Single and Dual-objective GANs via Classification Loss Functions: Addressing GAN Training Instabilities

Monica Welfert, Gowtham R. Kurri, Kyle Otstot, Lalitha Sankar

Abstract

Generative adversarial networks (GANs) are generative models that are trained to produce samples from an unknown (real) distribution. They consist of a generator (G) and a discriminator (D) that play an adversarial min-max game with each other. The generator's goal is to produce realistic synthetic data that can fool the discriminator, while the discriminator tries to distinguish between real and synthetic data. The choice of value functions optimized by G and D can lead to different types of GANs, with the most common ones falling in the broad class called f-GANs wherein the game simplifies to minimizing an f-divergence. Noting that the discriminator is a classifier, we introduce a loss function perspective of GANs. In particular, we reformulate the GAN value function using binary classification loss functions, oft-referred to as classification probability estimation (CPE) losses. For sufficiently large number of samples and capacities for G and D, we show that the resulting zero-sum game simplifies to minimizing an f-divergence under appropriate conditions on the loss function. We also show that our loss function approach is equivalent to the f-GAN formulation, under certain regularity conditions on f. In the finite sample and capacity setting, we define estimation and generalization errors to quantify the gap in the generator's performance relative to the optimal setting with infinite samples and obtain bounds on these errors. We specialize these results to α-GANs, defined using a particular tunable classification loss, α-loss, parameterized by $\alpha \in [0,\infty)$. Observing the desirable gradient behaviors demonstrated by α-loss, as well as the need to have different objectives for G and D, we introduce a class of dual-objective GANs in an effort to address the training instabilities of GANs. In particular, we model each objective using α-loss to obtain (α_D,α_G)-GANs, parameterized by $(\alpha_D,\alpha_G) \in [0,\infty)^2$. For sufficiently large number of samples and capacities for G and D, we show that the resulting non-zero sum game simplifies to minimizing an f-divergence under appropriate conditions on (α_D,α_G). We also introduce a more general dual-objective CPE loss formulation and define and obtain upper bounds on the estimation error for it in the finite sample and capacity setting. Finally, we highlight the value of tuning (α_D,α_G) in alleviating training instabilities for the synthetic 2D Gaussian mixture ring as well as the large publicly available Celeb-A and LSUN Classroom image datasets.

Index Terms

generative adversarial networks, CPE loss formulation, estimation error, training instabilities, dual objectives.

I. Introduction

GENERATIVE adversarial networks (GANs) have become a crucial data-driven tool for generating synthetic data. GANs are generative models trained to produce samples from an unknown (real) distribution using a finite number of training data samples. They consist 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 each other. The generator G_θ maps noise $Z \sim P_Z$ to a data sample in $\mathcal{X}$ via the mapping $z \mapsto G_\theta(z)$ and aims to mimic data from the real distribution P_r. The discriminator D_ω takes as input $x \in \mathcal{X}$ and classifies it as real or generated by computing a score $D_\omega(x) \in [0,1]$ which reflects the probability that x comes from P_r (real) as opposed to P_{G_θ} (synthetic). For a chosen value function $V(\theta,\omega)$, the adversarial game between G and D can be formulated as a zero-sum min-max problem given by

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

Goodfellow *et al.* [1] introduce the vanilla GAN for which

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

For this V_{VG}, they show that when the discriminator class $\{D_\omega\}_{\omega\in\Omega}$ is rich enough, (4) simplifies to minimizing the Jensen-Shannon divergence [2] between P_r and P_{G_θ}.

Various other GANs have been studied in the literature using different value functions, including f-divergence based GANs called f-GANs [3], IPM based GANs [4]–[6], etc. Observing that the discriminator is a classifier, recently, Kurri *et al.* [7],

This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, CIF-2031799, and CIF-1934766.
Manuscript received April 19, 2021; revised August 16, 2021.

[8] show that the value function in (4) can be written using a class probability estimation (CPE) loss $\ell(y,\hat{y})$ whose inputs are the true label $y\in\{0,1\}$ and predictor $\hat{y}\in[0,1]$ (soft prediction of y) as

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

Using this approach, they introduce α-GAN using the tunable CPE loss α-loss [9], [10], defined for $\alpha\in(0,\infty]$ as

$$\ell_\alpha(y,\hat{y}):=\frac{\alpha}{\alpha-1}\left(1-y\hat{y}^{\frac{\alpha-1}{\alpha}}-(1-y)(1-\hat{y})^{\frac{\alpha-1}{\alpha}}\right). \tag{2}$$

They show that the α-GAN formulation recovers 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] ($\alpha=\infty$). Further, for a large enough discriminator class, the min-max optimization for α-GAN in (4) simplifies to minimizing the Arimoto divergence [11], [12].

While each of the abovementioned GANs have distinct advantages, they continue to suffer from one or more types of training instabilities, including vanishing/exploding gradients, mode collapse, and sensitivity to hyperparameter tuning. In [1], Goodfellow *et al.* note that the generator's objective in the vanilla GAN can *saturate* early in training (due to the use of the sigmoid activation) when D can easily distinguish between the real and synthetic samples, i.e., when the output of D is near zero for all synthetic samples, leading to vanishing gradients. Further, a confident D induces a steep gradient at samples close to the real data, thereby preventing G from learning such samples due to exploding gradients. To alleviate these, [1] proposes a *non-saturating* (NS) generator objective:

$$V_{\text{VG}}^{\text{NS}}(\theta,\omega)=\mathbb{E}_{X\sim P_{G_\theta}}[-\log D_\omega(X)]. \tag{3}$$

This NS version of the vanilla GAN may be viewed as involving different objective functions for the two players (in fact, with two versions of the $\alpha=1$ CPE loss, i.e., log-loss, for D and G). However, it continues to suffer from mode collapse [13], [14]. While other dual-objective GANs have also been proposed (e.g., Least Squares GAN (LSGAN) [15], RényiGAN [16], NS f-GAN [3], hybrid f-GAN [17]), few have had success fully addressing training instabilities.

Recent results have shown that α-loss demonstrates desirable gradient behaviors for different α values [10]. It also assures learning robust classifiers that can reduce the confidence of D (a classifier) thereby allowing G to learn without gradient issues. To this end, we introduce a different α-loss objective for each player to address training instabilities. We propose a tunable dual-objective (α_D,α_G)-GAN, where the objective functions of D and G are written in terms of α-loss with parameters $\alpha_D\in(0,\infty]$ and $\alpha_G\in(0,\infty]$, respectively.

A. Our Contributions

Our key contributions are:

- For this non-zero sum game, we show that a Nash equilibrium exists. For appropriate (α_D,α_G) values, we derive the optimal strategies for D and G and prove that for the optimal D_{ω^*}, G minimizes an f-divergence and can therefore learn the real distribution P_r.
- Since α-GAN captures various GANs, including the vanilla GAN, it can potentially suffer from vanishing gradients due to a saturation effect. We address this by introducing a non-saturating version of the (α_D,α_G)-GAN and present its Nash equilibrium strategies for D and G.
- A natural question that arises is how to quantify the theoretical guarantees for dual-objective GANs, specifically for (α_D,α_G)-GANs, in terms of their estimation capabilities in the setting of limited capacity models and finite training samples. To this end, we define estimation error for (α_D,α_G)-GANs, present an upper bound on the error, and a matching lower bound under additional assumptions.
- Finally, we demonstrate empirically that tuning α_D and α_G significantly reduces vanishing and exploding gradients and alleviates mode collapse on a synthetic 2D-ring dataset. For the high-dimensional Stacked MNIST dataset, we show that our tunable approach is more robust in terms of mode coverage to the choice of GAN hyperparameters, including number of training epochs and learning rate, relative to both vanilla GAN and LSGAN.

B. Related Work

LS - FINISH THIS OR DO WE NEED THIS?

Raginsky, Ayfer, broad papers, generalization, ...

C. Outline

LS - does not need a separate subsection

D. Notation

II. Preliminaries: Overview of GANs and loss functions for classification

A. Background on GANs

We begin by presenting an overview of GANs in the literature. 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

$$\inf_{\theta\in\Theta}\sup_{\omega\in\Omega}V(\theta,\omega). \tag{4}$$

Goodfellow *et al.* [1] introduced the vanilla GAN using

$$\begin{aligned}V_{\text{VG}}(\theta,\omega)&=\mathbb{E}_{X\sim P_r}[\log D_\omega(X)]+\mathbb{E}_{Z\sim P_Z}[\log(1-D_\omega(G_\theta(Z)))]\\&=\mathbb{E}_{X\sim P_r}[\log D_\omega(X)]+\mathbb{E}_{X\sim P_{G_\theta}}[\log(1-D_\omega(X))],\end{aligned} \tag{5}$$

for which they showed that when the discriminator class $\{D_\omega\}$, parametrized by ω, is rich enough, (4) simplifies to finding $\inf_{\theta\in\Theta}2D_{\text{JS}}(P_r||P_{G_\theta})-\log 4$, 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 choosing the optimal discriminator

$$D_{\omega^*}(x)=\frac{p_r(x)}{p_r(x)+p_{G_\theta}(x)}, \tag{6}$$

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 by leveraging the variational characterization of f-divergences [18], Nowozin *et al.* [3] introduced f-GANs via the value function

$$V_f(\theta,\omega)=\mathbb{E}_{X\sim P_r}[D_\omega(X)]+\mathbb{E}_{X\sim P_{G_\theta}}[f^*(D_\omega(X))], \tag{7}$$

where[1] $D_\omega:\mathcal{X}\to\mathbb{R}$ and $f^*(t):=\sup_u\{ut-f(u)\}$ is the Fenchel conjugate of a convex lower semicontinuous function f defining an f-divergence $D_f(P_r||P_{G_\theta}):=\int_{\mathcal{X}}p_{G_\theta}(x)f\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx$ [19]–[21]. 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 $D_{\omega^*}(x)=f'\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)$. In order to respect the domain $\text{dom}(f^*)$ of the conjugate f^*, Nowozin *et al.* further decomposed (7) by assuming the discriminator D_ω can be represented in the form $D_\omega(x)=g_f(Q_\omega(x))$, yielding the value function

$$\widetilde{V}_f(\theta,\omega)=\mathbb{E}_{X\sim P_r}[g_f(Q_\omega(x))]+\mathbb{E}_{X\sim P_{G_\theta}}[f^*(g_f(Q_\omega(x)))], \tag{8}$$

where $Q_\omega:\mathcal{X}\to\mathbb{R}$ and $g_f:\mathbb{R}\to\text{dom}(f^*)$ is an output activation function specific to the f-divergence used.

Highlighting the problems with the continuity of various f-divergences (e.g., Jensen-Shannon, KL, reverse KL, total variation) over the parameter space Θ [13], Arjovsky *et al.* [4] proposed Wasserstein-GAN (WGAN) using the following Earth Mover's (also called Wasserstein-1) distance:

$$W(P_r,P_{G_\theta})=\inf_{\Gamma_{X_1X_2}\in\Pi(P_r,P_{G_\theta})}\mathbb{E}_{(X_1,X_2)\sim\Gamma_{X_1X_2}}\|X_1-X_2\|_2, \tag{9}$$

where $\Pi(P_r,P_{G_\theta})$ is the set of all joint distributions $\Gamma_{X_1X_2}$ with marginals P_r and P_{G_θ}. WGAN employs the Kantorovich-Rubinstein duality [22] using the value function

$$V_{\text{WGAN}}(\theta,\omega)=\mathbb{E}_{X\sim P_r}[D_\omega(X)]-\mathbb{E}_{X\sim P_{G_\theta}}[D_\omega(X)], \tag{10}$$

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 addresing GAN training instabilities unresolved [14].

[1]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.

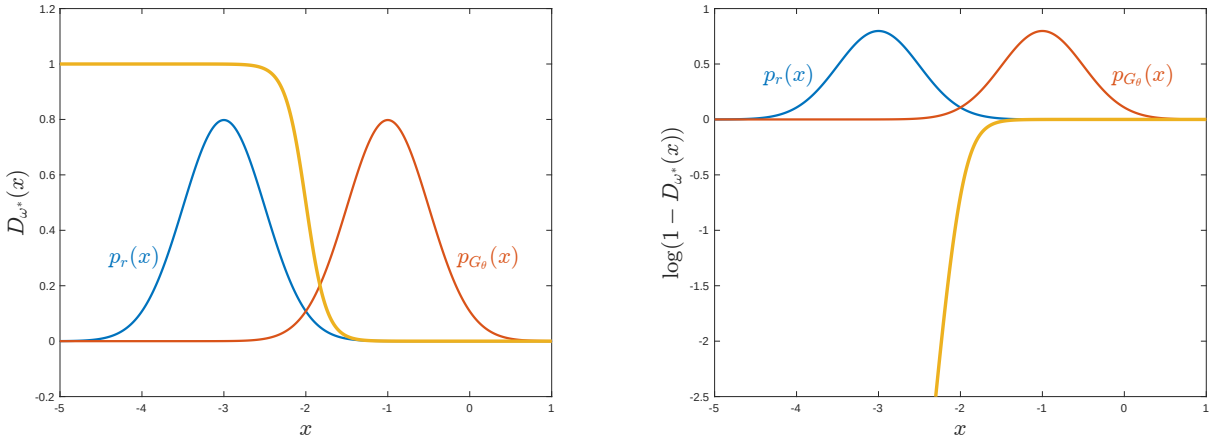


Fig. 1. A toy example of the vanilla GAN, where the real distribution $P_r = \mathcal{N}(-2, 0.5^2)$ (blue curve) and the assumed initial generated distribution $P_{G_\theta} = \mathcal{N}(2, 0.5^2)$ (orange curve). (a) A plot of the optimal discriminator output $D_{\omega^*}(x)$ in (6). (b) A plot of the generator's saturating loss $\log(1-D_{\omega^*}(x))$.

B. Background on Loss Functions for Classification

The ideal loss function for classification is the Bayes loss, also known as the 0-1 loss. However, the complexity of implementing such a non-convex loss has led to much interest in seeking surrogate loss functions for classification. Several surrogate losses with desirable properties have been proposed to train classifiers; the most oft-used and popular among them is log-loss, also referred to as cross-entropy loss. However, enhancing robustness of classifier has broadened the search for better surrogate losses or families of losses; one such family is the class probability estimator (CPE) losses that operate on a soft probability or risk estimate. Recently, it has been shown that a large class of known CPE losses can be captured by a tunable loss family called α-loss, which includes the well-studied exponential loss ($\alpha = 1/2$), log-loss ($\alpha = 1$), and soft 0-1 loss, i.e., the probability of error ($\alpha = \infty$). Formally, α-loss is defined as follows.

Definition 1 (Sypherd *et al.* [23])**.** *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*

$$\ell_\alpha(y, \hat{P}) \triangleq \frac{\alpha}{\alpha - 1} \left(1 - \hat{P}(y)^{\frac{\alpha-1}{\alpha}}\right). \tag{11}$$

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 [23]. Also, α-loss is convex in the probability term $\hat{P}(y)$. Regarding the history of (11), Arimoto first studied α-loss in finite-parameter estimation problems [24], and later Liao *et al.* independently introduced and used α-loss to model the inferential capacity of an adversary to obtain private attributes [25]. Most recently, Sypherd *et al.* studied α-loss in the classification setting [23], which is an impetus for this work.

C. Background on GAN Training Instabilities

In [1], Goodfellow *et al.* note that the generator's objective in the vanilla GAN can *saturate* early in training when the discriminator can easily distinguish between the real and synthetic samples, i.e., when the output of the discriminator is near zero for all synthetic samples, leading to vanishing gradients (see Fig. 1). Further, a confident discriminator induces a steep gradient at samples close to the real data, thereby preventing the generator from learning such samples due to exploding gradients (see again Fig. 1). To alleviate these, [1] propose a *non-saturating* (NS) generator objective:

$$V_{\text{VG}}^{\text{NS}}(\theta, \omega) = \mathbb{E}_{X \sim P_{G_\theta}}[-\log D_\omega(X)]. \tag{12}$$

This NS version of the vanilla GAN may be viewed as involving different objective functions for the two players. However, it continues to suffer from mode collapse (when the generator produces only a subset of the *modes* in the real data, i.e., when the generated data lacks diversity) [13], [14] due to failure to converge and sensitivity to hyperparameter initialization (e.g. learning rate) because of large gradients (see Fig. 2).

During vanilla GAN training, imbalanced performances between the generator and discriminator often coincide with the presence of exploding and vanishing gradients. When updating the generator weights during the backward pass of the network $D_\omega \circ G_\theta$, the gradients are computed by propagating the error signal from the output layer of D_ω to the input layer of G_θ, following the chain rule of derivatives. Each layer contributes to the gradient update by multiplying the incoming gradient with the local gradient of its activation function, and passing it to the preceding layer.

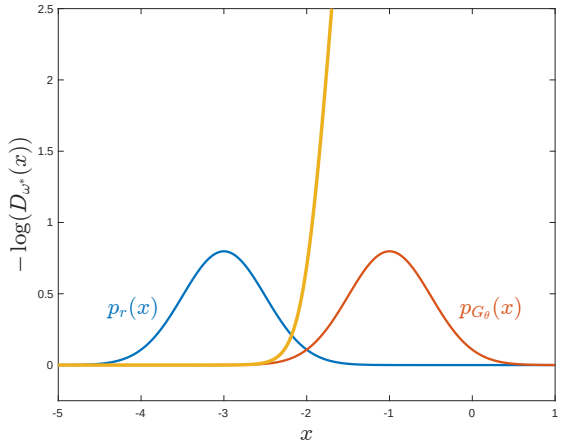


Fig. 2. A plot of the vanilla GAN generator's non-saturating loss $-\log(D_{\omega*}(x))$ for the same toy example as in Figure 1.

When the gradients become large, the successive multiplication of these gradients across the layers can result in an exponential growth, known as gradient explosion. Conversely, small gradients can lead to an exponential decay, referred to as gradient vanishing. In both cases, networks with multiple hidden layers are particularly susceptible to unstable weight updates, causing extremely large or small values that may overflow or underflow the numerical range of computations, respectively.

In the context of vanilla GANs, gradient explosion can occur when the generator successfully produces samples that are severely misclassified (close to 1) by the discriminator. During training, the generator is updated using the loss function $\log(1-D_\omega(x))$, which diverges to $-\infty$ as the discriminator output $D_\omega(x)$ approaches 1. Consequently, the gradients for the generator weights fail to converge to non-zero values, leading to the generated data potentially overshooting the real data in any direction. In severe cases of gradient explosion, the weight update can push the generated data towards a region far from the real data. As a result, the discriminator can easily assign zero probabilities to the generated data and ones to the real data. As the discriminator output approaches zero, the generator's loss function converges to zero, causing the gradients of the generator weights to vanish gradually. As shown in Figure 3(a), this phenomenon can prevent the generator from effectively correcting itself and improving its performance over time.

Addressing the issues of exploding and vanishing gradients, [1] proposed a solution in the form of a *non-saturating* (NS) generator objective:

$$V_{\text{VG}}^{\text{NS}}(\theta,\omega) = \mathbb{E}_{X\sim P_{G_\theta}}[-\log D_\omega(X)]. \tag{13}$$

The use of this non-saturating loss function allows the generated data to converge towards the real distribution. As the discriminator output approaches 1, the generator loss approaches zero, indicating that the generated data becomes more aligned with the real distribution. Additionally, with a high-performing discriminator, the generator receives steep gradients (as opposed to vanishing gradients) during the update process; this occurs because the generator loss diverges to $+\infty$ as the discriminator output approaches zero.

Although the non-saturating vanilla GAN (an industry standard) incorporates different objective functions for the generator and discriminator, it still suffers from mode collapse and oscillations [13], [14] which is discussed in the next section. These issues arise due to convergence problems and the sensitivity of the GAN to hyperparameter initialization, resulting from the presence of large gradients. Various alternative dual-objective GANs have been proposed, such as the Least Squares GAN (LSGAN) [15], RényiGAN [16], non-saturating f-GAN [3], and hybrid f-GAN [17]. However, these approaches have rarely been successful in fully addressing GAN training instabilities.

During GAN training, the primary goal of the generator is to produce high-quality samples that encompass the full range of diversity found in the real distribution. However, there is a potential drawback known as *mode collapse*, which occurs when the generator produces samples that closely resemble only a limited subset of the real data. In such cases, the generator lacks the incentive to capture the remaining modes since the discriminator struggles to effectively differentiate between the real and generated samples. One possible explanation for this phenomenon, as depicted in Figure 3(b), is that the generator and/or discriminator become trapped in a local minimum, impeding the necessary adjustments to mitigate mode collapse. In the figure, the cluster of generated data approaches a single mode in the real distribution, which forces the discriminator to adjust properly; if the discriminator landscape is sufficiently flat in the mode neighborhood, then the generator may struggle to adapt.

In the case of mode collapse, the generator or discriminator may converge prematurely, leading to a suboptimal standstill between the models. On the other hand, a generator training with the non-saturating value function $V_{\text{VG}}^{\text{NS}}$ may experience a complete failure to converge due to influence from outlier generated data. Illustrated in Figure 3(c), most of the generated

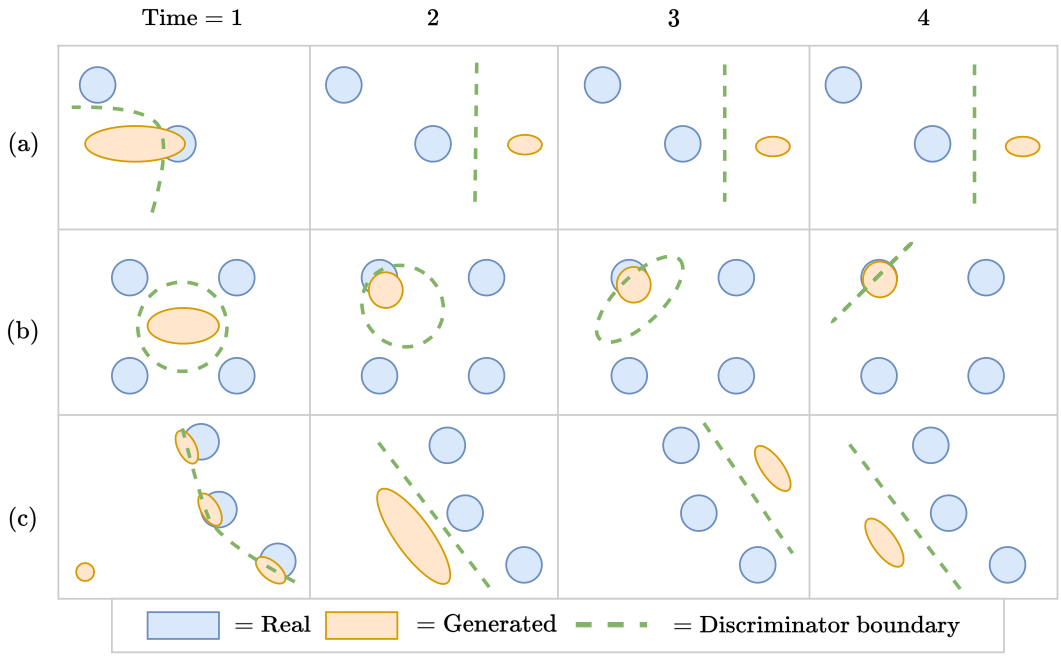


Fig. 3. An illustration of three common GAN failures– (a) exploding and vanishing gradients, (b) mode collapse, and (c) model oscillation– over 4 points in training time.

data occupies the real modes and could receive small gradients $\partial \ell_{\text{BCE}}(1, D_\omega(x))/\partial x$ if the discriminator landscape is locally flat. However, some outlier data is situated very far from the real distribution and consequently receive steep gradients. To address the high non-saturating loss from the outliers, the generator prioritizes directing the outlier data toward the real data over keeping the close data in place; as a result, the generator update reflects a compromise in Time$=2$ of Figure 3(c), where the outliers are resolved at the expense of nudging the other data away from the modes.

Although the generator succeeds at bringing down the average loss by eliminating these outliers, the discriminator is now able to confidently distinguish between the distributions, leading to near-zero probabilities assigned to the generated data. In turn, the generated samples all receive steep gradients which may result in oscillations around the real data. Gradually, the models lose their accumulated knowledge on the structure of the real distribution and essentially restart the training process.

III. Loss Function Perspective on GANs

Noting that a GAN involves a classifier (i.e., discriminator), it is well known that the value function $V_{\text{VG}}(\theta,\omega)$ in (5) considered by Goodfellow *et al.* [1] is related to binary cross-entropy loss. We first formalize this loss function perspective of GANs. In [26], Arora *et al.* observed that the log function in (5) can be replaced by any (monotonically increasing) concave function $\phi(x)$ (e.g., $\phi(x)=x$ for WGANs). In the context of using classification-based losses, we show that one can write $V(\theta,\omega)$ in terms of *any* class probability estimation (CPE) loss $\ell(y,\hat{y})$ whose inputs are the true label $y\in\{0,1\}$ and predictor $\hat{y}\in[0,1]$ (soft prediction of y). For a GAN, we have $(X|y=1)\sim P_r$, $(X|y=0)\sim P_{G_\theta}$, and $\hat{y}=D_\omega(x)$. With this, we define a value function

$$V(\theta,\omega)=\mathbb{E}_{X|y=1}[-\ell(y,D_\omega(X))]+\mathbb{E}_{X|y=0}[-\ell(y,D_\omega(X))] \tag{14}$$

$$=\mathbb{E}_{X\sim P_r}[-\ell(1,D_\omega(X))]+\mathbb{E}_{X\sim P_{G_\theta}}[-\ell(0,D_\omega(X))]. \tag{15}$$

For binary cross-entropy loss, i.e., $\ell_{\text{CE}}(y,\hat{y})\triangleq -y\log\hat{y}-(1-y)\log(1-\hat{y})$, notice that the expression in (15) is equal to V_{VG} in (5). For the value function in (15), we consider a GAN given by the min-max optimization problem:

$$\inf_{\theta\in\Theta}\sup_{\omega\in\Omega} V(\theta,\omega). \tag{16}$$

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

$$\phi(t)+\psi(t)\le\phi\left(\frac{1}{2}\right)+\psi\left(\frac{1}{2}\right),\ \text{for all } t\in[0,1], \tag{17}$$

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* [27] if $\psi(t) = \phi(1-t)$, for all $t \in [0,1]$. Notice that the value function considered by Arora *et al.* [26] is a special case of (15), i.e., (15) recovers the value function in [26, 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 (17), but not a necessary condition.

A. CPE loss GANs and f-divergences

We now 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 [28] and f-divergences first demonstrated by Nguyen *et al.* [29] and leveraging our CPE loss function perspective of GANs given in (15). This complements the connection established by Nowozin *et al.* [3] between the variational estimation approach of f-divergences [18] and f-divergence based GANs. We call a CPE loss function $\ell(y,\hat{y})$ *symmetric* [27] if $\ell(1,\hat{y}) = \ell(0,1-\hat{y})$ and an f-divergence $D_f(\cdot\|\cdot)$ *symmetric* [30], [31] 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* (15)*, the min-max optimization in* (4) *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 (2)), 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 [27], [28] for more details). We show after some manipulations that the inner optimization of the CPE loss GAN reduces to an f-divergence with

$$f(u) := -\inf_{t \in \mathbb{R}} \left(\tilde{\ell}(-t) + u\tilde{\ell}(t) \right). \tag{18}$$

For the converse, given a symmetric f-divergence, using [29, Corollary 3 and Theorem 1(b)], note that there exists a margin-based loss $\tilde{\ell}$ such that (18) 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 A for the detailed proof.

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 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.

We now specialize the loss function perspective of GANs to the GAN obtained by plugging in α-loss. We first write α-loss in (11) in the form of a binary classification loss to obtain

$$\ell_\alpha(y,\hat{y}) := \frac{\alpha}{\alpha-1} \left(1 - y\hat{y}^{\frac{\alpha-1}{\alpha}} - (1-y)(1-\hat{y})^{\frac{\alpha-1}{\alpha}} \right), \tag{19}$$

for $\alpha \in (0,1) \cup (1,\infty)$. Note that (19) recovers ℓ_{CE} as $\alpha \to 1$. Now consider a *tunable* α*-GAN* with a value function

$$\begin{aligned} V_\alpha(\theta,\omega) &= \mathbb{E}_{X \sim P_r}[-\ell_\alpha(1, D_\omega(X))] + \mathbb{E}_{X \sim P_{G_\theta}}[-\ell_\alpha(0, D_\omega(X))] \\ &= \frac{\alpha}{\alpha-1} \left(\mathbb{E}_{X \sim P_r}\left[D_\omega(X)^{\frac{\alpha-1}{\alpha}} \right] + \mathbb{E}_{X \sim P_{G_\theta}}\left[(1 - D_\omega(X))^{\frac{\alpha-1}{\alpha}} \right] - 2 \right). \end{aligned} \tag{20}$$

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

$$\lim_{\alpha \to \infty} V_\alpha(\theta,\omega) = \mathbb{E}_{X \sim P_r}[D_\omega(x)] - \mathbb{E}_{X \sim P_{G_\theta}}[D_\omega(x)] - 1 \tag{21}$$

is the value function (modulo a constant) used in Integral Probability Metric (IPM) based GANs[2], e.g., WGAN, McGan [32], Fisher GAN [33], and Sobolev GAN [34]. The resulting min-max game in α-GAN is given by

$$\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V_\alpha(\theta,\omega). \tag{22}$$

The following theorem provides the min-max solution, i.e., Nash equilibrium, to the two-player game in (22) for the non-parametric setting, i.e., when the discriminator set Ω is large enough.

Theorem 2. *For a fixed generator G_θ, the discriminator $D_{\omega^*}(x)$ optimizing the* sup *in* (22) *is given by*

$$D_{\omega^*}(x) = \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha}, \tag{23}$$

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

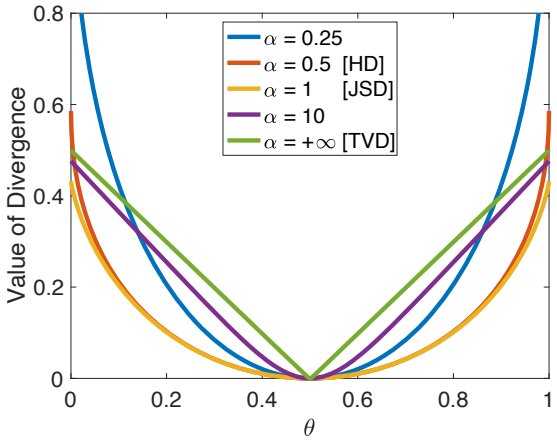


Fig. 4. A plot of D_{f_α} in (26) for several values of α where $p\sim\text{Ber}(1/2)$ and $q\sim\text{Ber}(\theta)$. Note that HD, JSD, and TVD, are abbreviations for Hellinger, Jensen-Shannon, and Total Variation divergences, respectively. As $\alpha\to 0$, the curvature of the divergence increases, 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}.

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). For this $D_{\omega^*}(x)$, (22) simplifies to minimizing a non-negative symmetric f_α-divergence $D_{f_\alpha}(\cdot||\cdot)$ as

$$\inf_{\theta\in\Theta} D_{f_\alpha}(P_r||P_{G_\theta})+\frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}}-2\right), \tag{24}$$

where

$$f_\alpha(u)=\frac{\alpha}{\alpha-1}\left((1+u^\alpha)^{\frac{1}{\alpha}}-(1+u)-2^{\frac{1}{\alpha}}+2\right), \tag{25}$$

for $u\geq 0$ and[3]

$$D_{f_\alpha}(P||Q)=\frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}(p(x)^\alpha+q(x)^\alpha)^{\frac{1}{\alpha}}\,dx-2^{\frac{1}{\alpha}}\right), \tag{26}$$

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

A detailed proof of Theorem 2 is in Appendix B.

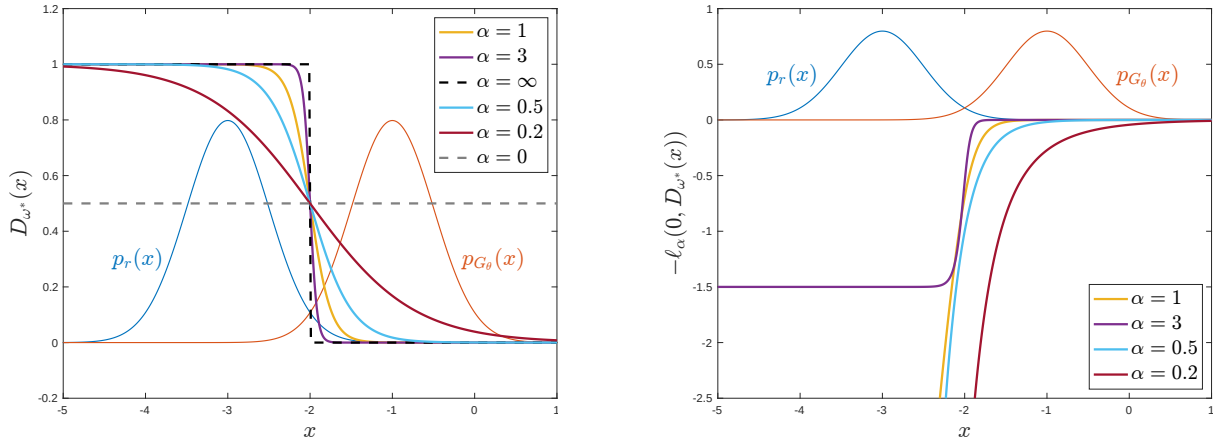
Remark 1. As $\alpha\to 0$, note that (23) implies a more cautious discriminator, i.e., if $p_{G_\theta}(x)\geq p_r(x)$, then $D_{\omega^*}(x)$ decays more slowly from $1/2$, and if $p_{G_\theta}(x)\leq p_r(x)$, $D_{\omega^*}(x)$ increases more slowly from $1/2$. Conversely, as $\alpha\to\infty$, (23) 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, (23) for $\alpha\to\infty$ induces a very confident discriminator. Regarding the generator's perspective, (24) 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 Fig. 4 for an illustration of this observation. See Fig. 5 for a toy example illustrating the effect of tuning α on the optimal discriminator and the generator's corresponding loss.

Note that the divergence $D_{f_\alpha}(\cdot||\cdot)$ (in (26)) that naturally emerges from the analysis of α-GAN was first proposed by Österriecher [11] in the context of statistics and was later referred to as the *Arimoto divergence* by Liese and Vajda [12]. Next we show that α-GAN recovers various well known f-GANs.

Theorem 3. *α-GAN recovers vanilla GAN, Hellinger GAN (H-GAN) [3], and Total Variation GAN (TV-GAN) [3] as $\alpha\to 1$, $\alpha=\frac{1}{2}$, and $\alpha\to\infty$, respectively.*

Proof sketch. We show the following: (i) as $\alpha\to 1$, (24) equals $\inf_{\theta\in\Theta} 2D_{\text{JS}}(P_r||P_{G_\theta})-\log 4$ recovering the vanilla GAN; (ii) for $\alpha=\frac{1}{2}$, (24) gives $2\inf_{\theta\in\Theta} D_{\text{H}^2}(P_r||P_{G_\theta})-2$ recovering Hellinger GAN (up to a constant); and (iii) as $\alpha\to\infty$, (24) equals $\inf_{\theta\in\Theta} D_{\text{TV}}(P_r||P_{G_\theta})-1$ recovering TV-GAN (modulo a constant). A detailed proof is in Appendix C.

[3]We note that the divergence D_{f_α} has been referred to as *Arimoto divergence* in the literature [11], [12], [35]. We refer the reader to Section **??** for more details.



(a) (b)

Fig. 5. A toy example of α-GAN, where the real distribution $P_r = \mathcal{N}(-2, 0.5^2)$ (blue curve) and the assumed initial generated distribution $P_{G_\theta} = \mathcal{N}(2, 0.5^2)$ (orange curve). (a) A plot of the optimal discriminator output $D_{\omega^*}(x)$ for $\alpha \in \{0, 0.2, 0.5, 1, 3, \infty\}$. As α decreases, D_{ω^*} becomes increasingly less confident in its predictions until it outputs $1/2$ for all x when $\alpha \to 0$. Conversely, as α increases, D_{ω^*} becomes increasingly more confident until it implements the Maximum Likelihood decision rule when $\alpha \to \infty$. (b) A plot of the generator's corresponding loss $-\ell_\alpha(0, D_{\omega^*}(x))$ for $\alpha \in \{0.2, 0.5, 1, 3\}$. As α decreases, the magnitude of the gradients of the loss increases, while increasing α saturates the gradients. Note that early in training, if the discriminator is very confident and outputs values close to 0 for the generated data, the generator will not have much gradient to continue learning, which can result in vanishing gradients. Decreasing α reduces the discriminator's confidence and provides more gradient for the generator to learn.

Next, we present an equivalence between f_α-GAN defined using the value function in (8) and α-GAN. Define $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. We first prove that there exists a mapping between the terms involved in the optimization of both GAN formulations in the following theorem.

Theorem 4. *For any $\alpha \in (0, \infty]$, let $\tilde{f}_\alpha$ be a slightly modified version of* (25) *defined for $u \geq 0$ as*

$$\tilde{f}_\alpha(u) = \frac{\alpha}{\alpha - 1} \left((1 + u^\alpha)^{\frac{1}{\alpha}} - (1 + u) \right). \tag{27}$$

Let $\tilde{f}_\alpha^*$ be the convex conjugate of $\tilde{f}_\alpha$ given by

$$\tilde{f}_\alpha^*(t) = \frac{\alpha}{\alpha - 1} \left(1 - (1 - s(t))^{\frac{\alpha - 1}{\alpha}} \right), \tag{28}$$

where

$$s(t) = \left(1 + \frac{\alpha - 1}{\alpha} t \right)^{\frac{\alpha}{\alpha - 1}}. \tag{29}$$

Let $g_{f_\alpha} : \overline{\mathbb{R}} \to dom(\tilde{f}_\alpha^*)$ be a bijective output activation function.

- *Given $v \in \overline{\mathbb{R}}$, there exists $d \in [0, 1]$ such that*

$$g_{f_\alpha}(v) = -\ell_\alpha(1, d) \quad \text{and} \quad \tilde{f}_\alpha^*(g_{f_\alpha}(v)) = \ell_\alpha(0, d). \tag{30}$$

- *Conversely, given $d \in [0, 1]$, there exists $v \in \overline{\mathbb{R}}$ such that* (30) *holds for the same function g_{f_α}.*

Proof sketch. The result follows from comparing the corresponding terms in the f-GAN value function in (8) (specifically for $f = \tilde{f}_\alpha$) and the α-GAN value function in (20). A detailed proof is in Appendix D.

Taking a closer look at the first equality in (30) and recalling that a margin-based loss is often obtained by composing a classification function (such as α-loss) and the logistic sigmoid function, we can derive an example of such a g_{f_α} using the margin-based α-loss [10] as

$$g_{f_\alpha}(v) = \frac{\alpha}{\alpha - 1} \left((1 + e^{-v})^{-\frac{\alpha - 1}{\alpha}} - 1 \right), \tag{31}$$

for $v \in \overline{\mathbb{R}}$ and $\alpha \neq 1$, where

$$g_{f_1}(v) = \lim_{\alpha \to 1} g_{f_\alpha}(v) = -\log(1 + e^{-v}) \tag{32}$$

for $v \in \overline{\mathbb{R}}$. The function g_{f_α} is monotonically increasing for any α, with range exactly matching $\text{dom}(f_\alpha^*)$. TODO: prove g_{f_α} is bijective.

The following corollary establishes the equivalence between f_α-GAN and α-GAN. Two optimization problems $\sup_{v\in A} g(v)$ and $\sup_{t\in B} h(t)$ are said to be equivalent [36], [37] if there exists a bijective function $k: A\to B$ such that

$$g(v)=h(k(v)) \text{ and } h(t)=g(k^{-1}(t)), \text{ for all } v\in A, t\in B. \tag{33}$$

In other words, two optimization problems are equivalent if a change of variable via the function k can transform one into the other.

Corollary 1. *For any $\alpha\in(0,\infty]$ and corresponding $\tilde{f}_\alpha$ defined in* (27)*, the optimization problems involved in $\tilde{f}_\alpha$-GAN (using* (8) *with $f=\tilde{f}_\alpha$) and α-GAN (using* (20)*) are equivalent for the choice*

$$g(Q_w)=\mathbb{E}_{X\sim P_r}\left[g_{f_\alpha}\left(Q_w(X)\right)\right]+\mathbb{E}_{X\sim P_{G_\theta}}\left[-\tilde{f}_\alpha^*\left(g_{f_\alpha}(Q_w(X))\right)\right]$$

with $A=\{Q_\omega: \mathcal{X}\to\overline{\mathbb{R}}\}$ and

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

with $B=\{D_\omega: \mathcal{X}\to[0,1]\}$ using $k: A\to B$ defined by

$$k(v)=s(g_{f_\alpha}(v))=\left(1+\left(\frac{\alpha-1}{\alpha}\right)g_{f_\alpha}(v)\right)^{\frac{\alpha}{\alpha-1}},$$

where s is defined in (29) *and g_{f_α} is a bijective output activation function mapping from $\overline{\mathbb{R}}$ to dom$(\tilde{f}_\alpha^*)$.*

Proof sketch. The result follows from (33) and Theorem 4. Proof details are in Appendix E.

The following theorem generalizes the equivalence demonstrated above between f_α-GAN and α-GAN to an equivalence between f-GANs (using the original value function in (**??**)) and CPE loss based GANs.

Theorem 5. *For any given symmetric f-divergence, the optimization problems involved in f-GAN and the CPE loss based GAN minimizing the same f-divergence are equivalent under the following regularity conditions on f:*

- *there exists a strictly convex CPE (partial) loss function ℓ such that*

$$f(u)=\sup_{t\in[0,1]} -u\ell(t)-\ell(1-t) \tag{34}$$

  (note that this condition without the requirement of strict convexity of ℓ is indeed guaranteed by [18, Theorem 2] for any convex function f resulting in a symmetric divergence), and
- *the function mapping $u\in\mathbb{R}_+$ to unique optimizer in* (34) *is bijective.*

Proof sketch. Observing that the inner optimization problem in the CPE-loss GAN formulation reduces to the pointwise optimization (34) and that of the f-GAN formulation reduces to the pointwise optimization

$$f(u)=\sup_{v\in\text{dom}f^*} uv-f^*(v), \tag{35}$$

it suffices to show that the variational forms of f in (34) and (35) are equivalent. We do this by showing that (34) is equivalent to the optimization problem

$$f(u)=\sup_{v\in\mathbb{R}_+} uf'(v)-[vf'(v)-f(v)], \tag{36}$$

which has been shown to be equivalent to (35) [38]. A detailed proof is in Appendix F.

Remark 2. Since α-loss, $\ell_\alpha(p)=\frac{\alpha}{\alpha-1}(1-p^{\frac{\alpha-1}{\alpha}})$, $p\in[0,1]$, is strictly convex for $\alpha\in(0,\infty)$, and the function mapping $u\in\mathbb{R}_+$ to unique optimizer in (34) with α-loss, i.e., $\frac{u^\alpha}{1+u^\alpha}$, is bijective, Theorem 5 implies that α-GAN is equivalent to $\tilde{f}_\alpha$-GAN with $\tilde{f}_\alpha$ defined in (27).

Remark 3. Though the CPE loss GAN and f-GAN formulations are equivalent, the following aspects differentiate the two:

- The f-GAN formulation focuses on the generator minimizing an f-divergence with no explicit emphasis on the role of the discriminator as a binary classifier in relation to the function f. With the CPE loss GAN formulation, we bring into the foreground the connection between the binary classification performed by the discriminator and the f-divergence minimization done by the generator.
- More importantly, the CPE loss function perspective of GANs allows us to prove convergence properties (Theorem 6), generalization error bounds (Theorem 7), and estimation error bounds (Theorem 8) as detailed in the following sections.

B. Convergence Guarantees for CPE-loss GANs

Building on the above one-to-one correspondence, we now present *convergence* results for CPE loss GANs, including 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 [39], 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.* [39] formally define any divergence that results from the inner optimization of a general GAN in (4) as an *adversarial divergence* [39, 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 [40]. Indeed, the divergence that results from the inner optimization of a CPE loss GAN (including α-GAN) in (16) 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.* [39] 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 [39, 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 (26) is a strict adversarial divergence. We briefly summarize the following terminology from Liu *et al.* [39] to present our results on convergence properties of CPE loss GANs. Let $\mathcal{P}(\mathcal{X})$ be the probability simplex of distributions over $\mathcal{X}$.

Definition 2 (Definition 11, [39])**.** *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 [39, 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 6. In fact, we prove that all symmetric f-divergences, including D_{f_α}, are equivalent in convergence.

Theorem 6. *Let $f_i:[0,\infty)\to\mathbb{R}$ be a convex function which is continuous at 0 and strictly convex at 1 such that $f_i(1)=0$, $uf_i(\frac{1}{u})=f_i(u)$, and $f_i(0)<\infty$, for $i\in\{1,2\}$. Then for a sequence of probability distributions $(P_n)_{n\in\mathbb{N}}\in\mathcal{P}(\mathcal{X})$ and a fixed distribution $P\in\mathcal{P}(\mathcal{X})$, we have $D_{f_1}(P_n||P)\to 0$ as $n\to\infty$ if and only if $D_{f_2}(P_n||P)\to 0$ as $n\to\infty$.*

Proof sketch. Note that it suffices to show that $D_f(\cdot\|\cdot)$ is equivalent to $D_{\text{TV}}(\cdot\|\cdot)$ for any function f satisfying the conditions in the theorem. To show this, we employ an elegant result by Feldman and Österreicher [**?**, Theorem 2] which gives lower and upper bounds on the Arimoto divergence in terms of TVD as

$$\gamma_f(D_{\text{TV}}(P||Q))\le D_f(P||Q)\le\gamma_f(1)D_{\text{TV}}(P||Q), \tag{37}$$

for an appropriately defined well-behaved (continuous, invertible, and bounded) function $\gamma_\alpha:[0,1]\to[0,\infty)$. We use the lower and upper bounds in (37) to show that $D_f(\cdot\|\cdot)$ is stronger than $D_{\text{TV}}(\cdot\|\cdot)$, and $D_f(\cdot\|\cdot)$ is weaker than $D_{\text{TV}}(\cdot\|\cdot)$, respectively. Proof details are in Appendix G.

Remark 4. We note that the proof techniques used in proving Theorem 6 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 H for details.

Theorems 1 through 6 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.* [26] show that under such limitations, convergence in divergence does not imply convergence in distribution, and have led to new metrics for evaluating GANs. To address these limitations, we consider two measures to evaluate the performance of GANs, namely generation and estimation errors, as detailed below.

C. Generalization and Estimation Error Bounds for CPE Loss GANs

Arora *et al.* [26] first defined *generalization* in GANs as the scenario when the divergence between the real distribution and the generated distribution is well-captured by the divergence between their empirical versions. In particular, a divergence or

distance[4] between distributions *generalizes* with m training samples and error ϵ if, for the learnt distribution P_G, the following holds with high probability:

$$\left|d(P_r,P_G)-d(\hat{P}_r,\hat{P}_G)\right|\leq\epsilon, \tag{38}$$

where $\hat{P}_r$ and $\hat{P}_G$ are the empirical versions of the real distribution (with m samples) and the generated distribution (with polynomial number of samples), respectively. It was shown by Arora *et al.* [26, Lemma 1] that the Jensen-Shannon divergence and Wasserstein distance don't generalize with any polynomial number of samples. However, they also showed that generalization does happen for a new notion of divergence, the *neural net divergence*, with a moderate number of training examples [26, Theorem 3.1]. To this end, they consider the following optimization problem

$$\inf_{\theta\in\Theta} d_{\mathcal{F}}(P_r,P_{G_\theta}), \tag{39}$$

where $d_{\mathcal{F}}(P_r,P_{G_\theta})$ is the neural net divergence defined as

$$d_{\mathcal{F}}(P_r,P_{G_\theta})=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]+\mathbb{E}_{X\sim P_{G_\theta}}[\phi(1-D_\omega(X))]\Big)-2\phi\left(\frac{1}{2}\right) \tag{40}$$

such that the class of discriminators $\mathcal{F}=\{D_\omega:\omega\in\Omega\}$ is L-Lipschitz with respect to the parameters ω, i.e., for every $x\in\mathcal{X}$, $|D_{\omega_1}(x)-D_{\omega_2}(x)|\leq L||\omega_1-\omega_2||$, for all $\omega_1,\omega_2\in\Omega$, and the function ϕ takes values in $[-\Delta,\Delta]$ and is L_ϕ-Lipschitz. Let p be the discriminator capacity (i.e., number of parameters) and $\epsilon>0$. For these assumptions, in [26, Theorem 3.1], Arora *et al.* prove that (40) generalizes. We summarize their result as follows: for the empirical versions $\hat{P}_r$ and $\hat{P}_G$ of two distributions P_r and P_G, respectively, with at least m random samples each, there exists a universal constant c such that when $m\geq\frac{cp\Delta^2\log(LL_\phi p/\epsilon)}{\epsilon^2}$, with probability at least $1-\exp(-p)$ (over the randomness of samples),

$$\left|d_{\mathcal{F}}(P_r,P_G)-d_{\mathcal{F}}(\hat{P}_r,\hat{P}_g)\right|\leq\epsilon. \tag{41}$$

Our first contribution is to show that we can generalize (40) and [26, Theorem 3.1] to incorporate any partial losses ϕ and ψ (not just those that are symmetric). To this end, we first define the *refined neural net divergence* as

$$\tilde{d}_{\mathcal{F}}(P_r,P_{G_\theta})=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]+\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))]\Big)-\phi\left(\frac{1}{2}\right)-\psi\left(\frac{1}{2}\right), \tag{42}$$

where the discriminator class is same as the above and the functions ϕ and ψ take values in $[-\Delta,\Delta]$ and are L_ϕ- and L_ψ-Lipschitz, respectively. Note that the functions ϕ and ψ should also satisfy (17) so as to respect the optimality of the uniformly random discriminator when $P_r=P_{G_\theta}$. The following theorem shows that the refined neural net divergence generalizes with a moderate number of training examples, thus extending [26, Theorem 3.1].

Theorem 7. *Let $\hat{P}_r$ and $\hat{P}_G$ be empirical versions of two distributions P_r and P_G, respectively, with at least m random samples each. For $\Delta,p,L,L_\phi,L_\psi,\epsilon>0$ defined above, there exists a universal constant c such that when $m\geq\frac{cp\Delta^2\log(L\max\{L_\phi,L_\psi\}p/\epsilon)}{\epsilon^2}$, we have that with probability at least $1-\exp(-p)$ (over the randomness of samples),*

$$\left|\tilde{d}_{\mathcal{F}}(P_r,P_G)-\tilde{d}_{\mathcal{F}}(\hat{P}_r,\hat{P}_g)\right|\leq\epsilon. \tag{43}$$

When $\phi(t)=t$ and $D_\omega=f_\omega$ can take values in $\mathbb{R}$ (not just in $[0,1]$), (40) yields the so-called *neural net (nn) distance*[5] [26], [41], [42] given by:

$$d_{\mathcal{F}_{nn}}(P_r,P_{G_\theta})=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P_r}[f_\omega(X)]-\mathbb{E}_{X\sim P_{G_\theta}}[f_\omega(X)]\Big), \tag{44}$$

where the discriminator[6] and generator $f_\omega(\cdot)$ and $G_\theta(\cdot)$, respectively, are neural networks. Using (44), Ji *et al.* [42] defined and studied the notion of *estimation error*, which quantifies the effectiveness of the generator (for a corresponding optimal discriminator model) in learning the real distribution with limited samples. In order to define estimation error for CPE-loss GANs (including α-GAN), we first introduce a *loss-inclusive neural net divergence*[7] $d^{(\ell)}_{\mathcal{F}_{nn}}$ to highlight the effect of the *loss* on the error. For training samples $S_x=\{X_1,\ldots,X_n\}$ and $S_z=\{Z_1,\ldots,Z_m\}$ from P_r and P_Z, respectively, we begin with the following minimization for GAN training:

$$\inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(\hat{P}_r,\hat{P}_{G_\theta}), \tag{45}$$

[4]For consistency with other works on generalization and estimation error, we refer to a semi-metric as a distance.

[5]This term was first introduced in [26] but with a focus on a discriminator D_ω taking values in $[0,1]$. Ji *et al.* [41], [42] generalized it to $D_\omega=f_\omega$ taking values in $\mathbb{R}$.

[6]In [42], f_ω indicates a discriminator function that takes values in $\mathbb{R}$.

[7]We refer to this measure as a divergence since it may not be a semi-metric for all choices of the loss ℓ.

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

$$d^{(\ell)}_{\mathcal{F}_{nn}}(\hat{P}_r,\hat{P}_{G_\theta})=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim\hat{P}_r}[\phi\big(D_\omega(X)\big)]+\mathbb{E}_{X\sim\hat{P}_{G_\theta}}[\psi\big(D_\omega(X)\big)]\Big)-\phi\left(\frac{1}{2}\right)-\psi\left(\frac{1}{2}\right), \tag{46}$$

where for brevity we henceforth use $\phi(\cdot):=-\ell(1,\cdot)$ and $\psi(\cdot):=-\ell(0,\cdot)$. As proven in Theorem 3, for $\ell=\ell_\alpha$ and $\alpha=\infty$, (46) reduces to the neural net total variation distance.

As a step towards obtaining bounds on the estimation error, we consider the following setup, analogous to that in [42]. For $x\in\mathcal{X}:=\{x\in\mathbb{R}^d:||x||_2\le B_x\}$ and $z\in\mathcal{Z}:=\{z\in\mathbb{R}^p:||z||_2\le B_z\}$, we consider discriminators and generators as neural network models of the form:

$$D_\omega: x\mapsto\sigma\big(\mathbf{w}_k^{\mathsf{T}} r_{k-1}(\mathbf{W}_{d-1}r_{k-2}(\dots r_1(\mathbf{W}_1(x))))\big) \tag{47}$$

$$G_\theta: z\mapsto\mathbf{V}_l s_{l-1}(\mathbf{V}_{l-1}s_{l-2}(\dots s_1(\mathbf{V}_1 z))), \tag{48}$$

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 $\mathbf{a}\in\mathbb{R}^t$, $r_i(\mathbf{a})=[r_i(a_1),\dots,r_i(a_t)]$ 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 [42, 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 homogeneous, i.e., $r_i(\lambda p)=\lambda r_i(p)$ and $s_j(\lambda p)=\lambda s_j(p)$, for any $\lambda\ge 0$ and $p\in\mathbb{R}$. Finally, as modelled in [42]–[45], we assume that the Frobenius norms of the parameter matrices are bounded, i.e., $||\mathbf{W}_i||_F\le M_i$, $i\in[1:k-1]$, $||\mathbf{w}_k||_2\le M_k$, and $||\mathbf{V}_j||_F\le N_j$, $j\in[1:l]$.

We define the estimation error for a CPE loss GAN as

$$d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_{\hat{\theta}^*}})-\inf_{\theta\in\Theta}d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}), \tag{49}$$

where $\hat{\theta}^*$ is the minimizer of (45) and present the following upper bound on the error. We also specialize these bounds for α-GANs, relying on the Rademacher complexity of this loss class to do so.

Theorem 8. *For 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*

$$\begin{aligned}&d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}})-\inf_{\theta\in\Theta}d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta})\\&\le\frac{L_\phi B_x U_\omega\sqrt{3k}}{\sqrt{n}}+\frac{L_\psi U_\omega U_\theta B_z\sqrt{3(k+l-1)}}{\sqrt{m}}\\&\quad+U_\omega\sqrt{\log\frac{1}{\delta}}\left(\frac{L_\phi B_x}{\sqrt{2n}}+\frac{L_\psi B_z U_\theta}{\sqrt{2m}}\right),\end{aligned} \tag{50}$$

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

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 (50)*, 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*

$$C_h(\alpha):=\begin{cases}\sigma(h)\sigma(-h)^{\frac{\alpha-1}{\alpha}}, & \alpha\in(0,1]\\ \left(\frac{\alpha-1}{2\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\frac{\alpha}{2\alpha-1}, & \alpha\in(1,\infty).\end{cases} \tag{51}$$

Proof sketch. Our proof involves the following steps:

- Building upon the proof techniques of Ji *et al.* [42, 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 [46, Lemma 26.9]. We remark that this differs considerably from the way the bounds on Rademacher complexities in [42, 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.* [23] where they showed that α-loss is Lipschitz for a logistic model with (72). 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 (72) 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 (50) is *also* a function of α. Proof details are in Appendix J.

We now focus on developing lower bounds on the estimation error. Due to the fact that oft-used techniques to obtain min-max lower bounds on the quality of an estimator (e.g., LeCam's methods, Fano's methods, etc.) require a semi-metric distance measure, we restrict our attention to a particular α-GAN, namely that for $\alpha=\infty$, to derive a matching lower bound on the estimation error. We consider the loss-inclusive neural net divergence in (46) with $\ell=\ell_\alpha$ for $\alpha=\infty$, which, for brevity, we henceforth denote as $d^{\ell_\infty}_{\mathcal{F}_{nn}}(\cdot,\cdot)$ As in [42], suppose the generator's class $\{G_\theta\}_{\theta\in\Theta}$ is rich enough such that the generator G_θ can learn the real distribution P_r and that the number m of training samples in S_z scales faster than the number n of samples in S_x[8]. Then $\inf_{\theta\in\Theta} d^{\ell_\infty}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta})=0$, so the estimation error simplifies to the single term $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P_r,P_{G_{\hat{\theta}*}})$. Furthermore, the upper bound in (50) reduces to $O(c/\sqrt{n})$ for some constant c (note that, in (51), $C_h(\infty)=1/4$). In addition to the above assumptions, also assume the activation functions r_i for $i\in[1:k-1]$ are either strictly increasing or ReLU. For the above setting, we derive a matching min-max lower bound (up to a constant multiple) on the estimation error.

Theorem 9. *For the setting above, let $\hat{P}_n$ be an estimator of P_r learned using the training samples $S_x=\{X_i\}_{i=1}^n$. Then,*

$$\inf_{\hat{P}_n}\sup_{P_r\in\mathcal{P}(\mathcal{X})}\mathbb{P}\left\{d^{\ell_\infty}_{\mathcal{F}_{nn}}(\hat{P}_n,P_r)\geq\frac{C(\mathcal{P}(\mathcal{X}))}{\sqrt{n}}\right\}>0.24,$$

where the constant $C(\mathcal{P}(\mathcal{X}))$ is given by

$$C(\mathcal{P}(\mathcal{X}))=\frac{\log(2)}{20}\Big[\sigma(M_k r_{k-1}(\dots r_1(M_1B_x))-\sigma(M_k r_{k-1}(\dots r_1(-M_1B_x))\Big]. \tag{52}$$

Proof sketch. To obtain min-max lower bounds, we first prove that $d^{\ell_\infty}_{\mathcal{F}_{nn}}$ is a semi-metric. The remainder of the proof is similar to that of [42, Theorem 2], replacing $d_{\mathcal{F}_{nn}}$ with $d^{\ell_\infty}_{\mathcal{F}_{nn}}$. Finally, we note that the additional sigmoid activation function after the last layer in D satisfies the monotonicity assumption as detailed in Appendix K. A challenge that remains to be addressed is to verify if $d^{\ell_\alpha}_{\mathcal{F}_{nn}}$ is a semi-metric for $\alpha<\infty$.

IV. Dual-objective GANs

As illustrated in Fig. 5, tuning $\alpha<1$ provides more gradient for the generator to learn early in training when the discriminator more confidently classifies the generated data as fake, alleviating vanishing gradients and also creates a smooth landscape for the generated data to descend towards the real data, alleviating exploding gradients. However, tuning $\alpha<1$ may provide too large of gradients for the generator when the generated samples approach the real samples, which can result in too much movement of the generated data, potentially repelling it from the real data. The following question therefore arises: Can we combine a less confident discriminator with a more stable generator loss? We show that we can do so by using different objectives for the discriminator and generator, resulting in the (α_D,α_G)-GAN.

A. (α_D,α_G)-GAN

LS - We should note that the ordering is similar to the min-max in terms of the game to be played.

We propose a dual-objective (α_D,α_G)-GAN with different objective functions for the generator and discriminator in which the discriminator maximizes $V_{\alpha_D}(\theta,\omega)$ while the generator minimizes $V_{\alpha_G}(\theta,\omega)$, where

$$V_\alpha(\theta,\omega)=\mathbb{E}_{X\sim P_r}[-\ell_\alpha(1,D_\omega(X))]+\mathbb{E}_{X\sim P_{G_\theta}}[-\ell_\alpha(0,D_\omega(X))], \tag{53}$$

for $\alpha=\alpha_D,\alpha_G\in(0,\infty]$. We recover the α-GAN [7], [8] value function when $\alpha_D=\alpha_G=\alpha$. The resulting (α_D,α_G)-GAN is given by

$$\sup_{\omega\in\Omega} V_{\alpha_D}(\theta,\omega) \tag{54a}$$

$$\inf_{\theta\in\Theta} V_{\alpha_G}(\theta,\omega). \tag{54b}$$

The following theorem presents the conditions under which the optimal generator learns the real distribution P_r when the discriminator set Ω is large enough.

Theorem 10. *For a fixed generator G_θ, the discriminator optimizing* (54a) *is given by*

$$D_{\omega^*}(x)=\frac{p_r(x)^{\alpha_D}}{p_r(x)^{\alpha_D}+p_{G_\theta}(x)^{\alpha_D}}, \tag{55}$$

[8]Since the noise distribution P_Z is known, one can generate an arbitrarily large number m of noise samples.

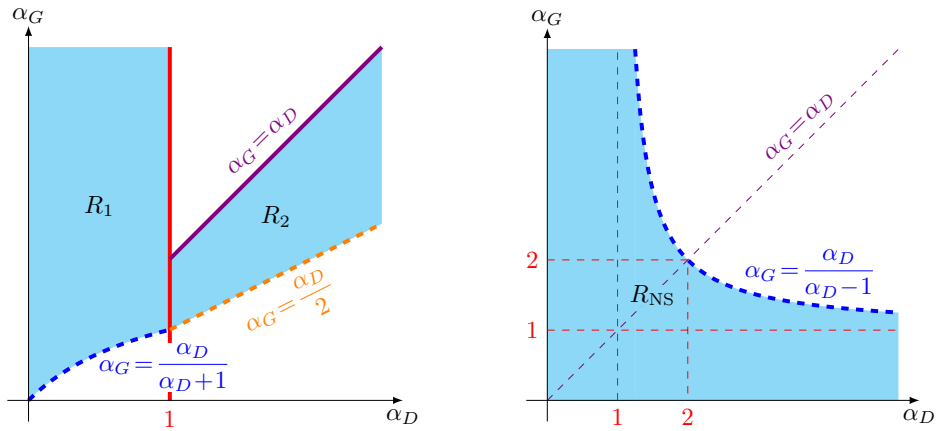


Fig. 6. (a) Plot of regions $R_1=\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,|\,\alpha_D\leq 1,\alpha_G>\frac{\alpha_D}{\alpha_D+1}\}$ and $R_2=\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,|\,\alpha_D>1,\frac{\alpha_D}{2}<\alpha_G\leq\alpha_D\}$ for which f_{α_D,α_G} is strictly convex. (b) Plot of region $R_{\text{NS}}=\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,|\,\alpha_D+\alpha_G>\alpha_D\alpha_G\}$ for which $f^{\text{NS}}_{\alpha_D,\alpha_G}$ is strictly convex.

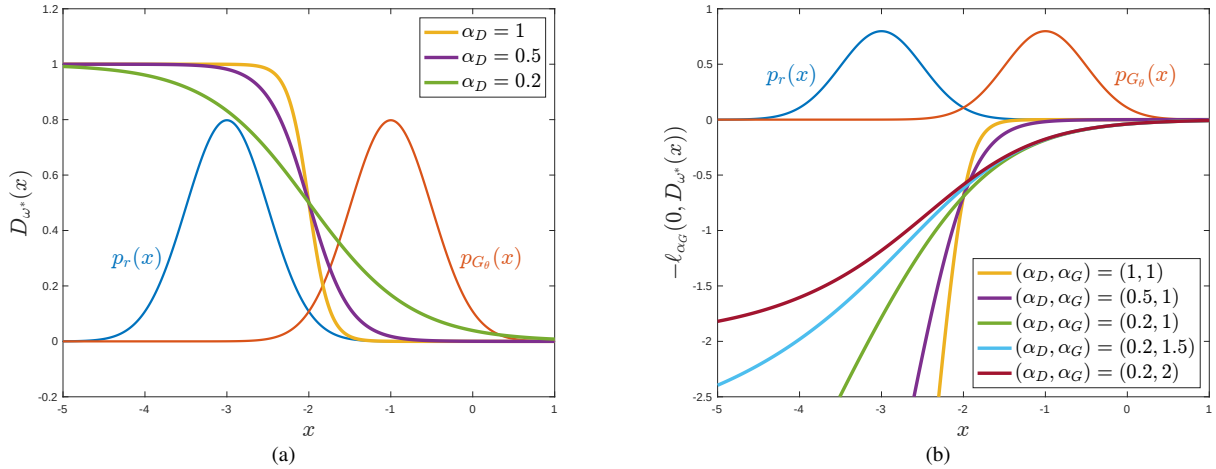


Fig. 7. (a) A plot of the optimal discriminator output $D_{\omega^*}(x)$ in (55) for several values of $\alpha_D\leq 1$ for the same toy example as in Figure 5. Tuning $\alpha_D<1$ decreases the confidence of the optimal discriminator D_{ω^*}. (b) A plot of the generator's loss $-\ell_{\alpha_G}(0,D_{\omega^*}(x))$ for several values of $(\alpha_D\leq 1,\alpha_G\geq 1)$. Tuning $\alpha_D<1$ provides more gradient for the generator to learn early in training when the discriminator more confidently classifies the generated data as fake, thereby alleviating vanishing gradients, while tuning $\alpha_G\geq 1$ creates a smooth landscape for the generated data to descend towards the real data, alleviating exploding gradients.

For this D_{ω^*} and the function $f_{\alpha_D,\alpha_G}:\mathbb{R}_+\to\mathbb{R}$ defined as

$$f_{\alpha_D,\alpha_G}(u)=\frac{\alpha_G}{\alpha_G-1}\left(\frac{u^{\alpha_D\left(1-\frac{1}{\alpha_G}\right)+1}+1}{(u^{\alpha_D}+1)^{1-\frac{1}{\alpha_G}}}-2^{\frac{1}{\alpha_G}}\right), \tag{56}$$

(54b) *simplifies to minimizing a non-negative symmetric f_{α_D,α_G}-divergence $D_{f_{\alpha_D,\alpha_G}}(\cdot||\cdot)$ as*

$$\inf_{\theta\in\Theta} D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta})+\frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right), \tag{57}$$

which is minimized iff $P_{G_\theta}=P_r$ for $(\alpha_D,\alpha_G)\in(0,\infty]^2$ such that $\left(\alpha_D\leq 1,\alpha_G>\frac{\alpha_D}{\alpha_D+1}\right)$ or $\left(\alpha_D>1,\frac{\alpha_D}{2}<\alpha_G\leq\alpha_D\right)$.

Proof sketch. We substitute the optimal discriminator of (54a) into the objective function of (54b) and translate it into the form

$$\int_{\mathcal{X}} p_{G_\theta}(x) f_{\alpha_D,\alpha_G}\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx+\frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right). \tag{58}$$

We then find the conditions on α_D and α_G for f_{α_D,α_G} to be strictly convex so that the first term in (58) is an f-divergence. Figure 6(a) illustrates the feasible (α_D,α_G)-region. A detailed proof can be found in Appendix L. See Fig. 7 for a toy example illustrating the value of tuning $\alpha_D<1$ and $\alpha_G\geq 1$.

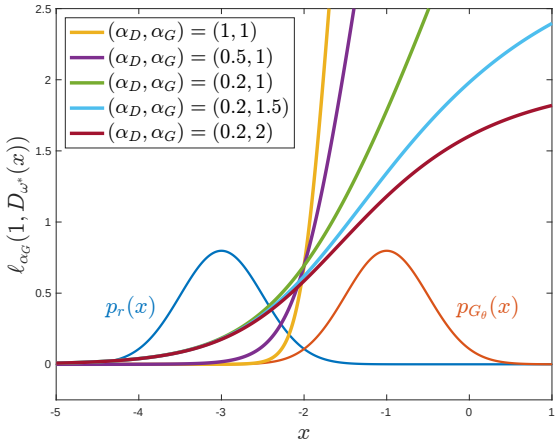


Fig. 8. A plot of the generator's NS loss $\ell_{\alpha_G}(1, D_{\omega^*}(x))$ for several values of $(\alpha_D \leq 1, \alpha_G \geq 1)$. Tuning $\alpha_D < 1$ and $\alpha_G = 1$ makes the loss less convex, which can help stabilize training by decreasing sensitivity to hyperparameter initialization and alleviating mode collapse; tuning $\alpha_G > 1$ results in a quasiconvex generator objective, which can further improve training stability.

Noting that α-GAN recovers various well-known GANs, including the vanilla GAN, which is prone to saturation, the (α_D, α_G)-GAN formulation using the generator objective function in (53) can similarly saturate early in training, potentially causing vanishing gradients. Thus, we propose the following NS alternative to the generator's objective in (53):

$$V^{\mathrm{NS}}_{\alpha_G}(\theta, \omega) = \mathbb{E}_{X \sim P_{G_\theta}}[\ell_{\alpha_G}(1, D_\omega(X))], \tag{59}$$

thereby replacing (54b) with

$$\inf_{\theta \in \Theta} V^{\mathrm{NS}}_{\alpha_G}(\theta, \omega). \tag{60}$$

Comparing (54b) and (60), note that the additional expectation term over P_r in (53) results in (54b) simplifying to a symmetric divergence for D_{ω^*} in (55), whereas the single term in (59) will result in (60) simplifying to an asymmetric divergence. The optimal discriminator for this NS game remains the same as in (55). The following theorem provides the solution to (60) under the assumption that the optimal discriminator can be attained.

Theorem 11. *For the same D_{ω^*} in* (55) *and the function $f^{NS}_{\alpha_D, \alpha_G} : \mathbb{R}_+ \to \mathbb{R}$ defined as*

$$f^{NS}_{\alpha_D, \alpha_G}(u) = \frac{\alpha_G}{\alpha_G - 1} \left(2^{\frac{1}{\alpha_G} - 1} - \frac{u^{\alpha_D \left(1 - \frac{1}{\alpha_G}\right)}}{\left(u^{\alpha_D} + 1\right)^{1 - \frac{1}{\alpha_G}}} \right), \tag{61}$$

(54b) *simplifies to minimizing a non-negative asymmetric $f^{NS}_{\alpha_D, \alpha_G}$-divergence $D_{f^{NS}_{\alpha_D, \alpha_G}}(\cdot || \cdot)$ as*

$$\inf_{\theta \in \Theta} D_{f^{NS}_{\alpha_D, \alpha_G}}(P_r || P_{G_\theta}) + \frac{\alpha_G}{\alpha_G - 1} \left(1 - 2^{\frac{1}{\alpha_G} - 1} \right), \tag{62}$$

which is minimized iff $P_{G_\theta} = P_r$ for $(\alpha_D, \alpha_G) \in (0, \infty]^2$ such that $\alpha_D + \alpha_G > \alpha_G \alpha_D$.

The proof mimics that of Theorem 10 and is detailed in Appendix L-B. Figure 6(b) illustrates the feasible (α_D, α_G)-region; in contrast to the saturating setting of Theorem 10, the NS setting constrains $\alpha \leq 2$ when $\alpha_D = \alpha_G = \alpha$. See Figure 8 for a toy example illustrating how tuning $\alpha_D < 1$ and $\alpha_G \geq 1$ can also alleviate training instabilities in the NS setting.

LS - add the gradient theorem here before segueing to general CPE-loss dual objective

B. CPE-loss based dual-objective GANs

Similarly to the single-objective loss function perspective in Section III, we can generalize the (α_D, α_G)-GAN formulation to incorporate general CPE losses. To this end, we introduce a dual-objective loss function perspective of GANs in which the discriminator maximizes $V_{\ell_D}(\theta, \omega)$ while the generator minimizes $V_{\ell_G}(\theta, \omega)$, where

$$V_\ell(\theta, \omega) = \mathbb{E}_{X \sim P_r}[-\ell(1, D_\omega(X))] + \mathbb{E}_{X \sim P_{G_\theta}}[-\ell(0, D_\omega(X))], \tag{63}$$

for any CPE losses $\ell = \ell_D, \ell_G$. The resulting CPE-loss dual-objective GAN is given by

$$\sup_{\omega \in \Omega} V_{\ell_D}(\theta, \omega) \tag{64a}$$

$$\inf_{\theta \in \Theta} V_{\ell_G}(\theta, \omega). \tag{64b}$$

The CPE losses ℓ_D and ℓ_G can be completely different losses, the same loss but with different parameter values, or the same loss with the same parameter values, in which case the above formulation reduces to the single-objective formulation in (16). For example, choosing $\ell_D=\ell_G=\ell_\alpha$, we recover the α-GAN formulation in (22); choosing $\ell_D=\ell_{\alpha_D}$ and $\ell_G=\ell_{\alpha_G}$, we obtain the (α_D,α_G)-GAN formulation in (54). The

C. Estimation error for CPE-loss dual-objective GANs

Theorem 10 assumes sufficiently large number of training samples and ample discriminator and generator capacity. However, in practice both the number of training samples and model capacity are usually limited. We consider the same setting as in Section III-C with finite training samples $S_x=\{X_1,\ldots,X_n\}$ and $S_z=\{Z_1,\ldots,Z_m\}$ from P_r and P_Z, respectively, and with neural networks chosen as the discriminator and generator models. The sets of samples S_x and S_z induce the empirical real and generated distributions $\hat{P}_r$ and $\hat{P}_{G_\theta}$, respectively. A useful quantity to evaluate the performance of GANs in this setting is again that of the estimation error. In Section III-C, we define estimation error for CPE-loss GANs. However, such a definition requires a common value function for both discriminator and generator, and therefore, does not directly apply to the dual-objective setting we consider here.

Our definition relies on the observation that estimation error inherently captures the effectiveness of the generator (for a corresponding optimal discriminator model) in learning with limited samples. We formalize this intuition below.

Since our proposed CPE-loss dual-objective GANs use different objective functions for the discriminator and generator, we start by defining the optimal discriminator ω^* for a generator model G_θ as

$$\omega^*(P_r,P_{G_\theta}) \coloneqq \operatorname*{argmax}_{\omega\in\Omega} V_{\ell_D}(\theta,\omega)\big|_{P_r,P_{G_\theta}}, \tag{65}$$

where the notation $|_{\cdot,\cdot}$ allows us to make explicit the distributions used in the value function. In keeping with the literature where the value function being minimized is referred to as the neural net (NN) distance (since D and G are modeled as neural networks) [8], [26], [42], we define the generator's NN distance $d_{\omega^*(P_r,P_{G_\theta})}$ as

$$d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta}) \coloneqq V_{\ell_G}(\theta,\omega^*(P_r,P_{G_\theta}))\big|_{P_r,P_{G_\theta}}. \tag{66}$$

The resulting minimization for training the CPE-loss dual-objective GAN using finite samples is

$$\inf_{\theta\in\Theta} d_{\omega^*(\hat{P}_r,\hat{P}_{G_\theta})}(\hat{P}_r,\hat{P}_{G_\theta}). \tag{67}$$

Denoting $\hat{\theta}^*$ as the minimizer of (67), we define the estimation error for CPE-loss dual-objective GANs as

$$d_{\omega^*(P_r,P_{G_{\hat{\theta}^*}})}(P_r,P_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta}). \tag{68}$$

LS - make a remark on why we take the generator's perspective? Remark on the dependence on the discriminator that cannot be eliminated – pointwise vs. uniform

We use the same notation as in Section III-C, detailed again in the following for easy reference. For $x\in\mathcal{X}\coloneqq\{x\in\mathbb{R}^d: ||x||_2\le B_x\}$ and $z\in\mathcal{Z}\coloneqq\{z\in\mathbb{R}^p: ||z||_2\le B_z\}$, we model the discriminator and generator as k- and l-layer neural networks, respectively, such that D_ω and G_θ can be written as:

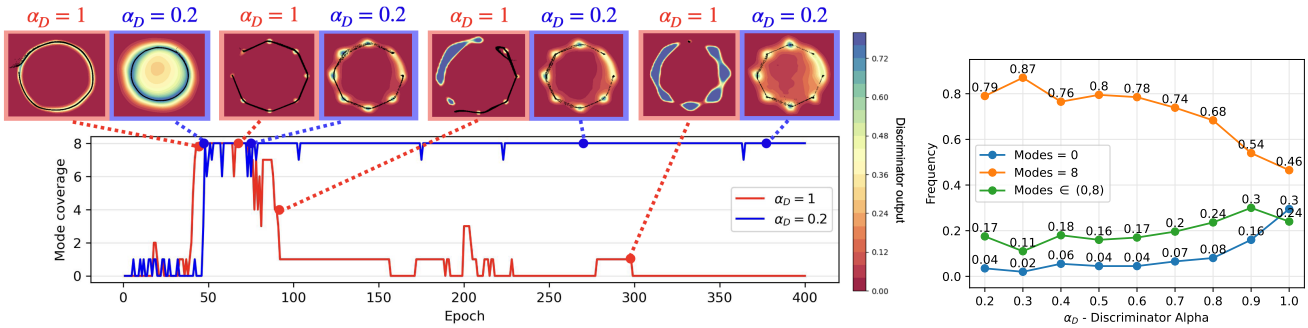
$$D_\omega : x\mapsto \sigma\left(\mathbf{w}_k^\mathsf{T} r_{k-1}(\mathbf{W}_{d-1} r_{k-2}(\ldots r_1(\mathbf{W}_1(x))))\right) \tag{69}$$

$$G_\theta : z\mapsto \mathbf{V}_l s_{l-1}(\mathbf{V}_{l-1} s_{l-2}(\ldots s_1(\mathbf{V}_1 z))), \tag{70}$$

where (i) $\mathbf{w}_k$ is a parameter vector of the output layer; (ii) for $i\in[1:k-1]$ and $j\in[1:l]$, $\mathbf{W}_i$ and $\mathbf{V}_j$ are parameter matrices; (iii) $r_i(\cdot)$ and $s_j(\cdot)$ are entry-wise activation functions of layers i and j, respectively, 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 (iv) $\sigma(\cdot)$ is the sigmoid function given by $\sigma(p)=1/(1+\mathrm{e}^{-p})$. 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 homogeneous, i.e., $r_i(\lambda p)=\lambda r_i(p)$ and $s_j(\lambda p)=\lambda s_j(p)$, for any $\lambda\ge 0$ and $p\in\mathbb{R}$. Finally, as is common in such analysis [42]–[45], we assume that the Frobenius norms of the parameter matrices are bounded, i.e., $||\mathbf{W}_i||_F\le M_i$, $i\in[1:k-1]$, $||\mathbf{w}_k||_2\le M_k$, and $||\mathbf{V}_j||_F\le N_j$, $j\in[1:l]$. We now present an upper bound on (68) in the following theorem.

Theorem 12. *For the setting described above, additionally assume that the functions $\phi(\cdot)\coloneqq\ell_G(1,\cdot)$ and $\psi(\cdot)\coloneqq\ell_G(1,\cdot)$ 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*

$$\begin{aligned} &d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta}) \\ &\le \frac{L_\phi B_x U_\omega\sqrt{3k}}{\sqrt{n}} + \frac{L_\psi U_\omega U_\theta B_z\sqrt{3(k+l-1)}}{\sqrt{m}} \\ &\quad + U_\omega\sqrt{\log\frac{1}{\delta}}\left(\frac{L_\phi B_x}{\sqrt{2n}} + \frac{L_\psi B_z U_\theta}{\sqrt{2m}}\right), \end{aligned} \tag{71}$$



(a) (b)

Fig. 9. (a) Plot of mode coverage over epochs for (α_D,α_G)-GAN training with the **saturating** objectives in (54). Fixing $\alpha_G=1$, we compare $\alpha_D=1$ (vanilla GAN) with $\alpha_D=0.2$. Placed above this plot are 2D visuals of the generated samples (in black) at different epochs; these show that both GANs successfully capture the ring-like structure, but the vanilla GAN fails to maintain the ring over time. We illustrate the discriminator output in the same visual as a heat map to show that the $\alpha_D=1$ discriminator exhibits more confident predictions (tending to 0 or 1), which in turn subjects G to vanishing and exploding gradients when its objective $\log(1-D)$ saturates as $D\to 0$ and diverges as $D\to 1$, respectively. This combination tends to repel the generated data when it approaches the real data, thus freezing any significant weight update in the future. In contrast, the less confident predictions of the $(0.2,1)$-GAN create a smooth landscape for the generated output to descend towards the real data. (b) Plot of success and failure rates over 200 seeds vs. α_D with $\alpha_G=1$ for the **saturating** (α_D,α_G)-GAN on the 2D-ring, which underscores the stability of $(\alpha_D<1,\alpha_G)$-GANs relative to vanilla GAN.

where the parameters $U_\omega \coloneqq M_k\prod_{i=1}^{k-1}(M_iR_i)$ *and* $U_\theta \coloneqq N_l\prod_{j=1}^{l-1}(N_jS_j)$.

In particular, when this bound is specialized to the case of (α_D,α_G)*-GAN by letting* $\phi(p)=\psi(1-p)=\frac{\alpha_G}{\alpha_G-1}\left(1-p^{\frac{\alpha_G-1}{\alpha_G}}\right)$, *the resulting bound is nearly identical to the terms in the RHS of* (71)*, except for substitutions* $L_\phi \leftarrow 4C_{Q_x}(\alpha_G)$ *and* $L_\psi \leftarrow 4C_{Q_z}(\alpha_G)$, *where* $Q_x \coloneqq U_\omega B_x$, $Q_z \coloneqq U_\omega U_\theta B_z$, *and*

$$C_h(\alpha) \coloneqq \begin{cases} \sigma(h)\sigma(-h)^{\frac{\alpha-1}{\alpha}}, & \alpha\in(0,1] \\ \left(\frac{\alpha-1}{2\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\frac{\alpha}{2\alpha-1}, & \alpha\in(1,\infty). \end{cases} \tag{72}$$

The proof is similar to that of Theorem 8 (and also [42, Theorem 1]). We observe that (71) does not depend on ℓ_D, an artifact of the proof techniques used, and is therefore most likely not the tightest bound possible. See Appendix M for proof details.

V. Illustration of Results

We illustrate the value of (α_D,α_G)-GAN as compared to the vanilla GAN (i.e., the $(1,1)$-GAN). Focusing on DCGAN architectures [47], we compare against LSGANs [15], the current state-of-the-art (SOTA) dual-objective approach. While WGANs [4] have also been proposed to address the training instabilities, their training methodology is distinctly different and uses a different optimizer (RMSprop), requires gradient clipping or penalty, and does not leverage batch normalization, all of which make meaningful comparisons difficult.

We evaluate our approach on three datasets: (i) a synthetic dataset generated by a two-dimensional, ring-shaped Gaussian mixture distribution (2D-ring) [48];(ii) the 64×64 Celeb-A image dataset [49]; and (iii) the 112×112 LSUN Classroom dataset [50]. For each dataset and pair of GAN objectives, we report several metrics that encapsulate the stability of GAN training over hundreds of random seeds. This allows us to clearly showcase the potential for tuning (α_D,α_G) to obtain stable and robust solutions for image generation.

A. 2D Gaussian Mixture Ring

The 2D-ring is an oft-used synthetic dataset for evaluating GANs. We draw samples from a mixture of 8 equal-prior Gaussian distributions, indexed $i\in\{1,2,\ldots,8\}$, with a mean of $(\cos(2\pi i/8),\ \sin(2\pi i/8))$ and variance 10^{-4}. We generate 50,000 training and 25,000 testing samples and the same number of 2D latent Gaussian noise vectors.

Both the D and G networks have 4 fully-connected layers with 200 and 400 units, respectively. We train for 400 epochs with a batch size of 128, and optimize with Adam [51] and a learning rate of 10^{-4} for both models. We consider three distinct settings that differ in the objective functions as: **(i)** (α_D,α_G)-GAN in (54); **(ii)** NS (α_D,α_G)-GAN's in (54a), (60); **(iii)** LSGAN with the 0-1 binary coding scheme (see Appendix N for details).

For every setting listed above, we train our models on the 2D-ring dataset for 200 random state seeds, where each seed contains different weight initializations for D and G. Ideally, a stable method will reflect similar performance across randomized initializations and also over training epochs; thus, we explore how GAN training performance for each setting varies across seeds and epochs. Our primary performance metric is *mode coverage*, defined as the number of Gaussians (0-8) that contain a generated sample within 3 standard deviations of its mean. A score of 8 conveys successful training, while a score of 0

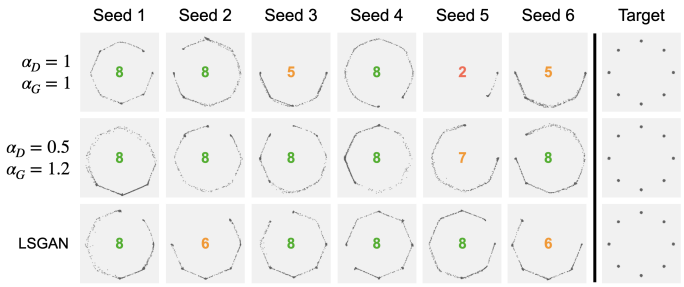


Fig. 10. Generated samples from two (α_D,α_G)-GANs trained with the **NS** objectives in (54a), (60), as well as the LSGAN. We provide 6 seeds to illustrate the stability in performance for each GAN across multiple runs.

conveys a significant GAN failure; on the other hand, a score in between 0 and 8 may be indicative of common GAN issues, such as mode collapse or failure to converge.

For the saturating setting, the improvement in stability of the $(0.2,1)$-GAN relative to the vanilla GAN is illustrated in Figure 9 as detailed in the caption. Vanilla GAN fails to converge to the true distribution 30% of the time while succeeding only 46% of the time. In contrast, the (α_D,α_G)-GAN with $\alpha_D<1$ learns a more stable G due to a less confident D (see also Figure 9(a)). For example, the $(0.3,1)$-GAN success and failure rates improve to 87% and 2%, respectively. For the NS setting in Figure 10, we find that tuning α_D and α_G yields more consistently stable outcomes than vanilla and LSGANs. Mode coverage rates over 200 seeds for saturating (Tables I and II) and NS (Table III) are in Appendix N.

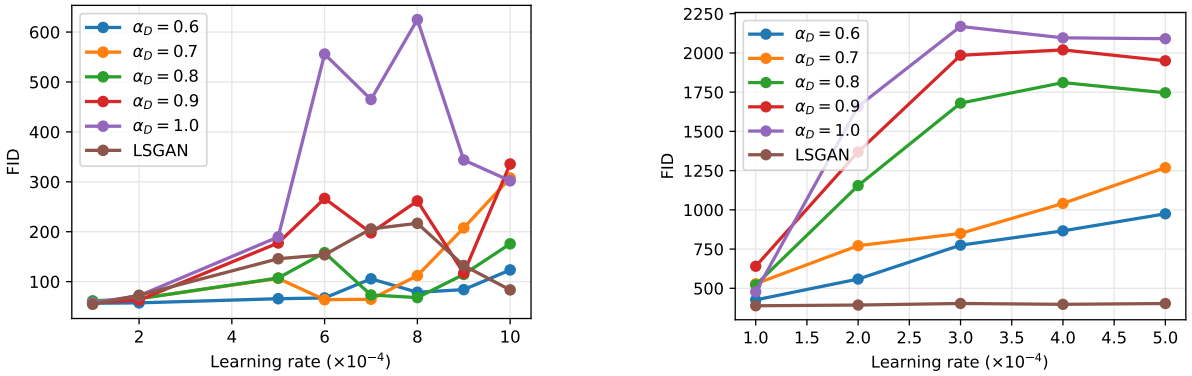
B. Celeb-A & LSUN Classroom

The Celeb-A dataset [49] is a widely recognized large-scale collection of over 200,000 celebrity headshots, encompassing images with diverse aspect ratios, camera angles, backgrounds, lighting conditions, and other variations. Similarly, the LSUN Classroom dataset [50] is a subset of the comprehensive Large-scale Scene Understanding (LSUN) dataset; it contains over 150,000 classroom images captured under diverse conditions and with varying aspect ratios. To ensure consistent input for the discriminator, we follow the standard practice of resizing the images to 64×64 for Celeb-A and 112×112 for LSUN Classroom. For both experiments, we randomly select 80% of the images for training and leave the remaining 20% for validation (evaluation of goodness metrics). Finally, for the generator, for each dataset, we generate a similar 80%-20% training-validation split of 100-dimensional latent Gaussian noise vectors, for a total matching the size of the true data.

For training, we employ the DCGAN architecture [47] that leverages deep convolutional neural networks (CNNs) for both D and G. In Appendix N, detailed descriptions of the D and G architectures can be found in Tables IV and V for the Celeb-A and LSUN Classroom datasets, respectively. Following SOTA methods, we focus on the non-saturating setting, utilizing appropriate objectives for vanilla GAN, (α_D,α_G)-GAN, and LSGAN. We consider a variety of learning rates, ranging from 10^{-4} to 10^{-3}, for Adam optimization. We evaluate our models every 10 epochs up to a total of 100 epochs and report the Fréchet Inception Distance (FID), an unsupervised similarity metric between the real and generated feature distributions extracted by InceptionNet-V3 [52]. For both datasets, we train each combination of objective function, number of epochs, and learning rate for 50 seeds. In the following subsections, we empirically demonstrate the dependence of the FID on learning rate and number of epochs for the vanilla GAN, (α_D,α_G)-GAN, and LSGAN. Achieving robustness to hyperparameter initialization is especially desirable in the unsupervised GAN setting as the choices that facilitate steady model convergence are not easily determined *a priori*.

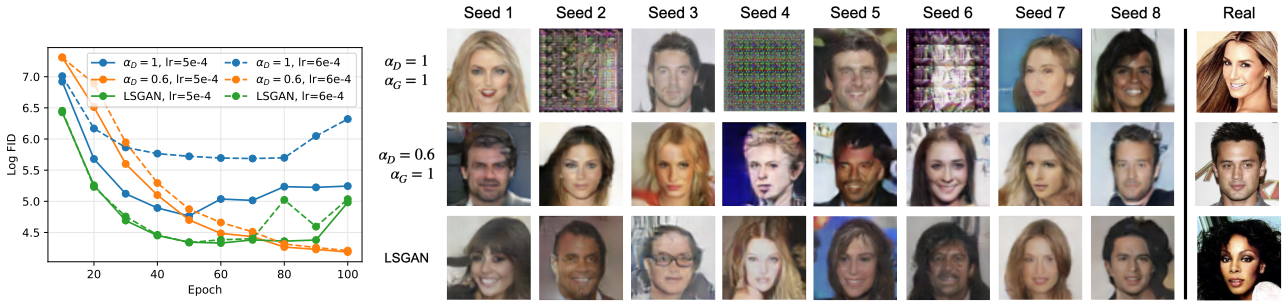
1) Celeb-A Results: In Figure 11(a), we examine the relationship between learning rate and FID for each GAN trained for 100 epochs on the Celeb-A dataset. When using learning rates of 1×10^{-4} and 2×10^{-4}, all GANs consistently perform well. However, when the learning rate increases,the vanilla $(1,1)$-GAN begins to exhibit instability across the 50 seeds. As the learning rate surpasses 5×10^{-4}, the performance of the vanilla GAN becomes even more erratic, underscoring the importance of GANs being robust to the choice of learning rate. Figure 11(a) also demonstrates that the GANs with $\alpha_D<1$ perform on par with, if not better than, the SOTA LSGAN. For instance, the $(0.6,1)$-GAN consistently achieves low FIDs across all tested learning rates.

In Figure 12(a), for different learning rates, we compare the dependence on the number of training epochs (hyperparameter) of the vanilla $(1,1)$-GAN, $(0.6,1)$-GAN, and LSGAN by plotting their FIDs every 10 epochs, up to 100 epochs, for two similar learning rates: 5×10^{-4} and 6×10^{-4}. We discover that the vanilla $(1,1)$-GAN performs significantly worse for the higher learning rate and deteriorates over time for both learning rates. Conversely, both the $(0.6,1)$-GAN and LSGAN consistently exhibit favorable FID performance for both learning rates. However, the $(0.6,1)$-GAN converges to a low FID, while the FID of the LSGAN slightly increases as training approaches 100 epochs. Finally, Fig. 12(b) displays a grid of generated Celeb-A faces, randomly sampled over 8 seeds for three GANs trained for 100 epochs with a learning rate of 5×10^{-4}. Here, we observe that the faces generated by the $(0.6,1)$-GAN and LSGAN exhibit a comparable level of quality to the rightmost column images,



(a) (b)

Fig. 11. (a) Plot of **Celeb-A** FID scores averaged over 50 seeds vs. learning rates for 6 different GANs, trained for 100 epochs. (b) Plot of **LSUN Classroom** FID scores averaged over 50 seeds vs. learning rates for 6 different GANs, trained for 100 epochs.



(a) (b)

Fig. 12. (a) Log-scale plot of **Celeb-A** FID scores over training epochs in steps of 10 up to 100 total, for three noteworthy GANs– $(1,1)$-GAN (vanilla), $(0.6,1)$-GAN, and LSGAN– and for two similar learning rates– 5×10^{-4} and 6×10^{-4}. Results show that the vanilla GAN performance is sensitive to learning rate choice, while the other two GANs achieve consistently low FIDs. (b) Generated Celeb-A faces from the same three GANs over 8 seeds when trained for 100 epochs with a learning rate of 5×10^{-4}. These samples show that the vanilla $(1,1)$-GAN training is sensitive to random model weight initializations, while the other two GANs demonstrate both robustness to random weight initializations as well as realistic face generation.

which are randomly sampled from the real Celeb-A dataset. On the other hand, the vanilla $(1,1)$-GAN shows clear signs of performance instability, as some seeds yield high-quality images while others do not.

2) LSUN Classroom Results: In Figure 11(b), we illustrate the relationship between learning rate and FID for GANs trained on the LSUN dataset for 100 epochs. In fact, when all GANs are trained with a learning rate of 1×10^{-4}, they consistently deliver satisfactory performance. However, increasing it to 2×10^{-4} leads to instability in the vanilla $(1,1)$-GAN across 50 seeds.

On the other hand, we observe that $\alpha_D<1$ contributes to stabilizing the FID across the 50 seeds even when trained with slightly higher learning rates. In Figure 11(b), we see that as α_D is tuned down to 0.6, the mean FIDs consistently decrease across all tested learning rates. These lower FIDs can be attributed to the increased stability of the network. Despite the gains in GAN stability achieved by tuning down α_D, Figure 11 demonstrates a noticeable disparity between the best (α_D,α_G)-GAN and the SOTA LSGAN. This suggests that there is still room for improvement in generating high-dimensional images with (α_D,α_G)-GANs.

In Appendix N, Figure 13(a), we illustrate the average FID throughout the training process for three GANs: $(1,1)$-GAN, $(0.6,1)$-GAN, and LSGAN, using two different learning rates: 1×10^{-4} and 2×10^{-4}. These findings validate that the vanilla $(1,1)$-GAN performs well when trained with the lower learning rate, but struggles significantly with the higher learning rate. In contrast, the $(0.6,1)$-GAN exhibits less sensitivity to learning rate, while the LSGAN achieves nearly identical scores for both learning rates. In Figure 13(b), we showcase the image quality generated by each GAN at epoch 100 with the higher learning rate. This plot highlights that the vanilla $(1,1)$-GAN frequently fails during training, whereas the $(0.6,1)$-GAN and LSGAN produce images that are more consistent in mimicking the real distribution. Finally, we present the FID vs. learning rate results for both datasets in Table VI in Appendix N. This allows yet another way to evaluate performance by comparing the percentage (out of 50 seeds) of FID scores below a desired threshold for each dataset, as detailed in the appendix.

LS - Comment on the results for the saturating setting that will be collated in the Appendix

VI. Conclusion

The conclusion goes here.

ACKNOWLEDGMENTS

This should be a simple paragraph before the References to thank those individuals and institutions who have supported your work on this article.

APPENDIX A
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 an increasing bijective link function $l:\mathbb{R}\to[0,1]$ as

$$\tilde{\ell}(t):=\ell(1,l(t)), \tag{73}$$

where the link l satisfies the following mild regularity conditions:

$$l(-t)=1-l(t), \tag{74}$$

$$l(0)=\frac{1}{2}, \tag{75}$$

$$l^{-1}(t)+l^{-1}(1-t)=0 \tag{76}$$

(e.g., sigmoid function, $\sigma(t)=1/(1+\mathrm{e}^{-t})$ satisfies this condition). Consider the inner optimization problem in (5) with the value function in (15) for this CPE loss ℓ.

$$\begin{aligned}
&\sup_{\omega}\int_{\mathcal{X}}(-p_r(x)\ell(1,D_\omega(x))-p_{G_\theta}(x)\ell(0,D_\omega(x)))\ dx\\
&=\int_{\mathcal{X}}\sup_{p_x\in[0,1]}(-p_r(x)\ell(1,p_x)-p_{G_\theta}(x)\ell(0,p_x))\ dx &(77)\\
&=\int_{\mathcal{X}}\sup_{p_x\in[0,1]}(-p_r(x)\ell(1,p_x)-p_{G_\theta}(x)\ell(1,1-p_x))\ dx &(78)\\
&=\int_{\mathcal{X}}\sup_{t_x\in\mathbb{R}}(-p_r(x)\ell(1,l(t_x))-p_{G_\theta}(x)\ell(1,1-l(t_x)))dx &(79)\\
&=\int_{\mathcal{X}}\sup_{t_x\in\mathbb{R}}(-p_r(x)\ell(1,l(t_x))-p_{G_\theta}(x)\ell(1,l(-t_x)))\ dx &(80)\\
&=\int_{\mathcal{X}}\sup_{t_x\in\mathbb{R}}(-p_r(x)\tilde{\ell}(t_x)-p_{G_\theta}(x)\tilde{\ell}(-t_x)\ dx &(81)\\
&=\int_{\mathcal{X}}p_{G_\theta}(x)\left(-\inf_{t_x\in\mathbb{R}}\left(\tilde{l}(-t_x)+\frac{p_r(x)}{p_{G_\theta}(x)}\tilde{l}(t_x)\right)\right)dx &(82)
\end{aligned}$$

where (78) follows because the CPE loss $\ell(y,\hat{y})$ is symmetric, (80) follows from (74), and (81) follows from the definition of the margin-based loss $\tilde{\ell}$ in (73). Now note that the function f defined as

$$f(u)=-\inf_{t\in\mathbb{R}}\left(\tilde{\ell}(-t)+u\tilde{\ell}(t)\right) \tag{83}$$

is convex since the infimum of affine functions is concave (observed earlier in [29] in a correspondence between margin-based loss functions and f-divergences). So, from (82), we get

$$\begin{aligned}
&\sup_{\omega}\int_{\mathcal{X}}(-p_r(x)\ell(1,D_\omega(x))-p_{G_\theta}(x)\ell(0,D_\omega(x)))\ dx\\
&\qquad=\int_{\mathcal{X}}p_{G_\theta}(x)f\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)\ dx &(84)\\
&\qquad=D_f(P_r\|P_{G_\theta}). &(85)
\end{aligned}$$

Thus, the resulting min-max optimization in (4) reduces to minimizing the f-divergence, $D_f(P_r\|P_{G_\theta})$ with f as given in (83).

For the converse statement, first note that given a symmetric f-divergence, it follows from [29, Theorem 1(b) and Corollary 3] that there exists a decreasing and convex margin-based loss function $\tilde{\ell}$ such that f can be expressed in the form (83). We may define an associated symmetric CPE loss $\ell(y,\hat{y})$ with

$$\ell(1,\hat{y}):=\tilde{\ell}(l^{-1}(\hat{y})), \tag{86}$$

where l^{-1} is the inverse of the same link function. Now repeating the steps as in (77)$-$(82), it is clear that the GAN based on this (symmetric) CPE loss results in minimizing the same symmetric f-divergence. It remains to verify that the symmetric CPE loss defined in (86) is such that $\ell(1,\hat{y})$ is decreasing so that the intuitive interpretation of vanilla GAN is retained and that it satisfies

(17) so that the optimal discriminator guesses uniformly at random when $P_r = P_{G_\theta}$. Note that $\ell'(1,\hat{y}) = \tilde{\ell}'(l^{-1}(\hat{y}))(l^{-1})'(\hat{y}) \leq 0$ since the margin-based loss $\tilde{\ell}$ is decreasing and the link function l (and hence its inverse) is increasing. So, $\ell(1,\hat{y})$ is decreasing. Observe that the loss function $\ell(1,\hat{y}) = \tilde{\ell}(l^{-1}(\hat{y}))$ may not be convex in y even though the margin-based loss function $\tilde{\ell}(\cdot)$ is convex. However, we show that the symmetric CPE loss associated with (86) indeed satisfies (17).

$$
\begin{aligned}
-\ell(1,t)-\ell(0,t) &= -\ell(1,t)-\ell(1,1-t) && (87)\\
&= -\tilde{\ell}(l^{-1}(t))-\tilde{\ell}(l^{-1}(1-t)) && (88)\\
&\leq -2\tilde{\ell}\left(\frac{1}{2}l^{-1}(t)+\frac{1}{2}l^{-1}(1-t)\right) && (89)\\
&= -2\tilde{\ell}(0) && (90)\\
&= -2\tilde{\ell}\left(l^{-1}\left(\frac{1}{2}\right)\right) && (91)\\
&= -\ell\left(1,\frac{1}{2}\right)-\ell\left(0,\frac{1}{2}\right), && (92)
\end{aligned}
$$

where (89) follows since the margin-based loss $\tilde{\ell}(\cdot)$ is convex, and (90) and (91) follow from (75) and (76), respectively.

APPENDIX B
PROOF OF THEOREM 2

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

$$
\sup_{\omega\in\Omega}\int_{\mathcal{X}} \frac{\alpha}{\alpha-1}\left(p_r(x)D_\omega(x)^{\frac{\alpha-1}{\alpha}}+p_{G_\theta}(x)(1-D_\omega(x))^{\frac{\alpha-1}{\alpha}}\right). \tag{93}
$$

Consider the function

$$
g(y) = \frac{\alpha}{\alpha-1}\left(ay^{\frac{\alpha-1}{\alpha}}+b(1-y)^{\frac{\alpha-1}{\alpha}}\right), \tag{94}
$$

for $a,b\in\mathbb{R}_+$ and $y\in[0,1]$. To show that the optimal discriminator is given by the expression in (23), 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

$$
y^* = \frac{a^\alpha}{a^\alpha+b^\alpha}. \tag{95}
$$

This gives (23). With this, the optimization problem in (22) can be written as $\inf_{\theta\in\Theta} C(G_\theta)$, where

$$
\begin{aligned}
C(G_\theta) &= \frac{\alpha}{\alpha-1}\times\\
&\left[\int_{\mathcal{X}}\left(p_r(x)D_{\omega^*}(x)^{\frac{\alpha-1}{\alpha}}+p_{G_\theta}(x)(1-D_{\omega^*}(x))^{\frac{\alpha-1}{\alpha}}\right)dx-2\right] && (96)\\
&= \frac{\alpha}{\alpha-1}\left[\int_{\mathcal{X}}\left(p_r(x)\left(\frac{p_r(x)^\alpha}{p_r(x)^\alpha+p_{G_\theta}(x)^\alpha}\right)^{\frac{\alpha-1}{\alpha}}+\right.\right.\\
&\quad \left.\left. p_{G_\theta}(x)\left(\frac{p_r(x)^\alpha}{p_r(x)^\alpha+p_{G_\theta}(x)^\alpha}\right)^{\frac{\alpha-1}{\alpha}}\right)dx-2\right] && (97)\\
&= \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}(p_r(x)^\alpha+p_{G_\theta}(x)^\alpha)^{\frac{1}{\alpha}}dx-2\right) && (98)\\
&= D_{f_\alpha}(P_r||P_{G_\theta})+\frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}}-2\right), && (99)
\end{aligned}
$$

where for the convex function f_α in (25),

$$
\begin{aligned}
D_{f_\alpha}(P_r||P_{G_\theta}) &= \int_{\mathcal{X}} p_{G_\theta}(x)f_\alpha\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx && (100)\\
&= \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}(p_r(x)^\alpha+p_{G_\theta}(x)^\alpha)^{\frac{1}{\alpha}}dx-2^{\frac{1}{\alpha}}\right). && (101)
\end{aligned}
$$

This gives us (24). 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 $P_r = P_{G_\theta}$.

APPENDIX C
PROOF OF THEOREM 3

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

$$\lim_{\alpha\to 1}\frac{\alpha}{\alpha-1}\left((a^{\alpha}+b^{\alpha})^{\frac{1}{\alpha}}-2^{\frac{1}{\alpha}-1}(a+b)\right)=a\log\left(\frac{a}{\frac{a+b}{2}}\right)+b\log\left(\frac{b}{\frac{a+b}{2}}\right). \tag{102}$$

Using this, we have

$$D_{f_1}(P_r||P_{G_\theta})\triangleq\lim_{\alpha\to 1}D_{f_\alpha}(P_r||P_{G_\theta}) \tag{103}$$

$$=\lim_{\alpha\to 1}\frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}(p_r(x)^{\alpha}+p_{G_\theta}(x)^{\alpha})^{\frac{1}{\alpha}}\,dx-2^{\frac{1}{\alpha}}\right) \tag{104}$$

$$=\lim_{\alpha\to 1}\Bigg[\frac{\alpha}{\alpha-1}\times \int_{\mathcal{X}}\left((p_r(x)^{\alpha}+p_{G_\theta}(x)^{\alpha})^{\frac{1}{\alpha}}-2^{\frac{1}{\alpha}-1}(p_r(x)+p_{G_\theta}(x))\right)dx\Bigg] \tag{105}$$

$$=\int_{\mathcal{X}}p_r(x)\log\frac{p_r(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx+ \int_{\mathcal{X}}p_{G_\theta}(x)\log\frac{p_{G_\theta}(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx \tag{106}$$

$$=:2D_{\text{JS}}(P_r||P_{G_\theta}), \tag{107}$$

where $D_{\text{JS}}(\cdot||\cdot)$ is the Jensen-Shannon divergence. Now, as $\alpha\to 1$, (24) equals $\inf_{\theta\in\Theta}2D_{\text{JS}}(P_r||P_{G_\theta})-\log 4$ recovering the vanilla GAN.

Substituting $\alpha=\frac{1}{2}$ in (26), we get

$$D_{f_{\frac{1}{2}}}(P_r||P_{G_\theta})=-\int_{\mathcal{X}}\left(\sqrt{p_r(x)}+\sqrt{p_{G_\theta}(x)}\right)^2dx+4 \tag{108}$$

$$=\int_{\mathcal{X}}\left(\sqrt{p_r(x)}-\sqrt{p_{G_\theta}(x)}\right)^2dx \tag{109}$$

$$=:2D_{\text{H}^2}(P_r||P_{G_\theta}), \tag{110}$$

where $D_{\text{H}^2}(P_r||P_{G_\theta})$ is the squared Hellinger distance. For $\alpha=\frac{1}{2}$, (24) 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

$$D_{f_1}(P_r||P_{G_\theta})\triangleq\lim_{\alpha\to\infty}D_{f_\alpha}(P_r||P_{G_\theta}) \tag{111}$$

$$=\lim_{\alpha\to\infty}\frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}}(p_r(x)^{\alpha}+p_{G_\theta}(x)^{\alpha})^{\frac{1}{\alpha}}\,dx-2^{\frac{1}{\alpha}}\right) \tag{112}$$

$$=\int_{\mathcal{X}}\max\{p_r(x),p_{G_\theta}(x)\}\;dx-1 \tag{113}$$

$$=\int_{\mathcal{X}}\max\{p_r(x)-p_{G_\theta}(x),0\}\;dx \tag{114}$$

$$=\int_{\mathcal{A}}(p_r(x)-p_{G_\theta}(x))\;dx \tag{115}$$

$$=\int_{\mathcal{A}}\frac{p_r(x)-p_{G_\theta}(x)}{2}\;dx+\int_{\mathcal{A}^c}\frac{p_{G_\theta}(x)-p_r(x)}{2}\;dx \tag{116}$$

$$=\frac{1}{2}\int_{\mathcal{X}}|p_r(x)-p_{G_\theta}(x)|\;dx \tag{117}$$

$$=:D_{\text{TV}}(P_r||P_{G_\theta}), \tag{118}$$

where $D_{\text{TV}}(P_r||P_{G_\theta})$ is the total variation distance between P_r and P_{G_θ}. Thus, as $\alpha\to\infty$, (24) equals $\inf_{\theta\in\Theta}D_{\text{TV}}(P_r||P_{G_\theta})-1$ recovering TV-GAN (modulo a constant).

APPENDIX D
PROOF OF THEOREM 4

We first derive the Fenchel conjugate f_α^* of f_α as follows:

$$f_\alpha^*(t)=\sup_u\left(ut-f_\alpha(u)\right)=\frac{\alpha}{\alpha-1}\sup_u\left(1+\left(1+\frac{\alpha-1}{\alpha}t\right)u-(1+u^\alpha)^{\frac{1}{\alpha}}\right). \tag{119}$$

The optimum u_* is obtained by setting the derivative of $ut-f_\alpha(u)$ to zero, yielding

$$1+\frac{\alpha-1}{\alpha}t=u_*^{\alpha-1}(1+u_*^\alpha)^{\frac{1}{\alpha}-1}=\left(\frac{u_*^\alpha}{1+u_*^\alpha}\right)^{\frac{\alpha-1}{\alpha}}, \tag{120}$$

i.e.,

$$u_*=u_*(t)=\left(\frac{s(t)}{1-s(t)}\right)^{\frac{1}{\alpha}} \tag{121}$$

with

$$s(t)=\left(1+\frac{\alpha-1}{\alpha}t\right)^{\frac{\alpha}{\alpha-1}}. \tag{122}$$

The verification that u_* is a global maximizer over $u\geq 0$ follows from

$$(ut-f_\alpha(u))''=-\left(\frac{u^\alpha}{1+u^\alpha}\right)^{\alpha-1}(1+u^\alpha)^{-2}\alpha u^{\alpha-1}<0$$

for all $u>0$. The relations (120) and (121) then lead to

$$\begin{aligned}f_\alpha^*(t)=u_*(t)t-f_\alpha(u_*(t))&=\frac{\alpha}{\alpha-1}\left(1+\left(1+\frac{\alpha-1}{\alpha}t\right)u_*(t)-(1+u_*(t)^\alpha)^{\frac{1}{\alpha}}\right)\\&=\frac{\alpha}{\alpha-1}\left(1-(1+u_*(t)^\alpha)^{\frac{1}{\alpha}-1}\right)\\&=\frac{\alpha}{\alpha-1}\left(1-(1-s(t))^{\frac{\alpha-1}{\alpha}}\right),\end{aligned} \tag{123}$$

where s is given by (122). The domain $\mathrm{dom}(f_\alpha^*)$ consists of values t such that $1+\frac{\alpha-1}{\alpha}t\geq 0$ and $s(t)\leq 1$, i.e., $t\in[-\frac{\alpha}{\alpha-1},0]$ for $\alpha>1$ and $t\leq 0$ for $\alpha\in(0,1)$. Also note that

$$f_1^*(t)=\lim_{\alpha\to 1}f_\alpha^*(t)=\lim_{\alpha\to 1}\frac{\alpha}{\alpha-1}\left(1-(1-s(t))^{\frac{\alpha-1}{\alpha}}\right)=-\log(1-e^t)$$

for $t\leq 0$, where s is again given by (122).

In the following we consider $\alpha\neq 1$ with results also valid for $\alpha=1$ by continuity. Let $v\in\overline{\mathbb{R}}$ and consider

$$d=s(g_{f_\alpha}(v))=\left(1+\frac{\alpha-1}{\alpha}g_{f_\alpha}(v)\right)^{\frac{\alpha}{\alpha-1}}. \tag{124}$$

We first show that $d\in[0,1]$ and then show that (30) is satisfied.

If $\alpha>1$, then $g_{f_\alpha}(v)\in[-\frac{\alpha}{\alpha-1},0]=\mathrm{dom}(f_\alpha^*)$. Therefore, $d\in[0,1]$. If $\alpha\in[0,1]$, then $g_{f_\alpha}(v)\in[-\infty,0]=\mathrm{dom}(f_\alpha^*)$. Therefore, $\left(1+\frac{\alpha-1}{\alpha}g_{f_\alpha}(v)\right)\in[1,\infty]$, and hence $d\in[0,1]$.

Using (124),

$$\ell_\alpha(1,d)=\frac{\alpha}{\alpha-1}\left(1-d^{\frac{\alpha-1}{\alpha}}\right)=\frac{\alpha}{\alpha-1}\left(1-s(g_{f_\alpha}(v))^{\frac{\alpha-1}{\alpha}}\right)=-g_{f_\alpha}(v),$$

and

$$\ell_\alpha(0,d)=\frac{\alpha}{\alpha-1}\left(1-(1-d)^{\frac{\alpha-1}{\alpha}}\right)=\frac{\alpha}{\alpha-1}\left(1-(1-s(g_{f_\alpha}(v)))^{\frac{\alpha-1}{\alpha}}\right)=f_\alpha^*(g_{f_\alpha}(v)).$$

Conversely, let $d\in[0,1]$ and consider

$$v=g_{f_\alpha}^{-1}(-\ell_\alpha(1,d))=g_{f_\alpha}^{-1}\left(\frac{\alpha}{\alpha-1}(d^{\frac{\alpha-1}{\alpha}}-1)\right). \tag{125}$$

We first show that $v\in\overline{\mathbb{R}}$ and then show that (30) is satisfied.

If $\alpha>1$, then $-\ell_\alpha(1,d)\in[-\frac{\alpha}{\alpha-1},0]=\mathrm{dom}(f_\alpha^*)$. Therefore, $v\in\overline{\mathbb{R}}$. If $\alpha\in[0,1]$, then $d^{\frac{\alpha-1}{\alpha}}\in[0,\infty]$ and $-\ell_\alpha(1,d)\in[-\infty,0]=\mathrm{dom}(f_\alpha^*)$. Hence, $v\in\overline{\mathbb{R}}$.

Using (125),

$$g_{f_\alpha}(v)=-\ell_\alpha(1,d),$$

and

$$s(g_{f_\alpha}(v))=\left(1+\frac{\alpha-1}{\alpha}g_{f_\alpha}(v)\right)^{\frac{\alpha}{\alpha-1}}=\left(1+\frac{\alpha-1}{\alpha}\left(\frac{\alpha}{\alpha-1}(d^{\frac{\alpha-1}{\alpha}}-1)\right)\right)^{\frac{\alpha}{\alpha-1}}=d,$$

so that

$$f_\alpha^*(g_{f_\alpha}(v))=\frac{\alpha}{\alpha-1}\left(1-(1-s(g_{f_\alpha}(v)))^{\frac{\alpha-1}{\alpha}}\right)=\frac{\alpha}{\alpha-1}\left(1-(1-d)^{\frac{\alpha-1}{\alpha}}\right)=\ell_\alpha(0,d).$$

APPENDIX E
PROOF OF COROLLARY 1

For $Q_\omega\in A$ define $D_\omega\in B$ such that $d=D_\omega(x)$ is obtained from (124) with $v=Q_\omega(x)$ for all $x\in\mathcal{X}$. By Theorem 4, $g(Q_\omega)=h(D_\omega)$. Conversely, for $D_\omega\in B$ define $Q_\omega\in A$ such that $v=Q_\omega(x)$ is obtained from (125) with $d=D_\omega(x)$ for all $x\in\mathcal{X}$. Again by Theorem 4, $h(D_\omega)=g(V_\omega)$.

To show that k is bijective, we first show that $s:\text{dom}(f_\alpha^*)\to[-\infty,1]$ defined in (122) is bijective. Let the function $s^{-1}:[-\infty,1]\to\text{dom}(f_\alpha^*)$ be defined by $s^{-1}(u)=\frac{\alpha}{\alpha-1}(u^{\frac{\alpha-1}{\alpha}}-1)$. Let $t\in\text{dom}(f_\alpha^*)$. Then

$$s^{-1}(s(t))=\frac{\alpha}{\alpha-1}\left[\left(\left(1+\frac{\alpha-1}{\alpha}t\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}}-1\right]=t.$$

Now, let $u\in[-\infty,1]$. Then

$$s(s^{-1}(u))=\left(1+\frac{\alpha-1}{\alpha}\left(\frac{\alpha}{\alpha-1}\left(u^{\frac{\alpha-1}{\alpha}}-1\right)\right)\right)^{\frac{\alpha}{\alpha-1}}=u.$$

Therefore, s^{-1} is the inverse of s, and hence s is bijective. As the composition of two bijective functions, k is also bijective.

APPENDIX F
PROOF OF THEOREM 5

As noted in the proof of Theorem 1, given a symmetric f-divergence, it follows from [29, Theorem 1(b) and Corollary 3] that there exists a CPE (partial) loss ℓ such that

$$f(u)=\sup_{t\in[0,1]}-u\ell(t)-\ell(1-t). \tag{126}$$

We assume that the loss l is strictly convex as mentioned in the theorem statement. Note that

$$f(u)=\sup_{v\in\text{dom}f^*}uv-f^*(v). \tag{127}$$

Noticing that the inner optimization problems in the CPE loss GAN and f-GAN formulations reduce to pointwise optimizations (126) and (127), respectively, it suffices to show that the variational forms of f in (126) and (127) are equivalent. To this end, we show that (126) is equivalent to the optimization problem

$$f(u)=\sup_{v\in\mathbb{R}_+}uf'(v)-[vf'(v)-f(v)] \tag{128}$$

which is known to be equivalent to (127) [38]. Let $k:\mathbb{R}_+\to[0,1]$ denote the bijective mapping from $u\in\mathbb{R}_+$ to the optimizer in (126). Fix a $v\in\mathbb{R}_+$. We have

$$f(v)=-v\ell(k(v))-\ell(1-k(v)), \tag{129}$$

and also note that

$$f'(v)=-\ell(k(v)). \tag{130}$$

Consider

$$vf'(v)-f(v)=-v\ell(k(v))+v\ell(k(v))+\ell(1-k(v)) \tag{131}$$

$$=\ell(1-k(v)), \tag{132}$$

where (131) follows from (130) and (129). Thus, with the change of variable $t=k(v)$, the objective function in (128) is equal to that of (126). Since the function k is invertible, for a fixed $t\in[0,1]$, we can also show that the change of variable $v=k^{-1}(t)$ in the objective function of (126) gives the objective function of (128).

APPENDIX G
PROOF OF THEOREM 6

Without loss of generality we take the functions f_1 and f_2 to be non-negative using the fact that $D_f(\cdot\|\cdot)=D_{f'}(\cdot\|\cdot)$ whenever $f'(x)=f(x)+c(x-1)$, for some $c\in\mathbb{R}$ (see [12, Theorem 2]). Note that it suffices to show that any symmetric f-divergence $D_f(\cdot\|\cdot)$ is equivalent to $D_{\text{TV}}(\cdot\|\cdot)$, i.e., $D_f(P_n||P)\to 0$ 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 any symmetric f-divergence which gives lower and upper bounds on it in terms of the total variation distance, D_{TV}. In particular, Feldman and Österreicher [**?**, Theorem 2] proved that for any symmetric f-divergence D_f, probability distributions P and Q, we have

$$\gamma_f(D_{\text{TV}}(P||Q))\le D_f(P||Q)\le\gamma_f(1)D_{\text{TV}}(P||Q), \tag{133}$$

where the function $\gamma_\alpha:[0,1]\to[0,\infty)$ defined by $\gamma_f(x)=(1+x)f\left(\frac{1-x}{1+x}\right)$ is convex, strictly increasing and continuous on $[0,1]$ such that $\gamma_f(0)=0$ and $\gamma_f(1)=2f(0)$.

We first prove the 'only if' part, i.e., $D_f(P_n||P)\to 0$ as $n\to\infty$ implies $D_{\text{TV}}(P_n||P)\to 0$ as $n\to\infty$. Suppose $D_f(P_n||P)\to 0$. From the lower bound in (133), it follows that $\gamma_f(D_{\text{TV}}(P_n||P))\le D_f(P_n||P)$, for each $n\in\mathbb{N}$. This implies that $\gamma_f(D_{\text{TV}}(P_n||P))\to 0$ as $n\to\infty$. We show below that γ_f is invertible and γ_f^{-1} is continuous. Then it would follow that $\gamma_f^{-1}\gamma_f(D_{\text{TV}}(P_n||P))=D_{\text{TV}}(P_n||P)\to\gamma_f^{-1}(0)=0$ as $n\to\infty$ proving that Arimoto divergence is stronger than the total variation distance. It remains to show that γ_f is invertible and γ_f^{-1} is continuous. Invertibility follows directly from the fact that γ_f is strictly increasing function. For the continuity of γ_α^{-1}, it suffices to show that $\gamma_f(C)$ 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. Now since γ_f is continuous, $\gamma_f(C)$ is compact because a continuous function of a compact set is compact. By Heine-Borel theorem, this gives that $\gamma_f(C)$ 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 $D_f(P_n||P)\to 0$. It follows from the upper bound in (133) that $D_f(P_n||P)\le D_{\text{TV}}(P_n||P)$, for each $n\in\mathbb{N}$. This implies that $D_f(P_n||P)\to 0$ as $n\to\infty$ which completes the proof.

APPENDIX H
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 $D_{\text{JS}}(P_n||P)\le(\log_e 2)D_{\text{TV}}(P_n||P)$, 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 I
PROOF OF THEOREM 7

The proof is along similar lines as that of [26, Theorem 3.1]. Below we argue that, with high probability, for every discriminator D_ω,

$$\left|\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]-\mathbb{E}_{X\sim P_{G_\theta}}[\phi(D_\omega(X))]\right|\le\frac{\epsilon}{2}, \tag{134}$$

$$\left|\mathbb{E}_{X\sim P_r}[\psi(D_\omega(X))]-\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))]\right|\le\frac{\epsilon}{2}. \tag{135}$$

Assuming ω^* to be an optimizer attaining $\tilde{d}_{\mathcal{F}}(P_r,P_{G_\theta})$, it would then follow that

$$\begin{aligned}
&\tilde{d}_{\mathcal{F}}(\hat{P}_r,\hat{P}_{G_\theta})\\
&=\sup_{\omega\in\Omega}\left|\mathbb{E}_{X\sim\hat{P}_r}[\phi(D_\omega(X))]+\mathbb{E}_{X\sim\hat{P}_{G_\theta}}[\psi(D_\omega(X))]\right| &(136)\\
&\ge\left|\mathbb{E}_{X\sim\hat{P}_r}[\phi(D_{\omega^*}(X))]+\mathbb{E}_{X\sim\hat{P}_{G_\theta}}[\psi(D_{\omega^*}(X))]\right| &(137)\\
&\ge\left|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*}(X))]+\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_{\omega^*}(X))]\right|\\
&\quad-\left|\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]-\mathbb{E}_{X\sim\hat{P}_r}[\phi(D_\omega(X))]\right|\\
&\quad-\left|\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))]-\mathbb{E}_{X\sim\hat{P}_{G_\theta}}[\psi(D_\omega(X))]\right| &(138)\\
&\ge\tilde{d}_{\mathcal{F}}(P_r,P_G)-\epsilon, &(139)
\end{aligned}$$

where (138) follows from the triangle inequality, (139) follows from (134) and (135). Similarly, we can prove the other direction, i.e., $\tilde{d}_{\mathcal{F}}(\hat{P}_r,\hat{P}_G)\le\tilde{d}_{\mathcal{F}}(P_r,P_G)+\epsilon$, which implies (43).

It remains to argue for the concentration bounds (134) and (135). Recall that the concentration bound in (134) was proved in [26, Proof of Theorem 3.1] by considering a $\frac{\epsilon}{8LL_\phi}$-net in Ω and leveraging the Lipschitzianity of the discriminator class $\mathcal{F}$ and the function ϕ. Using the exact same analysis, the concentration bound in (135) can be proved separately by considering a $\frac{\epsilon}{8LL_\psi}$-net. For both the bounds to hold simultaneously, it suffices to consider a $\frac{\epsilon}{8L\max\{L_\phi,L_\psi\}}$-net along the same lines as the last part of [26, Proof of Theorem 3.1], thus completing the proof.

APPENDIX J
PROOF OF THEOREM 8

We upper bound the estimation error in terms of the Rademacher complexities of appropriately defined *compositional* classes building upon the proof techniques of [42, Theorem 1]. We then bound these Rademacher complexities using a contraction lemma [46, Lemma 26.9]. Details are in order.

We first review the notion of Rademacher complexity.

Definition 3 (Rademacher complexity)**.** *Let $\mathcal{G}_\Omega := \{g_\omega : \mathcal{X} \to \mathbb{R} \,|\, \omega \in \Omega\}$ and $S = \{X_1 \ldots, X_n\}$ 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*

$$\mathcal{R}_S(\mathcal{G}_\Omega) = \mathbb{E}_{X,\epsilon} \sup_{\omega \in \Omega} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g_\omega(x_i) \right| \tag{140}$$

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

We write our discriminator model in (69) in the form

$$D_\omega(x) = \sigma(f_\omega(x)), \tag{141}$$

where f_ω is exactly the same discriminator model defined in [42, Equation (26)]. Now by following the similar steps as in [42, 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

$$\begin{aligned} & d^{(\ell)}_{\mathcal{F}_{nn}}(P_r, \hat{P}_{G_{\hat{\theta}^*}}) - \inf_{\theta \in \Theta} d^{(\ell)}_{\mathcal{F}_{nn}}(P_r, P_{G_\theta}) \\ & \leq 2 \sup_{\omega} \left| \mathbb{E}_{X \sim P_r} \phi(D_\omega(X)) - \frac{1}{n} \sum_{i=1}^{n} \phi(D_\omega(X_i)) \right| \\ & \quad + 2 \sup_{\omega, \theta} \left| \mathbb{E}_{Z \sim P_Z} \psi(D_\omega(g_\theta(Z))) - \frac{1}{m} \sum_{j=1}^{m} \psi(D_\omega(g_\theta(Z_j))) \right| \end{aligned} \tag{142}$$

Let us denote the supremums in the first and second terms in (142) by $F^{(\phi)}(X_1, \ldots, X_n)$ and $G^{(\psi)}(Z_1, \ldots, Z_m)$, respectively. We next bound $G^{(\psi)}(Z_1, \ldots, Z_m)$. 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 $z_1, \ldots, z_j, \ldots, z_m, z'_j$, using $\sup_r |h_1(r)| - \sup_r |h_2(r)| \leq \sup_r |h_1(r) - h_2(r)|$, we have

$$\begin{aligned} & G^{(\psi)}(z_1, \ldots, z_j, \ldots, z_m) - G^{(\psi)}(z_1, \ldots, z'_j, \ldots, z_m) \\ & \leq \sup_{\omega, \theta} \frac{1}{m} \left| \psi(D_\omega(g_\theta(z_j))) - \psi(D_\omega(g_\theta(z'_j))) \right| && (143) \\ & \leq \sup_{\omega, \theta} \frac{1}{m} \left| \psi(\sigma(f_\omega(g_\theta(z_j)))) - \psi(\sigma(f_\omega(g_\theta(z'_j)))) \right| && (144) \\ & \leq \frac{L_\psi}{4} \sup_{\omega, \theta} \frac{1}{m} \left| \sigma(f_\omega(g_\theta(z_j))) - \sigma(f_\omega(g_\theta(z'_j))) \right| && (145) \\ & \leq \frac{L_\psi}{4} \frac{2}{m} \left(M_k \prod_{i=1}^{k-1} (M_i R_i) \right) \left(N_l \prod_{j=1}^{l-1} (N_j S_j) \right) B_z && (146) \\ & = \frac{L_\psi Q_z}{2m}, && (147) \end{aligned}$$

where (144) follows from (141), (145) follows because $\psi(\sigma(\cdot))$ is $\frac{L_\psi}{4}$-Lipschitz, (146) follows by using the Cauchy-Schwarz inequality and the fact that $||Ax||_2 \leq ||A||_F ||x||_2$ (as observed in [42]), and (147) follows by defining

$$Q_z := \left(M_k \prod_{i=1}^{k-1} (M_i R_i) \right) \left(N_l \prod_{j=1}^{l-1} (N_j S_j) \right) B_z. \tag{148}$$

Using (147), the McDiarmid's inequality [46, Lemma 26.4] implies that, with probability at least $1-\delta$,

$$\begin{aligned}&G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m)\\&\leq \mathbb{E}_Z G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m)+\frac{L_\psi Q_z}{2}\sqrt{\log\frac{1}{\delta}/(2m)}.\end{aligned} \tag{149}$$

Following the standard steps similar to [42, Equation (20)], the expectation term in (149) can be upper bounded as

$$\begin{aligned}&\mathbb{E}_Z G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m)\\&\qquad\leq 2\mathbb{E}_{Z,\epsilon}\sup_{\omega,\theta}\left|\frac{1}{m}\sum_{j=1}^{m}\epsilon_j\psi(D_\omega(g_\theta(Z_j)))\right| &(150)\\&\qquad=: 2\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta}) &(151)\end{aligned}$$

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

$$G^{(\psi)}(Z_1,\ldots,Z_j,\ldots,Z_m)\leq 2\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta})+\sqrt{\log\frac{1}{\delta}}\frac{L_\psi Q_z}{2\sqrt{2m}}. \tag{152}$$

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

$$F^{(\phi)}(X_1,\ldots,X_n)\leq 2\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega})+\sqrt{\log\frac{1}{\delta}}\frac{L_\phi Q_x}{2\sqrt{2n}}, \tag{153}$$

where

$$\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega}):=\mathbb{E}_{X,\epsilon}\sup_{\omega}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi(D_\omega(X_i))\right|. \tag{154}$$

Combining (142), (152), and (153) using a union bound, we get, with probability at least $1-2\delta$,

$$\begin{aligned}&d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat{\theta}^*}})-\inf_{\theta\in\Theta}d^{(\ell)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta})\\&\qquad\leq 4\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega})+4\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta})\\&\qquad\quad+\sqrt{\log\frac{1}{\delta}}\left(\frac{L_\phi Q_x}{\sqrt{2n}}+\frac{L_\psi Q_z}{\sqrt{2m}}\right).\end{aligned} \tag{155}$$

Now we bound the Rademacher complexities in the RHS of (155). 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, [46])**.** *For each $i\in\{1,\ldots,n\}$, let $\gamma_i:\mathbb{R}\to\mathbb{R}$ be a ρ-Lipschitz function. Then, for $A\subset\mathbb{R}^n$,*

$$\mathcal{R}(\gamma\circ A)\leq\rho\mathcal{R}(A), \tag{156}$$

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

$$\begin{aligned}&\mathcal{R}_{S_x}(\mathcal{F}^{(\phi)}_{\Omega})\\&=\mathbb{E}_X[\mathcal{R}(\{(\phi(D_\omega(X_1)),\ldots,\phi(D_\omega(X_n))):\omega\in\Omega\})] &(157)\\&=\mathbb{E}_X[\mathcal{R}(\{(\phi(\sigma(f_\omega(X_1))),\ldots,\phi(\sigma(f_\omega(X_n)))):\omega\in\Omega\})] &(158)\\&\leq\frac{L_\phi}{4}\mathbb{E}_X[\mathcal{R}(\{(f_\omega(X_1),\ldots,(f_\omega(X_n)):\omega\in\Omega\})] &(159)\\&\leq\frac{L_\phi Q_x\sqrt{3k}}{4\sqrt{n}} &(160)\end{aligned}$$

where (158) follows from (141), (159) follows from Lemma 1 by substituting $\gamma(\cdot)=\phi(\sigma(\cdot))$, and (160) follows from [42, Proof of Corollary 1]. Using a similar approach, we obtain

$$\mathcal{R}_{S_z}(\mathcal{H}^{(\psi)}_{\Omega\times\Theta})\leq\frac{L_\psi Q_z\sqrt{3(k+l-1)}}{4\sqrt{m}}. \tag{161}$$

Substituting (160) and (161) into (155) gives (50).

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 [23, 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 (72). Now using the Cauchy-Schwarz inequality and the fact that $||Ax||_2\le||A||_F||x||_2$, it follows that

$$|f_\omega(\cdot)|\le Q_x, \tag{162}$$

$$|f_\omega(g_\theta(\cdot))|\le Q_z, \tag{163}$$

where $Q_x:=M_k\prod_{i=1}^{k-1}(M_iR_i)B_x$ and with Q_z as in (148). 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 (145) and (159) 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 (50):

$$\begin{aligned}
&d^{(\ell_\alpha)}_{\mathcal{F}_{nn}}(P_r,\hat{P}_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta}d^{(\ell_\alpha)}_{\mathcal{F}_{nn}}(P_r,P_{G_\theta})\\
&\le\frac{4C_{Q_x}(\alpha)Q_x\sqrt{3k}}{\sqrt{n}}+\frac{4C_{Q_z}(\alpha)Q_z\sqrt{3(k+l-1)}}{\sqrt{m}}\\
&\quad+2\sqrt{2\log\frac{1}{\delta}}\left(\frac{C_{Q_x}(\alpha)Q_x}{\sqrt{n}}+\frac{C_{Q_z}(\alpha)Q_z}{\sqrt{m}}\right).
\end{aligned}\tag{164}$$

APPENDIX K
PROOF OF THEOREM 9

Let $\phi(\cdot)=-\ell_\alpha(1,\cdot)$ and consider the following modified version of $d^{\ell_\alpha}_{\mathcal{F}_{nn}}(\cdot,\cdot)$ (defined in [8, eq. (13)]):

$$\begin{aligned}
&d^{\ell_\alpha}_{\mathcal{F}_{nn}}(P,Q)=\\
&\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[\phi(D_\omega(X))]+\mathbb{E}_{X\sim Q}[\phi(1-D_\omega(X))]\Big)-2\phi(1/2),
\end{aligned}$$

where

$$D_\omega(x)=\sigma\left(\mathbf{w}_k^{\mathsf{T}}r_{k-1}(\mathbf{W}_{d-1}r_{k-2}(\ldots r_1(\mathbf{W}_1(x))))\right):=\sigma(f_\omega(x)).$$

Taking $\alpha\to\infty$, we obtain

$$d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big). \tag{165}$$

We first prove that $d^{\ell_\infty}_{\mathcal{F}_{nn}}$ is a semi-metric.

Claim 1: For any distribution pair (P,Q), $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)\ge0$.

Proof. Consider a discriminator which always outputs 1/2, i.e., $D_\omega(x)=1/2$ for all x. Note that such a neural network discriminator exists, as setting $\mathbf{w}_k=0$ results in $D_\omega(x)=\sigma(0)=0$. For this discriminator, the objective function in (165) evaluates to $1/2-1/2=0$. Since $d^{\ell_\infty}_{\mathcal{F}_{nn}}$ is a supremum over all discriminators, we have $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)\ge0$.

Claim 2: For any distribution pair (P,Q), $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)=d^{\ell_\infty}_{\mathcal{F}_{nn}}(Q,P)$.

Proof.

$$\begin{aligned}
&d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)\\
&=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big)\\
&=\sup_{\mathbf{W}_1,\ldots,\mathbf{w}_k}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big)\\
&\overset{(i)}{=}\sup_{\mathbf{W}_1,\ldots,-\mathbf{w}_k}\Big(\mathbb{E}_{X\sim P}[\sigma(-f_\omega(x))]-\mathbb{E}_{X\sim Q}[\sigma(-f_\omega(x))]\Big)\\
&\overset{(ii)}{=}\sup_{\mathbf{W}_1,\ldots,\mathbf{w}_k}\Big(\mathbb{E}_{X\sim P}[1-\sigma(f_\omega(x))]-\mathbb{E}_{X\sim Q}[1-\sigma(f_\omega(x))]\Big)\\
&=\sup_{\mathbf{W}_1,\ldots,\mathbf{w}_k}\Big(\mathbb{E}_{X\sim Q}[\sigma(f_\omega(x))]-\mathbb{E}_{X\sim P}[\sigma(f_\omega(x))]\Big)\\
&=d^{\ell_\infty}_{\mathcal{F}_{nn}}(Q,P),
\end{aligned}$$

where (i) follows from replacing $\mathbf{w}_k$ with $-\mathbf{w}_k$ and (ii) follows from the sigmoid property $\sigma(-x)=1-\sigma(x)$ for all x.

Claim 3: For any distribution P, $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,P)=0$.

Proof.

$$d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,P)=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim P}[D_\omega(X)]\Big)=0.$$

Claim 4: For any distributions P,Q,R, $d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)\le d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,R)+d^{\ell_\infty}_{\mathcal{F}_{nn}}(R,Q)$.

Proof.

$$\begin{aligned}
&d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,Q)\\
&=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big)\\
&=\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim R}[D_\omega(X)]\\
&\qquad\qquad+\mathbb{E}_{X\sim R}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big)\\
&\le\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim P}[D_\omega(X)]-\mathbb{E}_{X\sim R}[D_\omega(X)]\Big)\\
&\quad+\sup_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim R}[D_\omega(X)]-\mathbb{E}_{X\sim Q}[D_\omega(X)]\Big)\\
&=d^{\ell_\infty}_{\mathcal{F}_{nn}}(P,R)+d^{\ell_\infty}_{\mathcal{F}_{nn}}(R,Q).
\end{aligned}$$

Thus, $d^{\ell_\infty}_{\mathcal{F}_{nn}}$ is a semi-metric. The remaining part of the proof of the lower bound follows along the same lines as that of [42, Theorem 2] by an application of Fano's inequality [53, Theorem 2.5] (that requires the involved divergence measure to be a semi-metric), replacing $d_{\mathcal{F}_{nn}}$ with $d^{\ell_\infty}_{\mathcal{F}_{nn}}$ and noting that the additional sigmoid activation function after the last layer in the discriminator satisfies the monotonicity assumption so that $C(\mathcal{P}(\mathcal{X}))>0$ (for $C(\mathcal{P}(\mathcal{X}))$ defined in (52)).

APPENDIX L
PROOF OF THEOREM 10

A. Proof of (a)

The proof to obtain (55) is the same as that for Theorem 2, where $\alpha=\alpha_D$. The generator's optimization problem in (54b) with the optimal discriminator in (55) can be written as $\inf_{\theta\in\Theta}V_{\alpha_G}(\theta,\omega^*)$, where

$$\begin{aligned}
&V_{\alpha_G}(\theta,\omega^*)\\
&=\frac{\alpha_G}{\alpha_G-1}\times\\
&\left[\int_{\mathcal{X}}\Big(p_r(x)D_{\omega^*}(x)^{\frac{\alpha_G-1}{\alpha_G}}+p_{G_\theta}(x)(1-D_{\omega^*}(x))^{\frac{\alpha_G-1}{\alpha_G}}\Big)dx-2\right]\\
&=\frac{\alpha_G}{\alpha_G-1}\left[\int_{\mathcal{X}}\left(p_r(x)\left(\frac{p_r(x)^{\alpha_D}}{p_r(x)^{\alpha_D}+p_{G_\theta}(x)^{\alpha_D}}\right)^{\frac{\alpha_G-1}{\alpha_G}}\right.\right.\\
&\qquad\left.\left.+p_{G_\theta}(x)\left(\frac{p_r(x)^{\alpha_D}}{p_r(x)^{\alpha_D}+p_{G_\theta}(x)^{\alpha_D}}\right)^{\frac{\alpha_G-1}{\alpha_G}}\right)dx-2\right]\\
&=\frac{\alpha_G}{\alpha_G-1}\times\\
&\left(\int_{\mathcal{X}}p_{G_\theta}(x)\left(\frac{(p_r(x)/p_{G_\theta}(x))^{\alpha_D(1-1/\alpha_G)+1}+1}{((p_r(x)/p_{G_\theta}(x))^{\alpha_D}+1)^{1-1/\alpha_G}}\right)dx-2\right)\\
&=\int_{\mathcal{X}}p_{G_\theta}(x)f_{\alpha_D,\alpha_G}\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx+\frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right),
\end{aligned}$$

where f_{α_D,α_G} is as defined in (56). Note that if f_{α_D,α_G} is strictly convex, the first term in the last equality above equals an f-divergence which is minimized if and only if $P_r=P_{G_\theta}$. Define the regions R_1 and R_2 as follows:

$$R_1:=\Big\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,\Big|\,\alpha_D\le 1,\alpha_G>\frac{\alpha_D}{\alpha_D+1}\Big\}$$

and

$$R_2:=\Big\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,\Big|\,\alpha_D>1,\frac{\alpha_D}{2}<\alpha_G\le\alpha_D\Big\}.$$

In order to prove that f_{α_D,α_G} is strictly convex for $(\alpha_D,\alpha_G)\in R_1\cup R_2$, we take its second derivative, which yields

$$\begin{aligned} &f''_{\alpha_D,\alpha_G}(u) \\ &= A_{\alpha_D,\alpha_G}(u)\Bigg[(\alpha_G+\alpha_D\alpha_G-\alpha_D)\left(u+u^{\alpha_D+\frac{\alpha_D}{\alpha_G}}\right) \\ &\qquad +(\alpha_G-\alpha_D\alpha_G)\left(u^{\frac{\alpha_D}{\alpha_G}}+u^{\alpha_D+1}\right)\Bigg], \end{aligned} \tag{166}$$

where

$$A_{\alpha_D,\alpha_G}(u)=\frac{\alpha_D}{\alpha_G}u^{\alpha_D-\frac{\alpha_D}{\alpha_G}-2}(1+u^{\alpha_D})^{\frac{1}{\alpha_G}-3}. \tag{167}$$

Note that $A_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$ and $\alpha_D,\alpha_G\in(0,\infty]$. Therefore, in order to ensure $f''_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$ it is sufficient to have

$$\alpha_G+\alpha_D\alpha_G-\alpha_D>\alpha_G(\alpha_D-1)B_{\alpha_D,\alpha_G}(u), \tag{168}$$

where

$$B_{\alpha_D,\alpha_G}(u)=\frac{u^{\frac{\alpha_D}{\alpha_G}}+u^{\alpha_D+1}}{u+u^{\alpha_D+\frac{\alpha_D}{\alpha_G}}} \tag{169}$$

for $u>0$. Since $B_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$, the sign of the RHS of (168) is determined by whether $\alpha_D\leq 1$ or $\alpha_D>1$. We look further into these two cases in the following:

Case 1: $\alpha_D\leq 1$. Then $\alpha_G(\alpha_D-1)B_{\alpha_D,\alpha_G}(u)\leq 0$ for all $u>0$ and $(\alpha_D,\alpha_G)\in(0,\infty]^2$. Therefore, we need

$$\alpha_G(1+\alpha_D)-\alpha_D>0\Leftrightarrow\alpha_G>\frac{\alpha_D}{\alpha_D+1}. \tag{170}$$

Case 2: $\alpha_D>1$. Then $\alpha_G(\alpha_D-1)B_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$ and $(\alpha_D,\alpha_G)\in(0,\infty]^2$. In order to obtain conditions on α_D and α_G, we determine the monotonicity of B_{α_D,α_G} by finding its first derivative as follows:

$$\begin{aligned} &B'_{\alpha_D,\alpha_G}(u) \\ &=\frac{(\alpha_G-\alpha_D)(u^{2\alpha_D}-1)+\alpha_D\alpha_G\left(u^{\alpha_D-\frac{\alpha_D}{\alpha_G}+1}-u^{\alpha_D+\frac{\alpha_D}{\alpha_G}-1}\right)}{\alpha_G u^{-\frac{\alpha_D}{\alpha_G}}\left(u+u^{\alpha_D+\frac{\alpha_D}{\alpha_G}}\right)^2}. \end{aligned}$$

Since the denominator of B'_{α_D,α_G} is positive for all $u>0$ and $(\alpha_D,\alpha_G)\in(0,\infty]^2$, we just need to check the sign of the numerator.

Case 2a: $\alpha_D>\alpha_G$. For $u\in(0,1)$,

$$u^{2\alpha_D}-1<0 \quad\text{and}\quad u^{\alpha_D-\frac{\alpha_D}{\alpha_G}+1}-u^{\alpha_D+\frac{\alpha_D}{\alpha_G}-1}>0,$$

so $B'_{\alpha_D,\alpha_G}(u)>0$. For $u>1$,

$$u^{2\alpha_D}-1>0 \quad\text{and}\quad u^{\alpha_D-\frac{\alpha_D}{\alpha_G}+1}-u^{\alpha_D+\frac{\alpha_D}{\alpha_G}-1}<0,$$

so $B'_{\alpha_D,\alpha_G}(u)<0$. For $u=1$, $B'_{\alpha_D,\alpha_G}(u)=0$. Hence, B'_{α_D,α_G} is strictly increasing for $u\in(0,1)$ and strictly decreasing for $u\geq 1$. Therefore, B_{α_D,α_G} attains a maximum value of 1 at $u=1$. This means B_{α_D,α_G} is bounded, i.e. $B_{\alpha_D,\alpha_G}\in(0,1]$ for all $u>0$. Thus, in order for (168) to hold, it suffices to ensure that

$$\alpha_G+\alpha_D\alpha_G-\alpha_D>\alpha_G(\alpha_D-1)\Leftrightarrow\alpha_G>\frac{\alpha_G}{2}. \tag{171}$$

Case 2b: $\alpha_D<\alpha_G$. For $u\in(0,1)$, $u^{2\alpha_D}-1<0$ and $u^{\alpha_D-\frac{\alpha_D}{\alpha_G}+1}-u^{\alpha_D+\frac{\alpha_D}{\alpha_G}-1}<0$, so $B'_{\alpha_D,\alpha_G}(u)<0$. For $u>1$, $u^{2\alpha_D}-1>0$ and $u^{\alpha_D-\frac{\alpha_D}{\alpha_G}+1}-u^{\alpha_D+\frac{\alpha_D}{\alpha_G}-1}>0$, so $B'_{\alpha_D,\alpha_G}(u)>0$. Hence, B'_{α_D,α_G} is strictly decreasing for $u\in(0,1)$ and strictly increasing for $u\geq 1$. Therefore, B_{α_D,α_G} attains a minimum value of 1 at $u=1$. This means that B_{α_D,α_G} is not bounded above, so it is not possible to satisfy (168) without restricting the domain of B_{α_D,α_G}.

Thus, for $(\alpha_D,\alpha_G)\in R_1\cup R_2$,

$$V_{\alpha_G}(\theta,\omega^*)=D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta})+\frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right).$$

This yields (57). Note that $D_{f_{\alpha_D,\alpha_G}}(P||Q)$ is symmetric since

$$\begin{aligned}
&D_{f_{\alpha_D,\alpha_G}}(Q||P)\\
&=\int_{\mathcal{X}} p(x) f_{\alpha_D,\alpha_G}\left(\frac{q(x)}{p(x)}\right)dx\\
&=\frac{\alpha_G}{\alpha_G-1}\times\\
&\left(\int_{\mathcal{X}} p(x)\left(\frac{(p(x)/q(x))^{-\alpha_D\left(1-\frac{1}{\alpha_G}\right)-1}+1}{((p(x)/q(x))^{-\alpha_D}+1)^{1-\frac{1}{\alpha_G}}}\right)dx-2^{\frac{1}{\alpha_G}}\right)\\
&=\frac{\alpha_G}{\alpha_G-1}\times\\
&\left(\int_{\mathcal{X}} p(x)\left(\frac{q(x)/p(x)+(p(x)/q(x))^{\alpha_D\left(1-\frac{1}{\alpha_G}\right)}}{(1+(p(x)/q(x))^{\alpha_D})^{1-\frac{1}{\alpha_G}}}\right)dx-2^{\frac{1}{\alpha_G}}\right)\\
&=\frac{\alpha_G}{\alpha_G-1}\times\\
&\left(\int_{\mathcal{X}} q(x)\left(\frac{1+(p(x)/q(x))^{\alpha_D\left(1-\frac{1}{\alpha_G}\right)}}{(1+(p(x)/q(x))^{\alpha_D})^{1-\frac{1}{\alpha_G}}}\right)dx-2^{\frac{1}{\alpha_G}}\right)\\
&=D_{f_{\alpha_D,\alpha_G}}(P||Q).
\end{aligned}$$

Since f_{α_D,α_G} is strictly convex and $f_{\alpha_D,\alpha_G}(1)=0$, $D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta})\geq 0$ with equality if and only if $P_r=P_{G_\theta}$. Thus, we have $V_{\alpha_G}(\theta,\omega^*)\geq\frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right)$ with equality if and only if $P_r=P_{G_\theta}$.

B. Proof of (b)

The generator's optimization problem in (54b) with the optimal discriminator in (55) can be written as $\inf_{\theta\in\Theta} V^{\text{NS}}_{\alpha_G}(\theta,\omega^*)$, where

$$\begin{aligned}
&V^{\text{NS}}_{\alpha_G}(\theta,\omega^*)\\
&=\frac{\alpha_G}{\alpha_G-1}\left[1-\int_{\mathcal{X}}\left(p_{G_\theta}(x)D_{\omega^*}(x)^{\frac{\alpha_G-1}{\alpha_G}}\right)dx\right]\\
&=\frac{\alpha_G}{\alpha_G-1}\left[1-\int_{\mathcal{X}} p_{G_\theta}(x)\left(\frac{p_r(x)^{\alpha_D}}{p_r(x)^{\alpha_D}+p_{G_\theta}(x)^{\alpha_D}}\right)^{\frac{\alpha_G-1}{\alpha_G}}dx\right]\\
&=\frac{\alpha_G}{\alpha_G-1}\left[1-\int_{\mathcal{X}} p_{G_\theta}(x)\frac{(p_r(x)/p_{G_\theta}(x))^{\alpha_D(1-1/\alpha_G)}}{((p_r(x)/p_{G_\theta}(x))^{\alpha_D}+1)^{1-1/\alpha_G}}dx\right]\\
&=\int_{\mathcal{X}} p_{G_\theta}(x) f^{\text{NS}}_{\alpha_D,\alpha_G}\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx+\frac{\alpha_G}{\alpha_G-1}\left(1-2^{\frac{1}{\alpha_G}-1}\right),
\end{aligned}$$

where $f^{\text{NS}}_{\alpha_D,\alpha_G}$ is as defined in (61). In order to prove that $f^{\text{NS}}_{\alpha_D,\alpha_G}$ is strictly convex for $(\alpha_D,\alpha_G)\in R_{\text{NS}}=\{(\alpha_D,\alpha_G)\in(0,\infty]^2\,|\,\alpha_D>\alpha_G(\alpha_D-1)\}$, we take its second derivative, which yields

$$\begin{aligned}
&f''_{\alpha_D,\alpha_G}(u)\\
&=A_{\alpha_D,\alpha_G}(u)\left[(\alpha_G-\alpha_D\alpha_G+\alpha_D)+\alpha_G(1+\alpha_D)u^{\alpha_D}\right],
\end{aligned}\tag{172}$$

where A_{α_D,α_G} is defined as in (167). Since $A_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$ and $(\alpha_D,\alpha_G)\in(0,\infty]^2$, to ensure $f''_{\alpha_D,\alpha_G}(u)>0$ for all $u>0$ it suffices to have

$$\frac{\alpha_G-\alpha_D\alpha_G+\alpha_D}{\alpha_G(1+\alpha_D)}>-u^{\alpha_D}$$

for all $u>0$. This is equivalent to

$$\frac{\alpha_G-\alpha_D\alpha_G+\alpha_D}{\alpha_G(1+\alpha_D)}>0,$$

which results in the condition

$$\alpha_D>\alpha_G(\alpha_D-1)$$

for $(\alpha_D,\alpha_G)\in(0,\infty]^2$. Thus, for $(\alpha_D,\alpha_G)\in R_{\text{NS}}$,

$$V_{\alpha_G}^{\text{NS}}(\theta,\omega^*)=D_{f^{\text{NS}}_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta})+\frac{\alpha_G}{\alpha_G-1}\left(1-2^{\frac{1}{\alpha_G}-1}\right).$$

This yields (62). Note that $D_{f^{\text{NS}}_{\alpha_D,\alpha_G}}(P||Q)$ is not symmetric since $D_{f^{\text{NS}}_{\alpha_D,\alpha_G}}(P||Q)\neq D_{f^{\text{NS}}_{\alpha_D,\alpha_G}}(Q||P)$. Since $f^{\text{NS}}_{\alpha_D,\alpha_G}$ is strictly convex and $f^{\text{NS}}_{\alpha_D,\alpha_G}(1)=0$, $D_{f^{\text{NS}}_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta})\geq 0$ with equality if and only if $P_r=P_{G_\theta}$. Thus, we have $V_{\alpha_G}^{\text{NS}}(\theta,\omega^*)\geq \frac{\alpha_G}{\alpha_G-1}\left(1-2^{\frac{1}{\alpha_G}-1}\right)$ with equality if and only if $P_r=P_{G_\theta}$.

APPENDIX M
PROOF OF THEOREM 12

By adding and subtracting relevant terms, we obtain

$$\begin{aligned}
&d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(P_r,P_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta})\\
&=d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(P_r,P_{G_{\hat\theta^*}})-d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(\hat P_r,P_{G_{\hat\theta^*}}) &\text{(173a)}\\
&\quad+\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(\hat P_r,P_{G_\theta})-\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta}) &\text{(173b)}\\
&\quad+d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(\hat P_r,P_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(\hat P_r,P_{G_\theta}). &\text{(173c)}
\end{aligned}$$

We upper-bound (173) in the following three steps. Let $\phi(\cdot)=-\ell_G(1,\cdot)$ and $\psi(\cdot)=-\ell_G(0,\cdot)$.
We first upper-bound (173a). Let $\omega^*(\hat\theta^*)=\omega^*(P_r,P_{G_{\hat\theta^*}})$. Using (66) yields

$$\begin{aligned}
&d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(P_r,P_{G_{\hat\theta^*}})-d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(\hat P_r,P_{G_{\hat\theta^*}})\\
&=\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat\theta^*}}}[\psi(D_{\omega^*(\hat\theta^*)}(X))]\\
&\quad-\left(\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat\theta^*}}}[\psi(D_{\omega^*(\hat\theta^*)}(X))]\right)\\
&\leq\left|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]-\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]\right|\\
&\leq\sup_{\omega\in\Omega}\left|\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]-\mathbb{E}_{X\sim \hat P_r}[\phi(D_\omega(X))]\right|. &\text{(174)}
\end{aligned}$$

Next, we upper-bound (173b). Let $\theta^*=\arg\min_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta})$ and $\omega^*(\theta^*)=\omega^*(P_r,P_{G_{\theta^*}})$. Then

$$\begin{aligned}
&\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(\hat P_r,P_{G_\theta})-\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta})\\
&\leq d_{\omega^*(\theta^*)}(\hat P_r,P_{G_{\theta^*}})-d_{\omega^*(\theta^*)}(P_r,P_{G_{\theta^*}})\\
&=\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\theta^*}}}[\psi(D_{\omega^*(\theta^*)}(X))]\\
&\quad-\left(\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\theta^*}}}[\psi(D_{\omega^*(\theta^*)}(X))]\right)\\
&=\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\theta^*)}(X))]-\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\theta^*)}(X))]\\
&\leq\sup_{\omega\in\Omega}\left|\mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))]-\mathbb{E}_{X\sim \hat P_r}[\phi(D_\omega(X))]\right|. &\text{(175)}
\end{aligned}$$

Lastly, we upper-bound (173c). Let $\tilde\theta=\arg\min_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(\hat P_r,P_{G_\theta})$ and $\omega^*(\tilde\theta)=\omega^*(P_r,P_{G_{\tilde\theta}})$. Then

$$\begin{aligned}
&d_{\omega^*(P_r,P_{G_{\hat\theta^*}})}(\hat P_r,P_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(\hat P_r,P_{G_\theta})\\
&=d_{\omega^*(\hat\theta^*)}(\hat P_r,P_{G_{\hat\theta^*}})-d_{\omega^*(\tilde\theta)}(\hat P_r,\hat P_{G_{\tilde\theta}})\\
&\quad+d_{\omega^*(\tilde\theta)}(\hat P_r,\hat P_{G_{\tilde\theta}})-d_{\omega^*(\tilde\theta)}(\hat P_r,P_{G_{\tilde\theta}})\\
&\leq d_{\omega^*(\hat\theta^*)}(\hat P_r,P_{G_{\hat\theta^*}})-d_{\omega^*(\hat\theta^*)}(\hat P_r,\hat P_{G_{\hat\theta^*}})\\
&\quad+d_{\omega^*(\tilde\theta)}(\hat P_r,\hat P_{G_{\tilde\theta}})-d_{\omega^*(\tilde\theta)}(\hat P_r,P_{G_{\tilde\theta}})\\
&=\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat\theta^*}}}[\psi(D_{\omega^*(\hat\theta^*)}(X))]\\
&\quad-\left(\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\hat\theta^*)}(X))]+\mathbb{E}_{X\sim \hat P_{G_{\hat\theta^*}}}[\psi(D_{\omega^*(\hat\theta^*)}(X))]\right)\\
&\quad+\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\tilde\theta)}(X))]+\mathbb{E}_{X\sim \hat P_{G_{\tilde\theta}}}[\psi(D_{\omega^*(\tilde\theta)}(X))]\\
&\quad-\left(\mathbb{E}_{X\sim \hat P_r}[\phi(D_{\omega^*(\tilde\theta)}(X))]+\mathbb{E}_{X\sim P_{G_{\tilde\theta}}}[\psi(D_{\omega^*(\tilde\theta)}(X))]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{X\sim P_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))] - \mathbb{E}_{X\sim \hat{P}_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))] \\
&\quad + \mathbb{E}_{X\sim \hat{P}_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))] - \mathbb{E}_{X\sim P_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))] \\
&\leq 2 \sup_{\omega\in\Omega,\theta\in\Theta} \Big| \mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))] - \mathbb{E}_{X\sim \hat{P}_{G_\theta}}[\psi(D_\omega(X))] \Big|. \qquad (176)
\end{aligned}
$$

Combining (174)-(176), we obtain the following bound for (173):

$$
\begin{aligned}
&d_{\omega^*(P_r,P_{G_{\hat{\theta}^*}})}(P_r,P_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d_{\omega^*(P_r,P_{G_\theta})}(P_r,P_{G_\theta}) \\
&\leq 2\sup_{\omega\in\Omega} \Big| \mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))] - \mathbb{E}_{X\sim \hat{P}_r}[\phi(D_\omega(X))] \Big| \\
&\quad + 2 \sup_{\omega\in\Omega,\theta\in\Theta} \Big| \mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))] - \mathbb{E}_{X\sim \hat{P}_{G_\theta}}[\psi(D_\omega(X))] \Big| \\
&= 2\sup_{\omega\in\Omega} \Big| \mathbb{E}_{X\sim P_r}[\phi(D_\omega(X))] - \frac{1}{n}\sum_{i=1}^{n} \phi(D_\omega(X_i)) \Big| \\
&\quad + 2 \sup_{\omega\in\Omega,\theta\in\Theta} \Big| \mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_\omega(X))] - \frac{1}{m}\sum_{j=1}^{m} \psi(D_\omega(X_j)) \Big|. \qquad (177)
\end{aligned}
$$

Note that (177) is exactly the same bound as that in (50). Hence, the remainder of the proof follows from the proof of Theorem 8, where $\phi(\cdot) = -\ell_G(1,\cdot)$ and $\psi(\cdot) = -\ell_G(0,\cdot)$. The specialization to (α_D,α_D)-GANs follows from setting $\ell_G = \ell_{\alpha_G}$.

APPENDIX N
ADDITIONAL EXPERIMENTAL RESULTS

A. Brief Overview of LSGAN

The Least Squares GAN (LSGAN) is a dual-objective min-max game introduced in [15]. The LSGAN objective functions, as the name suggests, involve squared loss functions for D and G which are written as

$$
\begin{aligned}
&\inf_{\omega\in\Omega} \frac{1}{2}\Big(\mathbb{E}_{X\sim P_r}[(D_\omega(X)-b)^2] + \mathbb{E}_{X\sim P_{G_\theta}}[(D_\omega(X)-a)^2] \Big) \\
&\inf_{\theta\in\Theta} \frac{1}{2}\Big(\mathbb{E}_{X\sim P_r}[(D_\omega(X)-c)^2] + \mathbb{E}_{X\sim P_{G_\theta}}[(D_\omega(X)-c)^2] \Big). \qquad (178)
\end{aligned}
$$

For appropriately chosen values of the parameters a, b, and c, (178) reduces to minimizing the Pearson χ^2-divergence between $P_r + P_{G_\theta}$ and $2P_{G_\theta}$. As done in the original paper [15], we use $a=0$, $b=1$ and $c=1$ for our experiments to make fair comparisons. The authors refer to this choice of parameters as the 0-1 binary coding scheme.

B. 2D Gaussian Mixture Ring

In Tables I and II, we report the success (8/8 mode coverage) and failure (0/8 mode coverage) rates over 200 seeds for a grid of (α_D,α_G) combinations for the *saturating* setting. Compared to the vanilla GAN performance, we find that tuning α_D below 1 leads to a greater success rate and lower failure rate. However, in this saturating loss setting, we find that tuning α_G away from 1 has no significant impact on GAN performance.

TABLE I
SUCCESS RATES FOR 2D-RING WITH THE SATURATING (α_D,α_G)-GAN OVER 200 SEEDS, WITH TOP 4 COMBINATIONS EMBOLDENED.

% OF SUCCESS (8/8 MODES)		α_D = 0.5	0.6	0.7	0.8	0.9	1.0
α_G	0.9	73	79	69	60	46	34
	1.0	**80**	**79**	74	68	54	47
	1.1	**79**	77	68	70	59	47
	1.2	**75**	74	71	65	57	46

In Table III, we detail the success rates for the NS setting. We note that for this dataset, no failures, and therefore, no vanishing/exploding gradients, occurred in the NS setting. In particular, we find that the $(0.5,1.2)$-GAN doubles the success rate of the vanilla $(1,1)$-GAN, which is more susceptible to mode collapse as illustrated in Figure 10. We also find that LSGAN achieves a success rate of 32.5%, which is greater than vanilla GAN but less than the best-performing (α_D,α_G)-GAN.

TABLE II
FAILURE RATES FOR 2D-RING WITH THE SATURATING (α_D,α_G)-GAN OVER 200 SEEDS, WITH TOP 3 COMBINATIONS EMBOLDENED.

% OF FAILURE (0/8 MODES)		α_D					
		0.5	0.6	0.7	0.8	0.9	1.0
α_G	0.9	11	10	12	13	29	49
	1.0	**5**	**5**	7	8	16	30
	1.1	7	9	13	12	13	26
	1.2	9	**5**	9	12	17	31

TABLE III
SUCCESS RATES FOR 2D-RING WITH THE NS (α_D,α_G)-GAN OVER 200 SEEDS, WITH TOP 5 COMBINATIONS EMBOLDENED.

% OF SUCCESS (8/8 MODES)		α_D							
		0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
α_G	0.8	35	24	19	19	14	16	18	10
	0.9	**39**	37	19	22	16	20	19	21
	1.0	34	35	29	28	26	22	20	32
	1.1	**40**	36	31	22	24	15	23	25
	1.2	**45**	38	34	25	26	28	20	22
	1.3	**44**	**39**	26	28	28	25	31	29

C. Celeb-A & LSUN Classroom

The discriminator and generator architectures used for the Celeb-A and LSUN Classroom datasets are described in Tables IV and V respectively. Each architecture consists of four CNN layers, with parameters such as kernel size (i.e., size of the filter, denoted as "Kernel"), stride (the amount by which the filter moves), and the activation functions applied to the layer outputs. Zero padding is also assumed. In both tables, "BN" represents batch normalization, a technique that normalizes the inputs to each layer using a batch of samples during model training. Batch normalization is commonly employed in deep learning to prevent cumulative floating point errors and overflows, and to ensure that all features remain within a similar range. This technique serves as a computational tool to address vanishing and/or exploding gradients.

In Table VI, we collate the FID results for both datasets as a function of the learning rates. This table captures the percentage (out of 50 seeds) of FID scores below a desired threshold, which is 80 for the CELEB-A dataset and 800 for the LSUN Classroom dataset.

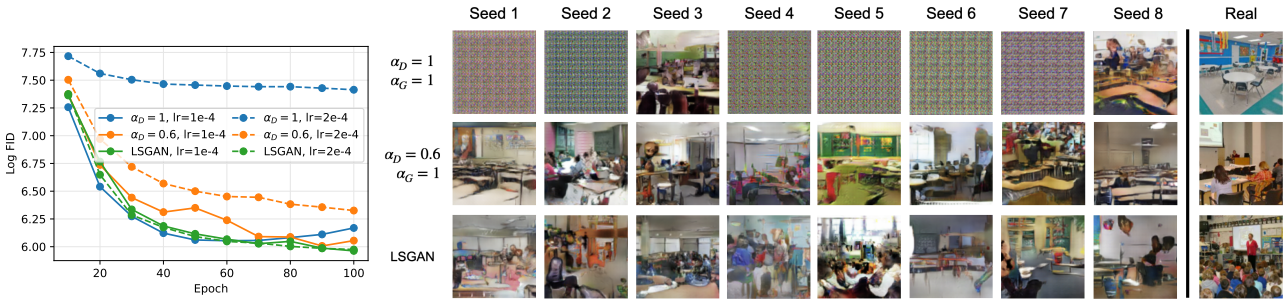
We first focus on the CELEB-A dataset: Table VI demonstrates that for a learning rate of 1×10^{-4}, all GANs (vanilla,

TABLE IV
DISCRIMINATOR AND GENERATOR ARCHITECTURES FOR CELEB-A.
THE FINAL SIGMOID ACTIVATION LAYER IS REMOVED FOR THE LSGAN DISCRIMINATOR.

DISCRIMINATOR						GENERATOR					
LAYER	OUTPUT SIZE	KERNEL	STRIDE	BN	ACTIVATION	LAYER	OUTPUT SIZE	KERNEL	STRIDE	BN	ACTIVATION
INPUT	3×64×64				LEAKY RELU	INPUT	100×1×1				RELU
CONVOLUTION	64×32×32	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	512×4×4	4×4	2	YES	RELU
CONVOLUTION	128×16×16	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	256×8×8	4×4	2	YES	RELU
CONVOLUTION	256×8×8	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	128×16×16	4×4	2	YES	RELU
CONVOLUTION	512×4×4	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	64×32×32	4×4	2	YES	RELU
CONVOLUTION	1×1×1	4×4	2		SIGMOID	CONVTRANSPOSE	3×64×64	4×4	2		TANH

TABLE V
DISCRIMINATOR AND GENERATOR ARCHITECTURES FOR LSUN CLASSROOM.
THE FINAL SIGMOID ACTIVATION LAYER IS REMOVED FOR THE LSGAN DISCRIMINATOR.

DISCRIMINATOR						GENERATOR					
LAYER	OUTPUT SIZE	KERNEL	STRIDE	BN	ACTIVATION	LAYER	OUTPUT SIZE	KERNEL	STRIDE	BN	ACTIVATION
INPUT	3×112×112				LEAKY RELU	INPUT	100×1×1				RELU
CONVOLUTION	64×56×56	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	512×7×7	7×7	2	YES	RELU
CONVOLUTION	128×28×28	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	256×14×14	4×4	2	YES	RELU
CONVOLUTION	256×14×14	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	128×28×28	4×4	2	YES	RELU
CONVOLUTION	512×7×7	4×4	2	YES	LEAKY RELU	CONVTRANSPOSE	64×56×56	4×4	2	YES	RELU
CONVOLUTION	1×1×1	7×7	2		SIGMOID	CONVTRANSPOSE	3×112×112	4×4	2		TANH



(a) (b)

Fig. 13. (a) Log-scale plot of **LSUN Classroom** FID scores over training epochs in steps of 10 up to 100 total, for three noteworthy GANs– $(1,1)$-GAN (vanilla), $(0.6,1)$-GAN, and LSGAN– and for two similar learning rates– 1×10^{-4} and 2×10^{-4}. Results show that the vanilla GAN performance is very sensitive to learning rate choice as the difference between training with 1×10^{-4} and 2×10^{-4} is drastic. On the other hand, the other two GANs achieve consistently lower FIDs, with the LSGAN performing the best. (b) Generated LSUN Classroom images from the same three GANs over 8 seeds when trained for 100 epochs with a learning rate of 2×10^{-4}. These samples show that the vanilla $(1,1)$-GAN training fails for most of seeds while the other two GANs perform fairly well across all seeds, thus exhibiting robustness to random weight initializations.

different (α_D,α_G)-GANs, and LSGANs) achieve an FID score below 80 at least 93% of the time. However, the instability of vanilla GAN is also evident in Table VI, where for a slightly higher learning rate of 6×10^{-4}, the $(1,1)$-GAN achieves an FID score below 80 only 60% of the time whereas at least one $(\alpha_D,\alpha_G=1)$-GAN consistently performs better than 76% over all chosen learning rates. We observe that tuning α_D below 1 contributes to stabilizing the FID scores over the 50 seeds while maintaining relatively low scores on average. This stability is emphasized in Table VI, in particular for the $(0.7,1)$-GAN, as it achieves an FID score below 80 at least 80% of the time for 7 out of the 10 the learning rates.

Table VI also illustrates similar results for the LSUN Classroom dataset. However, increasing it to 2×10^{-4} leads to instability in the vanilla $(1,1)$-GAN across 50 seeds.

TABLE VI
PERCENTAGE OUT OF 50 SEEDS OF FID SCORES BELOW 80 (CELEB-A) OR 800 (LSUN CLASSROOM) FOR EACH COMBINATION OF (α_D,α_G)-GAN AND LEARNING RATE, TRAINED FOR 100 EPOCHS. BEST RESULTS FOR EACH DATASET AND LEARNING RATE ARE **EMBOLDENED**.

GAN	CELEB-A								LSUN CLASSROOM				
	LEARNING RATE ($\times10^{-4}$)												
(α_D,α_G)	1	2	5	6	7	8	9	10	1	2	3	4	5
$(1,1)$	**100**	93.2	82.6	59.5	58.5	39.0	53.7	54.8	92.0	36.2	12.5	13.0	12.2
$(0.9,1)$	**100**	95.2	78.3	72.3	81.4	66.7	74.4	46.5	76.0	53.1	22.2	17.0	22.2
$(0.8,1)$	97.8	**97.6**	**88.9**	82.2	81.4	72.1	68.4	75.6	88.5	60.8	36.2	27.9	29.2
$(0.7,1)$	**100**	90.7	**88.9**	**91.5**	**86.4**	**81.2**	67.6	**80.0**	90.2	80.4	78.4	67.4	55.1
$(0.6,1)$	97.8	93.0	88.4	76.6	84.6	75.6	**76.9**	69.2	**95.7**	**90.4**	**85.1**	**78.3**	**66.0**

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] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

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

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

[8] G. R. Kurri, M. Welfert, T. Sypherd, and L. Sankar, "α-GAN: Convergence and estimation guarantees," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 276–281.

[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," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 6021–6051, 2022.

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

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

[15] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[17] B. Poole, A. A. Alemi, J. Sohl-Dickstein, and A. Angelova, "Improved generator objectives for gans," *arXiv preprint arXiv:1612.02780*, 2016.

[18] 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.

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

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

[21] 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.

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

[23] 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.

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

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

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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

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

[35] 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.

[36] A. Ruderman, M. Reid, D. García-García, and J. Petterson, "Tighter variational representations of f-divergences via restriction to probability measures," *arXiv preprint arXiv:1206.4664*, 2012.

[37] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.

[38] M. Shannon, "Properties of f-divergences and f-gan training," *arXiv preprint arXiv:2009.00757*, 2020.

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

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

[41] K. Ji and Y. Liang, "Minimax estimation of neural net distance," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[42] 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.

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

[44] 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.

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

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

[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[48] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "VEEGAN: Reducing mode collapse in GANs using implicit variational learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[49] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[50] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *CoRR*, vol. abs/1506.03365, 2015. [Online]. Available: http://arxiv.org/abs/1506.03365

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

[52] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a Nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[53] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2009.