

Hellinger loss function for Generative Adversarial Networks

Giovanni Saraceno^{*}, Anand N. Vidyashankar², and Claudio Agostinelli³

¹Department of Statistical Sciences, University of Padova, Italy

²Department of Statistics, George Mason University, VA, USA

³Department of Mathematics, University of Trento, Italy

Abstract

We propose Hellinger-type loss functions for training Generative Adversarial Networks (GANs), motivated by the boundedness, symmetry, and robustness properties of the Hellinger distance. We define an adversarial objective based on this divergence and study its statistical properties within a general parametric framework. We establish the existence, uniqueness, consistency, and joint asymptotic normality of the estimators obtained from the adversarial training procedure. In particular, we analyze the joint estimation of both generator and discriminator parameters, offering a comprehensive asymptotic characterization of the resulting estimators. We introduce two implementations of the Hellinger-type loss and we evaluate their empirical behavior in comparison with the classic (Maximum Likelihood-type) GAN loss. Through a controlled simulation study, we demonstrate that both proposed losses yield improved estimation accuracy and robustness under increasing levels of data contamination.

Keywords: Generative models, Generative Adversarial Networks, Hellinger distance, Outliers, Robustness.

1 Introduction

Deep learning models have been widely used across a variety of machine learning problems achieving great advances and have received increasing attention in data science and statistics. In fact, deep neural networks can be viewed as a non-linear and highly-parametrized generalization of statistical models [Yuan et al., 2020]. One of the tasks solved by deep neural networks is called generative modeling, in which we are interested in learning a model capable of describing the underlying probability distribution given a sample of data. With the learned model, we are able to generate new data. More recently proposed generative models proceed by an adversarial procedure, based on the idea that a data generator is good if generated data, labeled as “fake”, cannot be distinguished from real

*giovanni.saraceno@unipd.it

data. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [2014], are considered state of the art for generative models and have been developed in several fields from both practical applications and theoretical analysis. The purpose of GANs is to generate observations that are similar to samples collected by a target distribution p_* . GANs are conducted by an adversarial procedure that involves a family of generators and a family of discriminators, usually implemented by neural networks. In particular, the two networks are trained together in a minimax game: the generators transform low-dimensional observations drawn from a known density (usually normal or uniform) into fake observations trying to imitate p_* , while the goal of discriminators is to accurately discriminate between the samples from p_* and the generated “fake” data. GANs have been successfully applied in various domains, ranging from computer vision to natural language processing and medical imaging, achieving state-of-the-art performance.

Significant effort has been devoted to studying GANs from a statistical and theoretical perspective. The early theoretical work by Biau et al. [2020] provides an initial analysis of the asymptotic properties of the GAN estimators, showing that the original formulation is linked to the Jensen–Shannon divergence and establishing convergence results under regularity assumptions and smoothness conditions. Despite their success, GANs are known to be challenging to train, suffering issues like unstable convergence, vanishing gradients, and mode collapse. To address these problems, the original formulation has been extended by employing alternative divergence measures. In particular, [Nowozin et al., 2016] introduce the f GAN framework in which any f -divergence can be used as a training objective by way of a variational formulation, and [Arjovsky et al., 2017] consider the Wasserstein distance by introducing the Wasserstein GAN (WGAN), which is shown to improve training stability and mitigate vanishing gradients. These developments underscored that the choice of the divergence considered may be critical to GAN performance. Following this thought, several variants of GAN have been proposed, grounded in different statistical distances, e.g., the Least-Square GANs [Mao et al., 2017] and W_2 -GAN [Korotin et al., 2019]. Recent surveys, such as Chakraborty et al. [2024], further report the large amount of GAN variants, including the Cumulant GANs, introduced by Pantazis et al. [2023] which replace classical divergences with a framework based on cumulant generating functions, offering theoretical connections to Rényi divergences and improved gradient properties during training, and Relativistic GAN, that aim to address training instability and mode collapse by modifying the loss function or discriminator architecture.

At the same time, there is a growing interest in establishing a theoretical understanding of GANs. For example, following the initial asymptotic analysis of GAN estimators by Biau et al. [2020], Chakraborty and Bartlett [2024] have explored the generalization behavior and statistical efficiency of GANs in regimes where data lie on low-dimensional structures embedded in high-dimensional spaces. This is inspired by real-world data, such as images or sensory signals, which often possess a low intrinsic dimensionality despite their high ambient representation. Chakraborty and Bartlett [2024] provide rigorous convergence rate analyses for both GANs and their bidirectional variants (BiGANs). Collectively, these developments reflect an ongoing movement in the GAN literature toward frameworks that are not only efficient in practice but also with a strong theoretical basis.

Recent research has also focused on adapting adversarial training frameworks to enhance robustness against data contamination. Classical robust statistics suggest that using bounded divergence measures can yield estimators resistant to the influence of outliers. For example, Gao et al. [2019] analyze the relationship between GANs and classical

depth-based estimators, showing that the adversarial formulations can achieve optimal rates in robust location and scatter estimation problems when the discriminator is properly constructed. Gao et al. [2020] similarly study robust covariance matrix estimation through proper scoring rules, induced by variational approximation of f -divergences. Zhu et al. [2022] propose a general theoretical framework for GAN-based estimators of unknown parameters of the true distribution that satisfy robustness guarantees under broad distributional conditions, including sub-exponential classes. Zhang et al. [2023] introduce a robust GAN framework that trains the generator and discriminator against worst-case perturbations. Azimi et al. [2024] develop zGAN, an outlier-focused GAN for synthetic data, which explicitly generates realistic outliers and tail events to augment training data. While these studies indicate the potential of GANs in robust inference, they often consider specific aspects (e.g. robust loss design or rate optimality) in isolation. From this point of view, two key gaps can be identified. First, most theoretical analyses of GANs treat the generator and discriminator separately, for instance, assuming an optimal discriminator and focusing on the asymptotic properties of the generator. Second, the robustness of GAN estimators is rarely examined using statistical tools such as the influence function and the resistance to outliers.

In this work, we propose a Hellinger-type loss function for GAN training and investigate the theoretical properties of the corresponding estimators. The Hellinger distance is a symmetric, bounded divergence between two density functions with a long history in statistics and well known connection to robust estimation. Our contributions can be summarized as follows. We define a novel adversarial objective based on a Hellinger-type distance, aiming to reduce the influence of outliers on the estimators. We develop a comprehensive asymptotic theory analyzing the generator and discriminator parameters jointly. In particular, we establish the existence and uniqueness of the estimators, prove consistency, and derive their joint asymptotic normality. We investigate the robustness of the proposed Hellinger GAN estimator. Specifically, we derive the influence function of the joint estimator, which provides insights into its sensitivity to model contamination.

The remainder of the paper is organized as follows. Section 2 formally introduces the Hellinger-type loss function in the context of adversarial training and define the associated optimization objective. Section 3 presents the main theoretical results, including the existence, uniqueness, consistency, and joint asymptotic normality of the estimator under appropriate regularity conditions. Section 4 examines the robustness properties by deriving the influence functions. Section 5 contains the numerical experiments in a Gaussian setting in which we evaluate the performance of the Hellinger-type GAN loss in the presence of contamination. Section 6 presents results on the Fashion-MNIST dataset. Finally, Section 7 concludes the paper.

2 Hellinger-type Loss

From a mathematical point of view, we can represent the process of GANs as follows. Let X_1, \dots, X_n be i.i.d. observations sampled from some unknown density p_* on E , where E is a Borel subset of \mathbb{R}^d . The density p_* is supposed to be dominated by a fixed known measure μ on E and this condition holds for all densities we consider here. Let Z be a d' -random variable with density g , where $d' \ll d$. The generators can be represented by a parametric family of functions from $\mathbb{R}^{d'}$ to E , that is, $\mathcal{G} = \{G_\theta\}_{\theta \in \Theta}$, $\Theta \subset \mathbb{R}^p$. Each function G_θ is applied to the variable z , which is usually called latent variable or noise,

so that we can consider the natural family of density $\mathcal{Q} = \{q_\theta\}_{\theta \in \Theta}$ associated with the generators defined as $G_\theta(Z) \stackrel{\mathcal{L}}{=} q_\theta d\mu$, where these densities are possible candidates to represent p_* . On the other hand, the family of discriminators $\mathcal{D} = \{D_\alpha\}_{\alpha \in \Lambda}$, $\Lambda \subset \mathbb{R}^q$, can be described by a family of functions from E to $[0, 1]$. The value computed by the discriminator can be thought of as the probability that a given observation comes from the true density p_* . Notice that, since generators and discriminators are usually represented by neural networks, the dimensions p and q can be very large.

Let Z_1, \dots, Z_m be an i.i.d. sample distributed as the latent variable $Z \sim g$. According to the standard formulation of GANs, discriminators and generators are fine-tuned by optimizing the objective function

$$L_n(\theta, \alpha) = \frac{1}{n} \sum_{i=1}^n \ln(D_\alpha(X_i)) + \frac{1}{m} \sum_{j=1}^m \ln(1 - D_\alpha(G_\theta(Z_j)))$$

with respect to (θ, α) , where \ln indicates the natural logarithm. The corresponding population version is given by

$$L(\theta, \alpha) = \int \ln(D_\alpha(x)) p_*(x) d\mu(x) + \int \ln(1 - D_\alpha(G_\theta(z))) g(z) d\mu(z). \quad (1)$$

Therefore, this objective function represents the adversarial game between discriminators and generators: for a given θ , the discriminator is determined to be minimal in generated data $G_\theta(Z_j)$, $j = 1, \dots, m$, and maximal on samples X_i , $i = 1, \dots, n$; on the other hand, for a given α , the generator is chosen so that $D_\alpha(G_\theta(Z_j))$ are maximized. Hence, we want to find $(\hat{\theta}, \hat{\alpha})$ such that

$$(\hat{\theta}, \hat{\alpha}) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} L_n(\theta, \alpha). \quad (2)$$

Goodfellow et al. [2014] showed that the objective function given in equation (2), under appropriate conditions, reduces to the Jensen-Shannon divergence between the data generating density p_* and the family of parameterized densities. Following the idea of the connection between the loss function and a divergence, we propose to use a loss function constructed by using the Hellinger distance.

Definition 2.1. Consider the measurable space $(\mathcal{X}, \mathcal{F})$ and a σ -finite measure λ on $(\mathcal{X}, \mathcal{F})$. For every $u, v \in L^2(\mathcal{X})$ such that u and v are dominated by the measure λ , the squared Hellinger distance between u and v is given by

$$d_{HD}(u, v) = \int (u^{\frac{1}{2}} - v^{\frac{1}{2}})^2 d\lambda.$$

Notice that the Hellinger distance can be rewritten as

$$d_{HD}(u, v) = \int u d\lambda + \int v d\lambda - 2 \int |uv|^{\frac{1}{2}} d\lambda.$$

As special case, if u and v are density probability functions it simplifies to

$$d_{HD}(u, v) = 2 - 2 \int |uv|^{\frac{1}{2}} d\lambda.$$

Considering our setting, we want to construct a loss function such that the Hellinger distance between the functions D_α and $(1 - D_\alpha(G_\theta))$, for $\theta \in \Theta$ and $\alpha \in \Lambda$, is optimized. Defining $f(x, z) = p_*(x)g(z)$ and $\mu(x, z) = \mu(x) \times \mu(z)$, we can consider $d\lambda(x, z) =$

$f(x, z)d\mu(x, z)$. Hence, the Hellinger distance between the functions D_α and $(1 - D_\alpha(G_\theta))$ is given by

$$\begin{aligned} HD^2(D_\alpha, 1 - D_\alpha(G_\theta)) &= \int (D_\alpha^{\frac{1}{2}}(x) - (1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}})^2 d\lambda(x, z) \\ &= \int D_\alpha(x)d\lambda(x, z) + \int (1 - D_\alpha(G_\theta(z)))d\lambda(x, z) \\ &\quad - 2 \int D_\alpha^{\frac{1}{2}}(x)(1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}} d\lambda(x, z). \end{aligned}$$

In particular, since f is a product density and μ is a product measure we have $\int D_\alpha(x)d\lambda(x, z) = \int D_\alpha(x)p_*(x)d\mu(x)$ and $\int (1 - D_\alpha(G_\theta(z)))d\lambda(x, z) = \int (1 - D_\alpha(G_\theta(z)))g(z)d\mu(z)$. Hence, the objective function has the form

$$\begin{aligned} HD^2(\theta, \alpha) &= \int D_\alpha(x)p_*(x)d\mu(x) + \int (1 - D_\alpha(G_\theta(z)))g(z)d\mu(z) \\ &\quad - 2\gamma(\theta, \alpha) \\ &= h_1(\alpha) + h_2(\theta, \alpha) - 2\gamma(\theta, \alpha), \end{aligned} \tag{3}$$

where

$$\gamma(\theta, \alpha) = \int D_\alpha^{\frac{1}{2}}(x)p_*(x)d\mu(x) \int (1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}}g(z)d\mu(z). \tag{4}$$

We will denote $HD_n^2(\theta, \alpha)$ and $\gamma_n(\theta, \alpha)$ the corresponding sample versions.

2.1 Approximated objective function

Consider the same setting presented in the previous section. Notice that, the objective function given in (1) can be rewritten as

$$\begin{aligned} L(\theta, \alpha) &= \int \ln(D_\alpha(x))p_*(x)d\mu(x) + \int \ln(1 - D_\alpha(G_\theta(z)))g(z)d\mu(z) \\ &= \int \ln(D_\alpha(x))p_*(x)d\mu(x) \int g(z)d\mu(z) + \\ &\quad \int \ln(1 - D_\alpha \circ G_\theta(z))g(z)d\mu(z) \int p_*(x)d\mu(x) \\ &= 2 \int \ln(D_\alpha(x)(1 - D_\alpha(G_\theta(z))))^{\frac{1}{2}}f(x, z)d\mu(x, z) \\ &= 2 \int \ln(1 + (D_\alpha^{\frac{1}{2}}(x)(1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}} - 1))d\lambda(x, z). \end{aligned}$$

Therefore, by first order Taylor's series approximation, the function $L(\theta, \alpha)$ is approximated by the objective function

$$\begin{aligned} \tilde{L}(\theta, \alpha) &= -2 \left(1 - \int D_\alpha^{\frac{1}{2}}(x)(1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}} d\lambda(x, z) \right) \\ &= 2\gamma(\theta, \alpha) - 2. \end{aligned}$$

and hence $\tilde{L}_n(\theta, \alpha) = 2\gamma_n(\theta, \alpha) - 2$ is an approximation of $L_n(\theta, \alpha)$. Notice that the objective function $\tilde{L}(\theta, \alpha)$ resembles the formulation of Hellinger distance between two density functions of equations (3) and (4). In particular, the first two terms are referred to well-separated parts of the GAN process: while the first term is related to how the

discriminators evaluate the data, the second term does not depend on data but only on generated data from Z .

Finally, we are interested in finding (θ_n, α_n) such that

$$(\theta_n, \alpha_n) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD_n^2(\theta, \alpha), \quad (5)$$

or

$$\begin{aligned} (\theta_n, \alpha_n) &= \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} \tilde{L}_n(\theta, \alpha) \\ &= \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} \gamma_n(\theta, \alpha). \end{aligned} \quad (6)$$

2.2 Divergence-based Losses

The original GAN [Goodfellow et al., 2014], as described in (1), solves

$$\inf_{\theta} \sup_{\alpha} \left\{ \mathbb{E}_{X \sim p_*} [\ln D_{\alpha}(X)] + \mathbb{E}_{Z \sim g} [\ln(1 - D_{\alpha}(G_{\theta}(Z)))] \right\}.$$

Under the optimal discriminator $D_{\theta}^*(x) = p_*(x)/(p_*(x) + q_{\theta}(x))$, this reduces to minimizing the symmetric Jensen–Shannon divergence (JSD) between the data distribution p_* and the family of parametrized densities q_{θ} . The JSD is bounded and symmetric, but it can saturate, leading to vanishing gradients for the generator. The Wasserstein GAN (WGAN; Arjovsky et al., 2017) replaces the JSD objective by minimizing the 1-Wasserstein distance, that is

$$\inf_{\theta} \sup_{D: \|D\|_L \leq 1} \left\{ \mathbb{E}_{X \sim p_*} [D(X)] - \mathbb{E}_{Z \sim g} [D(G_{\theta}(Z))] \right\},$$

where the supremum is over all 1-Lipschitz functions D . By the Kantorovich–Rubinstein duality this maximization computes the Wasserstein metric enforcing a Lipschitz gradient constraint on D_{α} , and importantly WGANs yield non-vanishing gradients. More generally, Nowozin et al. [2016] (f -GAN) show that any f -divergence can be employed by introducing a variational discriminator T_{α} and its Fenchel–Legendre dual f^* . The f -GAN framework solves

$$\inf_{\theta} \sup_{T} \left\{ \mathbb{E}_{X \sim p_*} [T(X)] - \mathbb{E}_{Z \sim g} [f^*(T(G_{\theta}(Z)))] \right\},$$

which under optimal T_{α} is equivalent to minimizing the chosen f -divergence between p_* and q_{θ} . Different choices of f include the Kullback–Leibler, Pearson χ^2 , JSD, and the squared Hellinger distance. The stability and robustness properties depend strongly on the choice of f .

The proposed Hellinger-GAN instead directly targets the squared Hellinger distance between the discriminator on data and the discriminator on generated data, that is

$$\inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD^2(\theta, \alpha) = \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD^2(D_{\alpha}(\cdot), 1 - D_{\alpha}(G_{\theta}(\cdot))).$$

The squared Hellinger distance is symmetric and bounded, and its square root form may help to maintain the gradients globally finite.

In summary, divergence-based GANs vary in whether their losses are bounded, symmetric, and gradient-stable. Additionally, the existing literature develops asymptotic guarantees only for the generator parameters under optimal discriminator parameters. In our framework, we investigate the joint asymptotic behavior of both the generator and the discriminator parameters. The details of this joint analysis are developed in the following sections.

3 Asymptotic Properties

In this section we discuss existence and uniqueness of the proposed estimators together with their statistical asymptotic properties such as consistency and asymptotic normality of the estimators.

3.1 Existence and Uniqueness

We are interested in studying the properties of the objective function $\text{HD}^2(\theta, \alpha)$ solutions with respect to (θ, α) , which is equivalent to find the couple of parameters (θ_*, α_*) such that

$$(\theta_*, \alpha_*) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} \text{HD}^2(\theta, \alpha). \quad (7)$$

Our first result concerns the existence and uniqueness of (θ_*, α_*) . Consider the following assumptions.

- (H_G) G_θ is continuous with respect to θ , i.e. if $\theta_n \rightarrow \theta$ as $n \rightarrow \infty$, then $G_{\theta_n} \rightarrow G_\theta$ as $n \rightarrow \infty$. Θ is a compact subset of \mathbb{R}^p and the model $\{G_\theta\}_{\theta \in \Theta}$ is identifiable with respect to θ .
- (H_D) the function $(x, \alpha) \rightarrow D_\alpha(x)$ is C^1 , i.e. continuous and differentiable, with continuous differential. Λ is a compact subset of \mathbb{R}^q and the model $\{D_\alpha\}_{\alpha \in \Lambda}$ is identifiable with respect to α .

Theorem 3.1. (*Existence and uniqueness*) *If (H_D) and (H_G) hold, then there exists a unique (θ_*, α_*) such that*

$$(\theta_*, \alpha_*) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} \text{HD}^2(\theta, \alpha).$$

Proof. To prove the existence, we need to show that $\text{HD}^2(\theta, \alpha)$ is jointly continuous with respect to θ and α , that is, given $\alpha, \alpha_n \in \Lambda$ and $\theta, \theta_n \in \Theta$ such that $(\theta_n, \alpha_n) \rightarrow (\theta, \alpha)$, then

$$\text{HD}^2(\theta_n, \alpha_n) \rightarrow \text{HD}^2(\theta, \alpha) \text{ as } n \rightarrow \infty.$$

Note that

$$\begin{aligned} \text{HD}^2(\theta_n, \alpha_n) - \text{HD}^2(\theta, \alpha) &= h_1(\alpha_n) - h_1(\alpha) + h_2(\theta_n, \alpha_n) - h_2(\theta, \alpha) - 2\gamma(\theta_n, \alpha_n) + 2\gamma(\theta, \alpha) \\ &= H_1 + H_2 + H_3 \end{aligned}$$

First

$$\begin{aligned} H_1 &= h_1(\alpha_n) - h_1(\alpha) = \int D_{\alpha_n}(x)p_*(x)d\mu(x) - \int D_\alpha(x)p_*(x)d\mu(x) \\ &= \int (D_{\alpha_n}(x) - D_\alpha(x))p_*(x)d\mu(x). \end{aligned}$$

Notice that $(D_{\alpha_n}(x) - D_\alpha(x)) \leq 2$, then by the Dominated Convergence Theorem (DCT)

$$\lim_{n \rightarrow \infty} |H_1| \leq \int \lim_{n \rightarrow \infty} |D_{\alpha_n}(x) - D_\alpha(x)| p_*(x) d\mu(x).$$

This is equal to zero by the continuity of $D_\alpha(x)$ stated in assumption (H_D) . Considering the notation $D_{\alpha,\theta}(z) = D_\alpha(G_\theta(z))$,

$$\begin{aligned} H_2 &= h_2(\theta_n, \alpha_n) - h_2(\theta, \alpha) \\ &= \int (1 - D_{\alpha_n, \theta_n}(z)) g(z) d\mu(z) - \int (1 - D_{\alpha, \theta}(z)) g(z) d\mu(z) \\ &= \int [(1 - D_{\alpha_n, \theta_n}(z)) - (1 - D_{\alpha, \theta}(z))] q_\theta(z) d\mu(z) \\ &\quad + \int [(1 - D_{\alpha_n, \theta}(z)) - (1 - D_{\alpha, \theta}(z))] q_\theta(z) d\mu(z). \end{aligned}$$

Since $D_{\alpha,\theta}(z) \leq 1$, by the DCT we have

$$\begin{aligned} \lim_{n \rightarrow \infty} |H_2| &\leq \int \lim_{n \rightarrow \infty} |(1 - D_{\alpha_n, \theta_n}(z)) - (1 - D_{\alpha, \theta}(z))| g(z) d\mu(z) \\ &\quad + \int \lim_{n \rightarrow \infty} |(1 - D_{\alpha_n, \theta}(z)) - (1 - D_{\alpha, \theta}(z))| g(z) d\mu(z). \end{aligned}$$

The first term is equal to zero for the continuity of $D_\alpha(x)$ with respect to x by (H_D) and the continuity of $G_\theta(z)$ by (H_G) . By the continuity of $D_\alpha(x)$ with respect to α , also the second term is zero.

Finally, we prove that

$$\lim_{n \rightarrow \infty} |H_3| = 2 \lim_{n \rightarrow \infty} |\gamma(\theta_n, \alpha_n) - \gamma(\theta, \alpha)| = 0.$$

Notice that

$$\begin{aligned} \gamma(\theta_n, \alpha_n) - \gamma(\theta, \alpha) &= \int D_{\alpha_n}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int \left[(1 - D_{\alpha_n, \theta_n}(z))^{\frac{1}{2}} - (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} \right] g(z) d\mu(z) \\ &\quad + \int D_{\alpha_n}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} g(z) d\mu(z) \\ &\quad - \int D_{\alpha}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} g(z) d\mu(z) \end{aligned}$$

where we added and subtracted the quantity $(1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} g(z)$ in the second integral. Repeating the same operation with $(1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} g(z)$ we get

$$\begin{aligned} \gamma(\theta_n, \alpha_n) - \gamma(\theta, \alpha) &= \int D_{\alpha_n}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int \left[(1 - D_{\alpha_n, \theta_n}(z))^{\frac{1}{2}} - (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} \right] g(z) d\mu(z) \\ &\quad + \int D_{\alpha_n}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int \left[(1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} - (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} \right] g(z) d\mu(z) \\ &\quad + \int D_{\alpha_n}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} g(z) d\mu(z) \\ &\quad - \int D_{\alpha}^{\frac{1}{2}}(x) p_*(x) d\mu(x) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} g(z) d\mu(z). \end{aligned}$$

Finally, one more iteration with $D_\alpha^{\frac{1}{2}}(x)p_*(x)$ in the first integral leads to

$$\begin{aligned}\gamma(\theta_n, \alpha_n) - \gamma(\theta, \alpha) &= \int D_{\alpha_n}^{\frac{1}{2}}(x)p_*(x)d\mu(x) \int \left[(1 - D_{\alpha_n, \theta_n}(z))^{\frac{1}{2}} - (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} \right] g(z)d\mu(z) \\ &\quad + \int D_{\alpha_n}^{\frac{1}{2}}(x)p_*(x)d\mu(x) \int \left[(1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} - (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} \right] g(z)d\mu(z) \\ &\quad + \int \left(D_{\alpha_n}^{\frac{1}{2}}(x) - D_\alpha^{\frac{1}{2}}(x) \right) p_*(x)d\mu(x) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} g(z)d\mu(z) \\ &= J_{1n} + J_{2n} + J_{3n}\end{aligned}$$

We consider the limit as $n \rightarrow +\infty$ of these integrals separately. Note that $\int D_{\alpha_n}^{\frac{1}{2}}(x)p_*(x)d\mu(x) \leq 1$, $\forall n$, and $\left[(1 - D_{\alpha_n, \theta_n}(z))^{\frac{1}{2}} - (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} \right] \leq 2$. Then, for the DCT we have

$$\lim_{n \rightarrow +\infty} |J_{1n}| \leq \int \lim_{n \rightarrow +\infty} \left| (1 - D_{\alpha_n, \theta_n}(z))^{\frac{1}{2}} - (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} \right| g(z)d\mu(z)$$

and this is equal to zero by the continuity of $G_\theta(z)$ with respect to θ and the continuity of $D_\alpha(x)$ with respect to x . Similarly, by the DCT

$$\lim_{n \rightarrow +\infty} |J_{2n}| \leq \int \lim_{n \rightarrow +\infty} \left| (1 - D_{\alpha_n, \theta}(z))^{\frac{1}{2}} - (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} \right| g(z)d\mu(z) = 0$$

by the continuity of $D_\alpha(x)$ with respect to α . Finally, note that in J_{3n} the second integral does not depend on n , therefore it is not considered, and $\left(D_{\alpha_n}^{\frac{1}{2}}(x) - D_\alpha^{\frac{1}{2}}(x) \right) \leq 2$. Hence, for the DCT we have

$$\lim_{n \rightarrow +\infty} |J_{3n}| \leq \int \lim_{n \rightarrow +\infty} \left| D_{\alpha_n}^{\frac{1}{2}}(x) - D_\alpha^{\frac{1}{2}}(x) \right| p_*(x) = 0.$$

Considering the limit

$$\lim_{n \rightarrow +\infty} |\gamma(\theta_n, \alpha_n) - \gamma(\theta, \alpha)| \leq \lim_{n \rightarrow +\infty} (|J_{1n}| + |J_{2n}| + |J_{3n}|) = 0,$$

then

$$\lim_{n \rightarrow \infty} |HD^2(\theta_n, \alpha_n) - HD^2(\theta, \alpha)| \leq \lim_{n \rightarrow \infty} (|H_1| + |H_2| + |H_3|) = 0.$$

We proved that the set $\{(\theta_*, \alpha_*) | (\theta_*, \alpha_*) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD^2(\theta, \alpha)\}$ is not empty. It remains to prove the uniqueness. Assume that there exists $(\tilde{\theta}, \tilde{\alpha}) \in \Theta \times \Lambda$ such that

$$(\tilde{\theta}, \tilde{\alpha}) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD^2(\theta, \alpha).$$

This means that $HD^2(\tilde{\theta}, \tilde{\alpha}) = HD^2(\theta_*, \alpha_*)$. Hence by the identifiability assumption, $(\tilde{\theta}, \tilde{\alpha}) = (\theta_*, \alpha_*)$. \square

3.2 Consistency

We now want to prove the consistency property. Let $d\mu_*(x) = p_*(x)d\mu(x)$ and $d\mu_g(z) = g(z)d\mu(z)$ be the probability measures induced by the density p_* and g , respectively, and let $\mu_n(x)$ denote a sample-based estimator of $\mu_*(x)$; here, we are going to consider the empirical measure given by

$$\mu_n(x) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(x), \tag{8}$$

where $\delta_{X_i}(\cdot)$ denotes the Dirac measure at X_i . The sample version of HD^2 is given by

$$\begin{aligned} HD_n^2(\theta, \alpha) &= \int D_\alpha(x) d\mu_n(x) + \int (1 - D_{\alpha, \theta}(z)) d\mu_g(z) \\ &\quad - 2 \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ &= h_{1,n}(\alpha) + h_2(\theta, \alpha) - 2\gamma_n(\theta, \alpha). \end{aligned}$$

where

$$\gamma_n(\theta, \alpha) = \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int (1 - D_\alpha(G_\theta(z)))^{\frac{1}{2}} d\mu_g(z). \quad (9)$$

Remark 3.1. In the definition above, we adopt the empirical measure μ_n as a natural plug-in estimator of μ_* . However, this choice is not unique. In particular, one can also approximate μ_* by a smoothed estimator such as the kernel density estimator (KDE)

$$d\mu_n^{\text{KDE}}(x) = h_n(x) d\mu(x), \quad h_n(x) = \frac{1}{nc_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{c_n}\right) \quad (10)$$

with kernel function K and bandwidth parameter c_n . Using KDEs, the empirical measure leads to a direct empirical-process formulation which can be used to establish consistency and asymptotics. We have explored the theoretical properties of this alternative formulation and provide a detailed discussion in Section S1 of the Supplementary Material.

Let (θ_n, α_n) be the solution of the empirical objective function, that is

$$(\theta_n, \alpha_n) = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} HD_n^2(\theta, \alpha). \quad (11)$$

We have the following result.

Theorem 3.2. *If (H_D) and (H_G) hold, then*

$$(\theta_n, \alpha_n) \xrightarrow{a.s.} (\theta_*, \alpha_*) \text{ as } n \rightarrow \infty.$$

Proof. The proof is divided into two parts. We first start showing that $HD_n^2(\theta, \alpha)$ converges uniformly, almost surely, to $HD^2(\theta, \alpha)$, i.e.

$$\lim_{n \rightarrow \infty} \sup_{(\theta, \alpha) \in \Theta \times \Lambda} |HD_n^2(\theta, \alpha) - HD^2(\theta, \alpha)| = 0.$$

Note that, since $D_\alpha(x)$ is bounded, by the Strong Law of Large Numbers

$$\begin{aligned} |h_{1,n}(\alpha) - h_1(\alpha)| &= \left| \int D_\alpha(x) d\mu_n(x) - \int D_\alpha(x) d\mu_*(x) \right| \\ &= \left| \int D_\alpha(x) (d\mu_n(x) - d\mu_*(x)) \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Now consider γ_n . Notice that

$$\begin{aligned} |\gamma_n(\theta, \alpha) - \gamma(\theta, \alpha)| &\leq \left| \int D_\alpha^{\frac{1}{2}}(x) (d\mu_n(x) - d\mu_*(x)) \int (1 - D_{\alpha, \theta}(z))^{\frac{1}{2}} d\mu_g(z) \right| \\ &\leq \left| \int D_\alpha^{\frac{1}{2}}(x) (d\mu_n(x) - d\mu_*(x)) \right| \quad \forall \theta \in \Theta, \quad \forall \alpha \in \Lambda. \end{aligned}$$

Hence, in a similar way as above this term converges almost surely to zero as $n \rightarrow \infty$.

Now, recall that $\Theta \times \Lambda$ is compact, then we can extract a convergent subsequence $(\theta_{n_k}, \alpha_{n_k})$ from any sequence (θ_n, α_n) , i.e.

$$(\theta_{n_k}, \alpha_{n_k}) \longrightarrow (\tilde{\theta}, \tilde{\alpha}).$$

where $(\tilde{\theta}, \tilde{\alpha})$ is the limiting value. By the continuity of HD^2 , we have that

$$\lim_{n \rightarrow \infty} |\text{HD}^2(\theta_{n_k}, \alpha_{n_k}) - \text{HD}^2(\tilde{\theta}, \tilde{\alpha})| = 0.$$

Note that

$$\begin{aligned} |\text{HD}_{n_k}^2(\theta_{n_k}, \alpha_{n_k}) - \text{HD}^2(\tilde{\theta}, \tilde{\alpha})| &\leq |\text{HD}_{n_k}^2(\theta_{n_k}, \alpha_{n_k}) - \text{HD}^2(\theta_{n_k}, \alpha_{n_k})| \\ &\quad + |\text{HD}^2(\theta_{n_k}, \alpha_{n_k}) - \text{HD}^2(\tilde{\theta}, \tilde{\alpha})| \\ &\leq \sup_{(\theta, \alpha) \in \Theta \times \Lambda} |\text{HD}_{n_k}^2(\theta, \alpha) - \text{HD}^2(\theta, \alpha)| \\ &\quad + |\text{HD}^2(\theta_{n_k}, \alpha_{n_k}) - \text{HD}^2(\tilde{\theta}, \tilde{\alpha})|. \end{aligned}$$

and since we proved that both terms in the right hand side go to zero as $n \rightarrow \infty$, then

$$\text{HD}_{n_k}^2(\theta_{n_k}, \alpha_{n_k}) \longrightarrow \text{HD}^2(\tilde{\theta}, \tilde{\alpha}) \quad \mu - a.s..$$

Notice that $(\theta_{n_k}, \alpha_{n_k})$ is the optimizer of $\text{HD}_{n_k}^2(\theta, \alpha)$, then $(\tilde{\theta}, \tilde{\alpha})$ is the optimizer of $\text{HD}^2(\theta, \alpha)$. For the uniqueness, we have $(\tilde{\theta}, \tilde{\alpha}) \equiv (\theta_*, \alpha_*)$. \square

3.3 Joint Asymptotic Normality

We now investigate the joint asymptotic normality of (θ_n, α_n) . The optimizer (θ_n, α_n) is solution of the estimating equations given by

$$\nabla \text{HD}_n^2(\theta, \alpha) = 0$$

where ∇ denotes the gradient operator with respect to θ and α , i.e. $\nabla = \begin{pmatrix} \nabla_\alpha \\ \nabla_\theta \end{pmatrix}$. By Taylor's series expansion of $\text{HD}_n^2(\theta, \alpha)$ around (θ_*, α_*) , we get

$$\nabla \text{HD}_n^2(\theta_n, \alpha_n) = \nabla \text{HD}_n^2(\theta_*, \alpha_*) + \nabla^2 \text{HD}_n^2(\theta_*, \alpha_*)[(\theta_n, \alpha_n) - (\theta_*, \alpha_*)] \quad (12)$$

where ∇^2 denotes the matrix of second derivatives, i.e.

$$\nabla^2 = \begin{pmatrix} \nabla_\alpha^2 & \nabla_\theta \nabla_\alpha \\ \nabla_\alpha \nabla_\theta & \nabla_\theta^2 \end{pmatrix},$$

and $(\theta_n^*, \alpha_n^*) \in \mathcal{U}_n(\theta_*) \times \mathcal{V}_n(\alpha_*)$ with $\mathcal{U}_n(\theta_*) = \{\theta | \theta = t\theta_* + (1-t)\theta_n\}$ and $\mathcal{V}_n(\alpha_*) = \{\alpha | \alpha = t\alpha_* + (1-t)\alpha_n\}$. Note that $\nabla \text{HD}_n^2(\theta_n, \alpha_n) = 0$, hence

$$(\theta_n, \alpha_n) - (\theta_*, \alpha_*) = -[\nabla^2 \text{HD}_n^2(\theta_n^*, \alpha_n^*)]^{-1} \nabla \text{HD}_n^2(\theta_*, \alpha_*) \quad (13)$$

We now introduce some lemmas that will be useful in determining the joint asymptotic distribution of $\sqrt{n}((\theta_n, \alpha_n) - (\theta_*, \alpha_*))$.

Lemma 3.1. *Consider X_1, \dots, X_n i.i.d. observations such that $X_i \sim p_*$ and let $\mu_n(x)$ be the empirical estimator given in equation (8). Let f be a d -dimensional function uniformly bounded. We have that*

(i)

$$\int f(x) \left(d\mu_n(x) - d\mu_*(x) \right) \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty;$$

(ii)

$$\sqrt{n} \int f(x) \left(d\mu_n(x) - d\mu_*(x) \right) \xrightarrow{d} N_d(0, \Sigma_f) \quad \text{as } n \rightarrow \infty,$$

where $\Sigma_f = \text{Var}[f(X)]$.

Consider the following assumptions:

- (H₁) $(\alpha, x) \rightarrow D_\alpha(x)$ is of class \mathcal{C}^2 , uniformly bounded with uniformly bounded differential of first and second order.
- (H₂) $\forall z \in \mathbb{R}^{d'}, \theta \rightarrow G_\theta(z)$ is of class \mathcal{C}^2 , uniformly bounded with uniformly bounded differential.
- (H₃) The Hessian matrix of the objective function, $\nabla^2 \text{HD}_n^2(\theta, \alpha)$, is positive definite at the true parameter values (θ^*, α^*) .

Proposition 3.1. Assume that (H₁) – (H₃) hold, then

$$\sqrt{n} \nabla \text{HD}_n^2(\theta_*, \alpha_*) \longrightarrow N(0, S),$$

where S is a non-singular covariance matrix.

Proof. The derivative with respect to θ given in Appendix A.1 computed at (θ_*, α_*) can be rewritten as

$$\begin{aligned} \sqrt{n} \nabla_\theta \text{HD}_n^2(\theta_*, \alpha_*) &= \\ &- \sqrt{n} \int \nabla_\theta D_{\alpha_*, \theta_*}(z) d\mu_g(z) \\ &+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x) (d\mu_n(x) - d\mu_*(x)) \int \frac{\nabla_\theta D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} d\mu_g(z) \\ &+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x) d\mu_*(x) \int \frac{\nabla_\theta D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} d\mu_g(z) \\ &= T_1 + T_{2n} + T_3. \end{aligned}$$

By Lemma 3.1, we have that

$$T_{2n} \xrightarrow{d} N(0, S_1),$$

with

$$\Delta_1(X) = \left(D_{\alpha_*}^{\frac{1}{2}}(X) \right) \mathbb{E}_{Z \sim g} \left[\frac{\nabla_\theta D_{\alpha_*, \theta_*}(Z)}{(1 - D_{\alpha_*, \theta_*}(Z))^{1/2}} \right].$$

and

$$\begin{aligned} S_1 &= \text{Var}(\Delta_1(X)) \\ &= \mathbb{E}_{Z \sim g} \left[\frac{\nabla_\theta D_{\alpha_*, \theta_*}(Z)}{(1 - D_{\alpha_*, \theta_*}(Z))^{1/2}} \right] \text{Var} \left(D_{\alpha_*}^{\frac{1}{2}}(X) \right) \mathbb{E}_{Z \sim g} \left[\frac{\nabla_\theta D_{\alpha_*, \theta_*}(Z)}{(1 - D_{\alpha_*, \theta_*}(Z))^{1/2}} \right]^\top. \end{aligned}$$

Notice that $T_1 + T_3 = 0$ since it corresponds to $\nabla_\theta \text{HD}_n^2(\theta_*, \alpha_*) = 0$.

Let us now consider the derivative with respect to α given in Appendix A.1 computed at (θ_*, α_*) , that can be rewritten as

$$\begin{aligned}
\sqrt{n} \nabla_\alpha \text{HD}_n^2(\theta_*, \alpha_*) &= \\
&+ \sqrt{n} \int \nabla_\alpha D_{\alpha_*}(x) (d\mu_n(x) - d\mu_*(x)) \\
&+ \sqrt{n} \int \nabla_\alpha D_{\alpha_*}(x) d\mu_*(x) \\
&- \sqrt{n} \int \nabla_\alpha D_{\alpha_*, \theta_*}(z) d\mu_g(z) \\
&- \sqrt{n} \int \frac{\nabla_\alpha D_{\alpha_*}(x)}{D_{\alpha_*}(x)^{1/2}} (d\mu_n(x) - d\mu_*(x)) \int (1 - D_{\alpha_*, \theta_*}(z))^{\frac{1}{2}} d\mu_g(z) \\
&- \sqrt{n} \int \frac{\nabla_\alpha D_{\alpha_*}(x)}{D_{\alpha_*}(x)^{1/2}} d\mu_*(x) \int (1 - D_{\alpha_*, \theta_*}(z))^{\frac{1}{2}} d\mu_g(z) \\
&+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x) (d\mu_n(x) - d\mu_*(x)) \int \frac{\nabla_\alpha D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} d\mu_g(z) \\
&+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x) d\mu_*(x) \int \frac{\nabla_\alpha D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} d\mu_g(z) \\
&= J_{1,n} + J_2 + J_3 + J_{4,n} + J_5 + J_{6,n} + J_7.
\end{aligned}$$

Notice that $J_2 + J_3 + J_5 + J_7 = 0$ since it corresponds to $\nabla_\alpha \text{HD}^2(\theta_*, \alpha_*) = 0$. The remaining terms can be rewritten as

$$J_{1,n} + J_{4,n} + J_{6,n} = \sqrt{n} \int \Delta_2(x) (d\mu_n(x) - d\mu_*(x))$$

where

$$\begin{aligned}
\Delta_2(X) &= \nabla_\alpha D_{\alpha_*}(X) - \frac{\nabla_\alpha D_{\alpha_*}(X)}{D_{\alpha_*}(X)^{1/2}} \mathbb{E}_{Z \sim g} \left[(1 - D_{\alpha_*, \theta_*}(Z))^{\frac{1}{2}} \right] \\
&\quad + D_{\alpha_*}^{\frac{1}{2}}(X) \mathbb{E}_{Z \sim g} \left[\frac{\nabla_\alpha D_{\alpha_*, \theta_*}(Z)}{(1 - D_{\alpha_*, \theta_*}(Z))^{1/2}} \right]
\end{aligned}$$

By Lemma 3.1, we have that

$$J_{1,n} + J_{4,n} + J_{6,n} \xrightarrow{d} N(0, S_2) \quad \text{with} \quad S_2 = \text{Var}(\Delta_2(X)).$$

Then by the central limit theorem and the continuous mapping theorem, we have that

$$(T_{2n}; J_{1n} + J_{4n} + J_{6n}) \longrightarrow N(0, S)$$

where $S = \begin{bmatrix} S_1 & S_{12} \\ S_{12}^\top & S_2 \end{bmatrix}$ where $S_{12} = \text{Cov}(\Delta_1(X), \Delta_2(X))$. \square

Proposition 3.2. *Assume that $(H_1) - (H_3)$ are satisfied. Then*

$$\lim_{n \rightarrow \infty} \nabla^2 \text{HD}_n^2(\theta_n^*, \alpha_n^*) = \nabla^2 \text{HD}^2(\theta_*, \alpha_*)$$

Proof. The idea is to show the convergence element-wise, considering the second derivatives separately, which are reported in Appendix A.2. Consider the second derivative

with respect to θ . Notice that, using simple operations

$$\begin{aligned} A_2 &= \int D_{\alpha}^{\frac{1}{2}}(x)(d\mu_n(x) - d\mu_*(x)) \int \frac{\nabla_{\theta\theta^\top} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ &\quad + \int D_{\alpha}^{\frac{1}{2}}(x)d\mu_*(x) \int \frac{\nabla_{\theta\theta^\top} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z). \end{aligned}$$

Note that, by (H_2) , $D_{\alpha}^{\frac{1}{2}}(x)$ and $\nabla_{\theta\theta^\top} D_{\alpha,\theta}(z)$ are bounded and continuous around θ_* for all $z \in \mathbb{R}^{d'}$. Then, the first term converges to zero as $n \rightarrow \infty$ by the point (i) of the Lemma 3.1, while the second term does not depend on n . Similarly

$$\begin{aligned} A_3 &= \frac{1}{2} \int D_{\alpha}^{\frac{1}{2}}(x)(d\mu_n(x) - d\mu_*(x)) \int \frac{\nabla_\theta D_{\alpha,\theta}(z) \nabla_\theta^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z) \\ &\quad + \frac{1}{2} \int D_{\alpha}^{\frac{1}{2}}(x)d\mu_*(x) \int \frac{\nabla_\theta D_{\alpha,\theta}(z) \nabla_\theta^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z). \end{aligned}$$

By Lemma 3.1 the first term converges to zero. Analogous results can be obtained for the other terms as well as for the other derivatives. See full details in Section S2 of the Supplementary Material. \square

To conclude, we state the asymptotic normality of (θ_n, α_n) .

Theorem 3.3. *Assume that assumptions $(H_D), (H_G), (H_1) - (H_3)$ hold. Let (θ_n, α_n) be the sequence of estimators defined in 11 and let (θ^*, α^*) denote the unique minimax solution of equation 7. Then, as $n \rightarrow \infty$*

$$\sqrt{n}((\theta_n, \alpha_n) - (\theta_*, \alpha_*)) \xrightarrow{d} N(0, \Sigma)$$

where

$$\Sigma = J^{-1}S(J^{-1})^\top, \quad J = \nabla^2 \text{HD}^2(\theta_*, \alpha_*),$$

and S is the covariance matrix in Proposition 3.1.

Proof. Combining the results of Proposition 3.1 and Proposition 3.2 the theorem follows. \square

Remark 3.2. An alternative way to prove these asymptotic properties for the proposed Hellinger-based losses is to consider the profiled version of our estimator that focuses directly on the generator parameter. Specifically, we consider the profiled objectives

$$S_n(\theta) = \sup_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha), \quad S(\theta) = \sup_{\alpha \in \Lambda} \text{HD}^2(\theta, \alpha),$$

and study the profiled estimator $\hat{\theta}_n \in \arg \min_{\theta \in \Theta} S_n(\theta)$. Under the same regularity conditions (H_D) , (H_G) and $(H_1) - (H_3)$, we show that S_n converges to S uniformly on Θ , which yields almost sure consistency of $\hat{\theta}_n$ for the unique minimizer θ^* of S . Moreover, by combining envelope and implicit-function arguments, we derive a central limit theorem for the profiled estimator

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, \Sigma_\theta),$$

where Σ_θ coincides with the θ -marginal of the joint asymptotic covariance matrix in Theorem 3.3. The same conclusions hold for the approximated Hellinger objective. Full statements and proofs are reported in Section S4 of the Supplementary Material.

4 Influence Function

In this section, we compute the influence functions (IFs) associated with the classical GAN loss and the Hellinger-type loss proposed in this work. The IF offers valuable information on the robustness properties of adversarial training procedures, as it characterizes the local sensitivity of the adversarial game to infinitesimal contamination at a given point. Intuitively, it quantifies how a small perturbation in the data affects the resulting parameter estimates.

Let $p_\varepsilon(x) = (1 - \varepsilon)q_{\theta_0}(x) + \varepsilon h(x)$ denote the contaminated density at x , with the contaminating distribution $h(x)$ and let θ_0 denote the parameter values for which the model density q_{θ_0} coincides with the true distribution p_* . Given the underlying random vector $Z \sim g$, we have $X \sim p_\varepsilon(x)$ the contaminated random vector that generates the data. For $\varepsilon = 0$ we have $X = G_{\theta_0}(Z)$. For all $\varepsilon \in [0, 1)$, we define the Hellinger loss function under contamination given as

$$\text{HD}_\varepsilon^2(\theta, \alpha) = \int D_\alpha(x)p_\varepsilon(x)d\mu(x) + \int(1 - D_\alpha(x))q_\theta(x)d\mu(x) - 2\gamma_\varepsilon(\theta, \alpha)$$

where

$$\gamma_\varepsilon(\theta, \alpha) = \left(\int D_\alpha(x)^{1/2}p_\varepsilon(x)d\mu(x) \right) \left(\int(1 - D_\alpha(x))^{1/2}q_\theta(x)d\mu(x) \right) = C_1 C_2 .$$

For each $\varepsilon \in [0, 1)$, we define

$$(\theta_\varepsilon, \alpha_\varepsilon) = \arg \inf_{\theta} \sup_{\alpha} \text{HD}_\varepsilon^2(\theta, \alpha).$$

Notice that for $\varepsilon = 0$, $\text{HD}_0^2(\theta, \alpha)$ denotes the uncontaminated objective function, and (θ_*, α_*) is the corresponding solution. Observe that $p_\varepsilon(x)$ denotes the contaminated distribution, while $q_{\theta_\varepsilon}(x)$ is the model density evaluated at the parameter estimated under contamination.

Then, we define the influence functions for the Hellinger loss as

$$(IF(\theta), IF(\alpha)) = \left. \frac{\partial}{\partial \varepsilon} (\theta_\varepsilon, \alpha_\varepsilon) \right|_{\varepsilon=0} .$$

The optimizer $(\theta_\varepsilon, \alpha_\varepsilon)$ is solution of $\nabla \text{HD}_\varepsilon^2(\theta_\varepsilon, \alpha_\varepsilon) = 0$, for a fixed $\varepsilon \in [0, 1)$. Hence, in order to compute the influence function we can consider, for all $\varepsilon \in [0, 1)$

$$\frac{\partial}{\partial \varepsilon} \nabla \text{HD}_\varepsilon^2(\theta_\varepsilon, \alpha_\varepsilon) = 0 .$$

We have

$$\begin{aligned} \nabla_\alpha \text{HD}_\varepsilon^2(\alpha, \theta) &= \int \nabla_\alpha D_\alpha(x)p_\varepsilon(x)d\mu(x) - \int \nabla_\alpha D_\alpha(x)q_\theta(x)d\mu(x) - \nabla_\alpha \gamma_\varepsilon(\alpha, \theta) \\ &= A_\alpha + B_\alpha - 2(C_{1\alpha}C_2 + C_1C_{2\alpha}) \end{aligned}$$

and

$$\begin{aligned} \nabla_\theta \text{HD}_\varepsilon^2(\alpha, \theta) &= \int(1 - D_\alpha(x))s_\theta(x)q_\theta(x)d\mu(x) - \nabla_\theta \gamma_\varepsilon(\alpha, \theta) \\ &= B_\theta - 2C_1C_{2\theta} , \end{aligned}$$

where $C_{1\alpha} = \nabla_\alpha C_1$, $C_{2\alpha} = \nabla_\alpha C_2$, $C_{2\theta} = \nabla_\theta C_2$, and $s_\theta(x) = \nabla_\theta \log q_\theta(x)$ is the usual score function. We need to compute the derivatives with respect to ε of these terms and evaluate the expressions at $\varepsilon = 0$. For the first term

$$\begin{aligned}\frac{\partial}{\partial \varepsilon} A_\alpha &= \int \frac{\partial}{\partial \varepsilon} \nabla_\alpha D_{\alpha_\varepsilon}(x) p_\varepsilon(x) d\mu(x) + \int \nabla_\alpha D_{\alpha_\varepsilon}(x) \frac{\partial}{\partial \varepsilon} p_\varepsilon(x) d\mu(x) \\ &= \int \nabla_{\alpha\alpha^\top} D_{\alpha_\varepsilon}(x) p_\varepsilon(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon + \int \nabla_\alpha D_{\alpha_\varepsilon}(x) (h(x) - q_{\theta_0}(x)) d\mu(x) \\ \text{with } \varepsilon = 0 \text{ we obtain} \\ &= \int \nabla_{\alpha\alpha^\top} D_{\alpha_0}(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) + \int \nabla_\alpha D_{\alpha_0}(x) (h(x) - q_{\theta_0}(x)) d\mu(x) \\ &= A_{1\alpha} \text{IF}(\alpha) + A_{2\alpha}\end{aligned}$$

Following similar steps, the derivatives computed at $\varepsilon = 0$ are given as

$$\begin{aligned}\frac{\partial}{\partial \varepsilon} B_\alpha &= - \int \nabla_{\alpha\alpha^\top} D_{\alpha_0}(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) - \int \nabla_\alpha D_{\alpha_0}(x) s_{\theta_0}^\top(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\theta) \\ &= B_{1\alpha} \text{IF}(\alpha) + B_{2\alpha} \text{IF}(\theta); \\ \frac{\partial}{\partial \varepsilon} B_\theta &= - \int \nabla_\alpha D_{\alpha_0}(x) s_{\theta_0}(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) + \int (1 - D_{\alpha_0}(x)) \frac{\nabla_{\theta\theta^\top} q_{\theta_0}(x)}{q_{\theta_0}(x)} q_{\theta_0}(x) d\mu(x) \text{IF}(\theta) \\ &= B_{1\theta} \text{IF}(\alpha) + B_{2\theta} \text{IF}(\theta); \\ \frac{\partial}{\partial \varepsilon} C_{1\alpha} &= \int D_{\alpha_0}(x)^{-1/2} \left(\nabla_{\alpha,\alpha^\top} D_{\alpha_0}(x) - \frac{1}{2} D_{\alpha_0}(x)^{-1} \nabla_\alpha D_{\alpha_0}(x) \nabla_\alpha D_{\alpha_0}(x)^\top \right) q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) \\ &\quad + \int D_{\alpha_0}(x)^{-1/2} \nabla_\alpha D_{\alpha_0}(x) (h(x) - q_{\theta_0}(x)) d\mu(x) \\ &= C_{1\alpha a} \text{IF}(\alpha) + C_{1\alpha b}; \\ \frac{\partial}{\partial \varepsilon} C_2 &= - \frac{1}{2} \int (1 - D_{\alpha_0}(x))^{-1/2} \nabla_\alpha D_{\alpha_0}(x)^\top q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) \\ &\quad + \int (1 - D_{\alpha_0}(x))^{1/2} s_{\theta_0}(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\theta) \\ &= C_{2a} \text{IF}(\alpha) + C_{2b} \text{IF}(\theta); \\ \frac{\partial}{\partial \varepsilon} C_1 &= \frac{1}{2} \int D_{\alpha_0}^{-1/2}(x) \nabla_\alpha D_{\alpha_0}(x)^\top q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) + \int D_{\alpha_0}(x)^{1/2} (h(x) - q_{\theta_0}(x)) d\mu(x) \\ &= C_{1a} \text{IF}(\alpha) + C_{1b}; \\ \frac{\partial}{\partial \varepsilon} C_{2\alpha} &= - \frac{1}{2} \int (1 - D_{\alpha_0}(x))^{-3/2} \nabla_\alpha D_{\alpha_0}(x) \nabla_\alpha D_{\alpha_0}(x)^\top q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) \\ &\quad + \int (1 - D_{\alpha_0}(x))^{-1/2} \nabla_{\alpha,\alpha^\top} D_{\alpha_0}(x) q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) \\ &\quad + \int (1 - D_{\alpha_0}(x))^{-1/2} \nabla_\alpha D_{\alpha_0}(x) s_{\theta_0}(x)^\top q_{\theta_0}(x) d\mu(x) \text{IF}(\theta) \\ &= C_{2\alpha a} \text{IF}(\alpha) + C_{2\alpha b} \text{IF}(\alpha) + C_{2\alpha c} \text{IF}(\theta); \\ \frac{\partial}{\partial \varepsilon} C_{2\theta} &= \int (1 - D_{\alpha_0}(x))^{1/2} \frac{\nabla_{\theta\theta^\top} q_{\theta_0}(x)}{q_{\theta_0}(x)} q_{\theta_0}(x) d\mu(x) \text{IF}(\theta) \\ &\quad - \frac{1}{2} \int \frac{\nabla_\alpha D_{\alpha_0}(x)}{(1 - D_{\alpha_0}(x))^{\frac{1}{2}}} q_{\theta_0}(x) d\mu(x) \text{IF}(\alpha) \\ &= C_{2\theta a} \text{IF}(\theta) + C_{2\theta b} \text{IF}(\alpha).\end{aligned}$$

Combining all the terms together, we have the following two equations

$$\begin{aligned} 0 &= [A_{2\alpha} - 2C_{1ab}C_2 - 2C_{2\alpha}C_{1b}] \\ &\quad + [A_{1\alpha} + B_{1\alpha} - 2C_{1\alpha a}C_2 - 2C_{1\alpha}C_{2a} - 2C_{2\alpha}C_{1a} - 2C_1C_{2\alpha a} - 2C_1C_{2\alpha b}] \text{IF}(\alpha) \\ &\quad + [B_{2\alpha} - 2C_{1\alpha}C_{2b} - 2C_1C_{2\alpha c}] \text{IF}(\theta) \\ &= I_0 + I_\alpha \text{IF}(\alpha) + I_\theta \text{IF}(\theta), \end{aligned}$$

and

$$\begin{aligned} 0 &= -2C_{2\theta}C_{1b} + [B_{1\theta} - 2C_{2\theta}C_{1a} - 2C_1C_{2\theta b}] \text{IF}(\alpha) + [B_{2\theta} - 2C_1C_{2\theta a}] \text{IF}(\theta) \\ &= K_0 + K_\alpha \text{IF}(\alpha) + K_\theta \text{IF}(\theta). \end{aligned}$$

Solving the system we have

$$\begin{aligned} \text{IF}(\alpha) &= (I_\alpha - I_\theta K_\theta^{-1} K_\alpha)^{-1} (I_\theta K_\theta^{-1} K_0 - I_0) \\ \text{IF}(\theta) &= -K_\theta^{-1} (K_0 + K_\alpha \text{IF}(\alpha)) = -K_\theta^{-1} (K_0 + K_\alpha (I_\alpha - I_\theta K_\theta^{-1} K_\alpha)^{-1} (I_\theta K_\theta^{-1} K_0 - I_0)). \end{aligned}$$

The computation of the influence function for the standard loss function is reported in Section S3 of the Supplementary Material. Notice that the influence functions are calculated at the true parameter values (θ_*, α_*) . In the numerical experiments, we can set the true generator parameters θ_* , however, it is not straightforward for the discriminator parameters. In fact, we do not know the true values α_* and we can only consider the estimated parameter $\hat{\alpha}$ instead.

5 Numerical Experiments

We conducted numerical experiments to evaluate the empirical performance of the proposed GAN framework with Hellinger-type loss functions. We considered data generated from the normal distribution $\mathcal{N}(\mu_0, \sigma_0)$ with $\mu_0 = 10$ and $\sigma_0 = 1.5$. The generator is a parametric normal model with unknown mean and variance, that is $G_\theta(z) = \sigma z + \mu$ with $\theta = (\mu, \sigma)$ and $Z \sim N(0, 1)$. The discriminator is implemented as a feedforward neural network with one hidden layer of five nodes (16 parameters in total). In the data generating setting, we considered the sample size $n = 100000$, divided into batches of $n_B = 1000$ observations. This simplified setting is chosen since it allows for an explicit comparison of the parameter estimation quality and the influence of the loss function on the learning process. We also investigate the performance of the proposed loss functions in the case of contamination. Specifically, a proportion ε of data points is sampled from $\mathcal{N}(0, 1)$, at the percentage of contamination of $\varepsilon = 0, 1, 5, 10, 20$. For each setting and contamination level, we perform 100 independent replications, and each GAN is trained for 400 epochs using the Adam optimizer with learning rates 0.001, for both generator and discriminator.

We track the following evaluation metrics:

- Mean Squared Error (MSE) between the estimated generator parameters and the true values;
- The Root Mean Square Error (RMSE) for the generator parameters $\theta_* = (\mu_*, \sigma_*)$

$$\text{RMSEC}(\hat{\theta}) = \sqrt{\frac{1}{2} ((\hat{\mu} - \mu_*)^2 + (\hat{\sigma} - \sigma_*)^2)};$$

We compare the proposed approximated Hellinger loss in equation (6) and the complete Hellinger-type loss in equation (3), considering the KDE as sample-based estimator with bandwidth parameter $h = 0.001, 0.01, 0.5$, with the standard GAN loss given in equation (1) and the WGAN. Our goal is to assess the convergence behavior and fidelity of the generated samples.

5.1 Results

First, we examine the accuracy of parameter estimation for each GAN variant under varying contamination levels. Tables 1 and 2 report the best median (standard deviation) MSE achieved for the generator's parameters μ and σ , respectively, for each method. Similarly, Table 3 shows the best combined RMSE for (μ, σ) at the epoch with the lowest error. Here, the best epoch refers to the training epoch (out of 400) where the RMSE of the generator parameters was minimal; the corresponding MSE/RMSE values from that epoch are then averaged over replications. With this selection, we compare methods based on their best observed performance during training.

Table 1: Median (standard deviation) of best $MSE(\hat{\mu})$ ($\times 100$) across replications for the considered methods and percentage of contamination $\varepsilon = 0, 1, 5, 10, 20$.

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	0.001 (2.97)	0.001 (1.44)	0.035 (24.11)	0.019 (82.30)	141.103 (101.29)
WGAN	18.567 (13.24)	20.613 (17.38)	3.359 (21.12)	12.819 (8.57)	26.588 (29.01)
Approx. HD	0.002 (0.43)	0.001 (0.30)	0.002 (2.79)	0.006 (15.53)	93.003 (133.67)
HD ($c_n=0.0001$)	0.042 (1.59)	0.063 (689.19)	0.052 (18.19)	0.064 (86.67)	0.203 (117.61)
HD ($c_n=0.01$)	0.041 (1076.94)	0.097 (130.29)	0.223 (28.29)	1.697 (47.63)	70.500 (80.88)
HD ($c_n=0.5$)	0.211 (1399.96)	0.299 (0.53)	0.221 (272.86)	0.440 (0.73)	29.548 (129.62)

Table 2: Median (standard deviation) of best $MSE(\hat{\sigma})$ ($\times 100$) across replications for the considered methods and percentage of contamination $\varepsilon = 0, 1, 5, 10, 20$.

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	0.001 (74.14)	0.001 (31.64)	0.007 (340.75)	0.010 (486.29)	417.730 (706.51)
WGAN	2.945 (2.19)	5.052 (3.81)	6.567 (8.78)	15.325 (11.79)	56.548 (64.32)
Approx. HD	0.002 (19.06)	0.002 (0.05)	0.003 (58.65)	0.007 (189.64)	185.900 (612.16)
HD ($c_n=0.0001$)	0.026 (0.93)	0.058 (22.28)	0.057 (105.66)	0.057 (214.67)	0.226 (473.39)
HD ($c_n=0.01$)	0.101 (28.65)	0.094 (16.21)	0.081 (39.86)	0.362 (97.98)	42.283 (384.50)
HD ($c_n=0.5$)	0.342 (30.38)	0.305 (0.96)	0.166 (233.43)	0.085 (77.06)	11.991 (319.05)

In the clean data setting, most of the methods perform well in terms of median error. The standard GAN and the proposed approximate Hellinger loss show extremely low MSE for μ and σ . The Hellinger GAN using a KDE-based loss with a very small bandwidth ($h = 0.0001$) similarly attains a median error close to zero for both parameters. In contrast, the Wasserstein GAN shows a higher median error in the uncontaminated case, as well as the KDE-based Hellinger with a larger bandwidth. As the contamination level increases, the classical GAN loss becomes sensitive to even small fractions of outliers. By $\varepsilon = 5\%$ and 10% , the standard GAN's error even if shows low median MSE, begins

Table 3: Median (standard deviation) of best RMSE($\hat{\mu}, \hat{\sigma}$) ($\times 100$) across replications for the considered methods and percentage of contamination $\varepsilon = 0, 1, 5, 10, 20$.

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	0.328 (36.68)	0.300 (15.44)	1.625 (85.08)	1.232 (106.08)	167.839 (135.95)
WGAN	32.896 (10.13)	35.885 (11.80)	25.682 (18.53)	39.399 (13.71)	68.365 (18.82)
Approx. HD	0.514 (9.94)	0.410 (1.34)	0.595 (17.65)	0.926 (45.16)	118.810 (118.35)
HD ($c_n=0.0001$)	2.837 (4.66)	2.924 (59.71)	2.669 (37.20)	2.620 (68.38)	5.381 (102.47)
HD ($c_n=0.01$)	3.026 (87.58)	3.688 (27.93)	4.084 (20.53)	9.874 (32.16)	103.390 (76.43)
HD ($c_n=0.5$)	6.232 (109.08)	6.444 (3.71)	5.090 (52.67)	5.680 (22.73)	59.725 (72.82)

to fluctuate substantially across runs, considering the large standard deviations, and at $\varepsilon = 20\%$ its performance deteriorates drastically. The approximate Hellinger loss remains quite accurate up to moderate contamination but then degrades under higher contamination, showing a performance slightly better than the classical GAN. The WGAN shows a much more gradual increase in error as the contamination grows, with modest variability across runs suggesting consistent behavior even when outliers are present. Overall, while it sacrifices some absolute accuracy, WGAN has lower variability between replications. The Hellinger GAN with KDE loss shows robustness that depends on the choice of bandwidth c_n . A very small bandwidth, $c_n = 0.0001$, shows the lowest errors, even with 20% contamination. Additionally, in terms of the combined RMSE, the approximate Hellinger model achieves a median RMSE which is slightly better than standard GAN, while the Hellinger GAN with KDE and $c_n = 0.0001$ outperforms all the methods for high percentage of contamination. Tables S1–S3 in Section S5 of the Supplementary Material report the median MSE and RMSE of the generator parameters at the final epoch.

Figure 1 displays the evolution of the mean squared error (MSE) for the generator parameters μ (left column) and σ (right column) over the training epochs, for the percentage of contamination $\varepsilon = 0, 5, 10, 20$. This visualization provides insight into the training dynamics of the GAN losses considered. Hellinger-type losses show robustness benefits over standard GAN and WGAN losses, both in terms of final accuracy and training stability. In the uncontaminated case, most methods converge rapidly. For a higher level of contamination, the standard GAN shows larger errors for an increasing epoch, as well as WGAN, especially for σ . In contrast, the Approximate HD loss remains stable, with consistently low MSE throughout training even under 10% contamination. The performance of KDE-based HD loss remains solid when the bandwidth is very small ($c_n = 0.0001$).

6 Fashion MNIST dataset

Finally, we illustrate the proposed Hellinger losses on the higher dimensional Fashion MNIST image dataset. The training dataset is composed by 60,000 samples of 28×28 gray-scale images of clothing items from ten classes (e.g. T-shirt, trouser, coat, bag, boot). We keep the original resolution and rescale pixel intensities to lie in $[0, 1]$. The generator G_θ consists of three transpose-convolutional layers with batch normalization and ReLU activations that maps a latent vector $Z \sim \mathcal{N}(0, I_{100})$ to a 28×28 image, while the discriminator D_α is a two-layer convolutional network with spectral normalization that maps an image to a scalar in $[0, 1]$, using a sigmoid activation function. We use

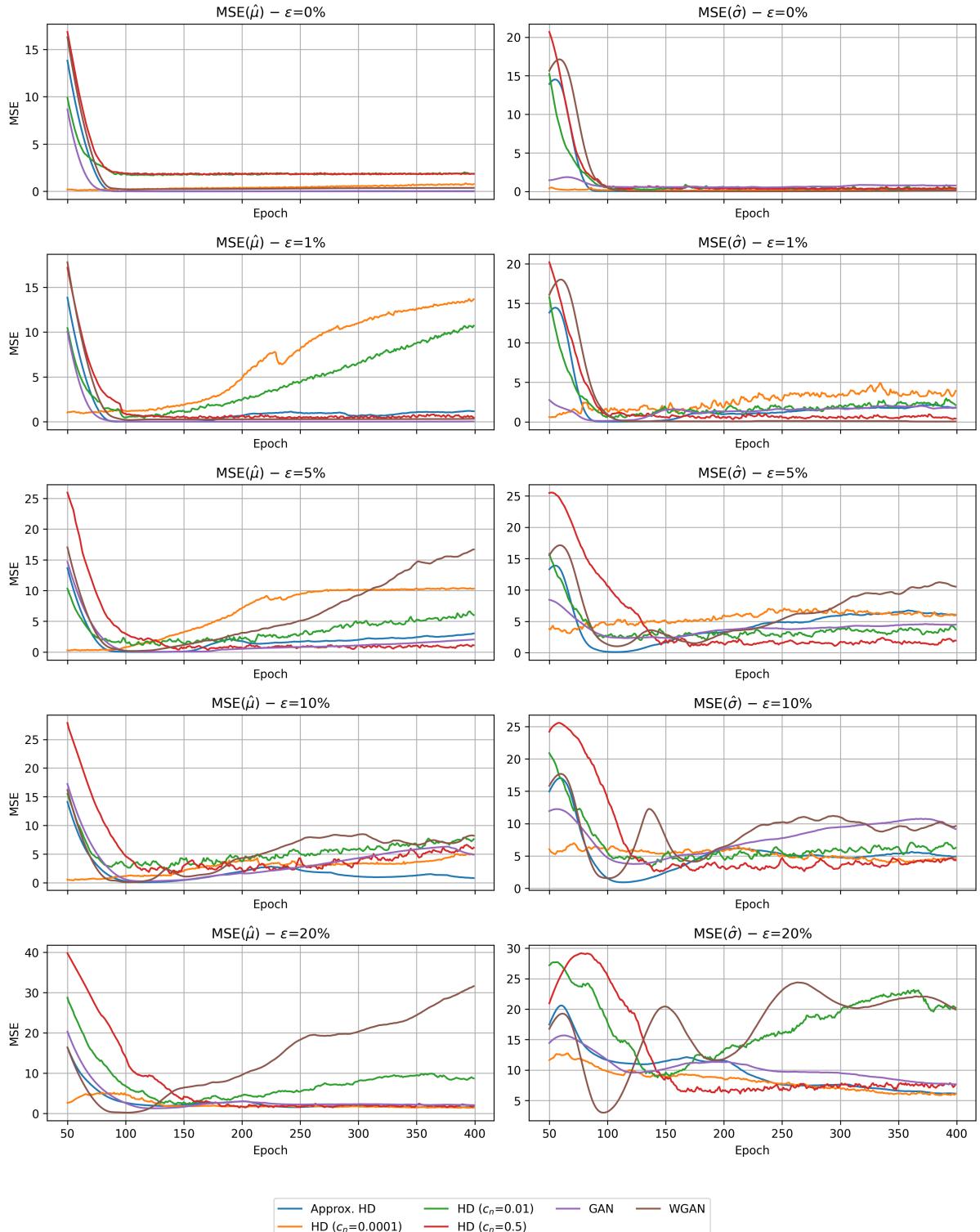


Figure 1: Mean Squared Error (MSE) per epoch for the different contamination percentage $\varepsilon = 0, 1, 5, 10, 20$ comparing the different methods.

Adam optimizer with learning rates of 0.0001 and update weight $\beta = 0.5$. We use a batch size of 128, sampled from the training dataset without replacement, and train the standard GAN, the approximated Hellinger GAN and the WGAN for 1000 epochs. All models share the same network architectures, optimizer and learning-rates, while only the loss function is changed.

Figure 2 reports 8×8 grids of samples generated after 1000 epochs by the considered losses. The standard GAN produces visually plausible items with clear silhouettes. Similarly, the Hellinger-based losses yield images of comparable visual quality while preserving a high degree of diversity across different classes and styles. By contrast, the WGAN configuration produces noticeably noisier samples. Many images show strong grid-like artifacts and saturated regions, and only a subset of silhouettes are clearly recognizable. This suggests that, with the present architecture and training setup, the Wasserstein objective is harder to optimize and yields a lower visual quality.



Figure 2: Fashion MNIST samples generated by the standard GAN (a), the approximated Hellinger GAN (b), the Hellinger GAN with KDE ($c_n = 0.0001$) (c) and WGAN (d) after 1000 epochs.

7 Conclusion

This work contributes to the statistical analysis of generative adversarial networks (GANs) by proposing and rigorously investigating a class of Hellinger-type loss functions within the GAN framework, motivated by the properties of symmetry, boundedness, and a natural connection to classical robust statistics. We define an adversarial objectives that operates on the discriminator output and we study the resulting estimator within a parametric M-estimation framework. Under mild regularity assumptions, we develop a comprehensive asymptotic theory for the joint estimation of generator and discriminator parameters $(\hat{\theta}_n, \hat{\alpha}_n)$ under the proposed Hellinger-type loss. Our results establish the existence, uniqueness, consistency, and asymptotic normality of the estimators under mild regularity assumptions.

We present controlled simulation experiments that highlight the advantages of the Hellinger-type GAN loss. In general, our simulation results demonstrate that in Gaussian settings the choice of loss function has an effect on training and robustness. The standard GAN is non-robust to even modest contamination, while the Wasserstein GAN is more resilient, maintaining bounded errors even as ε increases. The proposed Hellinger-type losses achieve competitive accuracy in the uncontaminated case and maintain substantially lower and more stable mean squared errors under contamination, especially for the complete Hellinger loss with a properly calibrated bandwidth. However, this depends on the choice of the bandwidth parameter c_n . The Fashion MNIST experiment, although not aimed at large-scale image generation, shows that the Hellinger-based losses can produce samples of comparable visual quality to the standard GANs, without sacrificing the robustness properties highlighted in the low-dimensional study.

This study highlights, among other aspects, that the choice of divergence in adversarial training affects the statistical properties of the resulting estimators. In general, distance-type losses can offer both practical and theoretical benefits in adversarial learning. From a broader perspective, this work wants to highlight the importance of integrating statistical principles into the design of generative models. While recent advances have focused heavily on architectural innovation and empirical benchmarks, our findings suggest that studying the asymptotic properties of GAN estimators under various loss functions may lead to more reliable, interpretable, and robust data generation techniques.

A

A.1 Derivatives of First order

Here, we report the first derivatives of the Hellinger loss, which are used in the proof of Proposition 3.1. Considering the following calculations

$$\begin{aligned} \nabla(1 - D_{\alpha,\theta}(z)) &= -\nabla D_{\alpha,\theta}(z), \\ \nabla_\alpha(D_{\alpha,\theta}^{1/2}(x)) &= \frac{1}{2} \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha^{1/2}(x)}, \\ \nabla_\theta((1 - D_{\alpha,\theta}(z))^{1/2}) &= -\frac{1}{2} \frac{\nabla_\theta D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{1/2}}, \\ \nabla_\alpha((1 - D_{\alpha,\theta}(z))^{1/2}) &= -\frac{1}{2} \frac{\nabla_\alpha D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{1/2}}, \end{aligned}$$

the derivatives of first order of the loss function $HD_n^2(\theta, \alpha)$ are given by

$$\begin{aligned}\nabla_\alpha HD_n^2(\theta, \alpha) &= \int \nabla_\alpha D_\alpha(x) d\mu_n(x) - \int \nabla_\alpha D_{\alpha,\theta}(z) d\mu_g(z) \\ &\quad - \int \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha(x)^{1/2}} d\mu_n(x) \int (1 - D_{\alpha,\theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ &\quad + \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{1/2}} d\mu_g(z)\end{aligned}$$

and

$$\begin{aligned}\nabla_\theta HD_n^2(\theta, \alpha) &= - \int \nabla_\theta D_{\alpha,\theta}(z) d\mu_g(z) \\ &\quad + \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_\theta D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{1/2}} d\mu_g(z).\end{aligned}$$

A.2 Derivatives of Second order

Here, we report the second derivatives of the Hellinger loss, which are used in the proof of Proposition 3.2. The second derivatives with respect to θ are given as

$$\begin{aligned}\nabla_{\theta\theta^\top} HD_n^2(\theta, \alpha) &= - \int \nabla_{\theta\theta^\top} D_{\alpha,\theta}(z) d\mu_g(z) \\ &\quad + \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_{\theta\theta^\top} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ &\quad + \frac{1}{2} \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_\theta D_{\alpha,\theta}(z) \nabla_\theta^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z) \\ &= A_1 + A_2 + A_3,\end{aligned}$$

and the second derivatives with respect to α are computed as

$$\begin{aligned}\nabla_{\alpha\alpha^\top} HD_n^2(\theta, \alpha) &= \int \nabla_{\alpha\alpha^\top} D_\alpha(x) d\mu_n(x) \\ &\quad - \int \nabla_{\alpha\alpha^\top} D_{\alpha,\theta}(z) d\mu_g(z) \\ &\quad - \int \frac{\nabla_{\alpha\alpha^\top} D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} d\mu_n(x) \int (1 - D_{\alpha,\theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ &\quad + \frac{1}{2} \int \frac{\nabla_\alpha D_\alpha(x) \nabla_\alpha^\top D_\alpha(x)}{D_\alpha^{\frac{3}{2}}(x)} d\mu_n(x) \int (1 - D_{\alpha,\theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ &\quad + \int \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} d\mu_n(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ &\quad + \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_{\alpha\alpha^\top} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ &\quad + \frac{1}{2} \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z) \nabla_\alpha^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z) \\ &= B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7,\end{aligned}$$

while the mixed derivatives are given by

$$\begin{aligned}
\nabla_{\theta\alpha} \text{HD}_n^2(\theta, \alpha) = & - \int \nabla_{\theta\alpha} D_{\alpha,\theta}(z) d\mu_g(z) \\
& + \frac{1}{2} \int \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} d\mu_n(x) \int \frac{\nabla_\theta D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\
& + \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_{\theta\alpha} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\
& + \frac{1}{2} \int D_\alpha^{\frac{1}{2}}(x) d\mu_n(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z) \nabla_\theta^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z) \\
= & C_1 + C_2 + C_3 + C_4.
\end{aligned}$$

Notice that, the term $\nabla_{\alpha\theta} \text{HD}_n^2$ differs from $\nabla_{\theta\alpha} \text{HD}_n^2$ computed above, only for the order of derivation in C_1 and C_3 .

References

- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.
- Azizjon Azimi, Bonu Boboeva, Ilyas Varshavskiy, Shuhrat Khalilbekov, Akhlitdin Nizamtdinov, Najima Noyoftova, and Sergey Shulgin. zgan: An outlier-focused generative adversarial network for realistic synthetic data generation, 2024. URL <https://arxiv.org/abs/2410.20808>.
- G. Biau, B. Cadre, M. Sangnier, and U. Tanielian. Some theoretical properties of gans. *The Annals of Statistics*, 48(3):1539 – 1566, 2020. doi: 10.1214/19-AOS1858.
- Saptarshi Chakraborty and Peter L. Bartlett. On the statistical properties of generative adversarial models for low intrinsic data dimension, 2024. URL <https://arxiv.org/abs/2401.15801>.
- Tanujit Chakraborty, Ujjwal Reddy K S, Shraddha M Naik, Madhurima Panja, and Bayapureddy Manvitha. Ten years of generative adversarial nets (gans): a survey of the state-of-the-art. *Machine Learning: Science and Technology*, 5(1):011001, jan 2024. doi: 10.1088/2632-2153/ad1f77. URL <https://dx.doi.org/10.1088/2632-2153/ad1f77>.
- Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial networks. In *International Conference on Learning Representations*, 2019.
- Chao Gao, Yuan Yao, and Weizhi Zhu. Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective. *Journal of Machine Learning Research*, 21(229):1–48, 2020. URL <http://jmlr.org/papers/v21/19-462.html>.
- 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*, NIPS’14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Bur-naev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, Yannis Stylianou, and Markos A. Katsoulakis. Cumulant gan. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9439–9450, 2023. doi: 10.1109/TNNLS.2022.3161127.

Y. Yuan, Y. Deng, Y. Zhang, and A. Qu. Deep learning from a statistical perspective. *Stat*, 9(1), jan 2020. ISSN 2049-1573, 2049-1573. doi: 10.1002/sta4.294.

S. Zhang, Z. Qian, and K. Huang. Robust generative adversarial network. *Machine Learning*, 112:5135—5161, 2023. doi: <https://doi.org/10.1007/s10994-023-06367-0>.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Robust estimation for non-parametric families via generative adversarial networks. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1100–1104. IEEE, 2022. doi: 10.1109/ISIT50566.2022.9834844.

Supplementary Material: “Hellinger loss function for Generative Adversarial Networks”

Giovanni Saraceno^{*}, Anand N. Vidyashankar², and Claudio Agostinelli³

¹Department of Statistical Sciences, University of Padova, Italy

²Department of Statistics, George Mason University, VA, USA

³Department of Mathematics, University of Trento, Italy

S1 Kernel density estimator μ_n^{KDE}

As pointed out in Remark 3.1 of the manuscript, the empirical measure μ_n is not the unique choice of sample-based estimator of μ_* . One possible alternative is to approximate μ_* by a smoothed estimator such as the kernel density estimator (KDE)

$$d\mu_n^{\text{KDE}}(x) = h_n(x)d\mu(x), \quad h_n(x) = \frac{1}{nc_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{c_n}\right)$$

with kernel function K and bandwidth parameter c_n . Considering the following additional assumptions, we can state analogous theorems for the consistency and asymptotic normality of (θ_n, α_n) .

(H_K) The kernel function K is such that $K(x) \rightarrow 0$ as $|x| \rightarrow \infty$, $\int |K(x)|dx < \infty$.

$(H_{n,1})$ $c_n \rightarrow 0$ and $nc_n^d \rightarrow \infty$ as $n \rightarrow \infty$.

$(H_{n,2})$ The parameter c_n satisfies $\sqrt{nc_n^2} \rightarrow 0$ as $n \rightarrow \infty$.

First, we introduce the following Lemma which is a modified version of Lemma 3.1 of the manuscript when the KDE μ_n^{KDE} is considered.

Lemma S1.1. *Consider X_1, \dots, X_n i.i.d. observations such that $X_i \sim p_*$ and let $h_n(x)$ be the kernel density estimator of p_* . Let f be a d -dimensional function uniformly bounded. Suppose that assumptions (H_K) and $(H_{n,1})$ hold, then, as $n \rightarrow \infty$,*

(i)

$$\sup_x |h_n(x) - p_*(x)| \rightarrow 0$$

in probability and a.s.

*giovanni.saraceno@unipd.it

(ii)

$$\int f(x)(h_n(x) - p_*(x))d\mu(x) \longrightarrow 0$$

in probability;

(iii)

$$\sqrt{n} \int f(x)((h_n(x) - p_*(x)))d\mu(x) \longrightarrow N_d(0, \Sigma_f)$$

in distribution, where $\Sigma_f = \text{Var}[f(X)] \int [K(x)]^2 dx$.

Proof. **Part (i)** The result follows from Glick's theorem [Glick, 1974].

Part (ii) Given the assumptions $(H_{n,1})$, $h_n(x)$ is a consistent estimator of $p_*(x)$. Since $f(x)$ is uniformly bounded, by the dominated convergence theorem

$$\begin{aligned} & \int |f(x)(h_n(x) - p_*(x))| d\mu(x) \\ & \leq \sup_x |f(x)| \int |h_n(x) - p_*(x)| d\mu(x) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

This implies that the integral itself tends to 0 in probability.

Part (iii) Consider the term

$$T_n = \sqrt{n} \int f(x)(h_n(x) - p_*(x))d\mu(x).$$

Substituting $h_n(x)$ into the expression for T_n , we get

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int f(x) \left(\frac{1}{c_n^d} K \left(\frac{x - X_i}{c_n} \right) - p_*(x) \right) d\mu(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i.$$

Note that $h_n(x)$ is asymptotically unbiased, then, as $c_n \rightarrow 0$, $\mathbb{E}[Y_i] \rightarrow 0$ and

$$\text{Var}(Y_i) \rightarrow \Sigma_f = \text{Var}(f(X)) \int [K(x)]^2 dx.$$

Then, by the central limit theorem

$$T_n \xrightarrow{d} N(0, \Sigma_f).$$

□

With this Lemma, we can prove the consistency property of the Hellinger Loss with KDE.

Theorem S1.1 (Consistency). *If (H_D) , (H_G) , (H_K) and $(H_{n,1})$ hold, then*

$$(\theta_n, \alpha_n) \xrightarrow{a.s.} (\theta_*, \alpha_*) \text{ as } n \rightarrow \infty.$$

Proof. The proof follows the same steps as in the proof of Theorem 3.1 of the manuscript. In the first step we show that $HD_n^2(\theta, \alpha)$ converges uniformly, almost surely, to $HD^2(\theta, \alpha)$. Note that $D_\alpha(x) \leq 1$, then

$$\begin{aligned} |h_{1,n}(\alpha) - h_1(\alpha)| &= \left| \int D_\alpha(x)h_n(x)d\mu(x) - \int D_\alpha(x)p_*(x)d\mu(x) \right| \\ &= \left| \int D_\alpha(x)(h_n(x) - p_*(x))d\mu(x) \right| \\ &\leq \int |h_n(x) - p_*(x)|d\mu(x), \end{aligned}$$

then

$$\lim_{n \rightarrow \infty} \int |h_n(x) - p_*(x)| d\mu(x) = 0 \quad a.s.$$

by point (i) of Lemma S1.1. In a similar way, we have

$$0 \leq \limsup_{n \rightarrow \infty} \sup_{(\theta, \alpha) \in \Theta \times \Lambda} |\gamma_n(\theta, \alpha) - \gamma(\theta, \alpha)| = 0.$$

The second step is the same as that of Theorem 3.1. \square

Finally, we introduce the following propositions to prove the analogous result of asymptotic normality.

Proposition S1.1. *Assume that assumptions (H_K) , $(H_{n,1})$, $(H_{n,2})$ and $(H_1) - (H_3)$ are satisfied. Then*

$$\sqrt{n} \nabla HD_n^2(\theta_*, \alpha_*) \longrightarrow N(0, S^{\text{KDE}}),$$

where S^{KDE} is a non-singular covariance matrix.

Proof. Recall that $D_\alpha(x)$ is bounded for each α . Then by point (ii) of Lemma S1.1

$$\begin{aligned} \int D_\alpha(x)(h_n(x) - p_*(x)) d\mu(x) &\longrightarrow 0, \\ \int D_{\alpha_*}^{\frac{1}{2}}(x)(h_n(x) - p_*(x)) d\mu(x) &\longrightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

The derivative with respect to θ given in Appendix A.1 computed at (θ_*, α_*) can be rewritten as

$$\begin{aligned} \sqrt{n} \nabla_\theta HD_n^2(\theta_*, \alpha_*) &= \\ &- \sqrt{n} \int \nabla_\theta D_{\alpha_*, \theta_*}(z) g(z) d\mu(z) \\ &+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x)(h_n(x) - p_*(x)) d\mu(x) \int \frac{\nabla_\theta D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} g(z) d\mu(z) \\ &+ \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x)p_*(x) d\mu(x) \int \frac{\nabla_\theta D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} g(z) d\mu(z) \\ &= T_1 + T_{2n} + T_3. \end{aligned}$$

By Lemma S1.1, we have that

$$T_{2n} \xrightarrow{d} N(0, S_1^{\text{KDE}}),$$

with

$$S_1^{\text{KDE}} = \text{Var}(\Delta_1(X)) \int [K(x)]^2 d(x) \quad \text{and} \quad \Delta_1(X) = D_{\alpha_*}^{\frac{1}{2}}(X) \int \frac{\nabla_\theta D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}} g(z) d\mu(z).$$

Notice that $T_1 + T_3 = 0$ since it corresponds to $\nabla_\theta HD^2(\theta_*, \alpha_*) = 0$.

Let us now consider the derivative with respect to α given in Appendix A.1 computed at (θ_*, α_*) , that can be rewritten as

$$\begin{aligned}
\sqrt{n} \nabla_\alpha HD_n^2(\theta_*, \alpha_*) &= \sqrt{n} \int \nabla_\alpha D_{\alpha_*}(x)(h_n(x) - p_*(x))d\mu(x) \\
&\quad + \sqrt{n} \int \nabla_\alpha D_{\alpha_*}(x)p_*(x)d\mu(x) \\
&\quad - \sqrt{n} \int \nabla_\alpha D_{\alpha_*, \theta_*}(z)g(z)d\mu(z) \\
&\quad - \sqrt{n} \int \frac{\nabla_\alpha D_{\alpha_*}(x)}{D_{\alpha_*}(x)^{1/2}}(h_n(x) - p_*(x))d\mu(x) \int (1 - D_{\alpha_*, \theta_*}(z))^{\frac{1}{2}}g(z)d\mu(z) \\
&\quad - \sqrt{n} \int \frac{\nabla_\alpha D_{\alpha_*}(x)}{D_{\alpha_*}(x)^{1/2}}p_*(x)d\mu(x) \int (1 - D_{\alpha_*, \theta_*}(z))^{\frac{1}{2}}g(z)d\mu(z) \\
&\quad + \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x)(h_n(x) - p_*(x))d\mu(x) \int \frac{\nabla_\alpha D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}}g(z)d\mu(z) \\
&\quad + \sqrt{n} \int D_{\alpha_*}^{\frac{1}{2}}(x)p_*(x)d\mu(x) \int \frac{\nabla_\alpha D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}}g(z)d\mu(z) \\
&= J_{1,n} + J_2 + J_3 + J_{4,n} + J_5 + J_{6,n} + J_7.
\end{aligned}$$

Notice that $J_2 + J_3 + J_5 + J_7 = 0$ since it corresponds to $\nabla_\alpha HD^2(\theta_*, \alpha_*) = 0$. The remaining terms can be rewritten as

$$J_{1,n} + J_{4,n} + J_{6,n} = \sqrt{n} \int \Delta_2(x)(h_n(x) - p_*(x))d\mu(x)$$

where

$$\begin{aligned}
\Delta_2(x) &= \nabla_\alpha D_{\alpha_*}(x) - \frac{\nabla_\alpha D_{\alpha_*}(x)}{D_{\alpha_*}(x)^{1/2}} \int (1 - D_{\alpha_*, \theta_*}(z))^{\frac{1}{2}}g(z)d\mu(z) \\
&\quad + D_{\alpha_*}^{\frac{1}{2}}(x) \int \frac{\nabla_\alpha D_{\alpha_*, \theta_*}(z)}{(1 - D_{\alpha_*, \theta_*}(z))^{1/2}}g(z)d\mu(z)
\end{aligned}$$

By Lemma S1.1, we have that

$$J_{1,n} + J_{4,n} + J_{6,n} \xrightarrow{d} N(0, S_2^{\text{KDE}}) \text{ with } S_2^{\text{KDE}} = \text{Var}(\Delta_2(X)) \int [K(x)]^2 dx.$$

Then by the central limit theorem and the continuous mapping theorem, we have that

$$(T_{2n}; J_{1n} + J_{4n} + J_{6n}) \longrightarrow N(0, S^{\text{KDE}})$$

where $S^{\text{KDE}} = \begin{pmatrix} S_1^{\text{KDE}} & S_{12}^{\text{KDE}} \\ S_{21}^{\text{KDE}} & S_2^{\text{KDE}} \end{pmatrix}$ where $S_{12}^{\text{KDE}} = \text{Cov}(\Delta_1(X), \Delta_2(X))$. \square

S2 Proof of Proposition 3.2

In this section, we provide additional details about the convergence of the other terms of the second derivatives of the Hellinger loss considered in Proposition 3.2. In the proof of Proposition 3.2 we showed the convergence of the second derivatives with respect to θ .

Let's now consider the second derivatives with respect to α reported in Appendix A.2. Using simple operations, it can be rewritten as

$$\begin{aligned} B_3 + B_4 + B_5 + B_6 + B_7 &= \int \Delta_B(x) (d\mu_n(x) - d\mu_*(x)) \\ &\quad + \int \Delta_B(x) d\mu_*(x) \end{aligned}$$

with

$$\begin{aligned} \Delta_B(x) = & - \frac{\nabla_{\alpha\alpha^\top} D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} \int (1 - D_{\alpha,\theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ & + \frac{1}{2} \frac{\nabla_\alpha D_\alpha(x) \nabla_\alpha^\top D_\alpha(x)}{D_\alpha^{\frac{3}{2}}(x)} \int (1 - D_{\alpha,\theta}(z))^{\frac{1}{2}} d\mu_g(z) \\ & + \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} \int \frac{\nabla_\alpha D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ & + D_\alpha^{\frac{1}{2}}(x) \int \frac{\nabla_{\alpha\alpha^\top} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ & + \frac{1}{2} D_\alpha^{\frac{1}{2}}(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z) \nabla_\alpha^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z). \end{aligned}$$

By assumption (H_2) , $\Delta_B(x)$ is bounded and continuous, then the first term converges to 0 as $n \rightarrow \infty$ by point (i) of Lemma 3.1. Finally, we consider the term of mixed derivatives, which can be rewritten as

$$\begin{aligned} C_2 + C_3 + C_4 &= \int \Delta_C(x) (d\mu_n(x) - d\mu_*(x)) \\ &\quad + \int \Delta_C(x) d\mu_*(x) \end{aligned}$$

with

$$\begin{aligned} \Delta_C(x) = & \frac{1}{2} \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha^{\frac{1}{2}}(x)} \int \frac{\nabla_\theta D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ & + D_\alpha^{\frac{1}{2}}(x) \int \frac{\nabla_{\theta\alpha} D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{1}{2}}} d\mu_g(z) \\ & + \frac{1}{2} D_\alpha^{\frac{1}{2}}(x) \int \frac{\nabla_\alpha D_{\alpha,\theta}(z) \nabla_\theta^\top D_{\alpha,\theta}(z)}{(1 - D_{\alpha,\theta}(z))^{\frac{3}{2}}} d\mu_g(z). \end{aligned}$$

By Lemma 3.1 the first term converges to zero.

S3 Influence Function for the standard loss function

In this section, we provide the computation of the influence function for the standard loss function used in GANs.

Let $p_\varepsilon(x) = (1 - \varepsilon)q_{\theta_0}(x) + \varepsilon h(x)$ denote the contaminated density at x , with the contaminating distribution $h(x)$ and θ_* denote the parameter values for which the model density q_{θ_*} coincides with the true distribution p_* . The underlying contaminated random

vector $Z \sim g$ is such that $X = G_{\theta_*}(Z)$, and for every ε we have $X \sim p_\varepsilon(x)$ the contaminated random vector that generates the data. We study the standard GAN log-loss function under contamination given as

$$L_\varepsilon(\theta, \alpha) = \int \ln(D_\alpha(x))p_\varepsilon(x)d\mu(x) + \int \ln(1 - D_\alpha(x))q_\theta(x)d\mu(x).$$

For all $\varepsilon \in [0, 1]$, we define

$$(\theta_\varepsilon, \alpha_\varepsilon) = \arg \inf_{\alpha} \sup_{\theta} L_\varepsilon(\theta, \alpha).$$

Notice that for $\varepsilon = 0$, $L_0(\theta, \alpha)$ denotes the uncontaminated objective function. Then, the influence functions for the standard loss is defined as

$$(IF(\theta), IF(\alpha)) = \left. \frac{\partial}{\partial \varepsilon} (\theta_\varepsilon, \alpha_\varepsilon) \right|_{\varepsilon=0}.$$

As similarly done in Section 4 of the manuscript, the optimizer $(\theta_\varepsilon, \alpha_\varepsilon)$ is solution of $\nabla L_\varepsilon(\theta_\varepsilon, \alpha_\varepsilon) = 0$, for a fixed $\varepsilon \in [0, 1)$, hence, in order to compute the influence function we can consider, for all $\varepsilon \in [0, 1)$

$$\frac{\partial}{\partial \varepsilon} \nabla L_\varepsilon(\theta_\varepsilon, \alpha_\varepsilon) = 0.$$

We have

$$\begin{aligned} \nabla_\alpha L_\varepsilon(\alpha, \theta) &= \int \frac{\nabla_\alpha D_\alpha(x)}{D_\alpha(x)} p_\varepsilon(x)d\mu(x) - \int \frac{\nabla_\alpha D_\alpha(x)}{1 - D_\alpha(x)} q_\theta(x)d\mu(x) \\ &= A_\alpha + B_\alpha \end{aligned}$$

and

$$\nabla_\theta L_\varepsilon(\alpha, \theta) = \int \ln(1 - D_\alpha(x))s_\theta(x)q_\theta(x)d\mu(x) = C_\theta,$$

where $s_\theta(x) = \nabla_\theta \log q_\theta(x)$ is the usual score function. We need to compute the derivatives with respect to ε of these terms and evaluate the expressions at $\varepsilon = 0$. For the first term

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} A_\alpha &= \int \frac{\frac{\partial}{\partial \varepsilon} \nabla_\alpha D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}(x)} p_\varepsilon(x)d\mu(x) - \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x) \frac{\partial}{\partial \varepsilon} D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}^2(x)} p_\varepsilon(x)d\mu(x) \\ &\quad + \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}(x)} \frac{\partial}{\partial \varepsilon} p_\varepsilon(x)d\mu(x) \\ &= \int \frac{\nabla_{\alpha\alpha^\top} D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}(x)} p_\varepsilon(x)d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon + \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x) \nabla_\alpha^\top D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}^2(x)} p_\varepsilon(x)d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon \\ &\quad + \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x)}{D_{\alpha_\varepsilon}(x)} (h(x) - q_{\theta_0}(x))d\mu(x) \end{aligned}$$

with $\varepsilon = 0$ we obtain

$$= A_1 \text{IF}(\alpha) + A_2 \text{IF}(\alpha) + A_3$$

Following similar steps, the derivatives computed at $\varepsilon = 0$ are given as

$$\begin{aligned}\frac{\partial}{\partial \varepsilon} B_\alpha &= - \int \frac{\frac{\partial}{\partial \varepsilon} \nabla_\alpha D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} q_{\theta_\varepsilon}(x) d\mu(x) + \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x) \frac{\partial}{\partial \varepsilon} (1 - D_{\alpha_\varepsilon}(x))}{(1 - D_{\alpha_\varepsilon}(x))^2} q_{\theta_\varepsilon}(x) d\mu(x) \\ &\quad - \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} \frac{\partial}{\partial \varepsilon} q_{\theta_\varepsilon}(x) d\mu(x) \\ &= - \int \frac{\nabla_{\alpha\alpha^\top} D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} q_{\theta_\varepsilon}(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon - \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x) \nabla_\alpha^\top D_{\alpha_\varepsilon}(x)}{(1 - D_{\alpha_\varepsilon}(x))^2} q_{\theta_\varepsilon}(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon \\ &\quad - \int \frac{\nabla_\alpha D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} s_{\theta_\varepsilon}(x) q_{\theta_\varepsilon}(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \theta_\varepsilon\end{aligned}$$

with $\varepsilon = 0$ we obtain

$$\begin{aligned}&= B_1 \text{IF}(\alpha) + B_2 \text{IF}(\alpha) + B_3 \text{IF}(\theta) \\ \frac{\partial}{\partial \varepsilon} C_\theta &= - \int \frac{\frac{\partial}{\partial \varepsilon} D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} s_{\theta_\varepsilon}(x) q_{\theta_\varepsilon}(x) d\mu(x) + \int \ln(1 - D_{\alpha_\varepsilon}(x)) \frac{\partial}{\partial \varepsilon} s_{\theta_\varepsilon}(x) q_{\theta_\varepsilon}(x) d\mu(x) \\ &\quad + \int \ln(1 - D_{\alpha_\varepsilon}(x)) s_{\theta_\varepsilon}(x) \frac{\partial}{\partial \varepsilon} q_{\theta_\varepsilon}(x) d\mu(x) \\ &= - \int \frac{\nabla_{\alpha\alpha^\top} D_{\alpha_\varepsilon}(x)}{1 - D_{\alpha_\varepsilon}(x)} s_{\theta_\varepsilon}(x) q_{\theta_\varepsilon}(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon \\ &\quad + \int \ln(1 - D_{\alpha_\varepsilon}(x)) \frac{\nabla_{\theta,\theta^\top} q_\theta(x)}{q_{\theta_\varepsilon}(x)} q_{\theta_\varepsilon}(x) d\mu(x) \frac{\partial}{\partial \varepsilon} \theta_\varepsilon \\ &= C_1 \text{IF}(\alpha) + C_2 \text{IF}(\theta)\end{aligned}$$

Combining all the terms together, we have the following two equations

$$\begin{aligned}0 &= A_3 + [A_1 + A_2 + B_1 + B_2] \text{IF}(\alpha) + B_3 \text{IF}(\theta) \\ &= D_0 + D_\alpha \text{IF}(\alpha) + D_\theta \text{IF}(\theta),\end{aligned}$$

and

$$0 = C_1 \text{IF}(\alpha) + C_2 \text{IF}(\theta)$$

Solving the system we have

$$\begin{aligned}\text{IF}(\alpha) &= -(D_\alpha - D_\theta C_\theta^{-1} C_\alpha)^{-1} D_0 \\ \text{IF}(\theta) &= -C_\theta^{-1} C_\alpha \text{IF}(\alpha) = C_\theta^{-1} C_\alpha (D_\alpha - D_\theta C_\theta^{-1} C_\alpha)^{-1} D_0.\end{aligned}$$

S4 Profiled Hellinger Distance Estimator

In this section, we develop the asymptotic theory for the profiled Hellinger distance estimator obtained by optimizing the generator parameter after profiling out the discriminator. We first establish the continuity of the empirical and population profiled criteria and the existence of optimizers, and show that profiling preserves the uniform convergence of $\text{HD}_n^2(\theta, \alpha)$ with its population counterpart. We derive a central limit theorem for the profiled estimator and relate it to the joint CLT for the full estimator. These results also extend to the approximated objective.

Lemma S4.1 (Supremum inequality). *Let (A, \mathcal{A}) be an index set and let $u, v : A \rightarrow \mathbb{R}$. Then*

$$\left| \sup_{a \in A} u(a) - \sup_{a \in A} v(a) \right| \leq \sup_{a \in A} |u(a) - v(a)|.$$

Proof. Fix $\varepsilon > 0$ and choose $a_\varepsilon \in A$ such that $\sup_a u(a) \leq u(a_\varepsilon) + \varepsilon$. Then

$$\sup_a u(a) - \sup_a v(a) \leq u(a_\varepsilon) - \sup_a v(a) + \varepsilon \leq u(a_\varepsilon) - v(a_\varepsilon) + \varepsilon \leq \sup_a |u(a) - v(a)| + \varepsilon.$$

Letting $\varepsilon \downarrow 0$ yields $\sup_a u(a) - \sup_a v(a) \leq \sup_a |u(a) - v(a)|$. Interchanging u and v gives the reverse inequality and hence the claim. \square

Lemma S4.2 (Continuity and existence of optimizers for the profiled criteria). *Assume (H_D) and (H_G) . For each $n \in \mathbb{N} \cup \{\infty\}$, define*

$$S_n(\theta) := \sup_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha), \quad \theta \in \Theta,$$

Then:

(i) HD_n^2 is continuous on $\Theta \times \Lambda$ for each n .

(ii) S_n is continuous on Θ for each n .

(iii) For each $\theta \in \Theta$, the supremum in $S_n(\theta)$ is attained: there exists $\alpha_n(\theta) \in \Lambda$ such that $S_n(\theta) = \text{HD}_n^2(\theta, \alpha_n(\theta))$.

Proof. (i) Under (H_D) and (H_G) , the maps $(x, \alpha) \mapsto D_\alpha(x)$ and $(z, \theta, \alpha) \mapsto D_\alpha(G_\theta(z))$ are continuous and bounded by 1. Inspecting the decomposition

$$\text{HD}_n^2(\theta, \alpha) = h_{1,n}(\alpha) + h_2(\theta, \alpha) - 2\gamma_n(\theta, \alpha),$$

with

$$\begin{aligned} h_{1,n}(\alpha) &= \int D_\alpha(x) h_n(x) d\mu(x), \\ h_2(\theta, \alpha) &= \int (1 - D_\alpha(G_\theta(z))) g(z) d\mu(z), \\ \gamma_n(\theta, \alpha) &= \left(\int D_\alpha^{1/2}(x) h_n(x) d\mu(x) \right) \left(\int (1 - D_\alpha(G_\theta(z)))^{1/2} g(z) d\mu(z) \right), \end{aligned}$$

we see that continuity of HD_n^2 follows in each term by dominated convergence, using the uniform bound $0 \leq D_\alpha \leq 1$ and the continuity of the integrands in (θ, α) . (The case $n = \infty$ is identical with h_n replaced by p^* .)

(ii) Since HD_n^2 is continuous on the compact set $\Theta \times \Lambda$, it is uniformly continuous. Thus, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|\text{HD}_n^2(\theta, \alpha) - \text{HD}_n^2(\theta', \alpha)| < \varepsilon$ whenever $\|\theta - \theta'\| < \delta$, for all $\alpha \in \Lambda$. Taking suprema over α yields $|S_n(\theta) - S_n(\theta')| \leq \varepsilon$, proving continuity of S_n .

(iii) By (ii) and compactness of Λ , Weierstrass' theorem gives $\alpha_n(\theta) \in \arg \max_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha)$. \square

Assumption S4.1 (H_n). *The conditions of Theorem 3.2 in the main paper hold, so that*

$$\sup_{(\theta, \alpha) \in \Theta \times \Lambda} |\text{HD}_n^2(\theta, \alpha) - \text{HD}^2(\theta, \alpha)| \longrightarrow 0 \quad \text{almost surely as } n \rightarrow \infty.$$

Proposition S4.1 (Uniform profiling over the discriminator). *Assume (H_n) . Then*

$$\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \xrightarrow{a.s.} 0,$$

where

$$S_n(\theta) = \sup_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha), \quad S(\theta) = \sup_{\alpha \in \Lambda} \text{HD}^2(\theta, \alpha).$$

Proof. Fix $\theta \in \Theta$ and apply Lemma S4.1 with $u(\alpha) = \text{HD}_n^2(\theta, \alpha)$ and $v(\alpha) = \text{HD}^2(\theta, \alpha)$ to get

$$|S_n(\theta) - S(\theta)| \leq \sup_{\alpha \in \Lambda} |\text{HD}_n^2(\theta, \alpha) - \text{HD}^2(\theta, \alpha)|.$$

Taking the supremum over $\theta \in \Theta$ yields

$$\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \leq \sup_{(\theta, \alpha) \in \Theta \times \Lambda} |\text{HD}_n^2(\theta, \alpha) - \text{HD}^2(\theta, \alpha)|.$$

By the first part of Theorem 3.2, the right-hand side converges to 0 almost surely under (H_n) . Hence the claim follows. \square

Theorem S4.1 (Consistency of the profiled estimator). *Assume (H_D) , (H_G) , and (H_n) . Let*

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} S_n(\theta), \quad \theta^* \in \arg \min_{\theta \in \Theta} S(\theta).$$

If Θ is compact and θ^ is the unique minimizer of S , then $\hat{\theta}_n \rightarrow \theta^*$ almost surely.*

Proof. By Lemma S4.2(ii) and compactness of Θ , the minima of S_n and S are attained. Proposition S4.1 gives $\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \rightarrow 0$ a.s.

Step 1 (modulus of separation of the minimum). We claim that for each $r > 0$ there exists $\eta(r) > 0$ such that

$$\inf\{S(\theta) : \theta \in \Theta, \|\theta - \theta^*\| \geq r\} \geq S(\theta^*) + \eta(r).$$

If not, we can find a sequence $(\theta_k) \subset \Theta$ with $\|\theta_k - \theta^*\| \geq r$ and $S(\theta_k) \downarrow S(\theta^*)$. By compactness of Θ there is a convergent subsequence $\theta_{k_\ell} \rightarrow \bar{\theta}$ with $\|\bar{\theta} - \theta^*\| \geq r$. Continuity of S (Lemma S4.2(ii) with $n = \infty$) gives $S(\bar{\theta}) = \lim_\ell S(\theta_{k_\ell}) = S(\theta^*)$, contradicting uniqueness of θ^* .

Step 2 (contradiction argument). Fix $r > 0$ and set $\eta = \eta(r) > 0$ from Step 1. Almost surely, for n large enough we have $\sup_{\theta} |S_n(\theta) - S(\theta)| < \eta/2$. Then

$$S_n(\hat{\theta}_n) \leq S_n(\theta^*) \leq S(\theta^*) + \eta/2. \tag{1}$$

If $\|\hat{\theta}_n - \theta^*\| \geq r$, then Step 1 implies $S(\hat{\theta}_n) \geq S(\theta^*) + \eta$, hence

$$S_n(\hat{\theta}_n) \geq S(\hat{\theta}_n) - \eta/2 \geq S(\theta^*) + \eta/2,$$

which contradicts the previous bound. Therefore $\|\hat{\theta}_n - \theta^*\| < r$ eventually almost surely. Since $r > 0$ is arbitrary, we conclude $\hat{\theta}_n \rightarrow \theta^*$ almost surely. \square

Remark S4.1 (Empirical and population optimizers are linked). For each n , Lemma S4.2(iii) gives

$$\alpha_n(\theta) \in \arg \max_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha),$$

so any saddle-point solution $(\theta_n, \alpha_n) \in \arg \min_{\theta \in \Theta} \arg \max_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha)$ satisfies

$$\theta_n \in \arg \min_{\theta \in \Theta} S_n(\theta).$$

Theorem S4.1 therefore implies the almost sure consistency of the generator estimate obtained by profiling over the discriminator parameter.

Proposition S4.2 (approximated objective). *Define*

$$\tilde{S}_n(\theta) := \sup_{\alpha \in \Lambda} \tilde{L}_n(\theta, \alpha)$$

and

$$\tilde{S}(\theta) := \sup_{\alpha \in \Lambda} \tilde{L}(\theta, \alpha),$$

where $\tilde{L}_n(\theta, \alpha) = 2\gamma_n(\theta, \alpha) - 2$ and $\tilde{L}(\theta, \alpha) = 2\gamma(\theta, \alpha) - 2$. Under (H_n) , $\sup_{\theta} |\tilde{S}_n(\theta) - \tilde{S}(\theta)| \rightarrow 0$ almost surely. If, in addition, Θ is compact and \tilde{S} has a unique minimizer θ^\dagger , then any $\hat{\theta}_n^{\text{approx}} \in \arg \min_{\theta \in \Theta} \tilde{S}_n(\theta)$ satisfies $\hat{\theta}_n^{\text{approx}} \rightarrow \theta^\dagger$ almost surely.

Proof. For fixed θ , Lemma S4.1 gives $|\sup_{\alpha} \gamma_n(\theta, \alpha) - \sup_{\alpha} \gamma(\theta, \alpha)| \leq \sup_{\alpha} |\gamma_n(\theta, \alpha) - \gamma(\theta, \alpha)|$. Taking \sup_{θ} and using the a.s. uniform convergence $\sup_{\theta, \alpha} |\gamma_n(\theta, \alpha) - \gamma(\theta, \alpha)| \rightarrow 0$ (from the proof of Theorem 3.2) yields the first claim. The consistency claim follows exactly as in Theorem S4.1. \square

S4.1 Central Limit Theorem

Definition S4.1 (Profiled objective and profiled estimator). For each $\theta \in \Theta$, let

$$S_n(\theta) := \sup_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha), \quad S(\theta) := \sup_{\alpha \in \Lambda} \text{HD}^2(\theta, \alpha).$$

A *profiled* generator estimator is any

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} S_n(\theta),$$

with a corresponding $\hat{\alpha}_n \in \arg \max_{\alpha \in \Lambda} \text{HD}_n^2(\hat{\theta}_n, \alpha)$.

The next assumption concerns block nonsingularity and interior solutions.

Assumption S4.2. (i) For each θ in a neighborhood of θ^* there is a unique interior maximizer $\alpha^!(\theta) \in \Lambda$ of $\alpha \mapsto \text{HD}^2(\theta, \alpha)$, with $\alpha^!(\theta^*) = \alpha^*$.
(ii) The block Hessian $\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)$ is nonsingular.
(iii) The profile Hessian

$$\widetilde{H}_{\theta\theta} := \nabla_{\theta\theta}^2 \text{HD}^2(\theta^*, \alpha^*) - \nabla_{\theta\alpha}^2 \text{HD}^2(\theta^*, \alpha^*) \left[\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*) \right]^{-1} \nabla_{\alpha\theta}^2 \text{HD}^2(\theta^*, \alpha^*)$$

is positive definite.

Lemma S4.3 (Envelope and implicit function identities). *Assume (H_D) , (H_G) and Assumption S4.2. Then $\alpha(\cdot)$ is C^1 in a neighborhood of θ^* and*

$$\nabla_\theta S(\theta) = \nabla_\theta \text{HD}^2(\theta, \alpha(\theta)), \quad \frac{d\alpha(\theta)}{d\theta} = -[\nabla_{\alpha\alpha}^2 \text{HD}^2]^{-1} \nabla_{\alpha\theta}^2 \text{HD}^2,$$

evaluated at $(\theta, \alpha(\theta))$. Moreover

$$\nabla_{\theta\theta}^2 S(\theta^*) = \widetilde{H}_{\theta\theta}.$$

Proof. The map $(\theta, \alpha) \mapsto \text{HD}^2(\theta, \alpha)$ is C^2 by (H_D) and (H_G) . By Assumption S4.2(i)–(ii), we have that $\nabla_\alpha \text{HD}^2(\theta, \alpha(\theta)) = 0$ and $\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)$ is invertible, so the implicit function theorem gives C^1 -smoothness of $\alpha'(\cdot)$ and the derivative formula. The envelope identity then follows by the chain rule together with $\nabla_\alpha \text{HD}^2(\theta, \alpha(\theta)) = 0$. Differentiating once more and substituting $d\alpha/d\theta$ yields the Schur-complement expression stated for $\nabla_{\theta\theta}^2 S(\theta^*)$. \square

Lemma S4.4 (First-order expansion of the empirical profile score). *Assume (H_D) , (H_G) , (H_n) , (H_1) – (H_3) and Assumption S4.2. Let $\alpha_n(\theta) \in \arg \max_{\alpha \in \Lambda} \text{HD}_n^2(\theta, \alpha)$. Then*

$$\begin{aligned} \sqrt{n} \nabla_\theta S_n(\theta^*) &= \\ &\sqrt{n} \left\{ \nabla_\theta \text{HD}_n^2(\theta^*, \alpha^*) - \nabla_{\theta\alpha}^2 \text{HD}^2(\theta^*, \alpha^*) [\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)]^{-1} \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha^*) \right\} + o_p(1). \end{aligned}$$

Consequently,

$$\sqrt{n} \nabla_\theta S_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, S_{\text{eff}}),$$

where

$$S_{\text{eff}} := \text{Var}\left(\Delta_\theta - \nabla_{\theta\alpha}^2 \text{HD}^2[\nabla_{\alpha\alpha}^2 \text{HD}^2]^{-1} \Delta_\alpha\right),$$

$$\Delta_\theta := \sqrt{n} \nabla_\theta \text{HD}_n^2(\theta^*, \alpha^*), \quad \Delta_\alpha := \sqrt{n} \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha^*).$$

All derivatives above are evaluated at (θ^*, α^*) .

Proof. By Danskin's/envelope theorem in the C^1 case,

$$\nabla_\theta S_n(\theta) = \nabla_\theta \text{HD}_n^2(\theta, \alpha_n(\theta))$$

because $\nabla_\alpha \text{HD}_n^2(\theta, \alpha_n(\theta)) = 0$. Evaluating at θ^* and Taylor expanding in α around α^* ,

$$\nabla_\theta S_n(\theta^*) = \nabla_\theta \text{HD}_n^2(\theta^*, \alpha^*) + \nabla_{\theta\alpha}^2 \text{HD}^2(\theta^*, \alpha^*) [\alpha_n(\theta^*) - \alpha^*] + r_n,$$

with $r_n = o_p(n^{-1/2})$. By standard M-estimation theory under (H_1) – (H_3) (see, e.g., van der Vaart, 1998, Thm. 5.41), the maximizer $\alpha_n(\theta^*)$ satisfies

$$\sqrt{n} (\alpha_n(\theta^*) - \alpha^*) = -H_{\alpha\alpha}^{-1} \sqrt{n} \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha^*) + o_p(1),$$

where $H_{\alpha\alpha} = \nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)$. Next, expand the first order condition of α at $(\theta^*, \alpha_n(\theta^*))$:

$$0 = \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha_n(\theta^*)) = \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha^*) + \nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*) [\alpha_n(\theta^*) - \alpha^*] + o_p(n^{-1/2}).$$

Since $\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)$ is nonsingular,

$$\alpha_n(\theta^*) - \alpha^* = -[\nabla_{\alpha\alpha}^2 \text{HD}^2(\theta^*, \alpha^*)]^{-1} \nabla_\alpha \text{HD}_n^2(\theta^*, \alpha^*) + o_p(n^{-1/2}).$$

Substitute into the first expansion, multiply by \sqrt{n} , and use Proposition 3.1 to obtain the stated linear representation and asymptotic normality with covariance S_{eff} . \square

The next lemma is concerned with the convergence of the empirical profile Hessian.

Lemma S4.5. *Under the assumptions of Lemma S4.4,*

$$\nabla_{\theta\theta}^2 S_n(\tilde{\theta}_n) \xrightarrow{p} \widetilde{H}_{\theta\theta}$$

for any sequence $\tilde{\theta}_n \rightarrow \theta^*$.

Proof. Let $S_n(\theta) = \text{HD}_n^2(\theta, \alpha_n(\theta))$ and differentiate, by the chain rule,

$$\nabla_{\theta\theta}^2 S_n(\theta) = \nabla_{\theta\theta}^2 \text{HD}_n^2(\theta, \alpha_n(\theta)) + \nabla_{\theta\alpha}^2 \text{HD}_n^2(\theta, \alpha_n(\theta)) \frac{d\alpha_n(\theta)}{d\theta}.$$

Differentiating the empirical first order condition of α , $\nabla_\alpha \text{HD}_n^2(\theta, \alpha_n(\theta)) = 0$ yields

$$\frac{d\alpha_n(\theta)}{d\theta} = -[\nabla_{\alpha\alpha}^2 \text{HD}_n^2(\theta, \alpha_n(\theta))]^{-1} \nabla_{\alpha\theta}^2 \text{HD}_n^2(\theta, \alpha_n(\theta))$$

whenever the inverse exists. By Proposition 3.2 (elementwise convergence of second derivatives) and consistency $\alpha_n(\tilde{\theta}_n) \rightarrow \alpha^*$, each empirical block converges in probability to its population counterpart, hence the whole expression converges to the Schur complement $\widetilde{H}_{\theta\theta}$. \square

Theorem S4.2 (CLT for the profiled generator estimator). *Assume (H_D) , (H_G) , (H_n) , (H_1) – (H_3) and Assumption S4.2. Let $\hat{\theta}_n \in \arg \min_{\theta \in \Theta} S_n(\theta)$ and suppose Θ is compact and S has a unique minimizer θ^* . Then*

$$\sqrt{n} (\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \widetilde{H}_{\theta\theta}^{-1} S_{\text{eff}} \widetilde{H}_{\theta\theta}^{-1}),$$

with $\widetilde{H}_{\theta\theta}$ and S_{eff} as in Lemmas S4.3 and S4.4.

Proof. By optimality, $0 = \nabla_\theta S_n(\hat{\theta}_n) = \nabla_\theta S_n(\theta^*) + \nabla_{\theta\theta}^2 S_n(\tilde{\theta}_n) (\hat{\theta}_n - \theta^*)$ for some $\tilde{\theta}_n$ on the line segment between $\hat{\theta}_n$ and θ^* . Multiply by \sqrt{n} and rearrange:

$$\sqrt{n} (\hat{\theta}_n - \theta^*) = -[\nabla_{\theta\theta}^2 S_n(\tilde{\theta}_n)]^{-1} \sqrt{n} \nabla_\theta S_n(\theta^*).$$

By Lemma S4.5, $\nabla_{\theta\theta}^2 S_n(\tilde{\theta}_n) \xrightarrow{p} \widetilde{H}_{\theta\theta}$, which is invertible by Assumption S4.2(iii), and by Lemma S4.4, $\sqrt{n} \nabla_\theta S_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, S_{\text{eff}})$. The result follows by Slutsky's theorem. \square

Remark S4.2 (Connection with Theorem 3.3). Let $H := \nabla^2 \text{HD}^2(\theta^*, \alpha^*)$ and write it in block form with respect to (α, θ) . The joint CLT of Theorem 3.3 gives $\sqrt{n}((\theta_n, \alpha_n) - (\theta^*, \alpha^*)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma = H^{-1} S(H^{-1})^\top$ with $S = \text{Var}(\sqrt{n} \nabla \text{HD}_n^2)$. A standard block inversion shows that the θ -marginal covariance equals $\widetilde{H}_{\theta\theta}^{-1} S_{\text{eff}} \widetilde{H}_{\theta\theta}^{-1}$, i.e. it coincides with the variance in Theorem S4.2. Thus the profiled CLT is exactly the θ -component of the joint CLT.

Remark S4.3 (Approximated objective). All statements above hold verbatim with HD_n^2 replaced by \tilde{L}_n (Section 2.1), since the derivative blocks and limit arguments used in Lemmas S4.3–S4.5 are the same.

Table S1: Median (standard deviation) of MSE ($\times 100$) of μ at the final epoch for the considered methods and percentage of contamination ε .

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	0.01 (1.9)	0.15 (11.6)	3.28 (1046.8)	54.89 (2124.0)	152.37 (490.5)
WGAN	34.27 (15.8)	25.03 (31.0)	116.46 (4283.5)	230.51 (1949.5)	578.34 (4954.6)
Approx. HD	0.99 (7.7)	0.73 (385.7)	21.96 (921.5)	2.67 (291.7)	146.72 (129.0)
HD ($c_n=0.0001$)	63.33 (68.4)	18.78 (6201.4)	11.00 (3734.6)	15.12 (2137.7)	7.22 (374.9)
HD ($c_n=0.01$)	26.69 (1072.7)	32.72 (7199.9)	83.15 (4426.2)	37.21 (5805.2)	147.95 (4804.0)
HD ($c_n=0.5$)	25.66 (1358.5)	20.29 (128.6)	17.75 (691.6)	93.31 (3533.9)	100.51 (514.8)

Table S2: Median (standard deviation) of MSE ($\times 100$) of σ at the final epoch for the considered methods and percentage of contamination ε .

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	11.18 (94.8)	3.06 (220.3)	28.52 (595.4)	331.04 (1939.5)	446.08 (836.9)
WGAN	5.00 (71.6)	2.18 (14.7)	69.05 (2662.6)	85.24 (2427.0)	1110.06 (2078.6)
Approx. HD	0.58 (58.5)	30.40 (209.5)	827.90 (708.0)	9.19 (619.8)	255.01 (665.0)
HD ($c_n=0.0001$)	10.31 (27.2)	20.93 (1312.4)	33.62 (1538.4)	37.02 (914.8)	29.62 (1096.0)
HD ($c_n=0.01$)	16.07 (55.1)	39.62 (1005.6)	112.32 (1313.8)	184.94 (2926.0)	675.86 (8350.9)
HD ($c_n=0.5$)	28.17 (2796.4)	30.00 (39.9)	65.55 (2393.2)	243.55 (1306.6)	584.61 (955.6)

Table S3: Median (standard deviation) of RMSE ($\times 100$) for (μ, σ) at the final epoch for the considered methods and percentage of contamination ε .

	$\varepsilon=0\%$	$\varepsilon=1\%$	$\varepsilon=5\%$	$\varepsilon=10\%$	$\varepsilon=20\%$
GAN	23.51 (43.5)	12.30 (68.8)	41.11 (134.6)	171.53 (206.3)	173.59 (141.4)
WGAN	44.20 (16.9)	36.71 (15.2)	105.47 (291.5)	125.57 (231.6)	351.97 (304.7)
Approx. HD	11.80 (28.1)	43.31 (85.9)	213.45 (138.3)	23.86 (124.0)	140.27 (114.6)
HD ($c_n=0.0001$)	65.24 (26.6)	57.23 (259.1)	55.11 (246.9)	47.47 (173.7)	42.98 (146.3)
HD ($c_n=0.01$)	54.12 (85.3)	65.00 (232.5)	107.42 (173.9)	110.96 (214.1)	202.90 (281.9)
HD ($c_n=0.5$)	54.58 (154.1)	50.09 (31.3)	65.56 (139.2)	131.67 (169.0)	190.86 (85.5)

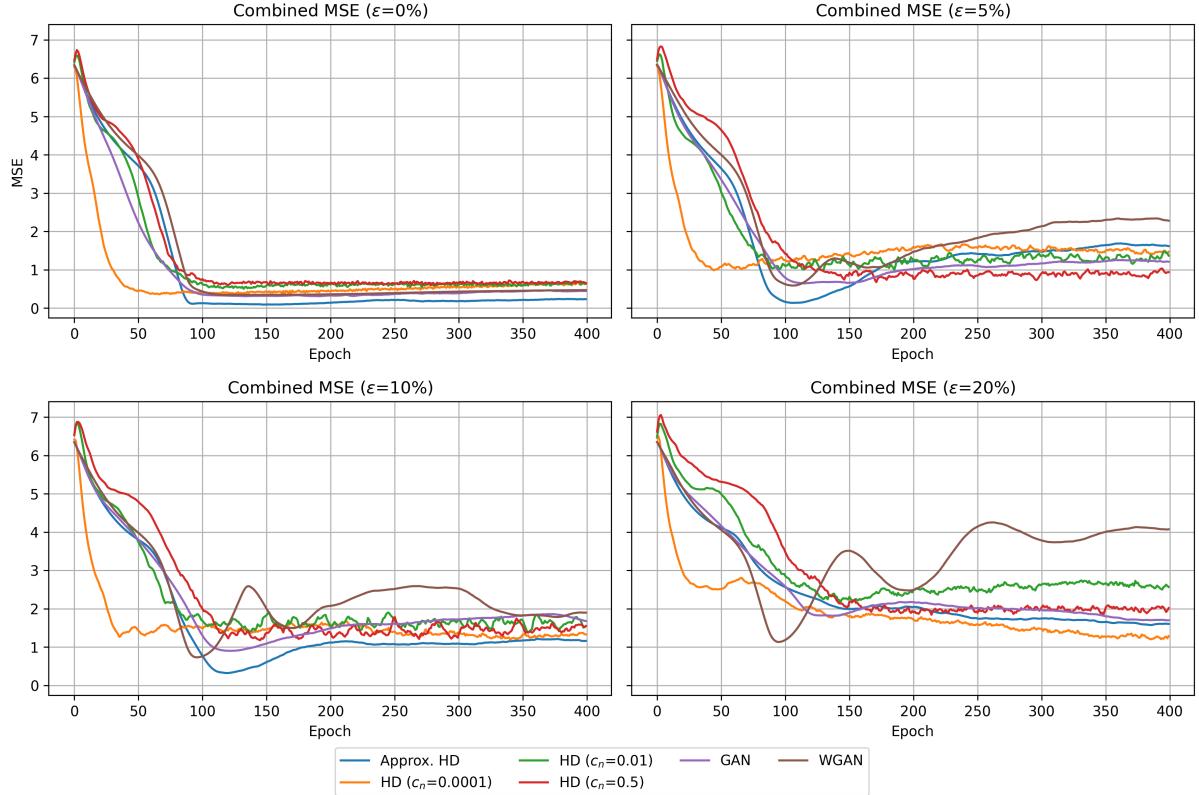


Figure S1: Mean Squared Error (MSE) per epoch for the different contamination percentage ε comparing the different methods.

S5 Additional Simulation Results

In this section we complement the simulation study results presented in Section 5 of the manuscript. Figure S2 reports, for each contamination proportion $\varepsilon \in \{0\%, 5\%, 10\%, 20\%\}$, the RMSE per epoch of the estimated location-scale parameters in the Gaussian model for all considered loss functions, namely the standard GAN, WGAN, the approximate Hellinger loss, and the KDE-based Hellinger losses with different bandwidth choices. Tables S1 and Table S2 report the median and standard deviation of the MSE for the location and scale parameters of the competing estimators at the final training epoch. While Table S3 shows the RMSE at the final epoch.

References

- N. Glick. Consistency conditions for probability estimators and integrals of density estimators. *Utilitas Mathematica*, 6:64 – 74, 1974.