Interpoint Inception Distance: Gaussian-Free Evaluation of Deep Generative Models

Dariusz Jajeśniak[0009−0005−4625−5898], Piotr Kościelniak[0000−0003−2389−0643], Przemysław Klocek[0009−0001−5427−5208], and Marcin Mazur[0000−0002−3440−8173]

Jagiellonian University, Faculty of Mathematics and Computer Science, Krakow, Poland
marcin.mazur@uj.edu.pl

Abstract. This paper introduces the Interpoint Inception Distance (IID) as a new approach for evaluating deep generative models. It is based on reducing the measurement of discrepancy between multidimensional feature distributions to one-dimensional interpoint comparisons. Our method provides a general tool for deriving a wide range of evaluation measures. The Cramér Interpoint Inception Distance (CIID) is notable for its theoretical properties, including a Gaussian-free structure of feature distribution and a strongly consistent estimator with unbiased gradients. Our experiments, conducted on both synthetic and large-scale real or generated data, suggest that CIID is a promising competitor to the Fréchet Inception Distance (FID), which is currently the primary metric for evaluating deep generative models.

Keywords: Deep generative model · Evaluation measure · Fréchet Inception Distance (FID) · Cramér Distance.

1 Introduction

In recent years, deep generative models (DGMs) have gained tremendous attention. These models are designed and trained to approximate a data distribution via a model distribution. After completing the training, the question arises as to how well this task was accomplished. The research on both of these issues necessitates an appropriate measure that quantifies the difference between the distribution of the training data and the model distribution (a training measure), or the distribution of the test data and the model distribution (an evaluation measure). Fig. 1 presents a diagrammatic representation of the general concept of training and evaluating deep generative models.

The purpose of a discrepancy measure in the training process is to construct an objective function that is optimized on the sets of real and fake data. Common choices include the Kulback-Leibler divergence (used in VAEs [17, 27]) and the Jensen-Shannon divergence (used in GANs [12]). However, using these measures in a learning process can be challenging due to computational problems such as complexity and vanishing gradient. Other common approaches include using the Optimal Transport (Wasserstein) Distance, as seen in WAEs [34], or a

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

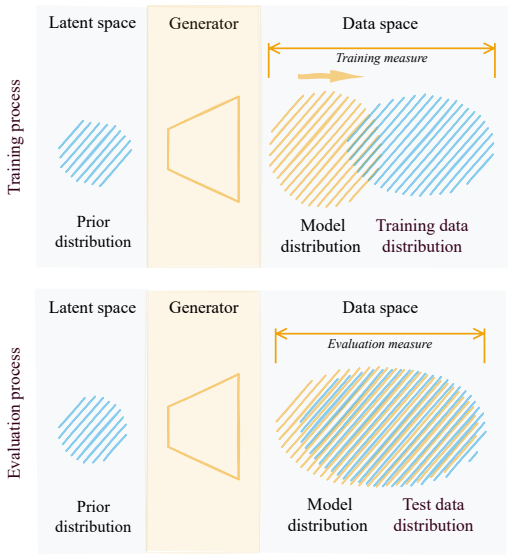

Fig. 1. General concept of training and evaluation of deep generative models. The figures were created using `diagrams.net` software.

kernel-based distance, as seen in CWAEs [19]. Various authors have made significant efforts to propose novel solutions that outperform state-of-the-art methods. However, finding non-adversarial methods that can compete with GANs is still a challenge. These strategies often require the use of certain techniques, such as hierarchical structure in Nouveau VAE (NVAE) [37], stage training in 2Stage-VAE [7], or Latent Trick in Latent Cramér-Wold (LCW) generator [18].

The evaluation of training results is another issue related to generative modeling that requires an appropriate measure. In recent years, this problem has become even more important as deep generative models have matured enough to be used in downstream tasks. Therefore, better and more nuanced evaluation techniques are necessary [26]. Commonly used measures for evaluating processes include Log-Likelihood (LL) [12], Inception Score (IS) [28], and Fréchet Inception Distance (FID) [14]. Additionally, approaches such as Precision and Recall [21] can provide insight into the model's misspecification. Each of these measures has limitations and weaknesses, as described in Section 2. It is important to note that a good evaluation measure should be consistent with human perceptual similarity judgment [5]. However, even for the most popular solutions, including those mentioned above, there are known examples that may not conform to the expected results. This phenomenon may occur despite a well-optimized objective and a good evaluation score [16]. In a critical study provided in [5], the author argues that there is no evaluation method for deep generative models that is sensitive to realistic fake samples, overfitting, mode collapse, transformations, and sample efficiency. Therefore, although many papers have been written on

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

the subject (see [5, 6]), it can still be difficult to determine the most appropriate measure for a fair comparison of models in certain cases.

The aim of this paper is to address the aforementioned challenge. We propose the *Interpoint Inception Distance (IID)* as a novel approach for evaluating deep generative models, based on the concept of reducing the measurement of discrepancy between multidimensional feature distributions to one-dimensional interpoint comparisons, as described by [23]. IID provides a general tool for deriving a wide range of evaluation measures, one of which, the *Cramér Interpoint Inception Distance (CIID)*, is particularly noteworthy for its desirable theoretical properties. Specifically, unlike FID, this method does not assume a Gaussian structure for the feature distribution and allows for a strongly consistent estimator with unbiased gradients, making it a relevant competitor to state-of-the-art solutions[1]. Based on the results of the experiments conducted on both synthetic and large-scale real or generated data, we have found that CIID could be a feasible substitute for FID, which is currently the primary metric for evaluating deep generative models [20].

This paper does not attempt to address all potential issues, but rather takes a first step toward improving the evaluation of deep generative models by measuring distributional discrepancy. We believe that the results obtained will have an impact on further studies in related fields, particularly in deepfake detection, which has become increasingly popular due to the development of generative models. Any improvement in generative modeling makes it harder to distinguish between what is real and what is fake. Therefore, it is crucial to evaluate generative models efficiently to reduce potential risks [6].

2 Related Work

Various measures that quantify differences between distributions have been defined in the literature. These include measures based on information theory, optimal transport theory, and kernel theory. In the following paragraphs, we briefly present our selection of the most significant examples with applications in evaluating deep generative models. For a comprehensive analysis and discussion of alternative methods, including adjustments and enhancements to those outlined in this section, refer to [5, 6].

Discrepancy Measures Based on the Information Theory In general, measures based on information theory rely on entropy $\mathrm{H}(\cdot)$ and/or cross-entropy $\mathrm{H}(\cdot,\cdot)$. The Log-Likelihood (or Evidence) [12, 33] is calculated using the following formula:

$$\mathrm{LL} = \mathbb{E}_{x\sim p_{\mathcal{X}}} \log p_G(x) = -\mathrm{H}(p_{\mathcal{X}}), \tag{1}$$

where $p_{\mathcal{X}}$ represents the real data distribution and p_G represents the model distribution induced on data space $\mathcal{X}$ by the generator network G (i.e., the

[1] Since our method relies on transferring distributions of real and fake data into the inception (feature) space, we only consider feature-based approaches such as FID. For the same reason, we do not discuss other concepts such as Precision and Recall.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

distribution of fake data). As likelihood in higher dimensions is intractable, generated data are often used to approximate $p_G(x)$. This requires the application of suitable estimation techniques, such as the Parzen window approach [33] or the reparametrization trick in VAE [27]. The Log-Likelihood can be considered a universal measure for the training and evaluation of deep generative models [35]. However, due to its low sample efficiency, reliable direct calculation requires large sample sizes. Therefore, for training purposes, it is often substituted with its lower bound, such as the Evidence Lower Bound (ELBO) in VAE, which allows for working with small batches. Furthermore, according to [33], this measure is generally uninformative about the quality of samples. This is because there are known models that produce great samples despite having a poor (low) Log-Likelihood, or vice versa.

The Inception Score [28] is commonly used to evaluate generated images. It can also be useful for training deep generative models, as demonstrated in [29], where a closely related objective for training Category-Aware Generative Adversarial Networks (CatGANs) was proposed. However, to obtain reliable results, it is necessary to evaluate the score on a large number of samples, at least 50k [28]. To calculate the Inception Score, we require the Inception v3 Network [30], which is pre-trained on the ImageNet dataset [8] to capture the desired features of the generated data. The Inception Score is calculated using the following formula:

$$\mathrm{IS} = \exp(\mathbb{E}_{x\sim p_G}\mathrm{KL}(p_L(\cdot|x)||p_L) = \exp(\mathrm{H}(p_L) - \mathbb{E}_{x\sim p_G}\mathrm{H}(\cdot|x)). \tag{2}$$

Here, $\mathrm{KL}(\cdot||\cdot)$ represents the Kullback-Leibler divergence, while $p_L(\cdot|x)$ represents the label (feature) distribution on the inception (feature) space conditioned on $x \in \mathcal{X}$ (so p_L is respective marginal label distribution). The Inception Score has been found to be reasonably correlated with the quality and diversity of generated images, as well as with human judgment [28]. However, it should be noted that the Inception Score does not take into account real data, which may result in models receiving better (higher) scores simply for producing sharp and diverse images, rather than those that follow the underlying distribution [40]. For a more detailed analysis, see [2].

Discrepancy Measures Based on the Optimal Transport Theory The Optimal Transport Distance determines the most cost-effective way to transport one probability measure into another. This is expressed by the following formula (see, e.g., [38]):

$$W_c(p,q) = \inf_{\gamma\in\Gamma(p,q)} \int_{\mathbb{R}^k\times\mathbb{R}^k} c(x,y)\, d\gamma(x,y), \tag{3}$$

where $\Gamma(p,q)$ is the family of joint probability distributions (known as couplings) having p and q as marginals, and $c(\cdot,\cdot)$ is a given transportation cost function. The state-of-the-art deep generative models, Wasserstein GAN [1] and Wasserstein Autoencoder (WAE) [34], aim to minimize the Wasserstein Distance between $p_{\mathcal{X}}$ and p_G, using either l_1^2 (for WGAN) or l_2^2 (for WAE) as the transportation cost function. However, it is important to note that computing directly

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

from Eq. (3) is difficult or even impossible. Therefore, WGAN adheres to the Kantorovich-Rubinstein duality [38] and minimizes a lower bound for $W_{l_1^2}$. Similarly, WAE utilizes Theorem 1 from [34] to create an objective function that comprises a reconstruction error term on $\mathcal{X}$ and an appropriate regularization term on the latent.

As demonstrated above, utilizing the Optimal Transport Distance directly in data space is not feasible due to its intractability in high dimensions. This limitation also applies to the evaluation of models. In this case, however, the Fréchet Inception Distance (FID) [14] can be used. FID is computed as the Wasserstein Distance $W_{l_2^2}$ between the data and model distributions, which are first transported into feature space by the Inception v3 network pre-trained on the ImageNet dataset, and then approximated by the nearest multidimensional Gaussians. While FID has become a standard for evaluating deep generative models trained on image datasets, it has some significant weaknesses. For example, it may overestimate "strange-looking" samples generated by well-optimized objectives (see, e.g., the figures in [16, Appendix E]). A biased estimator of FID that requires large samples makes it essentially unfeasible in the training process. The imposed Gaussian structure raises reasonable doubts about the reliability of this measure. Therefore, it is justifiable to search for improvements [4].

Discrepancy Measures Based on the Kernel Theory A kernel-based measure used to compare two probability distributions is the Maximum Mean Discrepancy (MMD) [24]. For a fixed characteristic kernel function k, it is defined as:

$$\mathrm{MMD}_k(p,q) = \mathbb{E}_{x,x'\sim p}\, k(x,x') - 2\mathbb{E}_{x\sim p, y\sim q}\, k(x,y) + \mathbb{E}_{y,y'\sim q}\, k(y,y'). \tag{4}$$

Because MMD has an unbiased estimator [13], even when used in data space, it has a low sample complexity. This makes it suitable for training models, including generative autoencoders (e.g., MMD-VAE [41]) and GANs (e.g., MMD Net [9]). Additionally, it performs well in a feature space [4, 40]. As an example, the Kernel Inception Distance (KID) is provided [4]. KID uses the polynomial characteristic kernel function and the Inception v3 Network, which is pre-trained on the ImageNet dataset. According to the experimental results presented in [4], KID can be considered a computationally efficient evaluation measure that does not require a Gaussian structure of feature distributions, unlike FID. However, both FID and KID do not differentiate between distributions with the same first three moments (refer to Fig. 2 and [36, Fig. 1]), which indicates room for improvement.

3 Interpoint Inception Distance

This section introduces the Interpoint Inception Distance (IID), a feature-based method for evaluating deep generative models. The approach aims to reduce the measurement of discrepancy between multidimensional distributions to one-dimensional interpoint comparisons [23]. The statement is general, allowing for the derivation of a broad range of evaluation measures. We present an instance of

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

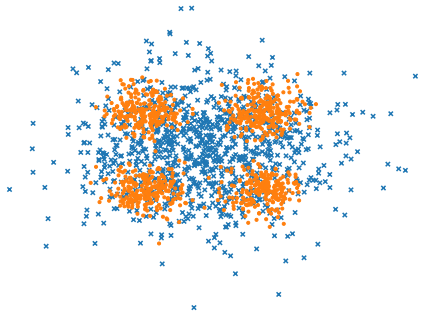

Fig. 2. Example of distributions with identical first three moments (see Section 4), resulting in FID and KID scores close to zero [26, 36].

the Cramér Interpoint Inception Distance (CIID), which has desirable theoretical properties, making it a suitable candidate for an effective evaluation metric.

General Statement Our approach is based on the following theorem, which is a rewrite of Theorem 1 and Remark 4 in [23].

Theorem 1. *Let X_1, X_2, X_3 and Y_1, Y_2, Y_3 be independent copies of two independent k-dimensional random variables X and Y, respectively. Let h be any real-valued nonnegative function*[2] *such that $h(x, y) = 0$ if and only if $x = y$. Then the following conditions are equivalent:*

(i) X and Y follow the same (k-dimensional) distribution,
(ii) $h(X_1, X_2)$, $h(Y_1, Y_2)$, and $h(X_3, Y_3)$ follow the same (univariate) distribution,
(iii) $h(X_1, X_2)$, $h(Y_1, Y_2)$, and $h(X_1, Y_1)$ follow the same (univariate) distribution.

Using Theorem 1, we define the Interpoint Distance (ID) with the following formula:

$$\begin{aligned} \mathrm{ID}(X, Y) = d(h(X_1, X_2), h(Y_1, Y_2)) + d(h(X_1, X_2), h(X_1, Y_1)) \\ + d(h(Y_1, Y_2), h(X_1, Y_1)), \end{aligned} \tag{5}$$

where d is an arbitrary one-dimensional statistical distance. Then, applying Eq. (5) to random variables representing features of real and fake data[3] yields a general evaluation measure called the Interpoint Inception Distance (IID). In this case, the function h measures discrepancies between the features of points

[2] Although it is not necessary for h to be symmetric, it can be interpreted as a type of semi-metric [39] on $\mathbb{R}^k$.

[3] This means that X and Y represent outputs of the Inception v3 Network, which was pre-trained on the ImageNet dataset.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

drawn from the real data distribution or the model distribution. In practice, we use the distance induced by the k-dimensional Euclidean norm $\|\cdot\|$, i.e., $h(x,y) = \|x-y\|$ for $x, y \in \mathbb{R}^k$. In the following paragraph, we discuss a possible choice for d that leads to a special case of IID of particular interest.

Cramér Interpoint Inception Distance (CIID) Our proposal is to use the Cramér Distance [31, 32] as d in the IID formula. The p-th Cramér Distance is defined by the following formula:

$$C^p(S,T) = \int_{-\infty}^{\infty} |F_S(t) - F_T(t)|^p \, dt, \tag{6}$$

where F_S and F_T represent the cumulative distribution functions (CDFs) for one-dimensional random variables S and T. This yields the p-th Cramér Interpoint Inception Distance (CIID^p).

The use of the CIID^p evaluation measure involves the application of an appropriate estimator, as described in the following paragraph, along with the corresponding theoretical analysis.

Estimation of CIID Random variables X and Y^θ (for $\theta \in \Theta$, where Θ is an open subset of the model's parameter space) are taken on the k-dimensional inception space to represent features of real data (following the data distribution $p_{\mathcal{X}}$) and fake data (following the model distribution p_G^θ), respectively. Sequences of independent copies of X and Y^θ are considered, namely

$$\mathrm{X}_{1,n} = (X_{1,1}, \ldots, X_{1,n}),\ \mathrm{X}_{2,n} = (X_{2,1}, \ldots, X_{2,n}), \tag{7}$$

and

$$\mathrm{Y}^\theta_{1,n} = (Y^\theta_{1,1}, \ldots, Y^\theta_{1,n}),\ \mathrm{Y}^\theta_{2,n} = (Y^\theta_{2,1}, \ldots, Y^\theta_{2,n}) \tag{8}$$

(these are interpreted as respective batch samples). The p-th Cramér Interpoint Inception Distance can be estimated using the following formula:

$$\begin{aligned}\widehat{\mathrm{CIID}}^p_n(X, Y^\theta) = {} & C^p(F_{\|\mathrm{X}_{1,n}-\mathrm{X}_{2,n}\|}, F_{\|\mathrm{Y}^\theta_{1,n}-\mathrm{Y}^\theta_{2,n}\|}) + C^p(F_{\|\mathrm{X}_{1,n}-\mathrm{X}_{2,n}\|}, F_{\|\mathrm{X}_{1,n}-\mathrm{Y}^\theta_{1,n}\|}) \\ & + C^p(F_{\|\mathrm{Y}^\theta_{1,n}-\mathrm{Y}^\theta_{2,n}\|}, F_{\|\mathrm{X}_{1,n}-\mathrm{Y}^\theta_{1,n}\|}),\end{aligned} \tag{9}$$

where

$$F_{\|\mathrm{X}_{1,n}-\mathrm{X}_{2,n}\|}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(\|X_{1,i} - X_{2,i}\|), \tag{10}$$

$$F_{\|\mathrm{X}_{1,n}-\mathrm{Y}^\theta_{1,n}\|}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(\|X_{1,i} - Y^\theta_{1,i}\|), \tag{11}$$

and

$$F_{\|\mathrm{Y}^\theta_{1,n}-\mathrm{Y}^\theta_{2,n}\|}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(\|Y^\theta_{1,i} - Y^\theta_{2,i}\|) \tag{12}$$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

are corresponding empirical cumulative distribution functions (ECDFs). (Here, I denotes a set characteristic function.)

Regarding $\widehat{\mathrm{CIID}}_n^p$, our main findings are that it is a strongly consistent estimator and possesses unbiased gradients for $p = 2$, as stated in the following theorem.

Theorem 2. *The Cramér Interpoint Inception Distance estimator defined in Eq.* (9) *satisfies the following conditions:*

$$\widehat{\mathrm{CIID}}_n^p(X, Y^\theta) \to \mathrm{CIID}^p(X, Y^\theta) \text{ if } n \to \infty \tag{13}$$

and

$$\mathbb{E}(\nabla_\theta \widehat{\mathrm{CIID}}_n^2(X, Y^\theta)) = \nabla_\theta \mathrm{CIID}^2(X, Y^\theta). \tag{14}$$

The proof of Theorem 2 follows directly from the lemma below.

Lemma 1. *Let* $\mathrm{S}_n = (S_1, \ldots, S_n)$ *and* $\mathrm{T}_n^\theta = (T_1^\theta, \ldots, T_n^\theta)$ *be sequences of independent copies of one-dimensional random variables* S *and* T^θ*, respectively. Then:*

$$C^2(F_{\mathrm{S}_n}, F_{\mathrm{T}_n^\theta}) \to C^2(F_S, F_{T^\theta}) \text{ if } n \to \infty \tag{15}$$

and

$$\mathbb{E}(\nabla_\theta C^2(F_{\mathrm{S}_n}, F_{\mathrm{T}_n^\theta})) = \nabla_\theta C^2(F_S, F_{T^\theta}). \tag{16}$$

Eq. (15) can be derived from the Glivenko-Cantelli Theorem [10] and the Lebesgue Convergence Theorem [11]. To prove Eq. (16), refer to [3].

4 Experiments

This section presents the experimental study that compares our proposed evaluation measure (CIID) with FID. We start by conducting experiments on synthetic data before transitioning to the case of large-scale real or generated data. The source code is available at https://github.com/djajesniak/CIID.

Experiments on Synthetic Data Let us consider two-dimensional distributions $p \sim N_2(0, I_2)$ and $q_m = q_m^1 \times q_m^2$ for $m \in [0, 1]$, where q_m^1 and q_m^2 are one-dimensional distributions both given by the density function

$$f_m(x) = \frac{1}{2} f_{-m,\sqrt{1-m^2}}(x) + \frac{1}{2} f_{m,\sqrt{1-m^2}}(x) \tag{17}$$

(here $f_{m,s}$ is a density function of the Gaussian $N(m, s)$). Then the first three moments of p and q_m are equal and $p = q_0$. Fig. 3 presents estimated FID and CID[4] values (calculated with different sample sizes) between p and $q_{0.95}$ and between p and p. It is clear that FID does not distinguish between p and $q_{0.95}$, while CID does (note that CID^1 is even sensitive to differences resulting from sample-based estimation). On the other hand, from Fig. 4 we learn that FID does not indicate the difference between p and q_m for any $m \in [0, 1]$, while CID does (but CID^2 starts to discriminate at $m \approx 0.6$).

[4] The Cramér Interpoint Distance (CID) is derived by using the Cramér Distance in the ID formula, instead of the IID formula as in the case of CIID.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

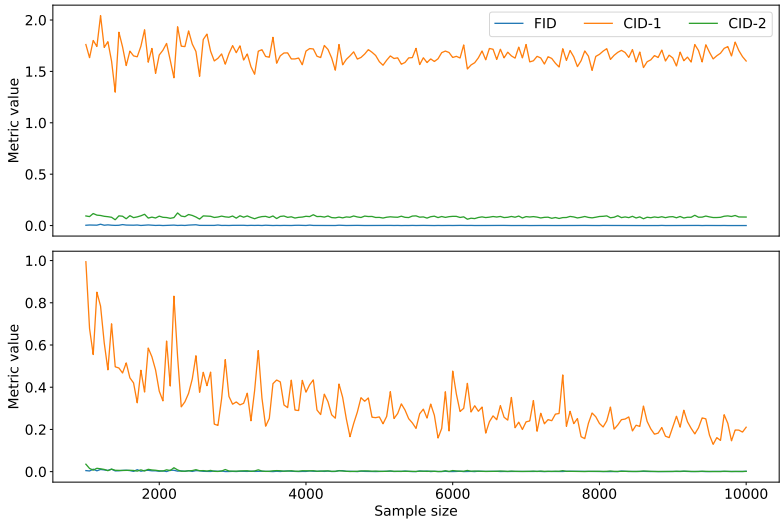

Fig. 3. Estimated FID and CID values (calculated using samples of different sizes) between the two-dimensional distributions p and $q_{0.95}$ (top) or p and p (bottom) that have the same first three moments.

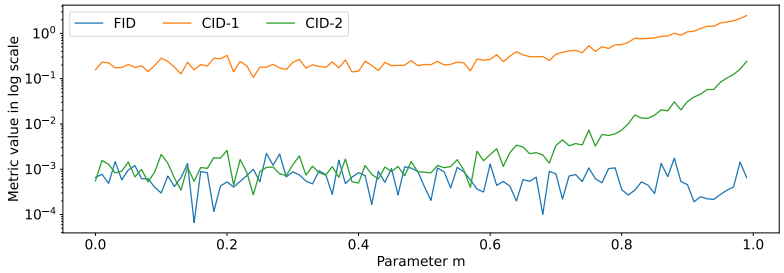

Fig. 4. Estimated FID and CID values on a logarithmic scale (calculated with a sample size of 10000) between the two-dimensional distributions p and q_m with the same first three moments.

Sensitivity to Implemented Disturbances We performed experiments inspired by those initially performed in [14] and then continued in [4]. We applied several different disturbances (i.e., "salt and pepper", Gaussian noise, black rectangles, Gaussian blur, and elastic transform) to 8k-sized data samples from the CelebA [22] and ImageNet [8] datasets, to examine resilience to the noise of CIID compared to FID. Figures 5, 6, 7, 8, and 9 present our experimental results. Each score was scaled to $[0, 1]$ to be plotted on one vertical axis. Generally, CIID behaved comparably to FID, but it seemed to better maintain small disturbances, as shown in Fig. 9, where the FID score for null disturbance is positive, which

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

may be attributed to the curse of dimensionality and the strong bias of its estimator.

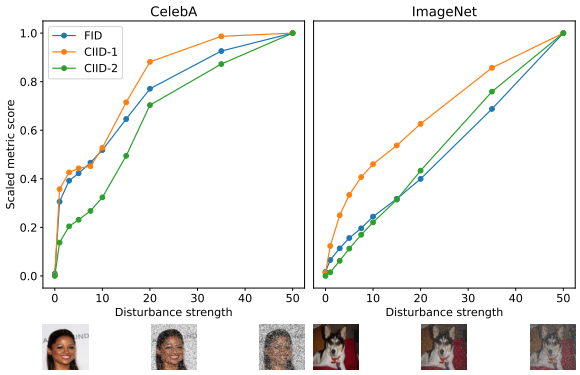

Fig. 5. Salt and pepper. The x-axis represents the percentage of pixels changed to black or white. On the y-axis is the current value of the metric divided by its maximum value.

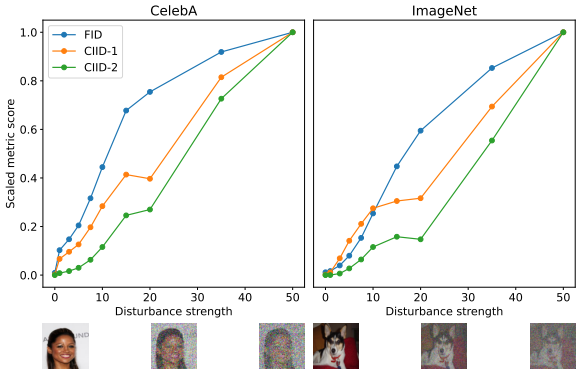

Fig. 6. Gaussian noise. The x-axis represents one-third of the standard deviation of the noise. On the y-axis is the current value of the metric divided by its maximum value.

Impact of Suboptimal Weights We conducted an empirical investigation employing a DCGAN [25] architecture trained on the CelebA dataset, to evaluate the effectiveness of CIID compared to FID throughout the network's training phase. We scrutinized the evaluation measures performance under suboptimal model weights. We run a training procedure 5 times computing evaluation scores on

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

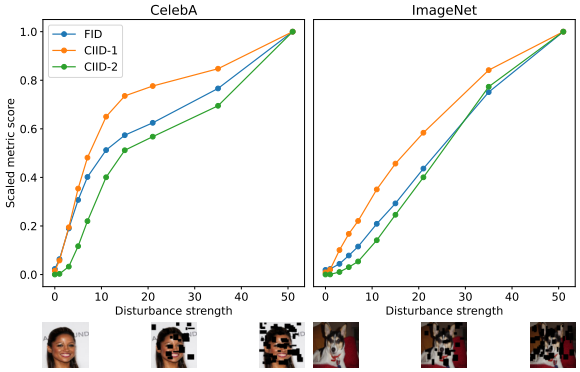

Fig. 7. Black rectangles. The x-axis represents the number of small black rectangles randomly added to the image. On the y-axis is the current value of the metric divided by its maximum value.

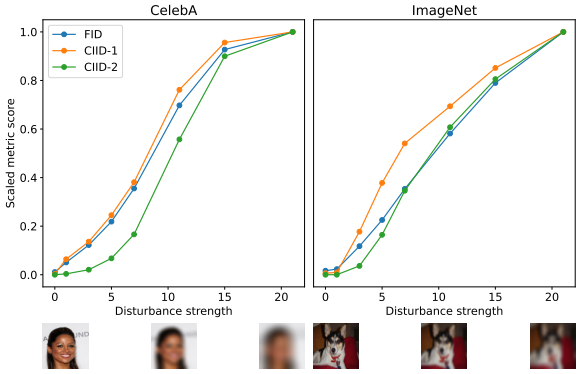

Fig. 8. Gaussian blur. The x-axis represents the strength of the disturbance. Kernel size is equal to $(3x, 3x)$ and sigma parameter to (x, x). On the y-axis is the current value of the metric divided by its maximum value.

8k-sized samples every 1000 batches and then taking the average. The mean values are presented in Fig. 10 (for generated samples refer to Fig. 11). During training, a similar behavior of FID and CIID was observed. Notably, CIID demonstrated a faster descent, reaching lower relative values[5] in comparison to FID.

Variability During the experiments, it was observed that CIID has lower variability than FID. To test it, we computed the CIID and FID ten times between samples from the CelebA dataset and sets of entirely black images (represented

[5] Compared to their values at epoch 0, which are maximal because they are represented as metric values between the original data samples and pure Gaussian noise.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

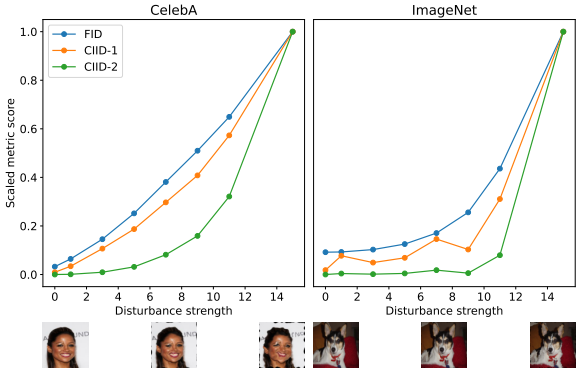

Fig. 9. Elastic transform. The x-axis represents the strength of the disturbance. The alpha parameter is equal to $10x$. On the y-axis is the current value of the metric divided by its maximum value.

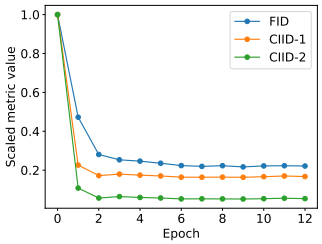

Fig. 10. Scaled values of the metrics during DCGAN training (averaged over 5 runs). The x-axis represents subsequent epochs and the y-axis denotes the current metric value divided by its maximum value.

by tensors containing only zeros). The coefficient of variation for each evaluation measure was then computed. The results showed a coefficient of variation of 0.00135 for FID, 0.00066 for CIID^1, and 0.00055 for CIID^2. This demonstrates that our proposed metrics have roughly half the variability of FID.

5 Conclusions

In this paper, the Interpoint Inception Distance (IID) is introduced as a new method for the evaluation of deep generative models. It is shown that one of its instances, the Cramér Interpoint Inception Distance (CIID), exhibits remarkable theoretical properties, such as a non-Gaussian feature distribution structure and an estimator that yields unbiased gradients and is strongly consistent. Experiments on synthetic and large-scale real or generated data suggest that CIID is a promising competitor to FID, distinguishing well between distributions with the same first three moments, having lower variability, and appearing more objective to small differences.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

Fig. 11. Examples of images generated after 1 epoch (left) and after 12 epochs (right) by DCGAN trained on the CelebA dataset.

Limitations and Future Directions Thus far, our solution has been validated on a limited number of experimental setups that only involve image data. However, we believe that it could also prove useful in the context of text or signal data. Moreover, we have not yet explored the use of CIID to account for *sample novelty*. This is a recent concept introduced in [15], which incorporates vulnerability to overfitting. On the other hand, the existence of an unbiased gradient estimator permits the justification of potential applications of our proposed measure in training processes. The aforementioned factors will serve to guide future directions of our research.

Acknowledgments. The work of D. Jajeśniak and M. Mazur was supported by the National Centre of Science (Poland) Grant No. 2021/41/B/ST6/01370. Some experiments were performed on servers purchased with funds from the flagship project entitled "Artificial Intelligence Computing Center Core Facility" from the DigiWorld Priority Research Area within the Excellence Initiative – Research University program at Jagiellonian University in Krakow. *D. Jajeśniak and M. Mazur acknowledge co-first authorship.*

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning. pp. 214–223. PMLR (2017)
2. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

3. Bellemare, M.G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., Munos, R.: The Cramer distance as a solution to biased Wasserstein gradients. arXiv preprint arXiv:1705.10743 (2017)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: International Conference on Learning Representations (2018)
5. Borji, A.: Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding **179**, 41–65 (2019)
6. Borji, A.: Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding **215**, 103329 (2022)
7. Dai, B., Wipf, D.: Diagnosing and enhancing VAE models. In: International Conference on Learning Representations (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
9. Dziugaite, G.K., Roy, D.M., Ghahramani, Z.: Training generative neural networks via maximum mean discrepancy optimization. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence. pp. 258–267 (2015)
10. Ferguson, T.S.: A course in large sample theory. Routledge (2017)
11. Gariepy, L.E.R., Evans, L.: Measure theory and fine properties of functions, revised edition. Studies in Advanced Mathematics, CRC Press, Boca Raton, FL (2015)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
13. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
15. Jiralerspong, M., Bose, J., Gemp, I., Qin, C., Bachrach, Y., Gidel, G.: Feature likelihood score: Evaluating the generalization of generative models using samples. Advances in Neural Information Processing Systems **36** (2024)
16. Jung, S., Keuper, M.: Internalized biases in Fréchet inception distance. In: NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications (2021)
17. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Knop, S., Mazur, M., Spurek, P., Tabor, J., Podolak, I.: Generative models with kernel distance in data space. Neurocomputing **487**, 119–129 (2022)
19. Knop, S., Spurek, P., Tabor, J., Podolak, I., Mazur, M., Jastrzebski, S.: Cramer-Wold auto-encoder. The Journal of Machine Learning Research **21**(1), 6594–6621 (2020)
20. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of ImageNet classes in Fréchet inception distance. arXiv preprint arXiv:2203.06026 (2022)
21. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738 (2015)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19

23. Maa, J.F., Pearl, D.K., Bartoszyński, R.: Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. The Annals of Statistics **24**(3), 1069–1074 (1996)
24. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al.: Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning **10**(1-2), 1–141 (2017)
25. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
26. Ravuri, S., Rey, M., Mohamed, S., Deisenroth, M.P.: Understanding deep generative models with generalized empirical likelihoods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24395–24405 (2023)
27. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 1278–1286 (2014)
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. Advances in Neural Information Processing Systems **29** (2016)
29. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
31. Székely, G.J.: E-statistics: The energy of statistical samples. Bowling Green State University, Department of Mathematics and Statistics Technical Report **3**(05), 1–18 (2003)
32. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. Journal of statistical planning and inference **143**(8), 1249–1272 (2013)
33. Theis, L., Oord, A.v.d., Bethge, M.: A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844 (2015)
34. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: International Conference on Learning Representations (2018)
35. Tolstikhin, I.O., Gelly, S., Bousquet, O., Simon-Gabriel, C.J., Schölkopf, B.: AdaGAN: Boosting generative models. Advances in Neural Information Processing Systems **30** (2017)
36. Tsitsulin, A., Munkhoeva, M., Mottin, D., Karras, P., Bronstein, A., Oseledets, I., Mueller, E.: The shape of data: Intrinsic distance for data distributions. In: International Conference on Learning Representations (2019)
37. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. Advances in Neural Information Processing Systems **33**, 19667–19679 (2020)
38. Villani, C.: Optimal transport, old and new, vol. 338. Springer (2009)
39. Wilson, W.A.: On semi-metric spaces. American Journal of Mathematics **53**(2), 361–373 (1931)
40. Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K.: An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755 (2018)
41. Zhao, S., Song, J., Ermon, S.: InfoVAE: Balancing learning and inference in variational autoencoders. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5885–5892 (2019)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63749-0_19