Some Theoretical Properties of GANs

Gérard Biau, Benoît Cadre, M. Sangnier, U. Tanielian

Some Theoretical Properties of GANs

G. Biau
Sorbonne Université, CNRS, LPSM
Paris, France
gerard.biau@upmc.fr

B. Cadre
Univ Rennes, CNRS, IRMAR
Rennes, France
benoit.cadre@ens-rennes.fr

M. Sangnier
Sorbonne Université, CNRS, LPSM
Paris, France
maxime.sangnier@upmc.fr

U. Tanielian
Sorbonne Université, CNRS, LPSM, Criteo
Paris, France
u.tanielian@criteo.com

Abstract

Generative Adversarial Networks (GANs) are a class of generative algorithms that have been shown to produce state-of-the art samples, especially in the domain of image creation. The fundamental principle of GANs is to approximate the unknown distribution of a given data set by optimizing an objective function through an adversarial game between a family of generators and a family of discriminators. In this paper, we offer a better theoretical understanding of GANs by analyzing some of their mathematical and statistical properties. We study the deep connection between the adversarial principle underlying GANs and the Jensen-Shannon divergence, together with some optimality characteristics of the problem. An analysis of the role of the discriminator family via approximation arguments is also provided. In addition, taking a statistical point of view, we study the large sample properties of the estimated distribution and prove in particular a central limit theorem. Some of our results are illustrated with simulated examples.

1 Introduction

The fields of machine learning and artificial intelligence have seen spectacular advances in recent years, one of the most promising being perhaps the success of Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014). GANs are a class of generative algorithms implemented by a system of two neural networks contesting with each other in a zero-sum game framework. This technique is now recognized as being capable of generating photographs that look authentic to human observers (e.g., Salimans et al., 2016), and its spectrum of applications is growing at a fast pace, with impressive results in the domains of inpainting, speech, and 3D modeling, to name but a few. A survey of the most recent advances is given by Goodfellow (2016).

The objective of GANs is to generate fake observations of a target distribution $p^\star$ from which only a true sample (e.g., real-life images represented using raw pixels) is available. It should be pointed out at the outset that the data involved in the domain are usually so complex that no exhaustive description of $p^\star$ by a classical parametric model is appropriate, nor its estimation by a traditional maximum likelihood approach. Similarly, the dimension of the samples is often very large, and this effectively excludes a strategy based on non-parametric density estimation techniques such as kernel or nearest neighbor smoothing, for example. In order to generate according to $p^\star$, GANs proceed by an adversarial scheme involving two components: a family of generators and a family of discriminators, which are both implemented by neural networks. The generators admit low-dimensional random observations with a known distribution (typically Gaussian or uniform) as input, and attempt to transform them into fake data that can match the distribution $p^\star$; on the other hand, the discriminators aim to accurately discriminate between the true observations from $p^\star$ and those produced by the generators. The generators and the discriminators are calibrated by optimizing an objective function in such a way that the distribution of the generated sample is as indistinguishable as possible from that of the original data. In pictorial terms, this process is often compared to a game of cops and robbers, in which a team of counterfeiters illegally produces banknotes and tries to make them undetectable in the eyes of a team of police officers, whose objective is of course the opposite. The competition pushes both teams to improve their methods until counterfeit money becomes indistinguishable (or not) from genuine currency.

From a mathematical point of view, here is how the generative process of GANs can be represented. All the densities that we consider in the article are supposed to be dominated by a fixed, known, measure μ on E, where E is a Borel subset of $\mathbb{R}^d$. This dominating measure is typically the Lebesgue or the counting measure, but, depending on the practical context, it can be a more complex measure. We assume to have at hand an i.i.d. sample $X_1, \ldots, X_n$, drawn according to some unknown density $p^\star$ on E. These random variables model the available data, such as images or video sequences; they typically take their values in a high-dimensional space, so that the ambient dimension d must be thought of as large. The generators as a whole have the form of a parametric family of functions from $\mathbb{R}^{d'}$ to E ($d' \ll d$), say $\mathscr{G} = \{G_\theta\}_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^p$. Each function G_θ is intended to be applied to a d'-dimensional random variable Z (sometimes called the noise—in most cases Gaussian or uniform), so that there is a natural family of densities associated with the generators, say $\mathscr{P} = \{p_\theta\}_{\theta\in\Theta}$, where, by definition, $G_\theta(Z) \overset{\mathscr{L}}{=} p_\theta \mathrm{d}\mu$. In this model, each density p_θ is a potential candidate to represent $p^\star$. On the other hand, the discriminators are described by a family of Borel functions from E to $[0,1]$, say $\mathscr{D}$, where each $D \in \mathscr{D}$ must be thought of as the probability that an observation comes from $p^\star$ (the higher $D(x)$, the higher the probability that x is drawn from $p^\star$). At some point, but not always, we will assume that $\mathscr{D}$ is in fact a parametric class, of the form $\{D_\alpha\}_{\alpha\in\Lambda}$, $\Lambda \subset \mathbb{R}^q$, as is certainly always the case in practice. In GANs algorithms, both parametric models $\{G_\theta\}_{\theta\in\Theta}$ and $\{D_\alpha\}_{\alpha\in\Lambda}$ take the form of neural networks, but this does not play a fundamental role in this paper. We will simply remember that the dimensions p and q are potentially very large, which takes us away from a classical parametric setting. We also insist on the fact that it is <u>not</u> assumed that $p^\star$ belongs to $\mathscr{P}$.

Let $Z_1, \ldots, Z_n$ be an i.i.d. sample of random variables, all distributed as the noise Z. The objective is to solve in θ the problem

$$\inf_{\theta \in \Theta} \sup_{D \in \mathscr{D}} \left[\prod_{i=1}^{n} D(X_i) \times \prod_{i=1}^{n} (1 - D \circ G_\theta(Z_i)) \right],$$

or, equivalently, to find $\hat{\theta} \in \Theta$ such that

$$\sup_{D \in \mathscr{D}} \hat{L}(\hat{\theta}, D) \leq \sup_{D \in \mathscr{D}} \hat{L}(\theta, D), \quad \forall \theta \in \Theta, \tag{1}$$

where

$$\hat{L}(\theta, D) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ln D(X_i) + \sum_{i=1}^{n} \ln(1 - D \circ G_\theta(Z_i))$$

(ln is the natural logarithm). In this problem, $D(x)$ represents the probability that an observation x comes from $p^\star$ rather than from p_θ. Therefore, for each θ, the discriminators (the police team) try to distinguish the original sample $X_1, \ldots, X_n$ from the fake one $G_\theta(Z_1), \ldots, G_\theta(Z_n)$ produced by the generators (the counterfeiters' team), by maximizing D on the X_i and minimizing it on the $G_\theta(Z_i)$. Of course, the generators have an exact opposite objective, and adapt the fake data in such a way as to mislead the discriminators' likelihood. All in all, we see that the criterion seeks to find the right balance between the conflicting interests of the generators and the discriminators. The hope is that the $\hat{\theta}$ achieving equilibrium will make it possible to generate observations $G_{\hat{\theta}}(Z_1), \ldots, G_{\hat{\theta}}(Z_n)$ indistinguishable from reality, i.e., observations with a law close to the unknown $p^\star$.

The criterion $\hat{L}(\theta, D)$ involved in (1) is the criterion originally proposed in the adversarial framework of Goodfellow et al. (2014). Since then, the success of GANs in applications has led to a large volume of literature on variants, which all have many desirable properties but are based on different optimization criteria—examples are MMD-GANs (Dziugaite et al., 2015), f-GANs (Nowozin et al., 2016), Wasserstein-GANs (Arjovsky et al., 2017), and an approach based on scattering transforms (Angles and Mallat, 2018). All these variations and their innumerable algorithmic versions constitute the galaxy of GANs. That being said, despite increasingly spectacular applications, little is known about the mathematical and statistical forces behind these algorithms (e.g., Arjovsky and Bottou, 2017; Liu et al., 2017; Zhang et al., 2018), and, in fact, nearly nothing about the primary adversarial problem (1). As acknowledged by Liu et al. (2017), basic questions on how well GANs can approximate the target distribution $p^\star$ remain largely unanswered. In particular, the role and impact of the discriminators on the quality of the approximation are still a mystery, and simple but fundamental questions regarding statistical consistency and rates of convergence remain open.

In the present article, we propose to take a small step towards a better theoretical understanding of GANs by analyzing some of the mathematical and statistical properties of the original adversarial problem (1). In Section 2, we study the deep connection between the population version of (1) and the Jensen-Shannon divergence, together with some optimality characteristics of the problem, often referred to in the literature but in fact poorly understood. Section 3 is devoted to a better comprehension of the role of the discriminator family via

approximation arguments. Finally, taking a statistical point of view, we study in Section 4 the large sample properties of the distribution $p_{\hat{\theta}}$ and $\hat{\theta}$, and prove in particular a central limit theorem for this parameter. Some of our results are illustrated with simulated examples. For clarity, most technical proofs are gathered in Section 5.

2 Optimality properties

We start by studying some important properties of the adversarial principle, emphasizing the role played by the Jensen-Shannon divergence. We recall that if P and Q are probability measures on E, and P is absolutely continuous with respect to Q, then the Kullback-Leibler divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int \ln \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P,$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P with respect to Q. The Kullback-Leibler divergence is always nonnegative, with $D_{\mathrm{KL}}(P \parallel Q)$ zero if and only if $P = Q$. If $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ exist (meaning that P and Q are absolutely continuous with respect to μ, with densities p and q), then the Kullback-Leibler divergence is given as

$$D_{\mathrm{KL}}(P \parallel Q) = \int p \ln \frac{p}{q} \mathrm{d}\mu,$$

and alternatively denoted by $D_{\mathrm{KL}}(p \parallel q)$. We also recall that the Jensen-Shannon divergence is a symmetrized version of the Kullback-Leibler divergence. It is defined for any probability measures P and Q on E by

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2} D_{\mathrm{KL}}\Big(P \,\Big\|\, \frac{P+Q}{2}\Big) + \frac{1}{2} D_{\mathrm{KL}}\Big(Q \,\Big\|\, \frac{P+Q}{2}\Big),$$

and satisfies $0 \leq D_{\mathrm{JS}}(P,Q) \leq \ln 2$. The square root of the Jensen-Shannon divergence is a metric often referred to as Jensen-Shannon distance (Endres and Schindelin, 2003). When P and Q have densities p and q with respect to μ, we use the notation $D_{\mathrm{JS}}(p,q)$ in place of $D_{\mathrm{JS}}(P,Q)$.

For a generator G_θ and an arbitrary discriminator $D \in \mathscr{D}$, the criterion $\hat{L}(\theta, D)$ to be optimized in (1) is but the empirical version of the probabilistic criterion

$$L(\theta, D) \stackrel{\text{def}}{=} \int \ln(D) p^\star \mathrm{d}\mu + \int \ln(1-D) p_\theta \mathrm{d}\mu.$$

We assume for the moment that the discriminator class $\mathscr{D}$ is not restricted and equals $\mathscr{D}_\infty$, the set of all Borel functions from E to $[0,1]$. We note however that, for all $\theta \in \Theta$,

$$0 \geq \sup_{D \in \mathscr{D}_\infty} L(\theta, D) \geq -\ln 2 \Big(\int p^\star \mathrm{d}\mu + \int p_\theta \mathrm{d}\mu \Big) = -\ln 4,$$

so that $\inf_{\theta \in \Theta} \sup_{D \in \mathscr{D}_\infty} L(\theta, D) \in [-\ln 4, 0]$. Thus,

$$\inf_{\theta \in \Theta} \sup_{D \in \mathscr{D}_\infty} L(\theta, D) = \inf_{\theta \in \Theta} \sup_{D \in \mathscr{D}_\infty : L(\theta, D) > -\infty} L(\theta, D).$$

This identity points out the importance of discriminators such that $L(\theta, D) > -\infty$, which we call θ-admissible. In the sequel, in order to avoid unnecessary problems of integrability, we only consider such discriminators, keeping in mind that the others have no interest.

Of course, working with $\mathscr{D}_\infty$ is somehow an idealized vision, since in practice the discriminators are always parameterized by some parameter $\alpha \in \Lambda$, $\Lambda \subset \mathbb{R}^q$. Nevertheless, this point of view is informative and, in fact, is at the core of the connection between our generative problem and the Jensen-Shannon divergence. Indeed, taking the supremum of $L(\theta, D)$ over $\mathscr{D}_\infty$, we have

$$\begin{aligned}\sup_{D \in \mathscr{D}_\infty} L(\theta, D) &= \sup_{D \in \mathscr{D}_\infty} \int \big[\ln(D) p^\star + \ln(1-D) p_\theta\big] \mathrm{d}\mu \\ &\leq \int \sup_{D \in \mathscr{D}_\infty} \big[\ln(D) p^\star + \ln(1-D) p_\theta\big] \mathrm{d}\mu \\ &= L(\theta, D^\star_\theta),\end{aligned}$$

where

$$D^\star_\theta \stackrel{\text{def}}{=} \frac{p^\star}{p^\star + p_\theta}. \tag{2}$$

By observing that $L(\theta, D^\star_\theta) = 2D_{\mathrm{JS}}(p^\star, p_\theta) - \ln 4$, we conclude that, for all $\theta \in \Theta$,

$$\sup_{D \in \mathscr{D}_\infty} L(\theta, D) = L(\theta, D^\star_\theta) = 2D_{\mathrm{JS}}(p^\star, p_\theta) - \ln 4.$$

In particular, $D^\star_\theta$ is θ-admissible. The fact that $D^\star_\theta$ realizes the supremum of $L(\theta, D)$ over $\mathscr{D}_\infty$ and that this supremum is connected to the Jensen-Shannon divergence between $p^\star$ and p_θ appears in the original article by Goodfellow et al. (2014). This remark has given rise to many developments that interpret the adversarial problem (1) as the empirical version of the minimization problem $\inf_\theta D_{\mathrm{JS}}(p^\star, p_\theta)$ over Θ. Accordingly, many GANs algorithms try to learn the optimal function $D^\star_\theta$, using for example stochastic gradient descent techniques and mini-batch approaches. However, it has not been known until now whether $D^\star_\theta$ is unique as a maximizer of $L(\theta, D)$ over all D. Our first result shows that this is indeed the case.

Theorem 2.1. *Let $\theta \in \Theta$ be such that $p_\theta > 0$ μ-almost everywhere. Then the function $D^\star_\theta$ is the unique discriminator that achieves the supremum of the functional $D \mapsto L(\theta, D)$ over $\mathscr{D}_\infty$, i.e.,*

$$\{D^\star_\theta\} = \operatorname*{arg\,max}_{D \in \mathscr{D}_\infty} L(\theta, D).$$

Proof. Let $D \in \mathscr{D}_\infty$ be a discriminator such that $L(\theta, D) = L(\theta, D^\star_\theta)$. In particular, $L(\theta, D) > -\infty$ and D is θ-admissible. We have to show that $D = D^\star_\theta$. Notice that

$$\int \ln(D) p^\star \mathrm{d}\mu + \int \ln(1-D) p_\theta \mathrm{d}\mu = \int \ln(D^\star_\theta) p^\star \mathrm{d}\mu + \int \ln(1-D^\star_\theta) p_\theta \mathrm{d}\mu. \tag{3}$$

Thus,

$$-\int \ln\Big(\frac{D^\star_\theta}{D}\Big) p^\star \mathrm{d}\mu = \int \ln\Big(\frac{1-D^\star_\theta}{1-D}\Big) p_\theta \mathrm{d}\mu,$$

i.e., by definition of $D_\theta^\star$,

$$-\int \ln\Big(\frac{p^\star}{D(p^\star+p_\theta)}\Big)p^\star \mathrm{d}\mu = \int \ln\Big(\frac{p_\theta}{(1-D)(p^\star+p_\theta)}\Big)p_\theta \mathrm{d}\mu. \tag{4}$$

Let $\mathrm{d}P^\star = p^\star \mathrm{d}\mu$, $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$,

$$\mathrm{d}\kappa = \frac{D(p^\star+p_\theta)}{\int D(p^\star+p_\theta)\mathrm{d}\mu}\mathrm{d}\mu, \quad \text{and} \quad \mathrm{d}\kappa' = \frac{(1-D)(p^\star+p_\theta)}{\int (1-D)(p^\star+p_\theta)\mathrm{d}\mu}\mathrm{d}\mu.$$

With this notation, identity (4) becomes

$$-D_{\mathrm{KL}}(P^\star \parallel \kappa) + \ln\Big[\int D(p^\star+p_\theta)\mathrm{d}\mu\Big] = D_{\mathrm{KL}}(P_\theta \parallel \kappa') - \ln\Big[\int (1-D)(p^\star+p_\theta)\mathrm{d}\mu\Big].$$

Upon noting that

$$\int (1-D)(p^\star+p_\theta)\mathrm{d}\mu = 2 - \int D(p^\star+p_\theta)\mathrm{d}\mu,$$

we obtain

$$D_{\mathrm{KL}}(P^\star \parallel \kappa) + D_{\mathrm{KL}}(P_\theta \parallel \kappa') = \ln\Big[\int D(p^\star+p_\theta)\mathrm{d}\mu\,\big(2 - \int D(p^\star+p_\theta)\mathrm{d}\mu\big)\Big].$$

Since $\int D(p^\star+p_\theta)\mathrm{d}\mu \in [0,2]$, we find that $D_{\mathrm{KL}}(P^\star \parallel \kappa) + D_{\mathrm{KL}}(P_\theta \parallel \kappa') \leq 0$, which implies

$$D_{\mathrm{KL}}(P^\star \parallel \kappa) = 0 \quad \text{and} \quad D_{\mathrm{KL}}(P_\theta \parallel \kappa') = 0.$$

Consequently,

$$p^\star = \frac{D(p^\star+p_\theta)}{\int D(p^\star+p_\theta)\mathrm{d}\mu} \quad \text{and} \quad p_\theta = \frac{(1-D)(p^\star+p_\theta)}{2-\int D(p^\star+p_\theta)\mathrm{d}\mu},$$

that is,

$$\int D(p^\star+p_\theta)\mathrm{d}\mu = \frac{D(p^\star+p_\theta)}{p^\star} \quad \text{and} \quad 1-D = \frac{p_\theta}{p^\star+p_\theta}\Big(2 - \int D(p^\star+p_\theta)\mathrm{d}\mu\Big).$$

We conclude that

$$1-D = \frac{p_\theta}{p^\star+p_\theta}\Big(2 - \frac{D(p^\star+p_\theta)}{p^\star}\Big),$$

i.e., $D = \frac{p^\star}{p^\star+p_\theta}$ whenever $p^\star \neq p_\theta$.

To complete the proof, it remains to show that $D = 1/2$ μ-almost everywhere on the set $A \overset{\text{def}}{=} \{p_\theta = p^\star\}$. Using the result above together with equality (3), we see that

$$\int_A \ln(D)p^\star \mathrm{d}\mu + \int_A \ln(1-D)p_\theta \mathrm{d}\mu = \int_A \ln(1/2)p^\star \mathrm{d}\mu + \int_A \ln(1/2)p_\theta \mathrm{d}\mu,$$

that is,

$$\int_A \big[\ln(1/4) - \ln(D(1-D))\big]p_\theta \mathrm{d}\mu = 0.$$

Observing that $D(1-D) \leq 1/4$ since D takes values in $[0,1]$, we deduce that $[\ln(1/4) - \ln(D(1-D))]p_\theta \mathbf{1}_A = 0$ μ-almost everywhere. Therefore, $D = 1/2$ on the set $\{p_\theta = p^\star\}$, since $p_\theta > 0$ μ-almost everywhere by assumption. □

By definition of the optimal discriminator $D_\theta^\star$, we have

$$L(\theta, D_\theta^\star) = \sup_{D \in \mathscr{D}_\infty} L(\theta, D) = 2D_{\mathrm{JS}}(p^\star, p_\theta) - \ln 4, \quad \forall \theta \in \Theta.$$

Therefore, it makes sense to let the parameter $\theta^\star \in \Theta$ be defined as

$$L(\theta^\star, D_{\theta^\star}^\star) \leq L(\theta, D_\theta^\star), \quad \forall \theta \in \Theta,$$

or, equivalently,

$$D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \leq D_{\mathrm{JS}}(p^\star, p_\theta), \quad \forall \theta \in \Theta. \tag{5}$$

The parameter $\theta^\star$ may be interpreted as the best parameter in Θ for approaching the unknown density $p^\star$ in terms of Jensen-Shannon divergence, in a context where all possible discriminators are available. In other words, the generator $G_{\theta^\star}$ is the ideal generator, and the density $p_{\theta^\star}$ is the one we would ideally like to use to generate fake samples. Of course, whenever $p^\star \in \mathscr{P}$ (i.e., the target density is in the model), then $p^\star = p_{\theta^\star}$, $D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) = 0$, and $D_{\theta^\star}^\star = 1/2$. This is, however, a very special case, which is of no interest, since in the applications covered by GANs, the data are usually so complex that the hypothesis $p^\star \in \mathscr{P}$ does not hold.

In the general case, our next theorem provides sufficient conditions for the existence and unicity of $\theta^\star$. For P and Q probability measures on E, we let $\delta(P,Q) = \sqrt{D_{\mathrm{JS}}(P,Q)}$, and recall that δ is a distance on the set of probability measures on E (Endres and Schindelin, 2003). We let $\mathrm{d}P^\star = p^\star \mathrm{d}\mu$ and, for all $\theta \in \Theta$, $\mathrm{d}P_\theta = p_\theta \mathrm{d}\mu$.

Theorem 2.2. *Assume that the model $\{P_\theta\}_{\theta\in\Theta}$ is identifiable, convex, and compact for the metric δ. Assume, in addition, that there exist $0 < m \leq M$ such that $m \leq p^\star \leq M$ and, for all $\theta \in \Theta$, $p_\theta \leq M$. Then there exists a unique $\theta^\star \in \Theta$ such that*

$$\{\theta^\star\} = \arg\min_{\theta\in\Theta} L(\theta, D_\theta^\star),$$

or, equivalently,

$$\{\theta^\star\} = \arg\min_{\theta\in\Theta} D_{\mathrm{JS}}(p^\star, p_\theta).$$

Proof. Observe that $\mathscr{P} = \{p_\theta\}_{\theta\in\Theta} \subset L^1(\mu) \cap L^2(\mu)$ since $0 \leq p_\theta \leq M$ and $\int p_\theta \mathrm{d}\mu = 1$. Recall that $L(\theta, D_\theta^\star) = \sup_{D\in\mathscr{D}_\infty} L(\theta, D) = 2D_{\mathrm{JS}}(p^\star, p_\theta) - \ln 4$. By identifiability of $\{P_\theta\}_{\theta\in\Theta}$, it is enough to prove that there exists a unique density $p_{\theta^\star}$ of $\mathscr{P}$ such that

$$\{p_{\theta^\star}\} = \arg\min_{p\in\mathscr{P}} D_{\mathrm{JS}}(p^\star, p).$$

Existence. Since $\{P_\theta\}_{\theta\in\Theta}$ is compact for δ, it is enough to show that the function

$$\begin{array}{ccc} \{P_\theta\}_{\theta\in\Theta} & \to & \mathbb{R}_+ \\ P & \mapsto & D_{\mathrm{JS}}(P^\star, P) \end{array}$$

is continuous. But this is clear since, for all $P_1, P_2 \in \{P_\theta\}_{\theta\in\Theta}$, $|\delta(P^\star, P_1) - \delta(P^\star, P_2)| \leq \delta(P_1, P_2)$ by the triangle inequality. Therefore, $\arg\min_{p\in\mathscr{P}} D_{\mathrm{JS}}(p^\star, p) \neq \emptyset$.

Unicity. For $a \in [m,M]$, we consider the function F_a defined by

$$F_a(x) = a\ln\Big(\frac{2a}{a+x}\Big) + x\ln\Big(\frac{2x}{a+x}\Big), \quad x \in [0,M],$$

with the convention $0\ln 0 = 0$. Clearly, $F_a''(x) = \frac{a}{x(a+x)} \geq \frac{m}{2M^2}$, which shows that F_a is β-strongly convex, with $\beta > 0$ independent of a. Thus, for all $\lambda \in [0,1]$, all $x_1, x_2 \in [0,M]$, and $a \in [m,M]$,

$$F_a(\lambda x_1 + (1-\lambda)x_2) \leq \lambda F_a(x_1) + (1-\lambda)F_a(x_2) - \frac{\beta}{2}\lambda(1-\lambda)(x_1-x_2)^2.$$

Thus, for all $p_1, p_2 \in \mathscr{P}$ with $p_1 \neq p_2$, and for all $\lambda \in (0,1)$,

$$\begin{aligned}
&D_{\mathrm{JS}}(p^\star, \lambda p_1 + (1-\lambda)p_2) \\
&\quad = \int F_{p^\star}(\lambda p_1 + (1-\lambda)p_2)\mathrm{d}\mu \\
&\quad \leq \lambda D_{\mathrm{JS}}(p^\star, p_1) + (1-\lambda)D_{\mathrm{JS}}(p^\star, p_2) - \frac{\beta}{2}\lambda(1-\lambda)\int (p_1-p_2)^2\mathrm{d}\mu \\
&\quad < \lambda D_{\mathrm{JS}}(p^\star, p_1) + (1-\lambda)D_{\mathrm{JS}}(p^\star, p_2).
\end{aligned}$$

In the last inequality, we used the fact that $\frac{\beta}{2}\lambda(1-\lambda)\int (p_1-p_2)^2\mathrm{d}\mu$ is positive and finite since $p_\theta \in L^2(\mu)$ for all θ. We conclude that the function $L^1(\mu) \supset \mathscr{P} \ni p \mapsto D_{\mathrm{JS}}(p^\star, p)$ is strictly convex. Therefore, its arg min is either the empty set or a singleton. □

Remark 2.1. *There are simple conditions for the model $\{P_\theta\}_{\theta\in\Theta}$ to be compact for the metric δ. It is for example enough to suppose that Θ is compact, $\{P_\theta\}_{\theta\in\Theta}$ is convex, and*

- *(i) For all $x \in E$, the function $\theta \mapsto p_\theta(x)$ is continuous on Θ;*
- *(ii) One has $\sup_{(\theta,\theta')\in\Theta^2} |p_\theta \ln p_{\theta'}| \in L^1(\mu)$.*

Let us quickly check that under these conditions, $\{P_\theta\}_{\theta\in\Theta}$ is compact for the metric δ. Since Θ is compact, by the sequential characterization of compact sets, it is enough to prove that if $\Theta \supset (\theta_n)_n$ converges to $\theta \in \Theta$, then $D_{\mathrm{JS}}(p_\theta, p_{\theta_n}) \to 0$. But,

$$D_{\mathrm{JS}}(p_\theta, p_{\theta_n}) = \int \Big[p_\theta \ln\Big(\frac{2p_\theta}{p_\theta + p_{\theta_n}}\Big) + p_{\theta_n}\ln\Big(\frac{2p_{\theta_n}}{p_\theta + p_{\theta_n}}\Big)\Big]\mathrm{d}\mu.$$

By the convexity of $\{P_\theta\}_{\theta\in\Theta}$, using (i) and (ii), the Lebesgue dominated convergence theorem shows that $D_{\mathrm{JS}}(p_\theta, p_{\theta_n}) \to 0$, whence the result.

Interpreting the adversarial problem in connection with the optimization program $\inf_{\theta\in\Theta} D_{\mathrm{JS}}(p^\star, p_\theta)$ is a bit misleading, because this is based on the assumption that all possible discriminators are available (and in particular the optimal discriminator $D_\theta^\star$). In the end this means assuming that we know the distribution $p^\star$, which is eventually not acceptable from a statistical perspective. In practice, the class of discriminators is always restricted to be a parametric family $\mathscr{D} = \{D_\alpha\}_{\alpha\in\Lambda}$, $\Lambda \subset \mathbb{R}^q$, and it is with this class that we

have to work. From our point of view, problem (1) is a likelihood-type problem involving two parametric families $\mathscr{G}$ and $\mathscr{D}$, which must be analyzed as such, just as we would do for a classical maximum likelihood approach. In fact, it takes no more than a moment's thought to realize that the key lies in the approximation capabilities of the discriminator class $\mathscr{D}$ with respect to the functions $D^\star_\theta$, $\theta \in \Theta$. This is the issue that we discuss in the next section.

3 Approximation properties

In the remainder of the article, we assume that $\theta^\star$ exists, keeping in mind that Theorem 2.2 provides us with precise conditions guaranteeing its existence and its unicity. As pointed out earlier, in practice only a parametric class $\mathscr{D} = \{D_\alpha\}_{\alpha\in\Lambda}$, $\Lambda \subset \mathbb{R}^q$, is available, and it is therefore logical to consider the parameter $\bar\theta \in \Theta$ defined by

$$\sup_{D\in\mathscr{D}} L(\bar\theta, D) \leq \sup_{D\in\mathscr{D}} L(\theta, D), \quad \forall \theta \in \Theta.$$

(We assume for now that $\bar\theta$ exists—sufficient conditions for this existence, relating to compactness of Θ and regularity of the model $\mathscr{P}$, will be given in the next section.) The density $p_{\bar\theta}$ is thus the best candidate to imitate $p_{\theta^\star}$, given the parametric families of generators $\mathscr{G}$ and discriminators $\mathscr{D}$. The natural question is then: is it possible to quantify the proximity between $p_{\bar\theta}$ and the ideal $p_{\theta^\star}$ via the approximation properties of the class $\mathscr{D}$? In other words, if $\mathscr{D}$ is growing, is it true that $p_{\bar\theta}$ approaches $p_{\theta^\star}$, and in the affirmative, in which sense and at which speed? Theorem 3.1 below provides a first answer to this important question, in terms of the difference $D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star})$. To state the result, we will need some assumptions.

Assumption (H_0) There exists a positive constant $\underline{t} \in (0, 1/2]$ such that

$$\min(D^\star_\theta, 1 - D^\star_\theta) \geq \underline{t}, \quad \forall \theta \in \Theta.$$

We note that this assumption implies that, for all $\theta \in \Theta$,

$$\frac{\underline{t}}{1-\underline{t}} p^\star \leq p_\theta \leq \frac{1-\underline{t}}{\underline{t}} p^\star.$$

It is a mild requirement, which implies in particular that for any θ, p_θ and $p^\star$ have the same support, independent of θ.

Let $\|\cdot\|_\infty$ be the supremum norm of functions on E. Our next condition guarantees that the parametric class $\mathscr{D}$ is rich enough to approach the discriminator $D^\star_{\bar\theta}$.

Assumption (H_ε) There exists $\varepsilon \in (0, \underline{t})$ and $D \in \mathscr{D}$, a $\bar\theta$-admissible discriminator, such that $\|D - D^\star_{\bar\theta}\|_\infty \leq \varepsilon$.

We are now equipped to state our approximation theorem. For ease of reading, its proof is postponed to Section 5.

Theorem 3.1. *Under Assumptions (H_0) and (H_ε), there exists a positive constant c (depending only upon $\underline{t}$) such that*

$$0 \leq D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \leq c\varepsilon^2. \tag{6}$$

This theorem points out that if the class $\mathscr{D}$ is rich enough to approximate the discriminator $D^\star_{\bar\theta}$ in such a way that $\|D - D^\star_{\bar\theta}\|_\infty \leq \varepsilon$ for some small ε, then replacing $D_{\mathrm{JS}}(p^\star, p_{\theta^\star})$ by $D_{\mathrm{JS}}(p^\star, p_{\bar\theta})$ has an impact which is not larger than a $\mathrm{O}(\varepsilon^2)$ factor. It shows in particular that the Jensen-Shannon divergence is a suitable criterion for the problem we are examining.

4 Statistical analysis

The data-dependent parameter $\hat\theta$ achieves the infimum of the adversarial problem (1). Practically speaking, it is this parameter that will be used in the end for producing fake data, via the associated generator $G_{\hat\theta}$. We first study in Subsection 4.1 the large sample properties of the distribution $p_{\hat\theta}$ via the criterion $D_{\mathrm{JS}}(p^\star, p_{\hat\theta})$, and then state in Subsection 4.2 the almost sure convergence and asymptotic normality of the parameter $\hat\theta$ as the sample size n tends to infinity. Throughout, the parameter sets Θ and Λ are assumed to be compact subsets of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. To simplify the analysis, we also assume that $\mu(E) < \infty$.

4.1 Asymptotic properties of $D_{\mathrm{JS}}(p^\star, p_{\hat\theta})$

As for now, we assume that we have at hand a parametric family of generators $\mathscr{G} = \{G_\theta\}_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^p$, and a parametric family of discriminators $\mathscr{D} = \{D_\alpha\}_{\alpha\in\Lambda}$, $\Lambda \subset \mathbb{R}^q$. We recall that the collection of probability densities associated with $\mathscr{G}$ is $\mathscr{P} = \{p_\theta\}_{\theta\in\Theta}$, where $G_\theta(Z) \overset{\mathscr{L}}{=} p_\theta \mathrm{d}\mu$ and Z is some low-dimensional noise random variable. In order to avoid any confusion, for a given discriminator $D = D_\alpha$ we use the notation $\hat L(\theta, \alpha)$ (respectively, $L(\theta, \alpha)$) instead of $\hat L(\theta, D)$ (respectively, $L(\theta, D)$) when useful. So,

$$\hat L(\theta, \alpha) = \sum_{i=1}^n \ln D_\alpha(X_i) + \sum_{i=1}^n \ln(1 - D_\alpha \circ G_\theta(Z_i)),$$

and

$$L(\theta, \alpha) = \int \ln(D_\alpha) p^\star \mathrm{d}\mu + \int \ln(1 - D_\alpha) p_\theta \mathrm{d}\mu.$$

We will need the following regularity assumptions:

Assumptions (H_{reg})

- (H_D) There exists $\kappa \in (0, 1/2)$ such that, for all $\alpha \in \Lambda$, $\kappa \leq D_\alpha \leq 1 - \kappa$. In addition, the function $(x, \alpha) \mapsto D_\alpha(x)$ is of class C^1, with a uniformly bounded differential.
- (H_G) For all $z \in \mathbb{R}^{d'}$, the function $\theta \mapsto G_\theta(z)$ is of class C^1, uniformly bounded, with a uniformly bounded differential.
- (H_p) For all $x \in E$, the function $\theta \mapsto p_\theta(x)$ is of class C^1, uniformly bounded, with a uniformly bounded differential.

Note that under (H_D), all discriminators in $\{D_\alpha\}_{\alpha\in\Lambda}$ are θ-admissible, whatever θ. All of these requirements are classic regularity conditions for statistical models, which imply in particular that the functions $\hat L(\theta, \alpha)$ and $L(\theta, \alpha)$ are continuous. Therefore, the compactness

of Θ guarantees that $\hat{\theta}$ and $\bar{\theta}$ exists. Conditions for the existence of $\theta^\star$ are given in Theorem 2.2.

We have known since Theorem 3.1 that if the available class of discriminators $\mathscr{D}$ approaches the optimal discriminator $D^\star_{\bar{\theta}}$ by a distance not more than ε, then $D_{\mathrm{JS}}(p^\star, p_{\bar{\theta}}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) = \mathrm{O}(\varepsilon^2)$. It is therefore reasonable to expect that, asymptotically, the difference $D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star})$ will not be larger than a term proportional to ε^2, in some probabilistic sense. This is precisely the result of Theorem 4.1 below. In fact, most articles to date have focused on the development and analysis of optimization procedures (typically, stochastic-gradient-type algorithms) to compute $\hat{\theta}$, without really questioning its convergence properties as the data set grows. Although our statistical results are theoretical in nature, we believe that they are complementary to the optimization literature, insofar as they offer guarantees on the validity of the algorithms.

In addition to the regularity hypotheses and Assumption (H_0), we will need the following requirement, which is a stronger version of (H_ε):

Assumption (H'_ε) There exists $\varepsilon \in (0, \underline{t})$ such that: for all $\theta \in \Theta$, there exists $D \in \mathscr{D}$, a θ-admissible discriminator, such that $\|D - D^\star_\theta\|_\infty \leq \varepsilon$.

We are ready to state our first statistical theorem.

Theorem 4.1. *Under Assumptions* (H_0), (H_{reg}), *and* (H'_ε), *one has*

$$\mathbb{E} D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) = \mathrm{O}\Big(\varepsilon^2 + \frac{1}{\sqrt{n}}\Big).$$

Proof. Fix $\varepsilon \in (0, \underline{t})$ as in Assumption (H'_ε), and choose $\hat{D} \in \mathscr{D}$, a $\hat{\theta}$-admissible discriminator, such that $\|\hat{D} - D^\star_{\hat{\theta}}\|_\infty \leq \varepsilon$. By repeating the arguments of the proof of Theorem 3.1 (with $\hat{\theta}$ instead of $\bar{\theta}$), we conclude that there exists a constant $c_1 > 0$ such that

$$2 D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) \leq c_1 \varepsilon^2 + L(\hat{\theta}, \hat{D}) + \ln 4 \leq c_1 \varepsilon^2 + \sup_{\alpha \in \Lambda} L(\hat{\theta}, \alpha) + \ln 4.$$

Therefore,

$$\begin{aligned}
2 D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) &\leq c_1 \varepsilon^2 + \sup_{\theta \in \Theta, \alpha \in \Lambda} |\hat{L}(\theta, \alpha) - L(\theta, \alpha)| + \sup_{\alpha \in \Lambda} \hat{L}(\hat{\theta}, \alpha) + \ln 4 \\
&= c_1 \varepsilon^2 + \sup_{\theta \in \Theta, \alpha \in \Lambda} |\hat{L}(\theta, \alpha) - L(\theta, \alpha)| + \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} \hat{L}(\theta, \alpha) + \ln 4 \\
&\quad \text{(by definition of } \hat{\theta}) \\
&\leq c_1 \varepsilon^2 + 2 \sup_{\theta \in \Theta, \alpha \in \Lambda} |\hat{L}(\theta, \alpha) - L(\theta, \alpha)| + \inf_{\theta \in \Theta} \sup_{\alpha \in \Lambda} L(\theta, \alpha) + \ln 4.
\end{aligned}$$

So,

$$\begin{aligned}
2D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) &\le c_1\varepsilon^2 + 2\sup_{\theta\in\Theta,\alpha\in\Lambda}|\hat{L}(\theta,\alpha) - L(\theta,\alpha)| + \inf_{\theta\in\Theta}\sup_{D\in\mathscr{D}_\infty} L(\theta, D) + \ln 4 \\
&= c_1\varepsilon^2 + 2\sup_{\theta\in\Theta,\alpha\in\Lambda}|\hat{L}(\theta,\alpha) - L(\theta,\alpha)| + L(\theta^\star, D^\star_{\theta^\star}) + \ln 4 \\
&\quad \text{(by definition of } \theta^\star\text{)} \\
&= c_1\varepsilon^2 + 2D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) + 2\sup_{\theta\in\Theta,\alpha\in\Lambda}|\hat{L}(\theta,\alpha) - L(\theta,\alpha)|.
\end{aligned}$$

Thus, letting $c_2 = c_1/2$, we have

$$D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \le c_2\varepsilon^2 + \sup_{\theta\in\Theta,\alpha\in\Lambda}|\hat{L}(\theta,\alpha) - L(\theta,\alpha)|. \tag{7}$$

Clearly, under Assumptions (H_D), (H_G), and (H_p), the process $(\hat{L}(\theta,\alpha) - L(\theta,\alpha))_{\theta\in\Theta,\alpha\in\Lambda}$ is subgaussian (e.g., van Handel, 2016, Chapter 5) for the distance $d = \|\cdot\|/\sqrt{n}$, where $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^p \times \mathbb{R}^q$. Let $N(\Theta\times\Lambda, \|\cdot\|, u)$ denote the u-covering number of $\Theta\times\Lambda$ for the distance $\|\cdot\|$. Then, by Dudley's inequality (van Handel, 2016, Corollary 5.25),

$$\mathbb{E}\sup_{\theta\in\Theta,\alpha\in\Lambda}|\hat{L}(\theta,\alpha) - L(\theta,\alpha)| \le \frac{12}{\sqrt{n}}\int_0^\infty \sqrt{\ln(N(\Theta\times\Lambda, \|\cdot\|, u))}\,\mathrm{d}u. \tag{8}$$

Since Θ and Λ are bounded, there exists $r > 0$ such that $N(\Theta\times\Lambda, \|\cdot\|, u) = 1$ for $u \ge r$ and

$$N(\Theta\times\Lambda, \|\cdot\|, u) = \mathrm{O}\left(\left(\frac{1}{u}\right)^{p+q}\right) \quad \text{for } u < r.$$

Combining this inequality with (7) and (8), we obtain

$$\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}) - D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \le c_3\left(\varepsilon^2 + \frac{1}{\sqrt{n}}\right),$$

for some positive constant c_3. The conclusion follows by observing that, by (5),

$$D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \le D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}}).$$

□

Theorem 4.1 is illustrated in Figure 1, which shows the approximate values of $\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$. We took $p^\star(x) = \frac{e^{-x/s}}{s(1+e^{-x/s})^2}$ (centered logistic density with scale parameter $s = 0.33$), and let $\mathscr{G}$ and $\mathscr{D}$ be two fully connected neural networks parameterized by weights and offsets. The noise random variable Z follows a uniform distribution on $[0,1]$, and the parameters of $\mathscr{G}$ and $\mathscr{D}$ are chosen in a sufficiently large compact set. In order to illustrate the impact of ε in Theorem 4.1, we fixed the sample size to a large $n = 100000$ and varied the number of layers of the discriminators from 2 to 5, keeping in mind that a larger number of layers results in a smaller ε. To diversify the setting, we also varied the number of layers of the

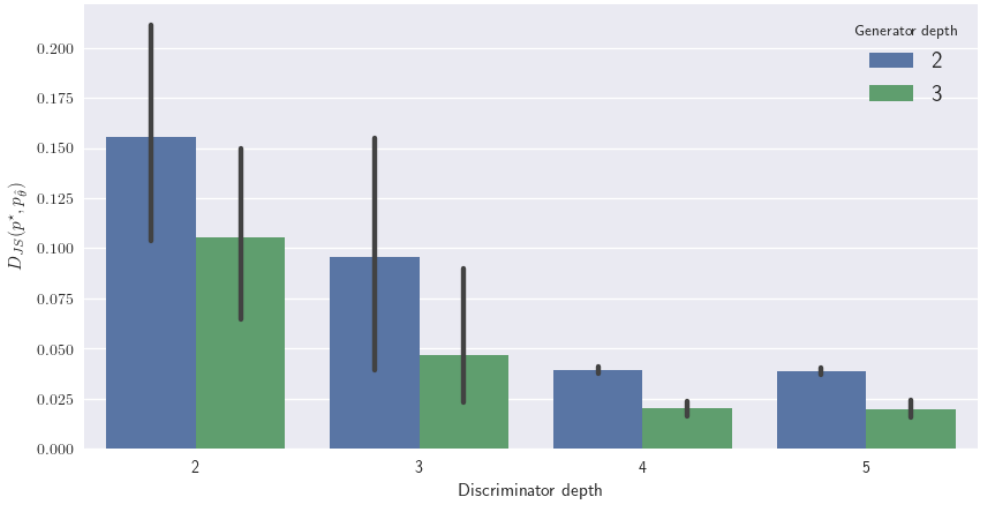
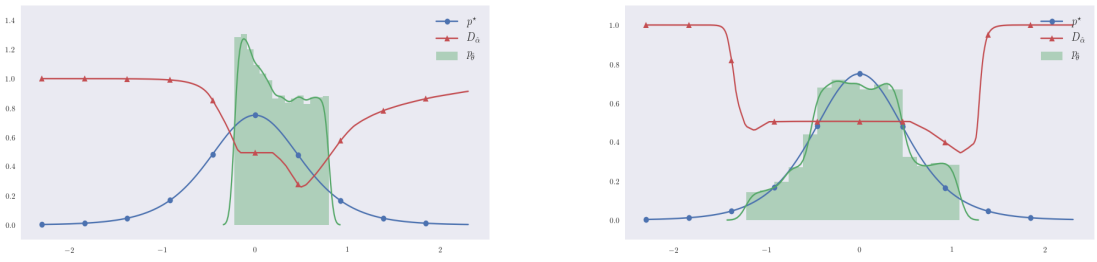


Figure 1: Bar plots of the Jensen-Shannon divergence $D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$ with respect to the number of layers (depth) of both the discriminators and generators. The height of each rectangle estimates $\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$.

generators from 2 to 3. The expectation $\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$ was estimated by averaging over 30 repetitions (the number of runs has been reduced for time complexity limitations). Note that we do not pay attention to the exact value of the constant term $D_{\mathrm{JS}}(p^\star, p_{\theta^\star})$, which is intractable in our setting.

Figure 1 highlights that $\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$ approaches the constant value $D_{\mathrm{JS}}(p^\star, p_{\theta^\star})$ as $\varepsilon \downarrow 0$, i.e., as the discriminator depth increases, given that the contribution of $1/\sqrt{n}$ is certainly negligible for $n = 100\,000$. Figure 2 shows the target density $p^\star$ vs. the histograms and kernel estimates of 100 000 data sampled from $G_{\hat{\theta}}(Z)$, in the two cases: (discriminator depth = 2, generator depth = 3) and (discriminator depth = 5, generator depth = 3). In accordance with the decrease of $\mathbb{E}D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$, the estimation of the true distribution $p^\star$ improves when ε becomes small.



(a) Discriminator depth = 2, generator depth = 3.

(b) Discriminator depth = 5, generator depth = 3.

Figure 2: True density $p^\star$, histograms, and kernel estimates (continuous line) of 100 000 data sampled from $G_{\hat{\theta}}(Z)$. Also shown is the final discriminator $D_{\hat{\alpha}}$.

Some comments on the optimization scheme. Numerical optimization is quite a tough point for GANs, partly due to nonconvex-concavity of the saddle point problem described in equation (1) and the nondifferentiability of the objective function. This motivates a very active line of research (e.g., Goodfellow et al., 2014; Nowozin et al., 2016; Arjovsky et al., 2017; Arjovsky and Bottou, 2017), which aims at transforming the objective into a more convenient function and devising efficient algorithms. In the present paper, since we are interested in original GANs, the algorithmic approach described by Goodfellow et al. (2014) is adopted, and numerical optimization is performed thanks to the machine learning framework `TensorFlow`, working with gradient descent based on automatic differentiation. As proposed by Goodfellow et al. (2014), the objective function $\theta \mapsto \sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha)$ is not directly minimized. We used instead an alternated procedure, which consists in iterating (a few hundred times in our examples) the following two steps:

(i) For a fixed value of θ and from a given value of α, perform 10 ascent steps on $\hat{L}(\theta,\cdot)$;

(ii) For a fixed value of α and from a given value of θ, perform 1 descent step on $\theta \mapsto -\sum_{i=1}^{n} \ln(D_\alpha \circ G_\theta(Z_i))$ (instead of $\theta \mapsto \sum_{i=1}^{n} \ln(1 - D_\alpha \circ G_\theta(Z_i))$).

This alternated procedure is motivated by two reasons. First, for a given θ, approximating $\sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha)$ is computationally prohibitive and may result in overfitting the finite training sample (Goodfellow et al., 2014). This can be explained by the shape of the function $\theta \mapsto \sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha)$, which may be almost piecewise constant, resulting in a zero gradient almost everywhere (or at best very low; see Arjovsky et al., 2017). Next, empirically, $-\ln(D_\alpha \circ G_\theta(Z_i))$ provides bigger gradients than $\ln(1 - D_\alpha \circ G_\theta(Z_i))$, resulting in a more powerful algorithm than the original version, while leading to the same minimizers.

In all our experiments, the learning rates needed in gradient steps were fixed and tuned by hand, in order to prevent divergence. In addition, since our main objective is to focus on illustrating the statistical properties of GANs rather than delving into optimization issues, we decided to perform mini-batch gradient updates instead of stochastic ones (that is, new observations of X and Z are not sampled at each step of the procedure). This is different of what is done in the original algorithm of Goodfellow et al. (2014).

We realize that our numerical approach—although widely adopted by the machine learning community—may fail to locate the desired estimator $\hat{\theta}$ (i.e., the exact minimizer in θ of $\sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha)$) in more complex contexts than those presented in the present paper. It is nevertheless sufficient for our objective, which is limited to illustrating the theoretical results with a few simple examples.

4.2 Asymptotic properties of $\hat{\theta}$

Theorem 4.1 states a result relative to the criterion $D_{\mathrm{JS}}(p^\star, p_{\hat{\theta}})$. We now examine the convergence properties of the parameter $\hat{\theta}$ itself as the sample size n grows. We would typically like to find reasonable conditions ensuring that $\hat{\theta} \to \bar{\theta}$ almost surely as $n \to \infty$. To reach this goal, we first need to strengthen a bit the Assumptions (H_{reg}), as follows:

Assumptions (H'_{reg})

(H'_D) There exists $\kappa \in (0,1/2)$ such that, for all $\alpha \in \Lambda$, $\kappa \leq D_\alpha \leq 1-\kappa$. In addition, the function $(x,\alpha) \mapsto D_\alpha(x)$ is of class C^2, with differentials of order 1 and 2 uniformly bounded.

(H'_G) For all $z \in \mathbb{R}^{d'}$, the function $\theta \mapsto G_\theta(z)$ is of class C^2, uniformly bounded, with differentials of order 1 and 2 uniformly bounded.

(H'_p) For all $x \in E$, the function $\theta \mapsto p_\theta(x)$ is of class C^2, uniformly bounded, with differentials of order 1 and 2 uniformly bounded.

It is easy to verify that under these assumptions the partial functions $\theta \mapsto \hat{L}(\theta,\alpha)$ (respectively, $\theta \mapsto L(\theta,\alpha)$) and $\alpha \mapsto \hat{L}(\theta,\alpha)$ (respectively, $\alpha \mapsto L(\theta,\alpha)$) are of class C^2. Throughout, we let $\theta = (\theta_1,\ldots,\theta_p)$, $\alpha = (\alpha_1,\ldots,\alpha_q)$, and denote by $\frac{\partial}{\partial\theta_i}$ and $\frac{\partial}{\partial\alpha_j}$ the partial derivative operations with respect to θ_i and α_j. The next lemma will be of constant utility. In order not to burden the text, its proof is given in Section 5.

Lemma 4.1. *Under Assumptions* (H'_{reg}), $\forall (a,b,c,d) \in \{0,1,2\}^4$ *such that* $a+b \leq 2$ *and* $c+d \leq 2$, *one has*

$$\sup_{\theta\in\Theta,\alpha\in\Lambda} \left| \frac{\partial^{a+b+c+d}}{\partial\theta_i^a \partial\theta_j^b \partial\alpha_\ell^c \partial\alpha_m^d} \hat{L}(\theta,\alpha) - \frac{\partial^{a+b+c+d}}{\partial\theta_i^a \partial\theta_j^b \partial\alpha_\ell^c \partial\alpha_m^d} L(\theta,\alpha) \right| \to 0 \quad \textit{almost surely},$$

for all $(i,j) \in \{1,\ldots,p\}^2$ *and* $(\ell,m) \in \{1,\ldots,q\}^2$.

We recall that $\bar{\theta} \in \Theta$ is such that

$$\sup_{\alpha\in\Lambda} L(\bar{\theta},\alpha) \leq \sup_{\alpha\in\Lambda} L(\theta,\alpha), \quad \forall \theta \in \Theta,$$

and insist that $\bar{\theta}$ exists under (H'_{reg}) by continuity of the function $\theta \mapsto \sup_{\alpha\in\Lambda} L(\theta,\alpha)$. Similarly, there exists $\bar{\alpha} \in \Lambda$ such that

$$L(\bar{\theta},\bar{\alpha}) \geq L(\bar{\theta},\alpha), \quad \forall \alpha \in \Lambda.$$

The following assumption ensures that $\bar{\theta}$ and $\bar{\alpha}$ are uniquely defined, which is of course a key hypothesis for our estimation objective. Throughout, the notation S° (respectively, ∂S) stands for the interior (respectively, the boundary) of the set S.

Assumption (H_1) The pair $(\bar{\theta},\bar{\alpha})$ is unique and belongs to $\Theta^\circ \times \Lambda^\circ$.

Finally, in addition to $\hat{\theta}$, we let $\hat{\alpha} \in \Lambda$ be such that

$$\hat{L}(\hat{\theta},\hat{\alpha}) \geq \hat{L}(\hat{\theta},\alpha), \quad \forall \alpha \in \Lambda.$$

Theorem 4.2. *Under Assumptions* (H'_{reg}) *and* (H_1), *one has*

$$\hat{\theta} \to \bar{\theta} \quad \textit{almost surely} \quad \textit{and} \quad \hat{\alpha} \to \bar{\alpha} \quad \textit{almost surely}.$$

Proof. We write

$$
\begin{aligned}
&|\sup_{\alpha\in\Lambda} L(\hat{\theta},\alpha) - \sup_{\alpha\in\Lambda} L(\bar{\theta},\alpha)| \\
&\quad\le |\sup_{\alpha\in\Lambda} L(\hat{\theta},\alpha) - \sup_{\alpha\in\Lambda} \hat{L}(\hat{\theta},\alpha)| + |\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha) - \inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda} L(\theta,\alpha)| \\
&\quad\le 2 \sup_{\theta\in\Theta,\alpha\in\Lambda} |\hat{L}(\theta,\alpha) - L(\theta,\alpha)|.
\end{aligned}
$$

Thus, by Lemma 4.1, $\sup_{\alpha\in\Lambda} L(\hat{\theta},\alpha) \to \sup_{\alpha\in\Lambda} L(\bar{\theta},\alpha)$ almost surely. In the lines that follow, we make more transparent the dependence of $\hat{\theta}$ in the sample size n and set $\hat{\theta}_n \overset{\text{def}}{=} \hat{\theta}$. Since $\hat{\theta}_n \in \Theta$ and Θ is compact, we can extract from any subsequence of $(\hat{\theta}_n)_n$ a subsequence $(\hat{\theta}_{n_k})_k$ such that $\hat{\theta}_{n_k} \to z \in \Theta$ (with $n_k = n_k(\omega)$, i.e., it is almost surely defined). By continuity of the function $\theta \mapsto \sup_{\alpha\in\Lambda} L(\theta,\alpha)$, we deduce that $\sup_{\alpha\in\Lambda} L(\hat{\theta}_{n_k},\alpha) \to \sup_{\alpha\in\Lambda} L(z,\alpha)$, and so $\sup_{\alpha\in\Lambda} L(z,\alpha) = \sup_{\alpha\in\Lambda} L(\bar{\theta},\alpha)$. Since $\bar{\theta}$ is unique by (H_1), we have $z = \bar{\theta}$. In conclusion, we can extract from each subsequence of $(\hat{\theta}_n)_n$ a subsequence that converges towards $\bar{\theta}$: this shows that $\hat{\theta}_n \to \bar{\theta}$ almost surely.

Finally, we have

$$
\begin{aligned}
&|L(\bar{\theta},\hat{\alpha}) - L(\bar{\theta},\bar{\alpha})| \\
&\quad\le |L(\bar{\theta},\hat{\alpha}) - L(\hat{\theta},\hat{\alpha})| + |L(\hat{\theta},\hat{\alpha}) - \hat{L}(\hat{\theta},\hat{\alpha})| + |\hat{L}(\hat{\theta},\hat{\alpha}) - L(\bar{\theta},\bar{\alpha})| \\
&\quad= |L(\bar{\theta},\hat{\alpha}) - L(\hat{\theta},\hat{\alpha})| + |L(\hat{\theta},\hat{\alpha}) - \hat{L}(\hat{\theta},\hat{\alpha})| + |\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha) - \inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda} L(\theta,\alpha)| \\
&\quad\le \sup_{\alpha\in\Lambda} |L(\bar{\theta},\alpha) - L(\hat{\theta},\alpha)| + 2 \sup_{\theta\in\Theta,\alpha\in\Lambda} |\hat{L}(\theta,\alpha) - L(\theta,\alpha)|.
\end{aligned}
$$

Using Assumptions (H'_D) and (H'_p), and the fact that $\hat{\theta} \to \bar{\theta}$ almost surely, we see that the first term above tends to zero. The second one vanishes asymptotically by Lemma 4.1, and we conclude that $L(\bar{\theta},\hat{\alpha}) \to L(\bar{\theta},\bar{\alpha})$ almost surely. Since $\hat{\alpha} \in \Lambda$ and Λ is compact, we may argue as in the first part of the proof and deduce from the unicity of $\bar{\alpha}$ that $\hat{\alpha} \to \bar{\alpha}$ almost surely. □

To illustrate the result of Theorem 4.2, we undertook a series of small numerical experiments with three choices for the triplet (true $p^\star$ + generator model $\mathscr{P} = \{p_\theta\}_{\theta\in\Theta}$ + discriminator family $\mathscr{D} = \{D_\alpha\}_{\alpha\in\Lambda}$), which we respectively call the **Laplace-Gaussian**, **Claw-Gaussian**, and **Exponential-Uniform** model. They are summarized in Table 1. We are aware that more elaborate models (involving, for example, neural networks) can be designed and implemented. However, once again, our objective is not to conduct a series of extensive simulations, but simply to illustrate our theoretical results with a few graphs to get some better intuition.

Figure 3 shows the densities $p^\star$. We recall that the claw density on $[0,\infty)$ takes the form

$$
p_{\text{claw}} = \frac{1}{2}\varphi(0,1) + \frac{1}{10}\big(\varphi(-1,0.1) + \varphi(-0.5,0.1) + \varphi(0,0.1) + \varphi(0.5,0.1) + \varphi(1,0.1)\big),
$$

where $\varphi(\mu,\sigma)$ is a Gaussian density with mean μ and standard deviation σ (this density is borrowed from Devroye, 1997).

Model	$p^\star$	$\mathscr{P} = \{p_\theta\}_{\theta\in\Theta}$	$\mathscr{D} = \{D_\alpha\}_{\alpha\in\Lambda}$
Laplace-Gaussian	$\frac{1}{2b}e^{-\frac{\lvert x\rvert}{b}}$	$\frac{1}{\sqrt{2\pi}\theta}e^{-\frac{x^2}{2\theta^2}}$	$\frac{1}{1+\frac{\alpha_1}{\alpha_0}e^{\frac{x^2}{2}(\alpha_1^{-2}-\alpha_0^{-2})}}$
	$b = 1.5$	$\Theta = [10^{-1}, 10^3]$	$\Lambda = \Theta\times\Theta$
Claw-Gaussian	$p_{\text{claw}}(x)$	$\frac{1}{\sqrt{2\pi}\theta}e^{-\frac{x^2}{2\theta^2}}$	$\frac{1}{1+\frac{\alpha_1}{\alpha_0}e^{\frac{x^2}{2}(\alpha_1^{-2}-\alpha_0^{-2})}}$
		$\Theta = [10^{-1}, 10^3]$	$\Lambda = \Theta\times\Theta$
Exponential-Uniform	$\lambda e^{-\lambda x}$	$\frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$	$\frac{1}{1+\frac{\alpha_1}{\alpha_0}e^{\frac{x^2}{2}(\alpha_1^{-2}-\alpha_0^{-2})}}$
	$\lambda = 1$	$\Theta = [10^{-3}, 10^3]$	$\Lambda = \Theta\times\Theta$

Table 1: Triplets used in the numerical experiments.

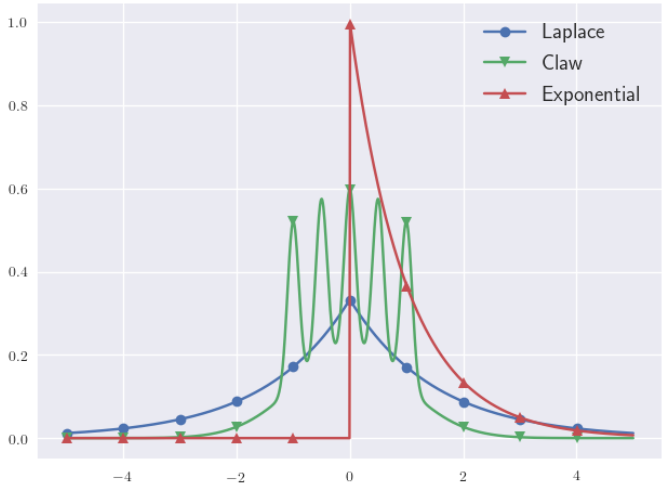


Figure 3: Probability density functions $p^\star$ used in the numerical experiments.

In the **Laplace-Gaussian** and **Claw-Gaussian** examples, the densities p_θ are centered Gaussian, parameterized by their standard deviation parameter θ. The random variable Z is uniform $[0,1]$ and the natural family of generators associated with the model $\mathscr{P} = \{p_\theta\}_{\theta\in\Theta}$ is $\mathscr{G} = \{G_\theta\}_{\theta\in\Theta}$, where each G_θ is the generalized inverse of the cumulative distribution function of p_θ (because $G_\theta(Z) \overset{\mathscr{L}}{=} p_\theta \mathrm{d}\mu$). The rationale behind our choice for the discriminators is based on the form of the optimal discriminator $D^\star_\theta$ described in (2): starting from

$$D^\star_\theta = \frac{p^\star}{p\star + p_\theta}, \quad \theta\in\Theta,$$

we logically consider the following ratio

$$D_\alpha = \frac{p_{\alpha_1}}{p_{\alpha_1} + p_{\alpha_0}}, \quad \alpha = (\alpha_0, \alpha_1) \in \Lambda = \Theta\times\Theta.$$

Figure 4 (**Laplace-Gaussian**), Figure 5 (**Claw-Gaussian**), and Figure 6 (**Exponential-Uniform**) show the boxplots of the differences $\hat{\theta} - \bar{\theta}$ over 200 repetitions, for a sample size n varying from 10 to 10000. In these experiments, the parameter $\bar{\theta}$ is obtained by averaging the $\hat{\theta}$ for the largest sample size n. In accordance with Theorem 4.2, the size of the boxplots shrinks around 0 when n increases, thus showing that the estimated parameter $\hat{\theta}$ is getting closer and closer to $\bar{\theta}$. Before analyzing at which rate this convergence occurs, we may have a look at Figure 7, which plots the estimated density $p_{\hat{\theta}}$ (for $n = 10000$) vs. the true density $p^\star$. It also shows the discriminator $D_{\hat{\alpha}}$, together with the initial density $p_{\theta_{\text{init}}}$ and the initial discriminator $D_{\alpha_{\text{init}}}$ fed into the optimization algorithm. We note that in the three models, $D_{\hat{\alpha}}$ is almost identically $1/2$, meaning that it is impossible to discriminate between the original observations and those generated by $p_{\hat{\theta}}$.

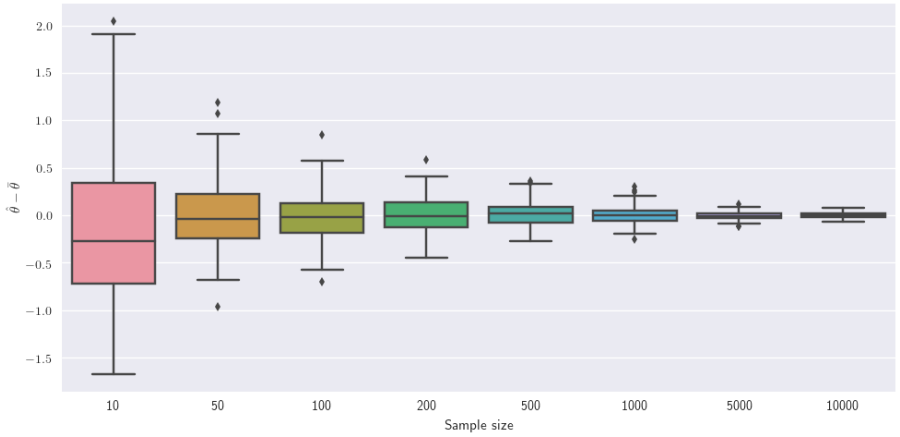


Figure 4: Boxplots of $\hat{\theta} - \bar{\theta}$ for different sample sizes (**Laplace-Gaussian** model, 200 repetitions).

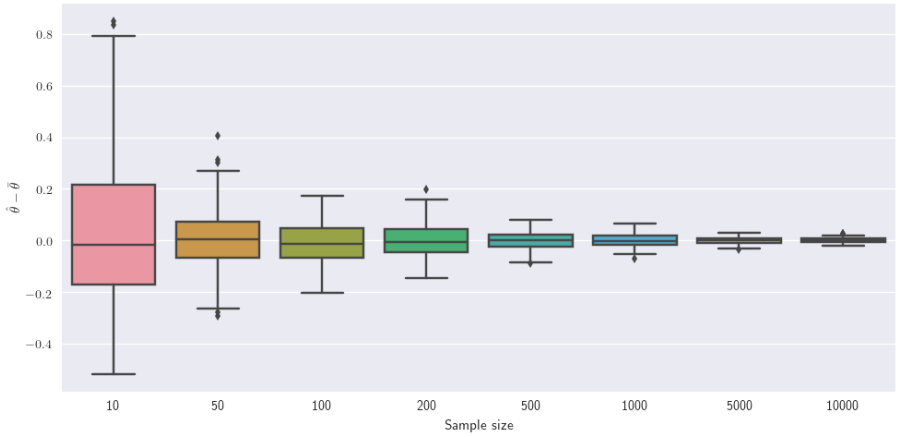


Figure 5: Boxplots of $\hat{\theta} - \bar{\theta}$ for different sample sizes (**Claw-Gaussian** model, 200 repetitions).

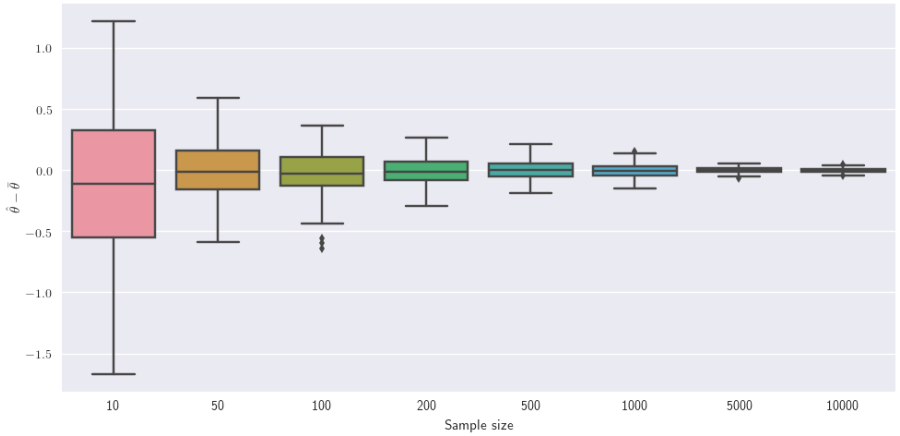


Figure 6: Boxplots of $\hat{\theta} - \bar{\theta}$ for different sample sizes (**Exponential-Uniform** model, 200 repetitions).

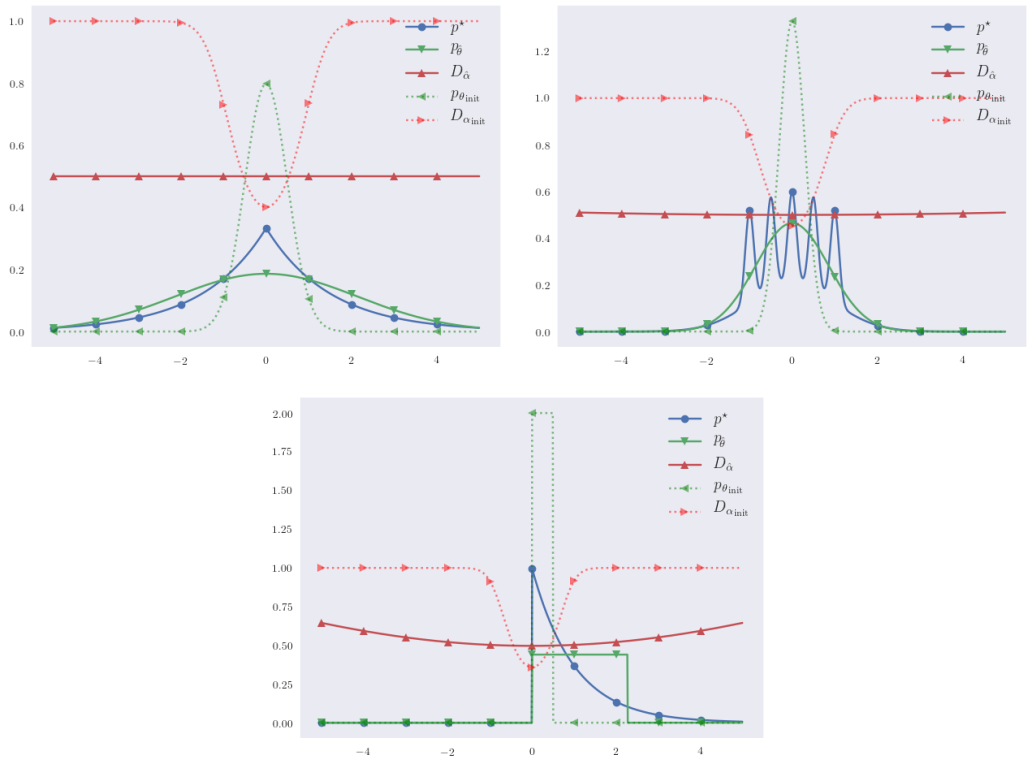


Figure 7: True density $p^\star$, estimated density $p_{\hat{\theta}}$, and discriminator $D_{\hat{\alpha}}$ for $n = 10000$ (from left to right: **Laplace-Gaussian**, **Claw-Gaussian**, and **Exponential-Uniform** model). Also shown are the initial density $p_{\theta_{\text{init}}}$ and the initial discriminator $D_{\alpha_{\text{init}}}$ fed into the optimization algorithm.

In line with the above, our next step is to state a central limit theorem for $\hat{\theta}$. Although simple to understand, this result requires additional assumptions and some technical pre-requisites. One first needs to ensure that the function $(\theta, \alpha) \mapsto L(\theta, \alpha)$ is regular enough

in a neighborhood of $(\bar{\theta}, \bar{\alpha})$. This is captured by the following set of assumptions, which require in particular the unicity of the maximizer of the function $\alpha \mapsto L(\theta, \alpha)$ for a θ around $\bar{\theta}$. For a function $F : \Theta \to \mathbb{R}$ (respectively, $G : \Theta \times \Lambda \to \mathbb{R}$), we let $HF(\theta)$ (respectively, $H_1G(\theta, \alpha)$ and $H_2G(\theta, \alpha)$) be the Hessian matrix of the function $\theta \mapsto F(\theta)$ (respectively, $\theta \mapsto G(\theta, \alpha)$ and $\alpha \mapsto G(\theta, \alpha)$) computed at θ (respectively, at θ and α).

Assumptions (H_{loc})

(H_U) There exists a neighborhood U of $\bar{\theta}$ and a function $\alpha : U \to \Lambda$ such that

$$\arg\max_{\alpha \in \Lambda} L(\theta, \alpha) = \{\alpha(\theta)\}, \quad \forall \theta \in U.$$

(H_V) The Hessian matrix $HV(\bar{\theta})$ is invertible, where $V(\theta) \stackrel{\text{def}}{=} L(\theta, \alpha(\theta))$.

(H_H) The Hessian matrix $H_2L(\bar{\theta}, \bar{\alpha})$ is invertible.

We stress that under Assumption (H_U), there is for each $\theta \in U$ a unique $\alpha(\theta) \in \Lambda$ such that $L(\theta, \alpha(\theta)) = \sup_{\alpha \in \Lambda} L(\theta, \alpha)$. We also note that $\alpha(\bar{\theta}) = \bar{\alpha}$ under (H_1). We still need some notation before we state the central limit theorem. For a function $f(\theta, \alpha)$, $\nabla_1 f(\theta, \alpha)$ (respectively, $\nabla_2 f(\theta, \alpha)$) means the gradient of the function $\theta \mapsto f(\theta, \alpha)$ (respectively, the function $\alpha \mapsto f(\theta, \alpha)$) computed at θ (respectively, at α). For a function $g(t)$, $J(g)_t$ is the Jacobian matrix of g computed at t. Observe that by the envelope theorem,

$$HV(\bar{\theta}) = H_1L(\bar{\theta}, \bar{\alpha}) + J(\nabla_1 L(\bar{\theta}, \cdot))_{\bar{\alpha}} J(\alpha)_{\bar{\theta}},$$

where, by the chain rule,

$$J(\alpha)_{\bar{\theta}} = -H_2L(\bar{\theta}, \bar{\alpha})^{-1} J(\nabla_2 L(\cdot, \bar{\alpha}))_{\bar{\theta}}.$$

Therefore, in Assumption(H_V), the Hessian matrix $HV(\bar{\theta})$ can be computed with the sole knowledge of L. Finally, we let

$$\ell_1(\theta, \alpha) = \ln D_\alpha(X_1) + \ln(1 - D_\alpha \circ G_\theta(Z_1)),$$

and denote by $\stackrel{\mathscr{L}}{\to}$ the convergence in distribution.

Theorem 4.3. *Under Assumptions* (H'_{reg}), (H_1), *and* (H_{loc}), *one has*

$$\sqrt{n}(\hat{\theta} - \bar{\theta}) \stackrel{\mathscr{L}}{\to} Z,$$

where Z is a Gaussian random variable with mean 0 *and variance*

$$\mathbf{V} = \text{Var}\big[-HV(\bar{\theta})^{-1} \nabla_1 \ell_1(\bar{\theta}, \bar{\alpha}) + HV(\bar{\theta})^{-1} J(\nabla_1 L(\bar{\theta}, \cdot))_{\bar{\alpha}} H_2L(\bar{\theta}, \bar{\alpha})^{-1} \nabla_2 \ell_1(\bar{\theta}, \bar{\alpha}) \big].$$

We note that the expression of the variance is relatively complex and, unfortunately, that it cannot be simplified, even for a dimension of the parameter equal to 1. Nevertheless, the take-home message is that the estimator $\hat{\theta}$ is asymptotically normal, with a convergence rate of $\sqrt{n}$. This is illustrated in Figures 8, 9, and 10, which respectively show the histograms and kernel estimates of the distribution of $\sqrt{n}(\hat{\theta} - \bar{\theta})$ for the **Laplace-Gaussian**, the **Claw-Gaussian**, and the **Exponential-Uniform** model in function of the sample size n (200 repetitions).

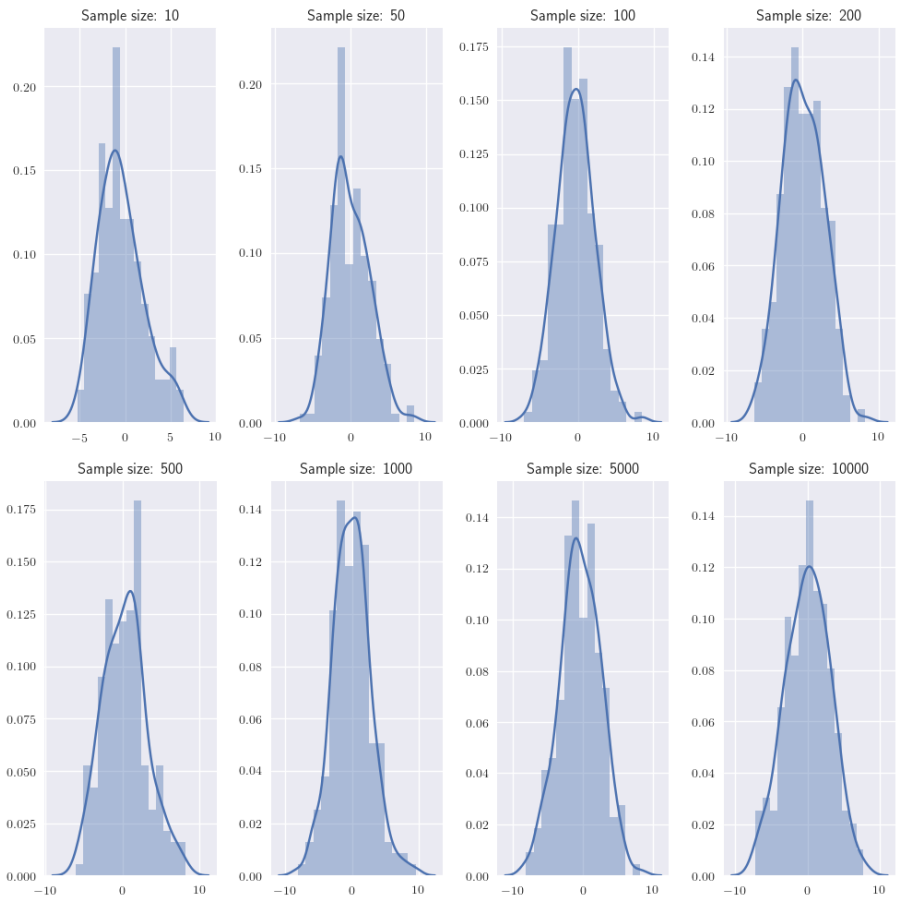


Figure 8: Histograms and kernel estimates (continuous line) of the distribution of $\sqrt{n}(\hat{\theta}-\bar{\theta})$ for different sample sizes n (**Laplace-Gaussian** model, 200 repetitions).

Proof. By technical Lemma 5.1, we can find under Assumptions (H'_{reg}) and (H_1) an open set $V \subset U \subset \Theta^\circ$ containing $\bar{\theta}$ such that, for all $\theta \in V$, $\alpha(\theta) \in \Lambda^\circ$. In the sequel, to lighten the notation, we assume without loss of generality that $V = U$. Thus, for all $\theta \in U$, we have $\alpha(\theta) \in \Lambda^\circ$ and $L(\theta,\alpha(\theta)) = \sup_{\alpha\in\Lambda} L(\theta,\alpha)$ (with $\alpha(\bar{\theta}) = \bar{\alpha}$ by (H_1)). Accordingly, $\nabla_2 L(\theta,\alpha(\theta)) = 0$, $\forall \theta \in U$. Also, since $H_2L(\bar{\theta},\bar{\alpha})$ is invertible by (H_H) and since the function $(\theta,\alpha) \mapsto H_2L(\theta,\alpha)$ is continuous, there exists an open set $U' \subset U$ such that $H_2L(\theta,\alpha)$ is invertible as soon as $(\theta,\alpha) \in (U',\alpha(U'))$. Without loss of generality, we assume that $U' = U$. Thus, by the chain rule, the function α is of class C^2 in a neighborhood $U' \subset U$ of $\bar{\theta}$, say $U' = U$, with Jacobian matrix given by

$$J(\alpha)_\theta = -H_2L(\theta,\alpha(\theta))^{-1} J\big(\nabla_2 L(\cdot,\alpha(\theta))\big)_\theta, \quad \forall \theta \in U.$$

We note that $H_2L(\theta,\alpha(\theta))^{-1}$ is of format $q \times q$ and $J(\nabla_2 L(\cdot,\alpha(\theta)))_\theta$ of format $q \times p$.

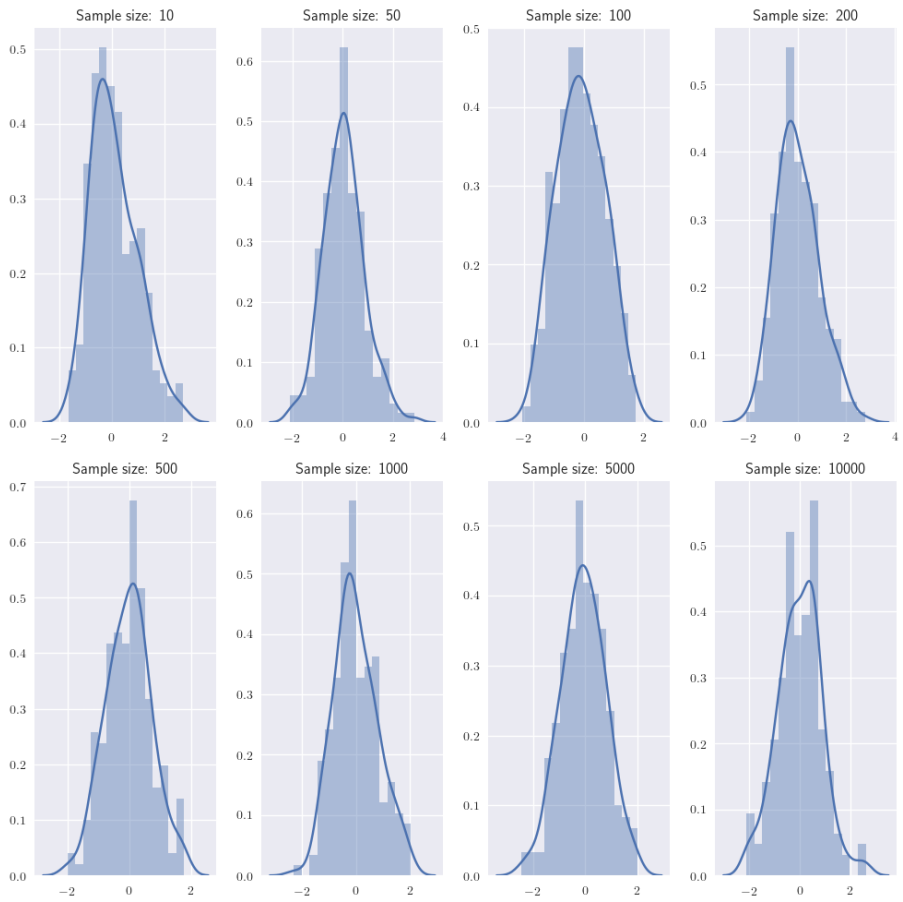


Figure 9: Histograms and kernel estimates (continuous line) of the distribution of $\sqrt{n}(\hat{\theta}-\bar{\theta})$ for different sample sizes n (**Claw-Gaussian** model, 200 repetitions).

Now, for each $\theta \in U$, we let $\hat{\alpha}(\theta)$ be such that $\hat{L}(\theta,\hat{\alpha}(\theta)) = \sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha)$. Clearly,

$$\begin{aligned}
|L(\theta,\hat{\alpha}(\theta)) - L(\theta,\alpha(\theta))| &\leq |L(\theta,\hat{\alpha}(\theta)) - \hat{L}(\theta,\hat{\alpha}(\theta))| + |\hat{L}(\theta,\hat{\alpha}(\theta)) - L(\theta,\alpha(\theta))| \\
&\leq \sup_{\alpha\in\Lambda} |L(\theta,\alpha) - \hat{L}(\theta,\alpha)| + |\sup_{\alpha\in\Lambda} \hat{L}(\theta,\alpha) - \sup_{\alpha\in\Lambda} L(\theta,\alpha)| \\
&\leq 2 \sup_{\alpha\in\Lambda} |\hat{L}(\theta,\alpha) - L(\theta,\alpha)|.
\end{aligned}$$

Therefore, by Lemma 4.1, $\sup_{\theta\in U} |L(\theta,\hat{\alpha}(\theta)) - L(\theta,\alpha(\theta))| \to 0$ almost surely. The event on which this convergence holds does not depend upon $\theta \in U$, and, arguing as in the proof of Theorem 4.2, we deduce that under (H_1), $\mathbb{P}(\hat{\alpha}(\theta) \to \alpha(\theta) \forall \theta \in U) = 1$. Since $\alpha(\theta) \in \Lambda^\circ$ for all $\theta \in U$, we also have $\mathbb{P}(\hat{\alpha}(\theta) \in \Lambda^\circ \forall \theta \in U) \to 1$ as $n \to \infty$. Thus, in the sequel, it will be assumed without loss of generality that, for all $\theta \in U$, $\hat{\alpha}(\theta) \in \Lambda^\circ$.

Still by Lemma 4.1, $\sup_{\theta\in\Theta,\alpha\in\Lambda} \|H_2\hat{L}(\theta,\alpha) - H_2 L(\theta,\alpha)\| \to 0$ almost surely. Since $H_2 L(\theta,\alpha)$ is invertible on $U \times \alpha(U)$, we have

$$\mathbb{P}\big(H_2\hat{L}(\theta,\alpha) \text{ invertible } \forall (\theta,\alpha) \in U \times \alpha(U)\big) \to 1.$$

Thus, we may and will assume that $H_2\hat{L}(\theta,\alpha)$ is invertible for all $(\theta,\alpha) \in U \times \alpha(U)$.

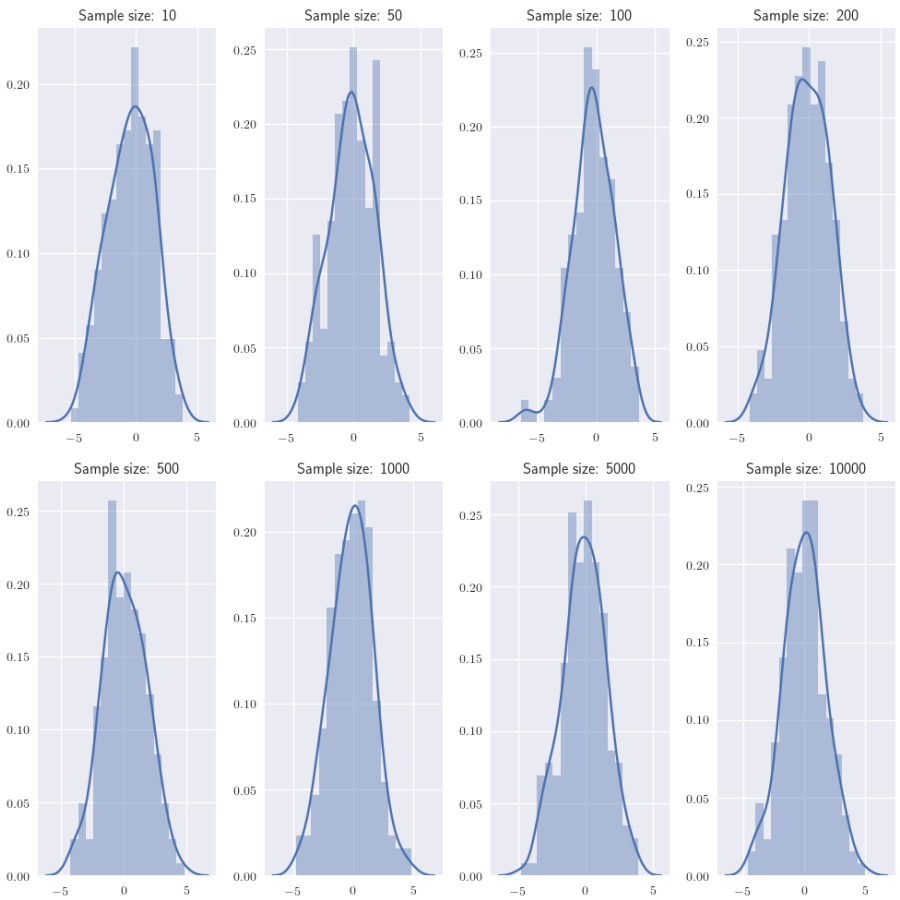


Figure 10: Histograms and kernel estimates (continuous line) of the distribution of $\sqrt{n}(\hat{\theta} - \bar{\theta})$ for different sample sizes n (**Exponential-Uniform** model, 200 repetitions).

Next, since $\hat{\alpha}(\theta) \in \Lambda^\circ$ for all $\theta \in U$, one has $\nabla_2 \hat{L}(\theta, \hat{\alpha}(\theta)) = 0$. Therefore, by the chain rule, $\hat{\alpha}$ is of class C^2 on U, with Jacobian matrix

$$J(\hat{\alpha})_\theta = -H_2\hat{L}(\theta, \hat{\alpha}(\theta))^{-1} J\big(\nabla_2 \hat{L}(\cdot, \hat{\alpha}(\theta))\big)_\theta, \quad \forall \theta \in U.$$

Let $\hat{V}(\theta) \stackrel{\text{def}}{=} \hat{L}(\theta, \hat{\alpha}(\theta)) = \sup_{\alpha \in \Lambda} \hat{L}(\theta, \alpha)$. By the envelope theorem, $\hat{V}$ is of class C^2, $\nabla \hat{V}(\theta) = \nabla_1 \hat{L}(\theta, \hat{\alpha}(\theta))$, and $H\hat{V}(\theta) = H_1\hat{L}(\theta, \hat{\alpha}(\theta)) + J(\nabla_1 \hat{L}(\theta, \cdot))_{\hat{\alpha}(\theta)} J(\hat{\alpha})_\theta$. Recall that $\hat{\theta} \to \bar{\theta}$ almost surely by Theorem 4.2, so that we may assume that $\hat{\theta} \in \Theta^\circ$ by (H_1). Moreover, we can also assume that $\hat{\theta} + t(\hat{\theta} - \bar{\theta}) \in U$, $\forall t \in [0,1]$. Thus, by a Taylor series expansion with integral remainder, we have

$$0 = \nabla \hat{V}(\hat{\theta}) = \nabla \hat{V}(\bar{\theta}) + \int_0^1 H\hat{V}(\hat{\theta} + t(\hat{\theta} - \bar{\theta}))\mathrm{d}t(\hat{\theta} - \bar{\theta}). \tag{9}$$

Since $\hat{\alpha}(\bar{\theta}) \in \Lambda^\circ$ and $\hat{L}(\bar{\theta}, \hat{\alpha}(\bar{\theta})) = \sup_{\alpha \in \Lambda} \hat{L}(\bar{\theta}, \alpha)$, one has $\nabla_2 \hat{L}(\bar{\theta}, \hat{\alpha}(\bar{\theta})) = 0$. Thus,

$$\begin{aligned} 0 &= \nabla_2 \hat{L}(\bar{\theta}, \hat{\alpha}(\bar{\theta})) \\ &= \nabla_2 \hat{L}(\bar{\theta}, \alpha(\bar{\theta})) + \int_0^1 H_2\hat{L}\big(\bar{\theta}, \alpha(\bar{\theta}) + t(\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta}))\big)\mathrm{d}t(\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta})). \end{aligned}$$

By Lemma 4.1, since $\hat{\alpha}(\bar{\theta}) \to \alpha(\bar{\theta})$ almost surely, we have

$$\hat{I}_1 \stackrel{\text{def}}{=} \int_0^1 H_2\hat{L}\big(\bar{\theta}, \alpha(\bar{\theta}) + t(\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta}))\big)\mathrm{d}t \to H_2L(\bar{\theta}, \bar{\alpha}) \quad \text{almost surely.}$$

Because $H_2L(\bar{\theta}, \bar{\alpha})$ is invertible, $\mathbb{P}(\hat{I}_1 \text{ invertible}) \to 1$ as $n \to \infty$. Therefore, we may assume, without loss of generality, that $\hat{I}_1$ is invertible. Hence,

$$\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta}) = -\hat{I}_1^{-1}\nabla_2\hat{L}(\bar{\theta}, \alpha(\bar{\theta})). \tag{10}$$

Furthermore,

$$\nabla\hat{V}(\bar{\theta}) = \nabla_1\hat{L}(\bar{\theta}, \hat{\alpha}(\bar{\theta})) = \nabla_1\hat{L}(\bar{\theta}, \alpha(\bar{\theta})) + \hat{I}_2(\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta})),$$

where

$$\hat{I}_2 \stackrel{\text{def}}{=} \int_0^1 J(\nabla_1\hat{L}(\bar{\theta}, \cdot))_{\alpha(\bar{\theta}) + t(\hat{\alpha}(\bar{\theta}) - \alpha(\bar{\theta}))}\mathrm{d}t.$$

By Lemma 4.1, $\hat{I}_2 \to J(\nabla_1 L(\bar{\theta}, \cdot))_{\alpha(\bar{\theta})}$ almost surely. Combining (9) and (10), we obtain

$$0 = \nabla_1\hat{L}(\bar{\theta}, \alpha(\bar{\theta})) - \hat{I}_2\hat{I}_1^{-1}\nabla_2\hat{L}(\bar{\theta}, \alpha(\bar{\theta})) + \hat{I}_3(\hat{\theta} - \bar{\theta}),$$

where

$$\hat{I}_3 \stackrel{\text{def}}{=} \int_0^1 H\hat{V}(\hat{\theta} + t(\hat{\theta} - \bar{\theta}))\mathrm{d}t.$$

By technical Lemma 5.2, we have $\hat{I}_3 \to HV(\bar{\theta})$ almost surely. So, by (H_V), it can be assumed that $\hat{I}_3$ is invertible. Consequently,

$$\hat{\theta} - \bar{\theta} = -\hat{I}_3^{-1}\nabla_1\hat{L}(\bar{\theta}, \alpha(\bar{\theta})) + \hat{I}_3^{-1}\hat{I}_2\hat{I}_1^{-1}\nabla_2\hat{L}(\bar{\theta}, \alpha(\bar{\theta})),$$

or, equivalently, since $\alpha(\bar{\theta}) = \bar{\alpha}$,

$$\hat{\theta} - \bar{\theta} = -\hat{I}_3^{-1}\nabla_1\hat{L}(\bar{\theta}, \bar{\alpha}) + \hat{I}_3^{-1}\hat{I}_2\hat{I}_1^{-1}\nabla_2\hat{L}(\bar{\theta}, \bar{\alpha}).$$

Using Lemma 4.1, we conclude that $\sqrt{n}(\hat{\theta} - \bar{\theta})$ has the same limit distribution as

$$S_n \stackrel{\text{def}}{=} -\sqrt{n}HV(\bar{\theta})^{-1}\nabla_1\hat{L}(\bar{\theta}, \bar{\alpha}) + \sqrt{n}HV(\bar{\theta})^{-1}J(\nabla_1 L(\bar{\theta}, \cdot))_{\bar{\alpha}}H_2L(\bar{\theta}, \bar{\alpha})^{-1}\nabla_2\hat{L}(\bar{\theta}, \bar{\alpha}).$$

Let

$$\ell_i(\theta, \alpha) = \ln D_\alpha(X_i) + \ln(1 - D_\alpha \circ G_\theta(Z_i)), \quad 1 \le i \le n.$$

With this notation, we have

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(-HV(\bar{\theta})^{-1}\nabla_1\ell_i(\bar{\theta}, \bar{\alpha}) + HV(\bar{\theta})^{-1}J(\nabla_1 L(\bar{\theta}, \cdot))_{\bar{\alpha}}H_2L(\bar{\theta}, \bar{\alpha})^{-1}\nabla_2\ell_i(\bar{\theta}, \bar{\alpha})\Big).$$

One has $\nabla V(\bar{\theta}) = 0$, since $V(\bar{\theta}) = \inf_{\theta\in\Theta} V(\theta)$ and $\bar{\theta} \in \Theta^\circ$. Therefore, under (H'_{reg}), $\mathbb{E}\nabla_1\ell_i(\bar{\theta}, \bar{\alpha}) = \nabla_1\mathbb{E}\ell_i(\bar{\theta}, \bar{\alpha}) = \nabla_1 L(\bar{\theta}, \bar{\alpha}) = \nabla V(\bar{\theta}) = 0$. Similarly, $\mathbb{E}\nabla_2\ell_i(\bar{\theta}, \bar{\alpha}) = \nabla_2\mathbb{E}\ell_i(\bar{\theta}, \bar{\alpha}) = \nabla_2 L(\bar{\theta}, \bar{\alpha}) = 0$, since $L(\bar{\theta}, \bar{\alpha}) = \sup_{\alpha\in\Lambda} L(\bar{\theta}, \alpha)$ and $\bar{\alpha} \in \Lambda^\circ$. Using the central limit theorem, we conclude that

$$\sqrt{n}(\hat{\theta} - \bar{\theta}) \stackrel{\mathscr{L}}{\to} Z,$$

where Z is a Gaussian random variable with mean 0 and variance

$$\mathbf{V} = \mathrm{Var}\big[-HV(\bar{\theta})^{-1}\nabla_1\ell_1(\bar{\theta}, \bar{\alpha}) + HV(\bar{\theta})^{-1}J(\nabla_1 L(\bar{\theta}, \cdot))_{\bar{\alpha}}H_2L(\bar{\theta}, \bar{\alpha})^{-1}\nabla_2\ell_1(\bar{\theta}, \bar{\alpha})\big].$$

□

5 Technical results

5.1 Proof of Theorem 3.1

Choose $\varepsilon \in (0, \underline{t})$ and $D \in \mathscr{D}$, a $\bar{\theta}$-admissible discriminator, such that $\|D - D^\star_{\bar{\theta}}\|_\infty \leq \varepsilon$. Observe that

$$\begin{aligned} L(\bar{\theta}, D) &= \int \ln(D) p^\star \mathrm{d}\mu + \int \ln(1-D) p_{\bar{\theta}} \mathrm{d}\mu \\ &= \int \ln\Big(\frac{D}{D^\star_{\bar{\theta}}}\Big) p^\star \mathrm{d}\mu + \ln\Big(\frac{1-D}{1-D^\star_{\bar{\theta}}}\Big) p_{\bar{\theta}} \mathrm{d}\mu + 2D_{\mathrm{JS}}(p^\star, p_{\bar{\theta}}) - \ln 4. \end{aligned} \tag{11}$$

Clearly,

$$\begin{aligned} \int \ln\Big(\frac{D}{D^\star_{\bar{\theta}}}\Big) p^\star \mathrm{d}\mu &= \int \ln\Big(1 + \Big[\frac{D}{D^\star_{\bar{\theta}}} - 1\Big]\Big) p^\star \mathrm{d}\mu \\ &= \int \ln\Big(1 + \frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}}\Big) p^\star \mathrm{d}\mu, \end{aligned}$$

where $\gamma_{\bar{\theta}} = D - D^\star_{\bar{\theta}}$. By a Taylor series expansion with remainder, we may write

$$\ln\Big(1 + \frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}}\Big) = \frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} - \frac{1}{2}\Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}}\Big)^2 + \frac{1}{3}\int_0^{\gamma_{\bar{\theta}}/D^\star_{\bar{\theta}}} \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} - u\Big)^2 \mathrm{d}u.$$

Whenever $\gamma_{\bar{\theta}} \leq 0$ (worst case), we have

$$\int_0^{\gamma_{\bar{\theta}}/D^\star_{\bar{\theta}}} \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} - u\Big)^2 \mathrm{d}u = -\int_{\gamma_{\bar{\theta}}/D^\star_{\bar{\theta}}}^0 \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} - u\Big)^2 \mathrm{d}u.$$

Observe that, for $\gamma_{\bar{\theta}}/D^\star_{\bar{\theta}} \leq u \leq 0$, since $\|\gamma_{\bar{\theta}}\|_\infty \leq \varepsilon$ by assumption and $D^\star_{\bar{\theta}} \geq \underline{t}$ by (H_0),

$$1 + u \geq 1 + \frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} \geq 1 - \frac{\varepsilon}{D^\star_{\bar{\theta}}} \geq 1 - \frac{\varepsilon}{\underline{t}} > 0.$$

Thus,

$$\int \ln\Big(\frac{D}{D^\star_{\bar{\theta}}}\Big) p^\star \mathrm{d}\mu \geq \int \Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}} - \frac{1}{2}\Big(\frac{\gamma_{\bar{\theta}}}{D^\star_{\bar{\theta}}}\Big)^2 - \frac{1}{9}\Big(\frac{|\gamma_{\bar{\theta}}|}{D^\star_{\bar{\theta}}}\Big)^3 \frac{1}{(1-\varepsilon/\underline{t})^3}\Big) p^\star \mathrm{d}\mu. \tag{12}$$

Similarly, we have

$$\begin{aligned} \int \ln\Big(\frac{1-D}{1-D^\star_{\bar{\theta}}}\Big) p_{\bar{\theta}} \mathrm{d}\mu &= \int \ln\Big(1 + \Big[\frac{1-D}{1-D^\star_{\bar{\theta}}} - 1\Big]\Big) p_{\bar{\theta}} \mathrm{d}\mu \\ &= \int \ln\Big(1 - \frac{\gamma_{\bar{\theta}}}{1-D^\star_{\bar{\theta}}}\Big) p_{\bar{\theta}} \mathrm{d}\mu. \end{aligned}$$

By a Taylor series with remainder,

$$\ln\Big(1 - \frac{\gamma_{\bar{\theta}}}{1-D^\star_{\bar{\theta}}}\Big) = -\frac{\gamma_{\bar{\theta}}}{1-D^\star_{\bar{\theta}}} - \frac{1}{2}\Big(\frac{\gamma_{\bar{\theta}}}{1-D^\star_{\bar{\theta}}}\Big)^2 + \frac{1}{3}\int_0^{-\gamma_{\bar{\theta}}/(1-D^\star_{\bar{\theta}})} \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar{\theta}}}{1-D^\star_{\bar{\theta}}} + u\Big)^2 \mathrm{d}u.$$

Whenever $\gamma_{\bar\theta} \geq 0$ (worst case), we have

$$\int_0^{-\gamma_{\bar\theta}/(1-D^\star_{\bar\theta})} \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}}+u\Big)^2 \mathrm{d}u = -\int_{-\gamma_{\bar\theta}/(1-D^\star_{\bar\theta})}^0 \frac{1}{(1+u)^3}\Big(\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}}+u\Big)^2 \mathrm{d}u.$$

But, for $-\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}} \leq u \leq 0$,

$$1+u \geq 1-\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}} \geq 1-\frac{\varepsilon}{1-D^\star_{\bar\theta}} \geq 1-\frac{\varepsilon}{\underline{t}} > 0.$$

Thus, we obtain

$$\int \ln\Big(\frac{1-D}{1-D^\star_{\bar\theta}}\Big) p_{\bar\theta}\mathrm{d}\mu \geq \int \Big(-\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}} - \frac{1}{2}\Big(\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}}\Big)^2 - \frac{1}{9}\Big(\frac{|\gamma_{\bar\theta}|}{1-D^\star_{\bar\theta}}\Big)^3 \frac{1}{(1-\varepsilon/\underline{t})^3}\Big) p_{\bar\theta}\mathrm{d}\mu. \tag{13}$$

Letting

$$\tau = \frac{1}{(1-\varepsilon/\underline{t})^3},$$

and combining (11), (12), and (13), we are led to

$$\begin{aligned}
L(\bar\theta, D) \geq & \int \Big(\frac{\gamma_{\bar\theta}}{D^\star_{\bar\theta}} - \frac{1}{2}\Big(\frac{\gamma_{\bar\theta}}{D^\star_{\bar\theta}}\Big)^2 - \frac{1}{9}\Big(\frac{|\gamma_{\bar\theta}|}{D^\star_{\bar\theta}}\Big)^3 \frac{1}{\tau}\Big) p^\star \mathrm{d}\mu \\
& + \int \Big(-\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}} - \frac{1}{2}\Big(\frac{\gamma_{\bar\theta}}{1-D^\star_{\bar\theta}}\Big)^2 - \frac{1}{9}\Big(\frac{|\gamma_{\bar\theta}|}{1-D^\star_{\bar\theta}}\Big)^3 \frac{1}{\tau}\Big) p_{\bar\theta}\mathrm{d}\mu \\
& + 2D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - \ln 4 \\
\geq & -\frac{\varepsilon^2}{2}\int \frac{p^\star}{D^{\star 2}_{\bar\theta}}\mathrm{d}\mu - \frac{\varepsilon^2}{2}\int \frac{p_{\bar\theta}}{(1-D^\star_{\bar\theta})^2}\mathrm{d}\mu - \frac{\varepsilon^3}{9\tau}\int \Big(\frac{p^\star}{D^{\star 3}_{\bar\theta}} + \frac{p_{\bar\theta}}{(1-D^\star_{\bar\theta})^3}\Big)\mathrm{d}\mu \\
& + 2D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - \ln 4 \\
= & -\frac{\varepsilon^2}{2}\Big(\int \frac{(p^\star + p_{\bar\theta})^2}{p^\star}\mathrm{d}\mu + \int \frac{(p^\star + p_{\bar\theta})^2}{p_{\bar\theta}}\mathrm{d}\mu\Big) \\
& - \frac{\varepsilon^3}{9\tau}\int \Big(\frac{(p^\star + p_{\bar\theta})^3}{p^{\star 2}} + \frac{(p^\star + p_{\bar\theta})^3}{p^2_{\bar\theta}}\Big)\mathrm{d}\mu + 2D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - \ln 4.
\end{aligned}$$

Using (H_0), we conclude that there exists a constant $c > 0$ (depending only upon $\underline{t}$) such that

$$L(\bar\theta, D) \geq -c\varepsilon^2 - \frac{c}{\tau}\varepsilon^3 + 2D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) - \ln 4,$$

i.e.,

$$2D_{\mathrm{JS}}(p^\star, p_{\bar\theta}) \leq c\varepsilon^2 + \frac{c}{\tau}\varepsilon^3 + L(\bar\theta, D) + \ln 4.$$

But

$$\begin{aligned} L(\bar{\theta}, D) &\leq \sup_{D \in \mathscr{D}} L(\bar{\theta}, D) \\ &\leq \sup_{D \in \mathscr{D}} L(\theta^\star, D) \\ &\quad \text{(by definition of } \bar{\theta}) \\ &\leq \sup_{D \in \mathscr{D}_\infty} L(\theta^\star, D) \\ &= L(\theta^\star, D^\star_{\theta^\star}) = 2D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) - \ln 4. \end{aligned}$$

Thus,

$$2D_{\mathrm{JS}}(p^\star, p_{\bar{\theta}}) \leq c\varepsilon^2 + \frac{c}{\tau}\varepsilon^3 + 2D_{\mathrm{JS}}(p^\star, p_{\theta^\star}).$$

This shows the right-hand side of inequality (6). To prove the left-hand side, just note that by inequality (5),

$$D_{\mathrm{JS}}(p^\star, p_{\theta^\star}) \leq D_{\mathrm{JS}}(p^\star, p_{\bar{\theta}}).$$

5.2 Proof of Lemma 4.1

To simplify the notation, we set

$$\Delta = \frac{\partial^{a+b+c+d}}{\partial \theta_i^a \partial \theta_j^b \partial \alpha_\ell^c \partial \alpha_m^d}.$$

Using McDiarmid's inequality (McDiarmid, 1989), we see that there exists a constant $c > 0$ such that, for all $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big| \sup_{\theta \in \Theta, \alpha \in \Lambda} |\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha)| - \mathbb{E} \sup_{\theta \in \Theta, \alpha \in \Lambda} |\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha)| \Big| \geq \varepsilon \Big) \leq 2e^{-cn\varepsilon^2}.$$

Therefore, by the Borel-Cantelli lemma,

$$\sup_{\theta \in \Theta, \alpha \in \Lambda} |\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha)| - \mathbb{E} \sup_{\theta \in \Theta, \alpha \in \Lambda} |\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha)| \to 0 \quad \text{almost surely.} \tag{14}$$

It is also easy to verify that under Assumptions (H'_{reg}), the process $(\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha))_{\theta \in \Theta, \alpha \in \Lambda}$ is subgaussian. Thus, as in the proof of Theorem 4.1, we obtain via Dudley's inequality that

$$\mathbb{E} \sup_{\theta \in \Theta, \alpha \in \Lambda} |\Delta \hat{L}(\theta, \alpha) - \Delta L(\theta, \alpha)| = \mathrm{O}\Big(\frac{1}{\sqrt{n}}\Big), \tag{15}$$

since $\mathbb{E}\Delta \hat{L}(\theta, \alpha) = \Delta L(\theta, \alpha)$. The result follows by combining (14) and (15).

5.3 Some technical lemmas

Lemma 5.1. *Under Assumptions* (H'_{reg}) *and* (H_1)*, there exists an open set* $V \subset \Theta^\circ$ *containing* $\bar\theta$ *such that, for all* $\theta \in V$*,* $\arg\max_{\alpha\in\Lambda} L(\theta,\alpha) \cap \Lambda^\circ \neq \emptyset$*.*

Proof. Assume that the statement is not true. Then there exists a sequence $(\theta_k)_k \subset \Theta$ such that $\theta_k \to \bar\theta$ and, for all k, $\alpha_k \in \partial\Lambda$, where $\alpha_k \in \arg\max_{\alpha\in\Lambda} L(\theta_k,\alpha)$. Thus, since Λ is compact, even if this means extracting a subsequence, one has $\alpha_k \to z \in \partial\Lambda$ as $k \to \infty$. By the continuity of L, $L(\bar\theta,\alpha_k) \to L(\bar\theta, z)$. But

$$\begin{aligned}|L(\bar\theta,\alpha_k) - L(\bar\theta,\bar\alpha)| &\le |L(\bar\theta,\alpha_k) - L(\theta_k,\alpha_k)| + |L(\theta_k,\alpha_k) - L(\bar\theta,\bar\alpha)| \\ &\le \sup_{\alpha\in\Lambda} |L(\bar\theta,\alpha) - L(\theta_k,\alpha)| + |\sup_{\alpha\in\Lambda} L(\theta_k,\alpha) - \sup_{\alpha\in\Lambda} L(\bar\theta,\alpha)| \\ &\le 2\sup_{\alpha\in\Lambda} |L(\bar\theta,\alpha) - L(\theta_k,\alpha)|,\end{aligned}$$

which tends to zero as $k \to \infty$ by (H'_D) and (H'_p). Therefore, $L(\bar\theta, z) = L(\bar\theta,\bar\alpha)$ and, in turn, $z = \bar\alpha$ by (H_1). Since $z \in \partial\Delta$ and $\bar\alpha \in \Delta^\circ$, this is a contradiction. □

Lemma 5.2. *Under Assumptions* (H'_{reg})*,* (H_1)*, and* (H_{loc})*, one has* $\hat I_3 \to HV(\bar\theta)$ *almost surely.*

Proof. We have

$$\hat I_3 = \int_0^1 H\hat V(\hat\theta + t(\hat\theta - \bar\theta))\mathrm{d}t = \int_0^1 \big(H_1\hat L(\hat\theta_t, \hat\alpha(\hat\theta_t)) + J(\nabla_1 \hat L(\hat\theta_t,\cdot))_{\hat\alpha(\hat\theta_t)} J(\hat\alpha)_{\hat\theta_t}\big)\mathrm{d}t,$$

where we set $\hat\theta_t = \hat\theta + t(\hat\theta - \bar\theta)$. Note that $\hat\theta_t \in U$ for all $t \in [0,1]$. By Lemma 4.1,

$$\begin{aligned}&\sup_{t\in[0,1]} \|H_1\hat L(\hat\theta_t, \hat\alpha(\hat\theta_t)) - H_1 L(\hat\theta_t, \hat\alpha(\hat\theta_t))\| \\ &\quad\le \sup_{\theta\in\Theta,\alpha\in\Lambda} \|H_1\hat L(\theta,\alpha) - H_1 L(\theta,\alpha)\| \to 0 \quad \text{almost surely.}\end{aligned}$$

Also, by Theorem 4.2, for all $t \in [0,1]$, $\hat\theta_t \to \bar\theta$ almost surely. Besides,

$$\begin{aligned}|L(\bar\theta,\hat\alpha(\hat\theta_t)) - L(\bar\theta,\alpha(\bar\theta))| &\le |L(\bar\theta,\hat\alpha(\hat\theta_t)) - L(\hat\theta_t,\hat\alpha(\hat\theta_t))| + |L(\hat\theta_t,\hat\alpha(\hat\theta_t)) - L(\bar\theta,\alpha(\bar\theta))| \\ &\le \sup_{\alpha\in\Lambda} |L(\bar\theta,\alpha) - L(\hat\theta_t,\alpha)| + 2\sup_{\theta\in\Theta,\alpha\in\Lambda} |\hat L(\theta,\alpha) - L(\theta,\alpha)|.\end{aligned}$$

Thus, via (H'_{reg}), (H_1), and Lemma 4.1, we conclude that almost surely, for all $t \in [0,1]$, $\hat\alpha(\hat\theta_t) \to \alpha(\bar\theta) = \bar\alpha$. Accordingly, almost surely, for all $t \in [0,1]$, $H_1 L(\hat\theta_t, \hat\alpha(\hat\theta_t)) \to H_1 L(\bar\theta,\bar\alpha)$. Since $H_1 L(\theta,\alpha)$ is bounded under (H'_D) and (H'_p), the Lebesgue dominated convergence theorem leads to

$$\int_0^1 H_1\hat L(\hat\theta_t, \hat\alpha(\hat\theta_t))\mathrm{d}t \to H_1 L(\bar\theta,\bar\alpha) \quad \text{almost surely.} \tag{16}$$

Furthermore,

$$J(\hat{\alpha})_\theta = -H_2\hat{L}(\theta,\hat{\alpha}(\theta))^{-1}J\big(\nabla_2\hat{L}(\cdot,\hat{\alpha}(\theta))\big)_\theta, \quad \forall(\theta,\alpha)\in U\times\alpha(U),$$

where U is the open set defined in the proof of Theorem 4.3. By the cofactor method, $H_2\hat{L}(\theta,\alpha)^{-1}$ takes the form

$$H_2\hat{L}(\theta,\alpha)^{-1} = \frac{\hat{c}(\theta,\alpha)}{\det(H_2\hat{L}(\theta,\alpha))},$$

where $\hat{c}(\theta,\alpha)$ is the matrix of cofactors associated with $H_2\hat{L}(\theta,\alpha)$. Thus, each component of $-H_2\hat{L}(\theta,\alpha)^{-1}J(\nabla_2\hat{L}(\cdot,\alpha))_\theta$ is a quotient of a multilinear form of the partial derivatives of $\hat{L}$ evaluated at (θ,α) divided by $\det(H_2\hat{L}(\theta,\alpha))$, which is itself a multilinear form in the $\frac{\partial^2\hat{L}}{\partial\alpha_i\partial\alpha_j}(\theta,\alpha)$. Hence, by Lemma 4.1, we have

$$\sup_{\theta\in U,\alpha\in\alpha(U)} \|H_2\hat{L}(\theta,\alpha)^{-1}J(\nabla_2\hat{L}(\cdot,\alpha))_\theta - H_2L(\theta,\alpha)^{-1}J(\nabla_2L(\cdot,\alpha))_\theta\| \to 0 \text{ almost surely.}$$

So, for all n large enough,

$$\begin{aligned}
&\sup_{t\in[0,1]} \|J(\hat{\alpha})_{\hat{\theta}_t} + H_2L(\hat{\theta}_t,\hat{\alpha}(\hat{\theta}_t))^{-1}J\big(\nabla_2L(\cdot,\hat{\alpha}(\hat{\theta}_t))\big)_{\hat{\theta}_t}\| \\
&\quad\le \sup_{\theta\in U,\alpha\in\alpha(U)} \|H_2\hat{L}(\theta,\alpha)^{-1}J(\nabla_2\hat{L}(\cdot,\alpha))_\theta - H_2L(\theta,\alpha)^{-1}J(\nabla_2L(\cdot,\alpha))_\theta\| \\
&\quad\to 0 \quad \text{almost surely.}
\end{aligned}$$

We know that almost surely, for all $t\in[0,1]$, $\hat{\alpha}(\hat{\theta}_t)\to\bar{\alpha}$. Thus, since the function $U\times\alpha(U)\ni(\theta,\alpha)\mapsto H_2L(\theta,\alpha)^{-1}J(\nabla_2L(\cdot,\alpha))_\theta$ is continuous, we have almost surely, for all $t\in[0,1]$,

$$H_2\hat{L}(\hat{\theta}_t,\hat{\alpha}(\hat{\theta}_t))^{-1}J\big(\nabla_2\hat{L}(\cdot,\hat{\alpha}(\hat{\theta}_t))\big)_{\hat{\theta}_t} \to H_2L(\bar{\theta},\bar{\alpha})^{-1}J(\nabla_2L(\cdot,\bar{\alpha}))_{\bar{\theta}}.$$

Therefore, almost surely, for all $t\in[0,1]$, $J(\hat{\alpha})_{\hat{\theta}_t}\to J(\alpha)_{\bar{\theta}}$. Similarly, almost surely, for all $t\in[0,1]$, $J(\nabla_1\hat{L}(\hat{\theta}_t,\cdot))_{\hat{\alpha}(\hat{\theta}_t)}\to J(\nabla_1L(\bar{\theta},\cdot))_{\bar{\alpha}}$. All involved quantities are uniformly bounded in t, and so, by the Lebesgue dominated convergence theorem, we conclude that

$$\int_0^1 J(\nabla_1\hat{L}(\hat{\theta}_t,\cdot))_{\hat{\alpha}(\hat{\theta}_t)}J(\hat{\alpha})_{\hat{\theta}_t}\mathrm{d}t \to J(\nabla_1L(\bar{\theta},\cdot))_{\bar{\alpha}}J(\alpha)_{\bar{\theta}} \quad \text{almost surely.} \tag{17}$$

Consequently, by combining (16) and (17),

$$\hat{I}_3 \to H_1L(\bar{\theta},\bar{\alpha}) + J(\nabla_1L(\bar{\theta},\cdot))_{\bar{\alpha}}J(\alpha)_{\bar{\theta}} = HV(\bar{\theta}) \quad \text{almost surely,}$$

as desired. □

Acknowledgments

We thank Flavian Vasile (Criteo) for stimulating discussions and insightful suggestions.

References

T. Angles and S. Mallat. Generative networks as inverse problems with scattering transforms. In *International Conference on Learning Representations*, 2018.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. Proceedings of Machine Learning Research, 2017.

L. Devroye. Universal smoothing factor selection in density estimation: Theory and practice. *TEST*, 6:223–320, 1997.

G.K. Dziugaite, D.M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267. AUAI Press, Arlington, 2015.

D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49:1858–1860, 2003.

I. Goodfellow. *NIPS 2016 Tutorial: Generative Adversarial Networks*. arXiv:1701.00160, 2016.

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and J. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., Red Hook, 2014.

S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5551–5559. Curran Associates, Inc., Red Hook, 2017.

C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, Cambridge, 1989.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc., Red Hook, 2016.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., Red Hook, 2016.

R. van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University, 2016.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discriminative-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.