Towards addressing GAN training instabilities: Dual-objective GANs with tunable parameters

Arizona State University, {}@asu.edu

***Abstract*—**

I. Introduction

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

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

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

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

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

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

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

Various other GANs have been studied in the literature (e.g., f-divergence based GANs known as f-GAN [3], IPM based GANs [4]–[6], Cumulant GAN [7], RényiGAN [8], to name a few) with different value functions. In each case, the corresponding min-max optimization problem simplifies to minimizing some measure of divergence between the real and generated distributions. Yet, a methodical way to compare and operationally interpret GAN value functions remains open.

This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, CIF-2031799, and CIF-1934766.

II. Problem Formulation

A. (α_D,α_G)-*GAN*

We propose a dual-objective (α_D,α_G)-GAN with different objective functions for the generator and discriminator. In particular, the discriminator maximizes $V_{\alpha_D}(\theta,\omega)$ while the generator minimizes $V_{\alpha_G}(\theta,\omega)$ defined as follows for $\alpha_G,\alpha_D \in (0,\infty]$:

$$\begin{aligned} &V_{\alpha_D}(\theta,\omega) \\ &= \mathbb{E}_{X\sim P_r}[-\ell_{\alpha_D}(1,D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell_{\alpha_D}(0,D_\omega(X))] \\ &= \frac{\alpha_D}{\alpha_D-1}\times \\ &\left(\mathbb{E}_{X\sim P_r}\left[D_\omega(X)^{\frac{\alpha_D-1}{\alpha_D}}\right] + \mathbb{E}_{X\sim P_{G_\theta}}\left[(1-D_\omega(X))^{\frac{\alpha_D-1}{\alpha_D}}\right] - 2\right), \end{aligned} \tag{4}$$

$$\begin{aligned} &V_{\alpha_G}(\theta,\omega) \\ &= \mathbb{E}_{X\sim P_r}[-\ell_{\alpha_G}(1,D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell_{\alpha_G}(0,D_\omega(X))] \\ &= \frac{\alpha_G}{\alpha_G-1}\times \\ &\left(\mathbb{E}_{X\sim P_r}\left[D_\omega(X)^{\frac{\alpha_G-1}{\alpha_G}}\right] + \mathbb{E}_{X\sim P_{G_\theta}}\left[(1-D_\omega(X))^{\frac{\alpha_G-1}{\alpha_G}}\right] - 2\right). \end{aligned} \tag{5}$$

We recover the original α-GAN value function when $\alpha_d = \alpha_g = \alpha$. The resulting (α_D,α_G)-GAN is given by

$$\sup_{\omega\in\Omega} V_{\alpha_D}(\theta,\omega) \tag{6a}$$

$$\inf_{\theta\in\Theta} V_{\alpha_G}(\theta,\omega). \tag{6b}$$

Non-saturating alternative to the generator's objective function in (5):

$$\begin{aligned} V_{\alpha_G}^{NS}(\theta,\omega) &= \mathbb{E}_{X\sim P_{G_\theta}}[\ell_{\alpha_G}(1,D_\omega(X))] \\ &= \frac{\alpha_G}{\alpha_G-1}\left(-\mathbb{E}_{X\sim P_{G_\theta}}\left[D_\omega(X)^{\frac{\alpha_G-1}{\alpha_G}}\right]+1\right). \end{aligned} \tag{7}$$

B. Estimation Error

We now consider a setting in which we have a limited number of 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 the discriminator and generator classes are neural networks; these limitations lead to estimation errors in training GANs [6], [9], [10]. Building on the

While [10] models the interplay between both the discriminator and generator in the estimation error bounds, those developed in [6], [9] do not explicitly capture the role of the generator. We adopt the approach in [10].

For $x \in \mathcal{X} := \{x \in \mathbb{R}^d : ||x||_2 \leq B_x\}$ and $z \in \mathcal{Z} := \{z \in \mathbb{R}^p : ||z||_2 \leq B_z\}$, we consider discriminators and generators as neural network models of the form:

$$D_\omega : x \mapsto \sigma\big(\mathbf{w}_k^\intercal r_{k-1}(\mathbf{W}_{d-1} r_{k-2}(\ldots r_1(\mathbf{W}_1(x))))\big) \tag{8}$$

$$G_\theta : z \mapsto \mathbf{V}_l s_{l-1}(\mathbf{V}_{l-1} s_{l-2}(\ldots s_1(\mathbf{V}_1 z))), \tag{9}$$

where, $\mathbf{w}_k$ is a parameter vector of the output layer; for $i \in [1:k-1]$ and $j \in [1:l]$, $\mathbf{W}_i$ and $\mathbf{V}_j$ are parameter matrices; $r_i(\cdot)$ and $s_j(\cdot)$ are entry-wise activation functions of layers i and j, i.e., for $\mathbf{a} \in \mathbb{R}^t$, $r_i(\mathbf{a}) = [r_i(a_1),\ldots,r_i(a_t)]$ and $s_i(\mathbf{a}) = [s_i(a_1),\ldots,s_i(a_t)]$; and $\sigma(\cdot)$ is the sigmoid function given by $\sigma(p) = 1/(1+\mathrm{e}^{-p})$ (note that σ does not appear in the discriminator in [10, 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 $\lambda \geq 0$ and $p \in \mathbb{R}$. Finally, as modelled in [10]–[13], we assume that the Frobenius norms of the parameter matrices are bounded, i.e., $||\mathbf{W}_i||_F \leq M_i$, $i \in [1:k-1]$, $||\mathbf{w}_k||_2 \leq M_k$, and $||\mathbf{V}_j||_F \leq N_j$, $j \in [1:l]$.

Let

$$\omega^* = \operatorname*{argmax}_{\omega\in\Omega}\Big(\mathbb{E}_{X\sim\hat{P}_r}[-\ell_{\alpha_D}(1,D_\omega(X))] + \mathbb{E}_{X\sim\hat{P}_{G_\theta}}[-\ell_{\alpha_D}(0,D_\omega(X))]\Big), \tag{10}$$

and define

$$d_{\omega^*(\theta)}(P_r,P_{G_\theta}) = V_{\alpha_G}(\theta,\omega^*). \tag{11}$$

Then the resulting minimization for the training of (α_D,α_G)-GAN is

$$\inf_{\theta\in\Theta} d_{\omega^*(\theta)}(\hat{P}_r,\hat{P}_{G_\theta}). \tag{12}$$

We define the estimation error for (α_D,α_G)-GAN as

$$d_{\omega^*(\hat{\theta}^*)}(P_r,P_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d_{\omega^*(\theta)}(P_r,P_{G_\theta}), \tag{13}$$

where $\hat{\theta}^*$ is the minimizer of (12).

III. Main Results

The following theorem provides the solution to the two-player game in (6) for the non-parametric setting, i.e., when the discriminator set Ω is large enough.

Theorem 1. *For a fixed generator G_θ, the discriminator $D_{\omega^*} : \mathcal{X} \to [0,1]$ optimizing* (6a) *is given by*

$$D_{\omega^*}(x) = \frac{p_r(x)^{\alpha_D}}{p_r(x)^{\alpha_D} + p_{G_\theta}(x)^{\alpha_D}}. \tag{14}$$

For this D_{ω^*}, (6b) *simplifies to minimizing a non-negative symmetric f_{α_D,α_G}-divergence $D_{f_{\alpha_D,\alpha_G}}(\cdot||\cdot)$ for $(\alpha_D,\alpha_G) \in R_1 \cup R_2$, where*

$$R_1 = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D \leq 1, \alpha_G > \frac{\alpha_D}{\alpha_D+1}\}$$

and

$$R_2 = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D > 1, \frac{\alpha_D}{2} < \alpha_G \leq \alpha_D\},$$

as

$$\inf_{\theta\in\Theta} D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) + \frac{\alpha_G}{\alpha_G-1}\Big(2^{\frac{1}{\alpha_G}} - 2\Big), \tag{15}$$

where

$$f_{\alpha_D,\alpha_G}(u) = \frac{\alpha_G}{\alpha_G-1}\left(\frac{u^{\alpha_D\left(1-\frac{1}{\alpha_G}\right)+1}+1}{(u^{\alpha_D}+1)^{1-\frac{1}{\alpha_G}}} - 2^{\frac{1}{\alpha_G}}\right), \tag{16}$$

for $u \geq 0$ and

$$D_{f_{\alpha_D,\alpha_G}}(P||Q) = \int_{\mathcal{X}} q(x) f_{\alpha_D,\alpha_G}\left(\frac{p(x)}{q(x)}\right) dx, \tag{17}$$

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

Theorem 2. *For the same D_{ω^*} in* (14)*,* (6b) *simplifies to minimizing a non-negative non-symmetric $f^{NS}_{\alpha_D,\alpha_G}$-divergence $D_{f^{NS}_{\alpha_D,\alpha_G}}(\cdot||\cdot)$ for $(\alpha_D,\alpha_G) \in R_{NS}$, where*

$$R_{NS} = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D > \alpha_G(\alpha_D - 1)\},$$

as

$$\inf_{\theta\in\Theta} D_{f^{NS}_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) + \frac{\alpha_G}{\alpha_G-1}\Big(1 - 2^{\frac{1}{\alpha_G}-1}\Big), \tag{18}$$

where

$$f^{NS}_{\alpha_D,\alpha_G}(u) = \frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}-1} - \frac{u^{\alpha_D\left(1-\frac{1}{\alpha_G}\right)}}{(u^{\alpha_D}+1)^{1-\frac{1}{\alpha_G}}}\right), \tag{19}$$

for $u \geq 0$ and

$$D_{f^{NS}_{\alpha_D,\alpha_G}}(P||Q) = \int_{\mathcal{X}} q(x) f^{NS}_{\alpha_D,\alpha_G}\left(\frac{p(x)}{q(x)}\right) dx, \tag{20}$$

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

Theorem 3. *In the setting previously described, 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*

$$\begin{aligned} &d_{\omega^*(\hat{\theta}^*)}(P_r,P_{G_{\hat{\theta}^*}}) - \inf_{\theta\in\Theta} d_{\omega^*(\theta)}(P_r,P_{G_\theta}) \\ &\leq \frac{4C_{Q_x}(\alpha_G) B_x U_\omega \sqrt{3k}}{\sqrt{n}} + \frac{4C_{Q_z}(\alpha_G) U_\omega U_\theta B_z \sqrt{3(k+l-1)}}{\sqrt{m}} \\ &\quad + U_\omega \sqrt{\log\frac{1}{\delta}}\left(\frac{4C_{Q_x}(\alpha_G) B_x}{\sqrt{2n}} + \frac{4C_{Q_z}(\alpha_G) B_z U_\theta}{\sqrt{2m}}\right), \end{aligned} \tag{21}$$

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

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

IV. Experimental Results

V. Conclusion

Appendix A
Proof of Theorem 1

Proof. The proof to obtain (14) is the same as that for [14, Theorem 1], where $\alpha = \alpha_D$. With this, the generator's optimization problem in (6b) can be written as $\inf_{\theta\in\Theta} V_{\alpha_G}(\theta,\omega^*)$, where

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

Define f_{α_D,α_G} as in (16). 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

$$\begin{aligned} &f''_{\alpha_D,\alpha_G}(u) \\ &= A_{\alpha_D,\alpha_G}(u)\Big[(\alpha_G+\alpha_D\alpha_G-\alpha_D)\left(u+u^{\alpha_D+\frac{\alpha_D}{\alpha_G}}\right) \\ &\qquad + (\alpha_G-\alpha_D\alpha_G)\left(u^{\frac{\alpha_D}{\alpha_G}}+u^{\alpha_D+1}\right)\Big], \end{aligned} \tag{23}$$

where

$$A_{\alpha_D,\alpha_G}(u) = \frac{\alpha_D}{\alpha_G} u^{\alpha_D - \frac{\alpha_D}{\alpha_G} - 2}(1+u^{\alpha_D})^{\frac{1}{\alpha_G}-3}. \tag{24}$$

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

$$\alpha_G+\alpha_D\alpha_G-\alpha_D > \alpha_G(\alpha_D-1)B_{\alpha_D,\alpha_G}(u), \tag{25}$$

where

$$B_{\alpha_D,\alpha_G}(u) = \frac{u^{\frac{\alpha_D}{\alpha_G}}+u^{\alpha_D+1}}{u+u^{\alpha_D+\frac{\alpha_D}{\alpha_G}}} \tag{26}$$

for $u > 0$. Since $B_{\alpha_D,\alpha_G}(u) > 0$ for all $u > 0$, the sign of the RHS of (25) is determined by whether $\alpha_D \le 1$ or $\alpha_D > 1$. We look further into these two cases in the following:

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

$$\alpha_G(1+\alpha_D)-\alpha_D > 0 \Leftrightarrow \alpha_G > \frac{\alpha_D}{\alpha_D+1}. \tag{27}$$

Case 2: $\alpha_D > 1$**.** If $\alpha_D > 1$, 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 try to understand the behavior of B_{α_D,α_G} by finding its first derivative as follows:

$$B'_{\alpha_D,\alpha_G}(u) = \tag{28}$$

$$\alpha_G(1+\alpha_D)-\alpha_D > 0 \Leftrightarrow \alpha_G > \frac{\alpha_D}{\alpha_D+1}. \tag{29}$$

Figure 1 illustrates the region for these conditions. Thus, for $(\alpha_D,\alpha_G) \in R_1 \cup R_2$,

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

where

$$D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) = \int_{\mathcal{X}} p_{G_\theta}(x) f_{\alpha_D,\alpha_G}\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx.$$

This yields (15). 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 $f_{\alpha_D,\alpha_G}(1) = 0$, $D_{f_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) \ge 0$ with equality if and only if $P_r = P_{G_\theta}$. Thus, we have $V_{\alpha_G}(\theta,\omega^*) \ge \frac{\alpha_G}{\alpha_G-1}\left(2^{\frac{1}{\alpha_G}}-2\right)$ with equality if and only if $P_r = P_{G_\theta}$.

□

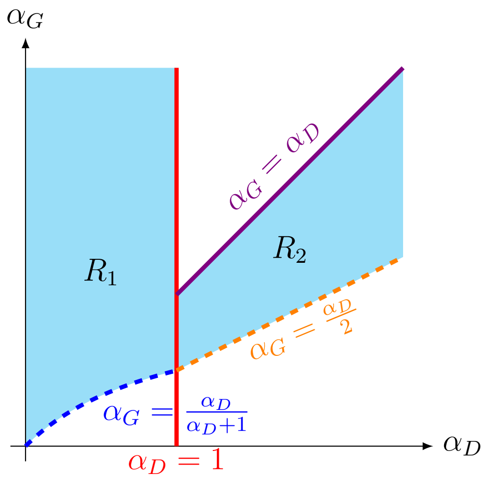

Fig. 1. Plot of region $R = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D \leq 1, \alpha_G > \frac{\alpha_D}{\alpha_D+1}\}$ for which f_{α_D,α_G} is strictly convex.

Appendix B
Proof of Theorem 2

Proof. Using (14), the generator's optimization problem in (6b) can be written as $\inf_{\theta\in\Theta} V^{NS}_{\alpha_G}(\theta,\omega^*)$, where

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

Define $f^{NS}_{\alpha_D,\alpha_G}$ as in (19). In order to prove that $f^{NS}_{\alpha_D,\alpha_G}$ is strictly convex for $(\alpha_D,\alpha_G) \in R_{NS} = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D > \alpha_G(\alpha_D-1)\}$, we take its second derivative, which yields

$$\begin{aligned}
&f''_{\alpha_D,\alpha_G}(u)\\
&= A_{\alpha_D,\alpha_G}(u)\left[(\alpha_G-\alpha_D\alpha_G+\alpha_D)+\alpha_G(1+\alpha_D)u^{\alpha_D}\right], \quad (30)
\end{aligned}$$

where A_{α_D,α_G} is defined as in (24). 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 $(\alpha_D,\alpha_G) \in (0,\infty]^2$. Figure 2 illustrates the region for this condition. Thus, for $(\alpha_D,\alpha_G) \in R_{NS}$,

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

where

$$D_{f^{NS}_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) = \int_{\mathcal{X}}p_{G_\theta}(x)f^{NS}_{\alpha_D,\alpha_G}\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)dx.$$

This yields (18). Note that $D_{f^{NS}_{\alpha_D,\alpha_G}}(P||Q)$ is not symmetric since $D_{f^{NS}_{\alpha_D,\alpha_G}}(P||Q) \neq D_{f^{NS}_{\alpha_D,\alpha_G}}(Q||P)$. Since $f^{NS}_{\alpha_D,\alpha_G}$ is strictly convex and $f^{NS}_{\alpha_D,\alpha_G}(1) = 0$, $D_{f^{NS}_{\alpha_D,\alpha_G}}(P_r||P_{G_\theta}) \geq 0$ with equality if and only if $P_r = P_{G_\theta}$. Thus, we have $V^{NS}_{\alpha_G}(\theta,\omega^*) \geq \frac{\alpha_G}{\alpha_G-1}\left(1-2^{\frac{1}{\alpha_G}-1}\right)$ with equality if and only if $P_r = P_{G_\theta}$.

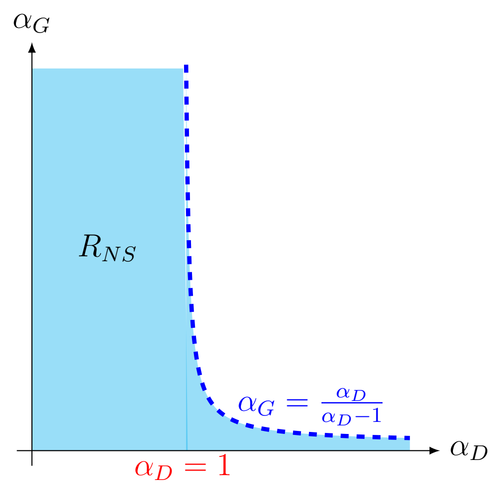

Fig. 2. Plot of region $R_{NS} = \{(\alpha_D,\alpha_G) \in (0,\infty]^2 \mid \alpha_D > \alpha_G(\alpha_D-1)\}$ for which $f^{NS}_{\alpha_D,\alpha_G}$ is strictly convex.

□

Appendix C
Proof of Theorem 3

By adding and subtracting relevant terms, we obtain

$$\begin{aligned}
&d_{\omega^*(\hat\theta^*)}(P_r,P_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta}d_{\omega^*(\theta)}(P_r,P_{G_\theta})\\
&= d_{\omega^*(\hat\theta^*)}(P_r,P_{G_{\hat\theta^*}})-d_{\omega^*(\hat\theta^*)}(\hat P_r,P_{G_{\hat\theta^*}}) \quad (31a)\\
&\quad+\inf_{\theta\in\Theta}d_{\omega^*(\theta)}(\hat P_r,P_{G_\theta})-\inf_{\theta\in\Theta}d_{\omega^*(\theta)}(P_r,P_{G_\theta}) \quad (31b)\\
&\quad+d_{\omega^*(\hat\theta^*)}(\hat P_r,P_{G_{\hat\theta^*}})-\inf_{\theta\in\Theta}d_{\omega^*(\theta)}(\hat P_r,P_{G_\theta}). \quad (31c)
\end{aligned}$$

We upper-bound (31) in the following three steps. Let $\phi(\cdot) = -\ell_{\alpha_G}(1,\cdot)$ and $\psi(\cdot) = -\ell_{\alpha_G}(0,\cdot)$.

We first upper-bound (31a). Using (11) yields

$$\begin{aligned}
&d_{\omega^*(\hat{\theta}^*)}(P_r,P_{G_{\hat{\theta}^*}})-d_{\omega^*(\hat{\theta}^*)}(\hat{P}_r,P_{G_{\hat{\theta}^*}})\\
&=\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]\\
&\quad-\Big(\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]\Big)\\
&=\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]-\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]\\
&\le\Big|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]-\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]\Big|\\
&\le\sup_{\omega\in\Omega}\Big|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega}(X))]-\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega}(X))]\Big|. \qquad (32)
\end{aligned}$$

Next, we upper-bound (31b). Let $\theta^* = \operatorname{argmin}_{\theta\in\Theta} d_{\omega^*(\theta)}(P_r,P_{G_\theta})$. Then

$$\begin{aligned}
&\inf_{\theta\in\Theta} d_{\omega^*(\theta)}(\hat{P}_r,P_{G_\theta})-\inf_{\theta\in\Theta} d_{\omega^*(\theta)}(P_r,P_{G_\theta})\\
&\le d_{\omega^*(\theta^*)}(\hat{P}_r,P_{G_{\theta^*}})-d_{\omega^*(\theta^*)}(P_r,P_{G_{\theta^*}})\\
&=\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\theta^*}}}[\psi(D_{\omega^*(\theta^*)}(X))]\\
&\quad-\Big(\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\theta^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\theta^*}}}[\psi(D_{\omega^*(\theta^*)}(X))]\Big)\\
&=\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\theta^*)}(X))]-\mathbb{E}_{X\sim P_r}[\phi(D_{\omega^*(\theta^*)}(X))]\\
&\le\sup_{\omega\in\Omega}\Big|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega}(X))]-\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega}(X))]\Big|. \qquad (33)
\end{aligned}$$

Lastly, we upper-bound (31c). Let $\tilde{\theta} = \operatorname{argmin}_{\theta\in\Theta} d_{\omega^*(\theta)}(\hat{P}_r,P_{G_\theta})$. Then

$$\begin{aligned}
&d_{\omega^*(\hat{\theta}^*)}(\hat{P}_r,P_{G_{\hat{\theta}^*}})-\inf_{\theta\in\Theta} d_{\omega^*(\theta)}(\hat{P}_r,P_{G_\theta})\\
&=d_{\omega^*(\hat{\theta}^*)}(\hat{P}_r,P_{G_{\hat{\theta}^*}})-d_{\omega^*(\tilde{\theta})}(\hat{P}_r,\hat{P}_{G_{\tilde{\theta}}})\\
&\quad+d_{\omega^*(\tilde{\theta})}(\hat{P}_r,\hat{P}_{G_{\tilde{\theta}}})-d_{\omega^*(\tilde{\theta})}(\hat{P}_r,P_{G_{\tilde{\theta}}})\\
&\le d_{\omega^*(\hat{\theta}^*)}(\hat{P}_r,P_{G_{\hat{\theta}^*}})-d_{\omega^*(\hat{\theta}^*)}(\hat{P}_r,\hat{P}_{G_{\hat{\theta}^*}})\\
&\quad+d_{\omega^*(\tilde{\theta})}(\hat{P}_r,\hat{P}_{G_{\tilde{\theta}}})-d_{\omega^*(\tilde{\theta})}(\hat{P}_r,P_{G_{\tilde{\theta}}})\\
&=\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]+\mathbb{E}_{X\sim P_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]\\
&\quad-\Big(\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\hat{\theta}^*)}(X))]+\mathbb{E}_{X\sim \hat{P}_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]\Big)\\
&\quad+\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\tilde{\theta})}(X))]+\mathbb{E}_{X\sim \hat{P}_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))]\\
&\quad-\Big(\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega^*(\tilde{\theta})}(X))]+\mathbb{E}_{X\sim P_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))]\Big)\\
&=\mathbb{E}_{X\sim P_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]-\mathbb{E}_{X\sim \hat{P}_{G_{\hat{\theta}^*}}}[\psi(D_{\omega^*(\hat{\theta}^*)}(X))]\\
&\quad+\mathbb{E}_{X\sim \hat{P}_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))]-\mathbb{E}_{X\sim P_{G_{\tilde{\theta}}}}[\psi(D_{\omega^*(\tilde{\theta})}(X))]\\
&\le 2\sup_{\omega\in\Omega,\theta\in\Theta}\Big|\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_{\omega}(X))]-\mathbb{E}_{X\sim \hat{P}_{G_\theta}}[\psi(D_{\omega}(X))]\Big|. \qquad (34)
\end{aligned}$$

Combining (32)-(34), we obtain the following bound for (31):

$$\begin{aligned}
&d_{\omega^*(\hat{\theta}^*)}(P_r,P_{G_{\hat{\theta}^*}})-\inf_{\theta\in\Theta} d_{\omega^*(\theta)}(P_r,P_{G_\theta})\\
&\le 2\sup_{\omega\in\Omega}\Big|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega}(X))]-\mathbb{E}_{X\sim \hat{P}_r}[\phi(D_{\omega}(X))]\Big|\\
&\quad+2\sup_{\omega\in\Omega,\theta\in\Theta}\Big|\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_{\omega}(X))]-\mathbb{E}_{X\sim \hat{P}_{G_\theta}}[\psi(D_{\omega}(X))]\Big|\\
&=2\sup_{\omega\in\Omega}\left|\mathbb{E}_{X\sim P_r}[\phi(D_{\omega}(X))]-\frac{1}{n}\sum_{i=1}^{n}\phi(D_{\omega}(X_i))\right|\\
&\quad+2\sup_{\omega\in\Omega,\theta\in\Theta}\left|\mathbb{E}_{X\sim P_{G_\theta}}[\psi(D_{\omega}(X))]-\frac{1}{m}\sum_{j=1}^{m}\psi(D_{\omega}(X_j))\right|. \qquad (35)
\end{aligned}$$

Note that (35) is exactly the same bound as that in [15, Equation (34)]. Hence, the remainder of the proof follows from [15, Theorem 3], where $\alpha=\alpha_G$.

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

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] G. R. Kurri, M. Welfert, T. Sypherd, and L. Sankar, "α-gan: Convergence and estimation guarantees," *arXiv preprint arXiv:2205.06393*, 2022.