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 GANs (Hellinger, vanilla, Total Variation) via the α-loss and corresponds to the minimization of the Arimoto divergence. We show that the Arimoto divergences induced by α-GAN equivalently converge, for all $\alpha \in \mathbb{R}_{>0} \cup \{\infty\}$. However, under restricted learning models and finite samples, we provide estimation bounds which indicate diverse GAN behavior as a function of α. Finally, we present empirical results on a toy dataset that highlight the practical 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

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

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 discriminator $D_{\omega^*}(x)$ maximizing (2) which has the form

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

This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, and CIF-2031799.

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], RényiGAN [7], 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 [8], we introduced a loss function [9] 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

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

(see Appendix A for more details on this). We specialize the setup in (4) to introduce α-GAN using α-loss, a tunable loss function parameterized by $\alpha \in \mathbb{R}_{>0} \cup \{\infty\}$ [10], [11], and with the loss function

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

In [8], 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 (TV) 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 [12], [13] given by

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

This results for the $D_{\omega^*}(x)$ maximizing (4) with

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

We build on [8] 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 [14] (Theorem 1). This not only complements the connection established between the variational form of f-divergence in [15] 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 a sufficiently large number of samples and ample discriminator capacity, we show that Arimoto divergences for all $\alpha \in \mathbb{R}_{>0} \cup \{\infty\}$ are *equivalent* in convergence (Theorem 2). This generalizes such an equivalence known [4], [16] 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 [17, 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

We now present our three main results here.

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 [18] and f-divergences first demonstrated by Nguyen *et al.* [14] 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 [15] and f-divergence based GANs. We call a CPE loss function $\ell(y,\hat{y})$ *symmetric* [9] if $\ell(1,\hat{y}) = \ell(0,1-\hat{y})$ and an f-divergence $D_f(\cdot\|\cdot)$ *symmetric* [19], [20] 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 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 GAN minimizing the same f-divergence.*

Proof sketch. Let ℓ be the symmetric CPE loss of a given CPE loss GAN; note that ℓ has a bivariate input $(y,\hat{y})$ (*e.g.* in (5)), where $y \in \{0,1\}$ and $\hat{y} \in [0,1]$. We define an associated margin-based loss function $\tilde{\ell}$ using a bijective link function (satisfying a mild regularity condition); note that a margin-based loss function has a univariate input $z \in \mathbb{R}$ (*e.g.*, the logistic loss $\tilde{l}^{\log}(z) = \log(1+e^{-z})$) and the bijective link function maps $z \to \hat{y}$ (see [9], [18] for more details). We show after some manipulations that the inner optimization of the CPE loss GAN reduces to an f-divergence with

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

For the converse, given a symmetric f-divergence, using [14, Corollary 3 and Theorem 1(b)], note that there exists a margin-based loss $\tilde{\ell}$ such that (9) holds. The rest of the argument follows from defining a symmetric CPE loss ℓ from this margin-based loss $\tilde{\ell}$ via the *inverse* of the same link function. See Appendix C for the detailed proof.

We note that this connection in Theorem 1 generalizes a previously given correspondence between α-GAN and the Arimoto divergence [8]. 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. Moreover, CPE loss based GANs, including α-GAN, inherit the intuitive and compelling interpretation of vanilla GANs that the discriminator should assign higher likelihood values to real samples and lower ones to generated samples (see Appendix A).

B. Convergence Properties of α-GAN

Building on the above one-to-one correspondence, we now present *convergence* results for a specific CPE loss based GAN, namely α-GAN, thereby providing a unified perspective on the convergence of a variety of f-divergences that arise when optimizing GANs. Here again, we assume a sufficiently large number of samples and ample discriminator capacity. In [16], Liu *et al.* address the following question in the context of convergence analysis of any GAN: 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 $\mathcal{X}$ is a compact metric space.

Liu *et al.* [16] formally define any divergence that results from the inner optimization of a general GAN in (1) as an *adversarial divergence* [16, 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 [21]. Indeed, the divergence that results from the inner optimization of CPE loss function GAN (4) (including α-GAN) is also an adversarial divergence. For *strict adversarial divergences* (a subclass of the adversarial divergences where the minimizer of the divergence is uniquely the real distribution), Liu *et al.* [16] show that convergence of the divergence to its 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 [16, Figure 1 and Corollary 12] based on a notion of *relative strength* between adversarial divergences. We note that the Arimoto divergence D_{f_α} in (7) is a strict adversarial divergence. We briefly summarize the following terminology from Liu *et al.* [16] to present our results on convergence properties of α-GAN. Let $\mathcal{P}(\mathcal{X})$ be the probability simplex of distributions over $\mathcal{X}$.

Definition 1 (Definition 11, [16])**.** *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 [16, 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 $\alpha \in \mathbb{R}_{>0} \cup \{\infty\}$ to the same equivalence class.

Theorem 2. *The Arimoto divergences for all $\alpha \in \mathbb{R}_{>0} \cup \{\infty\}$ 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, our proof of equivalence relies on the fact that TVD upper bounds JSD [2, Theorem 3]. See Appendix B for details.

Proof sketch. Noticing that $D_{f_\infty}(\cdot\|\cdot)$ is equal to TVD, denoted $D_{\mathrm{TV}}(\cdot\|\cdot)$ (see [22], [8, Theorem 2]), it suffices to show that $D_{f_\alpha}(\cdot\|\cdot)$ is equivalent to $D_{\mathrm{TV}}(\cdot\|\cdot)$, for $\alpha > 0$. To show this, we employ an elegant result by Österreicher and Vajda [22, Theorem 2] (with application in statistics) which gives lower and upper bounds on the Arimoto divergence in terms of TVD as

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

for an appropriately defined well-behaved (continuous, invertible, and bounded) function $\gamma_\alpha : [0,1] \to \mathbb{R}$. We use the lower and upper bounds in (10) to show that $D_{f_\alpha}(\cdot\|\cdot)$ is stronger than $D_{\mathrm{TV}}(\cdot\|\cdot)$, and $D_{f_\alpha}(\cdot\|\cdot)$ is weaker than $D_{\mathrm{TV}}(\cdot\|\cdot)$, respectively. Proof details are in Appendix D

Theorems 1 and 2 hold in the ideal setting of sufficient samples and discriminator capacity. In practice, however, GAN training is limited by both the number of training samples as well as the choice of G_θ and D_ω. In fact, recent results by Arora *et al.* [23] show that under such limitations, convergence in divergence does not imply convergence in distribution, and have led to new metrics for evaluating GANs. We now study one such quantity, namely estimation error.

C. Estimation Error Bounds for CPE Loss based GAN

We now consider a setting where we have a limited number of training samples[1] $S_x = \{X_1,\ldots,X_n\}$ and $S_z = \{Z_1,\ldots,Z_m\}$ from P_r and P_Z, respectively. Also, the discriminator and generator classes are typically neural networks; these limitations lead to estimation errors in training GANs [5], [24], [25]. While [25] models the interplay between both the discriminator and generator in the estimation error bounds, those developed in [5], [24] do not explicitly capture the role of the generator. We adopt the approach in [25]; to this end, we begin with the notion of neural net (nn) distance (first introduced in [23]) as defined for the setup in [25]:

$$\begin{aligned} &d_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \\ &= \sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P_r}[f_\omega(X)] - \mathbb{E}_{X\sim P_{G_\theta}}[f_\omega(X)]\Big), \end{aligned} \quad (11)$$

where the discriminator[2] and generator $f_\omega(\cdot)$ and $G_\theta(\cdot)$, respectively, are neural networks. We now introduce a loss-inclusive $d^{(\ell)}_{\mathcal{F}_{nn}}$ for CPE loss GANs (including α-GAN) to highlight the effect of the *loss* on the error. We begin with the following minimization for GAN training:

$$\inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(\hat{P}_r,\hat{P}_{G_\theta}), \quad (12)$$

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

$$\begin{aligned} &d^{(\ell)}_{\mathcal{F}_{nn}}(\hat{P}_r,\hat{P}_{G_\theta}) \\ &= \sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim\hat{P}_r}\phi\big(D_\omega(X)\big) + \mathbb{E}_{X\sim\hat{P}_{G_\theta}}\psi\big(D_\omega(X)\big)\Big), \end{aligned} \quad (13)$$

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:

$$D_\omega : x \mapsto \sigma\big(\mathbf{w}_k^{\mathsf{T}} r_{k-1}(\mathbf{W}_{d-1} r_{k-2}(\ldots r_1(\mathbf{W}_1(x)))\big) \quad (14)$$

$$G_\theta : z \mapsto \mathbf{V}_l s_{l-1}(\mathbf{V}_{l-1} s_{l-2}(\ldots s_1(\mathbf{V}_1 z))), \quad (15)$$

where, $\mathbf{w}_k$ is a parameter vector of the output layer; for $i \in [1:k-1]$ and $j \in [1:l]$, $\mathbf{W}_i$ and $\mathbf{V}_j$ are parameter matrices; $r_i(\cdot)$ and $s_j(\cdot)$ are entry-wise activation functions of layers i and j, i.e., for $\mathbf{a} \in \mathbb{R}^t$, $r_i(\mathbf{a}) = [r_i(a_1),\ldots,r_i(a_t)]$ and $s_i(\mathbf{a}) = [s_i(a_1),\ldots,s_i(a_t)]$; 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 [25, Equation (7)] as the discriminator considered in the neural net distance is not a soft classifier mapping to $[0,1]$). We assume that each $r_i(\cdot)$ and $s_j(\cdot)$ are R_i- and S_j-Lipschitz, respectively, and also that they are positive

[1] In practice, once a model is learned, one can generate any number of noise, and hence, synthetic samples; however, the number of real samples is the (finite sample) bottleneck for the goodness of the learned G_θ model.

[2] In [25], f_ω indicates a discriminator function that takes values in $\mathbb{R}$.

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 $p \in \mathbb{R}$. Finally, as modelled in [25]–[28], 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 $||\mathbf{V}_j||_F \leq N_j$, $j \in [1:l]$.

We define the estimation error for a CPE loss GAN as

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

where $\hat{\theta}^*$ is the minimizer of (12) and present the following upper bound on the error.

Theorem 3. *In the setting described above, additionally assume that 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 $S_x = \{X_i\}_{i=1}^n$ and $S_z = \{Z_j\}_{j=1}^m$, we have*

$$\begin{aligned} &d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \\ &\leq \frac{L_\phi B_x U_\omega \sqrt{3k}}{\sqrt{n}} + \frac{L_\psi U_\omega U_\theta B_z \sqrt{3(k+l-1)}}{\sqrt{m}} \\ &\quad + U_\omega \sqrt{\log\frac{1}{\delta}}\left(\frac{L_\phi B_x}{\sqrt{2n}} + \frac{L_\psi B_z U_\theta}{\sqrt{2m}}\right), \end{aligned} \tag{17}$$

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

In particular, when this bound is specialized to the case of α-GAN by letting $\phi(p) = \psi(1-p) = \frac{\alpha}{\alpha-1}\left(1-p^{\frac{\alpha-1}{\alpha}}\right)$, the resulting bound is nearly identical to the terms in the RHS of (17)*, except for substitutions $L_\phi \leftarrow 4C_{Q_x}(\alpha)$ and $L_\psi \leftarrow 4C_{Q_z}(\alpha)$, where $Q_x := U_\omega B_x$, $Q_z := U_\omega U_\theta B_z$, and*

$$C_h(\alpha) := \begin{cases} \sigma(h)\sigma(-h)^{\frac{\alpha-1}{\alpha}}, & \alpha \in (0,1] \\ \left(\frac{\alpha-1}{2\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} \frac{\alpha}{2\alpha-1}, & \alpha \in [1,\infty). \end{cases} \tag{18}$$

Proof sketch. Our proof involves the following steps:

- Building upon the proof techniques of Ji *et al.* [25, 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 [17, Lemma 26.9]. We remark that this differs considerably from the way the bounds on Rademacher complexities in [25, Corollary 1] are obtained because of the explicit role of the loss function in our setting.
- For the case of α-GAN, we extend a result by Sypherd *et al.* [11] where they showed that α-loss is Lipschitz for a logistic model with (18). 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 a sigmoid function acting on a *neural network* model. This is the reason behind the dependence of the Lipschitz constant on the neural network model parameters (in terms of Q_x and Q_z). Note that (18) is monotonically decreasing in α, indicating the bound saturates. However, one is not able to make definitive statements regarding the estimation bounds for relative values of α because the LHS in (17) is *also* a function of α. Proof details are in Appendix E.

III. Experimental Results

We now present experimental results of α-GAN trained over the set of $\alpha \in [0.2,20]$ for a simple dataset. The real training examples in this dataset consist of 25,600 unsigned seven-bit binary representations of uniformly-drawn *even* integers from 0 to 126; *i.e.,* P_r is the uniform distribution on even integers between 0 and 126. Note that we sometimes refer to even integer(s) as mode(s) (as is common in GAN literature).

We consider two settings: the first is a standard GAN training setup (**Base**) and the second (**Noisy**) differs from the first *only* in introducing noisy real samples. This may resemble a practical scenario where, unbeknownst to the practitioner implementing a GAN, the training data is mislabeled, *e.g.*, when a cat is labeled as a dog. Nevertheless in both settings, the goal of the generator G_θ is to learn the real distribution P_r. Overall, we find that α-GAN exhibits interesting characteristics as a function of α, and there is significant utility in tuning α away from $\alpha = 1$ (vanilla GAN). For both cases, we consider the same architectures for the generator and discriminator as detailed below. Our implementation builds on [29]; full experimental details (and further results) are in Appendix F.

Model and experimental details. The generator, with 7-length input and output, is 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; the discriminator takes a 7-length input and outputs a scalar with $D_\omega(x) = \sigma(W_d x + b_d)$, where $\omega = \{W_d, b_d\}$, $W_d \in \mathbb{R}^{1\times 7}$, and $b_d \in \mathbb{R}$. We use the following hyperparameter settings, which are fixed for all α: learning rate of 0.001, the Adam optimizer [30], standard normal noise *i.e.*, $P_Z = \mathcal{N}(\mathbf{0},\mathbb{I}_7)$, batch size of 256 for both the real and generator noise samples, and 1000 training epochs. After training, we feed each trained α-GAN generator the same set of 20,000 noise samples (also from P_Z) to evaluate its performance[3]. All results are averaged over 10 runs for each α, where the GAN is retrained in each run.

Noisy real data setup. We simulate noisy *real* training examples as follows: for a chosen percentage (%**Noise**) of corrupt samples that are sampled uniformly from the real training examples, we flip the least significant bit (LSB) in the binary representation of the even integer in order to make it *odd*. For every %**Noise** $\in \{10,15,20,30\}$, we train an α-GAN with the corresponding noisy real samples.

Evaluation metrics. We evaluate the performance of the **Base** and **Noisy** cases using the following four metrics: number of output modes, percentage of synthetic outputs that are odd, and both TVD and JSD between the empirical $\hat{P}_{G_\theta}$ and the uniform P_r. The number of output modes refers to the number of unique even integers between 0 and 126 (maximum of 64) output by the generator. We present these

[3]To eliminate additional randomness from test data, we use the same 20k samples, thereby illustrating the performance variations from changing α.

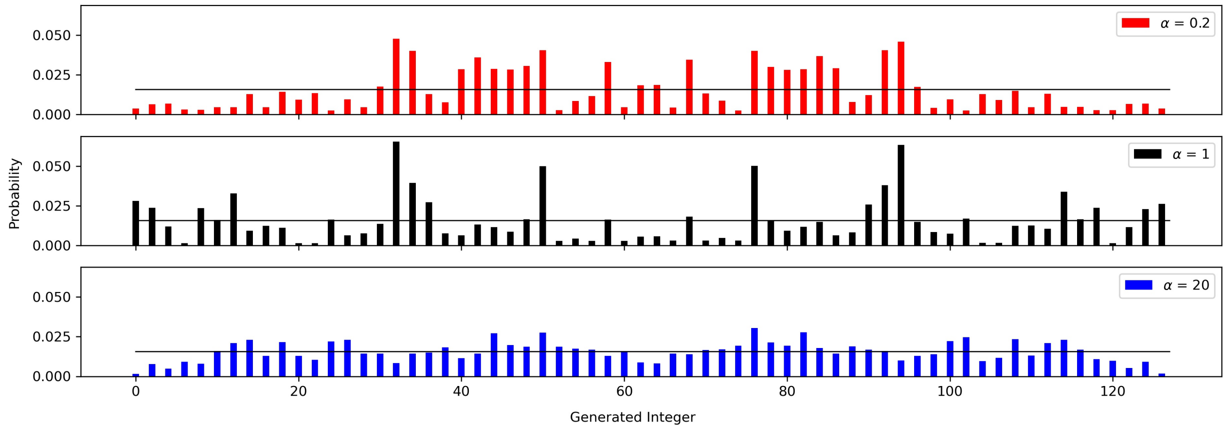

Fig. 1. Histograms of averaged (over 10 runs) $\hat{P}_{G_\theta}$ for $\alpha \in \{0.2, 1, 20\}$ in the **Base** setting. The thick black lines correspond to P_r (uniform over the evens $\in [0, 126]$). No odd integers were output for any α, indicating sufficient training.

	%Noise = 0				%Noise = 10				%Noise = 15				%Noise = 20				%Noise = 30			
α	#Modes	%Odd	TVD	JSD	#Modes	%Odd	TVD	JSD	#Modes	%Odd	TVD	JSD	#Modes	%Odd	TVD	JSD	#Modes	%Odd	TVD	JSD
0.2	61.0	0.0	0.503	0.269	62.8	0.0	0.426	0.203	62.4	0.0	0.489	0.260	63.5	0.0	0.439	0.217	62.0	0.0	0.455	0.238
0.5	60.2	0.0	0.592	0.366	60.5	0.0	0.599	0.364	58.5	0.0	0.610	0.393	58.7	0.0	0.622	0.391	60.3	0.0	0.582	0.353
0.7	60.0	0.0	0.581	0.360	59.0	0.0	0.622	0.403	57.2	0.0	0.632	0.418	57.2	0.0	0.620	0.403	60.0	0.0	0.611	0.377
1	58.1	0.0	0.609	0.383	59.3	0.0	0.630	0.404	60.7	0.0	0.627	0.408	58.0	0.0	0.624	0.410	59.3	0.0	0.609	0.395
4	55.5	0.0	0.613	0.392	55.4	3.4	0.672	0.464	55.9	3.9	0.608	0.412	56.4	4.0	0.599	0.391	55.4	1.7	0.603	0.389
10	62.5	0.0	0.414	0.198	62.1	5.5	0.500	0.276	61.9	3.7	0.492	0.272	62.6	1.4	0.500	0.274	61.7	5.1	0.485	0.268
20	62.4	0.0	0.357	0.154	62.0	8.4	0.348	0.169	63.1	3.8	0.343	0.146	62.8	6.1	0.341	0.154	61.2	6.5	0.369	0.181

TABLE I
RESULTS FOR α-GAN ON TOY DATASET IN **BASE** AND **NOISY** SETTINGS

metrics for $\alpha \in \{0.2, 0.5, 0.7, 1, 4, 10, 20\}$ in Table I. Figure 1 illustrates the (averaged) output probability distributions $\hat{P}_{G_\theta}$ for $\alpha = \{0.2, 1, 20\}$; note that $\alpha = 20$ exhibits the best overall performance, as it yields an averaged distribution closest to P_r. Histograms for the noisy settings can be found in Appendix F.

Interpretation of results in Figure 1 and Table I.

1) Our results suggest that for each α, α-GAN learns a mixture of Gaussians[4] with the mixture approaching the uniform distribution for larger α. The results in Fig. 1 confirm a conjecture raised in [8, Figure 2] that while different choices of α may ideally have equivalent convergence (now proved in Thm. 2), in practice, there will be significant differences in the output distributions for each α arising from how the different gradients for α-GAN affect convergence. Increasing the number of epochs did not change the observed behavior.
2) The result for $\alpha = 20$ in Fig. 1 is perhaps best explained by Sypherd *et al.* [11] where they show that in the standard supervised classification setting, models trained with $\alpha > 1$ approach the average probability of error (in estimating all the modes), thus yielding better overall performance. In the GAN setting, we know from [8] that as $\alpha \to \infty$, α-GAN approaches the TV GAN. Evidently in this scenario, GANs resembling TV GAN far outperform the vanilla GAN.

[4]We conjecture this because the latent noise driving the generator is Gaussian.

3) The results in Table I indicate that larger α perform the best with respect to average TVD and JSD. However, with increasing noise, larger α values also lead to more odd integer outputs; in other words, larger α learn the noisy distribution better. On the other hand, small α are more robust to the odd modes relative to large α while still performing better than $\alpha = 1$ in terms of average TVD and JSD. In summary, several questions yet remain on evaluating the role of α in learning from noisy data.

IV. CONCLUSION

Building on our prior work introducing α-GANs, we have introduced three new results here on the one-to-one correspondence between CPE losses and f-divergences, convergence properties of the Arimoto divergences induced by α-GANs, and the estimation error for CPE loss GANs including α-GAN. Our results on a toy dataset suggest that tuning α can enhance the quality of the synthetic data, in this case, with larger values of α offering more accuracy with respect to the real distribution. More work is needed to better understand the choice of α in limiting mode collapse. Our recent work suggests that tuning $\alpha < 1$ improves classification accuracy for imbalanced datasets [11]; we conjecture this will hold for α-GANs when the real data has an imbalance in samples for different modes. We believe the analysis here can help guide how α-GANs can address these challenges rigorously.

Appendix A
CPE Loss-Based GANs: Additional Observations

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

$$\phi(t) + \psi(t) \leq \phi(1/2) + \psi(1/2), \text{ for all } t \in [0,1], \tag{19}$$

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* [9] if $\psi(t) = \phi(1-t)$, for all $t \in [0,1]$. Notice that the value function considered by Arora *et al.* [23] is a special case of (4), i.e., (4) recovers the value function in [23, 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 (19), 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, i.e., $D_{\text{TV}}(P_n \| P) \to 0$ as $n \to \infty$ implies $D_{\text{JS}}(P_n \| P) \to 0$ as $n \to \infty$. Suppose $D_{\text{TV}}(P_n || P) \to 0$ as $n \to \infty$. Using the fact that the total variation distance upper bounds the Jensen-Shannon divergence [2, Theorem 3], we have $D_{\text{JS}}(P_n||P) \leq (\log_{\text{e}} 2) D_{\text{TV}}(P_n||P)$, for each $n \in \mathbb{N}$. This implies that $D_{\text{JS}}(P_n||P) \to 0$ as $n \to \infty$ since $D_{\text{TV}}(P_n||P) \to 0$ as $n \to \infty$. 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 triangle and Pinsker's inequalities.

Appendix C
Proof of Theorem 1

Consider a symmetric CPE loss $\ell(y,\hat{y})$, i.e., $\ell(1,\hat{y}) = \ell(0,1-\hat{y})$. We may define an associated margin-based loss using a bijective link function $l : \mathbb{R} \to [0,1]$ as

$$\tilde{\ell}(t) := \ell(1,l(t)), \tag{20}$$

where the link l satisfies a mild regularity condition

$$l(-t) = 1 - l(t) \tag{21}$$

(e.g., sigmoid function, $\sigma(t) = 1/(1+\text{e}^{-t})$ satisfies this condition). Consider the inner optimization problem in (2) with the value function in (4) for this CPE loss ℓ.

$$\begin{aligned}
&\sup_\omega \int_\mathcal{X} \left(-p_r(x)\ell(1,D_\omega(x)) - p_{G_\theta}(x)\ell(0,D_\omega(x))\right)\, dx \\
&= \int_\mathcal{X} \sup_{p_x \in [0,1]} \left(-p_r(x)\ell(1,p_x) - p_{G_\theta}(x)\ell(0,p_x)\right)\, dx \qquad (22)\\
&= \int_\mathcal{X} \sup_{p_x \in [0,1]} \left(-p_r(x)\ell(1,p_x) - p_{G_\theta}(x)\ell(1,1-p_x)\right)\, dx \qquad (23)\\
&= \int_\mathcal{X} \sup_{t_x \in \mathbb{R}} (-p_r(x)\ell(1,l(t_x)) - p_{G_\theta}(x)\ell(1,1-l(t_x)))dx \qquad (24)\\
&= \int_\mathcal{X} \sup_{t_x \in \mathbb{R}} (-p_r(x)\ell(1,l(t_x)) - p_{G_\theta}(x)\ell(1,l(-t_x)))\, dx \qquad (25)\\
&= \int_\mathcal{X} \sup_{t_x \in \mathbb{R}} (-p_r(x)\tilde{\ell}(t_x) - p_{G_\theta}(x)\tilde{\ell}(-t_x)\, dx \qquad (26)\\
&= \int_\mathcal{X} p_{G_\theta}(x) \left(-\inf_{t_x \in \mathbb{R}} \left(\tilde{l}(-t_x) + \frac{p_r(x)}{p_{G_\theta}(x)} \tilde{l}(t_x) \right) \right) dx \qquad (27)
\end{aligned}$$

where (23) follows because the CPE loss $\ell(y,\hat{y})$ is symmetric, (25) follows from (21), and (26) follows from the definition of the margin-based loss $\tilde{\ell}$ in (20). Now note that the function f defined as

$$f(u) = -\inf_{t \in \mathbb{R}} \left(\tilde{\ell}(-t) + u\tilde{\ell}(t) \right) \tag{28}$$

is convex since the infimum of affine functions is concave (observed earlier in [14] in a correspondence between margin-based loss functions and f-divergences). So, from (27), we get

$$\begin{aligned}
\sup_\omega \int_\mathcal{X} &\left(-p_r(x)\ell(1,D_\omega(x)) - p_{G_\theta}(x)\ell(0,D_\omega(x))\right)\, dx \\
&= \int_\mathcal{X} p_{G_\theta}(x) f\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)\, dx \qquad (29)\\
&= D_f(P_r \| P_{G_\theta}). \qquad (30)
\end{aligned}$$

Thus, the resulting min-max optimization in (1) reduces to minimizing the f-divergence, $D_f(P_r \| P_{G_\theta})$ with f as given in (28).

For the converse statement, first note that given a symmetric f-divergence, it follows from [14, Theorem 1(b) and Corollary 3] that there exists a margin-based loss function $\tilde{\ell}$ such that f can be expressed in the form (28). We may define an associated symmetric CPE loss $\ell(y,\hat{y})$ with

$$\ell(1,\hat{y}) := \tilde{\ell}(l^{-1}(\hat{y})), \tag{31}$$

where l^{-1} is the inverse of the same link function. Now repeating the steps as in (22) – (27), it is clear that the GAN based on this (symmetric) CPE loss results in minimizing the same symmetric f-divergence.

APPENDIX D
PROOF OF THEOREM 2

Noticing that $D_{f_\infty}(\cdot\|\cdot) = D_{\text{TV}}(\cdot\|\cdot)$ (see [22], [8, Theorem 2]), it suffices to show that $D_{f_\alpha}(\cdot\|\cdot)$ is equivalent to $D_{\text{TV}}(\cdot\|\cdot)$, for $\alpha > 0$, i.e., $D_{f_\alpha}(P_n||P) \to 0$ as $n \to \infty$ if and only if $D_{\text{TV}}(P_n||P) \to 0$ as $n \to \infty$. To this end, we employ a property of the Arimoto divergence D_{f_α} which gives lower and upper bounds on it in terms of the total variation distance, D_{TV}. In particular, Österreicher and Vajda [22, Theorem 2] proved that for any $\alpha > 0$, probability distributions P and Q, we have

$$\gamma_\alpha(D_{\text{TV}}(P||Q)) \le D_{f_\alpha}(P||Q) \le \gamma_\alpha(1) D_{\text{TV}}(P||Q), \quad (32)$$

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

We first prove the 'only if' part, i.e., $D_{f_\alpha}(P_n||P) \to 0$ as $n \to \infty$ implies $D_{\text{TV}}(P_n||P) \to 0$ as $n \to \infty$. Suppose $D_{f_\alpha}(P_n||P) \to 0$. From the lower bound in (32), it follows that $\gamma_\alpha(D_{\text{TV}}(P_n||P)) \le D_{f_\alpha}(P_n||P)$, for each $n \in \mathbb{N}$. This implies that $\gamma_\alpha(D_{\text{TV}}(P_n||P)) \to 0$ as $n \to \infty$. We show below that γ_α is invertible and γ_α^{-1} is continuous. Then it would follow that $\gamma_\alpha^{-1}\gamma_\alpha(D_{\text{TV}}(P_n||P)) = D_{\text{TV}}(P_n||P) \to \gamma_\alpha^{-1}(0) = 0$ as $n \to \infty$ proving that Arimoto divergence is stronger than the total variation distance. It remains to show that γ_α is invertible and γ_α^{-1} is continuous. Invertibility follows directly from the fact that γ_α is strictly monotone increasing function. For the continuity of γ_α^{-1}, it suffices to show that $\gamma_\alpha(C)$ 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 γ_α implies continuity and $\gamma_\alpha(C)$ is compact since a continuous function of a compact set is also compact. By Heine-Borel theorem, this gives that $\gamma_\alpha(C)$ is closed (and bounded) as desired.

We prove the 'if part' now, i.e., $D_{\text{TV}}(P_n||P) \to 0$ as $n \to \infty$ implies $D_{f_\alpha}(P_n||P) \to 0$. It follows from the upper bound in (32) that $D_{f_\alpha}(P_n||P) \le D_{\text{TV}}(P_n||P)$, for each $n \in \mathbb{N}$. This implies that $D_{f_\alpha}(P_n||P) \to 0$ as $n \to \infty$ which completes the proof.

APPENDIX E
PROOF OF THEOREM 3

We upper bound the estimation error in terms of the Rademacher complexities of appropriately defined *compositional* classes building upon the proof techniques of [25, Theorem 1]. We then bound these Rademacher complexities using a contraction lemma [17, Lemma 26.9]. Details are in order.

We first review the notion of Rademacher complexity.

Definition 2 (Rademacher complexity)**.** *Let $\mathcal{G}_\Omega := \{g_\omega : g_\omega$ is a function from $\mathcal{X}$ to $\mathbb{R}$, $\omega \in \Omega\}$ and $S = \{X_1,\dots,X_n\}$ be a set of random samples in $\mathcal{X}$ drawn independent and identically distributed (i.i.d.) from a distribution P_X. Then, the Rademacher complexity of $\mathcal{G}_\Omega$ is defined as*

$$\mathcal{R}_S(\mathcal{G}_\Omega) = \mathbb{E}_{X,\epsilon} \sup_{\omega\in\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g_\omega(x_i)\right| \quad (33)$$

where $\epsilon_1,\dots,\epsilon_n$ are independent random variables uniformly distributed on $\{-1,+1\}$.

We write our discriminator model in (14) in the form

$$D_\omega(x) = \sigma(f_\omega(x)), \quad (34)$$

where f_ω is exactly the same discriminator model defined in [25, Equation (26)]. Now by following the similar steps as in [25, Equations (16)-(18)] by replacing $f_\omega(\cdot)$ in the first and second expectation terms in the definition of $d_{\mathcal{F}_{nn}}(\cdot,\cdot)$ by $\phi(D_\omega(\cdot))$ and $-\psi(D_\omega(\cdot))$, respectively, we get

$$\begin{aligned} &d^{(\ell)}_{\mathcal{F}_{nn}}(P_r, \hat{P}_{G_{\hat\theta^*}}) - \inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(P_r, P_{G_\theta}) \\ &\le 2\sup_{\omega}\left|\mathbb{E}_{X\sim P_r}\phi(D_\omega(X)) - \frac{1}{n}\sum_{i=1}^{n}\phi(D_\omega(X_i))\right| \\ &\quad + 2\sup_{\omega,\theta}\left|\mathbb{E}_{Z\sim P_Z}\psi(D_\omega(g_\theta(Z))) - \frac{1}{m}\sum_{j=1}^{m}\psi(D_\omega(g_\theta(Z_j)))\right| \end{aligned} \quad (35)$$

Let us denote the supremums in the first and second terms in (35) by $F^{(\phi)}(X_1,\dots,X_n)$ and $G^{(\psi)}(Z_1,\dots,Z_m)$, respectively. We next bound $G^{(\psi)}(Z_1,\dots,Z_m)$. Note that $\psi(\sigma(\cdot))$ is $\frac{L_\psi}{4}$-Lipschitz since it is a composition of two Lipschitz functions $\psi(\cdot)$ and $\sigma(\cdot)$ which are L_ψ- and $\frac{1}{4}$-Lipschitz respectively. For any $z_1,\dots,z_j,\dots,z_m,z'_j$, using $\sup_r|h_1(r)| - \sup_r|h_2(r)| \le \sup_r|h_1(r) - h_2(r)|$, we have

$$\begin{aligned} &G^{(\psi)}(z_1,\dots,z_j,\dots,z_m) - G^{(\psi)}(z_1,\dots,z'_j,\dots,z_m) \\ &\le \sup_{\omega,\theta}\frac{1}{m}\left|\psi(D_\omega(g_\theta(z_j))) - \psi(D_\omega(g_\theta(z'_j)))\right| && (36) \\ &\le \sup_{\omega,\theta}\frac{1}{m}\left|\psi(\sigma(f_\omega(g_\theta(z_j)))) - \psi(\sigma(f_\omega(g_\theta(z'_j))))\right| && (37) \\ &\le \frac{L_\psi}{4}\sup_{\omega,\theta}\frac{1}{m}\left|\sigma(f_\omega(g_\theta(z_j))) - \sigma(f_\omega(g_\theta(z'_j)))\right| && (38) \\ &\le \frac{L_\psi}{4}\frac{2}{m}\left(M_k\prod_{i=1}^{k-1}(M_iR_i)\right)\left(N_l\prod_{j=1}^{l-1}(N_jS_j)\right)B_z && (39) \\ &= \frac{L_\psi Q_z}{2m}, && (40) \end{aligned}$$

where (37) follows from (34), (38) follows because $\psi(\sigma(\cdot))$ is $\frac{L_\psi}{4}$-Lipschitz, (39) follows by using the Cauchy-Schwarz inequality and the fact that $||Ax||_2 \le ||A||_F|||x||_2$ (as observed in [25]), and (40) follows by defining

$$Q_z := \left(M_k\prod_{i=1}^{k-1}(M_iR_i)\right)\left(N_l\prod_{j=1}^{l-1}(N_jS_j)\right)B_z. \quad (41)$$

Using (40), the McDiarmid's inequality [17, Lemma 26.4] implies that, with probability at least $1-\delta$,

$$
\begin{aligned}
&G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m) \\
&\leq \mathbb{E}_Z G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m) + \frac{L_\psi Q_z}{2}\sqrt{\log\frac{1}{\delta}/(2m)}. \quad (42)
\end{aligned}
$$

Following the standard steps similar to [25, Equation (20)], the expectation term in (42) can be upper bounded as

$$
\begin{aligned}
&\mathbb{E}_Z G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m) \\
&\quad \leq 2\mathbb{E}_{Z,\epsilon}\sup_{\omega,\theta}\left|\frac{1}{m}\sum_{j=1}^{m}\epsilon_j\psi(D_\omega(g_\theta(Z_j)))\right| \quad (43) \\
&\quad =: 2\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta}) \quad (44)
\end{aligned}
$$

So, we have, with probability at least $1-\delta$,

$$
G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m) \leq 2\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta}) + \sqrt{\log\frac{1}{\delta}}\frac{L_\psi Q_z}{2\sqrt{2m}}. \quad (45)
$$

Using a similar approach, we have, with probability at least $1-\delta$,

$$
F^{(\phi)}(X_1,\ldots,X_n) \leq 2\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega}) + \sqrt{\log\frac{1}{\delta}}\frac{L_\phi Q_x}{2\sqrt{2n}}, \quad (46)
$$

where

$$
\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega}) := \mathbb{E}_{X,\epsilon}\sup_{\omega}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi(D_\omega(X_i))\right|. \quad (47)
$$

Combining (35), (45), and (46) using a union bound, we get, with probability at least $1-2\delta$,

$$
\begin{aligned}
&d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \\
&\qquad \leq 4\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega}) + 4\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta}) \\
&\qquad\quad + \sqrt{\log\frac{1}{\delta}}\left(\frac{L_\phi Q_x}{\sqrt{2n}} + \frac{L_\psi Q_z}{\sqrt{2m}}\right). \quad (48)
\end{aligned}
$$

Now we bound the Rademacher complexities in the RHS of (48). We present the contraction lemma on Rademacher complexity required to obtain these bounds. For $A\subset\mathbb{R}^n$, let $\mathcal{R}(A) := \mathbb{E}_\epsilon\left[\sup_{a\in A}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i a_i\right|\right]$.

Lemma 1 (Lemma 26.9, [17])**.** *For each $i\in\{1,\ldots,n\}$, let $\gamma_i:\mathbb{R}\to\mathbb{R}$ be a ρ-Lipschitz function. Then, for $A\subset\mathbb{R}^n$,*

$$
\mathcal{R}(\gamma\circ A) \leq \rho\mathcal{R}(A), \quad (49)
$$

where $\gamma\circ A := \{(\gamma_1(a_1),\ldots,\gamma_n(a_n)) : a\in A\}$.

Note that $\phi(\sigma(\cdot))$ is $\frac{L_\phi}{4}$-Lipschitz since it is a composition of two Lipschitz functions $\phi(\cdot)$ and $\sigma(\cdot)$ which are L_ϕ- and $\frac{1}{4}$-Lipschitz respectively. Consider

$$
\begin{aligned}
&\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega}) \\
&= \mathbb{E}_X\left[\mathcal{R}(\{(\phi(D_\omega(X_1)),\ldots,\phi(D_\omega(X_n))) : \omega\in\Omega\})\right] \quad (50) \\
&= \mathbb{E}_X\left[\mathcal{R}(\{(\phi(\sigma(f_\omega(X_1))),\ldots,\phi(\sigma(f_\omega(X_n)))) : \omega\in\Omega\})\right] \quad (51) \\
&\leq \frac{L_\phi}{4}\mathbb{E}_X\left[\mathcal{R}(\{(f_\omega(X_1),\ldots,(f_\omega(X_n)) : \omega\in\Omega\})\right] \quad (52) \\
&\leq \frac{L_\phi Q_x\sqrt{3k}}{4\sqrt{n}} \quad (53)
\end{aligned}
$$

where (51) follows from (34), (52) follows from Lemma 1 by substituting $\gamma(\cdot)=\phi(\sigma(\cdot))$, and (53) follows from [25, Proof of Corollary 1]. Using a similar approach, we obtain

$$
\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta}) \leq \frac{L_\psi Q_z\sqrt{3(k+l-1)}}{4\sqrt{m}}. \quad (54)
$$

Substituting (53) and (54) into (48) gives (17).

A. Specialization to α-GAN

Let $\phi_\alpha(p)=\psi_\alpha(1-p)=\frac{\alpha}{\alpha-1}\left(1-p^{\frac{\alpha-1}{\alpha}}\right)$. It is shown in [11, Lemma 7] that $\phi_\alpha(\sigma(\cdot))$ is $C_h(\alpha)$-Lipschitz in $[-h,h]$, for $h>0$, with $C_h(\alpha)$ as given in (18). Now using the Cauchy-Schwarz inequality and the fact that $||Ax||_2 \leq ||A||_F||x||_2$, it follows that

$$
|f_\omega(\cdot)| \leq Q_x, \quad (55)
$$

$$
|f_\omega(g_\theta(\cdot))| \leq Q_z, \quad (56)
$$

where $Q_x := M_k\prod_{i=1}^{k-1}(M_iR_i)B_x$ and with Q_z as in (41). So, we have $f_\omega(\cdot)\in[-Q_x,Q_x]$ and $f_\omega(g_\theta(\cdot))\in[-Q_z,Q_z]$. Thus, we have that $\psi_\alpha(\sigma(\cdot))$ and $\phi_\alpha(\sigma(\cdot))$ are $C_{Q_z}(\alpha)$- and $C_{Q_x}(\alpha)$-Lipschitz, respectively. Now specializing the steps (38) and (52) with these Lipschitz constants, we get the following bound with the substitutions $\frac{L_\phi}{4}\leftarrow C_{Q_x}(\alpha)$ and $\frac{L_\psi}{4}\leftarrow 4C_{Q_z}(\alpha)$ in (17):

$$
\begin{aligned}
&d^{(\ell_\alpha)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d^{(\ell_\alpha)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \\
&\leq \frac{4C_{Q_x}(\alpha)Q_x\sqrt{3k}}{\sqrt{n}} + \frac{4C_{Q_z}(\alpha)Q_z\sqrt{3(k+l-1)}}{\sqrt{m}} \\
&\quad + 2\sqrt{2\log\frac{1}{\delta}}\left(\frac{C_{Q_x}(\alpha)Q_x}{\sqrt{n}} + \frac{C_{Q_z}(\alpha)Q_z}{\sqrt{m}}\right). \quad (57)
\end{aligned}
$$

Appendix F
Further Experimental Details

The GAN architecture is as follows: the generator, with 7-length input and output, is modeled as $G_\theta(z)=\sigma(W_gz+b_g)$, where $\theta=\{W_g,b_g\}$, $W_g\in\mathbb{R}^{7\times7}$, $b_g\in\mathbb{R}^7$, and $\sigma:\mathbb{R}\to(0,1)$ is the sigmoid function given by $\sigma(t)=(1+e^{-t})^{-1}$; the discriminator takes a 7-length input and outputs a scalar with $D_\omega(x)=\sigma(W_dx+b_d)$, where $\omega=\{W_d,b_d\}$, $W_d\in\mathbb{R}^{1\times7}$, and $b_d\in\mathbb{R}$.

In order to convert the generator's output $G_\theta(z)\in(0,1)^7$ into the corresponding binary representation $b\in\{0,1\}^7$ when evaluating the performance of the trained generator, we use a threshold, i.e. for $i \in \{1,2,\ldots,7\}$, $b_i = 1$ if $G_\theta(z)_i \geq 0.5$ and $b_i = 0$ otherwise.

Training was done on a computing cluster using NVIDIA V100 GPUs.

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] 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.

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

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

[10] 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.

[11] 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*, 2021.

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

[13] 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.

[14] 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.

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

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

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

[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, *Convex Statistical Distances*, ser. Teubner-Texte zur Mathematik. Teubner, 1987.

[20] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences," in *IEEE Information Theory Workshop*. IEEE, 2015, pp. 1–5.

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

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

[29] N. Bertagnolli, "Building a super simple GAN in pytorch," 2020, github.com/nbertagnolli/pytorch-simple-gan.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.