Bridging the Gap Between f-GANs and Bayes Hilbert Spaces

Linus Lach*, Alexander Fottner, Yarema Okhrin

Statistik
& Data Science
Prof. Dr. Yarema Okhrin

Content

- f-GANs
- Bayes Hilbert Spaces
- Mixed conjugates and pseudo f-divergences
- Bayes Hilbert space GANs
- References

f-GANs

f-Divergences (Csiszár, Shields, et al. 2004)

Quantify dissimilarity between two probability measures μ and ν that are absolutely continuous w.r.t to some σ-finite base measure λ.[1]

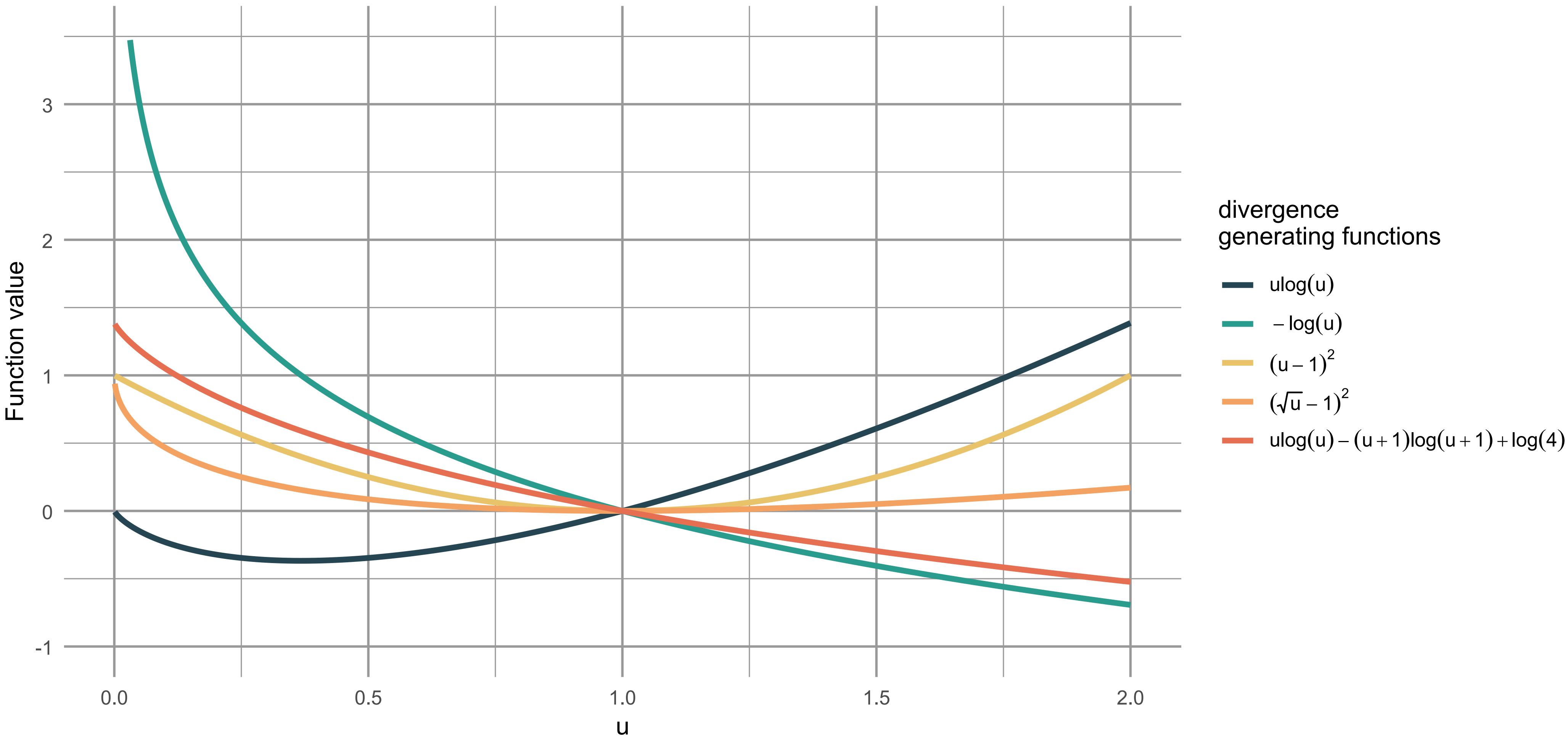
$$D_f(\mu, \nu) = \int_{\mathbb{R}} p_\nu(x) f\left(\frac{p_\mu(x)}{p_\nu(x)}\right) \lambda(\mathrm{d}x).$$

- $f : \mathbb{R}_+ \to \mathbb{R}$ convex,
- lower-semicontinuous, and
- satisfying $f(1) = 0$.

f is called **divergence generating function**

1. μ is absolutely continuous w.r.t. λ if every μ nullset is a λ nullset and vice versa

Examples of f-divergences generating functions



Estimating f-divergences is hard!

- Nonparametric estimation
- Only finite samples available
- Highdimensional setting

Solution: Find easy to estimate lower bound (Nguyen, Wainwright, and Jordan 2010):

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

Components of the lower bound

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

- $\mathcal{T}$ arbitrary class of measurable functions $T : \Omega \to \mathrm{dom}(f^*)$
- Fenchel conjugate f^* of f: $f^*(y) := \sup_{x \in \mathrm{dom}(f)} \{xy - f(x)\}$

Statistik & Data Science
Prof. Dr. Yarema Okhrin

Fenchel conjugates

Theorem (Rockafellar (1970))

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

This theorem also works for concave functions:

Theorem (Rockafellar (1970))

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a concave function. If f is upper semi-continuous, then the duality $f_{**}(x) = f(x)$ for all $x \in \mathbb{R}^n$ holds.

Here,

$$f_*(y) := \inf_{x \in \mathrm{dom}(g)} \{xy - f(x)\}.$$

Optimizing the lower bound

If f is not only convex and lower semicontinuous but $f \in \mathcal{C}^1$, then the bound

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

is tight and the supremum is attained at

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

Generative Adversarial Networks (GANs) in a nutshell

- Sample from an unknown distribution
- Alternately:
  - Train a generative model that creates samples from an unknown distribution
  - Train a discriminatory model (binary classifier) that incentives the generator to produce more realistic examples

Training objective for f-GANs

Nowozin, Cseke, and Tomioka (2016) extended the work of Goodfellow et al. (2014) to a generalized optimization problem:

$$\min_{\theta} \max_{\omega} \left[\mathbb{E}_{\mu}(T_{\omega}) - \mathbb{E}_{\nu_{\theta}}(f^* \circ T_{\omega}) \right]$$

θ (generator) and ω (discriminator) are each parameters of neural networks.

Bayes Hilbert Spaces

General idea

1. Construct a linear space for proportional σ-finite measures, including probability measures.
2. Consider the subspace of square-log-integrable densities.[1]
3. Define an inner product on this subspace using the centered log ratio for measures.
4. Obtain a Hilbert space structure that allows for a straight forward interpretation of distances between densities.

1. In this setting, measures can be identified with their corresponding Radon-Nikodym density.

Some definitions

- Define $\mathcal{M}(\lambda)$ as the set of measures on $(\Omega, \mathcal{B})$ that are equivalent to the base measure λ.
- Two measures $\mu, \nu \in \mathcal{M}(\lambda)$ are defined as B-equivalent denoted by $\mu =_{B(\lambda)} \nu$ if there exists a constant $c > 0$ such that $\mu(A) = c\nu(A)$ for all $A \in \mathcal{B}$. Then, $=_{B(\lambda)}$ is an equivalence relation on $\mathcal{M}(\lambda)$
- Finally, define $B(\lambda) := \mathcal{M}(\lambda)/ =_{B(\lambda)}$

Bayes Linear Space

For two measures $\mu, \nu \in B(\lambda)$ define the perturbation of μ by ν over some set $R \in \mathcal{B}$ as

$$(\mu \oplus \nu)(R) := \int_R \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \,\mathrm{d}\lambda$$

and the powering of μ by $\alpha \in \mathbb{R}$ as

$$(\alpha \odot \mu)(R) := \int_R \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^{\alpha} \mathrm{d}\lambda$$

Theorem (K. Van den Boogaart, Egozcue, and Pawlowsky-Glahn (2010))

$(B(\lambda), \oplus, \odot)$ is a linear space.

The centered log ratio

For $p \geq 1$ define

$$B^p(\lambda) = \left\{ \mu \in B(\lambda) : \int_{\mathbb{R}} \left| \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \right) \right|^p \mathrm{d}\lambda < +\infty \right\}.$$

The clr of $\mu \in B^p(\lambda)$ is defined as

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

Bayes Hilbert spaces

Theorem (K. G. Van den Boogaart, Egozcue, and Pawlowsky-Glahn (2014))

$B^2(\lambda)$ equipped with the inner product

$$\int \left(\log(f(x)) - \int \log(f(y))\lambda(\mathrm{d}y)\right) \left(\log(g(x)) - \int \log(g(y))\lambda(\mathrm{d}y)\right) \lambda(\mathrm{d}y)$$

denoted by $\langle f, g\rangle_{B^2(\lambda)}$ with f, g densities in $B^2(\lambda)$ is a separable Hilbert space.

Theorem (K. G. Van den Boogaart, Egozcue, and Pawlowsky-Glahn (2014))

$\mathrm{clr} : B^2(\lambda) \to L_0^2(\lambda)$ is an isometry of Hilbert spaces.

Mixed conjugates and pseudo f-divergences

Mixed conjugates

Consider $f:[0,\infty)\to\mathbb{R}$ continuous, satisfying $f(1)=0$, concave on some interval $[0,a)$, $a>0$, and convex on $[a,\infty)$. The mixed conjugate f_*^* of f is defined by

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

where

$$M := \left\{t\in \mathrm{dom}(f_*^*) : \operatorname*{argmax}_{x\in\mathrm{dom(f)}}(tx-f(x))\in[a,\infty)\right\},$$

$$N := \left\{t\in \mathrm{dom}(f_*^*) : \operatorname*{argmin}_{x\in\mathrm{dom(f)}}(tx-f(x))\in[0,a)\right\}.$$

Mixed conjugates and pseudo f-divergences

Lemma

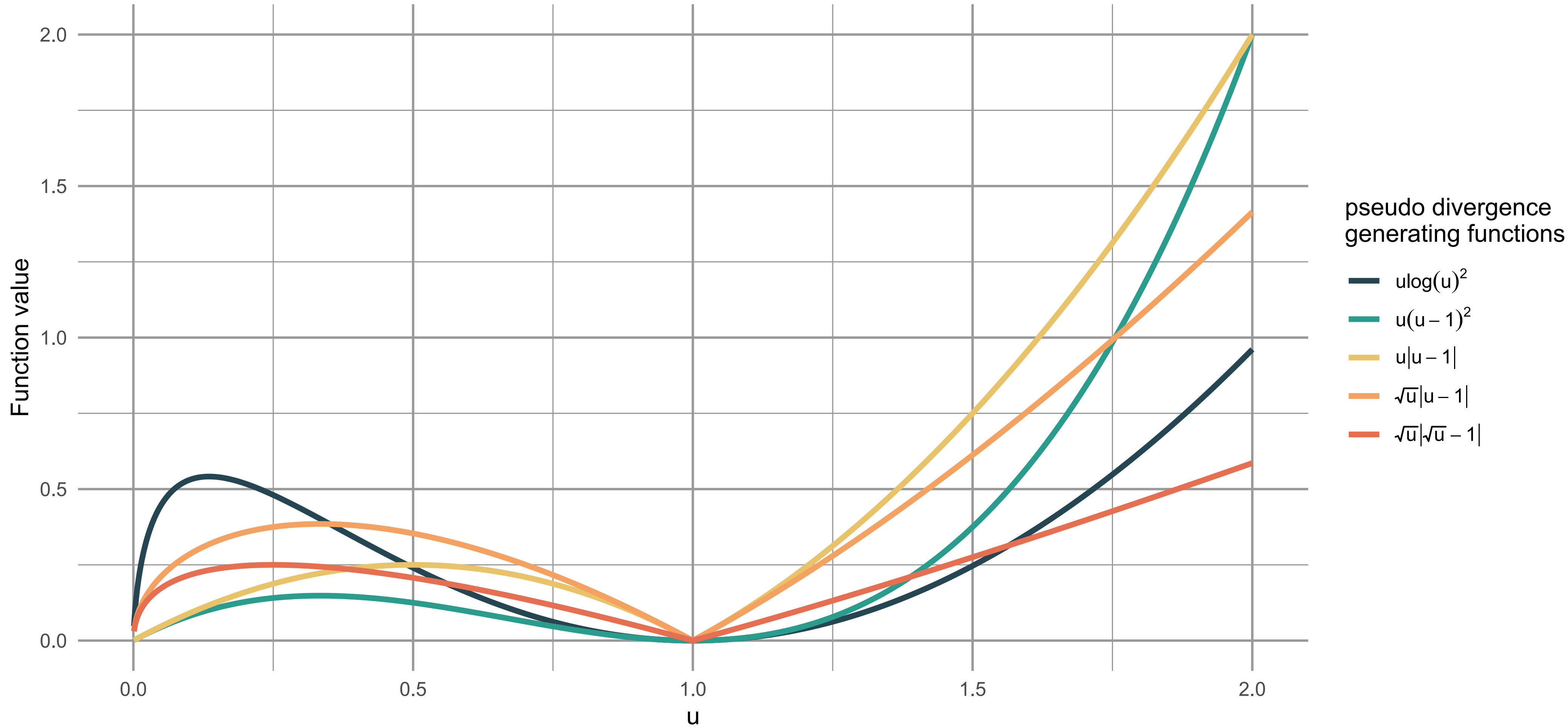
For any continuous function $f : [0, \infty) \to \mathbb{R}$ that is convex on $[a, \infty),\ a > 0$, concave on $[0, a)$, and satisfies $\lim_{x\to\infty} f(x) = +\infty$, the mixed conjugate satisfies:

1. $M \cap N = \emptyset$ and $M \cup N = \mathrm{dom}(f_*^*)$
2. $f_{**}^{**} = f$ for almost all $t \in \mathrm{dom}(f)$.

Lemma

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 $D_f(\mu, \nu) = 0 \iff \mu = \nu \quad \lambda - \mathrm{a.s.}$

Pseudo divergence generating functions



Lower bounds for pseudo f-divergences

Corollary

For any function f satisfying the conditions of the previous Lemma, the lower bound

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

holds.

Optimizing lower bounds for pseudo f-divergences

Theorem

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 well-defined. Furthermore, let $\mu,\nu\in B^2(\lambda)$ with $p_\mu(x):=\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x),\ p_\nu(x):=\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x)$. Then, $\tilde{T}(x):=f'\left(\frac{p_\mu(x)}{p_\nu(x)}\right)$ is an optimizer for

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

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

Bayes Hilbert space GANs

Bayes Hilbert space divergence

Consider the function

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

that is convcave on $[0, \exp(-1)]$ and convex for $x > \exp(-1)$.

Applying the previous lemma implies that f_{BHS} induces a pseudo f-divergence called f_{BHS}-divergence.

Connection to Bayes Hilbert spaces

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

Corollary

For $\tilde{T}(x) = f'_{BHS}\left(\frac{p_\mu(x)}{p_\nu(x)}\right)$ we have

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

Bayes Hilbert space GAN

Optimization problem:

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

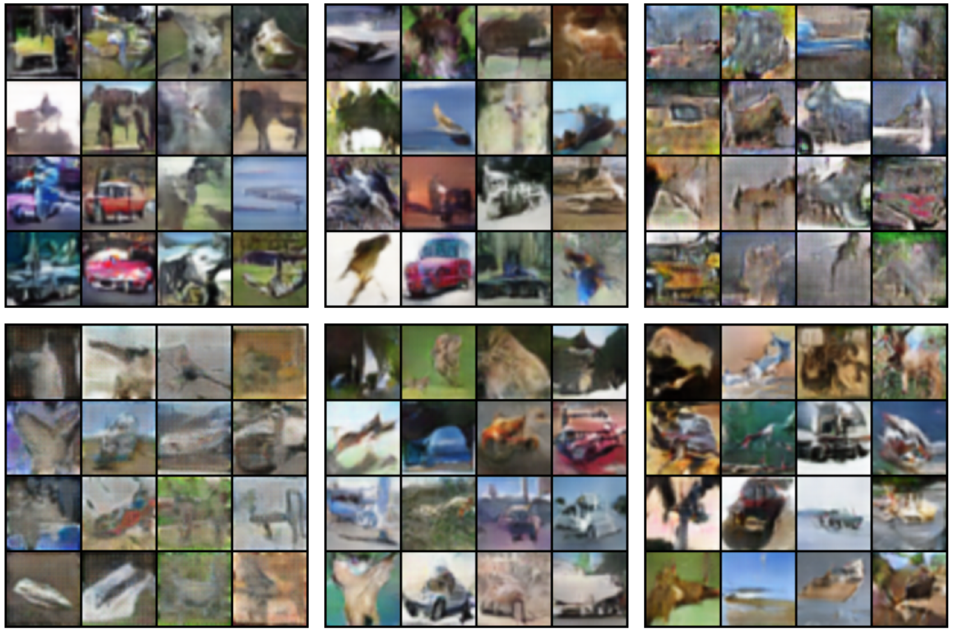
where

$$\bar{T}_{\omega}(x) = g_f \circ V_{\omega}(x).$$

V_{ω} denotes the discriminatory model with no restrictions on the output, i.e., $\mathrm{Im}(V_{\omega}) \subseteq \mathbb{R}$ and g_f a final output activation function depending on the domain of f_*^*.

Computational results

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



References

Csiszár, Imre, Paul C Shields, et al. 2004. “Information Theory and Statistics: A Tutorial.” *Foundations and Trends in Communications and Information Theory* 1 (4): 417–528.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Nets.” *Advances in Neural Information Processing Systems* 27.

Nguyen, XuanLong, Martin J Wainwright, and Michael I Jordan. 2010. “Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization.” *IEEE Transactions on Information Theory* 56 (11): 5847–61.

Nowozin, Sebastian, Botond Cseke, and Ryota Tomioka. 2016. “F-Gan: Training Generative Neural Samplers Using Variational Divergence Minimization.” *Advances in Neural Information Processing Systems* 29.

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

Van den Boogaart, Karl Gerald, Juan José Egozcue, and Vera Pawlowsky-Glahn. 2014. “Bayes Hilbert Spaces.” *Australian & New Zealand Journal of Statistics* 56 (2): 171–94.

Van den Boogaart, K., Juan José Egozcue, and Vera Pawlowsky-Glahn. 2010. “Bayes