A Characteristic Function Approach to Deep Implicit Generative Modeling

Abdul Fatir Ansari†, Jonathan Scarlett†‡, and Harold Soh†
†Department of Computer Science
‡Department of Mathematics
National University of Singapore
{afatir, scarlett, harold}@comp.nus.edu.sg

Abstract

Implicit Generative Models (IGMs) such as GANs have emerged as effective data-driven models for generating samples, particularly images. In this paper, we formulate the problem of learning an IGM as minimizing the expected distance between characteristic functions. Specifically, we minimize the distance between characteristic functions of the real and generated data distributions under a suitably-chosen weighting distribution. This distance metric, which we term as the characteristic function distance (CFD), can be (approximately) computed with linear time-complexity in the number of samples, in contrast with the quadratic-time Maximum Mean Discrepancy (MMD). By replacing the discrepancy measure in the critic of a GAN with the CFD, we obtain a model that is simple to implement and stable to train. The proposed metric enjoys desirable theoretical properties including continuity and differentiability with respect to generator parameters, and continuity in the weak topology. We further propose a variation of the CFD in which the weighting distribution parameters are also optimized during training; this obviates the need for manual tuning, and leads to an improvement in test power relative to CFD. We demonstrate experimentally that our proposed method outperforms WGAN and MMD-GAN variants on a variety of unsupervised image generation benchmarks.

1. Introduction

Implicit Generative Models (IGMs), such as Generative Adversarial Networks (GANs) [12], seek to learn a model $\mathbb{Q}_\theta$ of an underlying data distribution $\mathbb{P}$ using samples from $\mathbb{P}$. Unlike prescribed probabilistic models, IGMs do not require a likelihood function, and thus are appealing when the data likelihood is unknown or intractable. Empirically, GANs have excelled at numerous tasks, from unsupervised image generation [18] to policy learning [17].

The original GAN suffers from optimization instability and mode collapse, and often requires various ad-hoc tricks to stabilize training [31]. Subsequent research has revealed that the generator-discriminator setup in the GAN minimizes the Jensen-Shannon divergence between the real and generated data distributions; this divergence possesses discontinuities that results in uninformative gradients as $\mathbb{Q}_\theta$ approaches $\mathbb{P}$, which hampers training. Various works have since established desirable properties for a divergence that can ease GAN training, and proposed alternative training schemes [2, 34, 3], primarily using distances belonging to the Integral Probability Metric (IPM) family [29]. One popular IPM is the kernel-based metric Maximum Mean Discrepancy (MMD), and a significant portion of recent work has focussed on deriving better MMD-GAN variants [21, 5, 1, 22].

In this paper, we undertake a different, more elementary approach, and formulate the problem of learning an IGM as minimizing the expected distance between *characteristic functions* of real and generated data distributions. Characteristic functions are widespread in probability theory and have been used for two-sample testing [15, 11, 8], yet surprisingly, have not yet been investigated for GAN training. We find that this approach leads to a *simple* and *computationally-efficient* loss: the characteristic function distance (CFD). Computing CFD requires linear time in the number of samples (unlike the quadratic-time MMD), and our experimental results indicate that CFD minimization results in effective training.

This work provides both theoretical and empirical support for using CFD to train IGMs. We first establish that the CFD is continuous and differentiable almost everywhere with respect to the parameters of the generator, and that it satisfies continuity in the weak topology – key properties that make it a suitable GAN metric [3, 21]. We provide novel direct proofs that supplement the existing theory on GAN training metrics. Algorithmically, our key idea is simple: train GANs using empirical estimates of the CFD under optimized weighting distributions. We report on systematic experiments using synthetic distributions and four benchmark image datasets (MNIST, CIFAR10, STL10, CelebA).

Our experiments demonstrate that the CFD-based approach outperforms WGAN and MMD-GAN variants on quantitative evaluation metrics. From a practical perspective, we find the CFD-based GANs are simple to implement and stable to train. In summary, the key contributions of this work are:

- a novel approach to train implicit generative models using a loss derived from characteristic functions;
- theoretical results showing that the proposed loss metric is continuous and differentiable in the parameters of the generator, and satisfies continuity in the weak topology;
- experimental results showing that our approach leads to effective generative models favorable against state-of-the-art WGAN and MMD-GAN variants on a variety of synthetic and real-world datasets.

2. Probability Distances and GANs

We begin by providing a brief review of the Generative Adversarial Network (GAN) framework and recent distance-based methods for training GANs. A GAN is a generative model that implicitly seeks to learn the data distribution $\mathbb{P}_{\mathcal{X}}$ given samples $\{\mathbf{x}\}_{i=1}^{n}$ from $\mathbb{P}_{\mathcal{X}}$. The GAN consists of a generator network g_θ and a critic network f_ϕ (also called the discriminator). The generator $g_\theta : \mathcal{Z} \to \mathcal{X}$ transforms a latent vector $\mathbf{z} \in \mathcal{Z}$ sampled from a simple distribution (e.g., Gaussian) to a vector $\hat{\mathbf{x}}$ in the data space. The original GAN [12] was defined via an adversarial two-player game between the critic and the generator; the critic attempts to distinguish the true data samples from ones obtained from the generator, and the generator attempts to make its samples indistinguishable from the true data.

In more recent work, this two-player game is cast as minimizing a divergence between the real data distribution and the generated distribution. The critic f_ϕ evaluates some probability divergence between the true and generated samples, and is optimized to maximize this divergence. In the original GAN, the associated (implicit) distance measure is the Jensen-Shannon divergence, but alternative divergences have since been introduced, e.g., the 1-Wasserstein distance [3, 14], Cramer distance [4], maximum mean discrepancy (MMD) [21, 5, 1], and Sobolev IPM [28]. Many distances proposed in the literature can be reduced to the Integral Probability Metric (IPM) framework with different restrictions on the function class.

3. Characteristic Function Distance

In this work, we propose to train GANs using a distance metric based on characteristic functions (CFs). Letting $\mathbb{P}$ be the probability measure associated with a real-valued random variable X, the characteristic function $\varphi_{\mathbb{P}} : \mathbb{R}^d \to \mathbb{C}$ of X is given by

$$\varphi_{\mathbb{P}}(\mathbf{t}) = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[e^{i\langle \mathbf{t},\mathbf{x}\rangle}] = \int_{\mathbb{R}} e^{i\langle \mathbf{t},\mathbf{x}\rangle} d\,\mathbb{P}, \tag{1}$$

where $\mathbf{t} \in \mathbb{R}^d$ is the input argument, and $i = \sqrt{-1}$. Characteristic functions are widespread in probability theory, and are often used as an alternative to probability density functions. The characteristic function of a random variable completely defines it, i.e., for two distributions $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{P} = \mathbb{Q}$ if and only if $\varphi_{\mathbb{P}} = \varphi_{\mathbb{Q}}$. Unlike the density function, the characteristic function always exists, and is uniformly continuous and bounded: $|\varphi_{\mathbb{P}}(t)| \leq 1$.

The squared Characteristic Function Distance (CFD) [8, 16] between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is given by the weighted integrated squared error between their characteristic functions

$$\mathrm{CFD}^2_{\omega}(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\mathbf{t}) - \varphi_{\mathbb{Q}}(\mathbf{t})|^2 \,\omega(\mathbf{t};\eta) d\mathbf{t}, \tag{2}$$

where $\omega(\mathbf{t};\eta)$ is a weighting function, which we henceforth assume to be parametrized by η and chosen such that the integral in Eq. (2) converges. When $\omega(\mathbf{t};\eta)$ is the probability density function of a distribution on $\mathbb{R}^d$, the integral in Eq. (2) can be written as an expectation:

$$\mathrm{CFD}^2_{\omega}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbf{t}\sim\omega(\mathbf{t};\eta)}\left[|\varphi_{\mathbb{P}}(\mathbf{t}) - \varphi_{\mathbb{Q}}(\mathbf{t})|^2\right]. \tag{3}$$

By analogy to Fourier analysis in signal processing, Eq. (3) can be interpreted as the expected discrepancy between the Fourier transforms of two signals at frequencies sampled from $\omega(\mathbf{t};\eta)$. If $\mathrm{supp}(\omega) = \mathbb{R}^d$, it can be shown using the uniqueness theorem of characteristic functions that $\mathrm{CFD}_{\omega}(\mathbb{P},\mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$ [35].

In practice, the CFD can be approximated using empirical characteristic functions and finite samples from the weighting distribution $\omega(\mathbf{t};\eta)$. To elaborate, the characteristic function of a degenerate distribution $\delta_{\mathbf{a}}$ for $\mathbf{a} \in \mathbb{R}^d$ is given by $e^{i\langle \mathbf{t},\mathbf{a}\rangle}$ where $\mathbf{t} \in \mathbb{R}^d$. Given observations $\mathcal{X} := \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ from a probability distribution $\mathbb{P}$, the empirical distribution is a mixture of degenerate distributions with equal weights, and the corresponding empirical characteristic function $\hat{\varphi}_{\mathbb{P}}$ is a weighted sum of characteristic functions of degenerate distributions:

$$\hat{\varphi}_{\mathbb{P}}(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n} e^{i\langle \mathbf{t},\mathbf{x}_j\rangle}. \tag{4}$$

Let $\mathcal{X} := \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ and $\mathcal{Y} := \{\mathbf{y}_1,\ldots,\mathbf{y}_m\}$ with $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$ be samples from the distributions $\mathbb{P}$ and $\mathbb{Q}$ respectively, and let $\mathbf{t}_1,\ldots,\mathbf{t}_k$ be samples from $\omega(\mathbf{t};\eta)$. We define the empirical characteristic function distance (ECFD) between $\mathbb{P}$ and $\mathbb{Q}$ as

$$\mathrm{ECFD}^2_{\omega}(\mathbb{P},\mathbb{Q}) = \frac{1}{k}\sum_{i=1}^{k} |\hat{\varphi}_{\mathbb{P}}(\mathbf{t}_i) - \hat{\varphi}_{\mathbb{Q}}(\mathbf{t}_i)|^2, \tag{5}$$

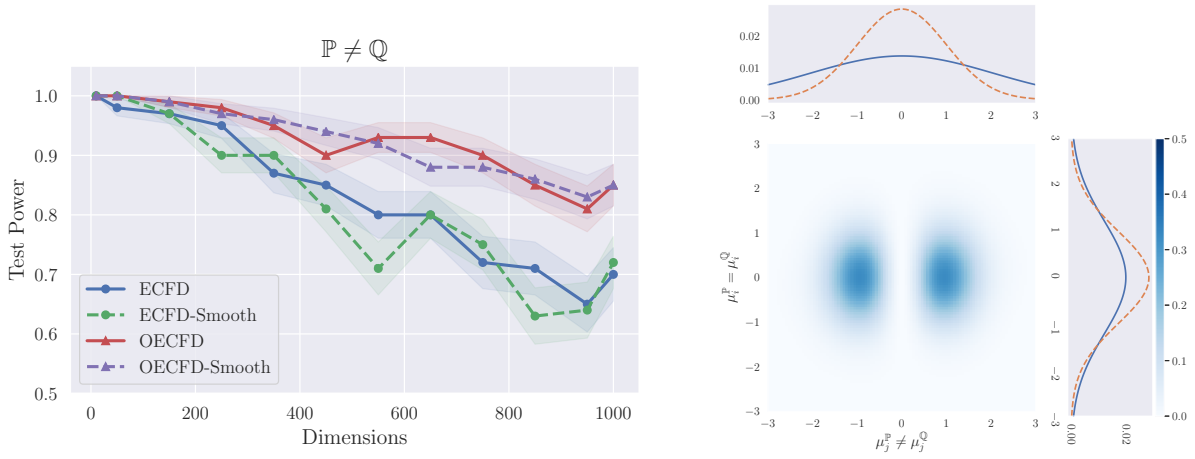

Figure 1: (left) Variation of test power with the number of dimensions for ECFD-based tests; (right) Change in the scale of the weighting distribution upon optimization.

where $\hat{\varphi}_{\mathbb{P}}$ and $\hat{\varphi}_{\mathbb{Q}}$ are the empirical CFs, computed using $\mathcal{X}$ and $\mathcal{Y}$ respectively.

A quantity related to CFD (Eq. 2) has been studied in [30] and [16], in which the discrepancy between the analytical and empirical characteristic functions of stable distributions is minimized for parameter estimation. The CFD is well-suited to this application because stable distributions do not admit density functions, making maximum likelihood estimation difficult. Parameter fitting has also been explored for other models such as mixture-of-Gaussians, stable ARMA process, and affine jump diffusion models [36].

More recently, [8] proposed fast ($O(n)$ in the number of samples n) two-sample tests based on ECFD, as well as a smoothed version of ECFD in which the characteristic function is convolved with an analytic kernel. The authors empirically show that ECFD and its smoothed variant have a better test-power/run-time trade-off compared to quadratic time tests, and better test power than the sub-quadratic time variants of MMD.

3.1. Optimized ECFD for Two-Sample Testing

The choice of $\omega(\mathbf{t}; \eta)$ is important for the success of ECFD in distinguishing two different distributions; choosing an appropriate distribution and/or set of parameters η allows better coverage of the frequencies at which the differences in $\mathbb{P}$ and $\mathbb{Q}$ lie. For instance, if the differences are concentrated at the frequencies far away from the origin and $\omega(\mathbf{t}; \eta)$ is Gaussian, the test power can be improved by suitably enlarging the variance of each coordinate of $\omega(\mathbf{t}; \eta)$.

To increase the power of ECFD, we propose to optimize the parameters η (e.g., the variance associated with a normal distribution) of the weighting distribution $\omega(\mathbf{t}; \eta)$ to maximize the power of the test. However, care should be taken when specifying how rich the class of functions $\omega(\cdot; \eta)$ is — the choice of *which* parameters to optimize and the associated constraints is important. Excessive optimization may cause the test to fixate on differences that are merely due to fluctuations in the sampling. As an extreme example, we found that optimizing $\mathbf{t}$'s directly (instead of optimizing the weighting distribution) severely degrades the test's ability to correctly accept the null hypothesis $\mathbb{P} = \mathbb{Q}$.

To validate our approach, we conducted a basic experiment using high-dimensional Gaussians, similar to [8]. Specifically, we used two multivariate Gaussians $\mathbb{P}$ and $\mathbb{Q}$ that have the same mean in all dimensions except one. As the dimensionality increases, it becomes increasingly difficult to distinguish between samples from the two distributions. In our tests, the weighting distribution $\omega(\mathbf{t}; \eta)$ was chosen to be a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^2))$, 10000 samples each were taken from $\mathbb{P}$ and $\mathbb{Q}$, and the number of frequencies (k) was set to 3. We optimized the parameter vector $\eta = \{\boldsymbol{\sigma}\}$ to maximize the ECFD using the Adam optimizer for 100 iterations with a batch-size of 1000.

Fig. 1a shows the variation of the test power (i.e., the fraction of times the null hypothesis $\mathbb{P} = \mathbb{Q}$ is rejected) with the number of dimensions. OEFCD refers to the optimized ECFD, and the "Smooth" suffix indicates the smoothed ECFD variant proposed by [8]. We see that optimization of η increases the power of ECFD and ECFD-Smooth, particularly at the higher dimensionalities. There do not appear to be significant differences between the optimized smoothed and non-smoothed ECFD variants. Moreover, the optimization improved the ability of the test to correctly distinguish the two different distributions, but did not hamper its ability to correctly accept the null hypothesis when the distributions are the same (see Appendix C).

To investigate how $\boldsymbol{\sigma}$ is adapted, we visualize two dimensions $\{i, j\}$ from the dataset where $\mu_i^{\mathbb{P}} = \mu_i^{\mathbb{Q}}$ and $\mu_j^{\mathbb{P}} \neq \mu_j^{\mathbb{Q}}$. Fig. 1b shows the absolute difference between the ECFs of $\mathbb{P}$ and $\mathbb{Q}$, with the corresponding dimensions of the weighting distribution plotted in both dimensions. The solid blue line shows the optimized distribution (for OECFD) while the dashed orange line shows the initial distribution (i.e., $\boldsymbol{\sigma} = 1$ for ECFD and ECFD-Smooth). In the dimension where the distributions are the same, σ has small deviation from the initial value. However, in the dimension where the distributions are different, the increase in variance is more pronounced to compensate for the spread of difference between the ECFs away from the origin.

4. Implicit Generative Modeling using CFD

In this section, we turn our attention to applying the (optimized) CFD for learning IGMs, specifically GANs. As in the standard GAN, our model is comprised of a generator $g_\theta : \mathcal{Z} \to \mathcal{X}$ and a critic $f_\phi : \mathcal{X} \to \mathbb{R}^m$, with parameter vectors θ and ϕ, and data/latent spaces $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Z} \subseteq \mathbb{R}^p$. Below, we write Θ, Φ, Π for the spaces in which the parameters θ, ϕ, η lie.

The generator minimizes the empirical CFD between the real and generated data. Instead of minimizing the distance between characteristic functions of raw high-dimensional data, we use a critic neural network f_ϕ that is trained to maximize the CFD between real and generated data distributions in a learned lower-dimensional space. This results in the following minimax objective for the IGM:

$$\inf_{\theta \in \Theta} \sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))}), \tag{6}$$

where $\psi = \{\phi, \eta\}$ (with corresponding parameter space Ψ), and η is the parameter vector of the weighting distribution ω. The optimization over η is omitted if we choose to not optimize the weighting distribution. In our experiments, we set $\eta = \{\boldsymbol{\sigma}\}$, with $\boldsymbol{\sigma}$ indicating the scale of each dimension of ω. Since evaluating the CFD requires knowledge of the data distribution, in practice, we optimize the empirical estimate ECFD_ω^2 instead of CFD_ω^2. We henceforth refer to this model as the Characteristic Function Generative Adversarial Network (CF-GAN).

4.1. CFD Properties: Continuity, Differentiability, and Weak Topology

Similar to recently proposed Wasserstein [3] and MMD [21] GANs, the CFD exhibits desirable mathematical properties. Specifically, CFD is continuous and differentiable almost everywhere in the parameters of the generator (Thm. 1). Moreover, as it is continuous in the weak topology (Thm. 2), it can provide a signal to the generator g_θ that is more informative for training than other "distances" that lack this property (e.g., Jensen-Shannon divergence). In the following, we provide proofs for the above claims under assumptions similar to [3].

The following theorem formally states the result of continuity and differentiability in θ almost everywhere, which is desirable for permitting training via gradient descent.

Theorem 1. *Assume that (i) $f_\phi \circ g_\theta$ is locally Lipschitz with respect to $(\theta, \mathbf{z})$ with constants $L(\theta, \mathbf{z})$ not depending on ϕ and satisfying $\mathbb{E}_{\mathbf{z}}[L(\theta, \mathbf{z})] < \infty$; (ii) $\sup_{\eta \in \Pi} \mathbb{E}_{\omega(\mathbf{t};\eta)}[\|\mathbf{t}\|] < \infty$. Then, the function $\sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))})$ is continuous in $\theta \in \Theta$ everywhere, and differentiable in $\theta \in \Theta$ almost everywhere.*

The following theorem establishes continuity in the weak topology, and concerns general convergent distributions as opposed to only those corresponding to $g_\theta(\mathbf{z})$. In this result, we let $\mathbb{P}^{(\phi)}$ be the distribution of $f_\phi(\mathbf{x})$ when $\mathbf{x} \sim \mathbb{P}$, and similarly for $\mathbb{P}_n^{(\phi)}$.

Theorem 2. *Assume that (i) f_ϕ is L_f-Lipschitz for some L_f not depending on ϕ; (ii) $\sup_{\eta \in \Pi} \mathbb{E}_{\omega(\mathbf{t})}[\|\mathbf{t}\|] < \infty$. Then, the function $\sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)})$ is continuous in the weak topology, i.e., if $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$, then $\sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) \to 0$, where $\xrightarrow{D}$ implies convergence in distribution.*

The proofs are given in the appendix. In brief, we bound the difference between characteristic functions using geometric arguments; we interpret e^{ia} as a vector on a circle, and note that $|e^{ia} - e^{ib}| \leq |a - b|$. We then upper-bound the difference of function values in terms of $\mathbb{E}_{\omega(\mathbf{t})}[\|\mathbf{t}\|]$ (assumed to be finite) and averages of Lipschitz functions of $\mathbf{x}, \mathbf{x}'$ under the distributions considered. The Lipschitz properties ensure that the function difference vanishes when one distribution converges to the other.

Various generators satisfy the locally Lipschitz assumption, e.g., when g_θ is a feed-forward network with ReLU activations. To ensure that f_ϕ is Lipschitz, common methods employed in prior work include weight clipping [3] and gradient penalty [14]. In addition, many common distributions satisfy $\mathbb{E}_{\omega(\mathbf{t})}[\|\mathbf{t}\|] < \infty$, e.g., Gaussian, Student-t, and Laplace with fixed $\boldsymbol{\sigma}$. When $\boldsymbol{\sigma}$ is unbounded and optimized, we normalize the CFD by $\|\boldsymbol{\sigma}\|$, which prevents $\boldsymbol{\sigma}$ from going to infinity.

An example demonstrating the necessity of Lipschitz assumptions in continuity results (albeit for a different metric) can be found in Example 1 of [1]. In the appendix, we discuss conditions under which Theorem 2 can be strengthened to an "if and only if" statement.

4.2. Relation to MMD and Prior Work

The CFD is related to the maximum mean discrepancy (MMD) [13]. Given samples from two distributions $\mathbb{P}$ and

$\mathbb{Q}$, the squared MMD is given by

$$\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) = \mathbb{E}\left[\kappa(x,x')\right] + \mathbb{E}\left[\kappa(y,y')\right] - 2\mathbb{E}\left[\kappa(x,y)\right] \tag{7}$$

where $x, x' \sim \mathbb{P}$ and $y, y' \sim \mathbb{Q}$ are independent samples, and κ is kernel. When the weighting distribution of the CFD is equal to the inverse Fourier transform of the kernel in MMD (i.e., $\omega(\mathbf{t}) = \mathcal{F}^{-1}\{\kappa\}$), the CFD and squared MMD are equivalent: $\mathrm{CFD}_\omega^2(\mathbb{P},\mathbb{Q}) = \mathrm{MMD}_\kappa^2(\mathbb{P},\mathbb{Q})$. Indeed, kernels with $\mathrm{supp}(\mathcal{F}^{-1}(\kappa)) = \mathbb{R}^d$ are called *characteristic kernels* [35], and when $\mathrm{supp}(\omega) = \mathbb{R}^d$, $\mathrm{MMD}_\kappa(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. Although formally equivalent under the above conditions, we find experimentally that optimizing empirical estimates of MMD and CFD result in different convergence profiles and model performance across a range of datasets. Also, unlike MMD, which takes quadratic time in the number of samples to approximately compute, the CFD takes $O(nk)$ time and is therefore computationally attractive when $k \ll n$.

Learning a generative model by minimizing the MMD between real and generated samples was proposed independently by [23] and [10]. The Generative Moment Matching Network (GMMN) [23] uses an autoencoder to first transform the data into a latent space, and then trains a generative network to produce latent vectors that match the true latent distribution. The MMD-GAN [21] performs a similar input transformation using a network f_ϕ that is adversarially trained to maximize the MMD between the true distribution $\mathbb{P}_\mathcal{X}$ and the generator distribution $\mathbb{Q}_\theta$; this results in a GAN-like min-max criterion. More recently, [5] and [1] have proposed different theoretically-motivated regularizers on the gradient of MMD-GAN critic that improve training. In our experiments, we compare against the MMD-GAN both with and without gradient regularization.

Very recent work [22] (IKL-GAN) has evaluated kernels parameterized in Fourier space, which are then used to compute MMD in MMD-GAN. In contrast to IKL-GAN, we derive the CF-GAN via characteristic functions rather than via MMD, and our method obviates the need for kernel evaluation. We also provide novel direct proofs for the theoretical properties of the optimized CFD that are *not* based on its equivalence to MMD. The IKL-GAN utilizes a neural network to sample random frequencies, whereas we use a simpler fixed distribution with a learned scale, reducing the number of hyperparameters to tune. Our method yields state-of-the-art performance, which suggests that the more complex setup in IKL-GAN may not be required for effective GAN training.

In parallel, significant work has gone into improving GAN training via architectural and optimization enhancements [27, 7, 18]; these research directions are orthogonal to our work and can be incorporated in our proposed model.

5. Experiments

In this section, we present empirical results comparing different variants of our proposed model: CF-GAN. We prefix O to the model name when the $\boldsymbol{\sigma}$ parameters were optimized along with the critic and omit it when $\boldsymbol{\sigma}$ was kept fixed. Similarly, we suffix GP to the model name when gradient penalty [14] was used to enforce Lipschitzness of f_ϕ. In the absence of gradient penalty, we clipped the weights of f_ϕ in $[-0.01, 0.01]$. When the parameters $\boldsymbol{\sigma}$ were optimized, we scaled the ECFD by $\|\boldsymbol{\sigma}\|$ to prevent $\boldsymbol{\sigma}$ from going to infinity, thereby ensuring $\mathbb{E}_{\omega(\mathbf{t})}\left[\|\mathbf{t}\|\right] < \infty$.

We compare our proposed model against two variants of MMD-GAN: (i) MMD-GAN [21], which uses MMD with a mixture of RBF kernels as the distance metric; (ii) MMD-GAN-GP$_{L2}$ [5], which introduces an additive gradient penalty based on MMD's IPM witness function, an L2 penalty on discriminator activations, and uses a mixture of RQ kernels. We also compare against WGAN [3] and WGAN-GP [14] due to their close relation to MMD-GAN [21, 5]. Our code is available online at https://github.com/crslab/OCFGAN.

5.1. Synthetic Data

We first tested the methods on two synthetic 1D distributions: a simple unimodal distribution ($\mathcal{D}_1$) and a more complex bimodal distribution ($\mathcal{D}_2$). The distributions were constructed by transforming $z \sim \mathcal{N}(0,1)$ using a function $h : \mathbb{R} \to \mathbb{R}$. For the unimodal dataset, we used the scale-shift function form used by [37], where $h(z) = \mu + \sigma z$. For the bimodal dataset, we used the function form used by planar flow [32], where $h(z) = \alpha z + \beta \tanh(\gamma \alpha z)$. We trained the various GAN models to approximate the distribution of the transformed samples. Once trained, we compared the transformation function $\hat{h}$ learned by the GAN against the true function h. We computed the mean absolute error (MAE) ($\mathbb{E}_z[|h(z) - \hat{h}(z)|]$) to evaluate the models. Further details on the experimental setup are in Appendix B.1.

Figs. 2a and 2b show the variation of the MAE with training iterations. For both datasets, the models with gradient penalty converge to better minima. In $\mathcal{D}_1$, MMD-GAN-GP and OCF-GAN-GP converge to the same value of MAE, but MMD-GAN-GP converges faster. During our experiments, we observed that the scale of the weighting distribution (which is intialized to 1) falls rapidly before the MAE begins to decrease. For the experiments with the scale fixed at 0.1 (CF-GAN-GP$_{\sigma=0.1}$) and 1 (CF-GAN-GP$_{\sigma=1}$), both models converge to the same MAE, but CF-GAN-GP$_{\sigma=1}$ takes much longer to converge than CF-GAN-GP$_{\sigma=0.1}$. This indicates that the optimization of the scale parameter leads to faster convergence. For the more complex dataset $\mathcal{D}_2$, MMD-GAN-GP takes significantly longer to converge compared to WGAN-GP and OCF-GAN-GP. OCF-GAN-GP converges fastest and to a better minimum,

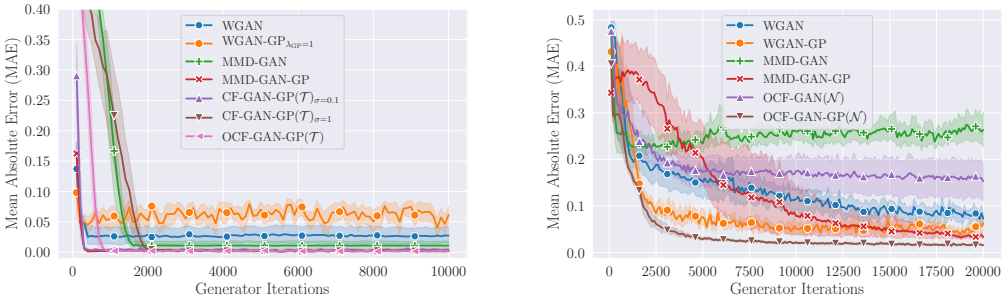

Figure 2: Variation of MAE for synthetic datasets $\mathcal{D}_1$ (left) and $\mathcal{D}_2$ (right) with generator iterations. The plots are averaged over 10 random runs.

followed by WGAN-GP.

5.2. Image Generation

A recent large-scale analysis of GANs [26] showed that different models achieve similar *best* performance when given *ample* computational budget, and advocates comparisons between distributions under practical settings. As such, we compare scores attained by the models from different initializations under fixed computational budgets. We used four datasets: 1) MNIST [20]: 60K grayscale images of handwritten digits; 2) CIFAR10 [19]: 50K RGB images; 3) CelebA [24]: ≈200K RGB images of celebrity faces; and 4) STL10 [9]: 100K RGB images. For all datasets, we center-cropped and scaled the images to 32×32.

Network and Hyperparameter Details Given our computational budget and experiment setup, we used a DCGAN-like generator g_θ and critic f_ϕ architecture for all models (similar to [21]). For MMD-GAN, we used a mixture of five RBF kernels (5-RBF) with different scales [21]. MMD-GAN-GP$_{L2}$ used a mixture of rational quadratic kernels (5-RQ). The kernel parameters and the trade-off parameters for gradient and L2 penalties were set according to [5]. We tested CF-GAN variants with two weighting distributions: Gaussian ($\mathcal{N}$) and Student's-t ($\mathcal{T}$) (with 2 degrees of freedom). For CF-GAN, we tested 3 scale parameters in the set $\{0.2, 0.5, 1\}$, and we report the best results. The number of frequencies (k) for computing ECFD was set to 8. Please see Appendix B.2 for implementation details.

Evaluation Metrics We compare the different models using three evaluation metrics: Fréchet Inception Distance (FID) [34], Kernel Inception Distance (KID) [5], and Precision-Recall (PR) for generative models [33]. Details on these metrics and the evaluation procedure can be found in Appendix B.2. In brief, the FID computes the Fréchet distance between two multivariate Gaussians and the KID computes the MMD (with a polynomial kernel of degree 3) between the real and generated data distributions. Both FID and KID give single value scores, and PR gives a two dimensional score which disentangles the quality of generated samples from the coverage of the data distribution. PR is defined by a pair F_8 (recall) and $F_{1/8}$ (precision) which represent the coverage and sample quality, respectively [33].

Results In the following, we summarize our main findings, and relegate the details to the Appendix. Table 1 shows the FID and KID values achieved by different models for CIFAR10, STL10, and CelebA datasets. In short, our model outperforms both variants of WGAN and MMD-GAN by a significant margin. OCF-GAN, using just one weighting function, outperforms both MMD-GANs that use a mixture of 5 different kernels.

We observe that the optimization of the scale parameter improves the performance of the models for both weighting distributions, and the introduction of gradient penalty as a means to ensure Lipschitzness of f_ϕ results in a significant improvement in the score values for all models. This is in line with the results of [14] and [5]. Overall, amongst the CF-GAN variants, OCF-GAN-GP with Gaussian weighting performs the best for all datasets.

The two-dimensional precision-recall scores in Fig. 3 provide further insight into the performance of different models. Across all the datasets, the addition of gradient penalty (OCF-GAN-GP) rather than weight clipping (OCF-GAN) leads to a higher improvement in recall compared to precision. This result supports recent arguments that weight clipping forces the generator to learn simpler functions, while gradient penalty is more flexible [14]. The improvement in recall with the introduction of gradient penalty is more noticeable for CIFAR10 and STL10 datasets compared to CelebA. This result is intuitive; CelebA is a more uniform and simpler dataset compared to CIFAR10/STL10, which contain more diverse classes of images, and thus

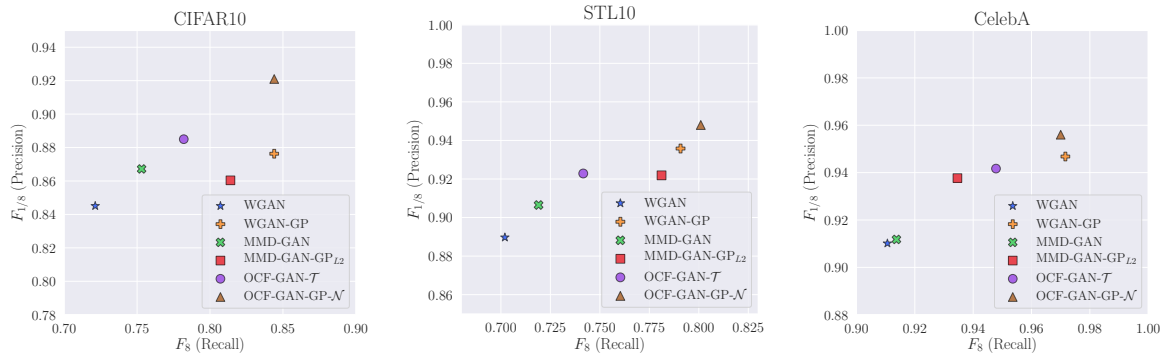

Figure 3: Precision-Recall scores (higher is better) for CIFAR10 (left), STL10 (center), and CelebA (right) datasets.

likely have modes that are more complex and far apart. Results on the MNIST dataset, where all models achieve good score values, are available in Appendix C, which also includes further experiments using the smoothed version of ECFD and the optimized smoothed version (no improvement over the unsmoothed versions on the image datasets).

Qualitative Results In addition to the quantitative metrics presented above, we also performed a qualitative analysis of the generated samples. Fig. 4 shows image samples generated by OCF-GAN-GP for different datasets. We also tested our method with a deep ResNet model on a 128×128 scaled version of CelebA dataset. Samples generated by this model (Fig. 5) show that OCF-GAN-GP scales to larger images and networks, and is able to generate visually appealing images comparable to state-of-the-art methods using similar sized networks. Additional qualitative comparisons can be found in Appendix C.

Impact of Weighting Distribution The choice of weighting distribution did not lead to drastic changes in model performance. The $\mathcal{T}$ distribution performs best when weight clipping is used, while $\mathcal{N}$ performs best in the case of gradient penalty. This suggests that the proper choice of distribution is dependent on both the dataset and the Lipschitz regularization used, but the overall framework is robust to reasonable choices.

We also conducted preliminary experiments using a uniform ($\mathcal{U}$) distribution weighting scheme. Even though the condition $\text{supp}(\mathcal{U}) = \mathbb{R}^m$ does not hold for the uniform distribution, we found that this does not adversely affect the performance (see Appendix C). The uniform weighting distribution corresponds to the sinc-kernel in MMD, which is known to be a non-characteristic kernel [35]. Our results suggest that such kernels could remain effective when used in MMD-GAN, but we did not verify this experimentally.

Figure 4: Image samples for the different datasets (top to bottom: CIFAR10, STL10, and MNIST) generated by OCF-GAN-GP (random samples without selection).

Impact of Number of Random Frequencies We conducted an experiment to study the impact of the number of random frequencies (k) that are sampled from the weighting distribution to compute the ECFD. We ran our best performing model (OCF-GAN-GP) with different values of k from the set $\{1, 4, 8, 16, 32, 64\}$. The FID and KID scores for this experiment are shown in Table 2. As expected, the score values improve as k increases. However, even for the

Table 1: FID and KID ($\times 10^3$) scores (lower is better) for CIFAR10, STL10, and CelebA datasets averaged over 5 random runs (standard deviation in parentheses).

Model	Kernel/ Weight	CIFAR10		STL10		CelebA	
		FID	KID	FID	KID	FID	KID
WGAN	–	44.11 (1.16)	25 (1)	38.61 (0.43)	23 (1)	17.85 (0.69)	12 (1)
WGAN-GP	–	35.91 (0.30)	19 (1)	27.85 (0.81)	15 (1)	10.03 (0.37)	6 (1)
MMD-GAN	5-RBF	41.28 (0.54)	23 (1)	35.76 (0.54)	21 (1)	18.48 (1.60)	12 (1)
MMD-GAN-GP$_{L2}$	5-RQ	38.88 (1.35)	21 (1)	31.67 (0.94)	17 (1)	13.22 (1.30)	8 (1)
CF-GAN	$\mathcal{N}_{(\sigma=0.5)}$	39.81 (0.93)	23 (1)	33.54 (1.11)	19 (1)	13.71 (0.50)	9 (1)
	$\mathcal{T}_{(\sigma=1)}$	41.41 (0.64)	22 (1)	35.64 (0.44)	20 (1)	16.92 (1.29)	11 (1)
OCF-GAN	$\mathcal{N}$	38.47 (1.00)	20 (1)	32.51 (0.87)	19 (1)	14.91 (0.83)	9 (1)
	$\mathcal{T}$	37.96 (0.74)	20 (1)	31.03 (0.82)	17 (1)	13.73 (0.56)	8 (1)
OCF-GAN-GP	$\mathcal{N}$	**33.08 (0.26)**	**17 (1)**	**26.16 (0.64)**	**14 (1)**	**9.39 (0.25)**	**5 (1)**
	$\mathcal{T}$	34.33 (0.77)	18 (1)	26.86 (0.38)	15 (1)	9.61 (0.39)	6 (1)

Table 2: FID and KID scores for on the MNIST dataset with varying numbers of frequencies used in OCF-GAN-GP.

# of freqs (k)	FID	KID $\times 10^3$
1	0.44 (0.03)	5 (1)
4	0.39 (0.05)	4 (1)
8	0.36 (0.03)	4 (1)
16	0.35 (0.02)	3 (1)
32	0.35 (0.03)	3 (1)
64	0.36 (0.07)	4 (1)

Figure 5: Image samples for the 128×128 CelebA dataset generated by OCF-GAN-GP with a ResNet generator (random samples without selection).

lowest number of frequencies possible ($k = 1$), the performance does not degrade too severely.

6. Discussion and Conclusion

In this paper, we proposed a novel weighted distance between characteristic functions for training IGMs, and showed that the proposed metric has attractive theoretical properties. We observed experimentally that the proposed model outperforms MMD-GAN and WGAN variants on four benchmark image datasets. Our results indicate that characteristic functions provide an effective alternative means for training IGMs.

This work opens additional avenues for future research. For example, the empirical CFD used for training may result in high variance gradient estimates (particularly with a small number of sampled frequencies), yet the CFD-trained models attain high performance scores with better convergence in our tests. The reason for this should be more thoroughly explored. Although we used the gradient penalty proposed by WGAN-GP, there is no reason to constrain the gradient to exactly 1. We believe that an exploration of the geometry of the proposed loss could lead to improvement in the gradient regularizer for the proposed method.

Apart from generative modeling, two sample tests such as MMD have been used for problems such as domain adaptation [25] and domain separation [6], among others. The optimized CFD loss function proposed in this work can be used as an alternative loss for these problems.

Acknowledgements This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2019-011) to H. Soh. J. Scarlett is supported by the Singapore National Research Foundation (NRF) under grant number R-252-000-A74-281.

References

[1] Michael Arbel, Dougal Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. In *NeurIPS*, 2018. 1, 2, 4, 5

[2] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv:1701.04862*, 2017. 1

[3] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 1, 2, 4, 5

[4] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv:1705.10743*, 2017. 2

[5] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 1, 2, 5, 6

[6] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, 2016. 8

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv:1809.11096*, 2018. 5

[8] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, 2015. 1, 2, 3

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 6

[10] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015. 5

[11] TW Epps and Kenneth J Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203, 1986. 1

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 4

[14] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of Wasserstein GANs. In *NIPS*, 2017. 2, 4, 5, 6

[15] CE Heathcote. A test of goodness of fit for symmetric random variables. *Australian Journal of Statistics*, 14(2):172–181, 1972. 1

[16] CR Heathcote. The integrated squared error estimation of parameters. *Biometrika*, 64(2):255–264, 1977. 2, 3

[17] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NIPS*, 2016. 1

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 5

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 6

[20] Yann LeCun, Léon Bottou, and Patrick Haffner. Gradient-based learning applied to document recognition. 2001. 6

[21] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *NIPS*, 2017. 1, 2, 4, 5, 6

[22] Chun-Liang Li, Wei-Cheng Chang, Youssef Mroueh, Yiming Yang, and Barnabás Póczos. Implicit kernel learning. In *AISTATS*, 2019. 1, 5

[23] Yujia Li, Kevin Swersky, and Richard S. Zemel. Generative moment matching networks. In *ICML*, 2015. 5

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 8

[26] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *NeurIPS*, 2018. 6

[27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018. 5

[28] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. *arXiv:1711.04894*, 2017. 2

[29] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. 1

[30] Albert S Paulson, Edward W Holcomb, and Robert A Leitch. The estimation of the parameters of the stable laws. *Biometrika*, 62(1):163–170, 1975. 3

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015. 1

[32] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015. 5

[33] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, pages 5228–5237, 2018. 6

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*, 2016. 1, 6

[35] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010. 2, 5, 7

[36] Jun Yu. Empirical characteristic function estimation and its applications. *Econometric reviews*, 2004. 3

[37] Manzil Zaheer, Chun-Liang Li, Barnabás Póczos, and Ruslan Salakhutdinov. GAN connoisseur : Can GANs learn simple 1D parametric distributions? 2018. 5