α-GAN: Convergence and Estimation Guarantees

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

***Abstract*—We present a unified approach to generative adversarial networks (GANs) through the lens of class probability estimation (CPE) losses. From this perspective, we prove a two-way correspondence between the min-max optimization of CPE loss function GANs and the minimization of associated f-divergences. In particular, we focus on α-GAN, which interpolates several known GANs (Hellinger, vanilla, and Total Variation) via the α-loss and corresponds to the minimization of the classical Arimoto divergence. We show that the Arimoto divergences induced by α-GAN equivalently converge, for all $\alpha \in (0, \infty]$. However, under restricted learning models and finite samples, we provide estimation bounds which indicate diverse GAN characteristics as a function of α. Furthermore, we present empirical results on a toy dataset that highlight the utility of tuning the α hyperparameter.**

I. Introduction

Generative adversarial networks (GANs) are *generative models* capable of producing new samples from an unknown (real) distribution using a finite number of training data samples. A GAN is composed of two modules, a generator G and a discriminator D, parameterized by vectors $\theta \in \Theta \subset \mathbb{R}^{n_g}$ and $\omega \in \Omega \subset \mathbb{R}^{n_d}$, respectively, which play an adversarial game with one another. The generator G_θ takes as input noise $Z \sim P_Z$ and maps it to a data sample in $\mathcal{X}$ via the mapping $z \mapsto G_\theta(z)$ with an aim of mimicking data from the real distribution P_r. For an input $x \in \mathcal{X}$, the discriminator classifies if it is real data or generated data by outputting $D_\omega(x) \in [0, 1]$, the probability that x comes from P_r (real) as opposed to P_{G_θ} (synthetic). The opposing goals of the generator and the discriminator lead to a zero-sum min-max game with a chosen value function $V(\theta, \omega)$ resulting in an optimization problem given by

$$\inf_{\theta\in\Theta} \sup_{\omega\in\Omega} V(\theta,\omega). \tag{1}$$

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

$$V_{\text{VG}}(\theta,\omega) = \mathbb{E}_{X\sim P_r}[\log D_\omega(X)] + \mathbb{E}_{X\sim P_{G_\theta}}[\log(1 - D_\omega(X))], \tag{2}$$

for which they showed that, when the discriminator class $\{D_\omega\}_{\omega\in\Omega}$ is rich enough, (1) simplifies to $\inf_{\theta\in\Theta} 2D_{\text{JS}}(P_r||P_{G_\theta}) - \log 4$, where $D_{\text{JS}}(P_r||P_{G_\theta})$ is the Jensen-Shannon divergence [2] between P_r and P_{G_θ}. This simplification is achieved, for any G_θ, by the optimal discriminator $D_{\omega^*}(x)$ maximizing (2) which has the form

$$D_{\omega^*}(x) = \frac{p_r(x)}{p_r(x) + p_{G_\theta}(x)}, \tag{3}$$

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).

Various other GANs have been studied in the literature (e.g., f-divergence based GANs known as f-GAN [3], IPM based GANs [4], [5], Cumulant GAN [6], to name a few) with different value functions. In each case, the corresponding min-max optimization problem simplifies to minimizing some measure of divergence between the real and generated distributions. Yet, a methodical way to compare and operationally interpret GAN value functions remains open.

Recently, in [7], we introduce a loss function [8] perspective of GANs where we show that a GAN can be formulated using *any* class probability estimation (CPE) loss $\ell(y,\hat{y})$ with inputs $y \in \{0,1\}$ (the true label) and predictor $\hat{y} \in [0,1]$ (soft prediction of y). We show that using CPEs, the value function (objective) in (1) can be written as

$$V(\theta,\omega) = \mathbb{E}_{X\sim P_r}[-\ell(1,D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell(0,D_\omega(X))] \tag{4}$$

(see extended version for more details on this []). We specialize the setup in (4) to introduce α-GAN using α-loss, a class of tunable loss functions parameterized by $\alpha \in (0,\infty]$ [9], [10], and with a loss function $\ell_\alpha(\cdot)$ given by

$$\ell_\alpha(y,\hat{y}) := \frac{\alpha}{\alpha-1}\left(1 - y\hat{y}^{\frac{\alpha-1}{\alpha}} - (1-y)(1-\hat{y})^{\frac{\alpha-1}{\alpha}}\right). \tag{5}$$

In [7], we show that the α-GAN formulation allows interpolating between various f-divergence based GANs including the Hellinger GAN [3] ($\alpha = 1/2$), the vanilla GAN [1] ($\alpha = 1$), and the Total Variation GAN [3] ($\alpha = \infty$), as well as IPM based GANs including WGAN [4] (for $\alpha = \infty$ and an appropriately constrained discriminator class). We also show that, for large enough discriminator capacity, the min-max optimization problem for α-GAN in (1) simplifies to

$$\inf_{\theta\in\Theta} D_{f_\alpha}(P_r||P_{G_\theta}) + \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}} - 2\right), \tag{6}$$

where $D_{f_\alpha}(P_r||P_{G_\theta})$ is the Arimoto divergence [11], [12] given by

$$D_{f_\alpha}(P||Q) = \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}} (p(x)^\alpha + q(x)^\alpha)^{\frac{1}{\alpha}} dx - 2^{\frac{1}{\alpha}}\right), \tag{7}$$

that results for the $D_{\omega^*}(x)$ maximizing (4) given by

$$D_{\omega^*}(x) = \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha}. \tag{8}$$

We build on [7] to investigate various aspects of CPE loss-based GANs including α-GAN as summarized below:

- We first establish a two-way correspondence between CPE loss function-based GANs and f-divergences building upon a correspondence between margin-based loss functions and f-divergences [13] (Theorem 1). This not only complements the connection established between variational form of f-divergence in [14] and the f-GAN formulation in [3] but, more crucially, also provides an easier way to implement a variety of f-GANs in practice.

- For sufficiently large number of samples and ample discriminator capacity, we show that Arimoto divergences for all $\alpha > 0$ are *equivalent* in convergence (Theorem 2). This generalizes such an equivalence known [4], [15] only for special cases, i.e., Jensen-Shannon divergence (JSD) for $\alpha = 1$, squared Hellinger distance for $\alpha = 1/2$, and total variation distance (TVD) for $\alpha = \infty$, thus providing a unified perspective on the convergence guarantees of several existing GANs. We present a simpler proof of the equivalence between JSD and TVD [4, Theorem 2(1)].
- When the generator and the discriminator models are neural networks of limited capacity, we present bounds on the estimation error for CPE loss GANs (including α-GAN) by leveraging a contraction lemma on Rademacher complexity [16, Lemma 26.9] (Theorem 3).
- Finally, we highlight the value of tuning α to generate distribution-accurate synthetic data for a toy dataset.

II. Main Results

A. Correspondence: CPE loss GANs and f-divergences

We first establish a precise correspondence between the family of GANs based on CPE loss functions and a family of f-divergences. We do this by building upon a relationship between margin-based loss functions and f-divergences first demonstrated by Nguyen *et al.* [13] and leveraging our CPE loss function perspective of GANs given in (4). This complements the connection established by Nowozin *et al.* [3] between the variational estimation approach of f-divergences [14] and f-divergence based GANs. We call a CPE loss function $\ell(y,\hat{y})$ *symmetic* [8] if $\ell(1,\hat{y}) = \ell(0,1-\hat{y})$ and an f-divergence $D_f(\cdot\|\cdot)$ *symmetric* [17], [18] if $D_f(P\|Q) = D_f(Q\|P)$. We assume GANs with sufficiently large number of samples and ample discriminator capacity.

Theorem 1. *For any symmetric CPE loss function based GAN with a value function in* (4)*, the min-max optimization in* (1) *reduces to minimizing an f-divergence. Conversely, for any GAN designed to minimize a symmetric f-divergence, there exists a (symmetric) CPE loss function based GAN minimizing the same f-divergence.*

Proof sketch. A detailed proof is in the extended version. We present a proof sketch here. Given any symmetric CPE loss function ℓ, we define a margin-based loss function $\tilde{\ell}$ using a bijective link function (satisfying a mild asumption) and show that the inner optimization of the CPE loss based GAN reduces to an f-divergence with

$$f(u) := -\inf_{t\in\mathbb{R}} \left(\tilde{\ell}(-t) + u\tilde{\ell}(t)\right). \tag{9}$$

For the converse statement, given any symmetric f-divergence, we note that there exists a margin-based loss function $\tilde{\ell}$ such that (9) holds [13, Corollary 3 and Theorem 1(b)]. The rest of the argument follows from defining a CPE loss function ℓ from this margin-based loss function $\tilde{\ell}$ via the inverse of the same link function.

We note that this connection in Theorem 1 generalizes a previously given correspondence between α-GAN and the Arimoto divergence [7]. A consequence of Theorem 1 is that it offers an interpretable way to design GANs and connect a desired measure of divergence to a corresponding loss function where the latter is easier to implement in practice. For example, CPE losses inherit the intuitive and compelling interpretation of vanilla GANs that the discriminator should assign higher likelihood values to real samples and lower ones to the generated samples.

B. Convergence Properties of α-GAN

We study *convergence* properties of α-GAN again under the assumption of sufficiently large number of samples and ample discriminator capacity. In [15], Liu *et al.* address a fundamental question in the context of convergence analysis of GANs, in general: For a sequence of generated distributions (P_n), does convergence of a divergence between the generated distribution P_n and a fixed real distribution P to the global minimum lead to some standard notion of distributional convergence of P_n to P? They answer this question in the affirmative provided the sample space $\mathcal{X}$ ($\mathcal{P}(\mathcal{X})$ is the probability simplex of distributions over $\mathcal{X}$) is a compact metric space.

To this end, Liu *et al.* [15] formally define any divergence that results from the inner optimization of a general GAN in (1) as an *adversarial divergence* [15, Definition 1], thus broadly capturing the divergences used by a number of existing GANs, including vanilla GAN [1], f-GAN [3], WGAN [4], and MMD-GAN [19]. Indeed, we remark that the divergence that results from the inner optimization of CPE loss function GAN as given in (4) (*e.g.,* α-GAN [20]) belongs to this family of adversarial divergences. For *strict adversarial divergences* (a subclass of the adversarial divergences where the minimizer of the divergence is uniquely the real distribution) [15, Definition 3], Liu *et al.* [15] show that convergence of divergence to the global minimum implies weak convergence of the generated distribution to the real distribution. Interestingly, this also leads to a structural result on the class of strict adversarial divergences [15, Figure 1 and Corollary 12] based on a notion of *relative strength* between adversarial divergences. For our purposes, we note that D_{f_α}, namely, the Arimoto divergence as given in (7), is a strict adversarial divergence. We briefly summarize the following terminology from Liu *et al.* [15] to present our results on convergence properties of α-GAN.

Definition 1 (Definition 11, [15])**.** *A* strict adversarial divergence τ_1 *is said to be stronger than another strict adversarial divergence τ_2 (or τ_2 is said to be weaker than τ_1) if for any sequence of probability distributions (P_n) and target distribution P (both in $\mathcal{P}(\mathcal{X})$), $\tau_1(P\|P_n) \to 0$ as $n \to \infty$ implies $\tau_2(P\|P_n) \to 0$ as $n \to \infty$. We say τ_1 is equivalent to τ_2 if τ_1 is both stronger and weaker than τ_2.*

Arjovsky *et al.* [4] proved that the Jensen-Shannon divergence (JSD) is equivalent to the total variation distance (TVD). Later, Liu *et al.* showed that the squared Hellinger distance is equivalent to both of these divergences, meaning that all three divergences belong to the same equivalence class (see [15, Figure 1]). Noticing that the squared Hellinger distance, JSD, and TVD correspond to Arimoto divergences $D_{f_\alpha}(\cdot||\cdot)$ 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 2, thereby adding the Arimoto divergences for all other $\alpha \in \mathbb{R}_+$ also to the same equivalence class.

Theorem 2. *The Arimoto divergences for all $\alpha \in (0,\infty]$ are equivalent in the sense of Definition 1. That is, for a sequence of probability distributions $(P_n) \in \mathcal{P}(\mathcal{X})$ and a fixed distribution $P \in \mathcal{P}(\mathcal{X})$, $D_{f_{\alpha_1}}(P_n||P) \to 0$ as $n \to \infty$ if and only if $D_{f_{\alpha_2}}(P_n||P) \to 0$ as $n \to \infty$, for any $\alpha_1 \neq \alpha_2$.*

Remark 1. We note that the proof techniques used in proving Theorem 2 give rise to a conceptually simpler proof of equivalence between JSD ($\alpha = 1$) and TVD ($\alpha = \infty$) proved earlier by Arjovsky *et al.* [4, Theorem 2(1)], where measure-theoretic analysis was used. In particular, this equivalence relies on the fact that TVD upper bounds JSD [2, Theorem 3]. See Appendix B for details.

Proof sketch. A detailed proof is in the extended version. Noticing that $D_{f_\infty}(\cdot||\cdot)$ is equal to TVD, denoted $D_{\text{TV}}(\cdot||\cdot)$ (see [21], [7, Theorem 2]), it suffices to show that $D_{f_\alpha}(\cdot||\cdot)$ is equivalent to $D_{\text{TV}}(\cdot||\cdot)$, for $\alpha > 0$. To show this, we employ an elegant result by Österreicher and Vajda [21, Theorem 2] which gives lower and upper bounds on the Arimoto divergence in terms of TVD as

$$\psi_\alpha(D_{\text{TV}}(P||Q)) \leq D_{f_\alpha}(P||Q) \leq \psi_\alpha(1) D_{\text{TV}}(P||Q), \quad (10)$$

for an appropriately defined function $\psi_\alpha : [0,1] \to \mathbb{R}$. We use the lower and upper bounds in (10) to show that $D_{f_\alpha}(\cdot||\cdot)$ is stronger than $D_{\text{TV}}(\cdot||\cdot)$, and $D_{f_\alpha}(\cdot||\cdot)$ is weaker than $D_{\text{TV}}(\cdot||\cdot)$, respectively.

TS - reminder to talk about how convergence will be different in practice (cite experiments and estimation theory). This will give nice transition to estimation error.

C. Estimation Error Bounds for CPE Loss based GAN

So far, we have assumed sufficiently large number of samples and ample discriminator capacity. However, in practice, we only have limited number of training samples $S_x = \{X_1,\dots,X_n\}$ and $S_z = \{Z_1,\dots,Z_m\}$ from the real and generated distributions P_r and P_Z, respectively, and these distributions are to be estimated from samples during the training [1]. Also, the discriminator and generator classes are typically neural networks; these limitations lead to estimation errors in training GANs [5], [22], [23]. Ji *et al.* [23] obtain bounds on an estimation error that capture the interplay between both the discriminator and generator, whereas the role of the generator is not explicitly captured in the estimation error bounds in [5], [22]. Ji *et al.* [23] quantify the estimation error using the neural net distance [24] defined by

$$d_{\mathcal{F}_{nn}}(P_r, P_{G_\theta}) = \sup_{\omega \in \Omega} \Big(\mathbb{E}_{X \sim P_r}[f_\omega(X)] - \mathbb{E}_{X \sim P_{G_\theta}}[f_\omega(X)] \Big),$$

where the discriminator and generator models $f_\omega(\cdot)$ and $G_\theta(\cdot)$ are neural networks. Leveraging the methodology in [23], we define and quantify the estimation error in training CPE loss function based GANs (specifically, α-GAN), thereby highlighting the effect of the loss on the error. We begin by considering the following minimization for GAN training:

$$\inf_{\theta \in \Theta} d^{\ell}_{\mathcal{F}_{nn}}(\hat{P}_r, \hat{P}_{G_\theta}), \quad (11)$$

where $\hat{P}_r$ and $\hat{P}_{G_\theta}$ are the empirical real and generated distributions estimated from S_x and S_z, respectively, and

$$\begin{aligned} &d^{\ell}_{\mathcal{F}_{nn}}(\hat{P}_r, \hat{P}_{G_\theta}) \\ &\quad = \sup_{\omega \in \Omega} \Big(\mathbb{E}_{X \sim \hat{P}_r} \phi\big(D_\omega(X)\big) + \mathbb{E}_{X \sim \hat{P}_{G_\theta}} \psi\big(D_\omega(X)\big) \Big) \end{aligned} \quad (12)$$

where, for brevity, we henceforth use $\phi(\cdot) := -\ell(1,\cdot)$ and $\psi(\cdot) := -\ell(0,\cdot)$. For $x \in \mathcal{X} := \{x \in \mathbb{R}^d : ||x||_2 \leq B_x\}$ and $z \in \mathcal{Z} := \{z \in \mathbb{R}^p : ||z||_2 \leq B_z\}$, we consider discriminators and generators as neural network models of the form:

$$D_\omega : x \mapsto \sigma\left(\mathbf{w}_k^\intercal r_{k-1}(\mathbf{W}_{d-1} r_{k-2}(\dots r_1(\mathbf{W}_1(x))))\right) \quad (13)$$

$$G_\theta : z \mapsto \mathbf{V}_l s_{l-1}(\mathbf{V}_{l-1} s_{l-2}(\dots s_1(\mathbf{V}_1 z))), \quad (14)$$

where, $\mathbf{w}_k$ is a parameter vector of the output layer; for $i \in [1:k-1]$ and $j \in [1:l]$, W_i and V_j are parameter matrices; $r_i(\cdot)$ and $s_j(\cdot)$ are entry-wise activation functions of layers i and j, i.e., for $\mathbf{a} \in \mathbb{R}^t$, $r_i(\mathbf{a}) = [r_i(a_1),\dots,r_i(a_t)]$ and $s_i(\mathbf{a}) = [s_i(a_1),\dots,s_i(a_t)]$; and $\sigma(\cdot)$ is the sigmoid function given by $\sigma(p) = 1/(1+\mathrm{e}^{-p})$ (note that σ does not appear in the discriminator in [23, Equation (7)] as the discriminator considered in the neural net distance is not a soft classifier). We assume that each $r_i(\cdot)$ and $s_j(\cdot)$ are R_i- and S_j-Lipschitz, respectively. Moreover, as adopted in [23], [25]–[27], we assume that the Frobenius norms of the parameter matrices are bounded, i.e., $||\mathbf{W}_i||_F \leq M_i$, $i \in [1:k-1]$, $||\mathbf{w}_k||_2 \leq M_k$, and $||\mathbf{V}_j||_F \leq N_j$, $j \in [1:l]$. Let $\hat{\theta}^*$ be the optimizer in (11).

We define the estimation error for a CPE loss GAN as

$$d^{\ell}_{\mathcal{F}_{nn}}(P_r, \hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta \in \Theta} d^{\ell}_{\mathcal{F}_{nn}}(P_r, P_{G_\theta}), \quad (15)$$

and present the following upper bound on the error.

Theorem 3. *In the setting described above, additionally assume the following.*

- *The activation functions $r_i(\cdot)$, $i \in [1:k-1]$ and $s_j(\cdot)$, $j \in [1:l-1]$ are positive homogeneous, i.e., $r_i(\lambda p) = \lambda r_i(p)$ and $s_j(\lambda p) = \lambda s_j(p)$, for any $\lambda \geq 0$ and $p \in \mathbb{R}$,*
- *The functions $\phi(\cdot)$ and $\psi(\cdot)$ are L_ϕ- and L_ψ-Lipschitz, respectively.*

Then, with probability at least $1-2\delta$ over the randomness of training samples $\{X_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^m$, we have

$$d^{\ell}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}*}}) - \inf_{\theta\in\Theta} d^{\ell}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \tag{16}$$

$$\leq \frac{4L_\phi B_x U_\omega \sqrt{3d}}{\sqrt{n}} + \frac{4L_\psi U_\omega U_\theta B_z \sqrt{3(k+l-1)}}{\sqrt{m}} \tag{17}$$

$$+ 2U_\omega \sqrt{2\log\frac{1}{\delta}} \left(\frac{L_\phi B_x}{\sqrt{n}} + \frac{L_\psi B_z U_\theta}{\sqrt{m}} \right), \tag{18}$$

where parameters $U_\omega := M_k \prod_{i=1}^{k-1}(M_i R_i)$ and $U_\theta := N_l \prod_{j=1}^{l-1}(N_j S_j)$.

Proof sketch. Our proof involves the following two steps.

- Building upon the proof techniques of Ji *et al.* [23, Theorem 1], we bound the estimation error in terms of Rademacher complexities of *compositional* function classes involving the CPE loss function.
- We then upper bound these Rademacher complexities leveraging a contraction lemma for Lipschitz loss functions [16, Lemma 26.9]. We remark that this considerably differs from the way the bounds on Rademacher complexities in [23, Corollary 1] are obtained because of the explicit role of the loss function in our setting.

Corollary 1. *Consider the setting of Theorem 3 for α-GAN, i.e., $\phi(p) = \psi(1-p) = \frac{\alpha}{\alpha-1}\left(1-p^{\frac{\alpha-1}{\alpha}}\right)$. Then, with probability at least $1-2\delta$ over the randomness of training samples $\{X_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^m$, we have*

$$d^{\ell}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}*}}) - \inf_{\theta\in\Theta} d^{\ell}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \tag{19}$$

$$\leq \frac{4C_{Q_x}(\alpha) B_x U_\omega \sqrt{3d}}{\sqrt{n}} + \frac{4C_{Q_z}(\alpha) U_\omega U_\theta B_z \sqrt{3(k+l-1)}}{\sqrt{m}} \tag{20}$$

$$+ 2U_\omega \sqrt{2\log\frac{1}{\delta}} \left(\frac{C_{Q_x}(\alpha) B_x}{\sqrt{n}} + \frac{C_{Q_z}(\alpha) B_z U_\theta}{\sqrt{m}} \right), \tag{21}$$

where parameters $U_\omega := M_k \prod_{i=1}^{k-1}(M_i R_i)$, $U_\theta := N_l \prod_{j=1}^{l-1}(N_j S_j)$, $Q_x := U_\omega B_x$, $Q_z := U_\omega U_\theta B_z$, and

$$C_h(\alpha) := \begin{cases} \sigma(h)\sigma(-h)^{\frac{\alpha-1}{\alpha}}, & \alpha\in(0,1] \\ \left(\frac{\alpha-1}{2\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} \frac{\alpha}{2\alpha-1}, & \alpha\in[1,\infty). \end{cases} \tag{22}$$

Proof. A detailed proof is in the extended version. We present a proof sketch here.

- The non-triviality of the proof arises from the fact that α-loss is not Lipschitz in general (it is Lipschitz only for $\alpha=\infty$). However, we note that when the input to the loss is from a logistic model in binary classification, Tyler *et al.* show that α-loss is Lipschitz leveraging its Lipschitzianity in an interval.
- Noting that similar to the logistic model, we also have a sigmoid in the outer layer of the discriminator, we generalize the preceding observation by proving that α-loss is Lipschitz when the input is equal to sigmoid function of the neural network model. This is the reason behind the dependence of the Lipcschitz constant on the neural network model parameters (in terms of Q_x and Q_z).

□

III. Proofs

A. Proof of Theorem 1

B. Proof of Theorem 2

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 $D_{\text{TV}}(P_n||P)\to 0$. 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 [21, Theorem 2] proved that for any $\alpha>0$, probability distributions P and Q, we have

$$\psi_\alpha(D_{\text{TV}}(P||Q)) \leq D_{f_\alpha}(P||Q) \leq \psi_\alpha(1) D_{\text{TV}}(P||Q), \tag{23}$$

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

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 (23), it follows that $\psi_\alpha(D_{\text{TV}}(P_n||P)) \leq D_{f_\alpha}(P_n||P)$, for each $n\in\mathbb{N}$. This implies that $\psi_\alpha(D_{\text{TV}}(P_n||P))\to 0$. We show below that ψ_α is invertible and ψ_α^{-1} is continuous. Then it would follow that $\psi_\alpha^{-1}\psi_\alpha(D_{\text{TV}}(P_n||P)) = D_{\text{TV}}(P_n||P) \to \psi_\alpha^{-1}(0) = 0$ proving that Arimoto divergence is stronger than the total variation distance. It remains to show that ψ_α is invertible and ψ_α^{-1} is continuous. Invertibility follows directly from the fact that ψ_α is strictly monotone increasing function. For the continuity of ψ_α^{-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 ψ_α 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 $D_{\text{TV}}(P_n||P)\to 0$. It follows from the upper bound in (23) that $D_{f_\alpha}(P_n||P) \leq D_{\text{TV}}(P_n||P)$, for each $n\in\mathbb{N}$. This implies that $D_{f_\alpha}(P_n||P)\to 0$ which completes the proof.

IV. Experimental Results

In this section, we implement α-GAN on a simple synthetic dataset in both noise-free and noisy settings for various $\alpha\in[0.2,20]$. Our implementation is adapted from a Github repository (github.com/nbertagnolli/pytorch-simple-gan) [cite].

Experimental Setup. Our experiments are performed on a synthetic dataset consisting of 25600 unsigned seven-bit binary representations of uniformly-drawn even integers from 0 to 126. In the noisy setting, we flip the last bit of a certain percentage of the data, making those integers odd. TS - it would also be good practice/ethical to say what computational resources we used (e.g., computing cluster, for 10 hours...)

The GAN architecture consists of: a generator with input and output of length 7 modeled as $G_\theta(z) = \sigma(W_g z + b_g)$, where $\theta = \{W_g, b_g\}$, $W_g \in \mathbb{R}^{7\times 7}$, $b_g \in \mathbb{R}^7$, and $\sigma : \mathbb{R} \to (0,1)$ is the sigmoid function given by $\sigma(t) = (1 + e^{-t})^{-1}$; and a discriminator with input of length 7 and output of length 1 modeled as $D_\omega(x) = \sigma(W_d x + b_d)$, where $\omega = \{W_d, b_d\}$, $W_d \in \mathbb{R}^{1\times 7}$, and $b_d \in \mathbb{R}$. We adopt the same learning rate, optimizer, and noise dimension settings as in []. We use standard normal noise, a batch size of 256 for both the real and generator noise samples, and 1000 training epochs. For a fair comparison, we generate all 25600 noise samples needed for training beforehand so that each α-GAN sees the same noise samples. After each α-GAN is trained, we feed each trained generator the same 20000 standard normal noise samples to evaluate its performance based on the following three metrics: number of modes outputted, number of odd integers outputted, and total variation distance (TVD) between the generated probability distribution and the true distribution (discrete uniform). The number of modes outputted refers to the number of unique even integers between 0 and 126 (maximum of 64) outputted by the generator.

Illustration of Results.

Noise (%)	$\alpha = 1$	$\alpha_* < 1$	$\alpha_* > 1$
0	(59, 0, 0.45)	(0.2, 59, 0, 0.45)	(1.2, 59, 0, 0.45)
5	(59, 0, 0.45)	(0.2, 59, 0, 0.45)	(1.2, 59, 0, 0.45)
10	(59, 0, 0.45)	(0.2,59, 0, 0.45)	(1.2,59, 0, 0.45)
15	(59, 0, 0.45)	(59, 0, 0.45)	(59, 0, 0.45)
20	(59, 0, 0.45)	(59, 0, 0.45)	(59, 0, 0.45)
30	(59, 0, 0.45)	(59, 0, 0.45)	(59, 0, 0.45)

TABLE I
Caption

Appendix A
Details Omitted From CPE Loss Function Perspective of GANs

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

$$\phi(t) + \psi(t) \leq \phi(1/2) + \psi(1/2), \text{ for all } t \in [0,1], \tag{24}$$

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* [8] if $\psi(t) = \phi(1-t)$, for all $t \in [0,1]$. Notice that the value function considered by Arora *et al.* [24] is a special case of (4), i.e., (4) recovers the value function in [24, Equation (2)] when the loss function $\ell(y,\hat{y})$ is symmetric. For symmetric losses, concavity of the function ϕ is a sufficient condition for satisfying (24), but not a necessary condition.

Appendix B
Equivalence of the Jensen-Shannon Divergence and the Total Variation Distance

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 $D_{\text{TV}}(P_n||P) \to 0$. Using the fact that the total variation distance upper bounds the Jensen-Shannon divergence [2, Theorem 3], we have $D_{\text{JS}}(P_n||P) \leq D_{\text{TV}}(P_n||P)$, for each $n \in \mathbb{N}$. This implies that $D_{\text{JS}}(P_n||P) \to 0$ since $D_{\text{TV}}(P_n||P) \to 0$. This greatly simplifies the corresponding proof of [4, 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 [4, Theorem 2(1)] using triangular and Pinsker's inequalities.

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

[3] 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.

[4] 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.

[5] T. Liang, "How well generative adversarial networks learn distributions," *arXiv preprint arXiv:1811.03179*, 2018.

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

[7] G. R. Kurri, T. Sypherd, and L. Sankar, "Realizing gans via a tunable loss function," in *2021 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–6.

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

[9] 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.

[10] 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.

[11] F. Österreicher, "On a class of perimeter-type distances of probability distributions," *Kybernetika*, vol. 32, no. 4, pp. 389–393, 1996.

[12] 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.

[13] 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.

[14] ——, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[15] S. Liu, O. Bousquet, and K. Chaudhuri, "Approximation and convergence properties of generative adversarial learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[17] F. Liese and I. Vajda, *Convex Statistical Distances*, ser. Teubner-Texte zur Mathematik. Teubner, 1987.

[18] I. Sason, “Tight bounds for symmetric divergence measures and a new inequality relating f-divergences,” in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[19] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, “Training generative neural networks via maximum mean discrepancy optimization,” *arXiv preprint arXiv:1505.03906*, 2015.

[20] G. R. Kurri, T. Sypherd, and L. Sankar, “Realizing gans via a tunable loss function,” *arXiv preprint arXiv:2106.05232*, 2021.

[21] 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.

[22] P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He, “On the discrimination-generalization tradeoff in gans,” *arXiv preprint arXiv:1711.02771*, 2017.

[23] K. Ji, Y. Zhou, and Y. Liang, “Understanding estimation and generalization error of generative adversarial networks,” *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 3114–3129, 2021.

[24] 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.

[25] B. Neyshabur, R. Tomioka, and N. Srebro, “Norm-based capacity control in neural networks,” in *Conference on Learning Theory*. PMLR, 2015, pp. 1376–1401.

[26] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” *Advances in neural information processing systems*, vol. 29, pp. 901–909, 2016.

[27] N. Golowich, A. Rakhlin, and O. Shamir, “Size-independent sample complexity of neural networks,” in *Conference On Learning Theory*. PMLR, 2018, pp. 297–299.