}$ denotes the Jensen-Shannon (JS) divergence. The optimization problem (2) can be interpreted as finding the closest generative model to the data distribution $P_X$ (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] is based on the first-order Wasserstein (the earthmover’s) distance. MMD-GAN [5, 6, 7] considers the maximum mean discrepancy. Energy-based GAN [8] uses the total variation distance. 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 convolutional neural nets. As shown in [9, 10], constraining the discriminator is in fact necessary to guarantee good generalization properties for a GAN’s learned model. Then, how does the minimum divergence interpretation illustrated in Figure 1a change after we constrain the discriminator? An existing approach used in [10, 11] is to view the maximum discriminator objective as a discriminator class $\mathcal{F}$-based distance between probability 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 [10] demonstrates a useful application of $\mathcal{F}$-based distances in analyzing GANs’ generalization properties, the connection between $\mathcal{F}$-based distances and the original divergence score remains unclear for a constrained $\mathcal{F}$. Then, what is the probabilistic interpretation of GAN minimax game in practice where a constrained discriminator is used? In this work, we address this question by interpreting the dual problem to the discriminator maximization problem. To analyze the dual problem, we develop a convex duality framework for divergence minimization problems with generalized moment matching constraints. We apply this convex duality framework to the $f$-divergence and Wasserstein distance families, providing interpretation for $f$-GAN, including vanilla GAN minimizing the JS-divergence, and Wasserstein GAN.

Specifically, we generalize [11]’s interpretation of the vanilla GAN problem (1), which only holds for an unconstrained discriminator set, to the more general case with linear space discriminator sets. Under this assumption, we interpret vanilla GAN as the following JS-divergence minimization between two sets of probability distributions (Figure 1b), the generative models and the discriminator moment-matching models,

$$\min_{G \in \mathcal{G}} \min_{Q \in \mathcal{P}_F(P_X)} \text{JSD}(P_G(Z), Q).$$

(3)

Here $\mathcal{P}_F(P_X)$ denotes the set of discriminator moment matching models that contains any distribution $Q$ satisfying moment matching constraints $E_Q[D(X)] = E_{P_X}[D(X)]$ for any discriminator $D \in \mathcal{F}$.
We prove that unlike the JS-divergence this hybrid divergence changes continuously and remedies the
where

As a byproduct, we apply the duality framework to the infimal convolution hybrid of f-divergence

The f-divergence family \([14]\) generalizes the KL and JS divergence measures. Given a convex lower

2.2 f-divergence

The f-divergence family \([14]\) generalizes the KL and JS divergence measures. Given a convex lower

2.3 Optimal transport cost, Wasserstein distance

The optimal transport cost for cost function \(c(x, x')\), which we denote by \(W_c\), is defined as

\[
W_c(P, Q) := \inf_{M \in \Pi(P, Q)} \mathbb{E}[c(X, X')], \tag{6}
\]

where \(\Pi(P, Q)\) contains all couplings with marginals \(P, Q\). The Kantorovich duality \([15]\) shows that

\[
W_c(P, Q) = \max_{D \in \mathcal{D}_{c\text{-concave}}} \mathbb{E}_P[D(X)] - \mathbb{E}_Q[D'(|X|)], \tag{7}
\]
where we use $D^c$ to denote $D$’s c-transform defined as $D^c(x) := \sup_{x'} D(x') - c(x, x')$ and call $D$ c-concave if $D$ is the c-transform of a valid function. An important special case is the first-order Wasserstein ($W_1$) distance corresponding to the norm cost $c(x, x') = \|x - x'\|$, i.e.

$$
W_1(P, Q) := \inf_{M \in \Pi(P, Q)} \mathbb{E}[\|X - X'\|].
$$

(8)

For the norm cost function, a function $D$ is c-concave if and only if $D$ is 1-Lipschitz, and the c-transform $D^c = D$ holds for a 1-Lipschitz $D$. Therefore, the Kantorovich duality result \cite{villani2003topics} implies

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

(9)

In this paper, we also consider and analyze the second-order Wasserstein ($W_2$) distance, corresponding to the norm-squared cost $c(x, x') = \|x - x'\|^2$, defined as

$$
W_2(P, Q) := \inf_{M \in \Pi(P, Q)} \mathbb{E}[\|X - X'\|^2]^{1/2}.
$$

(10)

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 \cite{gretton2012kernel} for the $f$-divergence family.

For a general divergence measure $d(P, Q)$, we define $d$’s convex conjugate for distribution $P$, which we denote by $d^*_P$, as the following operator mapping a real-valued function with domain $\mathcal{X}$ to a real number

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

(11)

Here the supremum is over all distributions on the support set $\mathcal{X}$. The following theorem connects this operation to divergence minimization problems under moment matching constraints. Next section, we discuss the application of this theorem in deriving several well-known GAN formulations for divergence measures discussed in Section \cite{goodfellow2014generative}.

**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,

$$
\min_{G \in \mathcal{G}} \max_{D \in \mathcal{F}} \mathbb{E}_{P_G}[D(X)] - d^*_P(D) = \min_{G \in \mathcal{G}} \min_Q \{ d(P_G(z), Q) + \max_{D \in \mathcal{F}} \mathbb{E}_{P_G}[D(X)] - \mathbb{E}_Q[D(X)] \}.
$$

(12)

**Proof.** We defer the proof to the Appendix. \hfill \Box

Theorem \cite{villani2003topics} interprets the LHS minimax problem in (12) as finding the closest generative model to a set of distributions penalized to share the same generalized moments specified by discriminators in $\mathcal{F}$ with $P_X$. The following corollary of Theorem \cite{villani2003topics} shows if we further assume that $\mathcal{F}$ is a linear space, then the additive penalty term penalizing the worst-case moment mismatch will turn to hard constraints in the discriminator optimization problem. This result reveals a divergence minimization problem between the generative models and the following set $\mathcal{P}_\mathcal{F}(P)$ which we call the discriminator moment matching models,

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

(13)

**Corollary 1.** In Theorem \cite{villani2003topics} suppose $\mathcal{F}$ is also a linear space, i.e. for any $D_1, D_2 \in \mathcal{F}$, $\lambda \in \mathbb{R}$ we have $D_1 + D_2 \in \mathcal{F}$ and $\lambda D_1 \in \mathcal{F}$. Then,

$$
\min_{G \in \mathcal{G}} \max_{D \in \mathcal{F}} \mathbb{E}_{P_G}[D(X)] - d^*_P(D) = \min_{G \in \mathcal{G}} \min_{Q \in \mathcal{P}_\mathcal{F}(P_X)} d(P_G(z), Q).
$$

(14)

In next section, we apply this duality framework to divergence measures discussed in Section \cite{goodfellow2014generative} and show how to derive various GAN problems through this convex duality 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 an f-divergence. Here we use \( f^* \) to denote \( f \)'s convex-conjugate [17], defined as \( f^*(u) := \sup_t ut - f(t) \). Theorem 2 applies to a general f-divergence \( d_f \) as long as the convex-conjugate \( f^* \) is a non-decreasing function, a condition met by all f-divergence examples discussed in [3] with the only exception of Pearson \( \chi^2 \)-divergence.

**Theorem 2.** Consider f-divergence \( d_f \) where the corresponding \( f \) has a non-decreasing convex-conjugate \( f^* \). In addition to the assumptions in Theorem 1, suppose that \( F \) is closed to adding constants, i.e. \( D + \lambda \in F \) for any \( D \in F \), \( \lambda \in \mathbb{R} \). Then, the minimax problem in the LHS of (12) and (14), reduces to

\[
\min_{G \in G} \max_{D \in F} \mathbb{E}[D(X)] - \mathbb{E}[f^*(D(G(Z)))]. \tag{15}
\]

**Proof.** We defer the proof to the Appendix. \( \square \)

The minimax problem (15) is the f-GAN problem introduced and discussed in [3]. Therefore, Theorem 2 reveals that f-GAN searches for the generative model minimizing the f-divergence to the discriminator moment matching models specified by discriminator set \( F \). The following example shows the application of this result to the vanilla GAN introduced in the original GAN work [1].

**Example 1.** Consider the JS-divergence, i.e. f-divergence corresponding to \( f_{\text{JS}}(t) = t \log t - (t + 1) \log \frac{t + 1}{2} \). Then, (15) up to additive and multiplicative constants reduces to

\[
\min_{G \in G} \max_{D \in F} \mathbb{E}[D(X)] + \mathbb{E}[\log(1 - \exp(D(G(Z)))]. \tag{16}
\]

Moreover, if for function set \( \tilde{F} \) the corresponding \( F = \{ D : D(x) = -\log(1 + \exp(D(x))) \}, \tilde{D} \in \tilde{F} \) is a convex set, then (16) will reduce to the following minimax game which is the vanilla GAN problem (1) with sigmoid activation applied to the discriminator output,

\[
\min_{G \in G} \max_{D \in F} \mathbb{E}[\log \frac{1}{1 + \exp(D(X))}] + \mathbb{E}[\log \frac{\exp(D(X))}{1 + \exp(D(X))}]. \tag{17}
\]

4.2 Optimal Transport Cost: Wasserstein GAN

**Theorem 3.** Let divergence \( d \) be optimal transport cost \( W_c \) where \( c \) is a non-negative lower semi-continuous cost function. Then, the minimax problem in the LHS of (12) and (14) reduces to

\[
\min_{G \in G} \max_{D \in F} \mathbb{E}[D(X)] - \mathbb{E}[D^c(G(Z))]. \tag{18}
\]

**Proof.** We defer the proof to the Appendix. \( \square \)

Therefore the minimax game between \( G \) and \( D \) in (18) can be viewed as minimizing the optimal transport cost between generative models and the distributions matching moments over \( F \) with \( P_X \)'s moments. The following example applies this result to the first-order Wasserstein distance and recovers the WGAN problem [2] with a constrained 1-Lipschitz discriminator.

**Example 2.** Let the optimal transport cost in (18) be the \( W_1 \) distance, and suppose \( F \) is a convex subset of 1-Lipschitz functions. Then, the minimax problem (18) will reduce to

\[
\min_{G \in G} \max_{D \in F} \mathbb{E}[D(X)] - \mathbb{E}[D(G(Z))]. \tag{19}
\]

Therefore, the moment-matching interpretation also holds for WGAN: for a convex set \( F \) of 1-Lipschitz functions WGAN finds the generative model with minimum \( W_1 \) distance to the distributions penalized to share the same moments over \( F \) with the data distribution. We discuss two more examples in the Appendix: 1) for the indicator cost \( c_I(x, x') = I(x \neq x') \) corresponding to the total variation distance we draw the connection to the energy-based GAN [8], 2) for the second-order cost \( c_2(x, x') = \|x - x'\|^2 \) 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 \( \mathcal{F}_{nn} = \{ f_w : w \in \mathcal{W} \} \), where \( f_w \) denotes a neural net function with fixed architecture and weights \( w \) 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 \( \text{conv}(\mathcal{F}_{nn}) \), 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 \( \text{conv}(\mathcal{F}_{nn}) \) computationally intractable. In the following theorem, we show that although a function in \( \text{conv}(\mathcal{F}_{nn}) \) 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 \( f_w \in \mathcal{F}_{nn} \) is L-Lipschitz and bounded as \( |f_w(x)| \leq M \). Also, assume that the \( k \)-dimensional random input \( X \) is norm-bounded as \( \|X\|_2 \leq R \). Then, any function in \( \text{conv}(\mathcal{F}_{nn}) \) can be uniformly approximated over the ball \( \|x\|_2 \leq R \) within \( \epsilon \)-error by a uniform combination \( f(x) = \frac{1}{m} \sum_{i=1}^{m} f_{w_i}(x) \) of \( m = O\left(\frac{M^k \log(LR/\epsilon)}{\epsilon^2}\right) \) functions \( (f_{w_i})_{i=1}^{m} \in \mathcal{F}_{nn} \).

**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 \( \text{conv}(\mathcal{F}_{nn}) \). Note that this approach is different from MIX-GAN [10] proposed for achieving equilibrium in GAN minimax 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 Infimal Convolution 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 an f-divergence \( d_f \), we define the divergence score \( d_{f,W_1} \), which we call the infimal convolution hybrid of \( d_f \) and \( W_1 \) divergence measures, as follows

\[
\min_{Q \in \mathcal{P}} \max_{P \in \mathcal{P}} \mathbb{E}[d_f(P, Q)] = \inf_Q W_1(P, Q) + d_f(Q, P).
\]

The above infimum is taken over all distributions on the support set \( \mathcal{X} \), finding the distribution \( Q^* \) minimizing the sum of the Wasserstein distance between \( P_1 \) and \( Q \) and the f-divergence from \( Q \) to \( P_2 \). Earlier in the introduction, we mentioned and discussed a special case of the above definition for the hybrid between the JS-divergence and Wasserstein distance changes continuously with the generative model.

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

**Proof.** We defer the proof to the Appendix.

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 L-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.

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

**Proof.** We defer the proof to the Appendix.
Theorem 1: Corresponding to the hybrid $d_{f,W_2}$, we show that the desired continuity property can also be achieved through the following infimal
which translates into training f-GAN with a Lipschitz-bounded discriminator. As another solution, Theorem 8.
Assume
Suppose
Theorem 7.
These observations are consistent with Theorems 5 and 6 showing that the discriminator loss will
duringly with generator parameters. Also, the samples generated by the Lipschitz-regularized DCGAN
become an estimate for the infimal convolution hybrid $d_{f,W_2}$ divergence.

To evaluate our theoretical results, we used the CelebA [20] and LSUN-bedroom [21] datasets.

The above result reduces minimizing the hybrid $d_{f,W_2}$ to
\[
d_{f,W_2}(P_1, P_2) := \inf_Q W_2^2(P_1, Q) + d_f(Q, P_2). \tag{22}
\]

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

Proof. We defer the proof to the Appendix.

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 $F$ satisfy the assumptions in Theorem\[\ref{thm:continuous}]. Then, the minimax problem in
Theorem\[\ref{thm:continuous}]
corresponding to the hybrid $d_{f,W_2}$ divergence reduces to
\[
\min_{G \in G} \max_{D \in F} E[D(G(X))] + E\left[ \min_u -f^*(D(G(Z) + u)) + \|u\|^2 \right]. \tag{23}
\]

Proof. We defer the proof to the Appendix.

The above result reduces minimizing the hybrid $d_{f,W_2}$ divergence to an f-GAN minimax game
with an additional 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 $G(Z)$ to $G(Z) + u$ will be proportional to $\|u\|^2$, constraining the
power of the third player who plays adversarially against the discriminator. To implement the
game between the three players, we can adversarially learn the discriminator while we are training
GAN, via the Wasserstein risk minimization (WRM) adversarial learning scheme discussed in \[\ref{g}].

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 [11] with the minimax formulation in [17] and DCGAN [23]
convolutional architecture for the neural net 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 per generator update.

Figure [2] shows how the discriminator loss evaluated over 2000 validation samples, which is an
estimate of the divergence measure, changed as we trained the DCGAN over LSUN samples. Using
standard DCGAN regularized by only batch normalization (BN) [25], we observed (Figure [2], top left) that the JS-divergence estimate always remained close to its maximum value $\log_2 2 = 1$ and also
correlated poorly with the visual quality of the generated samples. In this experiment, the vanilla GAN
training failed and led to mode collapse starting at about the 110,000th iteration. On the other hand,
after replacing BN with two different Lipschitz regularization techniques, spectral normalization (SN) [12]
and gradient penalty (GP) [13], to ensure that the discriminator is $1$-Lipschitz, the discriminator loss decreased in a continuous monotonic fashion (Figures [2] top right and [2] bottom left).

These observations are consistent with Theorems [5] and [6], showing that the discriminator loss
will become an estimate for the infimal convolution hybrid $d_{JSD,W_2}$ divergence which is behaving continu-
ously with 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 2 bottom right shows that a similar desired behavior with nice monotonic decrease in discriminator’s loss can also be achieved through minimizing the second-order hybrid divergence \(d_{JSD,W_2}\). In this experiment, we trained the discriminator in vanilla GAN via the Wasserstein risk minimization (WRM) adversarial learning scheme [19].

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 GP in Figures 3 top right and bottom left, we observed that the hybrid \(d_{JSD,W_1}\) changed nicely and monotonically, and correlated well with the quality of samples generated. Figure 3 bottom right shows that a similar desired behavior can also be obtained after minimizing the second-order infimal convolution hybrid \(d_{JSD,W_2}\) divergence. We defer the presentation of some random samples generated by the generators trained in these experiments to the Appendix.

8 Related Work

Theoretical studies of GAN have focused on three different aspects: approximation, generalization, and optimization. Regarding the approximation properties of GANs, [11] studies GANs’ approximation power through a moment-matching approach. The authors view the maximized discriminator objective as an \(F\)-based adversarial divergence, showing that the adversarial divergence between two distributions will be at its minimum value if the two distributions have the same generalized moments specified by \(F\). Our convex duality framework provides a dual interpretation for their results and draws the connection between the adversarial divergence and the original divergence scores. [26] studies the \(f\)-GAN problem through an information geometric approach and the connection between the Bregman divergence and the \(f\)-divergence.
Analyzing the generalization performance in GANs has been another problem of interest in the machine learning literature. [10] proves generalization guarantee results for GANs in terms of the $\mathcal{F}$-based distance measures. [27] uses an elegant approach based on birthday paradox to empirically study the generalizability of a GAN’s learned models. [28] develops a quantitative approach for examining diversity and generalization for a GAN’s learned distribution. [29] studies approximation-generalization trade-offs in GANs by analyzing the discriminative power in $\mathcal{F}$-based distances.

Regarding the optimization aspects of GANs, [30, 31] propose duality-based methods for improving optimization performance in training deep generative models. [32] suggests convolving the data distribution with a Gaussian distribution for regularizing the learning problem in f-GANs. Moreover, several other works including [33, 34, 35, 9, 36] explore the optimization and stability properties of GANs. We also note that the same convex analysis approach used in this paper for studying GANs has also provided several powerful frameworks for analyzing other 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


