Improving Training of Likelihood-based Generative Models with Gaussian Homotopy

Ba-Hien Tran, Giulio Franzese, Pietro Michiardi, Maurizio Filippone

Improving Training of Likelihood-based Generative Models with Gaussian Homotopy

Ba-Hien Tran [1] **Giulio Franzese** [1] **Pietro Michiardi** [1] **Maurizio Filippone** [1]

Abstract

Generative Models (GMs) have recently gained popularity thanks to their success in various domains. In computer vision, for instance, they are able to generate astonishing realistic-looking images. Likelihood-based GMs are fast at generating new samples, given that they need a single model evaluation per sample, but their sample quality is usually lower than score-based Diffusion Models (DMs). In this work, we verify that the success of score-based DMs is in part due to the process of data smoothing, by incorporating this in the training of likelihood-based GMs. In the literature of optimization, this process of data smoothing is referred to as Gaussian homotopy (GH), and it has strong theoretical grounding. Crucially, GH does not incur computational overheads, and it can be implemented by adding one line of code in any training loop. We report results on various GMs, including Variational Autoencoders and Normalizing Flows, applied to image datasets demonstrating that GH enables significant improvements in sample quality.

1. Introduction

Generative Models (GMs) have recently attracted considerable attention due to their tremendous success in various domains. Given a set of data points, GMs attempt to characterize their distribution so that it is then possible to draw new samples from the estimated distribution. Popular approaches include Variational Autoencoders (VAEs) (Kingma & Welling, 2014), Normalizing Flows (NFs) (Rezende & Mohamed, 2015), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and score-based Diffusion Models (DMs) (Ho et al., 2020; Song et al., 2021).

Although various approaches in GMs exhibit differences in optimization strategies and formulations, their underlying objectives share similarities, as they are related to some form of regularized optimal transport problem (Genevay et al., 2017; Onken et al., 2021; Chen et al., 2022). However, these different formulations give rise to diverse properties associated with GMs, and the advantages and disadvantages of each formulation can be understood through the concept of the GM tri-lemma (Xiao et al., 2022). The GM tri-lemma posits three desirable properties: (i) high sample quality, (ii) mode coverage, and (iii) fast sampling, and it has been argued that achieving all three simultaneously is challenging (Xiao et al., 2022).

Score-based DMs are currently dominating the state-of-the-art, offering high sample quality and good mode coverage. However, their formulation based on stochastic differential equations makes it computationally expensive to generate new samples. Likelihood-based GMs provide a complementary approach with lower sample quality and diversity but fast sampling, requiring only one model evaluation per sample. Recognizing the shared objective of all GMs, this paper aims to leverage the strengths of score-based DMs to enhance likelihood-based GMs without paying the price of costly sample generation.

One of the distinctive elements of score-based DMs is data smoothing, achieved through the pertubation of the data by Gaussian noise. In the optimization literature, adding noise to the data and reducing its level through iterations is also known as Gaussian homotopy (GH) (or continuation optimization), and it has the effect of annealing the smoothness of the loss landscape throughout optimization. GH has been shown to accelerate optimization, particularly for stochastic non-convex problems (Hazan et al., 2016). In the context of GMs, we view GH as a means to counter issues related to *manifold overfitting* (Loaiza-Ganem et al., 2022a). This problem occurs when data satisfies the so-called *manifold hypothesis* (Roweis & Saul, 2000), whereby data lies on a low-dimensional manifold of the input space, which is typically the case, for instance, for images. In this case, density estimation is problematic due to likelihood being infinite for any density with support on the data manifold (Loaiza-Ganem et al., 2022a). In this work, we show that GH guides the training procedure of likelihood-based GMs

[1]Department of Data Science, EURECOM, France. Correspondence to: Ba-Hien Tran <ba-hien.tran@eurecom.fr>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling

in a way which allows them to bypass the issues associated with manifold overfitting. We focus on VAEs and NFs, and provide experimental evidence on synthetic and real-world image data that GH consistently improves sample quality. Crucially, this strategy is extremely easy to implement, as it requires adding very little code to any existing training loop.

2. Related Work

Our work is positioned within the context of improving GMs through the addition of noise to the data. One popular approach is denoising autoencoders (Vincent et al., 2008), which reconstruct clean data from noisy samples. Recently, Meng et al. (2021) introduced a two-step approach to improve autoregressive generative models, where a smoothed version of the data is first modeled by adding a fixed level of noise, and then the original data distribution is recovered through an autoregressive denoising model. In a similar vein, Loaiza-Ganem et al. (2022b) recently attempted to use Tweedie's formula (Robbins, 1956) as a denosing step, but found that it does not improve the performance of NFs and VAEs. Our work is distinct from these approaches in that GH guides the estimated distribution towards the true data distribution in a progressive manner by means of annealing instead of fixing a noise level. Moreover, our approach does not require explicit denoising steps and can be readily applied to the optimization of any likelihood-based GMs without modifications.

3. Gaussian Homotopy for Generative Models

Given a dataset $\mathcal{D}$ consisting of N i.i.d samples $\mathcal{D} \triangleq \{\mathbf{x}_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$, we aim to estimate the unknown continuous generating distribution $p_{\text{data}}(\mathbf{x})$. In order to do so, we introduce a model $p_{\boldsymbol{\theta}}(\mathbf{x})$ with parameters $\boldsymbol{\theta}$ and attempt to estimate $\boldsymbol{\theta}$ based on the dataset $\mathcal{D}$. A common approach to estimate $\boldsymbol{\theta}$ is to maximize the likelihood of the data, which is equivalent to the following objective:

$$\mathcal{L}(\boldsymbol{\theta}) \triangleq -\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right]. \tag{1}$$

There are several approaches to parameterize the generative model $p_{\boldsymbol{\theta}}(\mathbf{x})$. In this work, we focus on two widely used likelihood-based GMs, which are NFs (Rezende & Mohamed, 2015) and VAEs (Kingma & Welling, 2014).

We propose a simple yet effective approach to improve likelihood-based GMs. Our method involves adding Gaussian noise to the data throughout training and gradually reducing its variance until recovering the original data. This procedure, which in the literature of optimization is referred to as Gaussian homotopy (GH), bears some similarity to the reverse process of score-based DMs, where a prior noise distribution is smoothly transformed into the data distribution (Song & Ermon, 2019; Song et al., 2021). However, we note that in score-based DMs the score network learns a model capable of handling all levels of noise, whereas in our case the smoothing process is in "one direction" only, from noise to data.

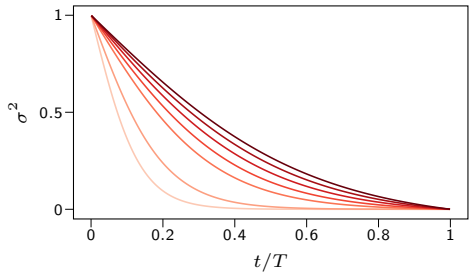

Figure 1: Illustration of sigmoid schedule (Jabri et al., 2022) with different temperatures. The temperature values from 0.2 to 0.9 are progressively shaded, with the lighter shade corresponding to lower temperatures.

Gaussian Homotopy. Starting from the target objective function, which in our case is $\mathcal{L}(\boldsymbol{\theta})$ in Eq. 1 (or a lower bound in the case of VAEs), GH constructs a family of functions $\mathcal{H}(\boldsymbol{\theta}, \gamma)$ parameterized by an auxiliary variable $\gamma \in [0, 1]$ so that $\mathcal{H}(\boldsymbol{\theta}, 0) = \mathcal{L}(\boldsymbol{\theta})$. The objective functions $\mathcal{H}(\boldsymbol{\theta}, \gamma)$ are defined so that their smoothness increases with γ, and the idea is to cast optimization of $\mathcal{L}(\boldsymbol{\theta})$ as a sequence of optimization problems involving $\mathcal{H}(\boldsymbol{\theta}, \gamma)$ with γ going from 1 to 0 with a given annealing schedule.

We implement $\mathcal{H}(\boldsymbol{\theta}, \gamma)$ with a simple transformation of the data involving the addition of Gaussian noise and rescaling in a *variance preserving* fashion (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kingma et al., 2021). Denoting by T the maximum number of iterations for which we train our model $p_{\boldsymbol{\theta}}(\mathbf{x})$, we can create a sequence of progressively less smoothed versions of the original data $\mathbf{x}$, which we refer to as $\tilde{\mathbf{x}}_t$. Here, t ranges from $t = 0$ (the most smoothed) to $t = T$ (the least smoothed). For any $t \in [0, T]$, the distribution of $\tilde{\mathbf{x}}_t$, conditioned on $\mathbf{x}$, is given as follows:

$$q(\tilde{\mathbf{x}}_t \,|\, \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}_t; \alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I}), \tag{2}$$

where $\alpha_t = \sqrt{1 - \sigma_t^2}$ and $\sigma_t^2 = \gamma(t/T)$, with $\gamma(\cdot)$ monotonically decreasing from 1 to 0 controlling the rate of smoothing. We employ a sigmoid schedule (Jabri et al., 2022, illustrated in Fig. 1) for $\gamma(\cdot)$, which has recently been

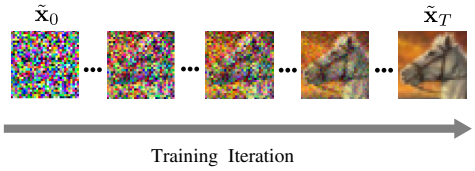

Figure 2: Illustration of Gaussian homotopy (GH).

Algorithm 1: Gaussian Homotopy

for $t \leftarrow 1, 2, ..., T$ **do**
  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ // Sample training data
  $\tilde{\mathbf{x}}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\varepsilon}$ // Smooth data with $\alpha_t, \sigma_t^2 \leftarrow \gamma(t/T)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\boldsymbol{\theta}_t \leftarrow \text{UPDATE}(\boldsymbol{\theta}_{t-1}, \tilde{\mathbf{x}}_t)$ // Train model

shown to be more effective in practice compared to other choices such as linear (Ho et al., 2020) or cosine schedules (Nichol & Dhariwal, 2021) for score-based DMs. The process of GH is illustrated in Fig. 2.

Intuitively, our procedure involves gradually transforming a Gaussian distribution with an identity covariance matrix into the distribution of the data. Algorithm 1 summarizes the proposed GH procedure, where the red line indicates a simple additional step required to include this perturbation compared with vanilla training.

4. GH Mitigates Challenges with Generative Modeling under the Manifold Hypothesis

In this section, we provide some insights as to why GH improves training of likelihood-based GMs under the manifold hypothesis.

4.1. The Manifold Hypothesis and Density Estimation in Low-Density Regions

The manifold hypothesis is a fundamental concept in manifold learning (Roweis & Saul, 2000; Tenenbaum et al., 2000; Bengio et al., 2012) stating that real-world high-dimensional data tend to lie on a manifold $\mathcal{M}$ characterized by a much lower dimensionality compared to the one of the input space (ambient dimensionality) (Narayanan & Mitter, 2010). This has been verified theoretically and empirically for many applications and datasets (Ozakin & Gray, 2009; Narayanan & Mitter, 2010; Pope et al., 2021; Tempczyk et al., 2022). For example, (Pope et al., 2021) report extensive evidence that natural image datasets have indeed very low intrinsic dimension relative to the high number of pixels in the images.

Under the manifold hypothesis, density estimation in the input space is challenging and ill-posed, with high-density regions on the manifold and nearly zero-density regions outside it (Meng et al., 2021). This implies a need for high Lipschitz constants in the target density. Scarce data in low-density regions hinders accurate density estimation in the tails, presenting significant challenges for training GMs (Cornish et al., 2020; Meng et al., 2021; Song & Ermon, 2019). Recently, score-based DMs have shown promise by gradually transforming a Gaussian distribution to the data distribution, suggesting that the mechanism associate with smoothing the data contributes to superior density estimation in low-density regions.

To demonstrate the challenges associated with accurate estimation in low-density regions, we consider a toy experiment where we use a REAL-NVP flow (Dinh et al., 2017) to model a two-dimensional mixture of Gaussians, which is a difficult test for NFs in general. Fig. 3 presents the true and estimated distributions, along with their corresponding scores; note that in the literature of GMs, the score refers to the gradient of log-density with respect to the input and not the parameters as in the Statistics literature (Hyvärinen, 2005). In regions of low data density, $p_{\boldsymbol{\theta}}(\mathbf{x})$ fails to accurately model the true density and scores, primarily due to the scarcity of data samples in these regions. This may be more problematic under the manifold hypothesis and for high-dimensional data such as images.

Conversely, the proposed addition of GH in the training process improves density estimation. Initially, the model has to deal with a simple coarse-grained version of the target density, which spans the entire support of the data, as shown in the top row of Figure 3. The low training loss in Figure 4 supports this observation. Subsequently, the method gradually reduces the level of noise allowing for a progressive refinement of the estimated versions of the target density. Each level of Gaussian noise guides the optimization process for the next, leading to the recovery of modes and effective density estimation in low-density regions. In contrast, the vanilla training procedure produces a poor estimate of the target density, which is evident from the trace-plot of the Maximum Mean Discrepancy (MMD) metric in Figure 4 and the visualization of the scores in Figure 3.

4.2. Manifold Overfitting

The manifold hypothesis suggests that overfitting on a manifold can occur when the model assigns an arbitrarily large likelihood in the vicinity of the manifold, even if it does not accurately capture the true distribution (Dai & Wipf, 2019; Loaiza-Ganem et al., 2022a). This issue is illustrated in Fig. 2 of Loaiza-Ganem et al. (2022a) and in Fig. 5 here, where we consider a von Mises distribution on the unit circle. In this experiment, the true data distribution is supported on a one-dimensional curve manifold in a two-dimensional space. Despite poor approximation of the true distribution, the model may achieve high likelihood by concentrating its density around the correct manifold.

In this work, we rely on the theoretical grounding of manifold overfitting established in Loaiza-Ganem et al. (2022a). In their work, the problem of manifold overfitting is formalized in Theorem 1. Their key message is that, a-priori, there is no reason to expect a likelihood-based model to converge to p_{data} out of all the possible $p^\dagger$ defined on the

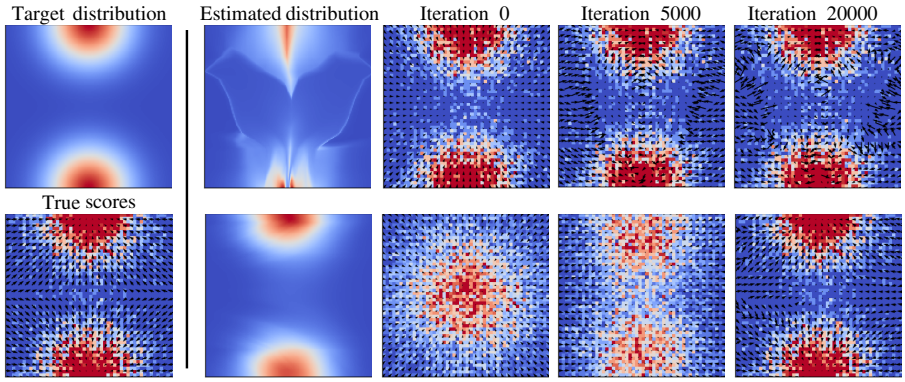

Figure 3: The first column shows the target distribution and the true scores. The second column depicts the estimated distributions of the GMM .The remaining columns show histogram of samples from the true (**top row**) and smoothed data (**bottom row**), and estimated scores.

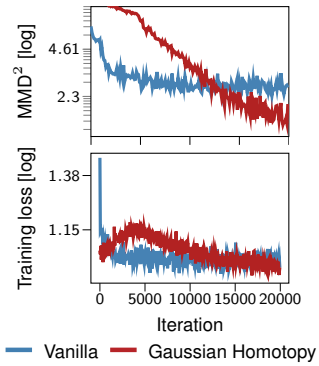

Figure 4: The learning curves of the GMM experiments.

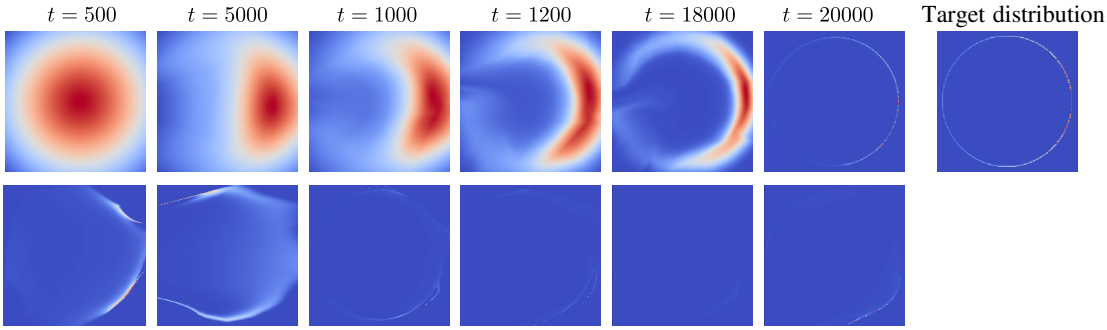

Figure 5: Progression of estimated densities for the von Mises distribution from the vanilla (**bottom**) and our GH (**top**) approaches.

manifold. Given any smooth probability measure $\mathbb{P}^\dagger$ defined on a manifold, their theorem claims existence of a sequence of measures converging weakly to $\mathbb{P}^\dagger$. The proof constructs such a sequence by convolving $\mathbb{P}^\dagger$ with a Gaussian kernel with progressively lower variance. Our GH approach relies on the same idea and enjoys the same theoretical property as the measure with associated p_{data} is included in the class of measures $\mathbb{P}^\dagger$ in their theorem. Intuitively, we can easily explain this as a successful technique to avoid manifold overfitting as follows: at iteration $t = 0$, we start with a target distribution obtained by convolving the desired data distribution p_{data} with a Gaussian kernel of large but finite variance $\sigma^2(0)$. Optimization is performed targeting this distribution, without experiencing manifold overfitting due to the non-degenerate dimensionality of the corrupted data. Subsequently, we iteratively reduce the variance of the Gaussian kernel. By iteratively repeating this procedure, we can reach the point where we are matching a distribution convolved with a Gaussian kernel with an arbitrarily small variance $\sigma^2(t)$, without ever experiencing manifold overfitting. This is demonstrated in the bottom row of Fig. 5, where GH guides the estimated density towards the target. Additionally, GH enables the estimated model not only to accurately learn the manifold but also to accurately capture the shape of the target density.

5. Experiments on Imaging Datasets

We evaluate our method on image generation tasks on CIFAR10 (Krizhevsky & Hinton, 2009) and CELEBA 64^2 (Liu et al., 2015) datasets, using a diverse set of likelihood-based GMs. We found that that further training the model on the original data after applying GH leads to better performance. Hence, in our approach we apply GH during the first half of the optimization phase, and we continue optimize the model using the original data in the second half. Nevertheless, to ensure a fair comparison, we adopt identical settings for the vanilla and for the proposed approach, including random seed, optimizer, and the total number of iterations.

It is worth noting that we did not experience issues with manifold overfitting when switching off GH in the second half of the optimization phase. We attribute this to a combination of factors including model capacity and stochastic optimization which prevents the models to assign zero density outside the data manifold. We will investigate this in greater detail in followup works; for now, we observe that in a typical stochastic optimization setting, GH has the effect of providing an effective mechanism to guide optimization.

We evaluate the quality of the generated images using the popular Fréchet Inception Distance (FID) score (Heusel

Table 1: Comparisions of FID score between vanilla and GH training on CIFAR10 and CELEBA dataset (*lower is better*).

Model	CIFAR10		CELEBA	
	Vanilla	GH	Vanilla	GH
REAL-NVP (Dinh et al., 2017)	131.15	121.75	81.25	79.68
GLOW (Kingma & Dhariwal, 2018)	74.62	64.87	97.59	70.91
VAE (Kingma & Welling, 2014)	191.98	155.13	80.19	72.97
VAE-IAF (Kingma et al., 2016)	193.58	156.39	80.34	73.56
IWAE (Burda et al., 2015)	183.04	146.70	78.25	71.38
β-VAE (Higgins et al., 2017)	112.42	93.90	67.78	64.59
HVAE (Caterini et al., 2018)	172.47	137.84	74.10	72.28

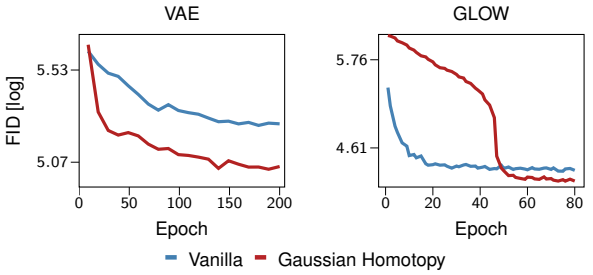

Figure 6: The progression of FID on CIFAR10 dataset.

et al., 2017). The results, reported in Table 1, indicate that the proposed GH strategy enables consistent improvements in performance compared to vanilla training, and this is consistent across all datasets and models. Furthermore, we observe that GH leads to faster convergence of the FID score for VAE-based models, as shown in Fig. 6.

6. Conclusion

In this work, we explored the impact of data smoothing on the performance of likelihood-based GMs, specifically focusing on NFs and VAEs. Data smoothing, implemented through Gaussian homotopy, is a well-known technique to improve optimization, it is easy to implement and it offers nice theoretical guarantees. We applied this idea to challenging generative modeling tasks involving imaging data and relatively large-scale architectures as a means to demonstrate systematic gains in performance in various conditions and input dimensions. Although we have not achieved competitive FID scores compared to score-based DMs, we believe that this work will serve as a basis for future research on performance enhancements in state-of-the-art models that combine DMs and likelihood-based GMs, and in alternative forms of data smooting to improve optimization of state-of-the-art GMs.

Acknowledgements

MF gratefully acknowledges support from the AXA Research Fund and the Agence Nationale de la Recherche (grant ANR-18-CE46-0002 and ANR-19-P3IA-0002).

References

Bengio, Y., Courville, A. C., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35: 1798–1828, 2012.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance Weighted Autoencoders. In *International Conference on Learning Representations*, 2015.

Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian Variational Auto-Encoder. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Chen, T., Liu, G.-H., and Theodorou, E. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations*, 2022.

Cornish, R., Caterini, A., Deligiannidis, G., and Doucet, A. Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2133–2143. PMLR, 13–18 Jul 2020.

Dai, B. and Wipf, D. Diagnosing and Enhancing VAE Models. In *International Conference on Learning Representations*, 2019.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density Estimation Using Real NVP. In *International Conference on Learning Representations*, 2017.

Genevay, A., Peyré, G., and Cuturi, M. GAN and VAE from an Optimal Transport Point of View. *arXiv preprint arXiv:1706.01807*, 2017.

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. Curran Associates, Inc., 2014.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. On Graduated Optimization for Stochastic Non-Convex Problems. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1833–1841, New York, New York, USA, 20–22 Jun 2016. PMLR.

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. Curran Associates, Inc., 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Hyvärinen, A. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Jabri, A., Fleet, D., and Chen, T. Scalable Adaptive Computation for Iterative Generation. *arXiv preprint arXiv:2212.11972*, 2022.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707. Curran Associates, Inc., 2021.

Kingma, D. P. and Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved Variational Inference with Inverse Autoregressive Flow. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Krizhevsky, A. and Hinton, G. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425.

Loaiza-Ganem, G., Ross, B. L., Cresswell, J. C., and Caterini, A. L. Diagnosing and Fixing Manifold Overfitting in Deep Generative Models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856.

Loaiza-Ganem, G., Ross, B. L., Wu, L., Cunningham, J. P., Cresswell, J. C., and Caterini, A. L. Denoising Deep Generative Models. In *I Can't Believe It's Not Better Workshop at NeurIPS 2022*, 2022b.

Meng, C., Song, J., Song, Y., Zhao, S., and Ermon, S. Improved Autoregressive Modeling with Distribution Smoothing. In *International Conference on Learning Representations*, 2021.

Narayanan, H. and Mitter, S. Sample Complexity of Testing the Manifold Hypothesis. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Nichol, A. Q. and Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. OT-Flow: Fast and Accurate Continuous Normalizing Flows via Optimal Transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

Ozakin, A. and Gray, A. Submanifold Density Estimation. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

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.

Rezende, D. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

Robbins, H. An Empirical Bayes Approach to Statistics. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1956*, volume 1, pp. 157–163, 1956.

Roweis, S. T. and Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290 (5500):2323–2326, 2000.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Tempczyk, P., Michaluk, R., Garncarek, L., Spurek, P., Tabor, J., and Golinski, A. LIDL: Local Intrinsic Dimension Estimation Using Approximate Likelihood. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21205–21231. PMLR, 17–23 Jul 2022.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In *International Conference on Learning Representations*, 2022.