On Relativistic f-Divergences

Alexia Jolicoeur-Martineau [1]

Abstract

We take a more rigorous look at Relativistic Generative Adversarial Networks (RGANs) and prove that the objective function of the discriminator is a statistical divergence for any concave function f with minimal properties ($f(0) = 0$, $f'(0) \neq 0$, $\sup_x f(x) > 0$). We devise additional variants of relativistic f-divergences. We show that the Wasserstein distance is weaker than f-divergences which are weaker than relativistic f-divergences. Given the good performance of RGANs, this suggests that Wasserstein GAN does not performs well primarily because of the weak metric, but rather because of regularization and the use of a relativistic discriminator. We introduce the minimum-variance unbiased estimator (MVUE) for Relativistic GANs and show that it does not perform better. We show that the estimator of Relativistic average GANs (RaGANs) is asymptotically unbiased and that the finite-sample bias is small; removing this bias does not improve performance.

1. Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are a very popular approach to approximately generate data from a complex probability distribution using only samples of data (without any information on the true data distribution). Most notably, it has been very successful at generating photo-realistic images (Karras et al., 2017; 2018). It consists in a game between two neural networks, the generator G and the discriminator D. The goal of D is to classify real from fake (generated) data. The goal of G is to generate fake data that appears to be real, thus "fooling" D into thinking that fake data is actually real.

There are many GAN variants and most of them consist of changing the loss function of D. To name a few: Standard GAN (SGAN) (Goodfellow et al., 2014), Least-Squares GAN (LSGAN) (Mao et al., 2017), Hinge-loss GAN (HingeGAN) (Miyato et al., 2018), Wasserstein GAN (WGAN) (Arjovsky et al., 2017).

For most GAN variants, training D is equivalent to estimating a divergence: SGAN estimates the Jensen–Shannon divergence (JSD), LSGAN estimates the Pearson χ^2 divergence, HingeGAN estimates the Reverse-KL divergence, and WGAN estimates the Wasserstein distance. Even more generally, f-GANs (Nowozin et al., 2016) estimate any f-divergence (which includes most of the popular divergences), while IPM-based GANs (Mroueh & Sercu, 2017) estimate any Integral probability metric (IPM) (Müller, 1997). Thus, intuitively, GANs can be thought of as estimating a diverge and then minimizing it (this is not technically correct; see Jolicoeur-Martineau (2018b)).

Recently, Jolicoeur-Martineau (2018a) showed that IPM-based GANs possess a unique type of discriminator which they call a Relativistic Discriminator (RD). They explained that one can construct f-GANs while using a RD and that doing so improves the stability of the training and quality of generated data. They called this approach Relativistic GANs (RGANs). They proposed two variants: Relativistic paired GANs (RpGANs)[1] and Relativistic Average GANs (RaGANs).

Jolicoeur-Martineau (2018a) provided mathematical and intuitive arguments as to why using a Relativistic Discriminator (RD) may be helpful. However, they did not prove that the loss functions are mathematically sensible. Furthermore, the estimators that they used are not the minimum-variance unbiased estimators (MVUE).

The contributions of this paper are the following:

1. We prove that the objective functions of the discriminator in RGANs are divergences (relativistic f-divergences).
2. We devise additional variants of Relativistic f-divergences.

[1] Mila, Université de Montréal . Correspondence to: Alexia Jolicoeur-Martineau <alexia.jolicoeur-martineau@mail.mcgill.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] We added the word "paired" to better distinguish the variant with paired real/fake data (originally called RGANs) and the general approach called Relativistic GANs (RGANs).

3. We show that the Wasserstein Distance is weaker than f-divergences which are weaker than relativistic f-divergences.

4. We present the minimum-variance unbiased estimator (MVUE) of RpGANs and show that using it hinders the performance of the generator.

5. We show that RaGANs are only asymptotically unbiased, but that the finite-sample bias is small. Removing this bias does not improve the performance of the generator.

2. Background

For the rest of the paper, we will refer to the "critic" $C(x)$ instead of the discriminator $D(x)$. The critic is the discriminator before applying the activation function ($D(x) = a(C(x))$, where a is an activation function and $C(x) \in \mathbb{R}$). Intuitively, the critic can be thought of as describing how realistic x is. In the case of SGAN and HingeGAN, a large $C(x)$ means that x is realistic, while a small $C(x)$ means that x is not realistic. We use this notation because Relativistic GANs are defined in terms of the critic rather than the discriminator.

2.1. Generative Adversarial Networks

GANs can be defined very generally in the following way:

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim\mathbb{P}}\left[f_1(C(x))\right] + \mathbb{E}_{y\sim\mathbb{Q}}\left[f_2(C(y))\right], \tag{1}$$

$$\sup_{G:Z\to\mathcal{X}} \mathbb{E}_{x\sim\mathbb{P}}\left[g_1(C(x))\right] + \mathbb{E}_{z\sim\mathbb{Z}}\left[g_2(C(G(z)))\right], \tag{2}$$

where f_1, f_2, g_1, g_2 : $\mathbb{R} \to \mathbb{R}$, $\mathbb{P}$ is the distribution of real data with support $\mathcal{X}$, $\mathbb{Z}$ is the latent distribution (generally a multivariate normal distribution), $C(x)$ is the critic evaluated at x, $G(z)$ is the generator evaluated at z, and $G(z) \sim \mathbb{Q}$, where $\mathbb{Q}$ is the distribution of fake data. See Brock et al. (2018) for details on how different choices of $\mathbb{Z}$ performs. The critic and the generator are generally trained with stochastic gradient descent (SGD) in alternating steps.

Most GANs can be separated in two classes: non-saturating and saturating loss functions. GANs with the saturating loss are such that g_1=$-f_1$ and g_2=$-f_2$, while GANs with the non-saturating loss are such that g_1=f_2 and g_2=f_1. In this paper, we will assume that the non-saturating loss is used as it generally works best in practice (Goodfellow et al., 2014) (Nowozin et al., 2016). Note that g_1 generally has no impact on training since its gradient with respect to G is zero; we can thus ignore it.

Although not always the case, the most popular GAN loss functions (SGAN, LSGAN with labels -1/1, HingeGAN, WGAN) are symmetric (i.e., $f_2(x) = f_1(-x)$). For simplicity, in this paper, we restrict ourselves to symmetric loss functions.

Non-saturating Symmetric GANs (SyGANs) can be represented more simply as:

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim\mathbb{P}}\left[f(C(x))\right] + \mathbb{E}_{y\sim\mathbb{Q}}\left[f(-C(y))\right], \tag{3}$$

$$\sup_{G:Z\to\mathcal{X}} \mathbb{E}_{z\sim\mathbb{Z}}\left[f(C(G(z)))\right], \tag{4}$$

for some function $f : \mathbb{R} \to \mathbb{R}$. For easier optimization, we generally want f to be concave with respect to the critic. This is the case in symmetric f-GANs.

In this paper, we restrict our relativistic divergences to symmetric cases with concave f. Although this may be somewhat constraining, not making these assumptions would be very problematic for GANs. By not assuming concavity, we could have an objective function that diverges to infinity (and thus an infinite divergence). This is particularly problematic for GANs because early in training, we expect $\mathbb{P}$ and $\mathbb{Q}$ to be perfectly separated (because of fully disjoint supports). This would cause the objective function to explode towards infinity and thereby causing severe instabilities. The Kullback–Leibler (KL) divergence is a good example of such a problematic divergence for GANs. If a single sample from the support of $\mathbb{Q}$ is not part of the support of $\mathbb{P}$, the divergence will be ∞. Also, note that the dual form of the KL divergence cannot be represented as a SyGAN with equation (3) since $f_1(x) = x$ and $f_2(x) = -e^{x-1}$ are not symmetric (Nowozin et al., 2016).

2.2. Integral Probability Metrics

Rather than using a concave function f to ensure a maximum on the objective function, IPM-based GANs instead force the critic to respect some constraint so that it does not grow too quickly. IPM-based GANs are defined in the following way:

$$\sup_{\substack{C:\mathcal{X}\to\mathbb{R}\\ C\in\mathcal{F}}} \mathbb{E}_{x\sim\mathbb{P}}\left[C(x)\right] - \mathbb{E}_{y\sim\mathbb{Q}}\left[C(y)\right], \tag{5}$$

$$\sup_{G:Z\to\mathcal{X}} \mathbb{E}_{z\sim\mathbb{Z}}\left[C(G(z))\right], \tag{6}$$

where $\mathcal{F}$ is a class of functions such that the IPM is not infinite. See Mroueh et al. (2017) for an extensive review of the choices of $\mathcal{F}$.

2.3. Relativistic GANs

Rather than training the critic on real and fake data separately, Relativistic GANs tries to maximize the critic's difference (CD). In Relativistic paired GANs (RpGANs), the CD is defined as $C(x)-C(y)$, while in Relativistic average

GANs (RaGANs), the CD is defined as $C(x) - \mathbb{E}_{y \sim \mathbb{Q}} C(y)$ (or vice-versa). The CD can be understood as how much more realistic real data is from fake data. The optimal size of the CD is determined by the choice of f. With a least-square loss, the CD must be exactly equal to 1. On the other hand, with a log-sigmoid loss, the CD is grown to around 2 or 3 (after-which the gradient of f vanishes to zero). This will be explained in more details in the next section. Again, we focus only on choices of f that have symmetry (as done with SyGANs).

Relativistic paired GANs (RpGANs) are defined in the following way:

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\substack{x\sim\mathbb{P}\\ y\sim\mathbb{Q}}} \left[f\left(C(x) - C(y) \right) \right], \tag{7}$$

$$\sup_{G:Z\to\mathcal{X}} \mathbb{E}_{\substack{x\sim\mathbb{P}\\ z\sim\mathbb{Z}}} \left[f\left(C(G(z)) - C(x) \right) \right]. \tag{8}$$

Relativistic average GANs (RaGANs) are defined in the following way:

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim\mathbb{P}} \left[f\left(C(x) - \mathbb{E}_{y\sim\mathbb{Q}} C(y) \right) \right] + \mathbb{E}_{y\sim\mathbb{Q}} \left[f\left(\mathbb{E}_{x\sim\mathbb{P}} C(x) - C(y) \right) \right], \tag{9}$$

$$\sup_{G:Z\to\mathcal{X}} \mathbb{E}_{z\sim\mathbb{Z}} \left[f\left(C(G(z)) - \mathbb{E}_{x\sim\mathbb{P}} C(x) \right) \right] + \mathbb{E}_{x\sim\mathbb{P}} \left[f\left(\mathbb{E}_{z\sim\mathbb{P}_z} C(G(z)) - C(x) \right) \right]. \tag{10}$$

3. Relativistic Divergences

We define statistical divergences in the following way:

Definition 3.1. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions and S be the set of all probability distributions with common support. A function $D : (S, S) \to \mathbb{R}_{>0}$ is a divergence if it respects the following two conditions:

$$D(\mathbb{P}, \mathbb{Q}) \geq 0$$
$$D(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}.$$

In other words, divergences are distances between probability distributions. The distribution of real data ($\mathbb{P}$) is fixed and our goal is to modify the distribution of fake data ($\mathbb{Q}$) so that the divergence decreases over time through the training process.

It is important to show that we use a divergence; this ensures that it is not possible to obtain a critic which cannot distinguish real from fake sample ($D(\mathbb{P}, \mathbb{Q}) = 0$) when the two distributions (real and fake) are not the same ($\mathbb{P} \neq \mathbb{Q}$). If we did not have a divergence, it could be possible to reach a situation where the generator cannot learn (since the critic returns the same value for real and fake samples) while the generator still isn't generating samples from the real distribution.

3.1. Main Theorem

As discussed in the introduction, in most GANs, the objective function of the critic at optimum is a divergence. We show that the objective function of the critic in RpGANs, RaGANs, and other variants also estimate a divergence. The theorem is as follows:

Theorem 3.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function such that $f(0) = 0$, f is differentiable at 0, $f'(0) \neq 0$, $\sup_x f(x) = M > 0$, and $\arg\sup_x f(x) > 0$. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions with support $\mathcal{X}$. Let $\mathbb{M} = \frac{1}{2}\mathbb{P} + \frac{1}{2}\mathbb{Q}$. Then, we have that

$$\mathrm{D}_f^{Rp}(\mathbb{P}, \mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} 2 \mathbb{E}_{\substack{x\sim\mathbb{P}\\ y\sim\mathbb{Q}}} \left[f\left(C(x) - C(y) \right) \right]$$

$$\mathrm{D}_f^{Ra}(\mathbb{P}, \mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim\mathbb{P}} \left[f\left(C(x) - \mathbb{E}_{y\sim\mathbb{Q}} C(y) \right) \right] + \mathbb{E}_{y\sim\mathbb{Q}} \left[f\left(\mathbb{E}_{x\sim\mathbb{P}} C(x) - C(y) \right) \right]$$

$$\mathrm{D}_f^{Ralf}(\mathbb{P}, \mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} 2 \mathbb{E}_{x\sim\mathbb{P}} \left[f\left(C(x) - \mathbb{E}_{y\sim\mathbb{Q}} C(y) \right) \right]$$

$$\mathrm{D}_f^{Rc}(\mathbb{P}, \mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{x\sim\mathbb{P}} \left[f\left(C(x) - \mathbb{E}_{m\sim\mathbb{M}} C(m) \right) \right] + \mathbb{E}_{y\sim\mathbb{Q}} \left[f\left(\mathbb{E}_{m\sim\mathbb{M}} C(m) - C(y) \right) \right]$$

are divergences.

We ask that the supremum of $f(x)$ is reached at some positive x (or at ∞). This is purely to ensure that a larger CD can be interpreted as leading to a larger divergence (rather than the opposite). This does not reduce the generality of Theorem 3.1. If $f(x)$ is maximized at $x < 0$, we have that $g(x) = f(-x)$ is maximized at $x > 0$ and one can simply use g instead of f.

We require that f is differentiable at zero and its derivative to be non-zero. This assumption may not be necessary, but it is needed for one of our main lemma which we use to prove that these objective functions are divergences.

Note that $\mathrm{D}_f^{Rp}(\mathbb{P}, \mathbb{Q})$ corresponds to RpGANs, $\mathrm{D}_f^{Ra}(\mathbb{P}, \mathbb{Q})$ corresponds to RaGANs, $\mathrm{D}_f^{Ralf}(\mathbb{P}, \mathbb{Q})$ corresponds to a simplified one-way version of RaGANs (RalfGANs), and $\mathrm{D}_f^{Rc}(\mathbb{P}, \mathbb{Q})$ corresponds to a new type of RGAN called Relativistic centered GANs (RcGANs). RalfGANs are not particularly interesting as they simply represent a simpler version of RaGANs. On the other hand, RcGANs are interesting as they center the critic scores using the mean of the whole mini-batch (rather than the mean of only real or only fake mini-batch samples). This divergence also has similarities to the Jensen–Shannon divergence (JSD) since the JSD is the sum of the KL-divergence between $\mathbb{P}$ and $\mathbb{M}$ to the KL-divergence between $\mathbb{Q}$ and $\mathbb{M}$.

A logical extension to RcGANs would be to standardize the critic scores; however, this would not lead to a divergence given that we could not control the size of the elements inside f. To make it a divergence, we would need a learnable scaling weight (as in batch norm (Ioffe & Szegedy, 2015)), but this would counter the effect of the standardization. Thus, standardizing and scaling would just correspond to an equivalent re-parametrization of D_f^{Rc}.

A sketch of the proof can be found below; the full proof is found in Appendix A.

3.2. Sketch of the Proof

Although the four divergences need separate proofs, a similar framework is used in each of them. Each proof consists of three steps. For clarity of notation, let $D_f(\mathbb{P},\mathbb{Q}) = \sup\limits_{C:\mathcal{X}\to\mathbb{R}} F(\mathbb{P},\mathbb{Q},C,f)$ be the divergence, where F is any of the objective functions in Theorem 3.1.

First, we show that $D_f(\mathbb{P},\mathbb{Q}) \geq 0$. This is easily proven by taking the simplest possible choice of critic, which does not depend on the probability distributions, i.e., $C^w(x) = k$ for all x. This critic always leads to $f(0)$ and thus to a objective function equal to 0. This means that

$$D_f(\mathbb{P},\mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} F(\mathbb{P},\mathbb{Q},C,f) \geq F(\mathbb{P},\mathbb{Q},C^w,f) = 0.$$

Second, we show that $\mathbb{P} = \mathbb{Q} \implies D_f(\mathbb{P},\mathbb{Q}) = 0$. This step generally relies on Jensen's inequality (for concave functions) which we use to show that $D_f(\mathbb{P},\mathbb{P}) \leq 0$. Given that $D_f(\mathbb{P},\mathbb{P}) \geq 0$ and $D_f(\mathbb{P},\mathbb{P}) \leq 0$, we have that $D_f(\mathbb{P},\mathbb{P}) = 0$.

Third, we show that $D_f(\mathbb{P},\mathbb{Q}) = 0 \implies \mathbb{P} = \mathbb{Q}$. This step is by far the most difficult to prove. Instead of showing it directly, we instead prove it by contraposition, i.e., we show that $\mathbb{P} \neq \mathbb{Q} \implies D_f(\mathbb{P},\mathbb{Q}) > 0$. To prove this, we use the fact that if $\mathbb{P} \neq \mathbb{Q}$, there must be values of the probability density functions, $p(x)$ and $q(x)$ respectively, such that $p(x) > q(x)$ (and vice versa). Let $T = \arg\sup_S \mathbb{P}(S) - \mathbb{Q}(S)$, we know that this set is not empty. Note that when $\mathbb{P}$ and $\mathbb{Q}$ have probability density functions $p(x)$ and $q(x)$ respectively, we have that $T = \{x | p(x) > q(x)\}$. To make the proof as simple as possible, we use the following sub-optimal critic:

$$C'(x) = \begin{cases} \nabla & \text{if } x \in T \\ 0 & \text{else,} \end{cases}$$

where $\nabla \neq 0$. This critic function is very simple, but, as we will show, there exists a $\nabla > 0$ such that this leads to an objective function greater than 0 which means that the divergence is also greater than 0.

With this critic in mind, our goal is to transform the problem into the following:

$$\begin{aligned} D_f(\mathbb{P},\mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} F(\mathbb{P},\mathbb{Q},C,f) &\geq F(\mathbb{P},\mathbb{Q},C',f) \\ &\geq L(\nabla) \\ &> 0, \end{aligned}$$

where $L(\nabla) = af(\nabla) + bf(-\nabla)$, for some $a > 0$ and $b > 0$ s.t. $a > b$. We have been able to show this with all divergences.

We want to find a $\nabla > 0$ large enough so that the positive term ($f(\nabla)$) is big, but small enough so that the negative term ($f(-\nabla)$) is not too big. The main caveat is that, by concavity, $f(\nabla) \leq |f(-\nabla)|$. This means that the negative term is always bigger in absolute value than the positive term. This is problematic, since a could be be very close to b and we want $af(\nabla) > bf(-\nabla)$ to get $L(\nabla) > 0$ which proves that we have a divergence. The solution is to choose ∇ to be very small. By continuity of the concave function, if we make ∇ small enough (very close to 0), we can reach a point where $(f(\nabla) \approx -f(-\nabla))$. In which case, if $a = b + \epsilon$, we have that

$$\begin{aligned} L(\nabla) = af(\nabla) + bf(-\nabla) &\approx af(\nabla) - bf(\nabla) \\ &= bf(\nabla) + \epsilon f(\nabla) - bf(\nabla) \\ &= \epsilon f(\nabla) \\ &> 0. \end{aligned}$$

In the actual proof, we show that there always exists a $\delta > 0$ small enough such that any $\nabla \in (0, \delta)$ leads to $L(\nabla) > 0$. This concludes the sketch of the proof.

3.3. Subtypes of Divergences

Figure 1 shows three examples of concave f with the necessary properties to be used in relativistic divergences; they are the concave functions used in SGAN, LSGAN (with labels 1/-1), and HingeGAN. Their respective mathematical functions are

$$f_S(z) = \log(\,\text{sigmoid}(z)) + \log(2), \tag{11}$$

$$f_{LS}(z) = -(z-1)^2 + 1, \tag{12}$$

$$f_{Hinge}(z) = -\max(0, 1-z) + 1. \tag{13}$$

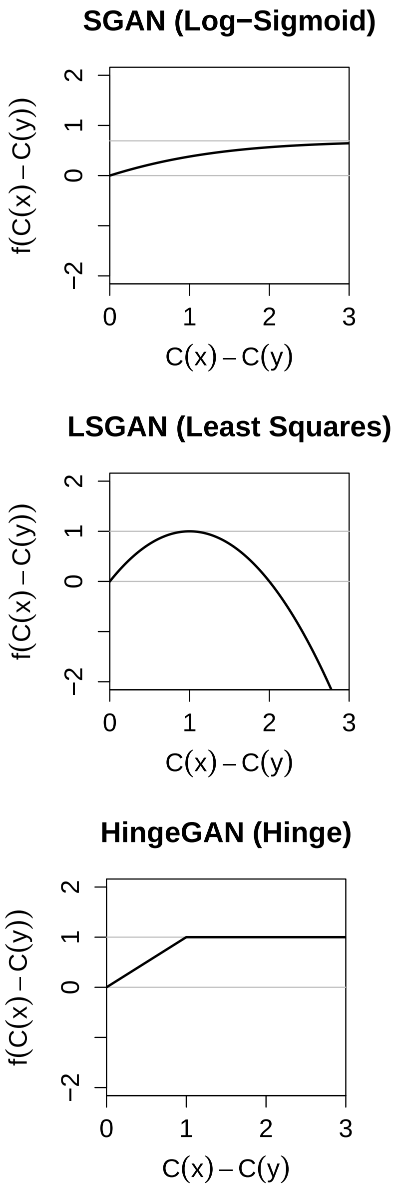

Figure 1. Plots of f with respect to the critic's difference (CD) using three appropriate choices of f for relativistic divergences. The bottom gray line represents $f(0) = 0$; the divergence is zero if all CDs are zero. The above gray line represents the maximum of f; the divergence is maximized if all CDs leads to that maximum.

Interestingly, we see that they form three different types of functions. Firstly, we have functions that grow exponentially less as x increases and thus reach their supremum at ∞. Secondly, we have functions that grow to a maximum and then forever decrease (thus penalizing large CDs). Thirdly, we have functions that grow to a maximum and then never change. SGAN is of the first type, LSGAN is of the second, and HingeGAN is of the third type.

This shows that for all three types, we have that the CD is only encouraged to grow until a certain point. With the first type, we never truly force the CD to stop growing, but the gradients vanish to zero. Thus, SGD effectively prevents the CDs from growing above a certain level (sigmoid saturates at around 2 or 3).

It is useful to keep in mind that Figure 1 also represents the concave functions used for SyGANs, in which case f applies to real and fake data separately ($f(x)$ and $f(-y)$).

3.4. Weakness of the Divergence

The paper by Arjovsky et al. (2017) on using the Wasserstein distance (and other IPMs) for GANs has been extremely influential. In this paper, the authors suggest that the Wasserstein distance is more appropriate than f-divergences for training a critic since it induces the weakest topology possible. Rather than giving a formal definition in terms of topologies, we use a simpler definition (as also done by Arjovsky et al. (2017)):

Definition 3.2. Let $\mathbb{P}$ be a probability distribution with support $\mathcal{X}$, $(\mathbb{P}_n)_{n\in\mathbb{N}}$ be a sequence of distributions converging to $\mathbb{P}$, and D_1 and D_2 be statistical divergences (per definition 3.1).

We say that D_1 is weaker than D_2 if we have that:

$$D_2(\mathbb{P}_n, \mathbb{P}) \to 0 \implies D_1(\mathbb{P}_n, \mathbb{P}) \to 0 \;\forall\, (\mathbb{P}_n)_{n\in\mathbb{N}},$$

but the converse is not true.

We say that D_1 is a weakest distance if we have that:

$$D_1(\mathbb{P}_n, \mathbb{P}) \to 0 \iff \mathbb{P}_n \xrightarrow{D} \mathbb{P} \;\forall\, (\mathbb{P}_n)_{n\in\mathbb{N}},$$

where $\xrightarrow{D}$ represents convergence in distribution.

Thus, intuitively, a weaker divergence can be thought of as converging more easily. Arjovsky et al. (2017) showed that the Wasserstein distance is a weakest divergence and that it is weaker than common f-divergences (as used in f-GANs and standard GANs). They also showed that the Wasserstein distance is continuous with respect to its parameters and they attributed this property to the weakness of the divergence.

Considering this argument, one would except that RaGANs would be weaker than RpGANs which would be weaker than Symmetric GANs since this is generally the order of their relative performance and stability (however, note that this is not always true and GANs can perform better than RaGANs). Instead, we found the opposite relationship:

Theorem 3.2. Let $\mathbb{P}$ be a probability distribution with support S, $(\mathbb{P}_n)_{n\in\mathbb{N}}$ be a sequence of distributions converging to $\mathbb{P}$, $f : \mathbb{R} \to \mathbb{R}$ be a concave function such that $f(0) = 0$,

f is differentiable at 0, $f'(0) \neq 0$, $\sup_x f(x) = M > 0$, and $\arg\sup_x f(x) > 0$. Then, we have that

$$\begin{aligned}&\mathrm{D}_f^W(\mathbb{P},\mathbb{Q}) \text{ is weakest,}\\&\mathrm{D}_f^W(\mathbb{P},\mathbb{Q}) \text{ is weaker than } \mathrm{D}_f^{Sy}(\mathbb{P},\mathbb{Q}),\\&\mathrm{D}_f^{Sy}(\mathbb{P},\mathbb{Q}) \text{ is weaker than } \mathrm{D}_f^{Rp}(\mathbb{P},\mathbb{Q}),\\&\mathrm{D}_f^{Rp}(\mathbb{P},\mathbb{Q}) \text{ is weaker than } \mathrm{D}_f^{Ra}(\mathbb{P},\mathbb{Q}),\end{aligned}$$

where D^W is the Wasserstein distance and D^{Sy} is the distance in Symmetric GANs (see equation 3).

The proof is in Appendix B.

Given the good performance of RaGANs, this suggests that the argument made by Arjovsky et al. (2017) is insufficient. It only focuses on a perfect sequence of converging distributions, but the generator training does not guarantee a converging sequence of fake data distributions. It ignores the complex dynamics and intricacies of the generator training, which are still not well understood. Furthermore, it assumes an optimal critic which is effectively unobtainable. In practice, obtaining a semi-optimal critic requires training the critic for multiple iterations before training the generator; this significantly increase the computational time.

Furthermore, it has been found that WGAN does not provide a good approximation of the Wasserstein distance and that better approximations of the Wasserstein distance lead to worse GANs (Mallasto et al., 2019). This provides further argument towards the idea that the weakness of the divergence is not a good indicator of a good divergence for GANs. As previously suggested (Jolicoeur-Martineau, 2018a), we hypothesize that what make WGAN good for GANs are likely 1) the constraint of the critic (a Lipschitz critic) and 2) the use of a relativistic discriminator, rather than the weakness of the divergence.

4. Estimators

4.1. RpGANs

To estimate RpGANs, Jolicoeur-Martineau (2018b) used the following estimator[2]:

$$\widehat{\mathrm{D}}_f^{\mathrm{Rp}}(\mathbb{P},\mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} \frac{2}{k}\sum_{i=1}^{k}\left[f(C(x_i)-C(y_i))\right],$$

where $x_1,\ldots,x_k$ and $y_1,\ldots,y_k$ are samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Although this is an unbiased estimator of $\mathrm{D}_f^{Rp}(\mathbb{P},\mathbb{Q})$, it is not the estimator with the minimal variance for a given mini-batch. Using the two-sample version (Lehmann, 1951) of the U-statistic theorem (Hoeffding, 1992) and given that the loss function is symmetric with respect to its arguments, one can show the following:

[2]Note that they actually used $\frac{1}{k}$ instead of $\frac{2}{k}$ because of how they defined the divergence.

Corollary 4.1. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions with support $\mathcal{X}$. Let $x_1,\ldots,x_k$ and $y_1,\ldots,y_k$ be i.i.d. samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then, we have that

$$\widehat{\mathrm{D}}_f^{\mathrm{Rp}*}(\mathbb{P},\mathbb{Q}) = \sup_{C:\mathcal{X}\to\mathbb{R}} \frac{2}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\left[f(C(x_i)-C(y_j))\right]$$

is the minimum-variance unbiased estimator (MVUE) of $\mathrm{D}_f^{Rp}(\mathbb{P},\mathbb{Q})$.

Although it is the MVUE, this estimator requires $O(k^2)$ operations instead of $O(k)$. In the experiments, we will show that using this estimator does not lead to good performance. Given the quadratic scaling and lack of performance gain, it may not be worth using.

4.2. RaGANs and RalfGANs

The divergences of RaGANs and RalfGANs assume that one knows the true expectation of the critic of real and fake data. However, in practice, we can only estimate the expectation. Although never explicitly mentioned, (Jolicoeur-Martineau, 2018a) simply replaced all expectations by the mini-batch mean:

$$\mathbb{E}\left[C(x)\right] \approx \frac{1}{k}\sum_{i=1}^{k} C(x_i),$$

where k is the size of the mini-batch.

Given the non-linear function applied after calculating the CD, the divergences of RaGANs are biased with finite batch size k. This means that RaGANs are only asymptotically unbiased. How large k must be for the bias to become negligible is unclear.

We attempted to find a close form for the bias with f_S, f_{LS}, and f_{Hinge} (equations 11, 12, 13 and Figure 1), but we were only able to find a closed form with f_{LS}. The bias with f_{LS} has a simple form and can be removed, as shown below:

Corollary 4.2. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions with support $\mathcal{X}$. Then, we have that

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \frac{1}{k}\left(\hat{\sigma}_{C(x)} + \hat{\sigma}_{C(y)} - \sum_{i=1}^{k}\left[\left(C(x_i) - \hat{\mu}_{C(y)} - 1\right)^2\right] - \sum_{j=1}^{k}\left[\left(\hat{\mu}_{C(x)} - C(y_j) - 1\right)^2\right]\right) + 2,$$

$$\sup_{C:\mathcal{X}\to\mathbb{R}} \frac{2}{k}\left(\hat{\sigma}_{C(y)} - \sum_{i=1}^{k}\left[\left(C(x_i) - \hat{\mu}_{C(y)} - 1\right)^2\right]\right) + 1,$$

$$\inf_{C:\mathcal{X}\to\mathbb{R}} \frac{1}{k}\left(\frac{1}{2}\hat{\sigma}_{C(x)} + \frac{1}{2}\hat{\sigma}_{C(y)} + \sum_{i=1}^{k}\left[\left(C(x_i) - \hat{\mu}_{C} - 1\right)^2\right] + \sum_{j=1}^{k}\left[\left(\hat{\mu}_{C} - C(y_j) - 1\right)^2\right]\right) - 2$$

are unbiased estimator of $\mathrm{D}^{Ra}_{f_{LS}}(\mathbb{P},\mathbb{Q})$, $\mathrm{D}^{Ralf}_{f_{LS}}(\mathbb{P},\mathbb{Q})$, and $\mathrm{D}^{Rc}_{f_{LS}}(\mathbb{P},\mathbb{Q})$ respectively. Furthermore,

$$\hat{\mu}_{C(x)} = \frac{1}{k}\sum_{i=1}^{k} C(x_i),$$

$$\hat{\mu}_{C(y)} = \frac{1}{k}\sum_{i=1}^{k} C(y_i),$$

$$\hat{\mu}_{C} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{C(x_i)+C(y_i)}{2}\right),$$

$$\hat{\sigma}_{C(x)} = \frac{1}{(k-1)}\sum_{i=1}^{k}\left(C(x_i) - \hat{\mu}_{C(x)}\right)^2,$$

$$\hat{\sigma}_{C(y)} = \frac{1}{(k-1)}\sum_{i=1}^{k}\left(C(y_i) - \hat{\mu}_{C(y)}\right)^2.$$

See Appendix C for the proof. This means that we can estimate the loss functions in RaLSGAN, RalfLSGAN, and RcLSGAN without bias. In the experiments, we will show that the bias is negligible with the usual choices of f (equations 11, 12, 13) and batch size (32 or higher).

5. Experiments

All experiments were done with the spectral GAN architecture for 32x32 images (Miyato et al., 2018) in Pytorch (Paszke et al., 2017). We used the standard hyperparameters: learning rate (lr) = .0002, batch size (k) = 32, and the ADAM optimizer (Kingma & Ba, 2014) with parameters $(\alpha_1, \alpha_2) = (.50, .999)$. We trained the models for 100k iterations with one critic update per generator update. For the datasets, we used CIFAR-10 (50k training images from 10 categories) (Krizhevsky, 2009), CelebA (200k of face images from celebrities) (Liu et al., 2015) and CAT (10k images of cats) (Zhang et al., 2008). All models were trained using the same seed (seed=1) with a single GPU. To evaluate the quality of generated outputs, we used the Fréchet Inception Distance (FID) (Heusel et al., 2017). For a review of the different evaluation metrics for GANs, please see Borji (2018). CAT was preprocessed by cropping all images to the faces of the cats, removing outliers (faces hidden by background), and removing images smaller than 32x32. CelebA images were center cropped to 160x160 before being resized to 32x32. See code for details; the code to reproduce the experiments is available on https://github.com/AlexiaJM/relativistic-f-divergences.

5.1. Bias

We approximated the bias of RaGANs and RcGANs by estimating the real/fake critic mean from 320 samples rather than the 32 mini-batch samples. For f_{LS}, we were able to calculate the true value of the bias (in expectation, see Corollary 4.2). Results on CIFAR-10 are shown in Figure 2.

For RAGANs, the approximation of the relative bias with f_{LS} was correct from 4k iterations and onwards. For all choices of f, we observed the same pattern of low approximated relative bias which stabilized after a certain number of iterations. We suspect that this may be due to the important instabilities of the first iterations when the discriminator is not optimal. At 15k iterations, all biases were stabilized. We calculated the average of the bias with different f starting at 15k iterations: .995 for the true relative bias with f_{LS}, .996 for the approximated relative bias with f_{LS}, .994 for the approximated relative bias with f_S, and .997 for the approximated relative bias with f_{Hinge}.

For RcGANs, the approximation of the bias with f_{LS} was correct from the very beginning of training. All biases were relatively stable over time with the exception of f_S which increased linearly over time (up to around 1.05). We calculated the average of the bias with different f: 1.007 for the true relative bias with f_{LS}, 1.007 for the approximated relative bias with f_{LS}, 1.03 for the approximated relative bias with f_S, and 1.007 for the approximated relative bias with f_{Hinge}.

Overall, this shows that the bias in the estimators of RaGANs and RcGANs tends to be small. Furthermore, with the exception of f_S, the bias is relatively stable over time. Thus, accounting for the bias, may not be necessary.

5.2. Divergences

To test the new relativistic divergences proposed (and verify whether removing the bias in RaGANs is useful), we ran experiments on CIFAR-10 using f_{LS}, on LSUN bedrooms

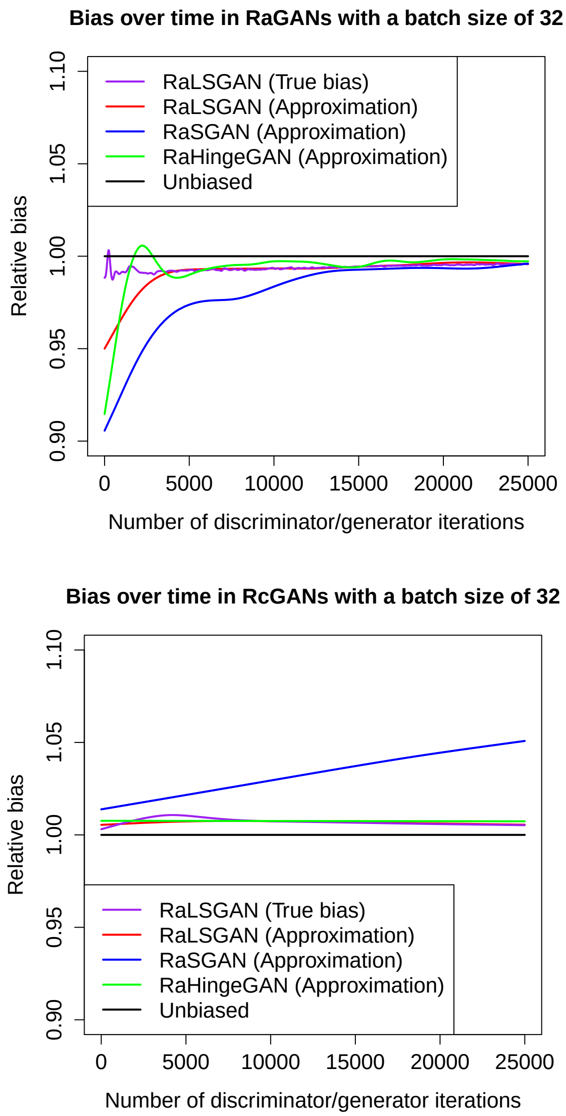

Figure 2. Plots of the relative bias (i.e., the biased estimate divided by the unbiased estimate) of relativistic average and centered f-divergences estimators over training time on CIFAR-10 with a mini-batch size of 32. Approximations of the bias were made using 320 independent samples.

using f_{Hinge}, and on CAT using f_{Hinge} (these choices of f were arbitrary). Results are shown in Table 1.

Using the MVUE for RpGAN resulted in the generator having a worse performance on CIFAR-10 with f_{LS} ($\beta = .37$, $p = .72$), CelebA with f_{Hinge} ($\beta = 2.08$, $p = .07$), and CAT with f_S ($\beta = 4.02$, $p = .003$). Similarly, using the unbiased estimator made the generator perform sightly worse for RaLSGAN ($\beta = 2.37$, $p = .04$) and RcLSGAN ($\beta = 1.33$, $p = .05$). These results are surprising as they suggest that using noisy or slightly biased estimators may

Table 1. Minimum (and standard deviation) of the FID calculated at 10k, 20k, ... , 100k iterations using different loss functions (see equations 11, 12, 13) and datasets.

Loss	CIFAR-10 f_{LS}	CelebA f_{Hinge}	CAT f_S
GAN	31.1 (8)	15.3 (52)	15.2 (11)
RpGAN	31.5 (8)	16.7 (4)	12.9 (2)
RpGAN$_{\text{MVUE}}$	30.2 (12)	21.9 (3)	18.2 (3)
RaGAN	**29.2** (**7**)	**15.9** (5)	**12.3** (**1**)
RaGAN$_{\text{unbiased}}$	30.3 (13)	-	-
RcGAN	31.7 (8)	18.1 (**3**)	16.5 (7)
RcGAN$_{\text{unbiased}}$	32.3 (9)	-	-

be beneficial.

6. Conclusion

Most importantly, we proved that the objective function of the critic in RGANs is a divergence. In addition, we showed that f-divergences are weaker than relativistic f-divergences. Thus, the weakness of the topology induced by a divergence alone cannot explain why WGAN performs well. Finally, we took a closer look at the estimators or RGANs and found that 1) the estimator of RpGANs used by Jolicoeur-Martineau (2018a) is not the minimum-variance unbiased estimator (MVUE) and 2) the estimators of RaGANs and RalfGANs are slightly biased with finite batch-sizes. Surprisingly, we found that neither using the MVUE with RpGANs or using an unbiased estimator with RaGANs and RalfGANs improved the performance. On the contrary, using better estimators always slightly decreased the quality of generated samples. This suggests that using noisy estimates of the divergences may be beneficial as a regularization mechanism. This could be explained by vanishing gradients when the discriminator becomes closer to optimality (Arjovsky & Bottou, 2017).

It still remains a mystery as to why RaGANs are better than RpGANs and the direct mechanism that leads to RGANs performing in a much more stable matter. Future work should attempt to better understand the effect of the critic's difference on training. Our experiments were limited to the generation of small images; thus, we encourage further experiments with the MVUE and the unbiased estimator of RaLSGAN.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Borji, A. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:1802.03446*, 2018.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

Hoeffding, W. A class of statistics with asymptotically normal distribution. In *Breakthroughs in Statistics*, pp. 308–334. Springer, 1992.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jolicoeur-Martineau, A. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018a.

Jolicoeur-Martineau, A. Gans beyond divergence minimization. *arXiv preprint arXiv:1809.02145*, 2018b.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Lehmann, E. L. Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pp. 165–179, 1951.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Mallasto, A., Montúfar, G., and Gerolin, A. How well do wgans estimate the wasserstein metric? *arXiv preprint arXiv:1910.03875*, 2019.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821. IEEE, 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mroueh, Y. and Sercu, T. Fisher gan. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2513–2523. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6845-fisher-gan.pdf.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

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

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Zhang, W., Sun, J., and Tang, X. Cat head detection-how to effectively exploit shape and texture features. In *European Conference on Computer Vision*, pp. 802–816. Springer, 2008.