Realizing GANs via a Tunable Loss Function

Gowtham R. Kurri, Tyler Sypherd, and Lalitha Sankar

Arizona State University, \{gkurri,tsypherd,lsankar\}@asu.edu

Abstract—We introduce a tunable GAN, called \(\alpha\)-GAN, parameterized by \(\alpha \in (0, \infty)\), which interpolates between various \(f\)-GANs and Integral Probability Metric based GANs (under constrained discriminator set). We construct \(\alpha\)-GAN using a supervised loss function, namely, \(\alpha\)-loss, which is a tunable loss function capturing several canonical losses. We show that \(\alpha\)-GAN is intimately related to the Arimoto divergence, which was first proposed by Österriecher (1996), and later studied by Liese and Vajda (2006). We also study the convergence properties of \(\alpha\)-GAN. We posit that the holistic understanding that \(\alpha\)-GAN introduces will have practical benefits of addressing both the issues of vanishing gradients and mode collapse.

I. INTRODUCTION

In [1], Goodfellow et al. introduced generative adversarial networks (GANs), a novel technique for training generative models to produce samples from an unknown (true) distribution using a finite number of real samples. A GAN involves two learning models (both represented by deep neural networks in practice): a generator model \(G\) that takes a random seed in a low-dimensional (relative to the data) latent space to generate synthetic samples (by implicitly learning the true distribution without explicit probability models), and a discriminator model \(D\) which classifies inputs (from either the true distribution or the generator) as real or fake. The generator wants to fool the discriminator while the discriminator wants to maximize the discrimination power between the true and generated samples. The opposing goals of \(G\) and \(D\) lead to a zero-sum min-max game in which a chosen value function is minimized and maximized over the model parameters of \(G\) and \(D\), respectively.

For the value function considered in vanilla GAN[1], when \(G\) and \(D\) are given enough training time and capacity, the min-max game is shown to have a Nash equilibrium leading to the generator minimizing the Jensen-Shannon divergence (JSD) between the true and the generated distributions. Subsequently, Nowozin et al. [4] showed that the GAN framework can minimize several \(f\)-divergences, including JSD, leading to \(f\)-GANs. Arguing that vanishing gradients are due to the sensitivity of \(f\)-divergences to mismatch in distribution supports, Arjovsky et al. [5] proposed Wasserstein GAN (WGAN) using a “weaker” Euclidean distance between distributions. This has led to a broader class of GANs based on integral probability metric (IPM) distances [6]. Yet neither the vanilla GAN nor the IPM GANs perform consistently well in practice due to a variety of issues that arise during training (e.g., mode collapse, vanishing gradients, oscillatory convergence, to name a few) [7]–[11], thus providing even less clarity on how to choose the value function.

In this work, we first formalize a supervised loss function perspective of GANs and propose a tunable \(\alpha\)-GAN based on \(\alpha\)-loss, a class of tunable loss functions [12], [13] parameterized by \(\alpha \in (0, \infty)\) that captures the well-known exponential loss (\(\alpha = 1/2\)) [14], the log-loss (\(\alpha = 1\)) [15], [16], and the 0-1 loss (\(\alpha = \infty\)) [17], [18]. Ultimately, we find that \(\alpha\)-GAN reveals a holistic structure in relating several canonical GANs, thereby unifying convergence and performance analyses. Our main contributions are as follows:

- We present a unique global Nash equilibrium to the min-max optimization problem induced by the \(\alpha\)-GAN, provided \(G\) and \(D\) have sufficiently large capacity and the models can be trained sufficiently long (Theorem 1). When the discriminator is trained to optimality (where its strategy under \(\alpha\)-loss is a tilted distribution), the generator seeks to minimize the Arimoto divergence of order \(\alpha\) (which has wide applications in statistics and information theory [19], [20]) between the true and the generated distributions, thereby providing an operational interpretation to the divergence. We note that our approach differs from Nowozin et al. \(f\)-GAN approach, please see Remark 1 for clarification.

- We show that \(\alpha\)-GAN interpolates between various \(f\)-GANs including vanilla GAN (\(\alpha = 1\)), Hellinger GAN (\(\alpha = 1/2\)), Total Variation GAN (\(\alpha = \infty\)), and IPM-based GANs including WGANs (when the discriminator set is appropriately constrained) by smoothly tuning the hyperparameter \(\alpha\) (see Theorem 2 and [14]). Thus, \(\alpha\)-GAN allows a practitioner to determine how much they want to resemble vanilla GAN, for instance, since certain datasets/distributions may favor certain GANs (or even interpolation between certain GANs). Analogous to results on \(\alpha\)-loss in classification [13], [21], where the model performance saturates quickly for \(\alpha \rightarrow \infty\), we expect a similar saturation for \(\alpha\)-GAN (see Figure 1). Thus, we posit that smooth tuning from JSD to IPM that results from increasing \(\alpha\) from 1 to \(\infty\) can address issues like mode collapse, vanishing gradients, etc.

- In Theorem 3 we reconstruct the Arimoto divergence using the margin-based form of \(\alpha\)-loss [21] and the variational formulation of Nguyen et al. [17], which sheds more light on the convexity of the generator function of
the divergence first proposed by Österreicher [22], and later studied by and Liese and Vajda [19].

- Finally, we study convergence properties of α-GAN in the presence of sufficiently large number of samples and discriminator capacity. We show that Arimoto divergences for all α > 0 are equivalent in convergence (Theorem 5) generalizing such an equivalence known in the literature [5], [23] only for special cases, i.e., for α = 1 (Jensen-Shannon divergence), α = 1/2 (squared Hellinger distance), and α = ∞ (total variation distance).

The remainder of the paper is organized as follows. We review α-loss and background on GANs in Section II. We present the loss function perspective of GANs in Section III. We propose and analyze tunable α-GAN in Section IV. Also, a connection between Arimoto divergence and the margin-based form of α-loss is examined in Section V. Finally, we study convergence properties of α-GAN in Section VI.

II. α-LOSS AND GANS

We first review a tunable class of loss functions, α-loss, that includes well-studied exponential loss (α = 1/2), log-loss (α = 1), and 0-1 loss (α = ∞). Then, we present an overview of some related GANs in the literature.

Definition 1 (Sypherd et al. [21]). For a set of distributions \( \mathcal{P}(Y) \) over \( Y \), α-loss \( \ell_\alpha : Y \times \mathcal{P}(Y) \to \mathbb{R}_+ \) for \( \alpha \in (0, 1) \cup (1, \infty) \) is defined as

\[
\ell_\alpha(y, \hat{P}) \triangleq \frac{\alpha}{\alpha - 1} \left(1 - \hat{P}(y)^{\frac{1}{\alpha}}\right).
\]

By continuous extension, \( \ell_1(y, \hat{P}) \triangleq -\log \hat{P}(y) \), \( \ell_\infty(y, \hat{P}) \triangleq 1 - \hat{P}(y) \), and \( \ell_0(y, \hat{P}) \triangleq \infty \).

Note that \( \ell_{1/2}(y, \hat{P}) = \hat{P}(y)^{-1} - 1 \), which is related to the exponential loss, particularly in the margin-based form [21]. Also, α-loss is convex in the probability term \( \hat{P}(y) \). Regarding the history of [1], Arimoto first studied α-loss in finite-parameter estimation problems [24], and later Liao et al. used α-loss to model the inferential capacity of an adversary to obtain private attributes [25]. Most recently, Sypherd et al. studied α-loss in the machine learning setting [21], which is an impetus for this work.

A. Background on GANs

Let \( P_r \) be a probability distribution over \( \mathcal{X} \subset \mathbb{R}^d \), which the generator wants to learn implicitly by producing samples by playing a competitive game with a discriminator in an adversarial manner. We parameterize the generator \( G \) and the discriminator \( D \) by vectors \( \theta \in \Theta \subset \mathbb{R}^n \) and \( \omega \in \Omega \subset \mathbb{R}^m \), respectively, and write \( G_\theta \) and \( D_\omega \) (\( \theta \) and \( \omega \) are typically the weights of neural network models for the generator and the discriminator, respectively). The generator \( G_\theta \) takes as input a \( d' \) (\( \leq d \))-dimensional latent noise \( Z \sim P_Z \) and maps it to a data point in \( \mathcal{X} \) via the mapping \( z \mapsto G_\theta(z) \). For an input \( x \in \mathcal{X} \), the discriminator outputs \( D_\omega(x) \in [0, 1] \), the probability that \( x \) comes from \( P_r \) (real) as opposed to \( P_{G_\theta} \) (synthetic). The generator and the discriminator play a two-player min-max game with a value function \( V(\theta, \omega) \), resulting in a saddle-point optimization problem given by

\[
\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V(\theta, \omega).
\]

Goodfellow et al. [1] introduced a value function

\[
V_G(\theta, \omega) = \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[\log (1 - D_\omega(G_\theta(Z)))]
\]

and showed that when the discriminator class \( \{D_\omega\} \), parametrized by \( \omega \), is rich enough, (2) simplifies to finding the \( \theta \) that minimizes \( \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] \) among all \( \omega \) (Jensen-Shannon divergence), and \( \sup_{\omega \in \Omega} V_G(\theta, \omega) \) is the Jensen-Shannon divergence [26] between \( P_r \) and \( P_{G_\theta} \). This simplification is achieved, for any \( G_\theta \), by choosing the optimal discriminator

\[
D_\omega(x) = \frac{p_r(x)}{p_r(x) + p_{G_\theta}(x)},
\]

where \( p_r \) and \( p_{G_\theta} \) are the corresponding densities of the distributions \( P_r \) and \( P_{G_\theta} \), respectively, with respect to a base measure \( dx \) (e.g., Lebesgue measure).

Generalizing this, Nowozin et al. [4] derived value function

\[
V_f(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[f^*(D_\omega(X))],
\]

where \( D_\omega : \mathcal{X} \to \mathbb{R} \) and \( f \) is the Fenchel conjugate of a convex lower semicontinuous function \( f \), for any \( f \)-divergence \( D_f(P_r||P_{G_\theta}) := \int_X P_{G_\theta}(x) f \left( \frac{p_r(x)}{p_{G_\theta}(x)} \right) dx \) [27]–[29] (not just the Jensen-Shannon divergence) leveraging its variational characterization [30]. In particular, \( \sup_{\omega \in \Omega} V_f(\theta, \omega) = D_f(P_r||P_{G_\theta}) \) when there exists \( \omega^* \in \Omega \) such that \( T_{\omega^*}(x) = f^* \left( \frac{p_r(x)}{p_{G_\theta}(x)} \right) \).

Highlighting the problems with the continuity of various \( f \)-divergences (e.g., Jensen-Shannon, KL, reverse KL, total variation) over the parameter space \( \Theta \) [10], Arjovsky et al. [5] proposed Wasserstein-GAN (WGAN) using the following Earth Mover’s (also called Wasserstein-1) distance:

\[
W(P_r, P_{G_\theta}) = \inf_{\Gamma_{X_1,X_2} \in \Pi(P_r, P_{G_\theta})} \mathbb{E}_{(X_1, X_2) \sim \Gamma_{X_1,X_2}} \|X_1 - X_2\|_2,
\]

where \( \Pi(P_r, P_{G_\theta}) \) is the set of all joint distributions \( \Gamma_{X_1,X_2} \) with marginals \( P_r \) and \( P_{G_\theta} \). WGAN employs the Kantorovich-Rubinstein duality [31] using the value function

\[
V_{\text{WGAN}}(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] - \mathbb{E}_{Z \sim P_Z}[D_\omega(X)],
\]

where the functions \( D_\omega : \mathcal{X} \to \mathbb{R} \) are all 1-Lipschitz, to simplify \( \sup_{\omega \in \Omega} V_{\text{WGAN}}(\theta, \omega) \) to \( W(P_r, P_{G_\theta}) \) when the class \( \Omega \) is rich enough. Although, various GANs have been proposed in the literature, each of them exhibits its own strengths and weaknesses in terms of convergence, vanishing gradients, mode collapse, computational complexity, etc. leaving the problem of instability unsolved [32].

\[3\]This is a slight abuse of notation in that \( D_\omega \) is not a probability here.
III. LOSS FUNCTION PERSPECTIVE OF GANS

Noting that a GAN involves a classifier (i.e., discriminator), it is well known that the value function \( V_{\varphi}(\theta,\omega) \) in (3) considered by Goodfellow et al. \cite{goodfellow2014} is related to cross-entropy loss. While perhaps it has not been explicitly articulated in the literature, we first formalize this loss function perspective of GANs. In \cite{arora2017}, Arora et al. observed that the log function in (3) can be replaced by any (monotonically increasing) concave function \( \phi(x) \) (e.g., \( \phi(x) = x \) for WGANs). More generally, we show that one can write \( V(\theta,\omega) \) in terms of \( \ell(\cdot) \) any classification loss \( \ell(y,\hat{y}) \) with inputs \( y \in \{0,1\} \) (the true label) and \( \hat{y} \in [0,1] \) (soft prediction of \( y \)). For a GAN, we have \( (X|y = 1) \sim P_r, (X|y = 0) \sim P_{G_{\theta}} \), and \( \hat{y} = D_\omega(x) \). With this, we define a value function

\[
V(\theta,\omega) = \mathbb{E}_{X|y = 1}[-\ell(y, D_\omega(X))] + \mathbb{E}_{X|y = 0}[-\ell(y, D_\omega(X))]
\]

\[
= \mathbb{E}_{X \sim P_r}[-\ell(1, D_\omega(X))] + \mathbb{E}_{X \sim P_{G_{\theta}}}[\ell(0, D_\omega(X))].
\]

For cross-entropy loss, i.e., \( \ell_{CE}(y, \hat{y}) \triangleq -y \log \hat{y} - (1 - y) \log (1 - \hat{y}) \), notice that the expression in (9) is equal to \( V_{\text{VG}} \) in (3). For the value function in (9), we consider a GAN given by the min-max optimization problem:

\[
\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V(\theta,\omega).
\]

Let \( \phi(\cdot) := -\ell(1,\cdot) \) and \( \psi(\cdot) := -\ell(0,\cdot) \) in the sequel. The functions \( \phi \) and \( \psi \) are assumed to be monotonically increasing and decreasing functions, respectively, so as to retain the intuitive interpretation of the vanilla GAN (the discriminator should output high values to real samples and low values to the generated samples). These functions should also satisfy the constraint

\[
\phi(t) + \psi(t) \leq \phi(1/2) + \psi(1/2), \quad \text{for all } t \in [0,1],
\]

so that the optimal discriminator guesses uniformly at random (i.e., outputs a constant value \( 1/2 \) irrespective of the input) when \( P_r = P_{G_{\theta}} \). A loss function \( \ell(y,\hat{y}) \) is said to be symmetric \cite{arora2017} if \( \psi(t) = \phi(1-t) \), for all \( t \in [0,1] \). Notice that the GAN considered by Arora et al. \cite{arora2017} (2)] is a special case of (10), in particular, (10) recovers the form of GAN in (3) (2)] when the loss function \( \ell(y,\hat{y}) \) is symmetric. For symmetric losses, concavity of the function \( \phi \) is a sufficient condition for satisfying (11), but not a necessary condition.

IV. TUNABLE \( \alpha \)-GAN

In this section, we examine the loss function perspective of GANs by focusing on the GAN obtained by plugging in \( \alpha \)-loss. We first write \( \alpha \)-loss in (1) in the form of a binary classification loss to obtain

\[
\ell_\alpha(y, \hat{y}) := \frac{\alpha}{\alpha - 1} \left( 1 - y \hat{y} \right)^{\frac{\alpha - 1}{\alpha}} - \left( 1 - y \right) \left( 1 - \hat{y} \right)^{\frac{\alpha - 1}{\alpha}},
\]

for \( \alpha \in (0,1) \cup (1,\infty) \). Note that (12) recovers \( \ell_{CE} \) as \( \alpha \to 1 \). Now consider a tunable \( \alpha \)-GAN with a value function

\[
V_\alpha(\theta,\omega) = \mathbb{E}_{X \sim P_r}[-\ell_\alpha(1, D_\omega(X))] + \mathbb{E}_{X \sim P_{G_{\theta}}}[\ell_\alpha(0, D_\omega(X))]
\]

\[
= \frac{\alpha}{\alpha - 1} \left( \mathbb{E}_{X \sim P_r} \left[ D_\omega(X)^{\frac{\alpha - 1}{\alpha}} \right] + \mathbb{E}_{X \sim P_{G_{\theta}}} \left[ (1 - D_\omega(X))^{\frac{\alpha - 1}{\alpha}} \right] - 2 \right).
\]

We can verify that \( \lim_{\alpha \to 1} V_\alpha(\theta,\omega) = V_{\text{VG}}(\theta,\omega) \) recovering the value function of the vanilla GAN. Also, notice that

\[
\lim_{\alpha \to \infty} V_\alpha(\theta,\omega) = \mathbb{E}_{X \sim P_r} [D_\omega(x)] - \mathbb{E}_{X \sim P_{G_{\theta}}} [D_\omega(x)] - 1
\]

is the value function (modulo a constant) used in Integral Probability Metric (IPM) based GAN\cite{alber2017}, e.g., WGAN \cite{arjovsky2017}, Fisher GAN \cite{banon2018}, and Sobolev GAN \cite{liu2018}. The resulting min-max game in \( \alpha \)-GAN is given by

\[
\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V_\alpha(\theta,\omega).
\]

The following theorem provides the min-max solution, i.e., Nash equilibrium, to the two-player game in (15) for the non-parametric setting, i.e., when the discriminator set \( \Omega \) is large enough.

Theorem 1 (min-max solution). For a fixed generator \( G_0 \), the discriminator \( D_\omega(x) \) optimizing the sup in (15) is given by

\[
D_\omega(x) = \frac{p_r(x)^{\alpha}}{p_r(x)^{\alpha} + p_{G_\omega}(x)^{\alpha}}.
\]

For this \( D_\omega(x) \), (15) simplifies to minimizing a non-negative symmetric \( f_\alpha \)-divergence \( D_{f_\alpha}(\cdot\|\cdot) \) as

\[
\inf_{\theta \in \Theta} D_{f_\alpha}(P_r\|P_{G_\omega}) + \frac{\alpha}{\alpha - 1} \left( \left( 2^{\frac{1}{\alpha}} - 2 \right) \right),
\]

where

\[
f_\alpha(u) = \frac{\alpha}{\alpha - 1} \left( u \right)^{\frac{1}{\alpha}} - (1 + u) - 2^{\frac{1}{\alpha}} + 2,
\]

for \( u \geq 0 \) and

\[
D_{f_\alpha}(P\|Q) = \frac{\alpha}{\alpha - 1} \left( \int_X (p(x)^{\alpha} + q(x)^{\alpha})^{\frac{1}{\alpha}} \, dx - 2^{\frac{1}{\alpha}} \right),
\]

which is minimized iff \( P_{G_\omega} = P_r \).

Remark 1. It can be inferred from (17) that when the discriminator is trained to optimality, the generator has to minimize the \( f_\alpha \)-divergence hinting at an application of \( f \)-GAN instead. Implementing \( \alpha \)-GAN directly via value function in (5) (for \( f_\alpha \)) involves finding convex conjugate of \( f_\alpha \), which is challenging in terms of computational complexity making it inconvenient

\footnote{Note that IPMs do not restrict the function \( D_\omega \) to be a probability.}

\footnote{We note that the divergence \( D_{f_\alpha} \) has been referred to as Arimoto divergence in the literature \cite{arimoto1972, aribo2017, maruyama2018}. We refer the reader to Section IV-A for more details.}
for optimization in the training phase of GANs. In contrast, our approach of using supervised losses circumvents this tedious effort and also provides an operational interpretation of $f$-divergence via losses. A related work where an $f$-divergence (in particular, $\alpha$-divergence [33]) shows up in the context of GANs, even when the problem formulation is not via $f$-GAN, is by Cai et al. [3]. However, our work differs from [3] in that the value function we use is well motivated via supervised loss functions of binary classification and also recovers the basic GAN [1] (among others).

**Remark 2.** As $\alpha \to 0$, note that (16) implies a more cautious discriminator, i.e., if $p_{G_\alpha}(x) \geq p_r(x)$, then $D_{\alpha}(x)$ decays more slowly from $1/2$, and if $p_{G_\alpha}(x) \leq p_r(x)$, $D_{\alpha}(x)$ increases more slowly from $1/2$. Conversely, as $\alpha \to \infty$, (16) simplifies to $D_{\alpha}(x) = \mathbb{I}\{p_r(x) > p_{G_\alpha}(x)\} + \frac{1}{2}\mathbb{I}\{p_r(x) = p_{G_\alpha}(x)\}$, where the discriminator implements the Maximum Likelihood (ML) decision rule, i.e., a hard decision whenever $p_r(x) \neq p_{G_\alpha}(x)$. In other words, (16) for $\alpha \to \infty$ induces a very confident discriminator. Regarding the generator’s perspective, (17) (and Figure 1) implies that the generator seeks to minimize the discrepancy between $P_r$ and $P_{G_\alpha}$ according to the geometry induced by $D_{\alpha}$. Thus, the optimization trajectory traversed by the generator during training is strongly dependent on the practitioner’s choice of $\alpha \in (0, \infty]$. Please refer to Figure 2 for an illustration of this observation.

A detailed proof of Theorem 1 is in Appendix A. For intuition on the construction of the function in (18), see Theorem 3. Next we show that $\alpha$-GAN recovers various well known $f$-GANs.

**Theorem 2** ($f$-GANs). $\alpha$-GAN recovers vanilla GAN, Hellinger GAN (H-GAN) [4], and Total Variation GAN (TV-GAN) [4] as $\alpha \to 1$, $\alpha = \frac{1}{2}$, and $\alpha \to \infty$, respectively.

---

**A. Reconstructing Arimoto Divergence**

It is interesting to note that the divergence $D_{\alpha}(\cdot|\cdot)$ (in (19)) that naturally emerges from the analysis of $\alpha$-GAN was first proposed by Österreicher [22] in the context of statistics and was later referred to as the Arimoto divergence by Liese and Vajda [19]. It was shown to have several desirable properties with applications in statistics and information theory [39], [40]. For example:

- A geometric interpretation of the divergence $D_{\alpha}$ in the context of hypothesis testing [22].
- $D_{\alpha}(P||Q)^{\alpha}$ defines a distance metric (satisfying the triangle inequality) on the set of probability distributions [20].

When the Arimoto divergence $D_{\alpha}$ was proposed, the convexity of the generating function $f_{\alpha}$ was proved via the traditional second derivative test [22] Lemma 1. We present an alternative approach to arriving at the Arimoto divergence by utilizing the margin-based form of $\alpha$-loss (see [21]) where the convexity of $f_{\alpha}$ (and also the symmetric property of $D_{\alpha}(\cdot|\cdot)$) arises in a rather natural manner, thereby reconstructing the Arimoto divergence through a distinct conceptual perspective.

We do this by noticing that the Arimoto divergence falls into the category of a broad class of $f$-divergences that can be obtained from margin-based loss functions. Such a connection between margin-based losses in classification and the corresponding $f$-divergences was introduced by Nguyen et al. [17] Theorem 1. They observed that, for a given margin-based loss function $\ell$, there is a corresponding $f$-divergence with the convex function $f$ defined as $f(u) := -\inf_t \left( u\ell(t) + \ell(-t) \right)$. The convexity of $f$ follows simply because the infimum of affine functions is concave, and this argument does not require

---

4In the binary classification context, the margin is represented by $t := y f(x)$, where $x \in \mathcal{X}$ is the feature vector, $y \in \{-1, +1\}$ is the label, and $f : \mathcal{X} \rightarrow \mathbb{R}$ is the prediction function produced by a learning algorithm.
Theorem 3. For the function \( f_\alpha \) in \((18)\), it holds that
\[
\alpha \left( u \tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) \right) - \frac{\alpha}{\alpha-1} \left( 2^{\frac{\alpha}{\alpha-1}} - 2 \right) \quad \text{for } u \geq 0.
\]
(21)

A detailed proof is in Appendix D.

V. CONVERGENCE PROPERTIES OF \( \alpha \)-GAN

We study convergence properties of \( \alpha \)-GAN under the assumption of sufficiently large number of samples and discriminator capacity (similar to \([23]\)). Liu et al. \([23]\) addressed a fundamental question in the context of convergence analysis of GANs, in general: For a sequence of generated distributions, does convergence of a divergence between the generated distribution and a fixed real distribution (that the generator wants to minimize) to the global minimum lead to some standard notion of distributional convergence to the real distribution? They answer this question in the affirmative provided \( \mathcal{X} \) is a compact metric space. They define adversarial divergences \([23]\) Definition 1] which capture the divergences used by a number of existing GANs that include vanilla GAN \([1]\), \( f \)-GAN \([4]\), WGAN \([5]\), and MMD-GAN \([41]\). For strict adversarial divergences (a subclass of the adversarial divergences where the minimizer of the divergence is uniquely the real distribution), they showed that convergence of divergence to global minimum implies weak convergence of the generated distribution to the real distribution. They also obtain a structural result on the class of strict adversarial divergences \([23]\) Corollary 12] based on a notion of relative strength between adversarial divergences.

We first borrow the following terminology from Liu et al. \([23]\) in order to study convergence properties of \( \alpha \)-GAN: A strict adversarial divergence \( \tau_1 \) is said to be stronger than another adversarial divergence \( \tau_2 \) (or \( \tau_2 \) is said to be weaker than \( \tau_1 \)) if for any sequence of probability distributions \( (P_n) \) and target distribution \( P \), \( \tau_1 (P \| P_n) \to 0 \) implies \( \tau_2 (P \| P_n) \to 0 \). We say \( \tau_1 \) is equivalent to \( \tau_2 \) if \( \tau_1 \) is both stronger and weaker than \( \tau_2 \). We say \( \tau_1 \) is strictly stronger than \( \tau_2 \) if \( \tau_1 \) is stronger than \( \tau_2 \) but not equivalent. We say \( \tau_1 \) and \( \tau_2 \) are not comparable if \( \tau_1 \) is neither stronger nor weaker than \( \tau_2 \).

Arjovsky et al. \([5]\) proved that Jensen-Shannon divergence is equivalent to total variation distance. Later Liu et al. showed that squared Hellinger distance is also equivalent to both these divergences, meaning that all the three divergences belong to the same equivalence class (see \([23]\) Figure 1). Noticing that squared Hellinger distance, Jensen-Shannon divergence, and total variation distance correspond to Arimoto divergences \( D_{\alpha} \) for \( \alpha = 1/2 \), \( \alpha = 1 \), and \( \alpha = \infty \), respectively, it is natural to ask the question: Are Arimoto divergences for all \( \alpha > 0 \) equivalent? We answer this question in the affirmative in Theorem 5 thereby adding the Arimoto divergences for all other \( \alpha \) also to the same equivalence class. As a motivation for a proof of this, we first give an alternative and more simpler proof for the equivalence of Jensen-Shannon divergence and total variation distance \([5]\) Theorem 2(1)].

Theorem 4. The Jensen-Shannon divergence is equivalent to the total variation distance.

Proof. We first show that the total variation distance is stronger than the Jensen-Shannon divergence. Consider a sequence of probability distributions \( (P_n) \) such that \( D_{TV} (P_n || P) \to 0 \). Using the fact that the total variation distance upper bounds the Jensen-Shannon divergence \([26]\) Theorem 3], we have \( D_{JS} (P_n || P) \leq (\log 2) D_{TV} (P_n || P) \), for each \( n \in \mathbb{N} \). This implies that \( D_{JS} (P_n || P) \to 0 \) since \( D_{TV} (P_n || P) \to 0 \). This greatly simplifies the corresponding proof of \([5]\) Theorem 2(1)] which uses measure-theoretic analysis, in particular, the Radon-Nikodym theorem. The proof for the other direction, i.e., the Jensen-Shannon divergence is stronger than the total variation distance, is exactly along the same lines as that of \([5]\) Theorem 2(1)] using triangular and Pinsker’s inequalities.

Theorem 5. Arimoto divergences for all \( \alpha > 0 \) are equivalent. That is, for a sequence of probability distributions \( (P_n) \), \( D_{\alpha_1} (P_n || P) \to 0 \) if and only if \( D_{\alpha_2} (P_n || P) \to 0 \) for any \( \alpha_1 \neq \alpha_2 \).

A detailed proof is in Appendix D.

VI. CONCLUSION

We have shown that a classical information-theoretic measure (Arimoto divergence) characterizes the ideal performance of a modern machine learning algorithm (\( \alpha \)-GAN) which interpolates between several canonical GANs. For future work, we will investigate \( \alpha \)-GAN in practice, with particular interest in its generalization guarantees and its efficacy to reduce mode collapse.

APPENDIX A

PROOF OF THEOREM A

For a fixed generator, \( G_\theta \), we first solve the optimization problem
\[
\sup_{\omega \in \Omega} \int_X \frac{\alpha}{\alpha-1} \left( p_r(x) D_{\omega}(x) \frac{\omega(x)}{\omega(x)} + p_{G_\theta}(x)(1 - D_{\omega}(x)) \frac{1 - \omega(x)}{1 - \omega(x)} \right).
\]
(22)

Consider the function
\[
g(y) = \frac{\alpha}{\alpha-1} \left( a y \frac{\omega(x)}{\omega(x)} + b(1 - y) \frac{1 - \omega(x)}{1 - \omega(x)} \right),
\]
(23)
for $a, b \in \mathbb{R}_+$ and $y \in [0, 1]$. To show that the optimal discriminator is given by the expression in \((16)\), it suffices to show that $g(y)$ achieves its maximum in $[0, 1]$ at $y^* = \frac{a^{\alpha}}{a^{\alpha} + b^{\alpha}}$.

Notice that for $\alpha > 1$, $y^{\frac{\alpha}{\alpha - 1}}$ is a concave function of $y$, meaning the function $g$ is concave. For $0 < \alpha < 1$, $y^{\frac{\alpha}{\alpha - 1}}$ is a convex function of $y$, but since $\frac{\alpha}{\alpha - 1}$ is negative, the overall function $g$ is again concave. Consider the derivative $g'(y^*) = 0$, which gives us

$$y^* = \frac{a^{\alpha}}{a^{\alpha} + b^{\alpha}}$$

\((24)\)

This gives \((16)\). With this, the optimization problem in \((15)\) can be written as $	ext{inf}_{\theta \in \Theta} C(\theta)$, where

$$C(\theta) = \frac{\alpha}{\alpha - 1} \times \left[ \int_X \left( p_r(x)D_{\alpha}^* x^{\frac{\alpha}{\alpha - 1}} + p_{G_\theta}(x)(1 - D_{\alpha}^* x^{\frac{\alpha}{\alpha - 1}}) \right) dx - 2 \right]$$

\((25)\)

\[= \frac{\alpha}{\alpha - 1} \left[ \int_X \left( p_r(x) \left( \frac{p_r(x)}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha} \right)^{\frac{\alpha}{\alpha - 1}} + p_{G_\theta}(x) \left( \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha} \right)^{\frac{\alpha}{\alpha - 1}} \right) dx - 2 \right] \]

\((26)\)

\[= \frac{\alpha}{\alpha - 1} \left( \int_X (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{2}} dx - 2 \right) \]

\((27)\)

\[= D_{f_\alpha}(P_r||P_{G_\theta}) + \frac{\alpha}{\alpha - 1} \left( \frac{2^\frac{1}{2} - 2}{2^\frac{1}{2}} \right) \]

\((28)\)

where for the convex function $f_\alpha$ in \((18)\),

$$D_{f_\alpha}(P_r||P_{G_\theta}) = \int_X p_{G_\theta}(x)f_\alpha \left( \frac{p_r(x)}{p_{G_\theta}(x)} \right) dx \]

\((29)\)

\[= \frac{\alpha}{\alpha - 1} \left( \int_X (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{2}} dx - 2^\frac{1}{2} \right) \]

\((30)\)

This gives \((17)\). Since $D_{f_\alpha}(P_r||P_{G_\theta}) \geq 0$ with equality if and only if $P_r = P_{G_\theta}$, we have $C(\theta) \geq \frac{\alpha}{\alpha - 1} \left( 2^\frac{1}{2} - 2 \right)$ with equality if and only if $P_r = P_{G_\theta}$.

**APPENDIX B**

**PROOF OF THEOREM 2**

First, using L'Hôpital's rule we can verify that, for $a, b > 0$,

$$\lim_{\alpha \to 1} \frac{\alpha}{\alpha - 1} \left( (a^\alpha + b^\alpha)^{\frac{1}{\alpha}} - 2^\frac{1}{2} - 1(a + b) \right) = a \log \left( \frac{a}{a + b} \right) + b \log \left( \frac{b}{a + b} \right).$$

\((31)\)

Using this, we have

$$D_{f_\alpha}(P_r||P_{G_\theta}) \overset{\text{def}}{=} \lim_{\alpha \to 1} D_{f_\alpha}(P_r||P_{G_\theta}) \]

\((32)\)

$$= \lim_{\alpha \to 1} \frac{\alpha}{\alpha - 1} \left( \int_X (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{2}} dx - 2^\frac{1}{2} \right)$$

\((33)\)

$$\int_X \left( (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{\alpha}} - 2^\frac{1}{2} - 1(p_r(x) + p_{G_\theta}(x)) \right) dx \]

\((34)\)

$$= \int_X p_r(x) \log \left( \frac{p_r(x)}{p_r(x) + p_{G_\theta}(x)} \right) dx +$$

\[\int_X p_{G_\theta}(x) \log \left( \frac{p_{G_\theta}(x)}{p_r(x) + p_{G_\theta}(x)} \right) dx \]

\((35)\)

$$= 2D_{JS}(P_r||P_{G_\theta}),$$

\((36)\)

where $D_{JS}(\cdot||\cdot)$ is the Jensen-Shannon divergence. Now, as $\alpha \to 1$, \((17)\) equals $\text{inf}_{\theta \in \Theta} 2D_{TV}(P_r||P_{G_\theta}) - \log 4$ recovering vanilla GAN.

Substituting $\alpha = \frac{1}{2}$ in \((19)\), we get

$$D_{f_\frac{1}{2}}(P_r||P_{G_\theta}) = -\int_X \left( \sqrt{p_r(x)} + \sqrt{p_{G_\theta}(x)} \right)^2 dx + 4$$

\((37)\)

$$= \int_X \left( \sqrt{p_r(x)} - \sqrt{p_{G_\theta}(x)} \right)^2 dx$$

\((38)\)

$$=: 2D_{RS}(P_r||P_{G_\theta}),$$

\((39)\)

where $D_{RS}(P_r||P_{G_\theta})$ is the squared Hellinger distance. For $\alpha = \frac{1}{2}$, \((17)\) gives $2\text{inf}_{\theta \in \Theta} D_{RS}(P_r||P_{G_\theta}) - 2$ recovering Hellinger GAN (up to a constant).

Noticing that, for $a, b > 0$, $\lim_{\alpha \to \infty} (a^\alpha + b^\alpha)^{\frac{1}{\alpha}} = \max\{a, b\}$ and defining $A := \{x \in X : p_r(x) \geq p_{G_\theta}(x)\}$, we have

$$D_{f_\alpha}(P_r||P_{G_\theta}) \overset{\text{def}}{=} \lim_{\alpha \to \infty} D_{f_\alpha}(P_r||P_{G_\theta})$$

\((40)\)

$$= \lim_{\alpha \to \infty} \frac{\alpha}{\alpha - 1} \left( \int_X (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{2}} dx - 2^\frac{1}{2} \right)$$

\((41)\)

$$= \int_X \max\{p_r(x), p_{G_\theta}(x)\} dx - 1$$

\((42)\)

$$= \int_A \max\{p_r(x) - p_{G_\theta}(x), 0\} dx$$

\((43)\)

$$= \int_A (p_r(x) - p_{G_\theta}(x)) dx$$

\((44)\)

$$= \int_A \frac{p_r(x) - p_{G_\theta}(x)}{2} dx + \int_{A^c} \frac{p_{G_\theta}(x) - p_r(x)}{2} dx$$

\((45)\)

$$= \frac{1}{2} \int_X |p_r(x) - p_{G_\theta}(x)| dx$$

\((46)\)

$$=: D_{TV}(P_r||P_{G_\theta}),$$

\((47)\)

where $D_{TV}(P_r||P_{G_\theta})$ is the total variation distance between $P_r$ and $P_{G_\theta}$. Thus, as $\alpha \to \infty$, \((17)\) equals $\text{inf}_{\theta \in \Theta} D_{TV}(P_r||P_{G_\theta}) - 1$ recovering TV-GAN (modulo a constant).
Appendix C
Proof of Theorem 5

Consider a sequence of probability distributions \((P_n)\). To prove the theorem, notice that it suffices to show that Arimoto divergence for any \(\alpha > 0\) is equivalent to the total variation distance, i.e., \(D_{f_\alpha}(P_n || P) \to 0\) if and only if \(D_{TV}(P_n || P) \to 0\). To this end, we employ a property of Arimoto divergence which gives lower and upper bounds on it in terms of the total variation distance. In particular, Österreicher and Vajda [20, Theorem 2] proved that for any \(\alpha > 0\), probability distributions \(P, Q\), we have

\[
\psi_\alpha(D_{TV}(P || Q)) \leq D_{f_\alpha}(P || Q) \leq \psi_\alpha(1)D_{TV}(P || Q),
\]

(48)

where the function \(\psi_\alpha : [0,1] \to \mathbb{R}\) defined by \(\psi_\alpha(p) = \frac{\alpha^\alpha}{\alpha - 1} \left( (1 + p)^{\alpha} + (1 - p)^{\alpha} \right) \frac{1}{\alpha - 1} - 2 \frac{\alpha}{\alpha - 1}\) for \(\alpha \in (0,1) \cup (1, \infty)\) is convex and strictly monotone increasing such that \(\psi_\alpha(0) = 0\) and \(\psi_\alpha(1) = \frac{\alpha}{\alpha - 1} \left( 2 - 2 \frac{\alpha}{\alpha - 1}\right)\).

We first prove the ‘only if’ part, i.e., Arimoto divergence is stronger than the total variation distance. Suppose \(D_{f_\alpha}(P_n || P) \to 0\). From the lower bound in (48), it follows that \(\psi_\alpha(D_{TV}(P_n || P)) \leq D_{f_\alpha}(P_n || P)\), for each \(n \in \mathbb{N}\). This implies that \(\psi_\alpha(D_{TV}(P_n || P)) \to 0\). We show below that \(\psi_\alpha\) is invertible and \(\psi_\alpha^{-1}\) is continuous. Then it would follow that

\[
\psi_\alpha^{-1} \psi_\alpha(D_{TV}(P_n || P)) = D_{TV}(P_n || P) \to \psi_\alpha^{-1}(0) = 0
\]

proving that Arimoto divergence is stronger than the total variation distance. It remains to show that \(\psi_\alpha\) is invertible and \(\psi_\alpha^{-1}\) is continuous. Invertibility follows directly from the fact that \(\psi_\alpha\) is strictly monotone increasing function. For the continuity of \(\psi_\alpha^{-1}\), it suffices to show that \(\psi_\alpha(C)\) is closed for a closed set \(C \subseteq [0,1]\). The closed set \(C\) is compact since a closed subset of a compact set \([0,1]\) in this case is also compact. Note that convexity of \(\psi_\alpha\) implies continuity and \(\psi_\alpha(C)\) is compact since a continuous function of a compact set is also compact. By Heine-Borel theorem, this gives that \(\psi_\alpha(C)\) is closed (and bounded) as desired.

For the ‘if part’, i.e., to prove that the total variation distance is stronger than Arimoto divergence, consider a sequence of probability distributions \((P_n)\) such that \(D_{TV}(P_n || P) \to 0\). It follows from the upper bound in (48) that \(D_{f_\alpha}(P_n || P) \leq D_{TV}(P_n || P)\), for each \(n \in \mathbb{N}\). This implies that \(D_{f_\alpha}(P_n || P) \to 0\) which completes the proof.

Appendix D
Proof of Theorem 3

We know from [12, Corollary 1] that for \(\eta \in [0,1]\),

\[
\inf_t \eta \tilde{\ell}_\alpha(t) + (1 - \eta)\tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha - 1} \left( 1 - (\eta^\alpha + (1 - \eta)^{\alpha}) \frac{\alpha}{\alpha - 1} \right).
\]

This implies that

\[
\inf_t \eta \tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha - 1} \left( 1 + \eta \frac{\alpha}{1 - \eta} - \left( \left( \frac{\eta}{1 - \eta} \right)^\alpha + 1 \right) \right)
\]

(49)

Now substituting \(u = \frac{\eta}{1 - \eta}\) and taking negation in (49), we get

\[
- \inf_t \eta \tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha - 1} \left( \frac{\alpha}{\alpha - 1} + 1 \right) - (1 + u),
\]

for \(u \geq 0\)

(50)

giving us (21).

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, 2014, pp. 2672–2680.

[2] J. H. Lim and J. C. Ye, “Geometric GAN,” arXiv preprint arXiv:1705.02894, 2017.

[3] L. Cai, Y. Chen, N. Cai, W. Cheng, and H. Wang, “Utilizing amari-alpha divergence to stabilize the training of generative adversarial networks,” Entropy, vol. 22, no. 4, p. 410, 2020.

[4] S. Nowozin, B. Cseke, and R. Tomioka, “f-GAN: Training generative neural samplers using variational divergence minimization,” in Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, p. 271–279.

[5] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 214–223.

[6] T. Liang, “How well generative adversarial networks learn distributions,” arXiv preprint arXiv:1811.03179, 2018.

[7] F. Huszár, “How (not) to train your generative model: Scheduled sampling, likelihood, adversary?” arXiv preprint arXiv:1511.05101, 2015.

[8] L. Metz, B. Bozo, D. Pfau, and J. Sohl-Dickstein, “Unrolled generative adversarial networks,” arXiv preprint arXiv:1611.02163, 2016.

[9] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” arXiv preprint arXiv:1606.03498, 2016.

[10] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” arXiv preprint arXiv:1701.04862, 2017.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[12] T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz, “A tunable loss function for binary classification,” in IEEE International Symposium on Information Theory, 2019, pp. 2479–2483.

[13] T. Sypherd, M. Diaz, L. Sankar, and G. Dasarathy, “On the α-loss landscape in the logistic model,” in IEEE International Symposium on Information Theory, 2020, pp. 2700–2705.

[14] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119 – 139, 1997.

[15] N. Merhav and M. Feder, “Universal prediction,” IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2124–2147, 1998.

[16] T. A. Courtade and R. D. Wesel, “Multiterminal source coding with an entropy-based distortion measure,” in IEEE International Symposium on Information Theory, 2011, pp. 2040–2044.

[17] X. Nguyen, M. J. Wainwright, and M. I. Jordan, “On surrogate loss functions and f-divergences,” The Annals of Statistics, vol. 37, no. 2, pp. 876–904, 2009.

[18] P. L. Bartlett, M. I. Jordan, and D. J. Mcallester, “Convexity, classification, and risk bounds,” Journal of the American Statistical Association, vol. 101, no. 473, pp. 138–156, 2006.

[19] F. Liele and I. Vajda, “On divergences and informations in statistics and information theory,” IEEE Transactions on Information Theory, vol. 52, no. 10, pp. 4394–4412, 2006.

[20] F. Österreicher and I. Vajda, “A new class of metric divergences on probability spaces and its applicability in statistics,” Annals of the Institute of Statistical Mathematics, vol. 55, no. 3, pp. 639–653, 2003.

[21] T. Sypherd, M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz, and L. Sankar, “A tunable loss function for robust classification: Calibration, landscape, and generalization,” arXiv preprint arXiv:1906.02314, 2019.

[22] F. Österreicher, “On a class of perimeter-type distances of probability distributions,” Kybernetika, vol. 32, no. 4, pp. 389–393, 1996.
[23] S. Liu, O. Bousquet, and K. Chaudhuri, “Approximation and convergence properties of generative adversarial learning,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[24] S. Arimoto, “Information-theoretical considerations on estimation problems,” Information and control, vol. 19, no. 3, pp. 181–194, 1971.

[25] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, “A tunable measure for information leakage,” in 2018 IEEE International Symposium on Information Theory (ISIT). IEEE, 2018, pp. 701–705.

[26] J. Lin, “Divergence measures based on the shannon entropy,” IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 145–151, 1991.

[27] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, “Generalization and equilibrium in generative adversarial nets (GANs),” in Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 224–232.

[28] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observation,” Studia Scientiarum Mathematicarum Hungarica, vol. 2, pp. 229–318, 1967.

[29] M. Wiatrak, S. V. Albrecht, and A. Nystrom, “Stabilizing generative adversarial networks: A survey,” arXiv preprint arXiv:1910.00927, 2019.

[30] M. D. Reid and R. C. Williamson, “Composite binary losses,” The Journal of Machine Learning Research, vol. 11, pp. 2387–2422, 2010.

[31] C. Villani, Optimal transport: old and new. Springer Science & Business Media, 2008, vol. 338.