Cumulant GAN

Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, Yannis Stylianou, *Fellow, IEEE*, and Markos A. Katsoulakis

***Abstract*— In this article, we propose a novel loss function for training generative adversarial networks (GANs) aiming toward deeper theoretical understanding as well as improved stability and performance for the underlying optimization problem. The new loss function is based on cumulant generating functions (CGFs) giving rise to Cumulant GAN. Relying on a recently derived variational formula, we show that the corresponding optimization problem is equivalent to Rényi divergence minimization, thus offering a (partially) unified perspective of GAN losses: the Rényi family encompasses Kullback–Leibler divergence (KLD), reverse KLD, Hellinger distance, and χ^2-divergence. Wasserstein GAN is also a member of cumulant GAN. In terms of stability, we rigorously prove the linear convergence of cumulant GAN to the Nash equilibrium for a linear discriminator, Gaussian distributions, and the standard gradient descent ascent algorithm. Finally, we experimentally demonstrate that image generation is more robust relative to Wasserstein GAN and it is substantially improved in terms of both inception score (IS) and Fréchet inception distance (FID) when both weaker and stronger discriminators are considered.**

***Index Terms*— Cumulant generating function (CGF), generative adversarial networks (GANs), image generation, Rényi divergence.**

I. Introduction

A Generative adversarial network (GAN) is a two-player zero-sum game between a discriminator and a generator, both being neural networks with high learning capacity. GANs [1] are powerful generative models capable of drawing new samples from an unknown distribution when only samples from that distribution are available. Their popularity stems from their ability to generate realistic samples from high-dimensional and complex distributions. In computer vision, GANs have been applied for (conditional) image generation [2]–[8], image synthesis from text (i.e., reverse captioning) [9], image-to-image translation [10], and image super-resolution [11]. In time-series data, GANs have been used for speech enhancement [12], speech synthesis [13], [14] as well as for natural language processing [15], [16] among other types of raw data. GANs have been also employed to protect classifiers against adversarial examples [17]–[19]. Several surveys and reviews on GANs are available in the literature [20]–[23]. Moreover, the concept of adversarial training which fairly designates that the loss function is learned (i.e., data-driven) and not predetermined by the user has been successfully applied in domain adaptation [10], [24]–[26] and representation disentanglement [27].

There are three ingredients that constitute a GAN: the architectures for both the generator and the discriminator, the training algorithm and the loss function which is further divided into the objective functional to be optimized and the function space where the discriminator belongs to. Over the years, the capacity of the neural networks has been increased resulting in significant gains in terms of naturalness and performance [6]–[8], [28]. Similarly, new normalization techniques such as spectral normalization [29] and new optimization algorithms [30] have been proposed. Two characteristic examples are progressive GAN training [6], [31] where the models are built in progressive levels of resolution and MelGAN [14] where weight normalization played a critical role for the generation of high-quality speech. Several heuristics have been also devised [32] to alleviate the difficulties of training GANs.

As already stated, the third ingredient in GANs' definition corresponds to the loss function. Since their introduction, GANs have been described as a tractable approach to minimize a divergence or a distance between the real data distribution and the model distribution. Indeed, the original formulation of GANs [1] can be seen as the minimization of the Shannon-Jensen divergence. f-GAN [33] is a generalization of vanilla GAN where a variational lower bound for the f-divergence is minimized. Least-Squares GAN [34] minimizes a softened version of the Pearson χ^2-divergence and hinge loss [35] proposes objective functionals aiming toward avoiding mode collapse issues. As it is well-documented, the training procedure of GANs often fails and several studies have suggested remedies to alleviate the observed hindrances. For instance, a recurring impediment with GAN training is the oscillatory

Manuscript received July 17, 2020; revised August 16, 2021 and December 8, 2021; accepted March 1, 2022. The work of Yannis Pantazis was supported in part by the Hellenic Foundation for Research and Innovation (HFRI) through the "Second Call for HFRI Research Projects to support Faculty Members and Researchers" under Project 4753 and in part by the project "Innovative Actions in Environmental Research and Development (PErAn)" (MIS 5002358) funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" under Grant NSRF 2014-2020. The work of Dipjyoti Paul was supported by the EU H2020 Research and Innovation Program under Grant MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu). The work of Michail Fasoulakis was supported by the Stavros Niarchos-FORTH Post-Doctoral Fellowship for the project Advancing Young Researchers' Human Capital in Cutting Edge Technologies in the Preservation of Cultural Heritage and the Tackling of Societal Challenges - ARCHERS. The work of Markos A. Katsoulakis was supported in part by the National Science Foundation (NSF) under Grant DMS-2008970, in part by the HDR-TRIPODS: Institute for Integrated Data Science: A Transdisciplinary Approach to Understanding Fundamental Trade-offs and Theoretical Foundations Program of NSF under Grant CISE-1934846, and in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA-9550-18-1-0214. *(Corresponding author: Yannis Pantazis.)*

Yannis Pantazis is with the Institute of Applied and Computational Mathematics, Foundation for Research & Technology–Hellas (FORTH), 70013 Heraklion, Greece (e-mail: pantazis@iacm.forth.gr).

Dipjyoti Paul and Yannis Stylianou are with the Department of Computer Science, University of Crete, 70013 Rethymno, Greece.

Michail Fasoulakis is with the Institute of Computer Science, FORTH, 70013 Heraklion, Greece.

Markos A. Katsoulakis is with the Department of Mathematics and Statistics, University of Massachusetts at Amherst, Amherst, MA 01003 USA.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2022.3161127.

Digital Object Identifier 10.1109/TNNLS.2022.3161127

behavior of the optimization algorithms due to the fact that the optimal solution is a saddle point of the loss function. Standard optimization algorithms such as stochastic gradient descent ascent (SGDA) may fail even for simple loss functions [30], [36].

One of the most successful approaches to improve stability is through the restriction of the function space of the discriminator. Wasserstein GAN (WGAN) [37] which has been further improved in [38] aims to minimize the Wasserstein distance which is equivalent to restricting to Lipschitz continuous functions. SNGAN [29] also restricts to Lipschitz continuous functions while MMD-GAN [39] restricts the discriminator to belong to a reproducible Hilbert kernel space. Recently, the dissociation between the objective functional and the function space has been presented in a rigorous mathematical framework [40]. In this article, we concentrate on the loss function and propose a new objective functional that further improves training stability and avoids mode collapse.

The novel objective functional is based on cumulant generating functions (CGFs) with the resulting model referred as Cumulant GAN. A key advantage of cumulants over expectations is that cumulants capture higher-order information about the underlying distributions which often results in more effective learning. Using this property, we rigorously prove that cumulant GAN converges exponentially fast when the gradient descent ascent algorithm is used for the special case with linear generator, linear discriminator, and Gaussian distributions. Despite being a simple case, this theoretical result offers a rigorous and valuable differentiation between WGAN, which fails to converge, and the proposed cumulant GAN that demonstrates linear convergence to the Nash equilibrium, when the same gradient descent ascent algorithm is used on both.

Interestingly, the optimization of cumulant GAN can be described as a weighting extension of the standard SGDA where the samples that confuse the discriminator the most receive a higher weight, thus, contributing more to the update of the neural network's parameters. Furthermore, by applying a recent variational representation formula [41], we show that cumulant GAN is capable of interpolating between several GAN formulations, thus, offering a partially unified mathematical framework. Indeed, the optimization of the proposed loss function is equivalent to the minimization of a divergence for a wide set of cumulant GAN's hyper-parameter values. It is also worth-noting that despite f-GAN's (partial) unification property [33], cumulant GAN and f-GAN formulations are not equivalent even when they minimize the same divergence and there is a subtle but important difference: the underlying variational representation which is eventually optimized is different. Ours is based on the Donsker–Varadhan representation formula while f-GAN is based on the Legendre transform of f-divergence. For KLD, Donsker–Varadhan formula is tighter than Legendre duality formula.[1] Additionally, our formulation is computationally more manageable because the hyper-parameters of cumulant GAN are of continuous nature while f-GAN requires different f's for different divergences.

[1]Simply by the fact that $x \geq e \log x$; see also [42].

Our numerical demonstrations aim to provide insights into cumulant GAN's representational ability and learnability advantages. Experiments on synthetic multi-modal data revealed the differences in the dynamics of learning for different hyper-parameter values of cumulant GAN. Even though the optimal solution is the same, the SGDA's dynamics for the training parameters driven by the chosen hyper-parameters' values resulted in very different distributional realizations with the two extremes being mode covering and mode selection. Moreover, using cumulant GAN, we were able to recover higher order statistics even when the discriminator is linear. Finally, we demonstrated increased robustness as well as improved performance on image generation for both CIFAR10 and ImageNet datasets. We performed relative comparisons with WGAN under standard as well as distressed settings which is a primary reason for training instabilities in GANs and demonstrated that not only cumulant GAN is more stable but also it is better up to 68% in terms of averaged inception score (IS) and up to 75% in terms of Fréchet inception distance (FID).

The article is organized as follows. Section II introduces the necessary background theory, while Section III defines cumulant GAN and highlights the derivation of several of its theoretical properties. In Section IV, numerical simulations on both synthetic and real datasets are presented, while Section V concludes the article.

II. Background

The proposed GAN is a substantial generalization of WGAN by means of CGFs. These concepts are briefly discussed in this section.

A. Wasserstein GAN

WGAN [37], [38] minimizes the Earth-Mover (a.k.a. one-Wasserstein) distance and primarily aims to avoid gradient saturation during the training process. Based on the Kantorovich–Rubinstein duality formula for Wasserstein distance, the loss function of WGAN can be written as

$$\min_{G} \max_{D \in \mathcal{D}} \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)] \tag{1}$$

where p_r and p_g correspond to the real data distribution and the implicitly defined model distribution, respectively. Namely, p_g denotes the distribution of $G(z)$, where G is the generator and $z \sim p_z(z)$ is a random input vector often following a standard normal or uniform distribution. $D(\cdot)$ is the discriminator (called critic in the WGAN setup) while $\mathcal{D}$ is the function space of all one-Lipschitz continuous functions. In WGAN, Lipschitz continuity is imposed by adding a (soft) regularization term on gradient values called gradient penalty (GP). It has been shown that GP regularization produces superior performance relative to weight clipping [38].

B. Cumulant Generating Functions

The CGF, also known as the log-moment generating function, is defined for a random variable with probability density function $p(x)$ as

$$\Lambda_{f,p}(\beta) = \log \mathbb{E}_p\left[e^{\beta f(x)}\right] \tag{2}$$

where f is a measurable function with respect to p. The standard CGF is obtained when $f(x) = x$. CGF is a convex function with respect to β and it contains information for all moments of p. CGF also encodes the tail behavior of distributions and plays a key role in the theory of Large Deviations for the estimation of rare events [43]. A power series expansion of the CGF reveals that the lower order statistics dominate when $|\beta| \ll 1$ while all statistics contribute to the CGF when $|\beta| \gg 1$. In statistical mechanics, CGF is the logarithm of the partition function, $-\beta^{-1}\Lambda_{f,p}(-\beta)$ is called the Helmholtz free energy where β is interpreted as the inverse temperature and f as the Hamiltonian [44]. Furthermore, it is straightforward to show that $\Lambda_{f,p}(0) = 0$ as well as $\Lambda'_{f,p}(0) = \mathbb{E}_p[f(x)]$, hence, the following limit for CGF holds:

$$\lim_{\beta \to 0} \beta^{-1}\Lambda_{f,p}(\beta) = \mathbb{E}_p[f(x)]. \tag{3}$$

We are now ready to introduce the new GAN.

III. Cumulant GAN

A. Definition

We define a novel GAN model by substituting the expectations in the loss function of WGAN with the respective CGFs. Thus, we propose to optimize the following minimax problem:

$$\begin{aligned} &\min_G \max_{D \in \mathcal{D}} \left\{(-\beta)^{-1}\Lambda_{D,p_r}(-\beta) - \gamma^{-1}\Lambda_{D,p_g}(\gamma)\right\} \\ &\equiv \min_G \max_{D \in \mathcal{D}} \underbrace{-\beta^{-1}\log \mathbb{E}_{p_r}\left[e^{-\beta D(x)}\right] - \gamma^{-1}\log \mathbb{E}_{p_g}\left[e^{\gamma D(x)}\right]}_{=L(\beta,\gamma)} \end{aligned} \tag{4}$$

where the hyper-parameters β and γ are two nonzero real numbers which control the learning dynamics as well as the optimal solution. Since the loss function is the difference of two CGFs, we call $L(\beta, \gamma)$ in (4) the cumulant loss function and the respective generative model as Cumulant GAN. Throughout this article, we assume the mild condition that both CGFs are finite for a neighborhood of $(0, 0)$, therefore, the cumulant loss is well defined for $|\beta| + |\gamma| < \epsilon$, for some $\epsilon > 0$.

The definition of the loss function is extended on the axes and the origin of the (β, γ)-plane using the limit in (3). Hence, the cumulant loss function is defined for all values of β and γ for which the new loss function is finite. It is straightforward to show that WGAN is a special case of cumulant GAN.

Proposition 1: Let $\mathcal{D}$ be the set of all 1-Lipschitz continuous functions. Then, cumulant GAN with $(\beta, \gamma) = (0, 0)$ is equivalent to WGAN.

Proof: The proposition is a consequence of the fact that

$$\lim_{\beta,\gamma \to 0} L(\beta, \gamma) = L(0, 0) = \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)].$$

□

Remark 1: The same proof applies when $\mathcal{D}$ is the set of all measurable and bounded functions and cumulant GAN with $(\beta, \gamma) = (0, 0)$ is equivalent to minimizing the Radon metric between the two distributions which corresponds to the origin in Fig. 1.

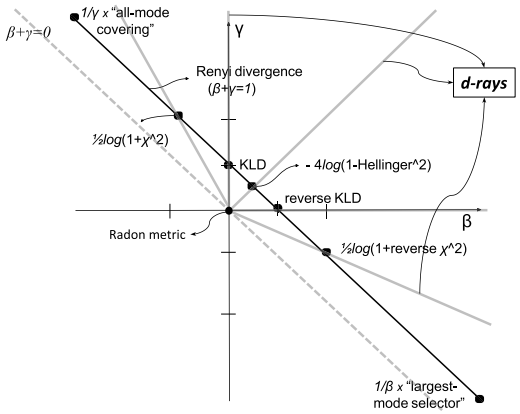


Fig. 1. Special cases of cumulant GAN. Line defined by $\beta + \gamma = 1$ has a point symmetry. The central point, $((1/2), (1/2))$, corresponds to the Hellinger distance. For each point, $(\alpha, 1 - \alpha)$, there is a symmetric one, i.e., $(1 - \alpha, \alpha)$, which has the same distance from the symmetry point. The respective divergences have reciprocal probability ratios (e.g., KLD & reverse KLD, χ^2-divergence & reverse χ^2-divergence, etc.). Each point on the ray starting at the origin and pass through the point $(\alpha, 1 - \alpha)$ also corresponds to (scaled) Rényi divergence of order α. These half-lines are called d-rays.

Next, we rigorously demonstrate that cumulant GAN can be seen as a unified and smooth interpolation between several well-known divergence minimization problems.

B. KLD, Reverse KLD and Rényi Divergence as Special Cases

A major inconvenience of many GAN formulations is their inability to interpret the loss function value and understand the properties of the obtained solution. Even when the stated goal is to minimize a divergence as in the original GAN and the f-GAN, the utilization of training tricks such as a nonsaturating generators may result in the minimization of something completely different as it was recently observed [45]. In contrast, the proposed cumulant loss function can be interpreted for several choices of its hyper-parameters. Below there is a list of values for β and γ that result to interpretable loss functions. Indeed, several well-known divergences are recovered when the function space for the discriminator is the set of all measurable and bounded functions. In the following, we make the convention that a forward divergence or simply divergence is a divergence that uses the probability ratio, (p_r/p_g), while a reverse divergence uses the reciprocal ratio.

Theorem 1: Let $\mathcal{D}$ be the set of all bounded and measurable functions. Then, the optimization of cumulant loss in (4) is equivalent to the minimization of

1) Kullback–Leibler divergence for $(\beta, \gamma) = (0, 1)$:
$$\min_G \max_{D \in \mathcal{D}} L(0, 1) \equiv \min_G D_{KL}\left(p_r || p_g\right).$$
2) Reverse KLD for $(\beta, \gamma) = (1, 0)$:
$$\min_G \max_{D \in \mathcal{D}} L(1, 0) \equiv \min_G D_{KL}\left(p_g || p_r\right).$$
3) Rényi divergence for $(\beta, \gamma) = (\alpha, 1-\alpha)$ with $\alpha \neq 0$ and $\alpha \neq 1$:
$$\min_G \max_{D \in \mathcal{D}} L(\alpha, 1 - \alpha) \equiv \min_G \mathcal{R}_\alpha\left(p_g || p_r\right)$$

as well as for $(\beta, \gamma) = (1-\alpha, \alpha)$ with $\alpha \neq 0$ and $\alpha \neq 1$:

$$\min_G \max_{D\in\mathcal{D}} L(1-\alpha, \alpha) \equiv \min_G \mathcal{R}_\alpha\big(p_r||p_g\big)$$

where $\mathcal{R}_\alpha(p||q)$ is the Rényi divergence defined by

$$\mathcal{R}_\alpha(p||q) = \frac{1}{\alpha(1-\alpha)} \log \mathbb{E}_q\left[\left(\frac{p}{q}\right)^\alpha\right]$$

when p and q are absolutely continuous with respect to each other and $\alpha > 0$.[2]

Proof:

1) Using the definition of $L(\beta, \gamma)$, we have:

$$\begin{aligned}\max_{D\in\mathcal{D}} L(0, 1) &= \max_{D\in\mathcal{D}}\big\{\mathbb{E}_{p_r}[D(x)] - \log \mathbb{E}_{p_g}\big[e^{D(x)}\big]\big\}\\ &= D_{KL}\big(p_r||p_g\big)\end{aligned}$$

where the last equation is the Donsker-Varadhan variational formula [43], [46].

2) Similarly

$$\begin{aligned}\max_{D\in\mathcal{D}} L(1, 0) &= \max_{D\in\mathcal{D}}\big\{-\log \mathbb{E}_{p_r}\big[e^{-D(x)}\big] - \mathbb{E}_{p_g}[D(x)]\big\}\\ &= \max_{D'=-D\in\mathcal{D}}\Big\{\mathbb{E}_{p_g}\big[D'(x)\big] - \log \mathbb{E}_{p_r}\Big[e^{D'(x)}\Big]\Big\}\\ &= D_{KL}\big(p_g||p_r\big)\end{aligned}$$

where we applied again the Donsker–Varadhan variational formula.

3) Generalizing a. and b. we now have:

$$\begin{aligned}&\max_{D\in\mathcal{D}} L(\alpha, 1-\alpha)\\ &\quad= \max_{D\in\mathcal{D}}\Big\{-\frac{1}{\alpha}\log \mathbb{E}_{p_r}\big[e^{-\alpha D(x)}\big]\\ &\qquad\qquad -\frac{1}{1-\alpha}\log \mathbb{E}_{p_g}\big[e^{(1-\alpha)D(x)}\big]\Big\}\\ &\quad= \max_{D'=-D\in\mathcal{D}}\Big\{\frac{1}{\alpha-1}\log \mathbb{E}_{p_g}\Big[e^{(\alpha-1)D'(x)}\Big]\\ &\qquad\qquad -\frac{1}{\alpha}\log \mathbb{E}_{p_r}\Big[e^{\alpha D'(x)}\Big]\Big\}\\ &\quad= \mathcal{R}_\alpha\big(p_g||p_r\big)\end{aligned}$$

where the last equation is an extension of the Donsker–Varadhan variational formula to Rényi divergence and was recently proved in (see [41, Th. 3.1]). For completeness, we provide a proof of the Rényi divergence variational representation in Appendix A of Supplementary Materials.

The proof for the case $L(1-\alpha, \alpha)$ is similar and agrees with the symmetry identity for the Rényi divergence, $\mathcal{R}_\alpha(p||q) = \mathcal{R}_{1-\alpha}(q||p)$. □

The Rényi divergence, $\mathcal{R}_\alpha$, interpolates between KLD ($\alpha \to 0$) and reverse KLD ($\alpha \to 1$). Interestingly, there are additional special cases that belong to the family of Rényi divergences. The following corollary states some of them, while Fig. 1 depicts schematically the obtained divergences and distances on the (β, γ)-plane.

[2]The definition is extended for $\alpha < 0$ using the symmetry identity $\mathcal{R}_\alpha(p||q) = \mathcal{R}_{1-\alpha}(q||p)$.

Corollary 1: Under the same assumption as in Theorem 1, the optimization of (4) is equivalent to the minimization of

1) *Hellinger distance* for $(\beta, \gamma) = ((1/2), (1/2))$:

$$\min_G \max_{D\in\mathcal{D}} L\left(\frac{1}{2}, \frac{1}{2}\right) \equiv \min_G -4\log\big(1 - D_H^2\big(p_g, p_r\big)\big)$$

where $D_H^2(p, q) = (1/2)\mathbb{E}_q[(((p/q))^{1/2} - 1)^2]$ is the square of the Hellinger distance [47].

2) χ^2-divergence for $(\beta, \gamma) = (-1, 2)$:

$$\min_G \max_{D\in\mathcal{D}} L(-1, 2) \equiv \min_G \frac{1}{2}\log\big(1 + \chi^2\big(p_r||p_g\big)\big)$$

and reverse χ^2-divergence for $(\beta, \gamma) = (2, -1)$:

$$\min_G \max_{D\in\mathcal{D}} L(2, -1) \equiv \min_G \frac{1}{2}\log\big(1 + \chi^2\big(p_g||p_r\big)\big)$$

where $\chi^2(p||q) = \mathbb{E}_q[((p/q) - 1)^2]$ is the χ^2-divergence[3] [47].

3) All-mode covering or worst case regret in minimum description length principle [48] for $(\beta, \gamma) = (\infty, -\infty)$:

$$\min_G \lim_{\alpha\to\infty} \alpha \max_{D\in\mathcal{D}} L(\alpha, 1-\alpha) \equiv \min_G \log\left(\operatorname*{ess\,sup}_{x\in\operatorname{supp}(p_g)} \frac{p_g(x)}{p_r(x)}\right)$$

where ess sup is the essential supremum of a function.

4) Largest-mode selector for $(\beta, \gamma) = (-\infty, \infty)$:

$$\min_G \lim_{\alpha\to\infty} \alpha \max_{D\in\mathcal{D}} L(1-\alpha, \alpha) \equiv \min_G \log\left(\operatorname*{ess\,sup}_{x\in\operatorname{supp}(p_r)} \frac{p_r(x)}{p_g(x)}\right).$$

Proof: All cases a.–d. follow from Theorem 1-c as special instances of Rényi divergence:

$$\begin{aligned}R_{1/2}(p\|q) &= -4\log(1 - D_H^2(p, q))\\ R_2(p\|q) &= \frac{1}{2}\log(1 + \chi^2(p||q)), \quad R_{-1}(p\|q) = R_2(q\|p)\end{aligned}$$

and

$$\lim_{\alpha\to\infty} \alpha R_\alpha(p\|q) = \log\left(\operatorname*{ess\,sup}_{x\in\operatorname{supp}(q)} \frac{p(x)'}{q(x)}\right).$$

We refer to [49] and [50] and the references therein for detailed proofs. □

The flexibility of the two hyper-parameters is significant since they offer a simple recipe to remedy some of the most frequent issues of GAN training. For instance, KLD tends to cover all the modes of the real distribution while reverse KLD tends to select a subset of them [45], [49]–[52] (see also Fig. 3 for a benchmark). Therefore, if mode collapse is observed during training, then, increasing γ with $\beta = 1-\gamma$ will push the generator toward generating a wider variety of samples. In the other limit, more realistic samples (e.g., less blurry images) but with less variability will be generated when β is increased while $\gamma = 1 - \beta$.

We also expand the interpretation of the (β, γ)-plane to the case where $\beta + \gamma > 0$ as the following proposition demonstrates. We show that the half-lines beginning at the origin and passing through the line $\beta + \gamma = 1$ define the same

[3]Forward χ^2-divergence is often called Pearson χ^2-divergence while the reverse χ^2-divergence is often called Neyman χ^2-divergence.

divergence therefore we call them divergence rays or d-rays for shorthand as depicted in Fig. 1 (gray half-lines).

Proposition 2: Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$ and $\delta \in [\delta_{\min}, \delta_{\max}]$ with $0 < \delta_{\min} < \delta_{\max} < \infty$ and $\mathcal{D}$ be the set of all bounded and measurable functions. Then, the optimization of cumulant loss in (4) is equivalent to the minimization of scaled Rényi divergence for $(\beta, \gamma) = (\delta(1-\alpha), \delta\alpha)$

$$\min_G \max_{D\in\mathcal{D}} L(\delta(1-\alpha), \delta\alpha) \equiv \min_G \frac{1}{\delta}\mathcal{R}_\alpha\big(p_r||p_g\big).$$

Proof: The maximization part of cumulant GAN becomes

$$\begin{aligned}
&\max_{D\in\mathcal{D}} L(\delta(1-\alpha), \delta\alpha) \\
&\quad= \max_{D\in\mathcal{D}}\Bigg\{-\frac{1}{\delta(1-\alpha)}\log\mathbb{E}_{p_r}\big[e^{-\delta(1-\alpha)D(x)}\big] \\
&\qquad\qquad -\frac{1}{\delta\alpha}\log\mathbb{E}_{p_g}\big[e^{\delta\alpha D(x)}\big]\Bigg\} \\
&\quad= \frac{1}{\delta}\max_{D'=\delta D\in\mathcal{D}}\Bigg\{\frac{1}{(\alpha-1)}\log\mathbb{E}_{p_r}\Big[e^{(\alpha-1)D'(x)}\Big] \\
&\qquad\qquad -\frac{1}{\alpha}\log\mathbb{E}_{p_g}\Big[e^{\alpha D'(x)}\Big]\Bigg\} \\
&\quad= \frac{1}{\delta}\mathcal{R}_\alpha\big(p_r||p_g\big)
\end{aligned}$$

which completes the proof since δ is positive and far from 0 or ∞ thus $\delta D \in \mathcal{D}$. □

Remark 2: From a practical perspective, the boundedness condition required in the above theoretical formulation can be easily enforced by considering a clipped discriminator with clipping factor M, i.e., $D_M(x) = M\tanh((D(x)/M))$. On the other hand, the set of all measurable functions is a very large class of functions and it might be difficult to be represented by a neural network. However, one can approximate measurable functions with continuous functions via Lusin's theorem [53] which states that every finite Lebesgue measurable function is approximated arbitrarily well by a continuous function except on a set of arbitrarily small Lebesgue measure. Therefore, a sufficiently large neural network can accurately approximate any measurable function.

C. Cumulant GAN as a Weighted Version of the SGDA Algorithm

The parameter estimation for the cumulant GAN is performed using the SGDA algorithm. Algorithm 1 presents the core part of SGDA's update steps where we exclude any regularization terms for clarity purposes. Namely, η and θ are the parameters of the discriminator and the generator, respectively, while λ is the learning rate. The proposed loss function is not the difference of two expected values; therefore, the order between differentiation and expectation approximation does matter. We choose to first approximate the expected values with the respective statistical averages as

$$\hat{L}_m(\beta, \gamma) = -\frac{1}{\beta}\log\sum_{i=1}^{m} e^{-\beta D(x_i)} - \frac{1}{\gamma}\log\sum_{i=1}^{m} e^{\gamma D(G(z_i))}. \quad (5)$$

Then, we apply the differentiation operator which results in a weighted version of SGDA as shown in Algorithm 1.

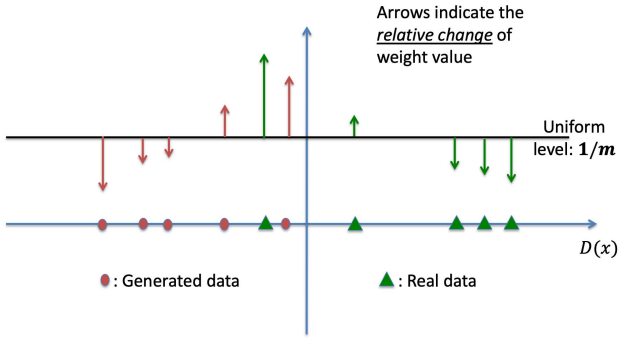


Fig. 2. Interpretation of cumulant GAN as a weighted variation of SGDA for $\beta, \gamma > 0$. Both real and generated samples for which the discriminator outputs a value closer to the decision boundary are assigned with larger weights because these are the samples which most probably confuse the discriminator.

Interestingly, several recent articles [52], [54]–[57] included a weighting perspective in their optimization approach.

Algorithm 1 Core of SGDA Iteration

Input: data batch: $\{x_i\}$, noise batch: $\{z_i\}$

for k steps **do**

$$\eta \leftarrow \eta + \lambda\left(\sum_{i=1}^{m} w_i^{\beta}\nabla_\eta D(x_i) - \sum_{i=1}^{m} w_i^{\gamma}\nabla_\eta D(G(z_i))\right)$$

end for

$$\theta \leftarrow \theta + \lambda\left(\sum_{i=1}^{m} w_i^{\gamma}\nabla_\theta D(G(z_i))\right)$$

The difference between WGAN and cumulant GAN for the update steps is the weights w_i^{β} and w_i^{γ}. In WGAN, the weights are constant and equal to $(1/m)$ while in cumulant GAN they are defined for any $i = 1, \ldots, m$ by

$$w_i^{\beta} = \frac{e^{-\beta D(x_i)}}{\sum_{j=1}^{m} e^{-\beta D(x_j)}}, \quad \text{and} \quad w_i^{\gamma} = \frac{e^{\gamma D(G(z_i))}}{\sum_{j=1}^{m} e^{\gamma D(G(z_j))}}.$$

The weights redistribute the sample distributions based on the assessment of the current discriminator. Fig. 2 qualitatively demonstrates the change of the weight relative to uniform weights for $\beta, \gamma > 0$. The weights place more emphasis on the real samples associated with the smallest $D(x_i)$ values. Similarly they place more emphasis on the synthetic samples that give the highest $D(G(z_i))$ values. A quantitative demonstration of the weights and how they evolve during the training process is presented in Appendix C (see Figs. 1–3) of Supplementary Materials.

The intuition behind the weighting mechanism is that samples that confuse the discriminator, i.e., the samples around the "fuzzy" decision boundary, are more valuable for the training process than samples that are easily distinguished, thus, they should weigh more. Essentially, the discriminator is updated with samples produced by a better generator than the current one, as well as with more challenging real samples. Similarly,

the generator is also updated using samples from a generator which is better than the current one. Overall, due to the use of the weights $w_i^{\beta}, w_i^{\gamma}$ in Algorithm 1, both generator and discriminator updates will be more affected by synthetic samples that are more indistinguishable from the real ones.

Additionally, the update of the discriminator is performed k times more than the generator's update offering two important advantages. First, more iterations for the discriminator implies that it better distinguishes the real data from the generated ones, making the weighting perspective more valid. Second, it better approximates the optimal discriminator, thus, the theory presented in the previous section becomes more credible in practice.

Remark 3: The Monte Carlo approximation in (5) is biased. However, it has been shown that it is consistent [52], hence, the error due to the statistical approximation decreases as the size of minibatch increases. Bias correction gradients using moving averages have been utilized in [42] for the estimation of CGF. However, the modification of the loss function and the lack of an interpretation analogous to the weights $w_i^{\beta}, w_i^{\gamma}$ are two key reasons to avoid inserting any bias-correction mechanism.

D. Convergence Guarantees for Linear Discriminator

Let $\mathcal{D}$ be the set of all linear functions (i.e., $D(x) = \eta^T x$ with $\eta, x \in \mathbb{R}^d$) and assume that the real data follow a Gaussian distribution with mean value $\mu \in \mathbb{R}^d$ and covariance matrix, I_d. The generator is defined by $G(z) = z + \theta$, where z is a standard d-dimensional Gaussian. The loss function for WGAN is[4]

$$\min_{\theta} \max_{\eta} \ \eta^T(\mu - \theta) \tag{6}$$

while the respective exact cumulant loss function from (4) is given by [58]

$$\min_{\theta} \max_{\eta} \frac{1}{-\beta}\left(-\beta\eta^T\mu + \frac{1}{2}\beta^2\eta^T\eta\right) - \frac{1}{\gamma}\left(\gamma\,\eta^T\theta + \frac{1}{2}\gamma^2\eta^T\eta\right) \equiv \min_{\theta} \max_{\eta} \ \eta^T(\mu - \theta) - \frac{\beta+\gamma}{2}\eta^T\eta. \tag{7}$$

It has been proven that the training dynamics oscillates without converging to the optimum for the WGAN loss function (6) if gradient descent ascent (GDA) is used [36] and more sophisticated algorithms such as training with optimism [30] or two-step extra-gradient approaches [59] are required to guarantee convergence. The use of CGFs transforms the optimization problem from just concave to a strongly concave problem for η. Actually, the cumulant loss function (7) is $((\beta+\gamma)/2)$-strongly concave. When the loss is both strongly convex and strongly concave, the GDA algorithm converges linearly (i.e., exponentially fast) to the optimal solution under efficient proximal mappings admission [60]. Our case, where the loss is not strongly-convex with respect to θ but it is strongly-concave for η, has also linear convergence when the coupling term between η and θ is full-column rank and the learning rates are properly chosen as it has been shown in [61]. The following theorem demonstrates that the training dynamics for the cumulant loss function (7) converges even when the GDA algorithm uses the same learning rate for both players which is the main difference between our result and [61]. Next, without loss of generality, we assume $\gamma = 0$.

Theorem 2: The GDA method with learning rate λ converges exponentially fast to the (unique) Nash equilibrium with rate $1-(1-\epsilon)\lambda\beta$ if $\beta \in (\lambda/\epsilon, 1)$ with $\lambda < \epsilon < 1$. Mathematically, for the tth iteration of GDA we have

$$||(\theta_t, \eta_t) - (\mu, 0)||_2^2 \leq c(1-(1-\epsilon)\lambda\beta)^t \tag{8}$$

where $(\theta^*, \eta^*) = (\mu, 0)$ is the Nash equilibrium while c is a computable positive constant.

Proof: The update step of GDA for the cumulant loss is given by

$$\eta_{t+1} = \eta_t + \lambda(\mu - \theta_t - \beta\eta_t)$$
$$\theta_{t+1} = \theta_t + \lambda\eta_t.$$

Define the energy function

$$E(\eta, \theta) = \eta^T\eta - \beta\eta^T(\mu - \theta) + (\mu - \theta)^T(\mu - \theta).$$

$E(\eta, \theta)$ is a second order polynomial for η; it is straightforward to show that if $0 < \beta < 1$ then $E(\eta, \theta) \geq 0$ for all η and θ and it is equal to 0 iff $\eta = \eta^* = 0$ and $\theta = \theta^* = \mu$. Additionally, it generally holds that

$$||(\theta, \eta) - (\mu, 0)||_2^2 \leq 2E(\eta, \theta)$$

since $2E(\eta, \theta) - ||(\theta, \eta) - (\mu, 0)||_2^2 = \eta^T\eta - 2\beta\eta^T(\mu - \theta) + (\mu - \theta)^T(\mu - \theta) \geq 0$ for all $0 < \beta < 1$.

Next, we show that $E(\eta_t, \theta_t)$ converges exponentially fast to 0. Since, $E(\eta, \theta) = \sum_{i=1}^{d} \eta_i^2 - \beta\eta_i(\mu_i - \theta_i) + (\mu_i - \theta_i)^2$, we can proceed with $d = 1$ without sacrificing the generality of the proof. Using symbolic calculations, we obtain

$$\begin{aligned} E(\eta_{t+1}, \theta_{t+1}) &= (1-(1-\epsilon)\lambda\beta)E(\eta_t, \theta_t) \\ &\quad - \lambda(\epsilon\beta - \lambda)\left[\eta_t^2 - \beta\eta_t(\mu - \theta_t) + (\mu - \theta_t)^2\right] \\ &\leq (1-(1-\epsilon)\lambda\beta)E(\eta_t, \theta_t) \end{aligned}$$

since $\eta_t^2 - \beta\eta_t(\mu - \theta_t) + (\mu - \theta_t)^2 \geq 0$ for $\beta < 1$ and $\beta > \lambda/\epsilon$. The iterative application of this inequality yields

$$E(\eta_{t+1}, \theta_{t+1}) \leq (1-(1-\epsilon)\lambda\beta)^{t+1}E(\eta_0, \theta_0).$$

Combining the above inequalities, we prove (8) with $c = 2E(\eta_0, \theta_0)$. □

Finally, we remark that similar to previous studies that prove linear convergence [60], [61], our proof utilizes the concept of energy functions (a.k.a. Lyapunov functionals), a tool from the theory of Dynamical Systems that has the potential to be transferable to more general optimization problems, too.

IV. Demonstrations

The source code of our demonstration examples is available online.[5]

[4]We did not add the GP in the current formulation since it has been shown that the convergence behavior of the gradient descent/ascend algorithm is not affected [30, Appendix B].

[5]https://github.com/dipjyoti92/CumulantGAN

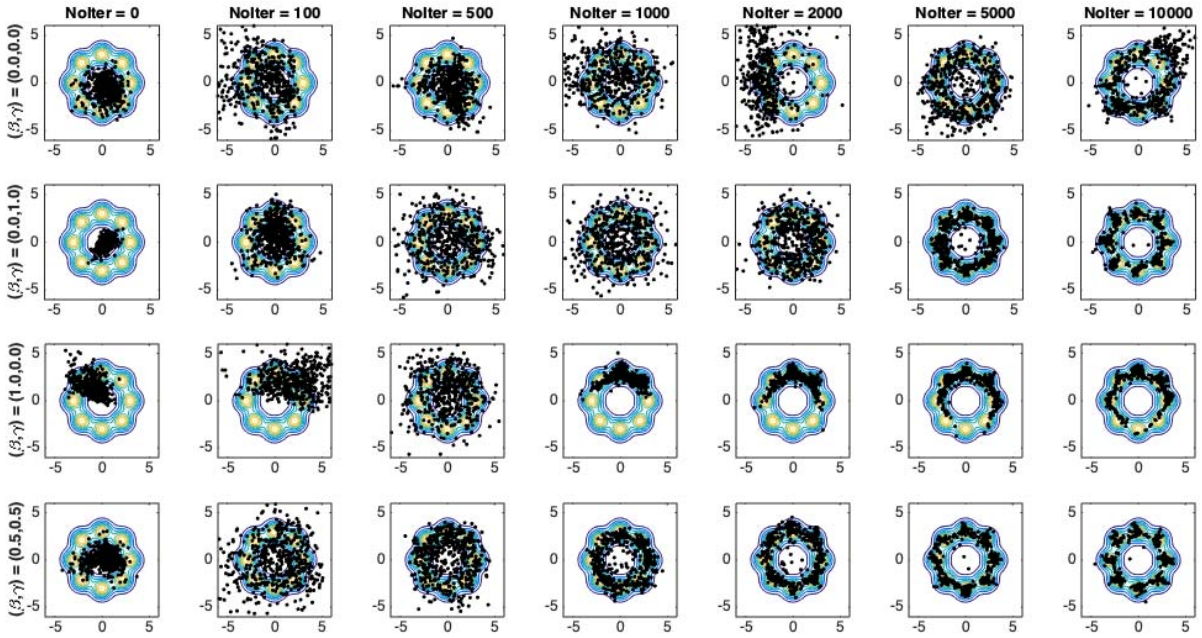


Fig. 3. Generated samples using the Wasserstein distance using clipping (1st row), KL divergence (2nd row), reverse KLD (3rd row), and Hellinger distance (last row). The boundedness condition is not enforced on this example but it is necessary to be satisfied when the hyper-parameters take negative values.

A. Traversing the (β, γ)-Plane: From Mode Covering to Mode Selection

As demonstrated in Section III-B and Fig. 1, the optimization of cumulant GAN for the set of bounded and measurable functions and various hyper-parameter values is equivalent to the minimization of a divergence. It is well known that different divergences result in fundamentally different behavior of the solution. For instance, KLD minimization tends to produce a distribution that covers all the modes while the reverse KLD tends to produce a distribution that is focused on a subset of the modes [49]–[51]. Taking the extreme cases, an all-mode covering is obtained as $\beta \to -\infty$ while largest mode selection is observed at the other limit direction.

Our first example aims at highlighting the above characteristics of divergences and additionally to verify that the suboptimal approximation of the function space of all bounded functions by a family of neural networks does not significantly affect the expected outcomes. Fig. 3 presents generated samples for various values of the (β, γ) pair at different stages of the training process as quantified by the number of iterations (denoted by "NoIter" in the Figure). The target distribution is a mixture of eight equiprobable and equidistant-from-the-origin Gaussian random variables. Both discriminator and generator are neural networks with two hidden layers with 32 units each and ReLU as activation function. Input noise for the generator is an 8-dimensional standard Gaussian. In all cases, the discriminator is updated $k = 5$ times followed by an update for the generator.

KLD minimization that corresponds to $(\beta, \gamma) = (0, 1)$ (second row) tends to cover all modes while reverse KLD that corresponds to $(\beta, \gamma) = (1, 0)$ (third row) tends to select a subset of them. This is particularly evident when the number of iterations is between 500 and 2000. Hellinger distance minimization (last row) produces samples with statistics that lie between KLD and reverse KLD minimization while Wasserstein distance minimization (first row) has a less controlled behavior. It is also noteworthy that reverse KLD was not able to fully cover all the modes after 10 K iterations. This behavior is not necessarily a drawback since the divergence of choice is primarily an application-specific decision. For instance, the lack of diversity might be sacrificed in image generation for the sake of sharpness of the synthetic images.

We note that despite demonstrating a single run, the plots in Fig. 3 are not cherry-picked. We have tested several architectures with more or fewer layers, as well as more or fewer units per layer, repeating each run several times, with qualitatively similar results which are presented in Appendix E-A (Figs. 4–6) of Supplementary Materials. We further tested and compared the performance of various hyper-parameter values of cumulant GAN on two additional data distributions and presented them in Appendix E-B (Figs. 7–12). The first dataset is a mixture of six equiprobable Student's t distributions while the second dataset is the Swiss-roll distribution. Overall, cumulant GAN with $(\beta, \gamma) = (0.5, 0.5)$ (i.e., Hellinger distance[6]) generated the most accurate results for all datasets and across various architectures. Finally, we experimented with d-rays and showed in Appendix E-C (see Figs. 13–15) that the training process is qualitatively similar in terms of "mode covering" versus "mode selection" across the divergence rays.

B. Learning the Covariance Matrix of a Multivariate Gaussian

A CGF can uniquely determine a distribution and contains information on all moments. Therefore, the use of simple

[6]Actually, we minimize $-4\log(1 - Hel^2)$, see Corollary 1.

discriminators which may fail under the WGAN loss might be sufficient under the cumulant loss to successfully train the generator. In this section, we provide an explicit example where the discriminator despite being a linear function the target is to learn the second-order statistic of a multivariate Gaussian distribution. Thus, the real data, $x \in \mathbb{R}^d$, follow a zero-mean Gaussian with covariance matrix Σ, the discriminator is given by $D(x) = \eta^T x$ while the generator is given by $G(z) = Az$ where A is a $d \times k$ matrix and z is a standard k-dimensional Gaussian. The aim is to obtain a solution, $\hat{\Sigma} = \hat{A}\hat{A}^T$, close to the true covariance matrix.

The loss function of WGAN is $L(0,0) = \eta^T \mathbb{E}_{p_r}[x] - \eta^T A \mathbb{E}_{p_z}[z] = 0$, therefore it is impossible here to learn the covariance matrix. On the other hand, the cumulant loss reads

$$L(\beta, \gamma) = -\frac{1}{2}\eta^T \left(\beta \Sigma + \gamma AA^T\right)\eta \tag{9}$$

allowing the possibility of a (β, γ) pair that makes the Nash equilibrium non-trivially informative regarding the covariance matrix. Indeed, we calculated the best response diagrams for $d = 1$ with fixed positive values of γ and inferred that suitable values are $\beta \ll -1$. Fig. 4 presents the average error of the covariance matrix evaluated using the Frobenius norm as a function of β. The covariance is computed using either the above exact loss function (upper plot) or the statistical approximation of the cumulant loss along with SGDA algorithm (lower plot) for three values of γ. We use 10 K samples for the latter case, average over ten iterations, and a different covariance matrix is used at each iteration. The true covariance matrix is rescaled so that its Frobenius norm equals to 1. We observe that the covariance matrix is learned satisfactorily when the exact loss function is used for large negative values of β. When the approximated, yet realistic, loss is used, the error between the true and the estimated covariance matrices increases after a certain value of $-\beta$ because tail statistics (requiring a large amount of samples) start to take control. Overall, the direct conclusion is that cumulant GAN is able to learn higher order statistics and produce samples with the correct covariance structure despite the fact that a very simple discriminator without any access to higher order statistics was deployed.

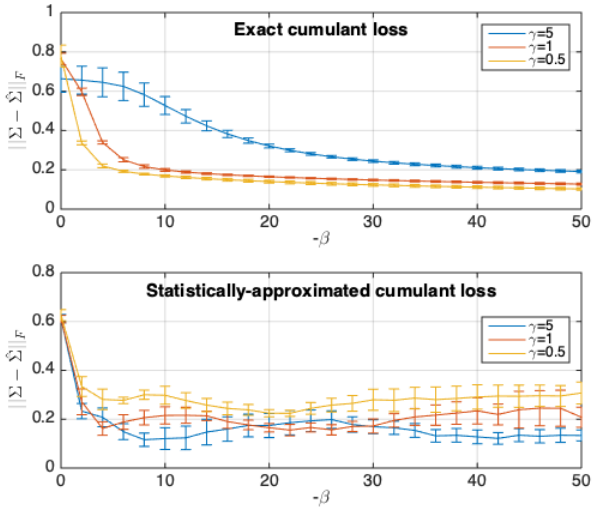


Fig. 4. Covariance estimation error for the exact cumulant loss function (upper plot) and for the statistically approximated cumulant loss function (lower plot).

C. Improved Image Generation

A series of experiments have been conducted on CIFAR-10 [62] and ImageNet [63] datasets demonstrating the effectiveness of cumulant GAN. In the experiments, we select pairs of (β, γ) that correspond to well-known divergences to highlight their effect on the training process as well as to facilitate connections with existing literature.

1) CIFAR-10 Dataset: CIFAR-10 is a well-studied dataset of $32 \times 32 \times 3$ RGB color images with ten classes. We evaluate the quality of the generated images using four different architectures: one with convolutional layers (CNN) and three with residual blocks (resnet). The generator for the CNN consists of one linear layer followed by three convolutional layers while the discriminator is a single convolutional layer followed by one linear layer. The generator for the three resnets consists of four residual blocks while the discriminator consists of two or three residual blocks. We train two versions with three residual blocks for the discriminator but with different channel dimension and learning rate. The complete description of the architectures can be found in Appendix F of Supplementary Materials. In all cases, we deliberately choose a weaker discriminator to challenge the training procedure.

Tables I and II report the averaged IS [32] and the averaged FID [64] along with their standard deviation over five runs for the four architectures. We test four different hyper-parameter values that correspond to minimization of Wasserstein distance, KLD, reverse KLD, and Hellinger distance [actually, $-4\log(1-\text{Hel}^2)$]. We use the two-sided GP for WGAN since it has been shown to provide better performance than the one-sided version [38]. However, the two-sided GP is not valid for cumulant GAN [40] therefore we enforce the one-sided version of the GP. In all cases, the optimization for the discriminator is realized over Lipschitz continuous functions. The implementation of cumulant GAN is based on available open-source code.[7] Following the reference code, we train the models with the Adam optimizer and the discriminator's parameters are updated $k = 5$ times more often than the parameters of the generator.

We remind that IS is a standard metric to evaluate the visual quality of generated image samples [32]. IS correlates with human judgment by feeding generated samples into a pre-trained Inception v3 classifier. Images with naturally looking objects are supposed to have low label (output) entropy whereas the entropy across images should be high. On the other hand, FID score uses the Inception v3 model activation layers (last pooling layer) to capture latent features calculated for a collection of real and generated images. The activation values are summarized as a multivariate Gaussian by calculating the mean and covariance of both real and generated images. The distance between these two distributions is then calculated using the Fréchet distance, also called the

[7] https://github.com/igul222/improved_wgan_training

TABLE I
MEAN IS ON CIFAR-10 AND IMAGENET

	CIFAR-10				ImageNet	
Architecture / Loss function	Conv layers Gen: 3 & Dis: 1	Residual blocks Gen: 4 & Dis: 2	Residual blocks (V1) Gen: 4 & Dis: 3	Residual blocks (V2) Gen: 4 & Dis: 3	Residual blocks Gen: 4 & Dis: 2	Residual blocks Gen: 4 & Dis: 4
Wasserstein	4.36 ± 0.10	4.58 ± 0.14	5.25 ± 0.23	6.45 ± 0.34	5.13 ± 0.45	8.88 ± 1.15
KLD	**4.81 ± 0.07**	7.63 ± 0.07	**7.42 ± 0.08**	7.28 ± 0.12	**8.86 ± 0.08**	10.03 ± 0.12
Reverse KLD	4.56 ± 0.13	**7.68 ± 0.08**	7.28 ± 0.05	**7.39 ± 0.08**	8.70 ± 0.33	**10.23 ± 0.13**
Hellinger	**4.82 ± 0.10**	**7.69 ± 0.06**	7.22 ± 0.08	**7.35 ± 0.09**	8.55 ± 0.23	**10.24 ± 0.58**

TABLE II
MEAN FID ON CIFAR-10 AND IMAGENET

	CIFAR-10				ImageNet	
Architecture / Loss function	Conv layers Gen: 3 & Dis: 1	Residual blocks Gen: 4 & Dis: 2	Residual blocks (V1) Gen: 4 & Dis: 3	Residual blocks (V2) Gen: 4 & Dis: 3	Residual blocks Gen: 4 & Dis: 2	Residual blocks Gen: 4 & Dis: 4
Wasserstein	173.66 ± 3.42	76.54 ± 4.04	71.77 ± 3.63	39.58 ± 13.85	137.78 ± 11.20	64.26 ± 16.06
KLD	**157.36 ± 2.02**	19.81 ± 0.55	**22.51 ± 1.75**	23.06 ± 1.20	**67.45 ± 1.74**	49.39 ± 0.75
Reverse KLD	**156.23 ± 6.51**	**18.74 ± 0.58**	24.62 ± 0.90	**21.54 ± 1.36**	72.96 ± 5.57	**45.91 ± 1.40**
Hellinger	**158.60 ± 2.96**	**18.66 ± 0.54**	24.59 ± 0.63	**21.15 ± 1.25**	69.88 ± 1.98	**45.80 ± 3.77**

two-Wasserstein distance. We use 50 K images to compute IS/FID scores. Higher IS means better generated image quality whereas the best generative model returns the lowest FID score.

We observe from the tables as well as from the panels of Fig. 5 which present the averaged IS as a function of the number of iterations for the four architectures that all hyper-parameter choices for cumulant GAN outperform the baseline WGAN. The relative improvement ranged from 4.6% (reverse KLD) up to 10.5% (Hellinger distance) for the CNN architecture while the relative improvement for the resnet with the weaker discriminator ranged from 66.6% (reverse KLD) up to 67.9% (Hellinger distance) revealing that cumulant GAN takes into consideration all discriminator's moments, i.e., all higher order statistics and not just the mean values, leading to better realization of the target distribution. Cumulant GAN achieves higher ISs than WGAN for the two versions of resnets with three residual blocks for the discriminator (lower panels in Fig. 5), too. As expected, cumulant GANs also perform better on FID metric with relative improvements up to 75.62% (Hellinger distance) for the resnet with the weaker discriminator and 68.64% (KLD) and 46.56% (Hellinger distance) for the two versions of resnets, respectively. All cumulant GAN variations (KLD, reverse KLD, and Hellinger) obtain similar results for both versions while the performance of WGAN is significantly affected by the choice of the hyper-parameter values, e.g., learning rate and channel dimension. This discrepancy in the performance highlights the enhanced robustness of cumulant GAN relative to WGAN implying that cumulant GAN may require less tuning to enjoy excellent performance. Finally, the samples generated by cumulant GAN also exhibit larger diversity and are visually better (we refer to Appendix G in Supplementary Materials).

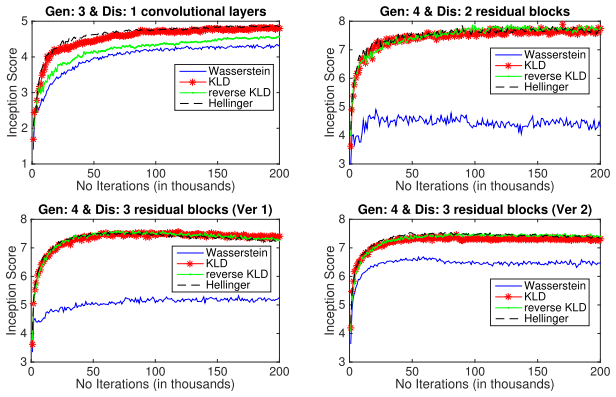


Fig. 5. IS for CIFAR-10 using various hyper-parameters of cumulant GAN and various architectures. In all cases, WGAN has a lower IS relative to the cumulant GAN with the hyper-parameter corresponding to Hellinger minimization achieving the best overall performance.

2) ImageNet Dataset: This large dataset consists of 64 × 64 × 3 color images with 1000 object classes. The large number of classes is challenging for GAN training due to the tendency to underestimate the entropy in the distribution [32]. We evaluate the performance on two different architectures which both have a generator with four residual blocks. The difference is in the number of residual blocks for the discriminator where we employ a weak discriminator with two residual blocks and a strong discriminator with four residual blocks. Fig. 6 presents the performance in terms of IS both for the baseline WGAN and for the variants of cumulant GAN when a weak discriminator (left panel) or a strong discriminator (right panel) is utilized. Improved ISs are obtained with cumulant GAN for both architectures. It is also important to note that our approach is much more effective than WGAN

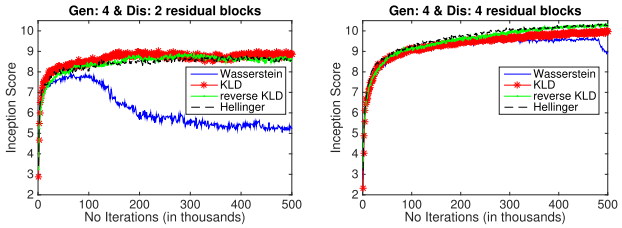


Fig. 6. Same as Fig. 5 but for ImageNet. Cumulant GAN achieves higher IS relative to WGAN for both weak (left panel) and strong (right panel) discriminator. Mode collapse has been mitigated in all cumulant GAN variants.

at avoiding mode collapse while still generating high-quality samples. The mean ISs along with the standard deviation over three repetitions are reported in Table I (rightmost columns). In terms of relative improvement, cumulant GAN is between 72.71% (KLD) to 69.59% (reverse KLD) better than WGAN for the weak discriminator and a similar trend is observed when the strong discriminator is used reaching IS as high as 10.24. As reported in Table II, the proposed cumulant GAN is superior relative to WGAN in generating high-quality images with low FID scores of 67.45 (KLD) for the weak discriminator and 45.80 (Hellinger distance) for the strong discriminator. By visual inspection of the generated images (Appendix G in Supplementary Materials), we conclude that all generators learn some basic and contiguous shapes with natural color and texture. Nevertheless, cumulant GAN provides better images with object specifications that are clearly more realistic.

Despite not being exhaustive, the presented examples demonstrated a preference of cumulant GAN over WGAN. In general, GAN optimization has essentially two critical components: the first being the function space where the discriminator lives while the other is the objective functional to be optimized. WGAN's breakthrough was the restriction of the function space to Lipschitz continuous functions that resulted in increased stability. However, there is no evidence that the best-performing loss function is the difference of two expectations as in WGAN. The presented examples revealed that there are better and more flexible options for the overall loss function and the proposed cumulant loss is one of them.

V. Conclusion and Future Directions

We proposed cumulant GAN by establishing a novel loss function based on the CGF of both real and generated distributions. The use of CGFs allows for an inclusive characterization of the distributions' statistics, making it possible to partially remove complexity from the discriminator. The net result is improved and more stable training of GANs. Furthermore, cumulant GAN has the capacity to smoothly interpolate between a wide range of divergences and distances by simply changing its two hyper-parameter values β & γ. Thus, it offers a flexible and comprehensive mechanism to choose–possibly adaptively–which objective to minimize. In a recent publication [65], the authors applied cumulant GAN for disentangled representation learning of speech signals and our plan is to further explore the improved capabilities of cumulant GAN in a variety of estimation and inference applications. Finally, the substitution of an expectation operator with the respective CGF does not have to be limited to WGAN. It can be applied to other GANs' loss function resulting in new GAN formulations. The theoretical and empirical ramifications of such extensions are left as future work.

References

[1] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[3] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 1486–1494.
[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[5] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn. (PMLR)*, vol. 70, 2017, pp. 2642–2651.
[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.
[7] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–35.
[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
[9] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, M. F. Balcan and K. Q. Weinberger, Eds. New York, NY, USA: PMLR, Jun. 2016, pp. 1060–1069.
[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[11] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[12] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.
[13] Y. Saito, S. Takamichi, H. Saruwatari, Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 26, no. 1, pp. 84–96, Jan. 2017.
[14] K. Kumar *et al.*, "MELGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14881–14892.
[15] T. Che *et al.*, "Maximum-likelihood augmented discrete generative adversarial networks," *CoRR*, vol. abs/1702.07983, 2017.
[16] W. Fedus, I. Goodfellow, and A. M. Dai, "MaskGAN: Better text generation via filling in the_," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17.
[17] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=BkJ3ibb0-
[18] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "APE-GAN: Adversarial perturbation elimination with GAN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3842–3846.
[19] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[20] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.
[21] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2022, doi: 10.1145/3439723.

[22] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," 2020, *arXiv:2001.06937*. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr2001.html#abs-2001-06937

[23] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," 2020, *arXiv:2006.05132*.

[24] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[25] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2962–2971, doi: 10.1109/CVPR.2017.316.

[26] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds., vol. 80, Jul. 2018, pp. 1989–1998. [Online]. Available: http://proceedings.mlr.press/v80/hoffman18a.html

[27] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2180–2188.

[28] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, Jun. 2019, pp. 7354–7363. [Online]. Available: http://proceedings.mlr.press/v97/zhang19d.html

[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[30] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training GANs with optimism," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–30.

[31] X. Qian, X. Cheng, G. Cheng, X. Yao, and L. Jiang, "Two-stream encoder GAN with progressive training for co-saliency detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 180–184, 2021.

[32] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2234–2242.

[33] S. Nowozin, B. Cseke, and R. Tomioka, "F-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[34] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[35] J. Hyun Lim and J. Chul Ye, "Geometric GAN," 2017, *arXiv:1705.02894*.

[36] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras, "Cycles in adversarial regularized learning," in *Proc. 29th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2018, pp. 2703–2717.

[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[39] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos, "MMD-GAN: Towards deeper understanding of moment matching network," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/dfd7468ac613286cdbb40872c8ef3b06-Paper.pdf

[40] J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, "(f, Γ)-divergences: Interpolating between f-divergences and integral probability metrics," *J. Mach. Learn. Res.*, vol. 23, no. 39, pp. 1–70, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-0100.html

[41] J. Birrell, P. Dupuis, M. A. Katsoulakis, L. Rey-Bellet, and J. Wang, "Variational representations and neural network estimation of Rényi divergences," *SIAM J. Math. Data Sci.*, vol. 3, no. 4, pp. 1093–1116, Jan. 2021, doi: 10.1137/20M1368926.

[42] M. I. Belghazi *et al.*, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[43] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, vol. 902. Hoboken, NJ, USA: Wiley, 2011.

[44] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations: A Mathematical Perspective*. Singapore: World Scientific, 2010.

[45] M. Shannon. (2020). *The Divergences Minimized by Non-Saturating GAN Training*. [Online]. Available: https://openreview.net/forum?id=BygY4grYDr

[46] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. IV," *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, Mar. 1983.

[47] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer, 2008.

[48] P. D. Grünwald *et al.*, *The Minimum Description Length Principle*, vol. 1. Cambridge, MA, USA: MIT Press, 2007.

[49] T. Minka *et al.*, "Divergence measures and message passing," Microsoft Res., Cambridge, U.K., Tech. Rep., 2005.

[50] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.

[51] J. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. Turner, "Black-box α-divergence minimization," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 1511–1520.

[52] Y. Li and R. E. Turner, "Rényi divergence variational inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1073–1081.

[53] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. New York, NY, USA: Wiley, 1999.

[54] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[55] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing, "On unifying deep generative models," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–22.

[56] R. D. Hjelm, A. P. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio, "Boundary-seeking generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17.

[57] Y. Pantazis, D. Paul, M. Fasoulakis, and Y. Stylianou, "Training generative adversarial networks with weights," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[58] B. Holmquist, "Moments and cumulants of the multivariate normal distribution," *Stochastic Anal. Appl.*, vol. 6, no. 3, pp. 273–278, Jan. 1988.

[59] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, vol. 108, 2020, pp. 1497–1507.

[60] G. H.-G. Chen and R. T. Rockafellar, "Convergence rates in forward-backward splitting," *SIAM J. Optim.*, vol. 7, no. 2, pp. 421–444, May 1997, doi: 10.1137/S1052623495290179.

[61] S. S. Du and W. Hu, "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, vol. 89, 2019, pp. 196–205.

[62] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[63] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

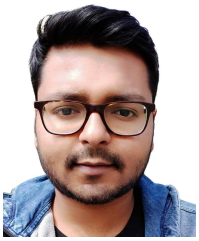
[64] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[65] D. Paul, S. Mukherjee, Y. Pantazis, and Y. Stylianou, "A universal multi-speaker multi-style text-to-speech via disentangled representation learning based on Rényi divergence minimization," in *Proc. INTERSPEECH*, 2021, pp. 1–12.

Yannis Pantazis received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Crete, Rethymno, Greece, in 2004, 2006, and 2010, respectively.

He was a Post-Doctoral Research Associate with the University of Massachusetts at Amherst, Amherst, MA, USA, from 2011 to 2015, and the University of Crete from 2016 to 2017. He has been a Researcher with the Institute of Applied and Computational Mathematics, Foundation of Research and Technology–Hellas, Heraklion, Greece, since 2018. His research interests include diverse areas of data science, such as machine learning, deep learning, generative models, speech/signal processing, information theory, uncertainty quantification, optimization theory, and dynamical systems.

Dr. Pantazis received the 1st Honorable Mention (Bronze Medal) on DARPA's Forecasting Chikungunya Virus Outbreak Competition in 2015.



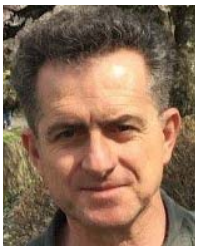
Dipjyoti Paul received the B.Tech. degree in electronics and communication engineering from the St. Thomas' College of Engineering and Technology under the West Bengal University of Technology, Kolkata, India, in 2013, and the M.Sc. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur (IIT Kharagpur), Kharagpur, India, in 2017. He is currently pursuing the Ph.D. degree in speech signal processing with the Department of Computer Science, University of Crete, Rethymno, Greece.

His research interests include diverse areas, such as machine/deep learning, signal/speech processing, computer vision, and optimization theory. He also has a broad interest in probabilistic machine learning methods and generative models.

Mr. Paul is also a member of ISCA.

Michail Fasoulakis received the B.Sc. degree in computer science and the M.Sc. degree in communications and signal processing from the Department of Computer Science, University of Crete, Rethymno, Greece, in 2008 and 2011, respectively, and the M.Sc. degree in computation and game theory from the University of Liverpool, Liverpool, U.K., in 2012, and the Ph.D. degree in computer science from the University of Warwick, Coventry, U.K., in 2017.

He was a Post-Doctoral Researcher with CWI, Amsterdam, The Netherlands, and the National Technical University of Athens, Athens, Greece. He is currently an Affiliated Researcher with the Institute of Computer Science (ICS), Foundation for Research and Technology–Hellas (FORTH), Heraklion, Greece, and the Department of Informatics, Athens University of Economics and Business, Athens, Greece. His research interests include (algorithmic) game theory and economics, algorithms, decision theory, operations research, machine learning, information theory, communications, and signal processing.



Yannis Stylianou (Fellow, IEEE) studied electrical engineering in NTUA Athens, Athens, Greece. He received the M.Sc. and Ph.D. degrees in signal processing from ENST-Paris, Paris, France, in 1992 and 1995, respectively.

From 1996 until 2001, he was with AT&T Labs Research (Murray Hill and Florham Park, NJ, USA), and until 2002, he was with Bell-Labs Lucent Technologies, Murray Hill, NJ, USA. He has been with the University of Crete, Rethymno, Greece, since 2002. From 2013 until 2018, he was the Group Leader of the Speech Technology Group, Toshiba Cambridge Research Laboratory, Cambridge, U.K. Since 2018, he has been with Apple, Cambridge. He is currently a Professor of speech processing with the University of Crete, and a Research Manager with Apple.

Dr. Stylianou was a member of the IEEE Speech and Language Technical Committee from 2007 to 2010 and from 2017 to 2019 and on the Board of the International Speech Communication Association (ISCA) from 2009 to 2013, of the IEEE Multimedia Communications Technical Committee. He is also a fellow of ISCA. He was a co-recipient of the 2018 ISCA Best Paper Award published in Speech Communication the period 2013–2017. He is or was on the Editorial Board of various journals, including the *Digital Signal Processing* journal (Elsevier) and the IEEE SIGNAL PROCESSING LETTERS.

Markos A. Katsoulakis received the bachelor's degree from the University of Athens, Athens, Greece, in 1987, and the Ph.D. degree in applied mathematics from Brown University, Providence, RI, USA, in 1993.

He is currently a Professor with the Department of Mathematics and Statistics, University of Massachusetts at Amherst, Amherst, MA, USA, where he is also the Director of the Center for Applied Mathematics and Computation. His research interests include probabilistic and multi-scale predictive modeling, uncertainty quantification, and information theory. He has also worked extensively in interdisciplinary collaborations involving chemical engineering, materials science, computer science, and atmospheric and oceanic sciences.

Dr. Katsoulakis was a member of the Editorial Board of the *SIAM Journal in Mathematical Analysis* from 2002 to 2014. He is a member of the editorial boards of the SIAM/ASA JOURNAL ON UNCERTAINTY QUANTIFICATION, the *SIAM Mathematical Modeling and Computation* Book Series, and the *Communications in Mathematical Sciences*. He is also a member of the Executive Committee of the UMass TRIPODS Institute for Theoretical Foundations of Data Science.