FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 50 (2025) No. 4
ISSN 0867-6356
DOI: 10.2478/fcds-2025-0018 e-ISSN 2300-3405

sciendo

f-divergence Analysis of Generative Adversarial Network

Mahmud Hasan*[1], Hailin Sang[2] *

Abstract. We aim to establish estimation bounds for various divergences, including total variation, Kullback-Leibler (KL) divergence, Hellinger divergence, and Pearson $\mathcal{X}^2$ divergence, within the GAN estimator. We derive an inequality based on empirical and population objective functions of the GAN model, achieving almost surely convergence rates. Subsequently, this inequality was employed to derive estimation bounds for total variation, Kullback-Leibler (KL) divergence, Hellinger divergence, and Pearson $\mathcal{X}^2$ divergence, leading to almost surely convergence rates and differences between the expected outputs of the discriminator on real data and generated data. Our study demonstrates better results compared to some existing ones, which are a specific case of the general objective function.

Keywords: discriminator, f-divergence, GAN, generator, neural network.

1. Introduction

In recent years, generative AI models have attracted significant attention across various communities due to their transformative potential in shaping the future of artificial intelligence. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014 [10], represent a prominent class of generative models based on an adversarial training framework between a generator and a discriminator neural network. The generator tries to produce realistic data samples, while the discriminator attempts to distinguish between real and generated data. This adversarial setup drives the generator to improve its outputs over time. GANs have achieved remarkable success in generating photorealistic human faces through advanced variants such as PG-GAN [15], StyleGAN [16], StyleGAN2 [17], StyleGAN3 [18], and Conditional GANs like cGAN [23]. GANs have been extensively applied in medical imaging tasks such as

*[1]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23219, USA, hasanm10@vcu.edu
[2] Department of Mathematics, University of Mississippi, University, MS 38677, USA, sang@olemiss.edu

tumor segmentation, MRI reconstruction, image denoising, and cross-modality synthesis using models like Pix2Pix [13], CycleGAN [35], SegAN [32], MedGAN [3], and DCGAN [26].

Due to the increasing applications of GANs, there is a growing need for foundational analysis, particularly in understanding estimation and generalization errors which remains an active area of research. Several theoretical advancements have been made in this direction. A detailed theoretical error analysis is provided in [12], while the notion that generalization can occur under a weaker metric known as the neural net distance is discussed in [4]. An f-divergence-based approach utilizing neural net distance is introduced by [21], and a gradient-based generalization framework grounded in neural net distance is presented in [14]. A recent study by the authors in [11] shows an improved theoretical error analysis based on the general objective function framework originally proposed in [4]. Our study in this paper focuses on deriving an f-divergence-based error bound for the GAN model under a general objective function, which will be discussed in more detail along with related existing work later in this section.

The goal of developing f-divergence bounds for GAN models comes from both theoretical and practical reasons. In theory, f-divergences give a general way to measure how different the model's distribution is from the real data distribution. This type of divergence includes many commonly used measures such as Kullback-Leibler, Jensen-Shannon, Hellinger, and total variation, which are often used when training different types of GANs. From an intuitive point of view, having bounds based on f-divergences helps us understand how close the generator is to learning the true data distribution, especially when we have a limited number of samples or the model is not perfect. These bounds are useful because they provide some guarantees on how much the generator can differ from the real data. They also help us understand the balance between sample size, model complexity, and training performance, which is important for building better and more stable GAN models.

We investigate the Generative Adversarial Network (GAN) models, where both the generator and discriminator are parameterized. Let the discriminator function class be given as

$$\mathcal{F} = \{f_w : \mathbb{R}^{p_0} \to \mathbb{R}\},$$

realized by a neural network with parameters $w \in \mathcal{W}$, which describe the weights of the network. Similarly, the generator neural network transformation class is defined as

$$\mathcal{G} = \{g_\theta(z) : \mathbb{R}^p \to \mathbb{R}^{p_0}\},$$

where $\theta \in \Theta$ are the generator weights.

Assume the random input variable Z follows the distribution $Z \sim \mu$, and the target variable X follows $X \sim \nu$. The distribution of the generators output, $g_\theta(Z)$, is denoted as g_θ^μ. To compare g_θ^μ and ν, we define the following objective function introduced in [10] and [4]:

$$d_{\mathcal{F},\phi}(g_\theta^\mu, \nu) := \sup_{w \in \mathcal{W}} |\mathbb{E}\phi(1 - f_w(g_\theta(Z))) + \mathbb{E}\phi(f_w(X))| - 2\phi(1/2), \tag{1}$$

where ϕ is referred to as the measuring function. We require ϕ to be monotone increasing. Common choices for ϕ in practice include $\phi(x) = x$, $\phi(x) = \log x$, and $\phi(x) = \log(\delta + (1-\delta)x)$ for some $0 < \delta < 1$.

The goal is to minimize this objective function:

$$\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^\mu, \nu).$$

Suppose we have n independent and identically distributed observations $X_i \sim \nu$ for $1 \le i \le n$, and the generator produces m independent and identically distributed terms $g_\theta(Z_j) \sim g_\theta^\mu$ for $1 \le j \le m$. Using empirical averages, we estimate the expectations in (1) and minimize the following empirical version:

$$d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n) = \sup_{w\in\mathcal{W}} \left| \frac{1}{m}\sum_{j=1}^{m} \phi(1 - f_w(g_\theta(Z_j))) + \frac{1}{n}\sum_{i=1}^{n} \phi(f_w(X_i)) \right| - 2\phi(1/2). \quad (2)$$

Here, $\hat{\mu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{Z_j}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ represent the empirical distributions of μ and ν, respectively. Equation (2) measures the distance between the empirical generator distribution $g_\theta^{\hat{\mu}_m}$ and the empirical target distribution $\hat{\nu}_n$. The error convergence rate for GANs in the context of f-divergence metrics, such as total variation (TV), Kullback-Leibler (KL) divergence, Hellinger square divergence, Pearson $\mathcal{X}^2$, has significant challenges. The concept of f-divergences traces back to the foundational work of Alfréd Rényi in 1961 [27], where he introduced a class of divergence measures to quantify differences between probability distributions. These divergences were later extended under the general f divergence framework by Csiszár [7] and others. Several well known examples of f divergences including the Kullback Leibler divergence [1], total variation distance [6], and Hellinger distance [5] are special cases of this framework. These f-divergence metrics can be used to define a comprehensive generalization error framework for GANs, despite the existing challenges and ongoing advancements in understanding GAN training. In this paper, we aim to present GANs convergence bound for various f-divergences by applying the objective functions defined in equations (1) and (2). These bounds can potentially help to explain the related errors. The objective function $d_{\mathcal{F},\phi}(g_\theta^\mu, \nu)$ in (1) aims to minimize as follows:

$$\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^\mu, \nu).$$

Let

$$\theta_1 = \arg\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n), \quad (3)$$

where $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in (2).

If $\rho_{\mu_\theta}(x)$ and $\rho_\nu(x)$ are the density functions for the two distributions g_θ^μ and ν, then for any convex function f, the f-divergence is defined as

$$d_f(g_\theta^\mu \,\|\, \nu) = \int \rho_\nu(x)\, f\left(\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}\right) dx. \quad (4)$$

Because θ_1 serves as the minimizer, $\rho_{\mu_{\theta_1}}(x)$ denotes the probability density function associated with the distribution $g^{\mu}_{\theta_1}$, where $g_{\theta_1}(Z)$ is the random variable. Consequently, the f-divergence for the estimator θ_1 can be expressed as:

$$d_f(g^{\mu}_{\theta_1}\|\nu) = \int \rho_\nu(x) f\left(\frac{\rho_{\mu_{\theta_1}(x)}}{\rho_\nu(x)}\right) dx. \tag{5}$$

The total variation (TV) distance [6] is a specific type of f-divergence that can be defined using function $f(x) = \frac{1}{2}|x-1|$ in the general equation for f-divergence in (4). Specifically, the TV distance between the distributions g^{μ}_{θ} and ν is given by:

$$\begin{aligned} d_f(g^{\mu}_{\theta}||\nu) = d_{TV}(g^{\mu}_{\theta}||\nu) &= \int \rho_\nu(x)\frac{1}{2}\left|\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)} - 1\right| dx \\ &= \frac{1}{2}\int |\rho_{\mu_\theta}(x) - \rho_\nu(x)| dx. \end{aligned}$$

The Kullback-Leibler (KL) divergence [1] between the distributions g^{μ}_{θ} and ν using the f-divergence in (4) with function $f(x) = x\log(x)$ can be expressed as:

$$\begin{aligned} d_f(g^{\mu}_{\theta}||\nu) = d_{KL}(g^{\mu}_{\theta}||\nu) &= \int \rho_\nu(x)\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}\log\left(\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}\right) dx \\ &= \int \rho_{\mu_\theta}(x)\log\left(\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}\right) dx. \end{aligned}$$

The Hellinger square divergence [5] between the distributions g^{μ}_{θ} and ν can be calculated using the given f-divergence in (4) for function $f(x) = (\sqrt{x}-1)^2$. The Hellinger square divergence for $f(x) = (\sqrt{x}-1)^2$ is derived as follows:

$$\begin{aligned} d_f(g^{\mu}_{\theta}||\nu) &= d^2_H(g^{\mu}_{\theta}||\nu) \\ &= \int \rho_\nu(x)\left(\sqrt{\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}} - 1\right)^2 dx \\ &= \int \rho_\nu(x)\left(\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)} - 2\sqrt{\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}} + 1\right) dx \\ &= \int \left(\rho_{\mu_\theta}(x) - 2\sqrt{\rho_\nu(x)\rho_{\mu_\theta}(x)} + \rho_\nu(x)\right) dx \\ &= \int \left(\sqrt{\rho_{\mu_\theta}(x)} - \sqrt{\rho_\nu(x)}\right)^2 dx. \end{aligned}$$

The χ^2 (Chi-squared) divergence [1] between the distributions g^{μ}_{θ} and ν is computed using the f-divergence in (4) function for $f(x) = (x-1)^2$ which can be written as follows:

$$d_f(g_\theta^\mu||\nu) = d_{\mathcal{X}^2}(g_\theta^\mu||\nu) = \int \rho_\nu(x)\left(\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)} - 1\right)^2 dx$$
$$= \int \frac{\left(\rho_{\mu_\theta}(x) - \rho_\nu(x)\right)^2}{\rho_\nu(x)} dx.$$

The f-divergences discussed above provide a framework for comparing probability distributions, and different f-divergences that can help to improve GAN model training performances. Those divergences can be written for the estimator θ_1 defined in (3) as follows:

$$d_{TV}(g_{\theta_1}^\mu||\nu) = \frac{1}{2}\int |\rho_{\mu_{\theta_1}}(x) - \rho_\nu(x)| dx.$$

$$d_{KL}(g_{\theta_1}^\mu||\nu) = \int \rho_{\mu_{\theta_1}}(x) \log\left(\frac{\rho_{\mu_{\theta_1}}(x)}{\rho_\nu(x)}\right) dx.$$

$$d_H^2(g_{\theta_1}^\mu||\nu) = \int \left(\sqrt{\rho_{\mu_{\theta_1}}(x)} - \sqrt{\rho_\nu(x)}\right)^2 dx.$$

$$d_{\mathcal{X}^2}(g_{\theta_1}^\mu||\nu) = \int \frac{\left(\rho_{\mu_{\theta_1}}(x) - \rho_\nu(x)\right)^2}{\rho_\nu(x)} dx.$$

Our study aims to determine convergence rates for the above f-divergence metrics using the objective function $d_{\mathcal{F},\phi}(g_\theta^\mu, \nu)$ defined in (1).

In prior research, such as [34], the authors investigated the KL divergence for the GAN estimator, using the objective function $d_{\mathcal{F}}(g_\theta^\mu, \nu)$ as defined in (20). Their study established that the upper bound of the KL divergence is determined by the Rademacher complexity of the discriminator class $\mathcal{F}$ for a sample size of n. Moreover, they noted that for the KL divergence to be bounded by $d_{\mathcal{F}}(g_\theta^\mu, \nu)$, the density ratio $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)$ must lie within the span of the set $\mathcal{F}$.

In [21], the author established estimation bounds for GANs concerning TV, KL divergence, and Hellinger square divergence as in [34]. The bounds derived in [21] depend on the sample sizes of the discriminator and generator, denoted by n and m, respectively. Notably, both studies, [34] and [21], concentrated on the objective function where $\phi(x) = x$.

The key question in the previous work is whether we can find the TV, KL, Hellinger divergence, and Pearson $\mathcal{X}^2$ divergence convergence rate for the general case of $\phi(x)$, utilizing the neural network structure described in [11]. Our investigation aims to broaden this scope to those divergences, with a similar approach presented in [21]. In our study, these divergence has almost surely convergence rates. Furthermore, the divergence bound in this paper applies to the general objective function in (1) for $\phi(x)$, while prior research in [21] and [34] focuses on the specific case where $\phi(x) = x$.

As a part of the detailed proof, we begin by establishing an inequality for the objective function $d_{\mathcal{F},\phi}(g_\theta^\mu, \nu)$ as defined in (1). The inequality asserts that $d_{\mathcal{F},\phi}(g_{\theta_1}^\mu, \nu)$

is bounded by the sum of three empirical and population objective functions, each corresponding to the discriminator, generator, and their combination. Subsequently, we show that these empirical objective functions resemble empirical processes. These processes can be used with Talagrand's inequality for getting almost sure convergence rates, as explained in [11], for two types of functions: $\mathcal{F}_1$ and $\mathcal{H}$, both structured as neural networks.

This paper organization is outlined as follows. In Section 2, we establish an inequality for the GAN estimator of the objective function $d_{\mathcal{F},\phi}(g_\theta^\mu, \nu)$ in (1) which will be applied for the main results. In Section 3, we outline the main results of convergence rates for TV, KL, Hellinger square, Pearson $\mathcal{X}^2$ divergence with the supporting proofs and discussions. Section 4 concludes the discussion and suggests directions for future research.

2. Inequality for GAN estimator

In this section, we derive an inequality that is bounded by three objective functions and later we have almost sure convergence rates for two empirical objective function. A similar inequality was derived in [21] and [12] for the vanilla GAN objective function. we derive the inequality for the GANs estimator θ_1, which is based on the following lemma. This lemma is utilized in a subsequent section to determine the convergence rate of various f-divergence metrics.

Lemma 2.1 *Let θ_1 be the solution to $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$, where $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is defined in (2). Denote $g_{\theta_1}^\mu$ as the probability distribution of $g_{\theta_1}(Z) \in \mathcal{G}$. Let $\mathcal{F}$ and $\mathcal{G}$ represent the discriminator and generator function classes as defined in [11]. Let $X \sim \nu$ be the target distribution, with $\|X\| < \infty$, and assume that the random input variable Z satisfies $\|Z\| < \infty$. Under these conditions, it follows that:*

$$d_{\mathcal{F},\phi}(g_{\theta_1}^\mu, \nu) \leq O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}} + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^\mu, \nu). \tag{6}$$

Proof. The inequality can be derived by employing the properties of supremum and infimum. Assuming the discriminator and generator sample sizes are n and m, respectively, and utilizing the property $\sup|f(\cdot) + g(\cdot)| \leq \sup|f(\cdot)| + \sup|g(\cdot)|$, the

derivation unfolds as follows:

$$
\begin{aligned}
& d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu) \\
&= \sup_{w\in\mathcal{W}} |\mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) + \mathbb{E}\phi(f_w(X))| - 2\phi(1/2) \\
&= \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z))) \\
&\quad + \mathbb{E}\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| - 2\phi(1/2) \\
&\le \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| \\
&\quad + \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| - 2\phi(1/2) \\
&\le \sup_{\theta\in\Theta, w\in\mathcal{W}} \Big|\mathbb{E}\phi(1 - f_w(g_{\theta}(Z))) - \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta}(Z)))\Big| \\
&\quad + \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| - 2\phi(1/2)
\end{aligned}
$$

$$
\begin{aligned}
&\le \sup_{\theta\in\Theta} d_{\mathcal{F}\circ\mathcal{G},\phi}(g_{\theta}^{\hat{\mu}_m}, g_{\theta}^{\mu}) + \\
&\quad \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(f_w(X)) - \hat{\mathbb{E}}_n\phi(f_w(X)) + \hat{\mathbb{E}}_n\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| \\
&\quad - 2\phi(1/2) \\
&\le \sup_{\theta\in\Theta} d_{\mathcal{F}\circ\mathcal{G},\phi}(g_{\theta}^{\hat{\mu}_m}, g_{\theta}^{\mu}) + \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(f_w(X)) - \hat{\mathbb{E}}_n\phi(f_w(X))\Big| \\
&\quad + \sup_{w\in\mathcal{W}} \Big|\hat{\mathbb{E}}_n\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta_1}(Z)))\Big| - 2\phi(1/2) \\
&= \sup_{\theta\in\Theta} d_{\mathcal{F}\circ\mathcal{G},\phi}(g_{\theta}^{\hat{\mu}_m}, g_{\theta}^{\mu}) + d_{\mathcal{F},\phi}(\hat{\nu}_n, \nu) \\
&\quad + \inf_{\theta\in\Theta} \sup_{w\in\mathcal{W}} \Big|\hat{\mathbb{E}}_n\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta}(Z)))\Big| - 2\phi(1/2) \\
&= \sup_{\theta\in\Theta} d_{\mathcal{F}\circ\mathcal{G},\phi}(g_{\theta}^{\hat{\mu}_m}, g_{\theta}^{\mu}) + d_{\mathcal{F},\phi}(\hat{\nu}_n, \nu) + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_{\theta}^{\hat{\mu}_m}, \hat{\nu}_n). \qquad (7)
\end{aligned}
$$

Here we defined the following:

$$
d_{\mathcal{F}\circ\mathcal{G},\phi}(g_{\theta}^{\hat{\mu}_m}, g_{\theta}^{\mu}) = \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(1 - f_w(g_{\theta}(Z))) - \hat{\mathbb{E}}_m\phi(1 - f_w(g_{\theta}(Z)))\Big|, \qquad (8)
$$

$$
d_{\mathcal{F},\phi}(\hat{\nu}_n, \nu) = \sup_{w\in\mathcal{W}} \Big|\mathbb{E}\phi(f_w(X)) - \hat{\mathbb{E}}_n\phi(f_w(X))\Big|, \qquad (9)
$$

and $d_{\mathcal{F},\phi}(g_{\theta}^{\hat{\mu}_m}, \hat{\nu}_n)$ is defined in (2). The inequality (7) demonstrates that the objective function of GAN estimation can be constructed into empirical versions of the objective

functions $d_{\mathcal{F}\circ\mathcal{G},\phi}(g_\theta^{\hat{\mu}_m}, g_\theta^{\mu})$, $d_{\mathcal{F},\phi}(\hat{\nu}_n, \nu)$, $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ which are associated with the discriminator class $\mathcal{F}$, generator class $\mathcal{G}$, and their composition $\mathcal{F}\circ\mathcal{G}$. We define as $u = (\theta, w)$ and $U = (\Theta, \mathcal{W})$

$$\mathcal{H} = \{h_u(Z) : h_u(Z) = \phi(1 - f_w(g_\theta(Z)))\} \tag{10}$$

and

$$\mathcal{F}_1 = \{f_1(X) : f_1(X) = \phi(f_w(X))\}. \tag{11}$$

Then, according to Theorem 3.1 and Theorem 3.2 in [11], we can write the following:

$$\begin{aligned}\sup_{\theta\in\Theta} d_{\mathcal{F}\circ\mathcal{G},\phi}(g_\theta^{\hat{\mu}_m}, g_\theta^{\mu}) &= \sup_{\theta\in\Theta, w\in\mathcal{W}} \left|\mathbb{E}\phi(1 - f_w(g_\theta(Z))) - \hat{\mathbb{E}}_m\phi(1 - f_w(g_\theta(Z)))\right| \\ &= \sup_{u\in U}\left|\frac{1}{m}\sum_{j=1}^{m}(h_u(Z_j) - \mathbb{E}h_u(Z_j))\right| \\ &= O_{a.s.}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}, \end{aligned} \tag{12}$$

$$\begin{aligned} d_{\mathcal{F},\phi}(\hat{\nu}_n, \nu) &= \sup_{w\in\mathcal{W}}\left|\mathbb{E}\phi(f_w(X)) - \hat{\mathbb{E}}_n\phi(f_w(X))\right| \\ &= \sup_{w\in\mathcal{W}}\left|\frac{1}{n}\sum_{i=1}^{n}(f_1(X_i) - \mathbb{E}f_1(X_i))\right| = O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}, \end{aligned} \tag{13}$$

and using the Lemma 3.1 in [11], we have,

$$\begin{aligned}\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n) &= \inf_{\theta\in\Theta}\sup_{w\in\mathcal{W}}\left|\hat{\mathbb{E}}_n\phi(f_w(X)) + \hat{\mathbb{E}}_m\phi(1 - f_w(g_\theta(Z)))\right| - 2\phi(1/2) \\ &\leq O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}} + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\mu}, \nu). \end{aligned} \tag{14}$$

Substituting (12), (13), and (14) in (7) the proof is complete. ■

3. Convergence rate of various f-divergence metrics

The analysis of f-divergence metrics for GAN estimators provides valuable insights into error analysis, model performance, and convergence, for improving the GAN training process and enhancing the quality of generated samples. We are doing this analysis by deriving different divergence bounds for the GAN estimator defined as:

$$\theta_1 = \arg\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n),$$

where $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in Equation (2).

We present the GAN convergence rate for various divergence measures, including total variation (TV), Kullback-Leibler (KL) divergence, Hellinger square divergence and Pearson's χ^2 divergence.

3.1. Total variation (TV) and Kullback-Leibler (KL) divergence bounds

The total variation (TV) divergence measures the difference between two probability distributions. Let $\rho_{\mu_{\theta_1}}(x)$ and $\rho_\nu(x)$ be the density functions for $g^\mu_{\theta_1}$ and ν, respectively. The total variation (TV) divergence and Kullback-Leibler (KL) divergence are defined for the estimator θ_1 as follows:

$$d_{TV}(g^\mu_{\theta_1}\|\nu) = \frac{1}{2}\int |\rho_{\mu_{\theta_1}}(x) - \rho_\nu(x)|\, dx, \tag{15}$$

$$d_{KL}\left(g^\mu_{\theta_1}\|\nu\right) = \int \rho_{\mu_{\theta_1}}(x) \log\left(\frac{\rho_{\mu_{\theta_1}}(x)}{\rho_\nu(x)}\right) dx. \tag{16}$$

Below, we present some findings from the literature regarding the KL divergence bound, along with the corresponding definitions.

Definition 3.1 *Pseudo-dimension [31]: Let $F = \{f : \Omega \to \mathbb{R}\}$ be a class of functions. The pseudo-dimension of F, denoted by $Pdim(F)$, is the largest integer m such that: $\exists(X_i, y_i) \in \Omega \times \mathbb{R}$, $i \in [m]$, for any $(b_1, \ldots, b_m) \in \{-1, 1\}^m$ there exists $f \in F$ such that $sign(f(X_i) - y_i) = b_i$, $\forall i \in [m]$. We use the following notation for the composition of function classes:*

$$F \circ G := \{f \circ g \,|\, f \in F, g \in G\}.$$

Definition 3.2 *[29] For each $g \in span\mathcal{F}$ that can be decomposed into $g = \sum_{i=1}^k w_i f_i + w_0$, the $\mathcal{F}$-variation norm $\|g\|_{\mathcal{F},1}$ of g is the infimum of $\sum_{i=1}^k |w_i|$ among all possible decompositions of g, that is,*

$$\|g\|_{\mathcal{F},1} = \inf\left\{\sum_{i=1}^k |w_i| : g = \sum_{i=1}^k w_i f_i + w_0, \forall k \in \mathbb{N}, w_0, w_i \in \mathbb{R}, f_i \in \mathcal{F}\right\}.$$

Proposition 3.1 *[34] Assume that g^μ_θ and ν have positive density functions $\rho_{\mu_\theta}(x)$ and $\rho_\nu(x)$, respectively. Then*

$$d_{KL}(\nu\|g^\mu_\theta) + d_{KL}(g^\mu_\theta\|\nu) = E_\nu\left[\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)\right] - E_{g^\mu_\theta}\left[\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)\right].$$

If $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right) \in span\mathcal{F}$, then

$$d_{KL}(\nu\|g^\mu_\theta) + d_{KL}(g^\mu_\theta\|\nu) \le \|\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)\|_{\mathcal{F},1} d_{\mathcal{F}}(g^\mu_\theta, \nu). \tag{17}$$

In this proposition, $\|.\|_{\mathcal{F},1}$ denotes the variation norm as defined in Definition (3.2), and $d_{\mathcal{F}}(g_\theta^\mu, \nu)$ represents the objective function defined in Equation (20) when $\phi(x) = x$. The result presented in (17) highlights the necessity for the log density ratio $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)$ to lie within the space spanned by $\mathcal{F}$ in order to bound the KL divergence with $d_{\mathcal{F}}(g_\theta^\mu, \nu)$.

Let

$$\hat{\theta} = \inf_{\theta\in\Theta} d_{\mathcal{F}}(g_\theta^\mu, \hat{\nu}_n)$$

where $d_{\mathcal{F}}(g_\theta^\mu, \hat{\nu}_n)$ is defined as

$$d_{\mathcal{F}}(g_\theta^{\mu_m}, \hat{\nu}_n) = \sup_{w\in\mathcal{W}} \left[f_w(g_\theta(Z_j)) - \frac{1}{n}\sum_{i=1}^{n} f_w(X_i) \right]. \tag{18}$$

Then, the KL divergence bound is investigated in [34] and [21] through the following corollary and theorem.

Corollary 3.1 *[34] Assume that g_θ^μ and ν have positive density functions $\rho_{\mu_\theta}(x)$ and $\rho_\nu(x)$, respectively. Assume $\mathcal{F}$ consists of bounded functions with $\Delta := \sup_{f\in\mathcal{F}} ||f||_\infty < \infty$. Further, assume the discriminator set $\mathcal{F}$ is compatible with the generator set $\mathcal{G}$ in the sense that $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right) \in span(\mathcal{F})$, $\forall g_\theta^\mu \in \mathcal{G}$, with a compatible coefficient defined as*
$\Lambda_{\mathcal{F},\mathcal{G}} := \sup_{g_\theta^\mu\in\mathcal{G}} ||\log(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)})||_{\mathcal{F},1} < \infty$. *Then*

$$d_{KL}(\nu||g_{\hat{\theta}}^\mu) \le \Lambda_{\mathcal{F},\mathcal{G}}(2R_n(\mathcal{F}) + 2\Delta\sqrt{\frac{2\log(1/\delta)}{n}} + \Delta \inf_{g_\theta^\mu\in\mathcal{G}} \sqrt{d_{KL}(\nu||g_\theta^\mu)} + \epsilon), \tag{19}$$

where $R_n(\mathcal{F})$ is the Rademacher complexity defined as:

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{w\in\mathcal{W}} \frac{2}{n}\sum_i \tau_i f_w(X_i)\right].$$

The bound in (19) depends on the compatibility coefficient $\Lambda_{\mathcal{F},\mathcal{G}}$, introducing a trade-off: $\mathcal{G}$ should be small with well-behaved density functions to ensure a small $\Lambda_{\mathcal{F},\mathcal{G}}$. On the other hand, it should be large enough to minimize the modeling error $\inf_{\nu\in\mathcal{G}} \mathbb{E}_\mu[d_{KL}(\nu||g_\theta^\mu)]$. In a related manner, the discriminator set should be sufficiently large to include all density ratios $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)$ within a ball of radius $\Lambda_{\mathcal{F},\mathcal{G}}$ spanned by $\mathcal{F}$. Additionally, it should be small enough to maintain a low Rademacher complexity $R_\mu^{(m)}(\mathcal{F})$.

If $\phi(x) = x$, the expression $d_{\mathcal{F},\phi}(g_\theta^\mu, \nu)$ from equation (1), with a slight abuse of notation, can be written as follows:

$$d_{\mathcal{F}}(g_\theta^\mu, \nu) = \sup_{w\in\mathcal{W}} \big[\mathbb{E}f_w(g_\theta(Z)) - \mathbb{E}f_w(X)\big]. \tag{20}$$

Similarly, the equation in (2) can be reformulated as:

$$d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n) = \sup_{w \in \mathcal{W}} \left[\hat{\mathbb{E}}_m f_w(g_\theta(Z)) - \hat{\mathbb{E}}_n f_w(X)\right],$$

i.e.,

$$d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n) = \sup_{w \in \mathcal{W}} \left[\frac{1}{m}\sum_{j=1}^{m} f_w(g_\theta(Z_j)) - \frac{1}{n}\sum_{i=1}^{n} f_w(X_i)\right]. \tag{21}$$

Theorem 3.1 *[21] Let θ_3 be the solution of $\inf_{\theta\in\Theta} d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ where $d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in* (21) *and K_1 is some constant for $\|f_w\| \le K_1$, m, and n denote the number of generator samples and target distribution samples. For total variation distance, and Kullback-Leibler divergence,*

$$\begin{aligned} 4\mathbb{E}d_{TV}^2(g_{\theta_3}^{\mu}||\nu) &\le 2\left[\mathbb{E}d_{KL}\left(\nu||g_{\theta_3}^{\mu}\right) + \mathbb{E}d_{KL}\left(g_{\theta_3}^{\mu}||\nu\right)\right] \\ &\le 2\sup_{\theta}\inf_{w}\left\{\left\|\log\frac{\rho_\nu(x)}{\rho_{\mu_\theta(x)}} - f_w(x)\right\|_\infty\right\} + \frac{K_1}{\sqrt{2}}\inf_{\theta}\sqrt{\left\|\log\frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}\right\|_\infty} \\ &\quad + C\sqrt{\frac{Pdim(\mathcal{F}\circ\mathcal{G})\ln m}{m}} + C\sqrt{\frac{Pdim(\mathcal{F})\ln n}{n}} + CK_1\sqrt{\frac{Pdim(\mathcal{F})\ln m}{m}}, \end{aligned} \tag{22}$$

where $C > 0$ is some universal constant independent of $Pdim(\mathcal{F})$, $Pdim(\mathcal{F}\circ\mathcal{G})$, m, n.

We establish bounds on the total variation (TV) and Kullback-Leibler (KL) divergence of the objective function in GAN estimation, represented by Theorem 3.2, for the general case of $\phi(x)$. This relationship is typically expressed using Pinsker's inequality. The proof further employs the inequality from Lemma 2.1.

Theorem 3.2 *Let θ_1 be the solution of $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ where $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in* (2)*. If $\rho_\nu(x)$ and $\rho_{\mu_{\theta_1}}(x)$ are the density functions for the distribution of ν and $g_{\theta_1}^{\mu}(Z)$, respectively, then for the total variation divergence and Kullback-Leibler divergence, we have:*

$$\begin{aligned} 4\mathbb{E}d_{TV}^2(g_{\theta_1}^{\mu}||\nu) &\le 2\mathbb{E}\left[d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right) + d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)\right] \\ &\le 2\sup_{\theta\in\Theta}\left\{\left\|\log\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)} - \phi(f_w(x))\right\|_\infty\right\} \\ &\quad + O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}} + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\mu}, \nu) + 2\phi(1/2). \end{aligned}$$

Proof. Using Pinsker's inequality proved in [30] Theorem 4.8, we have:

$$d_{TV}(g_{\theta_1}^{\mu}||\nu) = \frac{1}{2}\int|\rho_{\mu_{\theta_1}}(x) - \rho_\nu(x)|dx \le \sqrt{\frac{1}{2}d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right)}.$$

Then

$$2d_{TV}^2(g_{\theta_1}^{\mu}||\nu) \leq d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right). \tag{23}$$

Similarly,

$$d_{TV}(g_{\theta_1}^{\mu}||\nu) = \frac{1}{2}\int |\rho_{\mu_{\theta_1}}(x) - \rho_{\nu}(x)|dx \leq \sqrt{\frac{1}{2}d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)}.$$

Then,

$$d_{TV}(g_{\theta_1}^{\mu}||\nu) \leq \sqrt{\frac{1}{2}d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)},$$
$$2d_{TV}^2(g_{\theta_1}^{\mu}||\nu) \leq d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right). \tag{24}$$

Therefore adding (23) and (24), we have,

$$4d_{TV}^2(g_{\theta_1}^{\mu}||\nu) \leq \left[d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right) + d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)\right].$$

$$\begin{aligned}
4d_{TV}^2(g_{\theta_1}^{\mu}||\nu) &\leq \left[d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right) + d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)\right] \\
&= \int \rho_{\nu}(x)\log\frac{\rho_{\nu}(x)}{\rho_{\mu_{\theta_1}}(x)}dx + \int \rho_{\mu_{\theta_1}}(x)\log\frac{\rho_{\mu_{\theta_1}}(x)}{\rho_{\nu}(x)}dx \\
&= \int \rho_{\nu}(x)\log\frac{\rho_{\nu}(x)}{\rho_{\mu_{\theta_1}}(x)}dx - \int \rho_{\mu_{\theta_1}}(x)\log\frac{\rho_{\nu}(x)}{\rho_{\mu_{\theta_1}}(x)}dx \\
&= \int \log\frac{\rho_{\nu}(x)}{\rho_{\mu_{\theta_1}}(x)}\left(\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)\right)dx \\
&= \int \left(\log\frac{\rho_{\nu}(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x))\right)\left(\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)\right)dx \\
&+ \int \phi(f_w(x))\left(\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)\right)dx
\end{aligned}$$

$$
\begin{aligned}
&= \int \left(\log \frac{\rho_\nu(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x)) \right) \left(\rho_\nu(x) - \rho_{\mu_{\theta_1}}(x) \right) dx \\
&+ \mathbb{E}\phi(f_w(X)) - \mathbb{E}\phi(f_w(g_{\theta_1}(Z))) \\
&\leq \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x)) \right\|_\infty \left\| \rho_\nu - \rho_{\mu_{\theta_1}} \right\|_1 + \mathbb{E}\phi(f_w(X)) \\
&+ \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \mathbb{E}\phi(f_w(g_{\theta_1}(Z))), \text{ For } \phi(x) > 0 \\
&\leq \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x)) \right\|_\infty \left\| \rho_\nu - \rho_{\mu_{\theta_1}} \right\|_1 \\
&+ \sup_{w \in \mathcal{W}} \left[\mathbb{E}\phi(f_w(X)) + \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) \right] \\
&= \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x)) \right\|_\infty \left\| \rho_\nu - \rho_{\mu_{\theta_1}} \right\|_1 + d_{\mathcal{F},\phi}(g^\mu_{\theta_1}, \nu) + 2\phi(1/2).
\end{aligned}
$$

Given that both $g^\mu_{\theta_1}$ and ν represent probability distributions, it can be concluded that $\left\| \rho_\nu - \rho_{\mu_{\theta_1}} \right\|_1 \leq 2$. Utilizing Lemma 2.1 for $d_{\mathcal{F},\phi}(g^\mu_{\theta_1}, \nu)$ and calculating the expected value, we can express it as follows:

$$
\begin{aligned}
4\mathbb{E} d^2_{TV}(g^\mu_{\theta_1} || \nu) &\leq \mathbb{E}\left[d_{KL}\left(\nu || g^\mu_{\theta_1}\right) + d_{KL}\left(g^\mu_{\theta_1} || \nu\right) \right] \\
&\leq 2\mathbb{E}\left\{ \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_{\theta_1}}(x)} - \phi(f_w(x)) \right\|_\infty \right\} \\
&+ O_{a.s}\left(\frac{\log m}{m} \right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n} \right)^{\frac{1}{2}} + \inf_{\theta \in \Theta} d_{\mathcal{F},\phi}(g^\mu_\theta, \nu) + 2\phi(1/2) \\
&\leq 2 \sup_{\theta \in \Theta} \left\{ \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)} - \phi(f_w(x)) \right\|_\infty \right\} \\
&+ O_{a.s}\left(\frac{\log m}{m} \right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n} \right)^{\frac{1}{2}} + \inf_{\theta \in \Theta} d_{\mathcal{F},\phi}(g^\mu_\theta, \nu) + 2\phi(1/2).
\end{aligned} \tag{25}
$$

■

The term $\sup_{\theta \in \Theta} \left\{ \left\| \log \frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)} - \phi(f_w(x)) \right\|_\infty \right\}$ represents the maximum difference, measured using the sup norm, between the logarithm of the density ratio of the target and generated distributions and the discriminator as an input of ϕ, considering all possible generator parameters.

The term $\inf_{\theta \in \Theta} d_{\mathcal{F},\phi}(g^\mu_\theta, \nu)$ signifies the minimum difference between the generated distribution and the target distribution that can be achieved by tuning the parameters of the generator within the given parameter space Θ. It serves as a key measure of how well the GAN model can approximate the target distribution.

The rates $O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}+O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}$ denote the convergence rates to the sample sizes m and n, respectively. They indicate how quickly the GAN converges to its equilibrium state as the number of samples increases. Higher convergence rates imply faster convergence of the GAN model.

Remark 3.1 *In* (25)*, we derive the almost sure bound for total variation (TV) and KL divergence. This bound takes into consideration both the discriminator and generator sample sizes* m *and* n *and is applicable to the general form of* $\phi(x)$*. Furthermore, when* $\phi(x)=x$*, we obtain an improved result for TV and KL divergence, as shown in the following Corollary 3.2.*

Corollary 3.2 *Let* θ_1 *be the solution of* $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m},\hat{\nu}_n)$ *where* $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m},\hat{\nu}_n)$ *is given in* (2)*. Denote the density functions of* ν *and* $g_{\theta_1}^{\mu}(Z)$ *by* $\rho_\nu(x)$ *and* $\rho_{\mu_{\theta_1}}(x)$*, respectively. If* $\phi(x)=x$ *and* $\log\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}=f_w(x)$ *as in [21], then for both total variation divergence and Kullback-Leibler divergence, we have:*

$$\begin{aligned}4\mathbb{E}d_{TV}^2(g_{\theta_1}^{\mu}||\nu) &\le 2\mathbb{E}\left[d_{KL}\left(\nu||g_{\theta_1}^{\mu}\right)+d_{KL}\left(g_{\theta_1}^{\mu}||\nu\right)\right]\\ &\le O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}+O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\\ &+\inf_{\theta\in\Theta}\sup_{w\in\mathcal{W}}\left|\mathbb{E}f_w(X))-\mathbb{E}f_w(g_\theta(Z)))\right|. \qquad (26)\end{aligned}$$

By substituting $\phi(x)=x$ and $\log\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}=f_w(x)$ in (25), the proof of Corollary 3.2 is completed. The term $|\mathbb{E}f_w(X))-\mathbb{E}f_w(g_\theta(Z)))|$ captures the difference between the expected outputs of the discriminator on real data (X) and generated data ($g_\theta(Z)$). The term $\sup_{w\in\mathcal{W}}|\mathbb{E}[f_w(X)]-\mathbb{E}[f_w(g_\theta(Z))]|$ represents the maximum difference between the discriminator's responses to real and generated data across all possible choices of parameters w within the parameter space $\mathcal{W}$. The expression $\inf_{\theta\in\Theta}\sup_{w\in\mathcal{W}}|\mathbb{E}[f_w(X)]-\mathbb{E}[f_w(g_\theta(Z))]|$ represents the smallest possible difference between the maximum expected responses of the discriminator function f_w to real and generated data over all possible choices of parameters θ within the parameter space Θ.

Remark 3.2 *The bound in* (26) *states that the total variation divergence has a.s. results when the condition* $\log\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}=f_w(x)$ *holds. This result is better than the existing result outlined in Theorem 3.1. Theorem 3.1 demonstrates that the bound in* (22) *relies on the sample sizes* m *and* n *of the discriminator and generator, respectively, along with the pseudo-dimension under the condition* $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)=f_w(x)$*.*

The bound in (17) *depends on the variation norm of* $\log\left(\frac{\rho_\nu(x)}{\rho_{\mu_\theta}(x)}\right)$ *and* $d_{\mathcal{F}}(g_\theta^{\mu},\nu)$*. In* (19)*, the bound only considers the sample size of the discriminator and Rademacher complexity. It's worth noting that both results in* (17) *and* (19) *concern the GAN*

estimator $\hat{\theta}$ *in estimating* $\inf_{\theta\in\Theta} d_{\mathcal{F}}(g_\theta^\mu, \hat{\nu}_n)$*, while in our context, the GAN estimator* θ_1 *is used for* $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$.

3.2. Hellinger divergence bound

If $\rho_\nu(x)$ and $\rho_{\nu_{\theta_1}}(x)$ represent the probability density functions for the distributions of ν and $g_{\theta_1}^\mu$ respectively, then the Hellinger divergence is defined as follows:

$$d_H(g_{\theta_1}^\mu, \nu) = \left(\int \left(\sqrt{\rho_{\mu_{\theta_1}}(x)} - \sqrt{\rho_\nu(x)} \right)^2 \right)^{1/2}. \tag{27}$$

Our goal is to determine the Hellinger divergence bound for the general objective function used in GAN estimation, using a technique similar to that outlined in Theorem 3.2. In [21], one of the findings regarding the Hellinger divergence bound is presented in the following Theorem 3.3, which represents the specific case where $\phi(x) = x$.

Theorem 3.3 *[21] Let* θ_3 *be the solution of* $\inf_{\theta\in\Theta} d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ *where* $d_{\mathcal{F}}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ *is given in* (21) *and* K_1 *is some constant for* $\|f_w\| \le K_1$*,* m*, and* n *denote the number of generator samples and target distribution samples. Then,*

$$\begin{aligned}
&4\mathbb{E} d_H^2(g_{\theta_3}^\mu || \nu) \\
&\le 2 \sup_\theta \inf_w \left\{ \left\| \frac{\sqrt{\rho_{\mu_\theta}(x)} - \sqrt{\rho_\nu(x)}}{\sqrt{\rho_{\mu_\theta}(x)} - \sqrt{\rho_\nu(x)}} - f_w(x) \right\|_\infty \right\} \\
&+ 2K_1 \inf_\theta \sqrt{\left\| \frac{\sqrt{\rho_{\mu_\theta}(x)} - \sqrt{\rho_\nu(x)}}{\sqrt{\rho_{\mu_\theta}(x)} - \sqrt{\rho_\nu(x)}} \right\|_\infty} + C\sqrt{\frac{Pdim(\mathcal{F}\circ\mathcal{G}) \ln m}{m}} \\
&+ C\sqrt{\frac{Pdim(\mathcal{F}) \ln n}{n}} + C\sqrt{\frac{Pdim(\mathcal{F}) \ln m}{m}},
\end{aligned} \tag{28}$$

where $C > 0$ *is some universal constant independent of* $Pdim(\mathcal{F})$*,* $Pdim(\mathcal{F}\circ\mathcal{G})$*,* m *and* n.

We establish the Hellinger divergence bound of the objective function in GAN estimation, stated in Theorem 3.4, for the general case of $\phi(x)$.

Theorem 3.4 *Let* θ_1 *be the solution of* $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ *where* $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ *is given in* (2)*. If* $\rho_\nu(x)$ *and* $\rho_{\mu_{\theta_1}}(x)$ *are the density function for the distribution of* ν *and* $g_{\theta_1}^\mu$*. Then*

$$\begin{aligned}
\mathbb{E} d_H^2(g_{\theta_1}^\mu, \nu) \le\ & 2 \sup_{\theta\in\Theta} \left\{ \left\| \frac{(\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)})}{(\sqrt{\rho_{\mu_\theta(x)}} + \sqrt{\rho_\nu(x)})} - \phi(f_w(x)) \right\|_\infty \right\} \\
&+ O_{a.s}\left(\frac{\log m}{m} \right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n} \right)^{\frac{1}{2}} + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^\mu, \nu) + 2\phi(1/2).
\end{aligned} \tag{29}$$

Proof. We apply a minor simplification to the Hellinger divergence to align it with the objective function of GAN. Subsequently, we utilize the bound derived from Lemma 2.1 to complete the proof.

$$
\begin{aligned}
d_H^2(g_{\theta_1}^{\mu}, \nu) &= \int \left(\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}\right)^2 dx \\
&= \int \frac{\left(\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}\right)^2 \left(\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}\right)}{\left(\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}\right)} dx \\
&= \int \frac{\left(\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}\right) \left(\rho_\nu(x) - \rho_{\mu_{\theta_1}}(x)\right)}{\left(\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}\right)} dx \\
&= \int \left(\frac{\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}} - \phi(f_w(x)) \right) \left(\rho_\nu(x) - \rho_{\mu_{\theta_1}}(x)\right) dx \\
&+ \int \phi(f_w(x)) \left(\rho_\nu(x) - \rho_{\mu_{\theta_1}}(x)\right) dx \\
&= \int \left(\frac{\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}} - \phi(f_w(x)) \right) \left(\rho_\nu(x) - \rho_{\mu_{\theta_1}}(x)\right) dx \\
&+ \mathbb{E}\phi(f_w(X)) + \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) \\
&- \mathbb{E}\phi(f_w(g_{\theta_1}(Z))) \\
&\leq \left\| \frac{\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}} - \phi(f_w(x)) \right\|_\infty \left\| \rho_\nu(x) - \rho_{\mu_{\theta_1}}(x) \right\|_1 \\
&+ d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu) - \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \mathbb{E}\phi(f_w(g_{\theta_1}(Z))) \\
&\leq 2 \left\| \frac{\sqrt{\rho_\nu(x)} - \sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_{\theta_1}}(x)} + \sqrt{\rho_\nu(x)}} - \phi(f_w(x)) \right\|_\infty + d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu) + 2\phi(1/2).
\end{aligned}
$$

Using the Lemma 2.1 for $d_{\mathcal{F},\phi}(g^{\mu}_{\theta_1},\nu)$ and taking expectation on both sides we can write as follows:

$$\begin{aligned}\mathbb{E}d^2_H(g^{\mu}_{\theta_1},\nu) &\leq 2\mathbb{E}\left\{\left\|\frac{\sqrt{\rho_\nu(x)}-\sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_{\theta_1}}(x)}+\sqrt{\rho_\nu(x)}}-\phi(f_w(x))\right\|_\infty\right\}\\ &\quad+O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}+O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}+\inf_{\theta\in\Theta}d_{\mathcal{F},\phi}(g^{\mu}_{\theta},\nu)+2\phi(1/2)\\ &\leq 2\sup_{\theta\in\Theta}\left\{\left\|\frac{\sqrt{\rho_\nu(x)}-\sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_\theta}(x)}+\sqrt{\rho_\nu(x)}}-\phi(f_w(x))\right\|_\infty\right\}\\ &\quad+O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}+O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}+\inf_{\theta\in\Theta}d_{\mathcal{F},\phi}(g^{\mu}_{\theta},\nu)+2\phi(1/2). \quad (30)\end{aligned}$$

■

The result in (30) provides an a.s. bound, applicable to the general form of $\phi(x)$, taking into account both the discriminator and generator sample sizes. Under the condition $\frac{\sqrt{\rho_\nu(x)}-\sqrt{\rho_{\mu_{\theta_1}}(x)}}{\sqrt{\rho_{\mu_\theta}(x)}+\sqrt{\rho_\nu(x)}}=f_w(x)$, we obtain an improved result when $\phi(x)=x$, which is given by the following Corollary 3.3.

Corollary 3.3 *Let θ_1 be the solution of $\inf_{\theta\in\Theta}d_{\mathcal{F},\phi}(g^{\hat{\mu}_m}_{\theta},\hat{\nu}_n)$ where $d_{\mathcal{F},\phi}(g^{\hat{\mu}_m}_{\theta},\hat{\nu}_n)$ is given in* (2)*. Denote the density functions of ν and $g^{\mu}_{\theta_1}(Z)$ by $\rho_\nu(x)$ and $\rho_{\mu_{\theta_1}}(x)$, respectively. If $\phi(x)=x$ and $\frac{\sqrt{\rho_{\mu_\theta}(x)}-\sqrt{\rho_\nu(x)}}{\sqrt{\rho_{\mu_\theta}(x)}+\sqrt{\rho_\nu(x)}}=f_w(x)$ then for the Helinger divergence, we have:*

$$\begin{aligned}\mathbb{E}d^2_H(g^{\mu}_{\theta_1},\nu) &\leq O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}}+O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\\ &\quad+\inf_{\theta\in\Theta}\sup_{w\in\mathcal{W}}\left|\mathbb{E}f_w(X))-\mathbb{E}f_w(g_\theta(Z)))\right|. \quad (31)\end{aligned}$$

Remark 3.3 *Our bound in* (30) *represents a better result, considering both the discriminator and generator and the general case of $\phi(x)$. Also, the Corollary 3.3 is an almost sure result and the difference between the expected outputs of the discriminator on real data (X) and generated data ($g_\theta(Z)$) for $\phi(x)=x$. In the current result stated in* (28) *under the condition $\frac{\sqrt{\rho_{\mu_\theta}(x)}-\sqrt{\rho_\nu(x)}}{\sqrt{\rho_{\mu_\theta}(x)}+\sqrt{\rho_\nu(x)}}=f_w(x)$,* (28) *relies on pseudo-dimension, presenting a weaker bound compared to* (31)*.*

3.3. Pearson $\mathcal{X}^2$ divergence bound

In the context of GANs, the Pearson $\mathcal{X}^2$ divergence is not commonly used as an explicit divergence measure in the objective function. However, it is possible to derive

a connection between the Pearson divergence and the GAN objective function under certain assumptions. Directly optimizing the Pearson divergence in the GAN objective function is not common, and the use of other divergence measures is more prevalent in GAN training. The Pearson $\mathcal{X}^2$ divergence is defined as

$$d_{\mathcal{X}^2}(g_{\theta_1}^{\mu}, \nu) := \int \frac{(\rho_{\mu_{\theta_1}(x)} - \rho_{\nu}(x)^2}{\rho_{\nu}(x)} dx. \tag{32}$$

Theorem 3.5 *Let θ_1 be the solution of $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_{\theta}^{\hat{\mu}_m}, \hat{\nu}_n)$ where $d_{\mathcal{F},\phi}(g_{\theta}^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in* (2)*. If $\rho_{\nu}(x)$ and $\rho_{\mu_{\theta_1}}(x)$ are the density function for the distribution of ν and g_{θ}^{μ}. Then*

$$\begin{aligned}\mathbb{E}d_{\mathcal{X}^2}(g_{\theta}^{\mu}, \nu) \leq 2 \sup_{\theta\in\Theta} \left\{ \left\| 1 - \frac{\rho_{\mu_\theta}(x)}{\rho_{\nu}(x)} - \phi(f_w(x)) \right\|_{\infty} \right\} \\ + O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}} + \inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_{\theta}^{\mu}, \nu) + 2\phi(1/2). \end{aligned} \tag{33}$$

.

Proof. We employ the same technique of simplification as seen in the preceding Theorems 3.2 and 3.4 to establish the bounding of the Pearson $\mathcal{X}^2$ divergence. Starting from the definition of Pearson $\mathcal{X}^2$ divergence, we can express it as follows:

$$\begin{aligned} d_{\mathcal{X}^2}(g_{\theta_1}^{\mu}, \nu) &= \int \frac{(\rho_{\mu_{\theta_1}(x)} - \rho_{\nu}(x)^2}{\rho_{\nu}(x)} dx \\ &= \int \left(1 - \frac{\rho_{\mu_{\theta_1}}(x)}{\rho_{\nu}(x)}\right) (\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)) dx \\ &= \int \left(1 - \frac{\rho_{\mu_{\theta_1}}(x)}{\rho_{\nu}(x)} - \phi(f_w(x))\right) \left(\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)\right) \\ &+ \int \phi(f_w(x)) \left(\rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x)\right) \\ &\leq \left\| 1 - \frac{\rho_{\mu_{\theta_1}}(x)}{\rho_{\nu}(x)} - \phi(f_w(x)) \right\|_{\infty} \left|\left| \rho_{\nu}(x) - \rho_{\mu_{\theta_1}}(x) \right|\right|_1 + d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu) \\ &- \mathbb{E}\phi(1 - f_w(g_{\theta_1}(Z))) - \mathbb{E}\phi(f_w(g_{\theta_1}(Z))) \\ &\leq 2 \left\| 1 - \frac{\rho_{\mu_\theta}(x)}{\rho_{\nu}(x)} - \phi(f_w(x)) \right\|_{\infty} + d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu) + 2\phi(1/2). \end{aligned} \tag{34}$$

By applying the inequality in Lemma 2.1 for $d_{\mathcal{F},\phi}(g_{\theta_1}^{\mu}, \nu)$ and following similar steps outlined in Theorem 3.4, we can establish the proof. ■

The expression

$$\sup_{\theta\in\Theta} \left\{ \left\| 1 - \frac{\rho_{\mu_\theta}(x)}{\rho_{\nu}(x)} - \phi(f_w(x)) \right\|_{\infty} \right\}$$

represents the supremum, or the least upper bound, over all possible choices of the parameter θ within the parameter space Θ of the function

$$\left\| 1 - \frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)} - \phi(f_w(x)) \right\|_\infty .$$

The term captures the maximum difference, measured using the sup norm, between 1 minus the ratio of densities of the generated and target distributions, and the discriminator's as the input of ϕ, considering all possible generator parameters.

Corollary 3.4 *Let θ_1 be the solution of $\inf_{\theta\in\Theta} d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ where $d_{\mathcal{F},\phi}(g_\theta^{\hat{\mu}_m}, \hat{\nu}_n)$ is given in* (2)*. Denote the density functions of ν and $g_{\theta_1}^{\mu}(Z)$ by $\rho_\nu(x)$ and $\rho_{\mu_{\theta_1}}(x)$, respectively. If $\phi(x) = x$ and $1 - \frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)} = f_w(x)$ then for the Pearson $\mathcal{X}^2$ divergence bound, we have:*

$$\begin{aligned} \mathbb{E}[d_{\mathcal{X}^2}(g_\theta^{\mu}, \nu)] \leq O_{a.s}\left(\frac{\log m}{m}\right)^{\frac{1}{2}} + O_{a.s}\left(\frac{\log n}{n}\right)^{\frac{1}{2}} \\ + \inf_{\theta\in\Theta} \sup_{w\in\mathcal{W}} |\mathbb{E} f_w(X)) - \mathbb{E} f_w(g_\theta(Z)))| . \end{aligned} \tag{35}$$

Remark 3.4 *The bound on Pearson $\mathcal{X}^2$ divergence in* (33) *provides a.s. convergence rates, considering both discriminator and generator sample sizes, and is applicable to the general case of $\phi(x)$. Assuming that $1 - \frac{\rho_{\mu_\theta}(x)}{\rho_\nu(x)}$ equals $f_w(x)$, the expression* (35) *demonstrates a.s. convergence rates and the difference between the expected results of the discriminator for real data (X) and generated data ($g_\theta(Z)$).*

4. Conclusion

In this study, we have derived various f-divergence bounds that provide a better upper bound for the GAN model. We have derived the upper bounds for total variation, Kullback-Leibler (KL) divergence, Hellinger divergence, and Pearson X^2 divergence within the GAN estimator. Our investigation focuses on understanding the f-divergence analysis of the GAN estimator with parameterized discriminator and generator using a general empirical objective function. Some existing results in the literature correspond to specific cases of the error defined in this paper. We aim to explore various directions in future work, primarily finding the lower bound of the f-divergence bound to establish the optimality. As part of our future research, we aim to explore multiple directions. Additionally, we plan to investigate a general proof framework for the convexity of an arbitrary function f, which underlies many f-divergence measures. These theoretical insights can contribute to a deeper understanding of the convergence behavior and generalization properties of GAN models.

Acknowledgement. The research of Hailin Sang is partially supported by the Simons Foundation Grant No. 586789 and the NSF Grant No. OIA-2428880, USA.

References

[1] Agresti, A. (2002), *Categorical Data Analysis*, Wiley Series in Probability and Statistics, New York.

[2] Arjovsky, M., Chintala, S., Bottou, L. (2017), *Wasserstein Generative Adversarial Networks*, in Proceedings of the 34th International Conference on Machine Learning (ICML), 70, pp. 214-223.

[3] Armanious K., Jiang C., Fischer M., Küstner T., Hepp T., Nikolaou K., Gatidis S., Yang B., (2020), *MedGAN: Medical Image Translation using GANs for MRI Harmonization and Beyond*, Neural Networks.

[4] Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y. (2017), *Generalization and Equilibrium in Generative Adversarial Nets (GANs)*, in Proceedings of the 34th International Conference on Machine Learning (ICML), 70, pp. 224-232.

[5] Beran, R. (1977), *Minimum Hellinger Distance Estimates for Parametric Models*, The Annals of Statistics, pp. 445-463.

[6] Chung, J. K., Kannappan, P., Ng, C. T., Sahoo, P. K. (1989), *Measures of the Distance Between Probability Distributions*, Journal of Mathematical Analysis and Applications, pp. 280-292.

[7] Csiszar, I. (1967), *Information-Type Measures of Difference of Probability Distributions and Indirect Observations*, Studia Scientiarum Mathematicarum Hungarica, 2, pp. 299-318.

[8] Dziugaite, G. K., Roy, D. M., Ghahramani, Z. (2015), *Training Generative Neural Networks via Maximum Mean Discrepancy Optimization*, in Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, pp. 258-267.

[9] Goldberg, P. W., Jerrum, M. R. (1995), *Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers*, in Proceedings of the 6th Annual Conference on Computational Learning Theory, pp. 361-369.

[10] Goodfellow, I., Abadie, J. P., Mirza, M., Xu, B., Farley, D. W., Ozair, S., Courville, A., Bengio, Y. (2014), *Generative Adversarial Nets*, Advances in Neural Information Processing Systems (NIPS), 27, pp. 2672-2680.

[11] Hasan, M., Sang, H. (2023), *Error Analysis of Generative Adversarial Network*, arXiv preprint arXiv:2310.15387.

[12] Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., Yang, Y. (2022), *An Error Analysis of Generative Adversarial Networks for Learning Distributions*, Journal of Machine Learning Research (JMLR), pp. 5047-5089.

[13] Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2017), *Image-to-Image Translation with Conditional Adversarial Networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[14] Ji, K., Zhou, Y., Liang, Y. (2021), *Understanding Estimation and Generalization Error of Generative Adversarial Networks*, IEEE Transactions on Information Theory, 67, pp. 3114-3129.

[15] Karras, T., Aila, T., Laine, S., Lehtinen, J. (2018), *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, Proceedings of the International Conference on Learning Representations (ICLR).

[16] Karras, T., Laine, S., Aila, T. (2019), *A Style-Based Generator Architecture for Generative Adversarial Networks*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[17] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2020), *Analyzing and Improving the Image Quality of StyleGAN*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[18] Karras, T., Aittala, M., Laine, S., Herva, E., Aila, T. (2021), *Alias-Free Generative Adversarial Networks (StyleGAN3)*, Advances in Neural Information Processing Systems (NeurIPS).

[19] Li, Y., Swersky, K., Zemel, R. (2015), *Generative Moment Matching Networks*, in Proceedings of the 32nd International Conference on Machine Learning (ICML), 37, pp. 1718-1727.

[20] Liang, T. (2017), *How Well Can Generative Adversarial Networks (GAN) Learn Densities: A Nonparametric View*, arXiv preprint arXiv:1712.08244.

[21] Liang, T. (2021), *How Well Generative Adversarial Networks Learn Distributions*, Journal of Machine Learning Research (JMLR), 22, 1-41.

[22] MacKay, D. J. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

[23] Mirza, M., Osindero, S. (2014), *Conditional Generative Adversarial Nets*, arXiv preprint arXiv:1411.1784.

[24] Nowozin, S., Cseke, B., Tomioka, R. (2016), *f-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization*, Advances in Neural Information Processing Systems (NIPS), 29, 271-279.

[25] Oberman, A. M., Calder, J. (2018), *Lipschitz Regularized Deep Neural Networks Converge and Generalize*, arXiv preprint arXiv:1808.09540.

[26] Radford, A., Metz, L., Chintala, S. (2016), *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, Proceedings of the International Conference on Learning Representations (ICLR).

[27] Renyi, A. (1961), *On Measures of Entropy and Information*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.

[28] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H. (2016), *Generative Adversarial Text-to-Image Synthesis*, Proceedings of the 33rd International Conference on Machine Learning (ICML), 48, pp. 1060-1069.

[29] Trefethen, L. (2019), *Approximation Theory and Approximation Practice*, SIAM.

[30] Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press.

[31] Vidyasagar, M. (2003), *Vapnik-Chervonenkis Pseudo and Fat-Shattering Dimensions*, Communications and Control Engineering, Springer.

[32] Xue, Y., Xu, T., Zhang, H., Long, L. (2018), *SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation*, Neuroinformatics.

[33] Yi, X., Walia, E., Babyn, P. S. (2019), *Generative Adversarial Network in Medical Imaging: A Review*, Medical Image Analysis, 58, 101552.

[34] Zhang, P., Liu, Q., Zhou, D., Xu, T., He, X. (2018), *On the Discrimination-Generalization Trade-Off in GANs*, Proceedings of the 6th International Conference on Learning Representations (ICLR).

[35] Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017), *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2242-2251.

Received 19.05.2025, Accepted 16.10.2025