Realizing GANs via a Tunable Loss Function

Gowtham R. Kurri, Tyler Sypherd, and Lalitha Sankar
Arizona State University, {gkurri,tsypherd,lalithasankar}@asu.edu

2021 IEEE Information Theory Workshop (ITW) | 978-1-6654-0312-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITW48936.2021.9611499

***Abstract*—We introduce a tunable GAN, called α-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 α-GAN using a supervised loss function, namely, α-loss, which is a tunable loss function capturing several canonical losses. We show that α-GAN is intimately related to the Arimoto divergence, which was first proposed by Österriecher (1996), and later studied by Liese and Vajda (2006). We posit that the holistic understanding that α-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] [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 α-GAN based on α-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 α-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 α-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 α-loss is a tilted distribution), the generator seeks to minimize the *Arimoto divergence* (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 α-GAN interpolates between various f-GANs including vanilla GAN ($\alpha = 1$), Hellinger GAN [4] ($\alpha = 1/2$), Total Variation GAN [4] ($\alpha = \infty$), and IPM-based GANs including WGANs (when the discriminator set is appropriately constrained) by smoothly tuning the hyperparameter α (see Theorem 2 and (9)). Thus, α-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 α-loss in classification [13], [21], where the model performance saturates quickly for $\alpha \to \infty$, we expect a similar saturation for α-GAN (see Figure 1). Thus, we posit that smooth tuning from JSD to IPM that results from increasing α from 1 to ∞ can address issues like mode collapse, vanishing gradients, etc.
- Finally in Theorem 3, we reconstruct the Arimoto divergence using the margin-based form of α-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].

This work is supported in part by NSF grants CIF-1901243, CIF-2007688, and CIF-2134256.

[1]We refer to the GAN introduced by Goodfellow *et al.* [1] as *vanilla* GAN, as done in the literature [2], [3] to distinguish it from others introduced later.

978-1-6654-0312-2/21/$31.00 ©2021 IEEE

II. α-LOSS AND GANS

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

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

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

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 [23], and later Liao *et al.* used α-loss to model the inferential capacity of an adversary to obtain private attributes [24]. 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 $\theta \in \Theta \subset \mathbb{R}^{n_g}$ and $\omega \in \Omega \subset \mathbb{R}^{n_d}$, respectively, and write G_θ and D_ω (θ and ω are typically the weights of neural network models for the generator and the discriminator, respectively). The generator G_θ takes as input a $d'(\ll d)$-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_θ} (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

$$\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V(\theta, \omega). \tag{2}$$

Goodfellow *et al.* [1] introduced a value function

$$\begin{aligned} &V_{\text{VG}}(\theta, \omega) \\ &= \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[\log(1 - D_\omega(G_\theta(Z)))] \\ &= \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] + \mathbb{E}_{X \sim P_{G_\theta}}[\log(1 - D_\omega(X))] \end{aligned} \tag{3}$$

and showed that when the discriminator class $\{D_\omega\}$, parametrized by ω, is rich enough, (2) simplifies to finding the $\inf_{\theta \in \Theta} 2D_{\text{JS}}(P_r \| P_{G_\theta}) - \log 4$, where $D_{\text{JS}}(P_r \| P_{G_\theta})$ is the Jensen-Shannon divergence [25] between P_r and P_{G_θ}. This simplification is achieved, for any G_θ, by choosing the optimal discriminator

$$D_{\omega^*}(x) = \frac{p_r(x)}{p_r(x) + p_{G_\theta}(x)}, \tag{4}$$

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

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

$$V_f(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] + \mathbb{E}_{X \sim P_{G_\theta}}[f^*(D_\omega(X))], \tag{5}$$

where[2] $D_\omega : \mathcal{X} \to \mathbb{R}$ and $f^*(t) \triangleq \sup_u \{ut - f(u)\}$ is the Fenchel conjugate of a convex lower semincontinuous function f, for any f-divergence $D_f(P_r \| P_{G_\theta}) := \int_{\mathcal{X}} p_{G_\theta}(x) f\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right) dx$ [26]–[28] (not just the Jensen-Shannon divergence) leveraging its variational characterization [29]. 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 $T_{\omega^*}(x) = f'\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)$. Rényi divergence measures are also studied in the context of GANs [30]–[32].

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

$$W(P_r, P_{G_\theta}) = \inf_{\Gamma_{X_1 X_2} \in \Pi(P_r, P_{G_\theta})} \mathbb{E}_{(X_1, X_2) \sim \Gamma_{X_1 X_2}} \|X_1 - X_2\|_2, \tag{6}$$

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_θ}. WGAN employs the Kantorovich-Rubinstein duality [33] using the value function

$$V_{\text{WGAN}}(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] - \mathbb{E}_{X \sim P_{G_\theta}}[D_\omega(X)], \tag{7}$$

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 Ω is rich enough. Although, various GANs have been proposed in the literature, each of them exhibits their own strengths and weaknesses in terms of convergence, vanishing gradients, mode collapse, computational complexity, etc. leaving the problem of instability unsolved [34].

III. TUNABLE α-GAN

Noting that a GAN involves a classifier (i.e., discriminator), it is well known that the value function $V_{\text{VG}}(\theta, \omega)$ in (3) considered by Goodfellow *et al.* [1] is related to cross entropy loss. While perhaps it has not been explicitly articulated heretofore in the literature, we first formalize this loss function perspective of GANs and propose a tunable GAN based on α-loss generalizing vanilla GAN and various other GANs. In [35], Arora *et al.* observed that the log function in (3) can be replaced by any 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 a 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 observe that the value function

[2] This is a slight abuse of notation in that D_ω is not a probability here. However, we chose this for consistency in notation of discriminator across various GANs.

V_{VG} in (3) for the vanilla GAN can be expressed in terms of cross-entropy loss $\ell_{\text{CE}}(y,\hat{y}) \triangleq -y\log\hat{y} - (1-y)\log(1-\hat{y})$ as

$$V_{\text{VG}}(\theta,\omega) = \mathbb{E}_{X|y=1}[-\ell_{\text{CE}}(y, D_\omega(X))] + \mathbb{E}_{X|y=0}[-\ell_{\text{CE}}(y, D_\omega(X))] = \mathbb{E}_{X\sim P_r}[-\ell_{\text{CE}}(1, D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell_{\text{CE}}(0, D_\omega(X))].$$

Now we write α-loss in (1) analogous to ℓ_{CE} to obtain

$$\ell_\alpha(y,\hat{y}) := \frac{\alpha}{\alpha-1}\left(1 - y\hat{y}^{\frac{\alpha-1}{\alpha}} - (1-y)(1-\hat{y})^{\frac{\alpha-1}{\alpha}}\right), \quad (8)$$

for $\alpha \in (0,1)\cup(1,\infty)$. Note that (8) recovers ℓ_{CE} as $\alpha \to 1$. Now consider a *tunable α-GAN* with a value function

$$V_\alpha(\theta,\omega) = \mathbb{E}_{X\sim P_r}[-\ell_\alpha(1, D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell_\alpha(0, D_\omega(X))] = \frac{\alpha}{\alpha-1}\times \left(\mathbb{E}_{X\sim P_r}\left[D_\omega(X)^{\frac{\alpha-1}{\alpha}}\right] + \mathbb{E}_{X\sim P_{G_\theta}}\left[(1-D_\omega(X))^{\frac{\alpha-1}{\alpha}}\right] - 2\right).$$

We can verify that $\lim_{\alpha\to1} V_\alpha(\theta,\omega) = V_{\text{VG}}(\theta,\omega)$ recovering the value function of the vanilla GAN. Also, notice that

$$\lim_{\alpha\to\infty} V_\alpha(\theta,\omega) = \mathbb{E}_{X\sim P_r}[D_\omega(x)] - \mathbb{E}_{X\sim P_{G_\theta}}[D_\omega(x)] - 1 \quad (9)$$

is the value function (modulo a constant) used in Intergral Probability Metric (IPM) based GANs[3], e.g., WGAN, McGan [36], Fisher GAN [37], and Sobolev GAN [38]. The resulting min-max game in α-GAN is given by

$$\inf_{\theta\in\Theta}\sup_{\omega\in\Omega} V_\alpha(\theta,\omega). \quad (10)$$

The following theorem provides the min-max solution, i.e., Nash equilibrium, to the two-player game in (10) for the non-parametric setting, i.e., when the discriminator set Ω is large enough.

Theorem 1 (min-max solution)**.** *For a fixed generator G_θ, the discriminator $D_{\omega^*}(x)$ optimizing the* sup *in* (10) *is given by*

$$D_{\omega^*}(x) = \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha}. \quad (11)$$

For this $D_{\omega^*}(x)$, (10) *simplifies to minimizing a non-negative symmetric f_α-divergence $D_{f_\alpha}(\cdot||\cdot)$ as*

$$\inf_{\theta\in\Theta} D_{f_\alpha}(P_r||P_{G_\theta}) + \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}} - 2\right), \quad (12)$$

where

$$f_\alpha(u) = \frac{\alpha}{\alpha-1}\left((1+u^\alpha)^{\frac{1}{\alpha}} - (1+u) - 2^{\frac{1}{\alpha}} + 2\right), \quad (13)$$

for $u \geq 0$ and[4]

$$D_{f_\alpha}(P||Q) = \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}} (p(x)^\alpha + q(x)^\alpha)^{\frac{1}{\alpha}}\, dx - 2^{\frac{1}{\alpha}}\right), \quad (14)$$

which is minimized iff $P_{G_\theta} = P_r$.

For intuition on the construction of (13), see Theorem 3.

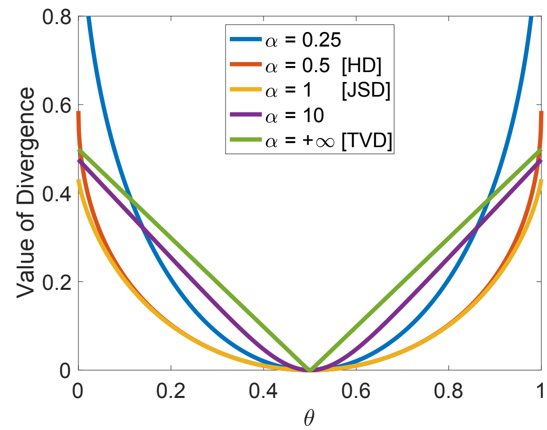

Fig. 1. A plot of D_{f_α} in (14) for several values of α where $p \sim \text{Ber}(1/2)$ and $q \sim \text{Ber}(\theta)$. Note that HD, JSD, and TVD, are abbreviations for Hellinger, Jensen-Shannon, and Total Variation divergences, respectively. As $\alpha \to 0$, the curvature of the divergence increases, placing increasingly more weight on $\theta \neq 1/2$. Conversely, for $\alpha \to \infty$, D_{f_α} quickly resembles D_{f_∞}, hence a saturation effect of D_{f_α}.

Remark 1. It can be inferred from (12) that when the discriminator is trained to optimality, the generator has to minimize the f_α-divergence hinting at an application of f-GAN instead. Implementing f_α-GAN directly via value function in (5) (for f_α) involves finding convex conjugate of f_α, which is challenging in terms of computational complexity making it inconvenient 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_α-divergence via losses. A related work where an f-divergence (in particular, α-divergence [39]) 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 (11) implies a more cautious discriminator, i.e., if $p_{G_\theta}(x) \geq p_r(x)$, then $D_{w^*}(x)$ decays more slowly from $1/2$, and if $p_{G_\theta}(x) \leq p_r(x)$, $D_{w^*}(x)$ increases more slowly from $1/2$. Conversely, as $\alpha \to \infty$, (11) simplifies to $D_{\omega^*}(x) = \mathbb{1}\{p_r(x) > p_{G_\theta}(x)\} + \frac{1}{2}\mathbb{1}\{p_r(x) = p_{G_\theta}(x)\}$, where the discriminator implements the Maximum Likelihood (ML) decision rule, i.e., a hard decision whenever $p_r(x) \neq p_{G_\theta}(x)$. In other words, (11) for $\alpha \to \infty$ induces a very confident discriminator. Regarding the generator's perspective, (12) (and Figure 1) implies that the generator seeks to minimize the discrepancy between P_r and P_{G_θ} according to the geometry induced by D_{f_α}. 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.

[3]Note that IPMs do not restrict the function D_ω to be a probability.

[4]We note that the divergence D_{f_α} has been referred to as *Arimoto divergence* in the literature [19], [20], [22]. We refer the reader to Section IV for more details.

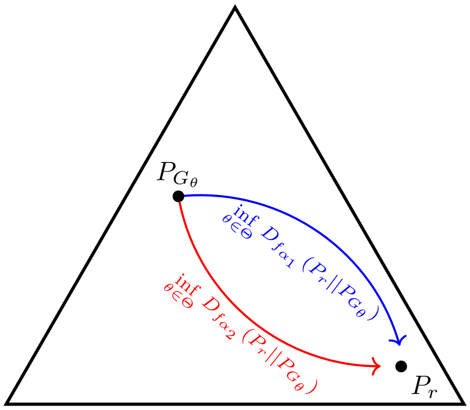

Fig. 2. An idealized illustration on the probability simplex of the infimum over θ in (12) for $\alpha_1, \alpha_2 \in (0, \infty]$ such that $\alpha_1 \neq \alpha_2$. The choice of α in the min-max game for the α-GAN in (10) defines the optimization trajectory taken by the generator (versus an optimal discriminator as specified in (11)) by distorting the underlying geometry according to D_{f_α}.

A detailed proof of Theorem 1 is in Appendix. Next we show that α-GAN recovers various well known f-GANs.

Theorem 2 (f-GANs)**.** *α-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 detailed proof is in Appendix.

IV. Reconstructing Arimoto Divergence

It is interesting to note that the divergence $D_{f_\alpha}(\cdot||\cdot)$ (in (14)) that naturally emerges from the analysis of α-GAN was first proposed by Österriecher [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 [40], [41]. For example:

- A geometric interpretation of the divergence D_{f_α} in the context of hypothesis testing [22].
- $D_{f_\alpha}(P||Q)^{\min\{\alpha,\frac{1}{2}\}}$ defines a distance metric (satisfying the triangle inequality) on the set of probability distributions [20].

When the Arimoto divergence D_{f_α} was proposed, the convexity of the generating function f_α 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[5] form of α-loss (see [21]) where the convexity of f_α (and also the symmetric property of $D_{f_\alpha}(\cdot||\cdot)$) 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 $\tilde{\ell}$, there is a corresponding f-divergence with the convex function f defined as $f(u) := -\inf_t \left(u\tilde{\ell}(t) + \tilde{\ell}(-t)\right)$. The convexity of f follows simply because the infimum of affine functions is concave, and this argument does not require $\tilde{\ell}$ to be convex[6]. Additionally, the f-divergence obtained is always symmetric because f satisfies $f(u) = uf(\frac{1}{u})$ since $\inf_t u\tilde{\ell}(t) + \tilde{\ell}(-t) = \inf_t \tilde{\ell}(t) + u\tilde{\ell}(-t)$.

The margin-based α-loss [12] for $\alpha \in (0,1) \cup (1, \infty)$, $\tilde{\ell}_\alpha : \bar{\mathbb{R}} \to \mathbb{R}_+$ is defined as

$$\tilde{\ell}_\alpha(t) \triangleq \frac{\alpha}{\alpha-1}\left(1 - \sigma(t)^{\frac{\alpha-1}{\alpha}}\right), \tag{15}$$

where $\sigma : \bar{\mathbb{R}} \to \mathbb{R}_+$ is the sigmoid function given by $\sigma(t) = (1 + \mathrm{e}^{-t})^{-1}$. With these preliminaries in hand, we have the following result.

Theorem 3. *For the function f_α in* (13)*, it holds that*

$$f_\alpha(u) = -\inf_t \left(u\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t)\right) - \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}} - 2\right) \quad \textit{for } u \geq 0. \tag{16}$$

A detailed proof is in Appendix.

V. Conclusion

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

Appendix

Proof of Theorem 1

For a fixed generator, G_θ, we first solve the optimization problem

$$\sup_{\omega\in\Omega} \int_{\mathcal{X}} \frac{\alpha}{\alpha-1}\left(p_r(x)D_\omega(x)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x)(1 - D_\omega(x))^{\frac{\alpha-1}{\alpha}}\right). \tag{17}$$

Consider the function

$$g(y) = \frac{\alpha}{\alpha-1}\left(ay^{\frac{\alpha-1}{\alpha}} + b(1-y)^{\frac{\alpha-1}{\alpha}}\right), \tag{18}$$

for $a, b \in \mathbb{R}_+$ and $y \in [0,1]$. To show that the optimal discriminator is given by the expression in (11), 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 $\alpha > 1$, $y^{\frac{\alpha-1}{\alpha}}$ is a concave function of y, meaning the function g is concave. For $0 < \alpha < 1$, $y^{\frac{\alpha-1}{\alpha}}$ 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

$$y^* = \frac{a^\alpha}{a^\alpha + b^\alpha}. \tag{19}$$

[5] In the binary classification context, the margin is represented by $t := yf(x)$, where $x \in \mathcal{X}$ is the feature vector, $y \in \{-1, +1\}$ is the label, and $f : \mathcal{X} \to \mathbb{R}$ is the prediction function produced by a learning algorithm.

[6] in fact α-loss in its margin-based form is only quasi-convex for $\alpha > 1$

This gives (11). With this, the optimization problem in (10) can be written as $\inf_{\theta\in\Theta} C(G_\theta)$, where

$$C(G_\theta) = \frac{\alpha}{\alpha-1}\times \left[\int_{\mathcal{X}} \left(p_r(x)D_{\omega^*}(x)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x)(1-D_{\omega^*}(x))^{\frac{\alpha-1}{\alpha}}\right)dx - 2\right] \tag{20}$$

$$= \frac{\alpha}{\alpha-1}\left[\int_{\mathcal{X}}\left(p_r(x)\left(\frac{p_r(x)^\alpha}{p_r(x)^\alpha+p_{G_\theta}(x)^\alpha}\right)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x)\left(\frac{p_r(x)^\alpha}{p_r(x)^\alpha+p_{G_\theta}(x)^\alpha}\right)^{\frac{\alpha-1}{\alpha}}\right)dx - 2\right] \tag{21}$$

$$= \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}}dx - 2\right) \tag{22}$$

$$= D_{f_\alpha}(P_r||P_{G_\theta}) + \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}}-2\right), \tag{23}$$

where for the convex function f_α in (13),

$$D_{f_\alpha}(P_r||P_{G_\theta}) = \int_{\mathcal{X}} p_{G_\theta}(x) f_\alpha\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx \tag{24}$$

$$= \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}}dx - 2^{\frac{1}{\alpha}}\right). \tag{25}$$

This gives us (12). Since $D_{f_\alpha}(P_r||P_{G_\theta}) \geq 0$ with equality if and only if $P_r = P_{G_\theta}$, we have $C(G_\theta) \geq \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}}-2\right)$ with equality if and only if $P_r = P_{G_\theta}$.

Proof of Theorem 2

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

$$\lim_{\alpha\to 1}\frac{\alpha}{\alpha-1}\left((a^\alpha+b^\alpha)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}-1}(a+b)\right) = a\log\left(\frac{a}{\frac{a+b}{2}}\right) + b\log\left(\frac{b}{\frac{a+b}{2}}\right). \tag{26}$$

Using this, we have

$$D_{f_1}(P_r||P_{G_\theta}) \triangleq \lim_{\alpha\to 1} D_{f_\alpha}(P_r||P_{G_\theta}) \tag{27}$$

$$= \lim_{\alpha\to 1}\frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}}dx - 2^{\frac{1}{\alpha}}\right) \tag{28}$$

$$= \lim_{\alpha\to 1}\left[\frac{\alpha}{\alpha-1}\times \int_{\mathcal{X}}\left(\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}-1}(p_r(x)+p_{G_\theta}(x))\right)dx\right] \tag{29}$$

$$= \int_{\mathcal{X}} p_r(x)\log\frac{p_r(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx + \int_{\mathcal{X}} p_{G_\theta}(x)\log\frac{p_{G_\theta}(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx \tag{30}$$

$$=: 2D_{\text{JS}}(P_r||P_{G_\theta}), \tag{31}$$

where $D_{\text{JS}}(\cdot||\cdot)$ is the Jensen-Shannon divergence. Now, as $\alpha\to 1$, (12) equals $\inf_{\theta\in\Theta} 2D_{\text{JS}}(P_r||P_{G_\theta}) - \log 4$ recovering vanilla GAN.

Substituting $\alpha = \frac{1}{2}$ in (14), we get

$$D_{f_{\frac{1}{2}}}(P_r||P_{G_\theta}) = -\int_{\mathcal{X}}\left(\sqrt{p_r(x)} + \sqrt{p_{G_\theta}(x)}\right)^2 dx + 4 \tag{32}$$

$$= \int_{\mathcal{X}}\left(\sqrt{p_r(x)} - \sqrt{p_{G_\theta}(x)}\right)^2 dx \tag{33}$$

$$=: 2D_{\text{H}^2}(P_r||P_{G_\theta}), \tag{34}$$

where $D_{\text{H}^2}(P_r||P_{G_\theta})$ is the squared Hellinger distance. For $\alpha = \frac{1}{2}$, (12) gives $2\inf_{\theta\in\Theta} D_{\text{H}^2}(P_r||P_{G_\theta}) - 2$ 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 $\mathcal{A} := \{x\in\mathcal{X} : p_r(x) \geq p_{G_\theta}(x)\}$, we have

$$D_{f_1}(P_r||P_{G_\theta}) \triangleq \lim_{\alpha\to\infty} D_{f_\alpha}(P_r||P_{G_\theta}) \tag{35}$$

$$= \lim_{\alpha\to\infty}\frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}}dx - 2^{\frac{1}{\alpha}}\right) \tag{36}$$

$$= \int_{\mathcal{X}}\max\{p_r(x), p_{G_\theta}(x)\}\ dx - 1 \tag{37}$$

$$= \int_{\mathcal{X}}\max\{p_r(x) - p_{G_\theta}(x), 0\}\ dx \tag{38}$$

$$= \int_{\mathcal{A}}\left(p_r(x) - p_{G_\theta}(x)\right)\ dx \tag{39}$$

$$= \int_{\mathcal{A}}\frac{p_r(x) - p_{G_\theta}(x)}{2}\ dx + \int_{\mathcal{A}^c}\frac{p_{G_\theta}(x) - p_r(x)}{2}\ dx \tag{40}$$

$$= \frac{1}{2}\int_{\mathcal{X}}\left|p_r(x) - p_{G_\theta}(x)\right|\ dx \tag{41}$$

$$=: D_{\text{TV}}(P_r||P_{G_\theta}), \tag{42}$$

where $D_{\text{TV}}(P_r||P_{G_\theta})$ is the total variation distance between P_r and P_{G_θ}. Thus, as $\alpha\to\infty$, (12) equals $\inf_{\theta\in\Theta} D_{\text{TV}}(P_r||P_{G_\theta}) - 1$ recovering TV-GAN (modulo a constant).

Proof of Theorem 3

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

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

This implies that

$$\inf_t \frac{\eta}{1-\eta}\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha-1}\left(1 + \frac{\eta}{1-\eta} - \left(\left(\frac{\eta}{1-\eta}\right)^\alpha + 1\right)^{\frac{1}{\alpha}}\right) \tag{43}$$

Now substituting u for $\frac{\eta}{1-\eta}$ and taking negation in (43), we get

$$-\inf_t u\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha-1}\left((u^\alpha+1)^{\frac{1}{\alpha}} - (1+u)\right), \quad \text{for } u\geq 0 \tag{44}$$

giving us (16).

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, p. 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. Poole, 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 J. D. Mcauliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[19] F. Liese 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. Arimoto, "Information-theoretical considerations on estimation problems," *Information and control*, vol. 19, no. 3, pp. 181–194, 1971.

[24] J. Liao, O. Kosut, L. Sankar, and F. du Pin Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8043–8066, 2019.

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

[26] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

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

[28] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[29] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[30] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. Katsoulakis, "Cumulant gan," *arXiv preprint arXiv:2006.06625*, 2020.

[31] H. Bhatia, W. Paul, F. Alajaji, B. Gharesifard, and P. Burlina, "Least k th-order and rényi generative adversarial networks," *Neural Computation*, vol. 33, no. 9, pp. 2473–2510, 2021.

[32] A. Sarraf and Y. Nie, "Rgan: Rényi generative adversarial network," *SN Computer Science*, vol. 2, no. 1, pp. 1–8, 2021.

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

[34] M. Wiatrak, S. V. Albrecht, and A. Nystrom, "Stabilizing generative adversarial networks: A survey," *arXiv preprint arXiv:1910.00927*, 2019.

[35] 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.

[36] Y. Mroueh, T. Sercu, and V. Goel, "McGan: Mean and covariance feature matching GAN," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 2527–2535.

[37] Y. Mroueh and T. Sercu, "Fisher GAN," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[38] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev GAN," *arXiv preprint arXiv:1711.04894*, 2017.

[39] S.-i. Amari, *α-Divergence and α-Projection in Statistical Manifold*. New York, NY: Springer New York, 1985, pp. 66–103.

[40] P. Cerone, S. S. Dragomir, and F. Österreicher, "Bound on extended f-divergences for a variety of classes," *Kybernetika*, vol. 40, no. 6, pp. 745–756, 2004.

[41] I. Vajda, "On metric divergences of probability measures," *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.