BRIDGING THE GAP BETWEEN f-DIVERGENCES AND BAYES HILBERT SPACES

Linus Lach*, **Alexander Fottner, Yarema Okhrin**
Department of Statistics and Data Science
University of Augsburg
{linus.lach,alexander.fottner,yarema.okhrin}@uni-a.de

ABSTRACT

We introduce a novel framework that generalizes f-divergences by incorporating divergence-generating functions that exhibit non-convex behavior in a neighborhood of the origin. Using this extension, we define a new class of pseudo f-divergences, encompassing a wider range of distances between distributions that traditional f-divergences cannot capture. Among these, we focus on a particular pseudo divergence, obtained by considering the induced metric of Bayes Hilbert spaces. Bayes Hilbert spaces are frequently used due to their inherent connection to Bayes's theorem as they allow sampling from potentially intractable posterior densities, a challenging task until now. In the more general context, we prove that pseudo f-divergences are well-defined and introduce a variational estimation framework that can be used in a statistical learning context. By applying this variational estimation framework to f-GANs, we achieve improved FID scores over existing f-GAN architectures and competitive results with the Wasserstein GAN, highlighting its potential for both theoretical research and practical applications in learning theory.

1 INTRODUCTION

A central challenge in learning theory is estimating statistical divergences between distributions from finite samples to quantify their dissimilarity. f-divergences are a popular class of divergences (Csiszár et al., 2004; Liese & Vajda, 2006) including the Kullback-Leibler (KL) and Pearson divergence. Their general form is given by $D_f(\mu, \nu) = \int f(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})\mathrm{d}\mu$, where f is a convex lower-semicontinuous function satisfying $f(1) = 0$. f-divergences and the subsequent estimation have been extensively studied in Nguyen et al. (2010); Rubenstein et al. (2019); Nguyen et al. (2009); Moon & Hero (2014); Keziou (2003). However, research on generalizing these divergences (Gimenez & Zou, 2022; Birrell et al., 2022; Chen et al., 2024), particularly by relaxing the convexity constraint, still needs to be explored.

We introduce a novel pseudo f-divergence framework that relaxes the convexity constraint on divergence-generating functions, offering greater flexibility and applicability in various machine-learning tasks. This flexibility is crucial in tasks like generative image modeling (Nowozin et al., 2016), or topic modeling on large text data sets (Hoffman et al., 2013), where the underlying data is diverse. Previously, the convexity requirement hindered connections between f-divergences and other dissimilarity measures like the Bayes Hilbert space metric (Van den Boogaart et al., 2014), frequently used in Bayesian statistics. In Bayesian statistics, inference typically requires sampling from intractable densities (Kroese et al., 2013). Sampling from intractable densities using divergences in a bayesian setting has been studied in Blei et al. (2017); Campbell & Broderick (2018); Campbell & Beronov (2019). However, these methods have only been applied using the KL-divergence (Blei et al., 2017; Campbell & Beronov, 2019) or lower dimensional synthetic data sets (Campbell & Broderick, 2018).

To highlight the application of general f-divergences in a Bayesian statistics context, we therefore explore a connection between f-divergences and Bayes Hilbert spaces, which provide an alternative structure for measuring distances between distributions. Bayes Hilbert spaces are distribution spaces with a vector structure, intrinsically connected to Bayes's theorem, where addition aligns with

Bayesian updating (Van den Boogaart et al., 2014). Due to their Hilbert space structure, they are equipped with an induced metric, providing a natural notion of distance between measures that can be incorporated into the proposed pseudo f-divergence framework. Another advantage of estimating the Bayes Hilbert space metric over traditional divergences arises from the Hilbert space structure. The Hilbert space structure implies that every Cauchy sequence converges within the space, which can be exploited in an optimization setting. By modeling the loss function as a sequence in a Bayes Hilbert space and using gradient-based approaches, convergence to a solution near the optimum is ensured, preventing instability near the optimum. However, defining a Hilbert space for probability measures is not straightforward (Van den Boogaart et al., 2010). First, probability measures must be identified as vectors before imposing a linear structure on them to obtain a linear space. This linear space can then be endowed with an inner product to define a Hilbert space. The linear space structure allows probability measures and their associated densities to be embedded in a subspace of L^2, enabling a clear understanding of the distances between distributions.

The connection between f-divergences and Bayes Hilbert spaces can be established using the clr transformation for distributions, which is defined as $\mathrm{clr}(\mu) = \log(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}) - \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda})\mathrm{d}\lambda$, where $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ is the Radon-Nikodym derivative of μ with respect to some continuous base measure λ. The induced metric of a Bayes Hilbert space, which involves the squared clr transformation, closely relates to a pseudo f-divergence with the pseudo-divergence-generating function $x \mapsto x\log(x)^2$. Since this function encompasses a local non-convexity on the interval $[0, \exp(-1)]$, we apply our newly developed framework, which allows us to estimate the so-called BHS divergence and also exploit the characteristics of Bayes Hilbert spaces to obtain a lower bound for the pseudo divergence that is attained.

1.1 Contributions

Our main contributions can be summarized as follows:

- We develop a novel theoretical framework bridging the gap between f-divergences and Bayes Hilbert spaces.
- We generalize f-divergences by allowing for locally non-convex divergence-generating functions.
- We present an efficient sampling algorithm for high dimensional distributions.
- We apply our extended pseudo-divergence estimation framework to f-GANs, outperforming traditional f-GANs achieving lower FID scores.

1.2 Related Work

f-divergences: Our work extends the existing body of research on f-divergences and their application in learning theory, particularly in the context of sampling from unknown distributions. A foundational approach to estimating f-divergences was developed by Nowozin et al. (2016); Nguyen et al. (2010), where the authors developed and utilized a variational estimation framework to estimate f-divergences with an application to GANs. Their framework allows for flexible divergence estimation but remains constrained by the convexity of divergence-generating functions. Our approach addresses this limitation by relaxing the convexity constraint, allowing for a broader range of pseudo f-divergences that can be tailored to specific tasks.

Besides f-divergences, Rényi divergences, which generalize the KL divergence by introducing a parameter $\alpha > 0, \alpha \neq 1$ that corresponds to a weight controlling the sensitivity of divergence regarding areas where the distributions are similar or dissimilar have been developed to provide greater flexibility (Van Erven & Harremos, 2014). While Rényi divergences provide greater flexibility compared to KL, they are still limited by the convexity constraint inherent in f-divergences.

A different approach to measure the dissimilarity between probability measures is offered by integral probability metric (IPM) such as the Wasserstein distance. The Wasserstein distance has gained popularity in generative modeling due to its robustness in comparing distributions with different supports (Sriperumbudur et al., 2012). However, as Yatracos (2022) has noted, the Wasserstein distance is unsuitable for estimating distributions with heavy tails, as it struggles to handle distributions whose tails are not exponentially bounded. Recent work by Chen et al. (2024) addresses this issue by proposing a penalized version of the Wasserstein distance, which combines it with α-divergences to estimate heavy-tailed distributions better. While promising, their method requires careful tuning of the penalty terms and relies on the Wasserstein metric's limitations, particularly in high-dimensional spaces.

In contrast, Birrell et al. (2022) combines IPMs and f-divergences introducing (f, Γ)-divergences that inherit the ability to compare distributions that are not absolutely continuous from IPMs, while also retaining the property of f-divergences to control heavy-tailed distributions. Although this hybrid approach broadens the applicability of divergence estimation, it still imposes the convexity constraint on the divergence-generating function.

Bayes Hilbert Spaces: Bayes spaces, and their extension to Bayes Hilbert spaces (Van den Boogaart et al., 2010; 2014; Egozcue et al., 2006) while still in their infancy, were introduced as a natural generalization of the theory of compositional data, originally popularized by Aitchison in Aitchison (1982; 1986); Van den Boogaart & Tolosana-Delgado (2013). These spaces address the challenge of defining meaningful distances between probability measures for specific data analysis tasks. The core idea is to introduce a vector space structure for measures, where the group operations align with Bayes updating (Van den Boogaart et al., 2014), thus creating a framework in which Bayesian reasoning can be incorporated directly into the space of measures.

Bayes Hilbert spaces are frequently used in functional regression, including Maier et al. (2021); Arata (2017); Machalova et al. (2016); Machalová et al. (2021); Talská et al. (2018), where Radon-Nikodym derivatives are modeled using spline functions within the Bayes Hilbert space.

However, one significant challenge with these approaches is their reliance on functional regression, where the targets or features are typically density functions. Since these densities are not directly observed in practice, the models must infer them from finite samples generated by the underlying distributions. This creates a gap between the theoretical framework of Bayes Hilbert spaces and practical learning scenarios.

Our work addresses this gap by incorporating a generative adversarial network (GAN) to approximate the underlying probability distribution while simultaneously minimizing the distance between the learned and true density in the Bayes Hilbert space. This integration allows us to bridge the functional models of Bayes spaces with finite-sample learning, offering a more practical and scalable solution to regression tasks involving density functions.

In parallel, Wynne (Wynne, 2023) explored the use of Bayes Hilbert spaces for Bayesian posterior approximation, concluding that these spaces provide a natural and appropriate framework for approximating posterior distributions. Our approach builds on this insight by extending its application to finite sample learning and divergence minimization in the Bayes Hilbert space.

1.3 Outline

We begin by reviewing key properties of f-divergences and establish theoretical assumptions on the underlying spaces. Next, we address challenges in quantifying dissimilarities between measures using f-divergences. Since estimating f-divergences relies on a variational framework involving Fenchel conjugates, we discuss how these conjugates are used in a dual representation of convex divergence-generating functions to derive an estimable lower bound. We then outline the conditions for realizing this bound and present its analytical form, which remains valid under milder assumptions.

In Section 3, we review Bayes Hilbert spaces as a framework for measuring distances between probability measures by embedding them into function spaces with a separable Hilbert space structure defined through the clr. The induced metric corresponds to the log-variance, which connects to f-divergences.

Finally, we apply this generalized divergence estimation framework to f-GANs, demonstrating that the more flexible pseudo-divergence approach outperforms traditional f-divergences.

Proofs of our theoretical results are provided in the Appendix.

2 f-DIVERGENCES

Statistical divergences measure the dissimilarity between probability distributions and are often used to estimate parameters in statistical models, such as generative algorithms designed to sample from unknown distributions. f-divergences are a widely used class of these measures defined as:

$$D_f(\mu,\nu) = \int_\Omega p_\nu(x) f\left(\frac{p_\mu(x)}{p_\nu(x)}\right) \lambda(\mathrm{d}x).$$

where μ and ν are assumed to be two probability distributions on the same measure space $(\Omega, \mathcal{B}, \lambda)$ ($\mathcal{B}$ denoting the Borel σ-field over Ω) that are both absolutely continuous with respect to the σ-finite base measure λ. p_μ and p_ν denote the Radon-Nikodym densities of μ and ν with respect to λ. Common choices for λ are the Lebesgue measure or counting measure. f is an arbitrary convex, lower-semicontinuous function $f : \mathbb{R}_+ \to \mathbb{R}$, satisfying $f(1) = 0$ called divergence generating function. By requiring that $f(1) = 0$, it is ensured that if $\mu = \nu$ a.s., $D_f(\mu,\nu) = 0$ and the lower-semicontinuity of f ensures that $\lim_{t \searrow 1} f(t) = f(0)$. Note, that f-divergences generally do not induce metrics as they are not symmetric in most cases, i.e., $D_f(\mu,\nu) \neq D_f(\nu,\mu)$. However, they can still effectively be used for measuring dissimilarity or pseudo distances between measures. Since the function f must be convex, which can be restrictive when defining new divergences, it potentially limits the range of distributions a generative model can learn effectively. A challenge with f-divergences is their difficulty in estimation from finite samples of μ and ν. To address this, particularly in estimating likelihood ratios for multivariate distributions, convex optimization techniques are employed (Nguyen et al., 2010). The optimization uses variational estimation involving the Fenchel conjugate of a continuous convex function defined as:

$$f^*(y) \coloneqq \sup_{x \in \mathrm{dom}(f)} \{x^T y - f(x)\}, \quad y \in \mathrm{dom}(f^*).$$

This implies the dual representation of f, known as the Fenchel or convex duality:

Theorem 2.1 (Rockafellar, 1970). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. If f is lower semicontinuous, then the duality $f^{**}(x) = f(x)$ for all $x \in \mathbb{R}^n$ holds.*

Leveraging the dual representation $f(x) = f^{**}(x) = \sup_{y \in \mathrm{dom}(f^*)}\{x^T y - f^*(y)\}$ in the sense of Theorem 2.1, a lower bound for the divergence of μ and ν can be derived (Nguyen et al., 2010):

$$D_f(\mu,\nu) \geq \sup_{T \in \mathcal{T}} \left\{ \mathbb{E}_\mu(T) - \mathbb{E}_\nu(f^* \circ T) \right\}. \tag{1}$$

Here, $\mathcal{T}$ denotes an arbitrary class of Borel measurable functions $T : \Omega \to \mathrm{dom}(f^*)$. By adding the additional assumption that f is continuously differentiable on its entire domain, the bound in equation 1 is tight and the supremum can be represented analytically as

$$\tilde{T}(x) \coloneqq f'\left(\frac{p_\mu(x)}{p_\nu(x)}\right).$$

By adapting the definition of convex conjugates to allow concave functions instead of convex functions, Theorem 2.1 can be adjusted to hold for upper-semicontinuous, concave functions as well. The concave conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f_*(y) \coloneqq \inf_{x \in \mathrm{dom}(f)} \{x^T y - f(x)\}, \quad y \in \mathrm{dom}(f_*).$$

Fenchels duality in the concave case then states $f_{**}(x) = f(x)$ for all $x \in \mathbb{R}^n$ (Rockafellar, 1970).

3 BAYES HILBERT SPACES

Another idea for quantifying distances between probability measures is the embedding into suitable spaces of functions such as Hilbert spaces. The induced metric of these spaces can then be used to measure the distance between those probability measures. This basic idea will be formulated mathematically in the following paragraph which is based on Van den Boogaart et al. (2014; 2010).

Denote by $\mathcal{M}$ the set of σ–finite measures on some Borel measurable space $(\Omega, \mathcal{B})$ that are equivalent to a σ–finite base measure λ on $\mathcal{B}$. Then, μ and ν are called B–equivalent, denoted by $\mu =_{B(\lambda)} \nu$, if there exists a constant $c > 0$ such that $\mu(A) = c\nu(A)$ for all $A \in \mathcal{B}$ with the convention $c \cdot (+\infty) = +\infty$. Define $B(\lambda)$ as the set of $=_B$–equivalent classes, i.e., the quotient space $\mathcal{M}(\lambda)/ =_B$. The relation $=_B$ is an equivalence relation on $\mathcal{M}(\lambda)$. Since the considered measures are equivalent to each other, the Radon–Nikodym theorem ensures that the densities $\mathrm{d}\mu/\mathrm{d}\nu$ exist for every $\mu, \nu \in \mathcal{M}(\lambda)$.

Given the equivalence class of $=_B$–equivalent measures, addition and multiplication can be defined: For $R \in \mathcal{B}$ define the perturbation of μ by ν, the powering of μ for a scalar $\alpha \in \mathbb{R}$, and the negative perturbation in $B(\lambda)$ as

$$(\mu \oplus \nu)(R) \coloneqq \int_R \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\,\mathrm{d}\lambda, \quad (\alpha \odot \mu)(R) \coloneqq \int_R \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^{\alpha} \mathrm{d}\lambda, \quad (\mu \ominus \nu)(R) \coloneqq \int_R \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\lambda.$$

In Bayes Hilbert space literature it is common to identify measures with their corresponding Radon–Nikodym densities. This is coherent in the sense that by setting $\mu =_B \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ for every $\mu \in B(\lambda)$ it holds that $\mu \oplus \nu =_B \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}$.

$(B(\lambda), \oplus, \odot)$ is a real linear space called Bayes linear space with base measure λ. The additive neutral element is given by $0_B \coloneqq \lambda$ and the additive inverse element $\ominus\mu \coloneqq \frac{\mathrm{d}\lambda}{\mathrm{d}\mu}$. To derive a Hilbert space based on this vector space, we now assume that the base measure λ is a probability measure. The space $B^p(\lambda),\ p \geq 1$ of λ–equivalent measures defined as

$$B^p(\lambda) \coloneqq \left\{\mu \in B(\lambda) : \int \left|\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\right|^p \mathrm{d}\lambda < +\infty\right\}$$

is a vector subspace of $B(\lambda)$. For details and more basic properties, see Van den Boogaart et al. (2010).

Let $\mu \in B^1(\lambda)$ with corresponding density $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$. Then, define the centered log-ratio transformation clr of μ as

$$\mathrm{clr}(\mu) \coloneqq \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right) - \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda.$$

The clr transformation has the desirable property of maintaining the relation of $\oplus$ with multiplication and $\odot$ with exponentiation respectively which is enforced by the properties of the logarithm. Furthermore, the map $\mathrm{clr} : B^1(\lambda) \to L_0^1(\lambda)$ is linear.

Recall, that $L_0^1(\lambda) = \{g \in L^1(\lambda) : \int g\mathrm{d}\lambda = 0\}$. Given the map clr, an inner product on B^2 can be defined which in turn induces a metric.

Let therefore p_μ and p_ν be two Radon-Nikodym densities in $B^2(\lambda)$, i.e., for $\mu, \nu \in B^2(\lambda)$ define $p_\mu(x) \coloneqq \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)$ and $p_\nu(x) \coloneqq \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x)$ for all $x \in \mathbb{R}^n$. Then, the inner product $\langle p_\mu(x), p_\nu(x)\rangle_{B^2(\lambda)}$ is defined as:

$$\int_\Omega \left(\log(p_\mu(x)) - \int \log(p_\mu(y))\,\lambda(\mathrm{d}y)\right)\left(\log(p_\nu(x)) - \int \log(p_\nu(y))\,\lambda(\mathrm{d}y)\right)\lambda(\mathrm{d}x).$$

$B^2(\lambda)$ is a separable Hilbert space and the map $\mathrm{clr} : B^2(\lambda) \to L_0^2(\lambda)$ is an isometry of Hilbert spaces.

The map $d_{B^2(\lambda)} : B^2(\lambda) \times B^2(\lambda) \to [0, \infty)$, $d_{B^2(\lambda)}(p_\mu, p_\nu) \coloneqq \left(\langle p_\mu \ominus p_\nu, p_\mu \ominus p_\nu\rangle_{B^2(\lambda)}\right)^{\frac{1}{2}}$ is an induced metric and the pair $(B^2(\lambda), d_{B^2(\lambda)})$ is a complete metric space.

If we want to construct an f-divergence that is closely related to the aforementioned metric, its explicit form is of major interest. We can derive such an analytical expression of the Bayes Hilbert space metric using the following Lemma.

Lemma 3.1. *Let p_μ and p_ν be two Radon–Nikodym densities in $B^2(\lambda)$. Then,*

$$d_{B^2(\lambda)}(p_\mu, p_\nu)^2 = \mathrm{Var}_\lambda(\log(\mu \ominus \nu)).$$

Therefore the Bayes Hilbert space metric directly corresponds to the log-variance loss studied in e.g. Richter et al. (2020); Nüsken & Richter (2021); Richter & Berner (2024), embedding those works in the area of Bayes Hilbert space learning.

4 BRIDGING THE GAP BETWEEN BAYES HILBERT SPACES AND f-DIVERGENCES

In this section, we show how a pseudo f-divergence can be formulated to directly relate to the Bayes Hilbert space metric d_{B^2}, enabling sampling from a probability distribution within a well-defined Bayes Hilbert space. This approach addresses a key challenge in Bayes Hilbert space learning, as highlighted in Maier et al. (2021): estimating densities through direct sampling from the learned distribution.

4.1 MIXED CONJUGATES

A key challenge is that f-divergences are only well-defined when f is lower semi-continuous and convex. Thus to establish the desired connection, we must first extend the definition of f-divergences to handle local subsets near the origin where f is not convex. Consider a continuous function $f : [0, \infty) \to \mathbb{R}$, which is concave on $[0, a)$ for some $a > 0$ and convex on $[a, \infty)$. To account for this local non-convexity, we define the mixed conjugate f_*^* of f as

$$f_*^*(t) := \sup_{x \in [a,\infty)} \{tx - f(x)\} \mathbb{1}_{\{t \in M\}} + \inf_{x \in [0,a)} \{tx - f(x)\} \mathbb{1}_{\{t \in N\}}, \tag{2}$$

where

$$M := \big\{ t \in \operatorname{dom}(f_*^*) : \operatorname*{argmax}_{x \in \operatorname{dom}(f)} (tx - f(x)) \in [a, \infty) \big\},$$
$$N := \big\{ t \in \operatorname{dom}(f_*^*) : \operatorname*{argmin}_{x \in \operatorname{dom}(f)} (tx - f(x)) \in [0, a) \big\}.$$

The set M (N) ensures that we consider the convex (concave) conjugate on the domain of f_*^* where f is convex (concave). Under mild conditions, this definition allows a disjoint characterization of the subdomains M and N. The motivation behind this mixed conjugate is straightforward: it enables a well-defined dual representation for locally non-convex functions.

Lemma 4.1. *For any continuous function $f : [0, \infty) \to \mathbb{R}$ that is convex on $[a, \infty)$, $a > 0$, concave on $[0, a)$, and satisfies $\lim_{x\to\infty} f(x) = +\infty$, the mixed conjugate satisfies:*

1. *$M \cap N = \emptyset$ and $M \cup N = \operatorname{dom}(f_*^*)$*
2. *$f_{**}^{**} = f$ for all $t \in \operatorname{dom}(f)$, where f_{**}^{**} denotes the biconjugate of f, i.e. $f_{**}^{**} = (f_*^*)_*^*$.*

The unboundedness assumption on f is not only a theoretical necessity in the proof of the lemma above but also carries significance in the context of divergence-generating functions. By assuming that f is unbounded, we permit the dissimilarity between probability measures to grow arbitrarily large.

At first glance, it is not clear why such pseudo-divergences are well-defined, however, the following Lemma proves their coherence.

Lemma 4.2. *Let $f : [0, \infty) \to [0, \infty)$ be concave on some interval $[0, a)$, $a > 0$ and convex on $[a, \infty)$. For any probability measures $\mu, \nu \in B^2(\lambda)$, $D_f(\mu, \nu)$ is well-defined in the sense that $D_f(\mu, \nu) \geq 0$ and $D_f(\mu, \nu) = 0 \iff \mu = \nu \quad \lambda - a.s.$*

To address the challenge of estimating likelihood ratios from finite samples, we build on the approach in Nguyen et al. (2010) by developing a generalized variational framework applicable to pseudo f-divergences.

Theorem 4.3. *Let $f : [0, \infty) \to \mathbb{R}$, convex on $[a, \infty)$ with $0 < a < \infty$, concave on $[0, a)$, assume that f is twice continuously differentiable, and that $((f_*^*)')^{-1}$ is λ-a.s. well-defined. Furthermore, let $\mu, \nu \in B^2(\lambda)$ with $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) := p_\mu(x)$, $\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x) := p_\nu(x)$, $x \in \mathbb{R}$. Then, $\tilde{T}(x) := f'\left(\frac{p_\mu(x)}{p_\nu(x)}\right)$ is an optimizer for*

$$\sup_{T \in C(M^*)} \big\{ (\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T)) \big\} + \inf_{T \in C(N^*)} \big\{ (\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T)) \big\}. \tag{3}$$

Here, $C(M^*)$ ($C(N^*)$) denotes the set of continuous functions $T : \Omega \to \operatorname{dom}(f_*^*)$ such that $f_*^* \circ T$ is convex (concave).

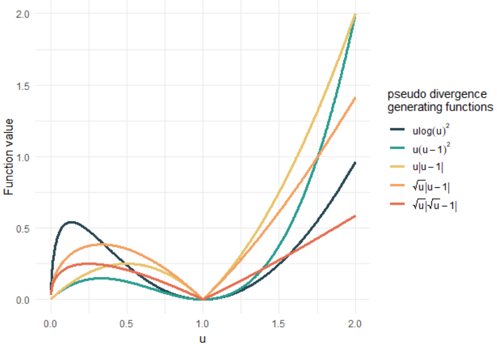

Figure 1: Plot of different pseudo-divergence-generating functions f

Table 1: Table of different pseudo-divergence-generating functions and their respective pseudo divergence

$f(u)$	$D_f(\mu,\nu)$
$u\log(u)^2$	$\mathbb{E}_\mu\left(\log\left(\frac{p_\mu(x)}{p_\nu(x)}\right)^2\right)$
$u(u-1)^2$	$\mathbb{E}_\mu\left(\left(\frac{p_\mu(x)}{p_\nu(x)}-1\right)^2\right)$
$u\|u-1\|$	$\mathbb{E}_\mu\left(\left\|\frac{p_\mu(x)}{p_\nu(x)}-1\right\|\right)$
$\sqrt{u}\|u-1\|$	$\mathbb{E}_\mu\left(\left\|\frac{\sqrt{p_\mu(x)}}{\sqrt{p_\nu(x)}}-\frac{\sqrt{p_\nu(x)}}{\sqrt{p_\mu(x)}}\right\|\right)$
$\sqrt{u}\|\sqrt{u}-1\|$	$\mathbb{E}_\mu\left(\left\|1-\frac{\sqrt{p_\nu(x)}}{\sqrt{p_\mu(x)}}\right\|\right)$

Theorem 4.3 can be applied to the extended class of divergence-generating functions. A brief selection of such pseudo-divergence-generating functions and their respective pseudo-divergence can be in Figure 1 and Table 1.

The optimization problem in equation 3 can be transformed to reduce the two objectives into a single equivalent objective with the same optimum. This transformation leverages the property $\inf(\cdot) = -\sup(\cdot)$ and substituting T with $\bar{T} = T(\mathbb{1}_{\{T\in C(M^*)\}} - \mathbb{1}_{\{T\in C(N^*)\}})$. For details, see Proof of Corollary A.1.

Corollary 4.4. *For any function f satisfying the conditions of Theorem 4.3*

$$D_f(\mu,\nu) \geq \sup_{\bar{T}\in C(M^*)\cup C(N^*)} \left\{\mathbb{E}_\mu(\bar{T}) - \mathbb{E}_\nu(f^*_* \circ \bar{T})\right\}. \tag{4}$$

Therefore, by applying the generalizing mixed conjugate to the variational estimation framework of Nguyen et al. (2010), we can still derive a lower bound for the pseudo f–divergence.

4.2 Bayes Hilbert Space Divergence

We can now utilize this newly established mixed conjugate framework to bridge a connection between the induced pseudo f–divergences and Bayes Hilbert spaces. One locally non-convex function we already considered in Table 1 is closely related to the centered log-ratio transformation and of special interest for the remainder of the paper. It is defined as

$$f_{BHS} : [0,\infty) \to \mathbb{R}, \quad x \mapsto x\log(x)^2. \tag{5}$$

It is easy to verify that f_{BHS} is concave on the interval $[0, e^{-1}]$ as well as $f_{BHS}(1) = 0$. We can therefore use f_{BHS} to define the pseudo f-divergence $D_{f_{BHS}}$.

Let $\mu \in B^2(\lambda)$, then we can derive a connection between this newly constructed divergence and the Bayes Hilbert space metric:

Lemma 4.5. *For f_{BHS} as defined in equation 5, it holds that*

$$D_{f_{BHS}}(\mu,\nu) = d_{B^2(\mu)}(p_\mu, p_\nu)^2 + \mathbb{E}_\mu\left(\log(\mu\ominus\nu)\right)^2.$$

Minimizing $D_{f_{BHS}}$ also minimizes $d_{B^2(\mu)}(p_\mu, p_\nu)$, ensuring that two probability measures close with respect to $D_{f_{BHS}}$ are similarly close in the corresponding Bayes Hilbert space. This aligns intuitively with statistical learning, where we assume μ as the true underlying measure and aim to learn a measure ν that closely approximates it. Consequently, μ generates a Bayes Hilbert space $B^2(\mu)$ in which the approximations must reside. Since densities relative to μ are rarely of primary interest, we convert them to more familiar densities, such as those with respect to Lebesgue measures.

The mixed conjugate of f_{BHS} is given by

$$f^*_{*,\mathrm{BHS}}(x) = \begin{cases} 2(-1-\sqrt{1+x})\exp(-1-\sqrt{1+x}) & x\in[-1,0) \\ 2(-1+\sqrt{1+x})\exp(-1+\sqrt{1+x}) & x \geq 0. \end{cases}$$

Here, $M = [0, \infty)$ and $N = [-1, 0)$. Applying Lemma 4.1 implies that $f^{**}_{**,\mathrm{BHS}} = f_{\mathrm{BHS}}$ and Theorem 4.3 implies that $\tilde{T}(x) = f'_{BHS}\left(\frac{p_\mu(x)}{p_\nu(x)}\right)$ is and optimizer for equation 3. Since $f^{*}_{*,\mathrm{BHS}} \circ \tilde{T}(x) = 2\log(x)x$, it follows that

$$\begin{aligned}
&\sup_{T \in C(M^*)} \left\{ (\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f^*_* \circ T))\mathbb{1}_{\{T \in M_C\}} \right\} + \inf_{T \in C(N^*)} \left\{ (\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f^*_* \circ T))\mathbb{1}_{\{T \in N_C\}} \right\} \\
&\qquad = \mathbb{E}_\mu(\tilde{T}) - \mathbb{E}_\nu(f^*_* \circ \tilde{T}) \\
&\qquad = D_{f_{\mathrm{BHS}}}(\mu, \nu).
\end{aligned}$$

Thus, the bound in Corollary 4.4 is tight for f_{BHS}. In summary, we have demonstrated that despite the local non-convexity of f_{BHS}, it defines a pseudo f-divergence closely related to the Bayes Hilbert space metric. A slight adaptation of the results from Nowozin et al. (2016); Nguyen et al. (2010) enables the minimization of a lower bound for $D_{f_{BHS}}$, with this bound being attained by the optimal estimate of $\tilde{T}$.

In the next section, we will apply these results to generative modeling, introducing a new pseudo f-GAN (Bayes Hilbert space GAN) that enables sampling from a learned Bayes Hilbert space distribution.

5 Computational Results

5.1 Setup

Using the PyTorch framework (Paszke et al., 2019), we implemented the Bayes Hilbert space GAN (BHSGAN) and compared its results to traditional f-GANs, shown in Table 2, as well as a Wasserstein-GAN (Arjovsky et al., 2017) on the MNIST (Deng, 2012) (Appendix A.2) and CIFAR10 (Krizhevsky et al., 2009) data sets. Furthermore, we trained the benchmarking GAN architectures to approximate a heavy-tailed distribution, specifically the lognormal distribution with parameters $(\mathrm{loc} = 0, \mathrm{shape} = 0.5, \mathrm{scale} = 1)$.

Table 2: Specification of f-GAN activation functions and Fenchel conjugates similar to Nowozin et al. (2016).

Name	g_f	$\mathrm{dom}(f^*_{(*)})$	(Mixed) Conjugate $f^*_{(*)}(t)$
Kullback-Leibler (KL)	v	$\mathbb{R}$	$\exp(t-1)$
Reverse KL	$-\exp(-v)$	$\mathbb{R}_-$	$-1 - \log(-t)$
Pearson	v	$\mathbb{R}$	$\frac{1}{4}t^2 + t$
GAN	$-\log(1 + \exp(-v))$	$\mathbb{R}_-$	$-\log(1 - \exp(t))$
BHS	$-1 + \exp(v)$	$[-1, \infty)$	$2\left(-1 \pm \sqrt{1+t}\right)\exp\left(-1 \pm \sqrt{1+t}\right)$

In terms of implementation, our approach closely resembles Nowozin et al. (2016) where the authors estimate generative models using variational divergence minimization (Nguyen et al., 2010). According to Corollary 4.4, we need to optimize equation 4 with respect to $\bar{T}$, which is equivalent to optimizing

$$\min_{\vartheta} \max_{\omega} F(\vartheta, \omega) := \min_{\vartheta} \max_{\omega} \mathbb{E}_\mu(\bar{T}_\omega) - \mathbb{E}_{\nu_\vartheta}(f^*_* \circ \bar{T}_\omega). \tag{6}$$

Here, ν_ϑ represents the generative model taking a random vector as an input and returning a sample to be evaluated by the discriminatory model. The discriminatory model $\bar{T}_\omega$ returns some real-valued constant we call *score*. A high value for the score indicates that the discriminatory model is confident of the sample being generated by the generative model. In contrast, a low score indicates that the discriminatory model evaluated the underlying sample to stem from the training data. ϑ and ω denote the parametrizations of the respective models. We optimize equation 6 by sampling from a finite training data set to estimate the underlying distribution. As underlying models for the approximation on the MNIST and synthetic heavy tailed data set, we chose basic feed-forward neural networks identical to the ones used in Nowozin et al. (2016).

For the CIFAR-10 dataset, we employ the DCGAN architecture introduced by Radford et al. (2015). We also utilize the same representation of the variational functions as in Nowozin et al. (2016), summarized as follows: To ensure that $\bar{T}_\omega$ respects the domain of f_*^*, we represent the variational function as $\bar{T}_\omega$ by $\bar{T}_\omega(x) = g_f \circ V_\omega(x)$. Here, V_ω denotes the discriminatory model with no restrictions on the output, meaning $\operatorname{Im}(V_\omega) \subseteq \mathbb{R}$ while g_f is a final output activation function tailored to the domain of f^* (see Table 2). This approach accommodates various f-divergences, particularly our newly introduced approach that allows for a locally non-convex divergence-generating function. In our specific case, where $f(x) = x\log(x)^2$, we choose $g_f(x) = -1 + \exp(v)$, to ensure the the output lies in $[-1, \infty)$. To train the generative adversarial network, we slightly adjust the approach in Radford et al. (2015). The specific adjustments, configurations, and algorithm can be found in the appendix.

5.2 Discussion of Results

To evaluate the performance of each GAN architecture in terms of estimating a heavy-tailed distribution, we estimate the shape and scale parameters of the generated distribution. These estimates are compared to the true parameters to assess each model's ability to capture the distribution's heavy-tailed nature.

Table 3: Comparison of GAN architectures with respect to shape and scale parameters.

Name	$\mathrm{shape} = 0.5$	$\mathrm{scale} = 1$
BHSGAN	0.5049 ± 0.0104	1.0937 ± 0.0227
Wasserstein	0.4220 ± 0.0071	1.6670 ± 0.0127
GAN	0.1952 ± 0.0028	1.3422 ± 0.0130
Pearson	0.6181 ± 0.0240	0.6863 ± 0.0183
KL	0.9951 ± 0.0167	1.0126 ± 0.0207
Reverse KL	0.8949 ± 0.0178	0.7376 ± 0.0120

Table 3 indicates that BHSGAN approximates the true shape parameter of the lognormal distribution more accurately than every other architecture. While the KL GAN slightly outperforms the BHS GAN in scale estimation, KL GAN consistently overestimates the shape parameter, with an average value nearly twice the true parameter. Thus, BHSGAN emerges as a robust choice for generating heavy-tailed distributions like the lognormal distribution.

In addition to the descriptive analysis, we provide and discuss a histogram and KDE estimate for each GAN architecture in Appendix A.2, which allows for a visual comparison of the different generated distributions.

However, the performance of the BHSGAN on the CIFAR10 dataset can be assessed by comparing the Fréchet Inception Distance (FID) (Heusel et al., 2017) of generated samples to different f-GAN and Wasserstein GAN architectures. Table 4 summarizes FID scores of the tested architectures.

Table 4: Mean FID Scores and two times the standard deviation for different architectures on the CIFAR10 data set over 100 seeds. Lower FID scores indicate better model performance.

Model	FID
BHSGAN	31.26 ± 0.08
KL GAN	37.50 ± 0.13
Reverse KL GAN	85.27 ± 0.19
Pearson GAN	33.60 ± 0.38
GAN	33.60 ± 0.10
WGAN	30.81 ± 0.12

The BHSGAN outperforms all f-GAN variants, proving its effectiveness in generating higher-quality and more diverse images. While the WGAN achieves a marginally better FID score, the difference is negligible and unlikely to be conclusive. Moreover, our approach guarantees that the approximated and true distributions remain closely aligned in the Bayes Hilbert space.

Upon inspecting Figure 2, the BHSGAN architecture stands out among the other f-GAN architectures by having the most refined picture structures. The KL, Reverse KL, and Pearson GAN produce less recognizable shapes using the same training procedure as the others.

As shown in Table 4, the WGAN and BHSGAN architectures produce the most recognizable samples. We can therefore conclude, that the BHSGAN outperforms traditional f-GANs and is on par with the Wasserstein architecture.

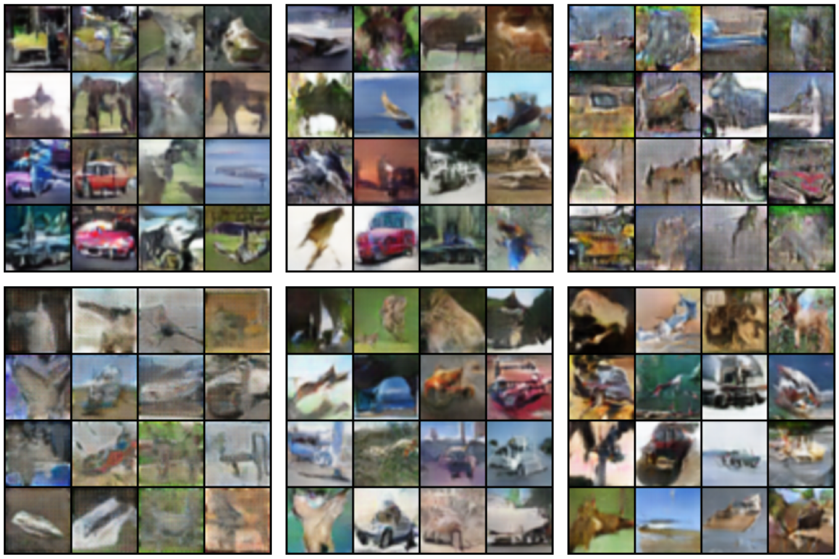

Figure 2: Experimental results for the CIFAR10 data set: Top row, from left to right: BHSGAN, KLGAN, and Reverse KLGAN. Bottom row, from left to right: Pearson GAN, original GAN, and Wasserstein GAN. Samples were drawn from the generators trained on the CIFAR10 dataset.

6 Conclusion

In this paper, we presented a novel framework for generalizing f-divergences, wherein we proved that the divergence-generating function of an f-divergence does not need to be convex on the whole domain, allowing for an even larger class of divergence-generating functions. We then showed, that there is a direct connection between the metric induced by the Bayes Hilbert space inner product and a pseudo f-divergence with pseudo-divergence-generating function f_{BHS}.

This connection of f-divergences and Bayes Hilbert spaces poses an interesting future direction as it circumvents one of the drawbacks of Bayes Hilbert space learning: learning from samples from distributions. Additionally, our framework enables sampling from estimated Bayes Hilbert space distributions in high dimensional settings, a feature that functional methods of estimation lack by design due to the curse of dimensionality. The framework therefore poses an interesting approach for sampling from posterior distributions in Bayesian statistics, which has to be addressed in detail in future research.

Since the primary goal of this publication is to extend the theoretical framework of f-divergences and bridge the gap to Bayes Hilbert spaces, we want to improve GAN performance in future work. Performance in terms of FID-Scores could be improved by considering a more advanced training procedure like progressively growing GAN (Karras et al., 2017) or state-of-the-art architectures like StyleGAN (Karras et al., 2019) and VQGAN (Esser et al., 2021). The proposed framework can also easily be implemented as an addition to existing f-GAN implementations as it only requires defining the new pseudo-divergence generating function f_{BHS} and adjusting the output activation function g_f specific to the domain of the generator.

7 Reproducibility Statement

We have taken several steps to ensure the reproducibility of our work. In Section 5, we provide detailed explanations of the framework and model architectures, with pseudocode for the key algorithm included in A.3. Proofs of technical results can be found in Section A.1.

References

J. Aitchison. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Springer Netherlands, 1986. ISBN 9780412280603.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Yoshiyuki Arata. A functional linear regression model in the space of probability density functions. Technical report, Research Institute of Economy, Trade and Industry (RIETI), 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of machine learning research*, 23(39):1–70, 2022.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. *Advances in Neural Information Processing Systems*, 32, 2019.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, pp. 698–706. PMLR, 2018.

Ziyu Chen, Hyemin Gu, Markos A Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Learning heavy-tailed distributions with wasserstein-proximal-regularized α-divergences. *arXiv preprint arXiv:2405.13962*, 2024.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

J. J. Egozcue, J. L. Díaz-Barrero, and V. Pawlowsky-Glahn. Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182, Jul 2006. ISSN 1439-7617. doi: 10.1007/s10114-005-0678-2.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

Jaime Roquero Gimenez and James Zou. A unified f-divergence framework generalizing vae and gan. *arXiv preprint arXiv:2205.05214*, 2022.

Jean-Pierre Gossez, EJ Lami Dozo, Jean Mawhin, and Lucien Waelbroeck. *Nonlinear Operators and the Calculus of Variations: Summer School Held in Bruxelles, 8-9 September 1975*, volume 543. Springer, 2006.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Amor Keziou. Dual representation of φ-divergences and applications. *Comptes rendus mathématique*, 336(10):857–862, 2003.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. URL `http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf`.

Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*. John Wiley & Sons, 2013.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Jitka Machalova, Karel Hron, and Gianna Serafina Monti. Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, 43(8):1419–1435, 2016.

Jitka Machalová, Renáta Talská, Karel Hron, and Aleš Gába. Compositional splines for representation of density functions. *Computational Statistics*, 36:1031–1064, 2021.

Eva-Maria Maier, Almond Stocker, Bernd Fitzenberger, and Sonja Greven. Additive density-on-scalar regression in bayes hilbert spaces with an application to gender economics. 2021.

Kevin R Moon and Alfred O Hero. Ensemble estimation of multivariate f-divergence. In *2014 IEEE International Symposium on Information Theory*, pp. 356–360. IEEE, 2014.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. On surrogate loss functions and f-divergences. 2009.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Nikolas Nüsken and Lorenz Richter. Solving high-dimensional hamilton–jacobi–bellman pdes using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *Partial Differential Equations and Applications*, 2(4):48, Jun 2021. ISSN 2662-2971. doi: 10.1007/s42985-021-00102-x. URL `https://doi.org/10.1007/s42985-021-00102-x`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Lorenz Richter and Julius Berner. Improved sampling via learned diffusions. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=h4pNROsO06`.

Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco J. R. Ruiz, and Ömer Deniz Akyildiz. Vargrad: a low-variance gradient estimator for variational inference. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O Tolstikhin. Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550 – 1599, 2012. doi: 10.1214/12-EJS722.

R. Talská, A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová. Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85, 2018. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2018.01.018.

K Van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. Bayes linear spaces. *SORT: statistics and operations research transactions, 2010, vol. 34, núm. 4, p. 201-222*, 2010.

K Gerald Van den Boogaart and Raimon Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

Karl Gerald Van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. Bayes hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

George Wynne. Bayes hilbert spaces for posterior approximation. *arXiv preprint arXiv:2304.09053*, 2023.

Yannis G Yatracos. Limitations of the wasserstein mde for univariate data. *Statistics and Computing*, 32(6):95, 2022.

A Appendix / supplemental material

A.1 Proofs of Theoretical Results

In this section, we provide proofs of several results we introduced throughout the publication.

While the following Lemma does not concern BHSGANs directly, we noticed that this connection has not yet been drawn in previous publications.

Proof of Lemma 3.1.

$$\begin{aligned} d_{B^2(\lambda)}(p_\mu, p_\nu)^2 &= \langle p_\mu \ominus p_\nu, p_\mu \ominus p_\nu \rangle_{B^2(\lambda)} \\ &= \langle \operatorname{clr}(p_\mu \ominus p_\nu), \operatorname{clr}(p_\mu \ominus p_\nu) \rangle_{L^2(\lambda)} \\ &= \int_\Omega \log(p_\mu(x) \ominus p_\nu(x))^2 \lambda(\mathrm{d}x) - \left(\int_\Omega \log(p_\mu(x) \ominus p_\nu(x)) \lambda(\mathrm{d}x) \right)^2 \\ &= \operatorname{Var}_\lambda(\log(\mu \ominus \nu)). \end{aligned}$$

□

The proof of Lemma 4.1 is straightforward but shows nevertheless the importance of the restriction to functions f that are unbounded on the convex part of the domain.

Proof of Lemma 4.1.

1. (a) Let $t^* \in M$. Then $\hat{x} := \hat{x}(t^*) = \operatorname*{argmax}_{x \in \operatorname{dom}(f)} (xt^* - f(x)) \in [a, \infty)$. Since $\hat{x} \in [a, \infty)$, the function $x \mapsto xt^* - f(x)$ is concave in $\hat{x}$ and monotonically decreasing for any $x > \hat{x}$. Thus, $\lim_{x \to \infty} xt^* - f(x) = -\infty$, i.e., $\operatorname*{argmin}_{x \in \operatorname{dom}(f)} xt^* - f(x) \in [a, \infty)$ implying that $t^* \notin N$. Analogously, $t_* \in N \implies t_* \notin M$.

   (b) In order to show that $M \cup N = \operatorname{dom}(f_*^*)$, let $\tilde{t} \in \operatorname{dom}(f_*^*)$. Then, there exists a $x_{\tilde{t}}$ such that $f_*^*(\tilde{t}) = x_{\tilde{t}}\tilde{t} - f(x_{\tilde{t}})$. Since $x_{\tilde{t}} \in \operatorname{dom}_f$, either $x_{\tilde{t}} \in [0, a)$ or $[a, \infty)$. As $\tilde{t}$ was chosen arbitrarily, $M \cup N = \operatorname{dom}(f_*^*)$ holds.

2. Since M and N are disjoint, we can write

$$\begin{aligned} f_{**}^{**} &= \sup_{x \in M} \{tx - f_*^*(x)\} \mathbb{1}_{\{t \in M^*\}} + \inf_{x \in N} \{tx - f_*^*(x)\} \mathbb{1}_{\{t \in N^*\}} \\ &= \sup_{x \in M} \{tx - f^*(x)\} \mathbb{1}_{\{t \in M^*\}} + \inf_{x \in N} \{tx - f_*(x)\} \mathbb{1}_{\{t \in N^*\}} \\ &= f \mathbb{1}_{\{t \in M^*\}} + f \mathbb{1}_{\{t \in N^*\}} \\ &= f. \end{aligned}$$

□

In Lemma 4.2 we restrict ourselves to the class of functions that have a nonnegative domain and range. The reason for this restriction is simply to ensure that the introduced measure of dissimilarity respects homogeneity.

Proof of Lemma 4.2. Let $\mu, \nu \in B^2(\lambda)$. Then,

$$\begin{aligned} D_f(\mu, \nu) &= \int_\Omega p_\nu(x) f\left(\frac{p_\mu(x)}{p_\nu(x)}\right) \lambda(\mathrm{d}x) \\ &= \int_\Omega \underbrace{p_\mu(x)}_{\geq 0} \underbrace{f\left(\frac{p_\mu(x)}{p_\nu(x)}\right)}_{\geq 0} \lambda(\mathrm{d}x) \\ &\geq 0 \end{aligned}$$

and

$$D_f(\mu,\nu) = \int_\Omega f\left(\frac{p_\mu(x)}{p_\nu(x)}\right)\nu(\mathrm{d}x) \overset{!}{=} 0 \iff \mu = \nu \quad \lambda\text{-a.s.}$$

□

In order to prove Theorem 4.3, we have to resort to a more technical argument involving the fundamental lemma of the calculus of variations to ensure that we can indeed derive an analytical representation for the optimizer of equation 3.

Proof of Theorem 4.3. In order to optimize equation 3, we have to maximize

$$(\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T))\, \mathbb{1}_{\{T \in C(M^*)\}}$$

and minimize

$$(\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T))\mathbb{1}_{\{T \in C(N^*)\}}$$

with respect to some $T \in C(\mathrm{dom}(f_*^*))$. Consider the term

$$\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T).$$

It can be rewritten as

$$\begin{aligned}\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^* \circ T) &= \int_0^\infty T(x)\mu(\mathrm{d}x) - \int_0^\infty (f_*^* \circ T)(x)\nu(\mathrm{d}x) \\ &= \underbrace{\int_0^\infty T(x)p_\mu(x) - (f_*^* \circ T)(x)p_\nu(x)\lambda(\mathrm{d}x)}_{:=E(T)}.\end{aligned}$$

Then, for an optimizer of equation 3

$$\left.\frac{\partial}{\partial\varepsilon}\right|_{\varepsilon=0} E(T+\varepsilon U) \overset{!}{=} 0$$

with $U \in C_0^\infty$ is a necessary condition for an extremum (Gelfand et al., 2000, Theorem 4.1).

Now,

$$E(T+\varepsilon U) = \int_0^\infty \big[(T+\varepsilon U)(x)p_\mu(x) - f_*^* \circ (T+\varepsilon U)(x)p_\nu(x)\big]\lambda(\mathrm{d}x)$$

and

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}E(T+\varepsilon U) &= \frac{\partial}{\partial\varepsilon}\int_0^\infty \big[(T+\varepsilon U)(x)p_\mu(x) - f_*^* \circ (T+\varepsilon U)(x)p_\nu(x)\big]\lambda(\mathrm{d}x) \\ &= \int_0^\infty \big[U(x)p_\mu(x) - (f_*^*)'(T(x)+\varepsilon U(x))U(x)p_\nu(x)\big]\lambda(\mathrm{d}x) \\ &= \int_0^\infty U(x)\big[p_\mu(x) - (f_*^*)'(T(x)+\varepsilon U(x))p_\nu(x)\big]\lambda(\mathrm{d}x) \\ &\overset{\varepsilon=0}{=} \int_0^\infty U(x)\big[p_\mu(x) - (f_*^*)'(T(x))p_\nu(x)\big]\lambda(\mathrm{d}x).\end{aligned} \tag{7}$$

Applying the fundamental lemma of the calculus of variations (Gelfand et al., 2000, Lemma 1.1) to equation 7 yields

$$\int_0^\infty U(x)\big[p_\mu(x) - (f_*^*)'(T(x))p_\nu(x)\big]\lambda(\mathrm{d}x) = 0 \implies p_\mu(x) - (f_*^*)'(T(x))p_\nu(x) = 0 \quad \text{a.e..}$$

Thus,

$$T(x) = \left(\left(f_*^{*\prime}\right)^{-1} \circ \frac{p_\mu}{p_\nu}\right)(x)$$

is a candidate for an optimizer. Note, that if $(f_*^{*\prime})^{-1}$ is well-defined, then $(f_*^{*\prime})^{-1}(x) = f'(x)$ (cf. Rockafellar (1970)). For the second derivative, we have

$$\frac{\partial^2}{\partial \varepsilon^2}\bigg|_{\varepsilon=0} E(T+\varepsilon U) = \int_0^\infty -(f_*^*)''(T(x))\underbrace{U(x)^2 p_\nu(x)}_{\geq 0}\lambda(\mathrm{d}x).$$

Depending on the properties of f_*^* the candidate

$$\tilde{T}(x) = \left(\left(f_*^{*\prime}\right)^{-1} \circ \frac{p_\mu}{p_\nu}\right)(x)$$

is thus a maximizer or minimizer. The restriction in the maximization part of equation 3 enforces that the optimizer $\tilde{T}$ is convex, implying that

$$(f_*^*)'' \circ \left(\left(f_*^{*\prime}\right)^{-1} \circ \frac{p_\mu}{p_\nu}\right)(x) \geq 0$$

and therefore

$$\frac{\partial^2}{\partial \varepsilon^2}\bigg|_{\varepsilon=0} E(T+\varepsilon U) \leq 0.$$

Conversely, to optimize the minimization part of equation 3, we consider the restriction to the concave part of f_*^*, implying that

$$(f_*^*)'' \circ \left(\left(f_*^{*\prime}\right)^{-1} \circ \frac{p_\mu}{p_\nu}\right)(x) \leq 0$$

and therefore

$$\frac{\partial^2}{\partial \varepsilon^2}\bigg|_{\varepsilon=0} E(T+\varepsilon U) \geq 0.$$

□

Corollary 4.4 plays an important role when considering the computational perspective of equation 3. Since we wanted to avoid optimizing a *min max* and *min min* game simultaneously, we derived a lower bound for the pseudo f divergence that can be solved as a simple *min max* game instead. The core idea is to rewrite the infimum as a supremum and combine both suprema to obtain a *min max* game framework.

Proof of Corollary 4.4.

$$\begin{aligned}
D_f(\mu,\nu) &= \int_\Omega p_\nu(x) f(\frac{p_\mu(x)}{p_\nu(x)})\lambda(\mathrm{d}x) \\
&= \int_\Omega \sup_{t\in M}\left\{t\frac{p_\mu(x)}{p_\nu(x)} - f_*^*(t)\right\}\lambda(\mathrm{d}x) + \int_\Omega \inf_{t\in N}\left\{t\frac{p_\mu(x)}{p_\nu(x)} - f_*^*(t)\right\}\lambda(\mathrm{d}x) && (8)\\
&\geq \sup_{T\in C(M^*)}\left\{\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^*\circ T)\right\} + \int_\Omega \inf_{T\in C(N^*)}\left\{t\frac{p_\mu(x)}{p_\nu(x)} - f^*(t)\right\}\lambda(\mathrm{d}x) && (9)\\
&= \sup_{T\in C(M^*)}\left\{\,(\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^*\circ T))\,\right\} - \sup_{T\in C(N^*)}\left\{(\mathbb{E}_\mu(T) - \mathbb{E}_\nu(f_*^*\circ T))\right\} && (10)\\
&= \sup_{T\in C(\mathrm{dom}(f_*^*))}\Big\{\mathbb{E}_\mu(T(\mathbb{1}_{\{T\in C(M^*)\}} - \mathbb{1}_{\{T\in C(N^*)\}}))- \\
&\qquad\qquad \mathbb{E}_\nu((f_*^*\circ T)(\mathbb{1}_{\{T\in C(M^*)\}} - \mathbb{1}_{\{T\in C(N^*)\}}))\Big\} \\
&= \sup_{\bar{T}\in C(M^*)\cup C(N^*)}\left\{\mathbb{E}_\mu(\bar{T}) - \mathbb{E}_\nu(f_*^*\circ\bar{T})\right\} && (11)
\end{aligned}$$

where $C(M^*)$ ($C(N^*)$) denotes the set of continuous functions such that $f_*^* \circ T$ is convex (concave). The second equality, equation 8, can be derived by applying the definition of the mixed conjugate.

equation 9 is derived in the same way as equation 1 and equation 10 is an application of Gossez et al. (2006) where we used that $-\inf = \sup$. In equation 11 we used the fact, that $C(M^*) \cap C(N^*) = \emptyset$. Defining $\bar{T} = T(\mathbb{1}_{\{T \in C(M^*)\}} - \mathbb{1}_{\{T \in C(N^*)\}})$ yields the desired result.

□

Proof of Lemma 4.5.

$$\begin{aligned}
D_{f_{BHS}}(\mu, \nu) &= \int_\Omega p_\nu(x) \frac{p_\mu(x)}{p_\nu(x)} \log\left(\frac{p_\mu(x)}{p_\nu(x)}\right)^2 \lambda(\mathrm{d}x) \\
&= \int_\Omega p_\mu(x) \log\left(\frac{p_\mu(x)}{p_\nu(x)}\right)^2 \lambda(\mathrm{d}x) \\
&= \mathbb{E}_\mu \left(\log(\mu \ominus \nu)^2\right) \\
&= d_{B^2(\mu)}(p_\mu, p_\nu)^2 + \mathbb{E}_\mu \left(\log(\mu \ominus \nu)\right)^2 .
\end{aligned}$$

□

A.2 Additional Experiments

In this section we present generated MNIST digits as well as histograms and KDEs for an estimated lognormal distribution for each benchmarking GAN architecture.

The findings of Table 3 can be validated further by considering Figure 3. BHS GAN approximates the true density well regarding mode and tail behavior. In comparison, the Wasserstein architecture fails to capture the left tail behavior completely, and the right tail shows some discrepancies in terms of monotonicity. While the KL GAN captures the true distributions' scale relatively well, the right tail seems to be much heavier than the tail of the true underlying distribution. Pearson GAN and Reverse KL GAN struggle with oscillations and deviations from the true density, especially in the tails. The basic GAN performs poorly in terms of alignment, with significant underestimation near the mode and noisy approximation in the tails.

The generated MNIST samples in Figure 4 visibly depict digits between 0 and 9, indicating that all GAN architectures learned the underlying distribution of the images well. Therefore we can conclude that the BHSGAN is on par with the other GAN architecture for the MNIST dataset. We limit ourselves to this high-level evaluation as a quantitative comparison for generative models on this kind of data is not meaningful (Nowozin et al., 2016).

A.3 Practical considerations and algorithms

During the experiments we found that training (pseudo) f-GANs is highly unstable, suffering from exploding and diminishing losses. To stabilize the training process we implemented two methods. Firstly, we introduced a gradient penalty identical to the one used for WGANs (Gulrajani et al., 2017) for all used GANs except the vanilla and reverse KL-GAN. Additionally, contrary to the popular approach of training the discriminatory model first, we decided to start by training the generative model first. We interpret this finding as a (prior) practical drawback of f-GANs, in the way that in theory, the supremum has to be found over all continuous functions, but in practice, this class of functions is too general to be learned during training. This means that if the discriminator is allowed to be too flexible it starts ”remembering” real images before the generator can generate good fake ones, resulting in exploding losses for the discriminator and vanishing ones for the generator. Accordingly, forcing the discriminator to be Lipschitz-continuous constrains the supremum and infimum in 3 to be found over all Lipschitz-continuous functions, which stabilizes the training process by hindering the discriminator from remembering real images too fast.

All experiments were conducted on an AORUS RTX 4090 eGPU 24GB.

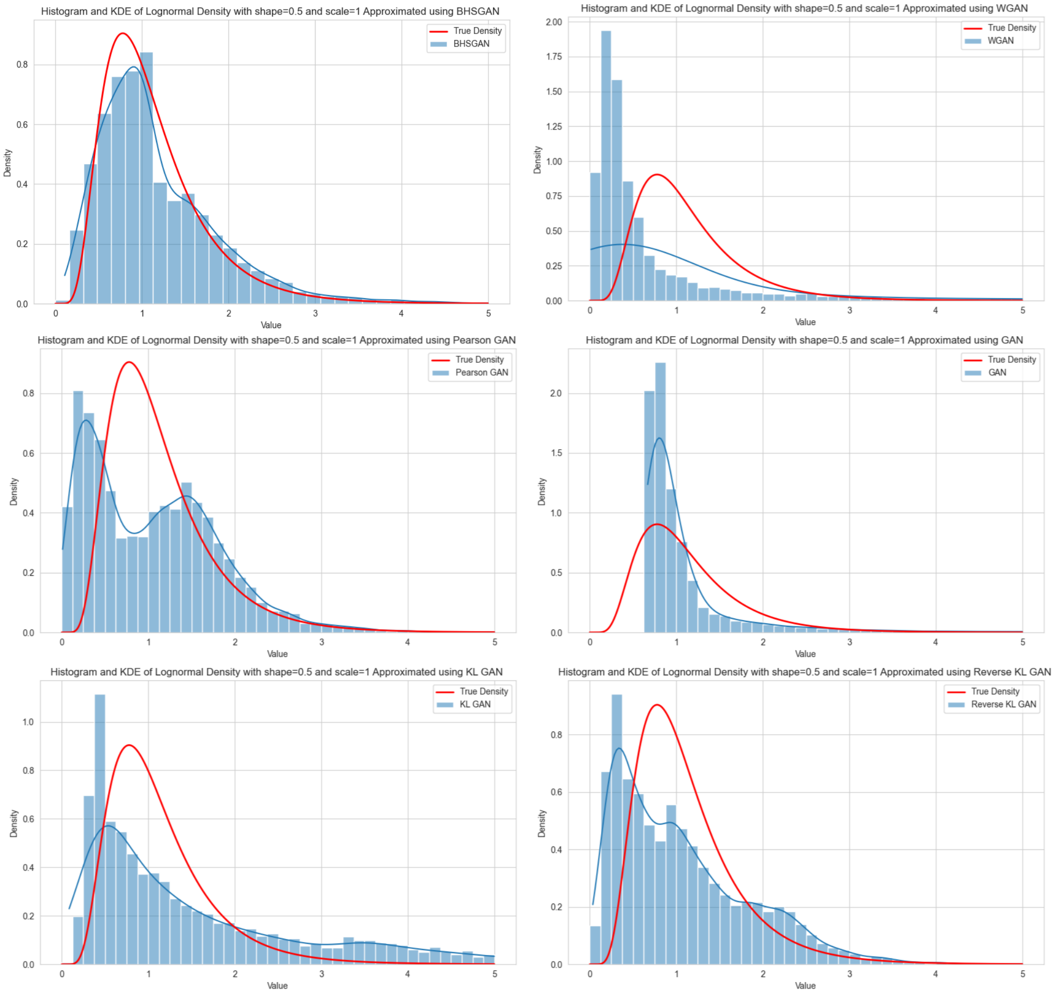

Figure 3: Histograms and KDE estimates for different GAN architectures

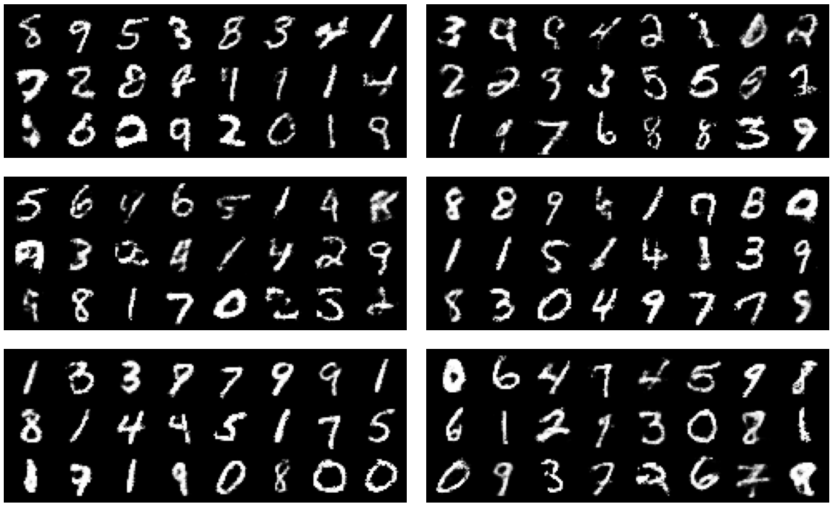

Figure 4: Experimental results for MNIST datasets: Left column, from top to bottom: BHSGAN, KLGAN, and Reverse KLGAN. Right column, from top to bottom: Pearson GAN, original GAN, and Wasserstein GAN. Samples were drawn from the generators trained on the MNIST dataset.

Algorithm 1 Default values for the training procedure in this paper were $\alpha_g = \alpha_d = 0.0002$ for all GANs except BHSGAN and REVKLGAN, where $\alpha_g = \alpha_d = 0.00005$, $n_d = 2$, $n_g = 2$, $m = 128$ for MNIST and $m = 64$ for CIFAR, $N_{\text{epochs}} = 50$, $d_z = 100$ and $\lambda = 10$ except for PEARSONGAN, where $\lambda = 20$

```
Input: $\alpha_g$, learning rate of generative model. $\alpha_d$, learning rate of discriminatory model. $n_d$, number of discriminator updates. $n_g$, number of generator updates. $m$, batch size. $N_{\text{epochs}}$, number of epochs. $d_z$, dimension of noise, $\lambda$, gradient penalty coefficient, $gp$, gradient penalty, $w_0$, initial parameters of discriminatory model $\vartheta_0$, initial parameters of generative model.
for $k = 1, ..., N_{\text{epochs}}$ do
  for each batch $= \{x^{(1)}, ..., x^{(m)}\}$ in training set do
    noise $= \{z^{(1)}, ..., z^{(m)}\} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{d_z \times d_z})$
    real_sample $= \{x^{(1)}, ..., x^{(m)}\}$
    for $n = 1, ..., n_g$ do
      $\mathcal{L}_{\nu_\vartheta} = -\frac{1}{m} \sum_{i=1}^{m} f_*^* \circ g_f \circ V_\omega \circ \nu_\vartheta(z^{(i)})$
      $\vartheta = \vartheta - \text{Adam}(\alpha_g, \vartheta, \mathcal{L}_{\nu_\vartheta})$
    end for
    for $n = 1, ..., n_d$ do
      $\mathcal{L}_{V_\omega} = \frac{1}{m} \sum_{i=1}^{m} \big(f_*^* \circ g_f \circ V_\omega \circ \nu_\vartheta(z^{(i)}) -$
                    $g_f \circ V_\omega(x^{(i)})\big) + \lambda \cdot gp$
      $\omega = \omega + \text{Adam}(\alpha_d, \omega, \mathcal{L}_{V_\omega})$
    end for
  end for
end for
```