PAC-Bayesian Generalization Bounds for Adversarial Generative Models

Sokhna Diarra Mbacke, Florence Clerc, Pascal Germain

PAC-Bayesian Generalization Bounds for Adversarial Generative Models

Sokhna Diarra Mbacke [1] **Florence Clerc** [2] **Pascal Germain** [1]

Abstract

We extend PAC-Bayesian theory to generative models and develop generalization bounds for models based on the Wasserstein distance and the total variation distance. Our first result on the Wasserstein distance assumes the instance space is bounded, while our second result takes advantage of dimensionality reduction. Our results naturally apply to Wasserstein GANs and Energy-Based GANs, and our bounds provide new training objectives for these two. Although our work is mainly theoretical, we perform numerical experiments showing non-vacuous generalization bounds for Wasserstein GANs on synthetic datasets.

1. Introduction

Deep Generative models have become a central research area in machine learning. Two of the most popular families of deep generative models are Variational Autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). GANs are known for producing impressive results in image generation (Brock et al., 2019; Karras et al., 2019), generating fake images indistinguishable from real ones. They also have been applied to video (Acharya et al., 2018), text (de Rosa & Papa, 2021) and protein generation (Repecka et al., 2021).

Motivation. In this work, we study the generalization properties of GANs using PAC-Bayesian theory. Considering the prevalence of GANs in machine learning, the question of generalization is important for numerous reasons. First, quantitatively measuring the discrepancy between the generator's distribution and the true distribution is a difficult problem. Indeed, there are known issues with the current evaluation metrics (Theis et al., 2016; Borji, 2019), and it can be quite challenging to detect when the generator only produces slight variations of the training samples. Moreover, having generalization bounds not only contributes to the theoretical understanding of GANs themselves, but also to the understanding of the structure of real-life datasets, if those can be provably approximated by GAN-generated data. In addition, given that GANs are used for data-augmentation in fields such as medical image classification (see e.g. Frid-Adar et al., 2018), theoretical guarantees can substantiate the soundness of such applications.

1.1. Notations and Preliminaries.

The set of K-Lipschitz functions defined on a space $\mathcal{X}$ is denoted Lip_K and the set of probability measures on $\mathcal{X}$ is denoted $\mathcal{M}_+^1(\mathcal{X})$. Integral Probability Metrics (IPM, see Müller, 1997) are a class of pseudometrics[1] defined on the space of probability measures. Given $P, Q \in \mathcal{M}_+^1(\mathcal{X})$ and a space $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$, the IPM induced by $\mathcal{F}$ is defined as

$$d_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left| \int f\, dP - \int f\, dQ \right|. \tag{1}$$

Examples of IPMs include the total variation distance d_{TV}, corresponding to the case $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} : -1 \leq f \leq 1\}$ and the Wasserstein distance W_1, corresponding to $\mathcal{F} = \mathrm{Lip}_1$.

Generative Adversarial Networks. GANs have two main components: the generator $g \in \mathcal{G}$ and the critic $f \in \mathcal{F}$, where both $\mathcal{G}$ and $\mathcal{F}$ are parameterized by neural networks. Given a n-sized training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ iid sampled from an unknown distribution P^* on a space $\mathcal{X}$, the generator is trained to produce samples that "look like" they came from P^* and the critic is trained to tell apart the real samples from the fake ones. The original GAN of Goodfellow et al. (2014) has been shown to minimize the Jensen-Shannon divergence (JS) between the true distribution P^* and the generator's distribution, denoted P^g.

The original GAN suffers from many problems such as training instability and mode collapse (Salimans et al., 2016).

[1]Université Laval [2]McGill University. Correspondence to: Sokhna Diarra Mbacke <sokhna-diarra.mbacke.1@ulaval.ca>, Florence Clerc <florence.clerc@mail.mcgill.ca>, Pascal Germain <pascal.germain@ift.ulaval.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]For the sake of readability, we will call also call pseudometrics distances in this work.

Upon providing some theoretical explanations for these issues, Arjovsky et al. (2017) introduce the Wasserstein GAN (WGAN), which replaces JS by the Wasserstein-1 distance (Villani, 2009) between P^* and P^g. Thanks to the Kantorovich-Rubinstein duality, the minimization of $W_1(P^*, P^g)$ is equivalent to the following objective:

$$\min_g \max_{f \in \text{Lip}_1} \left\{ \mathop{\mathbb{E}}_{\mathbf{x} \sim P^*} [f(\mathbf{x})] - \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim P^g} [f(\hat{\mathbf{x}})] \right\}. \tag{2}$$

In practice, however, Lip_1 is replaced by a family $\mathcal{F}$ of neural networks referred to as the *critic family*. This leads to the following objective

$$\min_g d_{\mathcal{F}}(P^*, P^g), \tag{3}$$

where $d_{\mathcal{F}}$ is sometimes referred to as the neural divergence or neural IPM (Arora et al., 2017; Biau et al., 2021), since $\mathcal{F}$ is a family of neural networks.

Another variant of GANs is the Energy-based GAN (Zhao et al., 2017), which views the critic as an energy function and uses a margin loss. More precisely, given a positive number m called the margin, EBGAN's critic and generator minimize respectively

$$\min_f \left\{ \mathop{\mathbb{E}}_{\mathbf{x} \sim P^*} f(\mathbf{x}) + \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim P^g} \max(0, m - f(\hat{\mathbf{x}})) \right\},$$

and

$$\min_g \left\{ \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim P^g} f(\hat{\mathbf{x}}) - \mathop{\mathbb{E}}_{\mathbf{x} \sim P^*} f(\mathbf{x}) \right\}.$$

Note that the critic is constrained to be non-negative. Arjovsky et al. (2017) showed that under an optimal critic, the EBGAN's generator minimizes (a constant scaling of) the total variation distance $d_{TV}(P^*, P^g)$.

Generalization. Since the true distribution P^* is unknown and the model has only access to its empirical counterpart P_n^*, the question of generalization naturally arises: How to certify that the learned distribution P^g is "close" to the true one P^*? The goal of this work is to study the generalization properties of GANs using PAC-Bayesian theory. More precisely, we prove non-vacuous PAC-Bayesian generalization bounds for generative models based on the Wasserstein distance and the total variation distance. Since we use the IPM formulation of these metrics, our results are naturally applicable to WGANs and EBGANs.

1.2. Related Works

There is a large body of works dedicated to the understanding of the generalization properties of GANs (Arora et al., 2017; Zhang et al., 2018; Liang, 2021; Singh et al., 2018; Uppal et al., 2019; Schreuder et al., 2021; Biau et al., 2021). Given a family of generators $\mathcal{G}$, a family of critics $\mathcal{F}$, and a discrepancy measure $\mathcal{D}$, the usual goal is to upper bound the quantity $\mathcal{D}(P^*, P^{\hat{g}})$, where $\hat{g}$ is an optimal solution to the empirical problem $\min_{g \in \mathcal{G}} \mathcal{D}(P_n^*, P^g)$. From a statistical perspective, the most common approach is to quantify the rate of convergence of $r(\hat{g}) := \mathcal{D}(P^*, P^{\hat{g}}) - \inf_{g \in \mathcal{G}} \mathcal{D}(P^*, P^g)$, as the size of the training set n goes to infinity. Assuming that the target distribution P^* has a smooth density, Singh et al. (2018); Liang (2021) and Uppal et al. (2019) provide rates of convergence dependent on the ambient dimension of the instance space $\mathcal{X}$ and the complexity of the critic family $\mathcal{F}$. Noting that the density assumption on P^* might be unrealistic in practice, Schreuder et al. (2021) prove rates of convergence assuming P^* is a smooth transformation of the uniform distribution on a low-dimensional manifold. This allows them to derive rates depending on the intrinsic dimension of the data, as opposed to its extrinsic dimension. Under simplicity assumptions on the critic family, Zhang et al. (2018) provide upper bounds for $r(\hat{g})$, when $\mathcal{D}$ is the negative critic loss $d_{\mathcal{F}}$. They first prove general bounds using the Rademacher complexity of $\mathcal{F}$, then bound this complexity in the case when $\mathcal{F}$ is a family of neural networks with certain constraints. More recently, Biau et al. (2021) developed upper bounds for $r(\hat{g})$, but assuming $\mathcal{D}$ is the Wasserstein-1 distance W_1. They argue that since the use of $d_{\mathcal{F}}$ in practice is purely motivated by optimization considerations, W_1 is a better way of assessing the generalization properties of WGANs.

One major distinction between this work and the ones cited above, is that our definition of the generalization error does not explicitly involve the modeling error $\inf_{g \in \mathcal{G}} \mathcal{D}(P^*, P^g)$. Instead, we define the generalization error as the discrepancy between the empirical loss and the expected population loss, allowing us to derive bounds that can be turned into an optimization objective to be minimized by a learning algorithm. Our approach to generalization is closer to the one taken by Arora et al. (2017), who study the generalization properties of GANs by defining the generalization error, for any generator g, as $|\mathcal{D}(P^*, P^g) - \mathcal{D}(P_n^*, P_n^g)|$, where $\mathcal{D}(P_n^*, P_n^g)$ is the discrepancy between the empirical training and generated distributions. They show that models minimizing W_1 do not generalize (in the sense that the generalization error cannot be made arbitrarily small, given a polynomial number of samples), while models minimizing $d_{\mathcal{F}}$ do, under certain conditions on $\mathcal{F}$. A distinction between our approach and the one taken by Arora et al. (2017) is that we define the empirical risk as the expectation $\mathbb{E}\, \mathcal{D}(P_n^*, P_n^g)$ with respect to the fake distribution P_n^g, since in practice, the samples defining P_n^g are drawn anew at each iteration. Moreover, we study distributions $\rho \in \mathcal{M}_+^1(\mathcal{G})$ over the set of generators, as well as individual generators $g \in \mathcal{G}$.

There are other differences between our approach and the ones above. First, our bounds do not depend on the complexity or smoothness of the critic family $\mathcal{F}$. In other words,

our generalization bounds apply systematically to any critic family $\mathcal{F}$, with no distinctions between the cases where $\mathcal{F}$ is a "small" subset of Lip_1 and where $\mathcal{F} = \text{Lip}_1$. The intuitive explanation is that the complexity of the critic family is naturally "embedded" in the empirical and population risks defined in the PAC-Bayesian framework. Second, because of the generality of the PAC-Bayesian theory, we make no assumptions on the structure of the critic family, and some of our bounds do not even make assumptions on the hypothesis space $\mathcal{G}$. The fact that these results can be directly applied to neural networks is a consequence of the generality of PAC-Bayes bounds. Moreover, our bounds provide novel training objectives, giving rise to models that use the training data to not only learn the distribution P^*, but also obtain a risk certificate valid on previously unseen data.

Aside from the study of the generalization properties of GANs, our work relates to the recent work of Ohana et al. (2022), who develop PAC-Bayes bounds for "adaptative" sliced Wasserstein distances. The sliced-Wasserstein distance (SW) (Rabin et al., 2011) is an optimization-focused alternative to the Wasserstein distance. Given distributions P and Q on a high-dimensional space, SW computes $W_1(P_1, Q_1)$ instead of $W_1(P, Q)$, where P_1 and Q_1 are projections of P and Q on a 1-dimensional space. Note that the bounds developed by Ohana et al. (2022) apply to the SW distance, whereas our bounds are developed for the Wasserstein distance between distributions on a high dimensional space. In addition, the bounds of Ohana et al. (2022) focus on the discriminative setting, that is, the models they study optimize to find the projections with the highest discriminative power. Then, they argue that these bounds can be applied to the study of generative models based on the distributional sliced-Wasserstein (Nguyen et al., 2021). In contrast, our results are specifically tailored to the generative modeling setting and provide upper bounds on the difference between the empirical risk of a critic and its population risk.

Finally, we mention a recent article (Chérief-Abdellatif et al., 2022) which uses PAC-Bayes to obtain generalization bounds on the *reconstruction loss* of VAEs. In short, Chérief-Abdellatif et al. (2022) clip the reconstruction loss in order to utilize McAllester's bound (McAllester, 2003), which applies to $[0, 1]$-bounded loss functions. Moreover, they omit the KL-loss, meaning they do not analyze a VAE per se, but simply a stochastic reconstruction machine. Hence, theirs is not a PAC-Bayesian analysis of a generative model, but of a reconstruction model.

1.3. Our Contributions

The primary objective of this work is to extend PAC-Bayesian theory to adversarial generative models. We develop novel PAC-Bayesian generalization bounds for generative models based on the Wasserstein distance and the total variation distance. First, assuming the instance space is bounded, we prove generalizations bounds for Wasserstein models dependent on the diameter of the instance space. Then, we show that one can obtain bounds dependent on the intrinsic dimension, assuming that the distributions are smooth transformations of a distribution on a low-dimensional space. Finally, we exhibit generalization bounds for models based on the total variation distance. To the best of our knowledge, ours are the first PAC-Bayes bounds developed for the generalization properties of generative models. Our results naturally apply to Wasserstein GANs and Energy-Based GANs. Moreover, our bounds provide new training objectives for WGANs and EBGANs, leading to models with statistical guarantees. It is noteworthy that we make no density assumptions on the true and generated distributions. Although our main motivation is theoretical, we perform numerical experiments showing non-vacuous generalization bounds for WGANs on synthetic datasets. We also report the results of preliminary experiments on the MNIST dataset.

2. PAC-Bayesian Theory

PAC-Bayesian theory (introduced by McAllester, 1999) applies Probably Approximately Correct (PAC) inequalities to *pseudo-Bayesian* learning algorithms—whose output could be framed as a *posterior* probability distribution over a class of candidate models— in order to provide generalization bounds for machine learning models. Here, the term generalization bound refers to upper bounds on the discrepancy between a model's empirical loss and its population loss (*i.e.*, the loss on the true data distribution). Optimizing these bounds lead to *self-certified* learning algorithms, that produce models whose behavior on the population is statistically guaranteed to be close to their behavior on the observed samples. PAC-Bayes has been applied to a wide variety of settings such as classification (Germain et al., 2009; Parrado-Hernández et al., 2012), linear regression (Germain et al., 2016; Shalaeva et al., 2020), meta-learning (Amit & Meir, 2018), variational inference for mixture models (Chérief-Abdellatif & Alquier, 2018) and online learning (Haddouche & Guedj, 2022). In recent years, PAC-Bayes has been used to obtain non-vacuous generalization bounds for neural networks (Dziugaite & Roy, 2018; Pérez-Ortiz et al., 2021). See Guedj (2019) and Alquier (2021) for recent surveys.

The wide variety of applications is due to the flexibility of the PAC-Bayesian framework. Indeed, the theory is very general, and requires few assumptions. We consider a training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, iid sampled from an unknown probability distribution P^* over an instance

space $\mathcal{X}$.[2] Given a hypothesis class $\mathcal{H}$ and a real-valued loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0, \infty)$, the empirical and population risks of each hypothesis $h \in \mathcal{H}$ are respectively defined as

$$\hat{\mathcal{R}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, \mathbf{x}_i) \;\text{ and }\; \mathcal{R}(h) = \mathop{\mathbb{E}}_{\mathbf{x}\sim P^*}\left[\ell(h, \mathbf{x})\right].$$

Instead of individual hypotheses $h \in \mathcal{H}$, PAC-Bayes focuses on a *posterior* probability distributions over hypotheses $\rho \in \mathcal{M}_+^1(\mathcal{H})$. These distributions can be seen as *aggregate* hypotheses. Similar to the risks for individual hypotheses, the empirical and true risks of an aggregate hypothesis $\rho \in \mathcal{M}_+^1(\mathcal{H})$ are respectively defined as

$$\hat{\mathcal{R}}_S(\rho) = \mathop{\mathbb{E}}_{h\sim\rho}\left[\hat{\mathcal{R}}_S(h)\right] \;\text{ and }\; \mathcal{R}(\rho) = \mathop{\mathbb{E}}_{h\sim\rho}\left[\mathcal{R}(h)\right].$$

The goal of PAC-Bayesian theory is to provide upper bounds on the discrepancy between $\mathcal{R}(\rho)$ and $\hat{\mathcal{R}}_S(\rho)$ which hold with high probability over the random draw of the training set S. As an example, consider the following general PAC-Bayes bound originally developed by Germain et al. (2009) and further formalized by Haddouche et al. (2021).

Theorem 2.1. *Let $\pi \in \mathcal{M}_+^1(\mathcal{H})$ be a prior distribution independent of the data, $D : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be a convex function, and $\delta \in (0, 1)$ be a real number. With probability at least $1 - \delta$ over the random draw of $S \sim P^{*\otimes n}$, the following holds for any $\rho \in \mathcal{M}_+^1(\mathcal{H})$ such that $\rho \ll \pi$ and $\pi \ll \rho$:*

$$D\left(\mathcal{R}(\rho), \hat{\mathcal{R}}_S(\rho)\right) \leq \mathrm{KL}(\rho \,||\, \pi) + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{h\sim\pi} \mathop{\mathbb{E}}_{S\sim P^{*\otimes n}} e^{D\left(\mathcal{R}(h), \hat{\mathcal{R}}_S(h)\right)}, \tag{4}$$

where $\mathrm{KL}(\rho \,||\, \pi)$ is the Kullback-Leibler divergence between distributions ρ and π.

The left-hand side of Equation (4) quantifies the discrepancy between the true risk $\mathcal{R}(\rho)$ and its empirical counterpart $\hat{\mathcal{R}}_S(\rho)$ for a given training set S, while the complexity term of the right-hand side involves the expectation with respect to $S \sim P^{*\otimes n}$. As the data distribution P^* is unknown, the latter term needs to be upper-bounded in order to obtain a finite and numerically computable bound.

Theorem 2.1 requires $\rho \ll \pi$, which is classic in PAC-Bayes bounds and necessary for the KL-divergence to be defined. However, it also requires $\pi \ll \rho$ which seems a bit more restrictive. As noted by Haddouche et al. (2021), one has to make sure that π and ρ have the same support, which is the case when they are from the same parametric family of distributions, such as Gaussian or Laplace. Although the KL-divergence appears in most PAC-Bayes bounds, some bounds have been developed with the Rényi divergence (Bégin et al., 2016) and IPMs (Amit et al., 2022).

Finally, note that Theorem 2.1 requires the prior distribution π to be independent of the training set S. Even though this restriction makes it easier to bound the exponential moment (Rivasplata et al., 2020), it may also lead to large values of the KL term in practice, since the posterior is likely to be far from the prior. A common strategy is to use a portion of the training data to learn the prior, while making sure this portion is not used in the numerical computation of the bound (Pérez-Ortiz et al., 2021).

Aside from bounds for aggregate hypotheses $\rho \in \mathcal{M}_+^1(\mathcal{H})$, PAC-Bayes bounds can be formulated for individual hypotheses $h \in \mathcal{H}$ as well. Such bounds hold with high probability over the random draw of a single predictor h sampled from the PAC-Bayesian posterior, and have appeared in, e.g., Catoni (2007). In some cases, the derandomization step is quite straightforward, as a result of the structure of the hypotheses. For instance, Germain et al. (2009) utilize the linearity of the hypotheses to express a randomized linear classifier as a single deterministic linear classifier. In the general case, however, it can be quite challenging and costly to derandomize PAC-Bayesian bounds (Neyshabur et al., 2018; Nagarajan & Kolter, 2019; Biggs & Guedj, 2022). Below, we present a result by Rivasplata et al. (2020), who provide a general theorem for derandomizing PAC-Bayes bounds.

Theorem 2.2. *With the definitions and assumptions of Theorem 2.1, given a measurable function $f : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$, the following holds with probability at least $1 - \delta$ over the random draws of $S \sim P^{*\otimes n}$ and $h \sim \rho$:*

$$f(S, h) \leq \log\frac{d\rho}{d\pi}(h) + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{h\sim\pi} \mathop{\mathbb{E}}_{S\sim P^{*\otimes n}} e^{f(S,h)}. \tag{5}$$

Removing the expectation with respect to the hypothesis space, is very useful in applications to neural networks (Viallard et al., 2021). Theorem 2.2 uses the Radon-Nikodym derivative of ρ with respect to π, which can lead to high variance when the bound is used as an optimization objective for neural networks. Viallard et al. (2021) empirically highlighted this phenomenon, and formulated a generic disintegrated bound where the Radon-Nikodym derivative is replaced by the Renyi-divergence between ρ and π.

3. PAC-Bayesian Bounds for Generative Models

This section presents our main results. We consider a metric space $(\mathcal{X}, d)$, an unknown probability measure P^* on $\mathcal{X}$ and a training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ iid sampled from P^*. The

[2] A vast majority of the PAC-Bayes literature is devoted to the prediction setting where each training instance is a pair (x, y) of some features x and a label y. We adopt slightly more general definitions that encompass unsupervised learning.

empirical counterpart of P^* defined by S is denoted P_n^*. We also consider a hypothesis space $\mathcal{G}$ such that each generator $g \in \mathcal{G}$ induces a probability measure P^g on $\mathcal{X}$, from which fake samples $S_g = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n\} \sim P^{g\otimes n}$ are generated. Thus,

$$P_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i} \text{ and } P_n^g = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\mathbf{x}}_i},$$

where $\delta_{\mathbf{x}_i}$ is the Dirac measure on sample $\mathbf{x}_i$.

3.1. Bounds for Wasserstein generative models

Let us consider a subset $\mathcal{F} \subseteq \text{Lip}_1$ that is *symmetric*, meaning $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$. We emphasize that $\mathcal{F}$ can be a small subset of Lip_1, or the whole set Lip_1. Given a generator $g \in \mathcal{G}$, we define its empirical risk as

$$\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) = \mathop{\mathbb{E}}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right], \tag{6}$$

where the expectation is taken with respect to the iid sample S_g that induces P_n^g and $d_{\mathcal{F}}(P_n^*, P_n^g)$ is the IPM induced by $\mathcal{F}$ (Equation 1).

The generalization error is defined as

$$\mathop{\mathbb{E}}_{S\sim P^{*\otimes n}}\left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right),$$

namely the difference between the population and empirical risks. These definitions can be extended to aggregate generators by taking the expectation according to $\rho \in \mathcal{M}_+^1(\mathcal{G})$. The following theorem provides bounds on the generalization error of both (i) aggregate and (ii) individual generators.

Theorem 3.1. *Let $\mathcal{F} \subseteq \text{Lip}_1$ be a symmetric set of real-valued functions on $\mathcal{X}$, $\Delta := \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} d(\mathbf{x},\mathbf{x}') < \infty$ be the diameter of $\mathcal{X}$, $P^* \in \mathcal{M}_+^1(\mathcal{X})$ be the true data-generating distribution and $S \in \mathcal{X}^n$ a n-sized iid sample from P^*. Consider a set of generators $\mathcal{G}$ such that each $g \in \mathcal{G}$ induces a distribution P^g on $\mathcal{X}$, a prior distribution π over $\mathcal{G}$, and real numbers $\lambda > 0$ and $\delta \in (0,1)$.*

(i) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S:

$$\mathop{\mathbb{E}}_{g\sim\rho}\mathop{\mathbb{E}}_{S}\left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathop{\mathbb{E}}_{g\sim\rho}\left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] \leq \frac{1}{\lambda}\left[\text{KL}(\rho\,||\,\pi) + \log\frac{1}{\delta}\right] + \frac{\lambda\Delta^2}{4n}. \tag{7}$$

(ii) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S and $g \sim \rho$:

$$\mathop{\mathbb{E}}_{S}\left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) \leq \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta}\right] + \frac{\lambda\Delta^2}{4n}. \tag{8}$$

Proof Idea. We provide a detailed outline of the proof here. The full details can be found in the supplementary material (Section A.2).

The proof of (i) relies on a technical lemma (Lemma A.3). It is possible to view $d_{\mathcal{F}}(P_n^*, P_n^g)$ as a function $\mathcal{X}^{2n} \to \mathbb{R}$ as P_n^* (resp. P_n^g) is the uniform distribution on n samples that were selected according to P^* (resp. P^g). Lemma A.3 states that $d_{\mathcal{F}}(P_n^*, P_n^g)$ has the bounded differences property with bounds Δ/n, meaning that if we were to change only one sample, the new value of $d_{\mathcal{F}}(P_n^*, P_n^g)$ would differ by at most Δ/n (see Definition A.1). The proof (provided in the appendix) uses properties of the sup and the fact that $\mathcal{F} \subset \text{Lip}_1$.

We then use a result used to prove McDiarmid's inequality (Lemma A.2, previously used by Ohana et al. (2022) for their bounds on the sliced Wasserstein distance) and Fubini's theorem to obtain that

$$\mathop{\mathbb{E}}_{S}[Y] \leq \exp\left[\frac{\lambda^2\Delta^2}{4n}\right],$$

where

$$Y := \mathop{\mathbb{E}}_{g\sim\pi}\mathop{\mathbb{E}}_{S_g}\left[\exp\left[\lambda\left(\mathop{\mathbb{E}}_{S,S_g}\left[d_{\mathcal{F}}(P_n^g, P_n^*)\right] - d_{\mathcal{F}}(P_n^g, P_n^*)\right)\right]\right].$$

Then, Markov's inequality combined with this result yields that with probability at least $1-\delta$ over the random draw of the training set S,

$$Y \leq \frac{1}{\delta}\exp\left[\frac{\lambda^2\Delta^2}{4n}\right].$$

The rest of the proof follows the main steps of the proof of Theorem 2.1, as presented by Haddouche et al. (2021). We use the Radon-Nikodym derivatives to change the expectation over $g \sim \pi$ into an expectation over $g \sim \rho$. Applying log (a monotone increasing function) to the inequality and then using Jensen's inequality for concave functions, with some further rewriting, yields (i).

In order to obtain (ii), we study $\xi = \log \mathbb{E}_S[Y]$. Similarly to what happens in the proof of (i), we have that $\xi \leq \frac{\lambda^2\Delta^2}{4n}$. However, using Jensen's inequality for convex functions, we can exchange the expectation over S_g and exp in the definition of Y to yield a new inequality. Combining it with previous result $\xi \leq \frac{\lambda^2\Delta^2}{4n}$, we obtain that

$$\log \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{g\sim\pi} e^{\lambda\left(\mathbb{E}_S\,\mathbb{E}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right] - \mathbb{E}_{S_g}\,d_{\mathcal{F}}(P_n^*, P_n^g)\right)} \leq \frac{\lambda^2\Delta^2}{4n}.$$

We then use the general desintegrated bound by Rivasplata et al. (2020) stated in Theorem 2.2. We take

$$f(S,g) = \lambda\left(\mathop{\mathbb{E}}_{S,S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right] - \mathop{\mathbb{E}}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right]\right).$$

Previously obtained inequality enables us to bound

$$\log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} \left[e^{f(S,g)} \right] \leq \frac{\lambda^2 \Delta^2}{4n},$$

which gives us the desired result and concludes the proof of (ii). □

Note that our desintegrated bound (8) still has the expectation with respect to the fake sample S_g. Unlike the usual PAC-Bayesian bounds which are mostly applicable to supervised learning, the loss we are bounding requires not only some data from the unknown distribution, but also some data depending on the hypotheses.

Theorem 3.1 requires the samples S_g from the generated distribution to have the same size n as the training set. In practice, this is not a problem, since the user can easily sample from P^g. One might wonder, however, if the bounds could be improved by increasing the number of fake samples. In our approach, the answer is no. Indeed, if the size of S_g is $m \neq n$, then we obtain bounds with last term $\frac{\lambda\Delta^2}{4n}$ replaced by $\frac{\lambda\Delta^2}{4\min(m,n)}$.

Although Theorem 3.1 provides upper bounds on the expected distance between empirical measures, it also implies upper bounds on the distance between the full distributions, as shown in the following corollary.

Corollary 3.2. *With the definitions and assumptions of Theorem 3.1, the following properties hold for any probability measure ρ such that $\rho \ll \pi$ and $\pi \ll \rho$.*

(i) With probability at least $1-\delta$ over the random draw of S:

$$\mathop{\mathbb{E}}_{g\sim\rho} d_{\mathcal{F}}(P^*, P^g) \leq \mathop{\mathbb{E}}_{g\sim\rho} \left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) \right] + \frac{1}{\lambda}\left[\mathrm{KL}(\rho \,||\, \pi) + \log\frac{1}{\delta} \right] + \frac{\lambda\Delta^2}{4n}.$$

(ii) With probability at least $1-\delta$ over the random draw of S and $g \sim \rho$:

$$d_{\mathcal{F}}(P^*, P^g) \leq \mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) + \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta} \right] + \frac{\lambda\Delta^2}{4n}.$$

The proof of Corollary 3.2 is in the supplementary material (Section A.2). As a special case, when $\mathcal{F} = \mathrm{Lip}_1$, Corollary 3.2 provides upper bounds on the Wasserstein distance between the full distributions P^* and P^g.

The manifold assumption. The bounds of Theorem 3.1 depend on the diameter of the instance space, which can be a handicap for real-world datasets such as image datasets. Indeed, the manifold hypothesis states that most high-dimensional real-world datasets lie in the vicinity of low-dimensional manifolds. There is a vast body of work dedicated to testing this assumption and estimating the intrinsic dimension of commonly used datasets (Fodor, 2002; Narayanan & Mitter, 2010; Fefferman et al., 2016; Pope et al., 2021). Moreover, latent variable generative models such as VAEs (Kingma & Welling, 2014), GANs (Goodfellow et al., 2014) and their variants exploit the manifold hypothesis by learning models which approximate distributions over high-dimensional spaces with transformations of low-dimensional latent distributions. This is also a main assumption of Schreuder et al. (2021), whose rates of convergence are dependent on the intrinsic dimension of the instance space. Taking a similar approach, we show that by assuming that the true distribution is a smooth transformation of a latent distribution over a low-dimensional hypercube, we can prove a PAC-Bayesian bound depending on the intrinsic dimension.

Before stating our next result, we recall the definition of a pushforward measure.

Definition 3.3 (Pushforward Measure)**.** Given measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, a probability measure $P_{\mathcal{Z}}$ over $\mathcal{Z}$, and a measurable function $g : \mathcal{Z} \to \mathcal{X}$, the pushforward measure defined by g and $P_{\mathcal{Z}}$ is the probability distribution $g\sharp P_{\mathcal{Z}}$ on $\mathcal{X}$ defined as

$$g\sharp P_{\mathcal{Z}}(A) = P_{\mathcal{Z}}(g^{-1}(A)),$$

for any measurable set $A \subseteq \mathcal{X}$. In more practical terms, sampling $\mathbf{x}$ from $g\sharp P_{\mathcal{Z}}$ means sampling a latent vector $\mathbf{z} \sim P_{\mathcal{Z}}$ first, then setting $\mathbf{x} = g(\mathbf{z})$. For example, a GAN's generator defines a pushforward distribution.

Theorem 3.4. *Let $P^* \in \mathcal{M}_+^1(\mathcal{X})$ be the true data-generating distribution and $S \in \mathcal{X}^n$ a n-sized iid sample from P^*. We consider a set of generators $\mathcal{G}$ such that each $g \in \mathcal{G}$ induces a distribution P^g on $\mathcal{X}$, a prior distribution π over $\mathcal{G}$, and real numbers $\lambda > 0$ and $\delta \in (0,1)$. We also consider a latent space $\mathcal{Z} = [0,1]^{d_{\mathcal{Z}}}$, a latent distribution $P_{\mathcal{Z}}$ on $\mathcal{Z}$, and a true generator $g^* : \mathcal{Z} \to \mathcal{X}$ such that $P^* = g^*\sharp P_{\mathcal{Z}}$ and each $g \in \mathcal{G}$ is a function $g : \mathcal{Z} \to \mathcal{X}$ with $P^g = g\sharp P_{\mathcal{Z}}$. Finally, we assume $\mathcal{G} \cup \{g^*\} \subseteq \mathrm{Lip}_K$ for some positive real number K.*

(i) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S:

$$\mathop{\mathbb{E}}_{g\sim\rho} \mathop{\mathbb{E}}_{S} \left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) \right] - \mathop{\mathbb{E}}_{g\sim\rho} \left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) \right] \leq \frac{1}{\lambda}\left[\mathrm{KL}(\rho \,||\, \pi) + \log\frac{1}{\delta} \right] + \frac{\lambda K^2 d_{\mathcal{Z}}}{4n}. \tag{9}$$

(ii) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S and $g \sim \rho$:

$$\mathop{\mathbb{E}}_{S}\left[\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) \leq \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta}\right] + \frac{\lambda K^2 d_{\mathcal{Z}}}{4n}. \tag{10}$$

The proof can be found in Section A.3 in the appendix. The proof is very similar to that of (i) of Theorem 3.1 but the technical lemma we rely on differs: instead of bounding small perturbations of $d_{\mathcal{F}}(P_n^*, P_n^g)$ using the diameter Δ, we bound those by $\frac{\lambda K^2 d_{\mathcal{Z}}}{n}$ (see Lemma A.5).

As noted by Schreuder et al. (2021), the Lipschitz assumption on the true generator g^* may be realistic in practice. Indeed, the generator learned by a GAN is a Lipschitz function of its input (Seddik et al., 2020) and GAN-generated data has been shown to be a good substitute for real-life data in many applications (Frid-Adar et al., 2018; Wang et al., 2018; Sandfort et al., 2019; Zhang et al., 2022).

A result similar to Corollary 3.2 can be proven for Theorem 3.4 (see Corollary A.6).

3.2. Bounds for Total-Variation generative models

In this section, we prove PAC-Bayesian generalization bounds for models based on the total variation distance. One such model is the EBGAN (Zhao et al., 2017). Indeed, Arjovsky et al. (2017) show that given an optimal critic, the EBGAN's generator minimizes a constant scaling of the total variation distance between the real and fake distributions.

Let us assume $\mathcal{F}$ is a symmetric set of functions $f : \mathcal{X} \to [-1, 1]$ and denote

$$\mathcal{D}_{\mathcal{F}}(P_n^*, P^g) = \mathop{\mathbb{E}}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right] .$$

When $\mathcal{F}$ is the set of all $[-1, 1]$-valued functions defined on $\mathcal{X}$, then $\mathcal{D}_{\mathcal{F}}(P_n^*, P^g)$ is the expected total variation distance between the real and fake empirical distributions.

Theorem 3.5. *Let $(\mathcal{X}, d)$ be a metric space, $P^* \in \mathcal{M}_+^1(\mathcal{X})$ be the true data-generating distribution and $S \in \mathcal{X}^n$ a n-sized iid sample from P^*. Consider a set of generators $\mathcal{G}$ such that each $g \in \mathcal{G}$ induces a distribution P^g on $\mathcal{X}$, a prior distribution π over $\mathcal{G}$ and real numbers $\lambda > 0$ and $\delta \in (0, 1)$.*

(i) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S:

$$\mathop{\mathbb{E}}_{g\sim\rho}\mathop{\mathbb{E}}_{S}\left[\mathcal{D}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathop{\mathbb{E}}_{g\sim\rho}\left[\mathcal{D}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] \leq \frac{1}{\lambda}\left[\mathrm{KL}(\rho \,||\, \pi) + \log\frac{1}{\delta}\right] + \frac{4\lambda}{n}. \tag{11}$$

(ii) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$ and $\pi \ll \rho$, the following holds with probability at least $1-\delta$ over the random draw of S and $g \sim \rho$:

$$\mathop{\mathbb{E}}_{S}\left[\mathcal{D}_{\mathcal{F}}\left(P_n^*, P^g\right)\right] - \mathcal{D}_{\mathcal{F}}\left(P_n^*, P^g\right) \leq \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta}\right] + \frac{4\lambda}{n}. \tag{12}$$

The proof of Theorem 3.5 is in the appendix (Section A.4). The proof is very similar to that of (i) of Theorem 3.1 but the technical lemma we rely on differs: instead of bounding small perturbations of $d_{\mathcal{F}}(P_n^*, P_n^g)$ using the diameter Δ, we bound those by $\frac{2}{n}$ (see Lemma A.7). A result similar to Corollary 3.2 for bounding the distance between the full distributions is also given by Corollary A.8.

Note that unlike the bounds for the Wasserstein distance, the bounds for the total variation distance do not involve the size of the latent or instance space. This is not surprising, since d_{TV} can be seen as a special case of W_1 when the underlying metric on $\mathcal{X}$ is $d = \mathbf{1}_{[x \neq y]}$. Results by Arjovsky et al. (2017) show that the topology induced by the total variation distance is as strong as the one induced by the Jensen Shannon divergence, implying that EBGANs may suffer from some of the issues of the original GAN. Therefore, we focus our experiments on WGANs.

3.3. Rate of convergence

The rate of convergence of the bounds proposed in this work depends on the choice of the hyperparameter λ. Choosing $\lambda = n$ leads to a fast rate of n^{-1}, but the bounds do not converge to 0. The optimal rate for a convergence to 0 is $n^{-1/2}$ and is obtained with $\lambda = \sqrt{n}$. Note that unlike previous results for WGANs (e.g. Biau et al., 2021; Schreuder et al., 2021), our optimal rate of convergence does not depend on the (intrinsic or extrinsic) dimension of the dataset. This is because our rates quantify the speed at which the empirical risk of a distribution P^g reaches its population risk. In contrast, the usual rate of $n^{-1/d}$, where d is the (intrinsic or extrinsic) dimension of the instance space $\mathcal{X}$ quantifies the speed at which the population risk of the distribution $P^{\hat{g}}$ minimizing the empirical problem $\min_{g\in\mathcal{G}} \mathcal{D}(P_n^*, P^g)$ reaches the best possible performance $\inf_{g\in\mathcal{G}} \mathcal{D}(P^*, P^g)$.

4. Experiments

4.1. Preliminary Discussion

Before presenting our experiments, we discuss some of the practical aspects of minimizing PAC-Bayesian bounds. First, we use probabilistic neural networks (Langford & Caruana, 2001) with a Gaussian distribution on each parameter.

Prior learning. As illustrated by Equation (7) the optimization of PAC-Bayes bounds requires a tradeoff between the empirical risk and the KL divergence $\mathrm{KL}(\rho \,||\, \pi)$. When using neural networks, controlling the KL divergence can be challenging, given the high dimensionality of the hypothesis class $\mathcal{H}$ in that case. If the prior π is independent from the data-generating distribution, then an optimal posterior ρ is likely to be very far from π, leading to a KL divergence that is orders of magnitude larger than the empirical risk. To circumvent this issue, it is common in the PAC-Bayes literature (Pérez-Ortiz et al., 2021) to use a portion of the training set to learn the prior π. Given a training set of size n, the prior's mean is learned on $n_0 < n$ samples, the posterior ρ is learned on all n samples, and the bound in computed on the remaining $n - n_0$ samples. Both π and ρ have diagonal covariance matrices, and the prior's covariance matrix is chosen, whereas the posterior's is learned. Note that there are other strategies for choosing a PAC-Bayesian prior, such as fixing the mean vector to $\mathbf{0}$ or random values from the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. However, learning the prior usually leads to a more balanced optimization objective and tighter risk certificates.

The impact of σ_0. In our experiments the hyperparameter, σ_0 plays two roles. First, the prior π is an isotropic Gaussian distribution with a covariance matrix $\sigma_0\mathbf{I}$, and second, the initial value of the posterior's covariance matrix is also $\sigma_0\mathbf{I}$. Note that the covariance matrix of the posterior ρ is a learned diagonal matrix Σ_ρ, but we use $\sigma_0\mathbf{I}$ as the initial value. Hence, σ_0 has a dual impact on the optimization. Since the KL divergence $\mathrm{KL}(\rho \,||\, \pi)$ gets larger as the prior π gets narrower, if σ_0 is too small, then the optimization may be too focused on the KL term, hence neglecting the empirical risk. However, because of the initial value of Σ_ρ, the variance of the posterior ρ is likely to remain close to the variance of the prior, which helps control the KL divergence. On the other hand, if σ_0 is too large, then minimizing $\mathrm{KL}(\rho \,||\, \pi)$ may require the posterior ρ to have a large variance as well, hence putting some weight on suboptimal generators and worsening the generative model's performance. This is illustrated in Figures 1 and 2: when $\sigma_0 = 0.1$, the model's empirical and true risks are relatively large, compared to the other values of σ_0. Figures 4 and 5 in the appendix show samples generated from the different models. One can observe that for both synthetic datasets, when $\sigma_0 = 0.1$, the models do not learn the data-generated distribution well.

Computational cost. The additional cost of training using our objective is very low. Indeed, we optimize the Gibbs posterior during training, instead of averaging over multiple generators $g \sim \rho$. This means that at each iteration, we compute the empirical risk using samples from the training set S and the distribution P^g given by a random generator $g \sim \rho$. Moreover, since both the prior and the posterior are Gaussian distributions with diagonal covariance matrices, the KL divergence is easily computed (see Pérez-Ortiz et al., 2021, Section 5.2).

Numerical computation of the bounds. The numerical computation of our (non-desintegrated) bounds requires the empirical risk, the KL divergence, and an additional term dependent on the data-generating process (for instance, Equation (7) requires the diameter of the instance space, while Equation (9) requires the intrinsic dimension and the Lipschitz constant of g^*). For real-life datasets, both the intrinsic dimension and the smoothness of the data-generating process are unknown. Although there exists estimations of the former for some datasets (e.g. Pope et al., 2021), to the best of our knowledge, there are no estimations of the latter in the literature. Finally, note that although the bounds for WGANs assume the critic family $\mathcal{F} \subseteq \mathrm{Lip}_1$, in practice, once can still optimize the bounds and obtain risk certificates when the critic network's Lipschitz constant K is larger, since $f \in \mathrm{Lip}_1$ if and only if $Kf \in \mathrm{Lip}_K$. Hence, in order to obtain valid risk certificates, one needs to scale the bounds accordingly, which requires the Lipschitz constant of the critic network to be known. This is not the case when using techniques such as the celebrated gradient penalty (Gulrajani et al., 2017).

4.2. Synthetic datasets

We perform experiments on two synthetic datasets: a mixture of 8 Gaussians arranged on a ring, and a mixture of 25 Gaussians arranged on a grid. These are standard synthetic datasets for GAN experiments, see, e.g, Dumoulin et al. (2017); Srivastava et al. (2017); Dieng et al. (2019). In order to formally ensure the diameter of the instance space is finite, we truncate the data so that the first dataset is contained in a disc of radius 3.2 and the second dataset in a square of side 8.2, both centered at the origin. We optimized the right-hand side of Equation (7) plus $\mathbb{E}_{g\sim\rho}\left[\mathcal{W}_\mathcal{F}\left(P_n^*, P^g\right)\right]$, estimating the latter expectation by randomly sampling 100 generators from ρ. In our chosen models, both the generator and critic are fully connected networks, and we use the Björk orthonormalization algorithm (Björck & Bowie, 1971) to enforce Lipschitz continuity on the critic. We performed experiments using both ReLU and GroupSort activations (Anil et al., 2019), and we report the results using GroupSort as it leads to more stability.

The standard deviation of the prior π is denoted σ_0

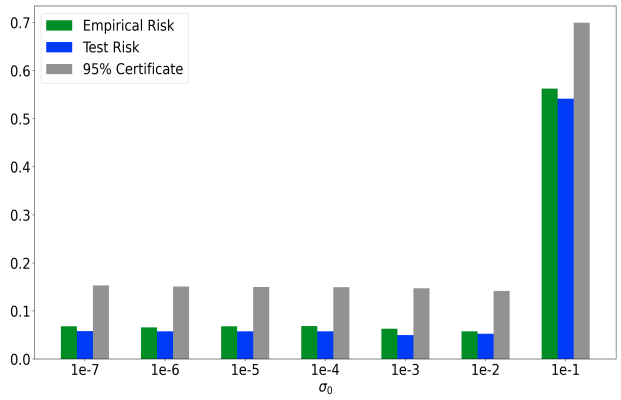

Figure 1. Negative critic losses and risk certificates of a model trained on a mixture of 8 Gaussian distributions arranged on a ring. The x-axis shows the value of the prior parameters' std σ_0. See Appendix (Fig. 4) for illustrations of the generated samples.

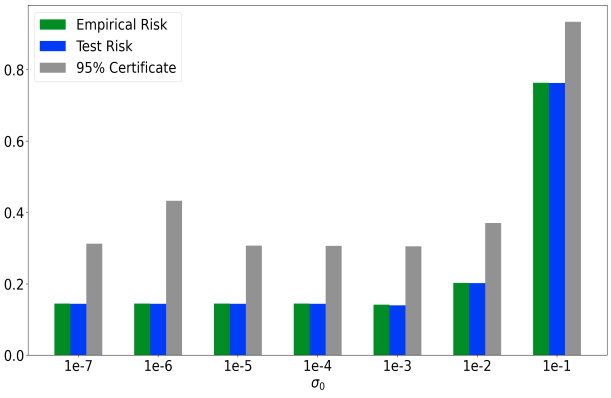

Figure 2. Negative critic losses and risk certificates of a model trained on a mixture of 25 Gaussian distributions arranged on a grid. The x-axis shows the value of the prior parameters' std σ_0. See Appendix (Fig. 5) for illustrations of the generated samples.

and we performed a sweep over the values $\sigma_0 \in \{10^{-7}, 10^{-6}, 10^{-5}, 0.0001, 0.001, 0.01, 0.1\}$, and fix the hyperparameter $\lambda = \frac{n}{1024}$, where n is the size of the training set. The standard deviation of the posterior is learned, and we use σ_0 as a starting point. Samples from the learned distributions are displayed in the appendix (Figures 4 and 5).

Figures 1 and 2 show the risks (negative critic losses) on the training and the test sets, as well as the risk certificate given by Equation (7), for the different values of the hyperparameter σ_0. The expectations with respect to $g \sim \rho$ are approximated by averaging over 100 generators independently sampled from ρ.

We observe that the learned generator has similar empirical and test risks. This is a known asset of learning by optimizing a PAC-Bayesian bound, as it prevents overfitting the training samples. We even notice that some model instances have an empirical risk slightly larger than their test risk, a phenomenon rarely observed when training a discriminative (prediction) model. In our generative setting, this indicates that the critic's ability to distinguish the real samples from the fake ones is consistent, whether the real samples are from the training set or the test set. The computed risk certificates lie in the same order of magnitude than the test loss, which qualifies them as *non-vacuous*.

4.3. Experiments on MNIST

We performed preliminary experiments on the MNIST dataset (Deng, 2012) using the standard DCGAN architecture (Radford et al., 2016), which requires the images to be re-sized to 64 x 64 pixels. Here, we used gradient penalty (Gulrajani et al., 2017) to enforce Lipschitz continuity on the critic. Similar to the experiments on synthetic datasets, we used different values of σ_0 and computed the FID scores on 2000 random samples from each model. See Section B.2 for more details.

5. Conclusion and Future Works

Recent years have seen a growing interest in PAC-Bayesian theory, as a framework for deriving statistical guarantees for a variety of machine learning models (Guedj, 2019). Despite the long list of topics for which PAC-Bayesian bounds have been developed, generative models were missing from this list. In this work, we developed PAC-Bayesian bounds for adversarial generative models. We showed that these bounds can be numerically computed and provide non-vacuous risk certificates for synthetic datasets.

In future works, we will explore risk certificates on real-life datasets. Unlike synthetic datasets for which we can have all the information such as the intrinsic and extrinsic dimensions, real-life datasets come with the challenge that some information is unknown. Computing the bounds of Theorem 3.1 would require the use of the diameter of the instance space, which is clearly irrelevant to the structure of the dataset. On the other hand, the bounds of Theorem 3.4 require some information about the smoothness of the data generating process. In future works, we will explore empirical estimations of that quantity.

Acknowledgements

This research is supported by the Canada CIFAR AI Chair Program, and the NSERC Discovery grant RGPIN-2020-07223. F. Clerc is funded by IVADO through the DEEL project and by a grant from NSERC.

References

Acharya, D., Huang, Z., Paudel, D. P., and Van Gool, L. Towards high resolution video generation with progressive growing of sliced Wasserstein GANs. *arXiv preprint arXiv:1810.02419*, 2018.

Alquier, P. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pp. 205–214. PMLR, 2018.

Amit, R., Epstein, B., Moran, S., and Meir, R. Integral probability metrics PAC-bayes bounds. In *Advances in Neural Information Processing Systems*, 2022.

Anil, C., Lucas, J., and Grosse, R. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pp. 291–301. PMLR, 2019.

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

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pp. 224–232. PMLR, 2017.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. PAC-Bayesian bounds based on the Rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 435–444. PMLR, 2016.

Biau, G., Sangnier, M., and Tanielian, U. Some theoretical insights into Wasserstein GANs. *Journal of Machine Learning Research*, 2021.

Biggs, F. and Guedj, B. On margins and derandomisation in PAC-Bayes. In *International Conference on Artificial Intelligence and Statistics*, pp. 3709–3731. PMLR, 2022.

Björck, Å. and Bowie, C. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Borji, A. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. ISSN 1077-3142.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Catoni, O. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Institute of Mathematical Statistic, 2007.

Chérief-Abdellatif, B.-E. and Alquier, P. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

Chérief-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. On PAC-Bayesian reconstruction guarantees for VAEs. In *International Conference on Artificial Intelligence and Statistics*, pp. 3066–3079. PMLR, 2022.

de Rosa, G. H. and Papa, J. P. A survey on text generation using generative adversarial networks. *Pattern Recognition*, 119:108098, 2021.

Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Dieng, A. B., Ruiz, F. J., Blei, D. M., and Titsias, M. K. Prescribed generative adversarial networks. *arXiv preprint arXiv:1910.04302*, 2019.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

Dziugaite, G. K. and Roy, D. M. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Fodor, I. K. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

Fomin, V., Anmol, J., Desroziers, S., Kriss, J., and Tejani, A. High-level library to help with training neural networks in pytorch. https://github.com/pytorch/ignite, 2020.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, pp. 353–360, 2009.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Guedj, B. A primer on PAC-Bayesian learning. In *Proceedings of the French Mathematical Society*, volume 33, pp. 391–414. Société Mathématique de France, 2019.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haddouche, M. and Guedj, B. Online PAC-Bayes learning. In *Advances in Neural Information Processing Systems*, 2022.

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kingma, D. P. and Welling, M. Auto-encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

Langford, J. and Caruana, R. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Liang, T. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22 (228):1–41, 2021.

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

McAllester, D. A. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

McDiarmid, C. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

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

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Narayanan, H. and Mitter, S. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Ohana, R., Nadjahi, K., Rakotomamonjy, A., and Ralaivola, L. Shedding a PAC-Bayesian light on adaptive sliced-Wasserstein distances. *arXiv preprint arXiv:2206.03230*, 2022.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(1): 3507–3531, 2012.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286. PMLR, 2014.

Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. PAC-Bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16833–16845, 2020.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Sandfort, V., Yan, K., Pickhardt, P. J., and Summers, R. M. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019.

Schreuder, N., Brunel, V.-E., and Dalalyan, A. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pp. 1051–1071. PMLR, 2021.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. In *International Conference on Machine Learning*, pp. 8573–8582. PMLR, 2020.

Shalaeva, V., Esfahani, A. F., Germain, P., and Petreczky, M. Improved PAC-Bayesian bounds for linear regression. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 5660–5667, 2020.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. Veegan: Reducing mode collapse in GANs using implicit variational learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and 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.

Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016.

Uppal, A., Singh, S., and Póczos, B. Nonparametric density estimation & convergence rates for GANs under besov ipm losses. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Viallard, P., Germain, P., Habrard, A., and Morvant, E. A general framework for the disintegration of PAC-Bayesian bounds. *arXiv preprint arXiv:2102.08649*, 2021.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wang, Y.-X., Girshick, R., Hebert, M., and Hariharan, B. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7278–7286, 2018.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Ying, Y. McDiarmid's inequalities of Bernstein and Bennett forms. *City University of Hong Kong*, pp. 318, 2004.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

Zhang, Y., Gupta, A., Saunshi, N., and Arora, S. On predicting generalization using GANs. In *International Conference on Learning Representations*, 2022.

Zhao, J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017.

A. Proofs

A.1. Preliminaries

We start this section with the following definition.

Definition A.1 (Bounded differences)**.** A function $f : \mathcal{X}^n \to \mathbb{R}$ is said to have the *bounded differences property* if for some non-negative constants $c_1, \ldots, c_n$, we have for any $1 \leq i \leq n$,

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

In other words, if we change the i^{th} argument of f while keeping all the others fixed, the value of the function cannot change by more than c_i.

The following lemma is used to prove a special case of McDiarmid's inequality (McDiarmid, 1989).

Lemma A.2. *Let $f : \mathcal{X}^n \to \mathbb{R}$ be a function that has the bounded differences property with constants $c_i, 1 \leq i \leq n$. Then, denoting $Z = f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we have that*

$$\mathbb{E}\left[\exp\left[\lambda\left(\mathbb{E}\left[Z\right] - Z\right)\right]\right] \leq \exp\left[\lambda^2 \nu / 8\right], \tag{13}$$

where $\nu = \sum_{i=1}^n c_i^2$.

Below, we include a summary of the proof by Ying (2004) (with minor modifications) for completeness.

Proof. The proof relies on the clever use of the following functions: for each $1 \leq k \leq n$, we define a function $g_k : \mathcal{X}^k \to \mathbb{R}$ by

$$g_k(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \mathop{\mathbb{E}}_{\mathbf{x}_k,\ldots,\mathbf{x}_n}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right] - \mathop{\mathbb{E}}_{\mathbf{x}_{k+1},\ldots,\mathbf{x}_n}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right], \text{ when } k < n,$$
$$g_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathop{\mathbb{E}}_{\mathbf{x}_n}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right] - f(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

For every k, the function g_k satisfies the following results:

$$\mathop{\mathbb{E}}_{\mathbf{x}_k}\left[g_k(\mathbf{x}_1, \ldots, \mathbf{x}_k)\right] = 0 \qquad \text{and} \qquad 0 \leq b_k - a_k \leq c_k,$$

where we have denoted $a_k = \inf_{\mathbf{x}_k} g_k(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and $b_k = \sup_{\mathbf{x}_k} g_k(\mathbf{x}_1, \ldots, \mathbf{x}_k)$. These results allow us to conclude using Hoeffding's lemma that for every k

$$\int_{\mathcal{X}} e^{\lambda g_k(\mathbf{x}_1,\ldots,\mathbf{x}_k)}\, dP^*(\mathbf{x}_k) \leq e^{\lambda^2 c_k^2/8}.$$

Finally, we use the fact that

$$\mathbb{E}\left[Z\right] - Z = \sum_{k=1}^n g_k(\mathbf{x}_1, \ldots, \mathbf{x}_k)$$

to rewrite $\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z]-Z)}\right]$ using Fubini's thorem. We get the desired result by induction using previously-shown inequality. □

A.2. Proof of Theorem 3.1

Lemma A.3. *Let P, Q be probability measures on $\mathcal{X}$ and P_n, Q_n be the empirical distributions corresponding to the iid samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n \sim Q$ respectively, meaning*

$$P_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}(\mathbf{x}) \quad \text{and} \quad Q_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{y}_i}(\mathbf{x})$$

for any $\mathbf{x} \in \mathcal{X}$.

Let $\mathcal{F} \subseteq \mathrm{Lip}_1$ be a symmetric subset of Lip_1. Recall the definition of the IPM defined by $\mathcal{F}$:

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left\{ \int f \, dP - \int f \, dQ \right\}.$$

Then the empirical IPM $d_{\mathcal{F}}(P_n, Q_n)$, seen as a function $\mathcal{X}^{2n} \to \mathbb{R}$, has the bounded differences property with $c_i = \frac{\Delta}{n}$ and $\Delta = \mathrm{diam}(\mathcal{X})$, for all $1 \leq i \leq 2n$.

Proof. We show, without loss of generality, that $c_n = \frac{\Delta}{n}$. We have

$$\begin{aligned}
&d_{\mathcal{F}}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) - d_{\mathcal{F}}(\mathbf{x}_1, \ldots, \mathbf{x}'_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) \\
&= \frac{1}{n} \left\{ \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} f(\mathbf{x}_i) - \sum_{i=1}^{n} f(\mathbf{y}_i) \right] - \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n-1} f(\mathbf{x}_i) + f(\mathbf{x}'_n) - \sum_{i=1}^{n} f(\mathbf{y}_i) \right] \right\} \\
&\leq \frac{1}{n} \left\{ \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} f(\mathbf{x}_i) - \sum_{i=1}^{n} f(\mathbf{y}_i) - \sum_{i=1}^{n-1} f(\mathbf{x}_i) - f(\mathbf{x}'_n) + \sum_{i=1}^{n} f(\mathbf{y}_i) \right] \right\} \\
&= \frac{1}{n} \sup_{f \in \mathcal{F}} \left[f(\mathbf{x}_n) - f(\mathbf{x}'_n) \right] \\
&\leq \frac{\Delta}{n}.
\end{aligned}$$

The first inequality (second to third lines) follows from a property of the supremum and the last inequality follows from $\mathcal{F} \subseteq \mathrm{Lip}_1$ and $\mathrm{diam}(\mathcal{X}) = \Delta$. □

When $\mathcal{F} = \mathrm{Lip}_1$, then Lemma A.3 states that the Wasserstein distance between empirical measures has the bounded differences property, which follows from a result by Weed & Bach (2019).

Combining Lemmas A.2 and A.3 yields the following result.

Proposition A.4. *Let P and Q be two probability measures on $\mathcal{X}$ and P_n, Q_n be their empirical counterparts corresponding to S_P and S_Q respectively. Then*

$$\mathbb{E}\left[\exp\left[\lambda \left(\mathbb{E}\left[d_{\mathcal{F}}(P_n, Q_n) \right] - d_{\mathcal{F}}(P_n, Q_n) \right) \right] \right] \leq \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right],$$

where both expectations are taken over $(S_P, S_Q) \sim P^{\otimes n} \times Q^{\otimes n}$.

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1.

(i) For a given generator $g \in \mathcal{G}$, Proposition A.4 implies

$$\mathop{\mathbb{E}}_{S, S_g} \exp\left[\lambda \left(\mathop{\mathbb{E}}_{S, S_g} \left[d_{\mathcal{F}}(P_n^g, P_n^*) \right] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \leq \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right],$$

where we write $\mathbb{E}_{S, S_g}$ instead of $\mathbb{E}_{(S, S_g) \sim P^{* \otimes n} \times P^{g \otimes n}}$ in order to simplify the notation. Taking the average with respect to the prior $\pi \in \mathcal{M}_+^1(\mathcal{G})$ and using Fubini's theorem, we get

$$\mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g \sim \pi} \mathop{\mathbb{E}}_{S_g} \left[\exp\left[\lambda \left(\mathop{\mathbb{E}}_{S, S_g} \left[d_{\mathcal{F}}(P_n^g, P_n^*) \right] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right] \leq \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right]. \tag{14}$$

Now, defining

$$Y \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{g\sim\pi} \mathop{\mathbb{E}}_{S_g} \left[\exp\left[\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right],$$

we have that Y is a positive random variable and Markov's inequality implies

$$\mathbb{P}\left[Y \geq \frac{1}{\delta} \mathbb{E}\left[Y\right] \right] \leq \delta$$

for any real number $\delta \in (0, 1)$. Taking complementary event, we get that with probability at least $1 - \delta$ over the random draw of $S \sim P^{*\otimes n}$,

$$Y \leq \frac{1}{\delta} \mathbb{E}\left[Y\right] \leq \frac{1}{\delta} \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right],$$

where the last inequality follows from (14). So we've just shown that with probability at least $1 - \delta$ over the random draw of the training set $S \sim P^{*\otimes n}$,

$$\mathop{\mathbb{E}}_{g\sim\pi} \mathop{\mathbb{E}}_{S_g} \left[\exp\left[\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right] \leq \frac{1}{\delta} \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right].$$

Now, assume $\rho \in \mathcal{M}_+^1(\mathcal{G})$ is such that $\pi \ll \rho$ and $\rho \ll \pi$. We can change the expectation with respect to π into an expectation with respect to ρ using the Radon-Nikodym derivative $\frac{d\pi}{d\rho}$ to obtain

$$\mathop{\mathbb{E}}_{g\sim\rho} \left[\frac{d\pi}{d\rho} \mathop{\mathbb{E}}_{S_g} \left[\exp\left[\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right] \right] \leq \frac{1}{\delta} \exp\left[\frac{\lambda^2 \Delta^2}{4n} \right].$$

Taking the logarithm on both sides and using Jensen's inequality yields

$$\mathop{\mathbb{E}}_{g\sim\rho} \left[\mathop{\mathbb{E}}_{S_g} \left[\log\left(\frac{d\pi}{d\rho} \right) + \lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right] \leq \log \frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{4n},$$

which is equivalent to

$$- \mathop{\mathbb{E}}_{g\sim\rho} \left[\log\left(\frac{d\rho}{d\pi} \right) \right] + \mathop{\mathbb{E}}_{g\sim\rho} \left[\mathop{\mathbb{E}}_{S_g} \left[\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right) \right] \right] \leq \log \frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{4n},$$

since $\rho \ll \pi$ and $\frac{d\pi}{d\rho} = \left(\frac{d\rho}{d\pi} \right)^{-1}$. This last inequality can be re-written as follows:

$$\lambda \left(\mathop{\mathbb{E}}_{g\sim\rho} \mathop{\mathbb{E}}_{S_g} \left[\mathop{\mathbb{E}}_{S} [d_{\mathcal{F}}(P_n^g, P_n^*)] - d_{\mathcal{F}}(P_n^g, P_n^*) \right] \right) \leq \mathrm{KL}(\rho \,||\, \pi) + \log \frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{4n},$$

or, using the linearity of the expectation and the definition $\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right) = \mathbb{E}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right]$,

$$\lambda \left(\mathop{\mathbb{E}}_{g\sim\rho} \mathop{\mathbb{E}}_{S} [\mathcal{W}_{\mathcal{F}}\left(P_n^*, P^g\right)] - \mathop{\mathbb{E}}_{g\sim\rho} \mathcal{W}_{\mathcal{F}}\left(P_n^g, P_n^*\right) \right) \leq \mathrm{KL}(\rho \,||\, \pi) + \log \frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{4n}.$$

The proof above uses the ideas of Germain et al. (2009) and Haddouche et al. (2021). We provided details for completeness and clarity.

(ii) Denote

$$\xi = \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} \mathop{\mathbb{E}}_{S_g} \left[e^{\lambda \left(\mathbb{E}_S \mathbb{E}_{S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] - d_{\mathcal{F}}(P_n^*, P_n^g) \right)} \right].$$

First, using Fubini's theorem and Proposition A.4 we have

$$\begin{aligned}
\xi &= \log \mathop{\mathbb{E}}_{g\sim\pi} \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{S_g} \left[e^{\lambda \left(\mathbb{E}_S \mathbb{E}_{S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] - d_{\mathcal{F}}(P_n^*, P_n^g) \right)} \right] \\
&\leq \log \mathop{\mathbb{E}}_{g\sim\pi} \left[e^{\frac{\lambda^2 \Delta^2}{4n}} \right] \\
&= \frac{\lambda^2 \Delta^2}{4n}.
\end{aligned}$$

Then, using the convexity of the exponential and Jensen's inequality, we obtain

$$\xi = \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} \mathop{\mathbb{E}}_{S_g} \left[e^{\lambda\left(\mathbb{E}_S \mathbb{E}_{S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)] - d_{\mathcal{F}}(P_n^*, P_n^g)\right)} \right]$$

$$\geq \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} e^{\lambda \mathbb{E}_{S_g}\left(\mathbb{E}_S \mathbb{E}_{S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)] - d_{\mathcal{F}}(P_n^*, P_n^g)\right)}$$

$$= \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} e^{\lambda\left(\mathbb{E}_S \mathbb{E}_{S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathbb{E}_{S_g} d_{\mathcal{F}}(P_n^*, P_n^g)\right)}.$$

The combination of these two inequalities yields

$$\log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} e^{\lambda\left(\mathbb{E}_S \mathbb{E}_{S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathbb{E}_{S_g} d_{\mathcal{F}}(P_n^*, P_n^g)\right)} \leq \frac{\lambda^2 \Delta^2}{4n}. \tag{15}$$

Now, a result by Rivasplata et al. (2020) states that for any measurable function f, the following holds with probability at least $1-\delta$ over the random draw of $S \sim P^{*\otimes n}$ and $g \sim \rho$:

$$f(S,g) \leq \log \frac{d\rho}{d\pi}(g) + \log \frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} \left[e^{f(S,g)} \right].$$

Taking

$$f(S,g) = \lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathop{\mathbb{E}}_{S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] \right),$$

we get

$$\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathop{\mathbb{E}}_{S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] \right) \leq \log \frac{d\rho}{d\pi}(g) + \log \frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{g\sim\pi} e^{\lambda\left(\mathbb{E}_{S,S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathbb{E}_{S_g}[d_{\mathcal{F}}(P_n^*, P_n^g)]\right)}.$$

Combining this result with (15), we obtain

$$\lambda \left(\mathop{\mathbb{E}}_{S,S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] - \mathop{\mathbb{E}}_{S_g} [d_{\mathcal{F}}(P_n^*, P_n^g)] \right) \leq \log \frac{d\rho}{d\pi}(g) + \log \frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{4n}.$$

□

Remark. As stated in the main paper, increasing the number of fake samples S_g from n to m worsens the bounds. This is because in that case, the constants $c_i = \frac{\Delta}{n}$ of Lemma A.3 become $c_i = \max(\frac{\Delta}{n}, \frac{\Delta}{m})$, leading to a worse bound.

Next, we prove Corollary 3.2.

Proof of Corollary 3.2. Denote $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $S_g = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ the iid datasets corresponding to the empirical distributions P_n^* and P_n^g respectively. The properties of the supremum imply

$$\sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{S,S_g} \left[\int f\, dP_n^g - \int f\, dP_n^* \right] \right\} \leq \mathop{\mathbb{E}}_{S,S_g} \left[\sup_{f\in\mathcal{F}} \left\{ \int f\, dP_n^g - \int f\, dP_n^* \right\} \right].$$

Moreover, since S and S_g are iid datasets, we have

$$\mathop{\mathbb{E}}_{S} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) \right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim P^*} [f(\mathbf{x})] \quad \text{and} \quad \mathop{\mathbb{E}}_{S_g} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{y}_i) \right] = \mathop{\mathbb{E}}_{\mathbf{y}\sim P^g} [f(\mathbf{y})].$$

Therefore,

$$d_{\mathcal{F}}(P^*, P^g) = \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{S,S_g} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) - \frac{1}{n} \sum_{i=1}^n f(\mathbf{y}_i) \right] \leq \mathop{\mathbb{E}}_{S,S_g} \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) - \frac{1}{n} \sum_{i=1}^n f(\mathbf{y}_i) \right\} = \mathop{\mathbb{E}}_{S,S_g} d_{\mathcal{F}}(P_n^*, P^g).$$

Combining this inequality with Theorems 3.1-(i) and 3.1-(ii) yields the desired results. □

A.3. Proof of Theorem 3.4

The proof of Theorem 3.4 is similar to the proof of Theorem 3.1. The only difference is that instead of Lemma A.3, we use the following result.

Lemma A.5. *Let $\mathcal{Z} = [0,1]^{d_\mathcal{Z}}$ and $P_\mathcal{Z}$ be a probability measure on $\mathcal{Z}$. Let P, Q be probability measures on $\mathcal{X}$ such that $P = g_1 \sharp P_\mathcal{Z}$ and $Q = g_2 \sharp P_\mathcal{Z}$ with $g_1, g_2 \in \mathrm{Lip}_K$, $K \geq 1$. Let P_n, Q_n be the empirical distributions corresponding to the iid samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n \sim Q$ respectively. Then the function $W_\mathcal{F} : \mathcal{X}^{2n} \to \mathbb{R}$, defined as*

$$W_\mathcal{F}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) = W_\mathcal{F}(P_n, Q_n),$$

has the bounded differences property with $c_i = \frac{K\sqrt{d_\mathcal{Z}}}{n}$, for all $1 \leq i \leq 2n$.

Proof. First, let $w_1, \ldots, w_n, w'_n, z_1, \ldots, z_n \sim P_\mathcal{Z}$ such that for all $1 \leq i \leq n$,

$$\mathbf{x}_i = g_1(w_i), \ \mathbf{x}'_n = g_1(w'_n) \quad \text{and} \quad \mathbf{y}_i = g_2(\mathbf{z}_i). \tag{16}$$

We have

$$\begin{aligned}
&W_\mathcal{F}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) - W_\mathcal{F}(\mathbf{x}_1, \ldots, \mathbf{x}'_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) \\
&= \frac{1}{n}\left\{\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^n f(\mathbf{x}_i) - \sum_{i=1}^n f(\mathbf{y}_i)\right] - \sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{n-1} f(\mathbf{x}_i) + f(\mathbf{x}'_n) - \sum_{i=1}^n f(\mathbf{y}_i)\right]\right\} \\
&\leq \frac{1}{n}\left\{\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^n f(\mathbf{x}_i) - \sum_{i=1}^n f(\mathbf{y}_i) - \sum_{i=1}^{n-1} f(\mathbf{x}_i) - f(\mathbf{x}'_n) + \sum_{i=1}^n f(\mathbf{y}_i)\right]\right\} \\
&= \frac{1}{n}\sup_{f\in\mathcal{F}}\left[f(\mathbf{x}_n) - f(\mathbf{x}'_n)\right] \\
&\leq \frac{K\sqrt{d_\mathcal{Z}}}{n}.
\end{aligned}$$

In order to prove the last inequality, we just need to show that for any $f \in \mathcal{F}$, $f(\mathbf{x}_n) - f(\mathbf{x}'_n) \leq K\sqrt{d_\mathcal{Z}}$. Let $f \in \mathcal{F}$. Using (16) and the assumptions $\mathcal{F} \subseteq \mathrm{Lip}_1$, $g_1 \in \mathrm{Lip}_K$ and $\mathcal{Z} = [0,1]^{d_\mathcal{Z}}$, which implies $\mathrm{diam}(\mathcal{Z}) = \sqrt{d_\mathcal{Z}}$, we have

$$f(\mathbf{x}_n) - f(\mathbf{x}'_n) = f(g_1(w_n)) - f(g_1(w'_n)) \leq K\sqrt{d_\mathcal{Z}}.$$

□

The following result is similar to Corollary 3.2 and bounds the distance between the full distributions.

Corollary A.6. *With the definitions and assumptions of Theorem 3.4, the following properties hold for any probability measure ρ such that $\rho \ll \pi$ and $\pi \ll \rho$.*

(i) With probability at least $1-\delta$ over the random draw of S:

$$\mathop{\mathbb{E}}_{g\sim\rho} d_\mathcal{F}(P^*, P^g) \leq \mathop{\mathbb{E}}_{g\sim\rho}\left[\mathcal{W}_\mathcal{F}\left(P_n^*, P^g\right)\right] + \frac{1}{\lambda}\left[\mathrm{KL}(\rho\,||\,\pi) + \log\frac{1}{\delta}\right] + \frac{\lambda K^2 d_\mathcal{Z}}{4n}. \tag{17}$$

(ii) With probability at least $1-\delta$ over the random draw of S and $g \sim \rho$:

$$d_\mathcal{F}(P^*, P^g) \leq \mathcal{W}_\mathcal{F}\left(P_n^*, P^g\right) + \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta}\right] + \frac{\lambda K^2 d_\mathcal{Z}}{4n}. \tag{18}$$

A.4. Proof of Theorem 3.5

The proof of Theorem 3.5 requires the following result.

Lemma A.7. *Let P, Q be probability measures on $\mathcal{X}$ and P_n, Q_n be the empirical distributions corresponding to the iid samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n \sim Q$ respectively. Then the empirical total variation distance has the bounded differences property with $c_i = \frac{2}{n}$, for all $1 \leq i \leq 2n$.*

Proof. We have

$$
\begin{aligned}
&\mathcal{D}_{\mathcal{F}}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) - \mathcal{D}_{\mathcal{F}}(\mathbf{x}_1, \ldots, \mathbf{x}'_n, \mathbf{y}_1, \ldots, \mathbf{y}_n) \\
&= \frac{1}{n}\left\{\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{n} f(\mathbf{x}_i) - \sum_{i=1}^{n} f(\mathbf{y}_i)\right] - \sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{n-1} f(\mathbf{x}_i) + f(\mathbf{x}'_n) - \sum_{i=1}^{n} f(\mathbf{y}_i)\right]\right\} \\
&\leq \frac{1}{n}\left\{\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{n} f(\mathbf{x}_i) - \sum_{i=1}^{n} f(\mathbf{y}_i) - \sum_{i=1}^{n-1} f(\mathbf{x}_i) - f(\mathbf{x}'_n) + \sum_{i=1}^{n} f(\mathbf{y}_i)\right]\right\} \\
&= \frac{1}{n}\sup_{f\in\mathcal{F}}\left[f(\mathbf{x}_n) - f(\mathbf{x}'_n)\right] \\
&\leq \frac{2}{n}.
\end{aligned}
$$

The last inequality follows from $-1 \leq f \leq 1$, for any $f \in \mathcal{F}$. □

The following result is similar to Corollaries 3.2 and A.6. It bounds on the distance between the full distributions.

Corollary A.8. *With the definitions and assumptions of Theorem 3.5, the following properties hold for any probability measure ρ such that $\rho \ll \pi$ and $\pi \ll \rho$.*

(i) With probability at least $1 - \delta$ over the random draw of S:

$$
\mathop{\mathbb{E}}_{g\sim\rho} d_{\mathcal{F}}(P^*, P^g) \leq \mathop{\mathbb{E}}_{g\sim\rho}\left[\mathcal{D}_{\mathcal{F}}(P_n^*, P^g)\right] + \frac{1}{\lambda}\left[\mathrm{KL}(\rho\,||\,\pi) + \log\frac{1}{\delta}\right] + \frac{4\lambda}{n}. \tag{19}
$$

(ii) With probability at least $1 - \delta$ over the random draw of S and $g \sim \rho$:

$$
d_{\mathcal{F}}(P^*, P^g) \leq \mathcal{D}_{\mathcal{F}}(P_n^*, P^g) + \frac{1}{\lambda}\left[\log\frac{d\rho}{d\pi}(g) + \log\frac{1}{\delta}\right] + \frac{4\lambda}{n}. \tag{20}
$$

B. Samples from the experiments

B.1. Synthetic datasets

We used two datasets: a Gaussian mixture with eight components arranged on a ring, and a Gaussian mixture with nine components on a grid. Figure 3 shows real samples from the actual training sets, and Figures 4 and 5 show samples from the trained models.

B.2. MNIST dataset

In our experiments with the MNIST dataset (Deng, 2012), we used the standard DCGAN architecture (Radford et al., 2016) for the generator and the critic. We experimented with different values for the hyperparameter σ_0 to train the probabilistic models. We computed the FID scores (Heusel et al., 2017) for the different models using 2000 random samples and the off-the-shelf implementation provided in the Pytorch-ignite library (Fomin et al., 2020), with a inception network (Szegedy et al., 2016) pre-trained on Imagenet. Since the Inception network requires 3-channel images, we transformed the original MNIST images by copying the single channel twice. The scores obtained for different models are displayed in Table B.2 and random (not cherry-picked) samples are displayed on Figures 6 to 11.

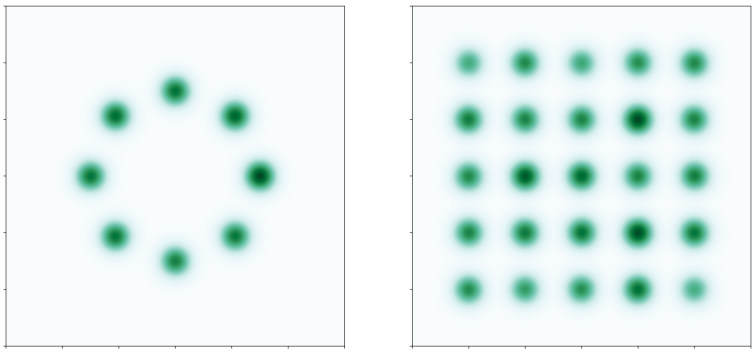

Figure 3. Real samples from the respective datasets. The image on the left represents samples from the Gaussian mixture with 8 components arranged on a ring, and the image on the right shows samples from the Gaussian mixture with 25 components arranged on a grid.

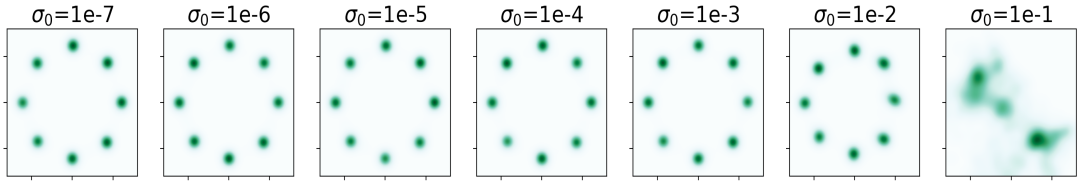

Figure 4. Samples from the models trained on the Gaussian ring

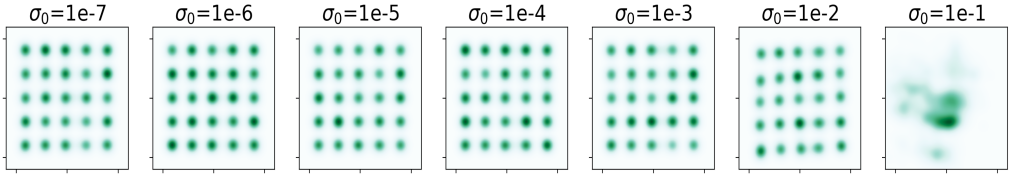

Figure 5. Samples from the models trained on the Gaussian grid

Table 1. FID scores from the various models trained on MNIST

σ_0	SCORE
10^{-7}	113.16
10^{-6}	106.59
10^{-5}	111.15
10^{-4}	107.54
10^{-3}	112.51
10^{-2}	189.30

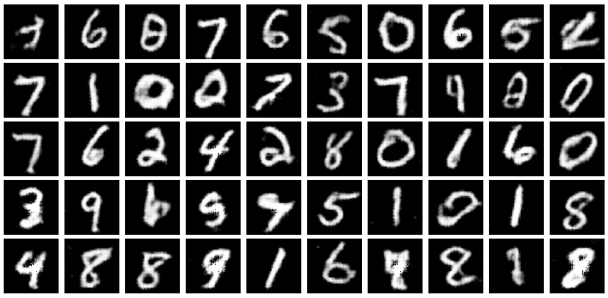

Figure 6. $\sigma_0 = 10^{-2}$

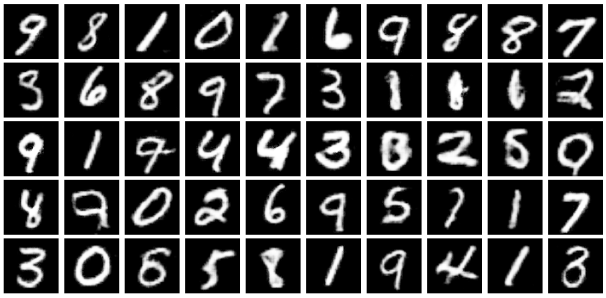

Figure 7. $\sigma_0 = 10^{-3}$

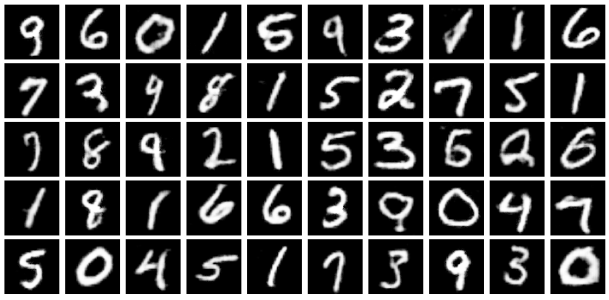

Figure 8. $\sigma_0 = 10^{-4}$

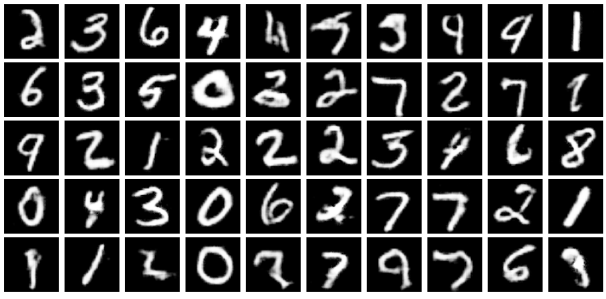

Figure 9. $\sigma_0 = 10^{-5}$

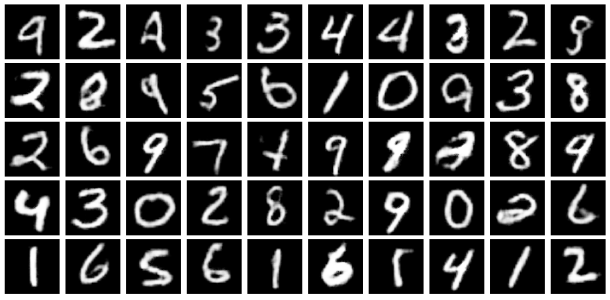

Figure 10. $\sigma_0 = 10^{-6}$

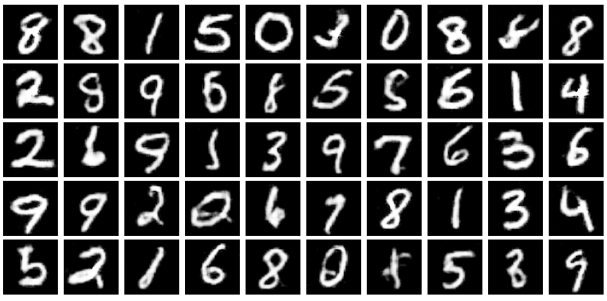

Figure 11. $\sigma_0 = 10^{-7}$