A Characteristic Function Approach to Deep Implicit Generative Modeling

Supplementary Material

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

A. Proofs

A.1. Proof of Theorem 1

Let $\mathbb{P}_{\mathcal{X}}$ be the data distribution, and let $\mathbb{P}_{g_\theta(\mathcal{Z})}$ be the distribution of $g_\theta(\mathbf{z})$ when $\mathbf{z} \sim \mathbb{P}_{\mathcal{Z}}$, with $\mathbb{P}_{\mathcal{Z}}$ being the latent distribution. Recall that the characteristic function of a distribution $\mathbb{Q}$ is given by

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

The quantity $\mathrm{CFD}^2_\omega(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))})$ can then be written as

$$\mathrm{CFD}^2_\omega(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))}) = \mathbb{E}_{\mathbf{t}\sim\omega(\mathbf{t};\eta)}\left[|\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})|^2\right], \tag{2}$$

where we denote the characteristic functions of $\mathbb{P}_{f_\phi(\mathcal{X})}$ and $\mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))}$ by $\varphi_{\mathcal{X}}$ and φ_θ respectively, with an implicit dependence of ϕ. For notational simplicity, we henceforth denote $\mathrm{CFD}^2_\omega(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))})$ by $\mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta)$.

Since the difference of two functions' maximal values is always upper bounded by the maximal gap between the two functions, we have

$$\left|\sup_{\psi\in\Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \sup_{\psi\in\Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'})\right| \le \sup_{\psi\in\Psi} |\mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'})| \tag{3}$$

$$\le |\mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'})| + \epsilon \tag{4}$$

where $\psi^* = \{\phi^*, \eta^*\}$ denotes any parameters that are within ϵ of the supremum on the right-hand side of (4), and where $\epsilon > 0$ may be arbitrarily small. Such ψ^* always exists by the definition of supremum. Subsequently, we define $h_\theta = f_{\phi^*} \circ g_\theta$ for compactness.

Let ω^* denote the distribution $\omega(\mathbf{t})$ associated with η^*. We further upper bound the right-hand side of (4) as follows:

$$|\mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'})| = \left|\mathbb{E}_{\omega^*(\mathbf{t})}\left[|\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})|^2\right] - \mathbb{E}_{\omega^*(\mathbf{t})}\left[|\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_{\theta'}(\mathbf{t})|^2\right]\right| \tag{5}$$

$$\overset{(a)}{\le} \mathbb{E}_{\omega^*(\mathbf{t})}\left[\left||\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})|^2 - |\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_{\theta'}(\mathbf{t})|^2\right|\right], \tag{6}$$

where (a) uses the linearity of expectation and Jensen's inequality.

Since any characteristic function is bounded by $|\varphi_{\mathbb{P}}(\mathbf{t})| \le 1$, the value of $|\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})|$ for any θ is upper bounded by

2. Since the function $f(u) = u^2$ is (locally) 4-Lipschitz over the restricted domain $[0, 2]$, we have

$$\left| |\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})|^2 - |\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_{\theta'}(\mathbf{t})|^2 \right| \leq 4 \left| |\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_\theta(\mathbf{t})| - |\varphi_{\mathcal{X}}(\mathbf{t}) - \varphi_{\theta'}(\mathbf{t})| \right| \tag{7}$$

$$\overset{(b)}{\leq} 4 \left| \varphi_\theta(\mathbf{t}) - \varphi_{\theta'}(\mathbf{t}) \right| \tag{8}$$

$$= 4 \left| \mathbb{E}_{\mathbf{z}} \left[e^{i\langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle} \right] - \mathbb{E}_{\mathbf{z}} \left[e^{i\langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle} \right] \right| \tag{9}$$

$$\overset{(c)}{\leq} 4 \mathbb{E}_{\mathbf{z}} \left[\left| e^{i\langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle} - e^{i\langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle} \right| \right], \tag{10}$$

where (b) uses the triangle inequality, and (c) uses Jensen's inequality.

In Eq. (10), let $\left| e^{i\langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle} - e^{i\langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle} \right| =: \left| e^{ia} - e^{ib} \right|$, which can be interpreted as the length of the chord that subtends an angle of $|a - b|$ at the center of a unit circle centered at origin. The length of this chord is given by $2 \sin \frac{|a-b|}{2}$, and since $2 \sin \frac{|a-b|}{2} \leq |a - b|$, we have

$$\left| e^{i\langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle} - e^{i\langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle} \right| \leq \left| \langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle - \langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle \right| \tag{11}$$

$$\overset{(d)}{\leq} \|\mathbf{t}\| \cdot \|h_\theta(\mathbf{z}) - h_{\theta'}(\mathbf{z})\|, \tag{12}$$

where (d) uses the Cauchy-Schwarz inequality.

Furthermore, using the assumption $\sup_{\eta \in \Pi} \mathbb{E}_{\omega(\mathbf{t})} \left[\|\mathbf{t}\| \right] < \infty$, we get

$$\mathbb{E}_{\omega^*(\mathbf{t})} \left[\mathbb{E}_{\mathbf{z}} \left[\left| e^{i\langle \mathbf{t}, h_\theta(\mathbf{z}) \rangle} - e^{i\langle \mathbf{t}, h_{\theta'}(\mathbf{z}) \rangle} \right| \right] \right] \leq \mathbb{E}_{\omega^*(\mathbf{t})} \left[\|\mathbf{t}\| \right] \mathbb{E}_{\mathbf{z}} \left[\|h_\theta(\mathbf{z}) - h_{\theta'}(\mathbf{z})\| \right] \tag{13}$$

with the first term being finite.

By assumption, h is locally Lipschitz, i.e., for any pair $(\theta, \mathbf{z})$, there exists a constant $L(\theta, \mathbf{z})$ and an open set $U_{\theta, \mathbf{z}}$ such that $\forall (\theta', \mathbf{z}') \in U_{\theta, \mathbf{z}}$ we have $\|h_\theta(\mathbf{z}) - h_{\theta'}(\mathbf{z}')\| \leq L(\theta, \mathbf{z}) \|\theta - \theta'\|$. Setting $\mathbf{z}' = \mathbf{z}$ and taking the expectation, we obtain

$$\mathbb{E}_{\omega^*(\mathbf{t})} \left[\|\mathbf{t}\| \right] \mathbb{E}_{\mathbf{z}} \left[\|h_\theta(\mathbf{z}) - h_{\theta'}(\mathbf{z})\| \right] \leq \mathbb{E}_{\omega^*(\mathbf{t})} \left[\|\mathbf{t}\| \right] \mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] \|\theta - \theta'\| \tag{14}$$

for all θ' sufficiently close to θ.

Recall also that $\mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] < \infty$ by assumption. Combining Eqs. (6), (10), and (14), we get

$$\left| \mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \mathrm{D}_{\psi^*}(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'}) \right| \leq 4 \mathbb{E}_{\omega^*(\mathbf{t})} \left[\|\mathbf{t}\| \right] \mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] \|\theta - \theta'\|, \tag{15}$$

and combining with (4) gives

$$\left| \sup_{\psi \in \Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \sup_{\psi \in \Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'}) \right| \leq 4 \mathbb{E}_{\omega^*(\mathbf{t})} \left[\|\mathbf{t}\| \right] \mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] \|\theta - \theta'\| + \epsilon \tag{16}$$

$$\leq 4 \left(\sup_{\eta \in \Pi} \mathbb{E}_{\omega(\mathbf{t})} \left[\|\mathbf{t}\| \right] \right) \mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] \|\theta - \theta'\| + \epsilon. \tag{17}$$

Taking the limit $\epsilon \to 0$ on both sides gives

$$\left| \sup_{\psi \in \Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta) - \sup_{\psi \in \Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\theta'}) \right| \leq 4 \left(\sup_{\eta \in \Pi} \mathbb{E}_{\omega(\mathbf{t})} \left[\|\mathbf{t}\| \right] \right) \mathbb{E}_{\mathbf{z}} \left[L(\theta, \mathbf{z}) \right] \|\theta - \theta'\|, \tag{18}$$

which proves that $\sup_{\psi \in \Psi} \mathrm{D}_\psi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_\theta)$ is locally Lipschitz, and therefore continuous. In addition, Radamacher's theorem [3] states any locally Lipschitz function is differentiable almost everywhere, which establishes the differentiability claim.

A.2. Proof of Theorem 2

Let $\mathbf{x}_n \sim \mathbb{P}_n$ and $\mathbf{x} \sim \mathbb{P}$. To study the behavior of $\sup_{\psi \in \Psi} \mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)})$, we first consider

$$\mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) = \mathbb{E}_{\omega(\mathbf{t})} \left[\left| \mathbb{E}_{\mathbf{x}_n} \left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n) \rangle} \right] - \mathbb{E}_{\mathbf{x}} \left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}) \rangle} \right] \right|^2 \right] \tag{19}$$

Since $\left|\mathbb{E}_{\mathbf{x}_n}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle}\right] - \mathbb{E}_{\mathbf{x}}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x})\rangle}\right]\right| \in [0,2]$, using the fact that $u^2 \leq 2|u|$ for $u \in [-2,2]$, we have

$$\begin{aligned}
&\mathbb{E}_{\omega(\mathbf{t})}\left[\left|\mathbb{E}_{\mathbf{x}_n}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle}\right] - \mathbb{E}_{\mathbf{x}}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x})\rangle}\right]\right|^2\right] \\
&\quad\leq 2\,\mathbb{E}_{\omega(\mathbf{t})}\left[\left|\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle} - e^{i\langle \mathbf{t}, f_\phi(\mathbf{x})\rangle}\right]\right|\right] &(20)\\
&\quad\overset{(a)}{\leq} 2\,\mathbb{E}_{\omega(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[\left|e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle} - e^{i\langle \mathbf{t}, f_\phi(\mathbf{x})\rangle}\right|\right]\right] &(21)\\
&\quad\overset{(b)}{\leq} 2\,\mathbb{E}_{\omega(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[\min\left\{2, |\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle - \langle \mathbf{t}, f_\phi(\mathbf{x})\rangle|\right\}\right]\right] &(22)\\
&\quad\overset{(c)}{\leq} 2\,\mathbb{E}_{\omega(\mathbf{t})}\left[\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[\min\left\{2, \|\mathbf{t}\| \cdot \|f_\phi(\mathbf{x}_n) - f_\phi(\mathbf{x})\|\right\}\right]\right], &(23)
\end{aligned}$$

where (a) uses Jensen's inequality, (b) uses the geometric properties stated following Eq. (10) and the fact that $|e^{ia} - e^{ib}| \leq 2$, and (c) uses the Cauchy-Schwarz inequality.

For brevity, let $T_{\max} = \sup_{\eta\in\Pi} \mathbb{E}_{\omega(\mathbf{t})}\left[\|\mathbf{t}\|\right]$, which is finite by assumption. Interchanging the order of the expectations in Eq. (23) and applying Jensen's inequality (to $\mathbb{E}_{\omega(\mathbf{t})}$ alone) and the concavity of $f(u) = \min\{2, u\}$, we can continue the preceding upper bound as follows:

$$\begin{aligned}
&\mathbb{E}_{\omega(\mathbf{t})}\left[\left|\mathbb{E}_{\mathbf{x}_n}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x}_n)\rangle}\right] - \mathbb{E}_{\mathbf{x}}\left[e^{i\langle \mathbf{t}, f_\phi(\mathbf{x})\rangle}\right]\right|^2\right] \\
&\quad\leq 2\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[\min\left\{2, T_{\max}\|f_\phi(\mathbf{x}_n) - f_\phi(\mathbf{x})\|\right\}\right] &(24)\\
&\quad\overset{(d)}{\leq} 2\mathbb{E}_{\mathbf{x}_n,\mathbf{x}}\left[\min\left\{2, T_{\max}L_f\|\mathbf{x}_n - \mathbf{x}\|\right\}\right], &(25)
\end{aligned}$$

where (d) defines L_f to be the Lipschitz constant of f_ϕ, with is independent of ϕ by assumption.

Observe that $g(u) = \min\{2, T_{\max}L_f|u|\}$ is a bounded Lipschitz function of u. By the Portmanteau theorem ([5], Thm. 13.16), convergence in distribution $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$ implies that $\mathbb{E}[g(\|\mathbf{x}_n - \mathbf{x}\|)] \to 0$ for any such g, and hence (25) yields $\sup_{\psi\in\Psi} \mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) \to 0$ (upon taking $\sup_{\psi\in\Psi}$ on both sides), as required.

A.3. Discussion on an "only if" Counterpart to Theorem 2

Theorem 2 shows that, under some technical assumptions, the function $\sup_{\psi\in\Psi} \mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)})$ satisfies continuity in the weak toplogy, i.e.,

$$\mathbb{P}_n \xrightarrow{D} \mathbb{P} \implies \sup_{\psi\in\Psi} \mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) \to 0.$$

where $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$ denotes convergence in distribution.

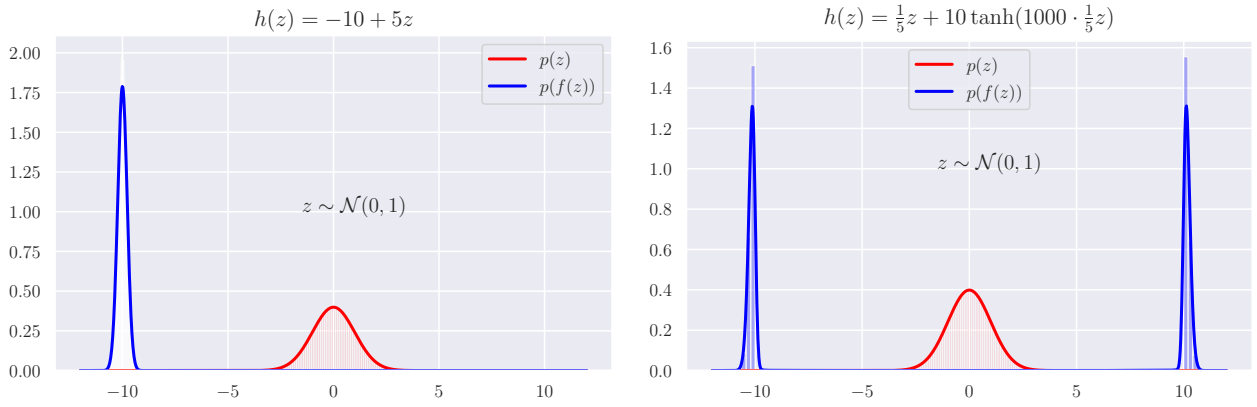
Here we discuss whether the opposite is true: Does $\sup_{\psi\in\Psi} \mathrm{CFD}^2_\omega(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) \to 0$ imply that $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$? In general, the answer is negative. For example:

- If Φ only contains the function $\phi(x) = 0$, then $\mathbb{P}^{(\phi)}$ is always the distribution corresponding to deterministically equaling zero, so *any* two distributions give zero CFD.
- If $\omega(t)$ has bounded support, then two distributions $\mathbb{P}_1, \mathbb{P}_2$ whose characteristic functions only differ for t values outside that support may still give $\mathbb{E}_{\omega(t)}\left[|\varphi_{\mathbb{P}_1}(t) - \varphi_{\mathbb{P}_2}(t)|^2\right] = 0$.

In the following, however, we argue that the answer is positive when $\{f_\phi\}_{\phi\in\Phi}$ is "sufficiently rich" and $\{\omega\}_{\eta\in\Pi}$ is "sufficiently well-behaved".

Rather than seeking the most general assumptions that formalize these requirements, we focus on a simple special case that still captures the key insights, assuming the following:

- There exists $L > 0$ such that $\{f_\phi\}_{\phi\in\Phi}$ includes all *linear* functions that are L-Lipschitz;
- There exists $\eta \in \Pi$ such that $\omega(\mathbf{t})$ has support $\mathbb{R}^m$, where m is the output dimension of f_ϕ.



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$

Figure 1: The PDFs of $\mathcal{D}_1$ and $\mathcal{D}_2$ (in blue) estimated using Kernel Density Estimation (KDE) along with the true distribution $p(z)$ (in red).

To give examples of these, note that neural networks with ReLU activations can implement arbitrary linear functions (with the Lipschitz condition amounting to bounding the weights), and note that the second assumption is satisfied by any Gaussian $\omega(\mathbf{t})$ with a fixed positive-definite covariance matrix.

In the following, let $\mathbf{x}_n \sim \mathbb{P}_n$ and $\mathbf{x} \sim \mathbb{P}^{(\phi)}$. We will prove the contrapositive statement:

$$\mathbb{P}_n \overset{D}{\not\to} \mathbb{P}^{(\phi)} \implies \sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_n, \mathbb{P}^{(\phi)}) \not\to 0.$$

By the Cramér-Wold theorem [2], $\mathbb{P}_n \overset{D}{\not\to} \mathbb{P}^{(\phi)}$ implies that we can find constants $c_1, \ldots, c_d$ such that

$$\sum_{i=1}^{d} c_i \mathbf{x}_n^{(i)} \overset{D}{\not\to} \sum_{i=1}^{d} c_i \mathbf{x}^{(i)}, \tag{26}$$

where $\mathbf{x}^{(i)}, \mathbf{x}_n^{(i)}$ denote the i-th entries of $\mathbf{x}, \mathbf{x}_n$, with d being their dimension.

Recall that we assume $\{f_\phi\}_{\phi \in \Phi}$ includes all linear functions from $\mathbb{R}^d$ to $\mathbb{R}^m$ with Lipschitz constant at most $L > 0$. Hence, we can select $\phi \in \Phi$ such that every entry of $f_\phi(x)$ equals $\frac{1}{Z}\sum_{i=1}^{d} c_i x^{(i)}$, where Z is sufficiently large so that the Lipschitz constant of this f_ϕ is at most L. However, for this ϕ, (26) implies that $f_\phi(\mathbf{x}_n) \overset{D}{\not\to} f_\phi(\mathbf{x})$, which in turn implies that $|\varphi_{\mathbb{P}_n^{(\phi)}}(t) - \varphi_{\mathbb{P}^{(\phi)}}(t)|$ is bounded away from zero for all t in some set $\mathcal{T}$ of positive Lebesgue measure.

Choosing $\omega(\mathbf{t})$ to have support $\mathbb{R}^m$ in accordance with the second technical assumption above, it follows that $\mathbb{E}_{\omega(t)}\left[|\varphi_{\mathbb{P}_1^{(\phi)}}(t) - \varphi_{\mathbb{P}_2^{(\phi)}}(t)|^2\right] \not\to 0$ and hence $\sup_{\psi \in \Psi} \mathrm{CFD}_\omega^2(\mathbb{P}_n^{(\phi)}, \mathbb{P}^{(\phi)}) \not\to 0$.

B. Implementation Details

B.1. Synthetic Data Experiments

The synthetic data was generated by first sampling $z \sim \mathcal{N}(0, 1)$ and then applying a function h to the samples. We constructed distributions of two types: a scale-shift unimodal distribution $\mathcal{D}_1$ and a "scale-split-shift" bimodal distribution $\mathcal{D}_2$. The function h for the two distributions are defined as follows:

- $\mathcal{D}_1$: $h(z) = \mu + \sigma z$; we set $\mu = -10$ and $\sigma = \frac{1}{5}$. This shifts the mean of the distribution to -10, resulting in the $\mathcal{N}(-10, \frac{1}{5^2})$ distribution. Fig. 1a shows the PDF (and histogram) of the original distribution $p(z)$ and the distribution of $h(z)$, which is approximated using Kernel Density Estimation (KDE).

- $\mathcal{D}_2$: $h(z) = \alpha z + \beta \tanh(\gamma \alpha z)$; we set $\alpha = \frac{1}{5}$, $\beta = 10$, $\gamma = 100$. This splits the distribution into two modes and shifts the two modes to -10 and $+10$. Fig. 1b shows the PDF (and histogram) of the original distribution $p(z)$ and the distribution of $h(z)$, which is approximated using KDE.

For the two cases described above, there are two transformation functions that will lead to the same distribution. In each case, the second transformation function is given by:

- $\mathcal{D}_1$: $g(z) = \mu - \sigma z$
- $\mathcal{D}_2$: $g(z) = -\alpha z + \beta \tanh(-\gamma \alpha z)$

As there are two possible correct transformation functions (h and g) that the GANs can learn, we computed the Mean Absolute Error (MAE) as follows

$$\text{MAE} = \min\left(\mathbb{E}_z\left[|h(z) - \hat{h}(z)|\right], \mathbb{E}_z\left[|g(z) - \hat{h}(z)|\right]\right), \tag{27}$$

where $\hat{h}$ is the transformation learned by the generator. We estimated the expectations in Eq. (27) using 5000 samples.

For the generator and critic network architectures, we followed [9]. Specifically, the generator is a multi-layer perceptron (MLP) with 3 hidden layers of sizes 7, 13, 7, and the Exponential linear unit (ELU) non-linearity between the layers. The critic network is also an MLP with 3 hidden layers of sizes 11, 29, 11, and the ELU non-linearity between the layers. The inputs and outputs of both networks are one-dimensional. We used the RMSProp optimizer with a learning rate of 0.001 for all models. The batch size was set to 50, and 5 critic updates were performed per generator iteration. We trained the models for 10000 and 20000 generator iterations for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. For all the models that rely on weight clipping, clipping in the range $[-0.01, 0.01]$ for $\mathcal{D}_2$ resulted in poor performance, so we modified the range to $[-0.1, 0.1]$.

We used a mixture of 5 RBF kernels for MMD-GAN [6], and a mixture of 5 RQ kernels and gradient penalty (as defined in [1]) for MMD-GAN-GP. For the CF-GAN variants, we used a single weighting distribution (Student-t and Gaussian for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively). The gradient penalty trade-off parameter (λ_{GP}) for WGAN-GP was set to 1 for $\mathcal{D}_1$ as the value of 10 led to erratic performance.

B.2. Image Generation

CF-GAN Following [6], a decoder was also connected to the critic in CF-GAN to reconstruct the input to the critic. This encourages the critic to learn a representation that has a high mutual information with the input. The auto-encoding objective is optimized along with the discriminator, and the final objective is given by

$$\inf_{\theta} \sup_{\psi} \text{CFD}^2_{\omega}(\mathbb{P}_{f_\phi(\mathcal{X})}, \mathbb{P}_{f_\phi(g_\theta(\mathcal{Z}))}) - \lambda_1 \mathbb{E}_{\mathbf{u} \in \mathcal{X} \cup g_\theta(\mathcal{Z})}\left[\mathcal{D}(\mathbf{u}, f^d_\phi(f_\phi(\mathbf{u})))\right], \tag{28}$$

where f^d_ϕ is the decoder network, λ_1 is the regularization parameter, and $\mathcal{D}$ is the discrepancy between the two data-points (e.g., squared error, cross-entropy, etc.). Although the decoder is interesting from an auto-encoding perspective of the representation learned by f_ϕ, we found that the removal of the decoder did not impact the performance of the model; this can be seen by the results of OCF-GAN-GP, which does not use a decoder network.

We also reduced the feasible set [6] of f_ϕ, which amounts to an additive penalty of $\lambda_2 \min\left(\mathbb{E}[f_\phi(\mathbf{x})] - \mathbb{E}[f_\phi(g_\theta(\mathbf{z}))], 0\right)$. We observed in our experiments that this led to improved stability of training, especially for the models that use weight clipping to enforce Lipschitz condition. For more details, we refer the reader to [6].

Network and Hyperparameter Details We used DCGAN-like generator g_θ and critic f_ϕ architectures, same as [6] for all models. Specifically, both g_θ and d_ϕ are fully convolutional networks with the following structures:

- g_θ: upconv(256) $\rightarrow$ bn $\rightarrow$ relu $\rightarrow$ upconv(128) $\rightarrow$ bn $\rightarrow$ relu $\rightarrow$ upconv(64) $\rightarrow$ bn $\rightarrow$ relu $\rightarrow$ upconv(c) $\rightarrow$ tanh;
- f_ϕ: conv(64) $\rightarrow$ leaky-relu(0.2) $\rightarrow$ conv(128) $\rightarrow$ bn $\rightarrow$ leaky-relu(0.2) $\rightarrow$ conv(256) $\rightarrow$ bn $\rightarrow$ leaky-relu(0.2) $\rightarrow$ conv(m),

where conv, upconv, bn, relu, leaky-relu, and tanh refer to convolution, up-convolution, batch-normalization, ReLU, LeakyReLU, and Tanh layers respectively. The decoder f_{ϕ^d} (whenever used) is also a DCGAN-like decoder. The generator takes a k-dimensional Gaussian latent vector as the input and outputs a 32×32 image with c channels. The value of

k was set differently depending on the dataset: MNIST (10), CIFAR10 (32), STL10 (32), and CelebA (64). The output dimensionality of the critic network (m) was set to 10 (MNIST) and 32 (CIFAR10, STL10, CelebA) for the MMD-GAN and CF-GAN models and 1 for WGAN and WGAN-GP. The batch normalization layers in the critic were omitted for WGAN-GP and OCF-GAN-GP (as suggested by [4]).

RMSProp optimizer was used with a learning rate of 5×10^{-5}. All models were optimized with a batch size of 64 for 125000 generator iterations (50000 for MNIST) with 5 critic updates per generator iteration. We tested CF-GAN variants with two weighting distributions: Gaussian ($\mathcal{N}$) and Student-t ($\mathcal{T}$) (with 2 degrees of freedom). We also conducted preliminary experiments using Laplace ($\mathcal{L}$) and Uniform ($\mathcal{U}$) weighting distributions (see Table 1). For CF-GAN, we tested with 3 scale parameters for $\mathcal{N}$ and $\mathcal{T}$ from the set $\{0.2, 0.5, 1\}$, and we report the best results. The trade-off parameter for the auto-encoder penalty (λ_1) and feasible-set penalty (λ_2) were set to 8 and 16 respectively, as in [6]. For OCF-GAN-GP, the trade-off for the gradient penalty was set to 10, same as WGAN-GP. The number of random frequencies k used for computing ECFD for all CF-GAN models was set to 8. For MMD-GAN, we used a mixture of five RBF kernels $k_\sigma(x, x') = \exp\left(\frac{\|x-x'\|^2}{2\sigma^2}\right)$ with different scales (σ) in $\Sigma = \{1, 2, 4, 8, 16\}$ as in [6]. For MMD-GAN-GP$_{L2}$, we used a mixture of rational quadratic kernels $k_\sigma(x, x') = \left(1 + \frac{\|x-x'\|^2}{2\alpha}\right)^{-\alpha}$ with α in $\mathcal{A} = \{0.2, 0.5, 1, 2, 5\}$; the trade-off parameters of the gradient and L2 penalties were set according to [1].

Evaluation Metrics We compared the different models using three evaluation metrics: Fréchet Inception Distance (FID) [8], Kernel Inception Distance (KID) [1], and Precision-Recall (PR) for Generative models [7]. All evaluation metrics use features extracted from the `pool3` layer (2048 dimensional) of an Inception network pre-trained on ImageNet, except for MNIST, for which we used a LeNet5 as the feature extractor. FID fits Gaussian distributions to Inception features of the real and fake images and then computes the Fréchet distance between the two Gaussians. On the other hand, KID computes the MMD between the Inception features of the two distributions using a polynomial kernel of degree 3. This is equivalent to comparing the first three moments of the two distributions.

Let $\{x_i^r\}_{i=1}^n$ be samples from the data distribution $\mathbb{P}_r$ and $\{x_i^g\}_{i=1}^m$ be samples from the GAN generator distribution $\mathbb{Q}_\theta$. Let $\{z_i^r\}_{i=1}^n$ and $\{z_i^g\}_{i=1}^m$ be the feature vectors extracted from the Inception network for $\{x_i^r\}_{i=1}^n$ and $\{x_i^g\}_{i=1}^m$ respectively. The FID and KID are then given by

$$\mathrm{FID}(\mathbb{P}_r, \mathbb{Q}_\theta) = ||\mu_r - \mu_g||^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{29}$$

$$\begin{aligned}\mathrm{KID}(\mathbb{P}_r, \mathbb{Q}_\theta) =& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left[\kappa(z_i^r, z_j^r)\right] \\ &+ \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} \left[\kappa(z_i^g, z_j^g)\right] \\ &- \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\kappa(z_i^r, z_j^g)\right],\end{aligned} \tag{30}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the sample mean & covariance matrix of the inception features of the real and generated data distributions, and κ is a polynomial kernel of degree 3, i.e.,

$$\kappa(x, y) = \left(\frac{1}{m}\langle x, y\rangle + 1\right)^3, \tag{31}$$

where m is the dimensionality of the feature vectors.

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. For more details about PR, we refer the reader to [7]. In brief, PR is defined by a pair F_8 (recall) and $F_{1/8}$ (precision), which represent the coverage and sample quality respectively [7].

We used 50000 (10000 for PR) random samples from the different GANs to compute the FID and KID scores. For MNIST and CIFAR10, we compared against the standard test sets, while for CelebA and STL10, we compared against 50000 random images sampled from the dataset. Following [1], we computed FID using 10 bootstrap resamplings and KID by sampling 1000 elements (without replacement) 100 times.

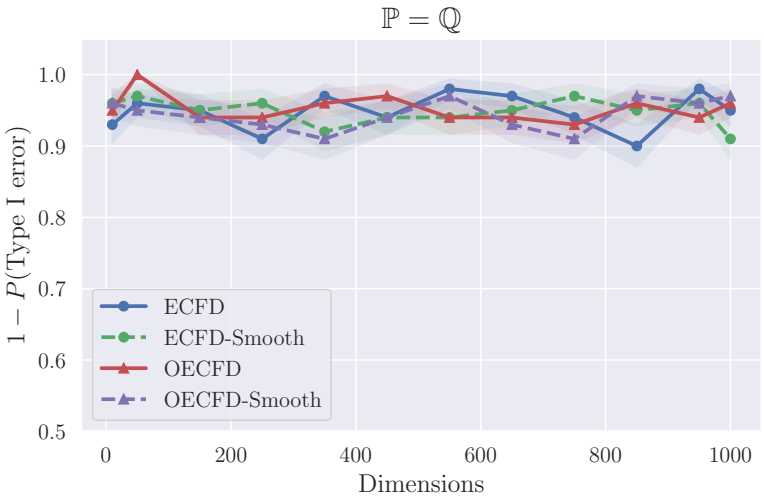


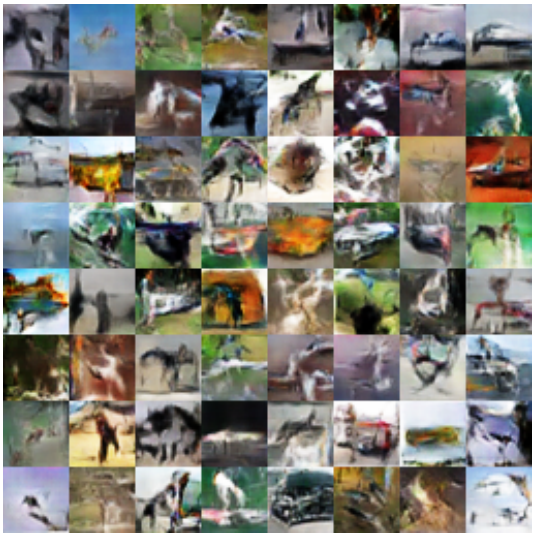
Figure 2: Probability of correctly accepting the null hypothesis $\mathbb{P} = \mathbb{Q}$ for various numbers of dimensions and different variants of ECFD.

C. Additional Results

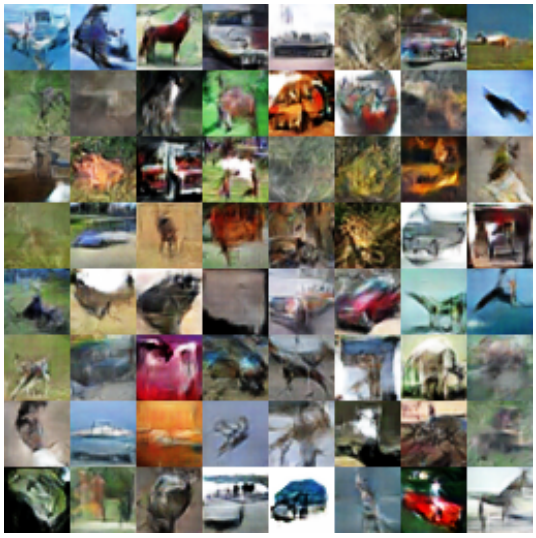
Fig. 2 shows the probability of accepting the null hypothesis $\mathbb{P} = \mathbb{Q}$ when it is indeed correct for different two sample tests based on ECFs. As mentioned in the main text, the optimization of the parameters of the weighting distribution does not hamper the ability of the test to correctly recognize the cases that $\mathbb{P} = \mathbb{Q}$.

Table 1 shows the FID and KID scores for various models for the CIFAR10, STL10, and CelebA datasets, including results for the smoothed version of ECFD and Laplace ($\mathcal{L}$) & Uniform ($\mathcal{U}$) weighting distributions. The FID and KID scores for the MNIST dataset are shown in Table 2.

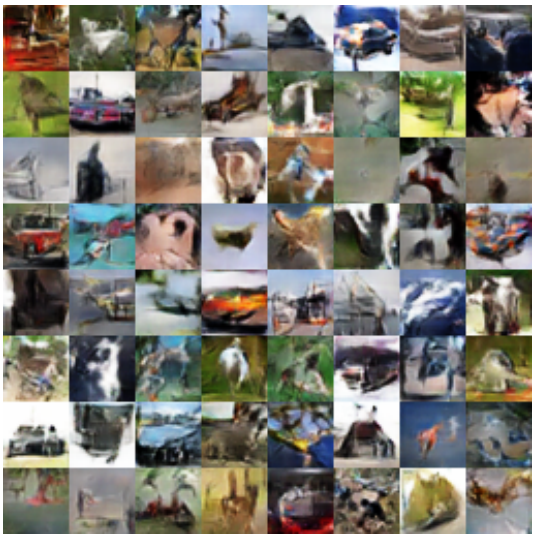
Figures 3, 4, and 5 show random images generated by different GAN models for CIFAR10, CelebA, and STL10 datasets respectively. The images generated by models that do not use gradient penalty (WGAN and MMD-GAN) are less sharp and have more artifacts compared to their GP counterparts. Fig. 6 shows random images generated from OCF-GAN-GP($\mathcal{N}$) trained on the MNIST dataset with a different number of random frequencies (k). It is interesting to note that the change in sample quality is imperceptible even when $k = 1$. Figure 7 shows additional samples from OCF-GAN-GP with a ResNet generator trained on CelebA 128×128.



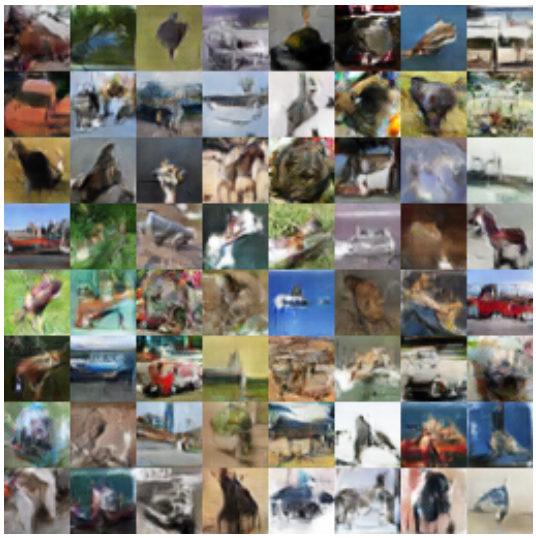
(a) WGAN



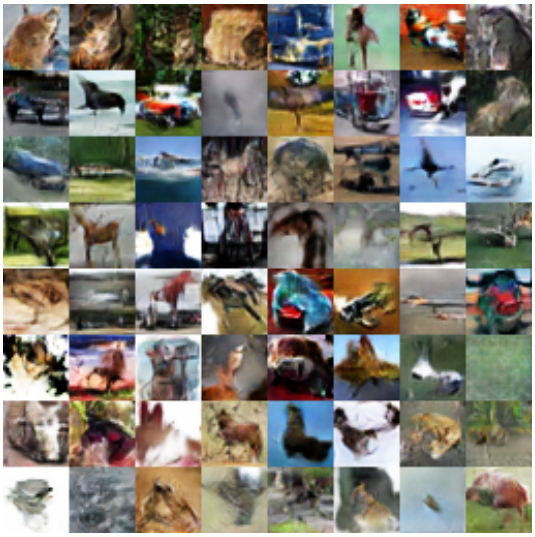
(b) WGAN-GP



(c) MMD-GAN



(d) MMD-GAN-GP



(e) OCF-GAN-GP

(f) CIFAR10 Test Set

Figure 3: Image samples from the different models for the CIFAR10 dataset.

(a) WGAN

(b) WGAN-GP

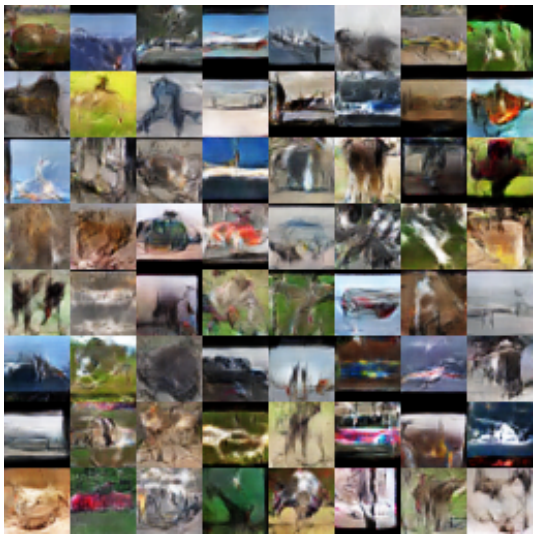
(c) MMD-GAN

(d) MMD-GAN-GP

(e) OCF-GAN-GP

(f) CelebA Real Samples

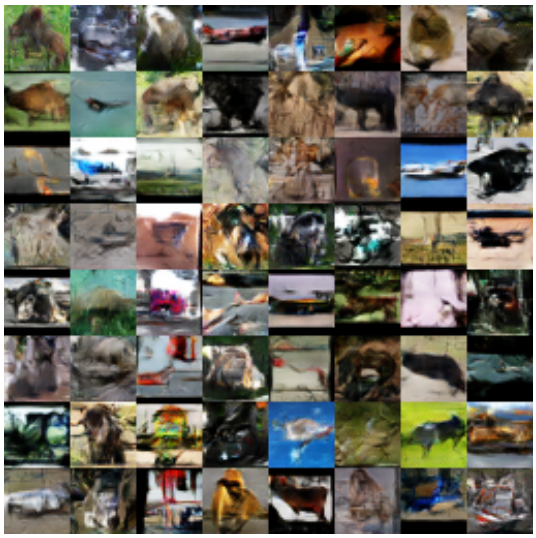
Figure 4: Image samples from the different models for the CelebA dataset.



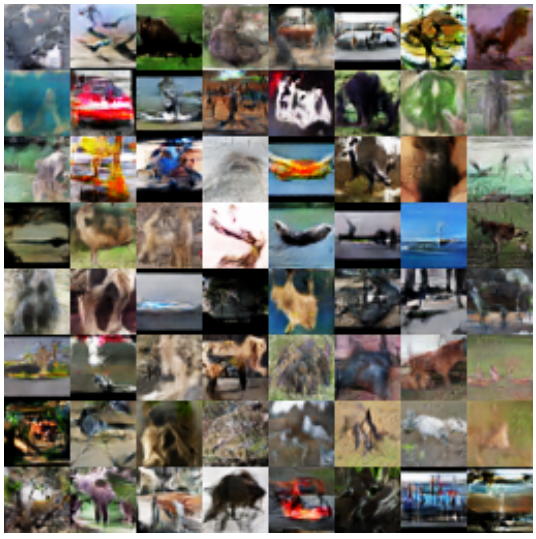
(a) WGAN

(b) WGAN-GP

(c) MMD-GAN



(d) MMD-GAN-GP



(e) OCF-GAN-GP

(f) STL10 Test Set

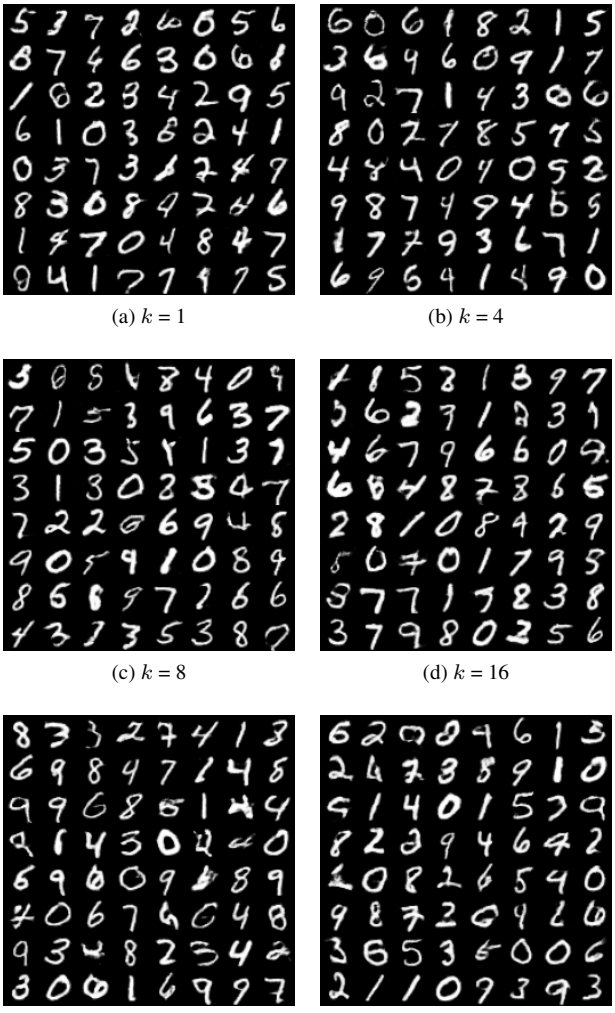
Figure 5: Image samples from the different models for the STL10 dataset.

Table 1: FID and KID ($\times 10^3$) scores (lower is better) for CIFAR10, STL10, and CelebA datasets. Results are averaged over 5 random runs wherever the standard deviation is indicated 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	$\mathcal{N}_{(\hat{\sigma})}$	38.47 (1.00)	20 (1)	32.51 (0.87)	19 (1)	14.91 (0.83)	9 (1)
	$\mathcal{T}_{(\hat{\sigma})}$	37.96 (0.74)	20 (1)	31.03 (0.82)	17 (1)	13.73 (0.56)	8 (1)
	$\mathcal{L}_{(\hat{\sigma})}$	36.90	20	32.09	18	14.96	10
	$\mathcal{U}_{(\hat{\sigma})}$	37.79	21	31.80	18	14.94	10
CF-GAN-Smooth	$\mathcal{N}_{(\sigma=0.5)}$	41.17	24	32.98	19	13.42	9
OCF-GAN-Smooth	$\mathcal{N}_{(\sigma)}$	38.97	21	32.60	18	14.97	9
OCF-GAN-GP	$\mathcal{N}_{(\hat{\sigma})}$	**33.08 (0.26)**	**17 (1)**	**26.16 (0.64)**	**14 (1)**	**9.39 (0.25)**	**5 (1)**
	$\mathcal{T}_{(\hat{\sigma})}$	34.33 (0.77)	18 (1)	26.86 (0.38)	15 (1)	9.61 (0.39)	6 (1)
	$\mathcal{L}_{(\hat{\sigma})}$	36.06	19	29.31	16	11.65	7
	$\mathcal{U}_{(\hat{\sigma})}$	35.14	18	27.62	15	10.29	6

Table 2: FID and KID scores (lower is better) achieved by the various models for the MNIST dataset. Results are averaged over 5 random runs and the standard deviation is indicated in parentheses.

Model	Kernel/Weight	MNIST	
		FID	KID $\times 10^3$
WGAN	–	1.69 (0.09)	20 (2)
WGAN-GP	–	0.26 (0.02)	2 (1)
MMD-GAN	5-RBF	0.68 (0.18)	10 (5)
MMD-GAN-GP$_{L2}$	5-RQ	0.51 (0.04)	6 (2)
CF-GAN	$\mathcal{N}_{(\sigma=1)}$	0.98 (0.33)	16 (10)
	$\mathcal{T}_{(\sigma=0.5)}$	0.85 (0.19)	12 (4)
OCF-GAN	$\mathcal{N}_{(\hat{\sigma})}$	0.60 (0.12)	7 (3)
	$\mathcal{T}_{(\hat{\sigma})}$	0.78 (0.11)	9 (1)
OCF-GAN-GP	$\mathcal{N}_{(\hat{\sigma})}$	0.35 (0.02)	3 (1)
	$\mathcal{T}_{(\hat{\sigma})}$	0.48 (0.06)	6 (1)



(a) $k = 1$ (b) $k = 4$

(c) $k = 8$ (d) $k = 16$

(e) $k = 32$ (f) $k = 64$

Figure 6: Image samples from OCF-GAN-GP for the MNIST dataset trained using different numbers of random frequencies (k).

Figure 7: Image samples for the 128×128 CelebA dataset generated by OCF-GAN-GP with a ResNet generator.

References

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

[2] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936. 4

[3] Herbert Federer. *Geometric measure theory*. Springer, 2014. 2

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

[5] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013. 3

[6] 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. 5, 6

[7] 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

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

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