DEEP MMD GRADIENT FLOW WITHOUT ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a gradient flow procedure for generative modeling by transporting particles from an initial source distribution to a target distribution, where the gradient field on the particles is given by a noise-adaptive Wasserstein Gradient of the Maximum Mean Discrepancy (MMD). The noise-adaptive MMD is trained on data distributions corrupted by increasing levels of noise, obtained via a forward diffusion process, as commonly used in denoising diffusion probabilistic models. The result is a generalization of MMD Gradient Flow, which we call Diffusion-MMD-Gradient Flow or DMMD. The divergence training procedure is related to discriminator training in Generative Adversarial Networks (GAN), but does not require adversarial training. We obtain competitive empirical performance in unconditional image generation on CIFAR10, MNIST, CELEB-A (64 x64) and LSUN Church (64 x 64). Furthermore, we demonstrate the validity of the approach when MMD is replaced by a lower bound on the KL divergence.

1 INTRODUCTION

In recent years, generative models have achieved impressive capabilities on image Saharia et al. (2022), audio Le et al. (2023) and video generation Ho et al. (2022) tasks but also protein modeling Watson et al. (2022) and 3d generation Poole et al. (2022). Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Rombach et al., 2022) underpin these new methods. In these models, we learn a backward denoising diffusion process via denoising score matching (Hyvärinen, 2005; Vincent, 2011). This backward process corresponds to the time-reversal of a forward noising process. At sampling time, starting from random Gaussian noise, diffusion models produce samples by discretizing the backward process.

One challenge that arises when applying these models in practice is that the Stein score (that is, the gradient log of the current noisy density) becomes ill-behaved near the data distribution (Yang et al., 2023): the diffusion process needs to be slowed down at this point, which incurs a large number of sampling steps near the data distribution. Indeed, if the manifold hypothesis holds Tenenbaum et al. (2000); Fefferman et al. (2016); Brown et al. (2022) and the data is supported on a lower dimensional space, it is expected that the score will explode for noise levels close to zero, to ensure that the backward process concentrates on this lower dimensional manifold Bortoli (2023); Pidstrigach (2022); Chen et al. (2022). While strategies exist to mitigate these issues, they trade-off the quality of the output against inference speed, see for instance (Song et al., 2023; Xu et al., 2023; Sauer et al., 2023).

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) represent an alternative popular generative modelling framework (Brock et al., 2019; Karras et al., 2020a). Candidate samples are produced by a *generator*: a neural net mapping low dimensional noise to high dimensional images. The generator is trained in alternation with a *discriminator*, which is a measure of discrepancy between the generator and target images. An advantage of GANs is that image generation is fast once the GAN is trained (Xiao et al., 2022), although image samples are of lower quality than for the best diffusion models (Ho et al., 2020; Rombach et al., 2022). When learning a GAN model, the main challenge arises due to the presence of the generator, which must be trained adversarially alongside the discriminator. This requires careful hyperparameter tuning (Brock et al., 2019; Karras et al., 2020b; Liu et al., 2020), without which GANs may suffer from training instability and mode collapse (Arora et al., 2017; Kodali et al., 2017; Salimans et al., 2016).

Nonetheless, the process of GAN design has given rise to a strong understanding of discriminator functions, and a wide variety of different divergence measures have been applied. These fall broadly into two categories: the integral probability metrics (among which, the Wasserstein distance (Arjovsky et al., 2017; Gulrajani et al., 2017; Genevay et al., 2018) and the Maximum Mean Discrepancy (Li et al., 2017; Bińkowski et al., 2021; Arbel et al., 2018)) and the f-divergences (Goodfellow et al., 2014; Nowozin et al., 2016; Mescheder et al., 2018; Brock et al., 2019). While it would appear that f-divergences ought to suffer from the same shortcomings as diffusions when the target distribution is supported on a submanifold Arjovsky et al. (2017), the divergences used in GANs are in practice variational lower bounds on their corresponding f-divergences (Nowozin et al., 2016), and in fact behave closer to IPMs in that they do not require overlapping support of the target and generator samples, and can metrize weak convergence (Arbel et al., 2021, Proposition 14) and (Zhang et al., 2018) (there remain important differences, however: notably, f-divergences and their variational lower bounds need not be symmetric in their arguments).

A natural question then arises: is it possible to define a Wasserstein gradient flow (Ambrosio et al., 2008; Santambrogio, 2015) using a GAN discriminator as a divergence measure? In this setting, the divergence (discriminator) provides a gradient field directly onto a set of particles (rather than to a generator), transporting them to the target distribution. Contributions in this direction include the MMD flow Arbel et al. (2019); Hertrich et al. (2023), which defines a Wasserstein Gradient Flow on the Maximum Mean Discrepancy (Gretton et al., 2012); and the KALE (KL approximate lower-bound estimator) flow Glaser et al. (2021), which defines a Wasserstein gradient flow on a KL lower bound of the kind used as a GAN discriminator based on an f-divergence (Nowozin et al., 2016). We describe the MMD and its corresponding Wasserstein gradient flow in Section 2. These approaches employ fixed function classes (namely, reproducing kernel Hilbert spaces) for the divergence, and are thus not suited to high dimensional settings such as images. Moreover, we show in this work that even for simple examples in low dimensions, an adaptive discriminator ensures faster convergence of a source distribution to the target, see Section 3.

A number of more recent approaches employ trained neural net features in divergences for a subsequent gradient flow (e.g. Fan et al., 2022; Franceschi et al., 2023). Broadly speaking, these works used adversarial means to train a *series* of discriminator functions, which are then applied in sequence to a population of particles. While more successful on images than kernel divergences, the approaches retain two shortcomings: they still require adversarial training (on their own prior output), with all the challenges that this entails; and their empirical performance falls short in comparison with modern diffusions and GANs (see related work in Section 6 for details).

In the present work, we propose a novel Wasserstein Gradient flow on a noise-adaptive MMD divergence measure, leveraging insights from both GANs and diffusion models. To *train the discriminator*, we start with clean data, and use a forward diffusion process from (Ho et al., 2020) to produce noisy versions of the data with given levels of noise (data with high levels of noise are analogous to the output of a poorly trained generator, whereas low noise is analogous to a well trained generator). The added noise is always Gaussian. For a given level of noise, we train a noise conditional MMD discriminator to distinguish between the clean and the noisy data, using a single network across all noise levels. This allows us to have better control over the discriminator training procedure than would be achievable with a GAN generator at different levels of refinement, where this control is implicit and hard to characterize.

To *draw new samples*, we propose a novel noise-adaptive version of MMD gradient flow (Arbel et al., 2019). Starting from Gaussian distribution, we move them in the direction of the target distribution by following MMD Gradient flow (Arbel et al., 2019), adapting our MMD discriminator to the corresponding level of noise. See Section 4 for details. This allows us to have a fine grained control over the sampling process. As a final challenge, MMD gradient flows have previously required large populations of interacting particles for the generation of novel samples, which is expensive (quadratic in the number of particles) and impractical. In Section 5, we propose a scalable approximate sampling procedure for a case of a linear base kernel, which allows *single* samples to be generated with a very little loss in quality, at cost independent of the number of particles used in training. The MMD is an instance of an integral probability metric, however many GANs have been designed using discriminators derived from f-divergences. Section D demonstrates how our approach can be applied to such divergences, using a lower bound on the KL divergence as an illustration. Section 6 contains a review of alternative approaches to using GAN discriminators for sample generation. Finally, in Section 7, we show that our method, Diffusion-MMD-gradient flow (DMMD), yields competitive performance in generative modeling on 2-D datasets as well as in unconditional image generation on CIFAR10 (Krizhevsky et al., 2009), MNIST, CELEB-A, LSUN Church.

2 BACKGROUND

In this section, we define the MMD as a GAN discriminator, then describe Wasserstein gradient flow as it applies for this divergence measure.

MMD GAN. Let $\mathcal{X} \subset \mathbb{R}^D$ and $\mathcal{P}(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$. Let $P \in \mathcal{P}(\mathcal{X})$ be the *target* (data) distribution and $Q_\psi \in \mathcal{P}(\mathcal{X})$ be a distribution associated with a *generator* parameterized by $\psi \in \mathbb{R}^L$. Let $\mathcal{H}$ be Reproducing Kernel Hilbert Space (RKHS), see (Schölkopf & Smola, 2018) for details, for some kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) between Q_ψ and P is defined as $\mathrm{MMD}(Q_\psi, P) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \{\mathbb{E}_{Q_\psi}[f(X)] - \mathbb{E}_P[f(X)]\}$. We refer to the function $f_{Q_\psi,P}$ that attains the supremum as the *witness function*,

$$f_{Q_\psi,P}(z) \propto \int k(x,z)\mathrm{d}Q_\psi(x) - \int k(y,z)\mathrm{d}P(y), \tag{1}$$

which will be essential in defining our gradient flow. Given $X^N = \{x_i\}_{i=1}^N \sim Q_\psi^{\otimes N}$ and $Y^M = \{y_i\}_{i=1}^M \sim P^{\otimes M}$, the empirical witness function is known in closed form, $\hat{f}_{Q_\psi,P}(x) \propto \frac{1}{N}\sum_{i=1}^N k(x_i, x) - \frac{1}{M}\sum_{j=1}^M k(y_j, x)$, and an unbiased estimate of MMD^2 (Gretton et al., 2012) is likewise straightforward. In the MMD GAN (Bińkowski et al., 2021; Li et al., 2017), the kernel is

$$k(x, y) = k_{\text{base}}(\phi(x; \theta), \phi(y; \theta)), \tag{2}$$

where k_{base} is a base kernel and $\phi(\cdot; \theta) : \mathcal{X} \to \mathbb{R}^K$ are neural networks *discriminator* features with parameters $\theta \in \mathbb{R}^H$. We use the modified notation $\mathrm{MMD}_u^2[X^N, Y^M; \theta]$ to highlight the functional dependence on the discriminator parameters. The MMD is an Integral Probability Metric (IPM) (Muller, 1997), and thus well defined on distributions with disjoint support: this argument was made in favor of IPMs by Arjovsky et al. (2017). Note further that the Wasserstein GAN discriminators of Arjovsky et al. (2017); Gulrajani et al. (2017) can be understood in the MMD framework, when the base kernel is linear. Indeed, it was observed by Genevay et al. (2018) that requiring closer approximation to a true Wasserstein distance resulted in decreased performance in GAN image generation, likely due to the the exponential dependence of sample complexity on dimension for the exact computation of the Wasserstein distance; this motivates an interpretation of these discriminators simply as IPMs using a class of linear functions of learned features. We further note that the variational lower bounds used in approximating f-divergences for GANs share the property of being well defined on distribtions with disjoint support Nowozin et al. (2016); Arbel et al. (2021), although they need not be symmetric in their arguments. Finally, while Q_ψ and θ are trained adversarially in GANs, our setting will only require us to learn the discriminator parameter θ.

Wasserstein gradient flows. Instead of a GAN generator, we can move a sample of particles along the Wasserstein Gradient flow associated with the discriminator (Ambrosio et al., 2008). Let $\mathcal{P}_2(\mathcal{X})$ be a set of probability distributions on $\mathcal{X}$ with a finite second moment equipped with the 2-Wasserstein distance. Let $\mathcal{F}(\nu) : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$ be a functional defined over $\mathcal{P}_2(\mathcal{X})$ with a property that $\arg\inf_\nu \mathcal{F}(\nu) = P$. We consider the problem of transporting mass from an initial distribution $\nu_0 = Q$ to a target distribution $\mu = P$, finding a continuous path $(\nu_t)_{t\geq 0}$ starting from ν_0 that converges to μ. This problem is studied in Optimal Transport theory (Villani, 2008; Santambrogio, 2015). This path can be discretized as a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ such that $X_n \sim \nu_n$,

$$X_{n+1} = X_n - \gamma \nabla \mathcal{F}'(\nu_n)(X_n), \quad X_0 \sim Q, \tag{3}$$

where $\eta > 0$ and $\mathcal{F}'(\nu_n)(X_n)$ is the first variation of $\mathcal{F}$ associated with the Wasserstein gradient, see (Ambrosio et al., 2008; Arbel et al., 2019) for precise definitions. As $n \to \infty$ and $\gamma \to 0$, depending on the conditions on $\mathcal{F}$, the process (3) will convergence to the gradient flow as a continuous time limit (Ambrosio et al., 2008).

MMD gradient flow. For a choice $\mathcal{F}(\nu) = \mathrm{MMD}^2[\nu, P]$ and a fixed kernel, conditions for convergence of the process in (3) to P are given by Arbel et al. (2019). Moreover, the first variation of $\mathcal{F}'(\nu) = f_{\nu,P} \in \mathcal{H}$ is the witness function defined earlier.[1] Using (1)-(3), the discretized MMD gradient flow for any $n \in \mathbb{N}$ is given by

$$X_{n+1} = X_n - \gamma \nabla f_{\nu_n,P}(X_n), \qquad X_0 \sim Q. \tag{4}$$

This provides an algorithm to (approximately) sample from the target distribution P. We remark that Arbel et al. (2019); Hertrich et al. (2023) used a kernel with fixed hyperparameters. In the next section, we will argue that even for RBF kernels (where only the bandwidth is chosen), faster convergence will be attained using kernels that adapt during the gradient flow. Details of kernel choice for alternative approaches are given in related work (Section 6).

3 A MOTIVATION FOR ADAPTIVE KERNELS

In this section, we demonstrate the benefit of using an *adaptive* kernel when performing MMD gradient flow. We show that even in the simple setting of Gaussian sources and targets, an adaptive kernel improves the convergence of the flow. Let $k_\alpha(x, y) = \alpha^{-d} \exp[-\|x - y\|^2/(2\alpha^2)]$ be the normalized Gaussian kernel. For any $\mu \in \mathbb{R}^d$ and $\sigma > 0$ we denote by $\pi_{\mu,\sigma}$ the Gaussian distribution with mean μ and covariance matrix $\sigma^2\mathrm{Id}$. We denote MMD_α the MMD associated with k_α.

Proposition 3.1. *For any $\mu_0 \in \mathbb{R}^d$ and $\sigma > 0$, let $\alpha^\star$ be given by*

$$\alpha^\star = \mathrm{argmax}_{\alpha \geq 0} \|\nabla_{\mu_0} \mathrm{MMD}^2_\alpha(\pi_{0,\sigma}, \pi_{\mu_0,\sigma})\|.$$

Then, we have that

$$\alpha^\star = \mathrm{ReLU}(\|\mu_0\|^2/(d + 2) - 2\sigma^2)^{1/2}. \tag{5}$$

The result is proved in Appendix H. The quantity $\|\nabla_{\mu_0} \mathrm{MMD}^2_\alpha(\pi_{0,\sigma}, \pi_{\mu_0,\sigma})\|$ represents how much the mean of the Gaussian $\pi_{\mu_0,\sigma}$ is displaced by a flow w.r.t. MMD^2_α. We want $\|\nabla_{\mu_0} \mathrm{MMD}^2_\alpha(\pi_{0,\sigma}, \pi_{\mu_0,\sigma})\|$ as large as possible as it denotes the *maximum displacement possible*.

We show that $\alpha^\star$ maximizing this displacement is given by (5). It is notable that assuming that when $\sigma > 0$ is fixed, this quantity depends on $\|\mu_0\|$, i.e. the distance between the two distributions. This observation justifies our approach of following an *adaptive* MMD flow at inference time. We further highlight the phase transition behaviour of Proposition 3.1: once the Gaussians are sufficiently close, the optimal kernel width is zero (note that this phase transition would not be observed in the simpler Dirac GAN example of Mescheder et al. (2018), where the source and target distributions are Dirac masses with no variance). This phase transition suggests that the flow associated with MMD benefits *less* from adaptivity as the supports of the distributions overlap. We exploit this observation by introducing an optional denoising stage to our procedure; see the end of Section 4.

In practice, it is not desirable to approximate the distributions of interest by Gaussians, and richer neural network kernel features $\phi(x; \theta)$ are used (see Section 7). Approaches to optimize the MMD parameters for GAN training are described by Arbel et al. (2018), which serve as proxies for convergence speed: it is not sufficient simply to maximize the MMD, since the witness function should remain Lipschitz to ensure convergence (Arbel et al., 2018, Proposition 2). It is achieved in practice by controlling the gradient of the witness function; we take a similar approach in Section 4.

4 DIFFUSION MAXIMUM MEAN DISCREPANCY GRADIENT FLOW

In this section, we present *Diffusion Maximum Mean Discrepancy gradient flow* (DMMD), a new generative model with a training procedure of MMD discriminator which does not rely on adversarial training, and leverages ideas from diffusion models. The sampling part of DMMD consists in following a noise adaptive variant of MMD gradient flow.

Adversarial-free training of noise conditional discriminators. In order to train a discriminator without adversarial training, we propose to use insights from GANs training. In a GAN setting, at

[1] In the case of variational lower bounds on f-divergences, the witness function is still well defined, and the first variation takes the same form in respect of this witness function: see Glaser et al. (2021) for the case of the KL divergence.

the beginning of the training, the generator is randomly initialized and therefore produces samples close to random noise. This would produce a coarse discriminator since it is trained to distinguish clean data from random noise. As the training progresses and the generator improves so does the discriminative power of the discriminator. This behavior of the discriminator is central in the training of GANs (Goodfellow et al., 2014). We propose a way to replicate this gradually improving behavior without adversarial training and instead relying on principles from diffusion models (Ho et al., 2020).

The forward process in diffusion models allows us to generate a probability path $P_t, t \in [0,1]$, such that $P_0 = P$, where P is our target distribution and $P_1 = \mathrm{N}(0, \mathrm{Id})$ is a Gaussian noise. Given samples $x_0 \sim P_0 = P$, the samples $x_t|x_0$ are given by

$$x_t = \alpha_t x_0 + \beta_t \epsilon, \quad \epsilon \in \mathrm{N}(0, \mathrm{Id}), \tag{6}$$

with $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$[2]. From the form of the $x_t|x_0$, we observe that for low noise level t, the samples x_t are very close to the original data x_0, whereas for the large values of x_t they are close to a unit Gaussian random variable. Using the GANs terminology, x_t could be thought as the output of a generator such that for high/low noise level t, it would correspond to *undertrained* / *well-trained* generator. Using this insight, for each noise level $t \in [0,1]$, we define a discriminator $\mathrm{MMD}^2(P_t, P; t, \theta)$ using the kernel of type (2) with noise-conditional discriminator features $\phi(x; t; \theta)$ parameterized by a Neural Network with learned parameters θ. We consider the following noise-conditional loss function

$$\mathcal{L}(\theta, t) = -\mathrm{MMD}^2(P_t, P; t, \theta) \tag{7}$$

where the minus sign comes from the fact that our aim is to maximize the squared MMD. In addition, we regularize this loss with ℓ_2-penalty (Bińkowski et al., 2021) denoted $\mathcal{L}_{\ell_2}(\theta, t)$ as well as with the gradient penalty (Bińkowski et al., 2021; Gulrajani et al., 2017) denoted $\mathcal{L}_{\nabla}(\theta, t)$, see Appendix B.2 for the precise definition of these two losses. The total noise-conditional loss is then given as

$$\mathcal{L}_{\text{tot}}(\theta, t) = \mathcal{L}(\theta, t) + \lambda_{\ell_2}\mathcal{L}_{\ell_2}(\theta, t) + \lambda_{\nabla}\mathcal{L}_{\nabla}(\theta, t), \tag{8}$$

for a suitable choice of hyperparameters $\lambda_{\ell_2} \geq 0, \lambda_{\nabla} \geq 0$. Finally, the total loss is given as $\mathcal{L}_{\text{tot}}(\theta) = \mathbb{E}_{t \sim U[0,1]}\left[\mathcal{L}_{\text{tot}}(\theta, t)\right]$, where $U[0,1]$ is a uniform distribution. In practice, we use sampled-based unbiased estimator of MMD, see Appendix B.2. The procedure is described in Algorithm 1.

Adaptive gradient flow sampling. In order to produce samples from P, we use the adaptive MMD gradient flow with noise conditional discriminators $\mathrm{MMD}^2[P_t, P; t; \theta^\star]$, where $\theta^\star$ are the discriminator parameters obtained using Algorithm 1. Let $t_i = t_{\text{min}} + i\Delta t, i = 0, \dots, T$ be the noise discretisation, where $\Delta t = (t_{\text{max}} - t_{\text{min}})/T$ such that $t_0 = t_{\text{min}}, t_T = t_{\text{max}}$ for some $t_{\text{min}} = \epsilon$ and $t_{\text{max}} = 1 - \epsilon$, where $\epsilon \ll 1$. We sample N_{p} initial particles $\{Z^i | Z^i \sim \mathrm{N}(0, \mathrm{Id})\}_{i=1}^{N_{\text{p}}}$. For each t, we follow MMD gradient flow (4) for N_{s} steps with learning rate $\eta > 0$

$$Z_t^{i,n+1} = Z_t^{i,n} - \eta \nabla f_{\nu_{N_{\text{p}},n}^t, P}(Z_t^{i,n}, t; \theta^\star). \tag{9}$$

Here $\nu_{N_{\text{p}},n}^t = 1/N_{\text{p}} \sum_{i=1}^{N_{\text{p}}} \delta_{Z_{i,n}^t}$ is the empirical distribution of particles $\{Z_t^{i,n}\}_{i=1}^{N_{\text{p}}}$ at the noise level t and the iteration n, δ is a Dirac mass measure. The function $f_{\nu_{N_{\text{p}},n}^t, P}(z, t; \theta^\star)$ is adapted from equation (1) where ν is replaced by this empirical distribution. After following the gradient flow (9) for N_{s} steps, we initialize a new gradient flow with initial particles $Z_{t-\Delta t}^{i,0} = Z_t^{i,N_{\text{s}}}$ for each $i = 1, \dots, N_{\text{p}}$, with the decreased level of noise $t - \Delta t$. The recurrence is initialized with $Z_{t_{\text{max}}}^{i,0} = Z^i$ where $\{Z^i\}_{i=1}^{N_{\text{p}}}$ are the initial particles. This procedure corresponds to running $T+1$ consecutive MMD gradient flows for N_{s} iterations each, gradually decreasing the noise level t from t_{max} to t_{min}. The resulting particles $\{Z_{t_{\text{min}}}^{i,N_{\text{s}}}\}_{i=1}^{N_{\text{p}}}$ are used as samples from P. See Algorithm 2.

In practice, we sample (once) a large batch N_c of $\{X_0^j\}_{j=1}^{N_c} \sim P^{\otimes N_c}$ from the data distribution and denote by $\hat{P}_{N_{\text{c}}}(X_0)$ the corresponding empirical distribution. Then we use the empirical witness function $f_{\nu_{N_{\text{p}},n}^t, \hat{P}_{N_{\text{c}}}(X_0)}(z, t; \theta^\star)$ given by

$$\frac{1}{N_{\text{p}}} \sum_{i=1}^{N_{\text{p}}} k_{\text{base}}(\phi(Z_t^{n,i}, t; \theta^\star), \phi(z, t; \theta^\star)) - \frac{1}{N_{\text{c}}} \sum_{j=1}^{N_{\text{c}}} k_{\text{base}}(\phi(X_0^j, t; \theta^\star), \phi(z, t; \theta^\star)). \tag{10}$$

[2]Different schedules (α_t, β_t) are available in the literature. We focus on Variance Preserving SDE ones Song et al. (2020) here

Algorithm 1 Train noise-conditional MMD discriminator

Input: Dataset $\mathcal{D} = \{x_i\}_{i=1}^N$
Discriminator features $\phi(x, t; \theta)$ with parameters $\theta \in \mathbb{R}^K$
$\lambda_\nabla \geq 0, \lambda_{\ell_2} \geq 0$ - gradient and ℓ_2 penalty coefficients
$\gamma > 0$ – learning rate
N_{iter} – number of iterations, B – batch size
N_{noise} – number of noise levels per batch
for $i = 1$ **to** N_{iter} **do**
  Sample a batch B of clean particles $X_0 \sim P(X_0)$
  for $n = 1$ **to** N_{noise} **do**
    Sample noise level $t_n \sim U[0, 1]$
    Sample $X_{t_n} \sim p(X_{t_n}|X_0, t_n)$
    Let the clean and noisy features be
    $\phi_{t_n}^{X_0} = \phi(X_0, t_n; \theta)$
    $\phi_{t_n}^{X_{t_n}} = \phi(X_{t_n}, t_n; \theta)$
    For linear base kernel (11), use optimized (19) to compute MMD loss (7)
    Compute the loss $\mathcal{L}_{\text{tot}}(\theta, t_n)$ using (8)
  end for
  Compute total loss
  $\mathcal{L}_{\text{tot}}(\theta) = \frac{1}{N_{\text{noise}}} \sum_{n=1}^{N_{\text{noise}}} \mathcal{L}_{\text{tot}}(\theta, t_n)$
  Update discriminator features
  $\theta \leftarrow \text{ADAM}(\theta, \mathcal{L}_{\text{tot}}(\theta), \gamma)$
end for

Algorithm 2 Noise-adaptive MMD gradient flow

Inputs: T – number of noise levels
$t_{\text{max}}, t_{\text{min}}$ – maximum/minimum noise levels
N_{s} – number of gradient flow steps per noise level
$\eta > 0$ – gradient flow learning rate
N_{p} – number of noisy particles
Batch of clean particles $X_0 \sim \mathcal{P}_0$.
Steps: Sample initial particles $Z \sim \text{N}(0, \text{Id})$
Set $\Delta t = (t_{\text{max}} - t_{\text{min}})/T$
for $i = T$ **to** 0 **do**
  Set the noise level $t = t_{\text{min}} + i\Delta t$
  Set $Z_t^0 = Z$
  for $n = 0$ **to** $N_{\text{s}} - 1$ **do**
    Use (10) to compute $f_{\nu^t_{N_{\text{p}},n}, \hat{P}_{N_{\text{c}}}(X_0)}(Z_t^n, t; \theta^\star)$
    $Z_t^{n+1} = Z_t^n - \eta \nabla f_{\nu^t_{N_{\text{p}},n}, \hat{P}_{N_{\text{c}}}(X_0)}(Z_t^n, t; \theta^\star)$
  end for
  Set $Z = Z_t^N$
end for
Output Z

Final denoising. In diffusion models (Ho et al., 2020), it is common to use a denoising step at the end to improve samples quality. We found empirically that a few MMD gradient flow steps at the end of the sampling with a higher learning rate η allowed to reduce noise and improve performance.

5 Scalable DMMD with linear kernel

The computational complexity of the MMD estimate on two sets of N samples is $O(N^2)$, so as as of the witness function (10) for N clean and noisy particles. Using linear base kernel (see (2))

$$k_{\text{base}}(x, y) = \langle x, y \rangle, \tag{11}$$

allows to reduce the computation complexity of both quantities down to $O(N)$, see Appendix B.3. We consider the average noise conditional discriminator features on the *whole* dataset

$$\bar{\phi}(X_0, t; \theta^\star) = \tfrac{1}{N} \textstyle\sum_{i=1}^N \phi(X_0^i, t; \theta^\star). \tag{12}$$

Using linear kernel (11) allows us to use average features (12) in the second term of (10). In practice, we can precompute these features for T timesteps and store them in memory for later use for sampling purposes. The associated storage cost is $O(TK)$ where K is the dimensionality of these features.

Approximate sampling procedure. MMD gradient flow (9) requires us to use multiple interacting particles Z to produce samples, where the interaction is captured by the first term in (10). In practice this means that the performance will depend on the number of these particles. In this section, we propose an approximation to MMD gradient flow with a linear base kernel (11) which allows us to sample particles *independently*, therefore removing the need for multiple particles. For a linear kernel, the interaction term in (10) for a particle Z, equals to $\langle \frac{1}{N_{\text{p}}} \sum_{i=1}^{N_{\text{p}}} \phi(Z_t^{n,i}, t; \theta^\star), \phi(Z, t; \theta^\star) \rangle$. For a large number of particles N_{p}, the contribution of each particle $Z_{n,i}^t$ on the interaction term with Z will be small. For a sufficiently large N_{p}, we hypothesize that $\frac{1}{N_{\text{p}}} \sum_{i=1}^{N_{\text{p}}} \phi(Z_t^{n,i}, t; \theta^\star) \approx$

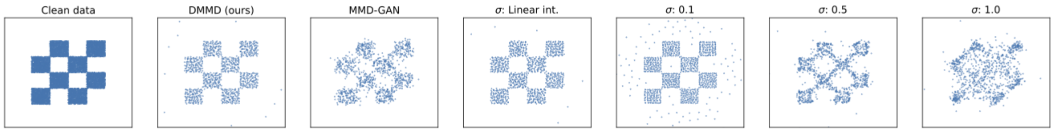

Figure 1: Samples from MMD Gradient flow with different parameters for the RBF kernel.

$\frac{1}{N}\sum_{j=1}^{N}\phi(X_t^j, t; \theta^\star)$, where N is the size of the dataset and X_t^j are produced by the forward diffusion process (6) applied to each X_0^j. In Section 7, we test this approximation in practice. Using this approximation, we consider an approximate witness function

$$\hat{f}_{P_t,P}(z) = \langle \phi(z, t; \theta^*), \bar{\phi}(X_t, t; \theta^\star) - \bar{\phi}(X_0, t; \theta^\star) \rangle, \tag{13}$$

with $\bar{\phi}(X_t, t; \theta^\star)$ precomputed using (12). In practice, we sample *single* particle $Z \sim \mathrm{N}(0, \mathrm{Id})$ and follow noise-adaptive MMD gradient flow with (13), i.e. $Z_t^{n+1} = Z_t^n - \eta \nabla \hat{f}_{P_t,P}(Z_t^n)$. The corresponding algorithm is described in Appendix B.4.

6 Related Work

Adversarial training and MMD-GAN. Integral Probability Metrics (IPMs) are good candidates to define discriminators in the context of generative modeling, since they are well defined even in the case of distributions with non-overlapping support (Muller, 1997). Moreover, implementations of f-divergence discriminators in GANs rely on variational lower bounds (Nowozin et al., 2016): as noted earlier, these share useful properties of IPMs in theory and in practice (notably, they remain well defined for distributions with disjoint support, and may metrize weak convergence for sufficiently rich witness function classes (Arbel et al., 2021, Proposition 14) and (Zhang et al., 2018)). Several works (Arjovsky et al., 2017; Gulrajani et al., 2017; Genevay et al., 2018; Li et al., 2017; Bińkowski et al., 2021) have exploited IPMs as discriminators for the training of GANs, where the IPMs are MMDs using (linear or nonlinear) kernels defined on learned neural net features, making them suited to high dimensional settings such as image generation. Interpreting the IPM-based GAN discriminator as a squared MMD yields an interesting theoretical insight: Franceschi et al. (2022) show that training a GAN with an IPM objective implicitly optimizes MMD^2 in the Neural Tangent Kernel (NTK) limit (Jacot et al., 2020). IPM GAN discriminators are trained jointly with the generator in a min-max game. Adversarial training is challenging, and can suffer from instability, mode collapse, and misconvergence (Xiao et al., 2022; Bińkowski et al., 2021; Li et al., 2017; Arora et al., 2017; Kodali et al., 2017; Salimans et al., 2016). Note that once a GAN has been trained, the samples can be refined via MCMC sampling in the generator latent space (e.g., using kinetic Langevin dynamics; see Ansari et al., 2021; Che et al., 2021; Arbel et al., 2021).

Discriminator flows for generative modeling. Wasserstein Gradient flows (Ambrosio et al., 2008; Santambrogio, 2015) applied to a GAN discriminator are informally called *discriminator flows*, see (Franceschi et al., 2023). A number of recent works have focused on replacing a GAN generator by a discriminator flow. Fan et al. (2022) propose a discretisation of JKO (Jordan et al., 1998) scheme to define a Kullback-Leibler (KL) divergence gradient flow. Other approaches have used a discretized interactive particle-based approach instead of JKO, similar to (3). Heng et al. (2023); Franceschi et al. (2023) build such a flow based on f-divergences, whereas Franceschi et al. (2023) focuses on MMD gradient flow. In all these works, an explicit generator is replaced by a corresponding discriminator flow. The sampling process during training is as follows: Let Y_k be the samples produced at training iteration k by the gradient flow $\mathcal{F}_k$ induced by the discriminator $\mathcal{D}_k$ applied to samples Y_{k-1} from the previous iteration. We denote this by $Y_k \leftarrow \mathcal{F}_k(\mathcal{D}_k, Y_{k-1})$. Then, the discriminator at iteration $k+1$ is trained on samples Y_k. A challenge of this process is that the training sample for the next discriminator will be determined by the previous discriminators, and thus the generation process is still adversarial: particle transport minimizes the previous discriminator value, and the subsequent discriminator is maximized on these particles. Consequently, it is difficult to control or predict the overall sample trajectory from the initial distribution to the target, which might explain the performance shortfall of these methods in image generation settings. By contrast, we have explicit control over the training particle trajectory via the forward noising diffusion process.

On top of that, these approaches (except for Heng et al., 2023) require to store all intermediate discriminators $\mathcal{D}_1, \ldots, \mathcal{D}_N$ throughout training (N is the total number of training iterations). These discriminators are then used to produce new samples by applying the sequence of gradient flows $\mathcal{F}_N(\mathcal{D}_N, \cdot) \circ \ldots \circ \mathcal{F}_1(\mathcal{D}_1, \cdot)$ to Y_0 sampled from the initial distribution. This creates a large memory overhead. An alternative is to use pretrained features obtained elsewhere or a fixed kernel with empirically selected hyperparameters (see Hertrich et al., 2023; Hagemann et al., 2023; Altekrüger et al., 2023), however this limits the applicability of the method. To the best of our knowledge, our approach is the first to demonstrate the possibility to train a discriminator without adversarial training, such that this discriminator can then be used to produce samples with a gradient flow. Unlike the alternatives, our approach does not require to store intermediate discriminators.

MMD **for diffusion refinement/regularization.** MMD has been used to either regularize training of diffusion models (Li & van der Schaar, 2024) or to finetune them (Aiello et al., 2023) for fast sampling. The MMD kernel in these works has the form (2) with Inception features (Szegedy et al., 2014). Our method removes the need to use pretrained features by training th MMD discriminator.

Diffusion models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) represent a powerful new family of generative models due to their strong empirical performance in many domains (Saharia et al., 2022; Le et al., 2023; Ho et al., 2022; Watson et al., 2022; Poole et al., 2022). Unlike GANs, diffusion models do not require adversarial training. At training time, a denoiser is learned for multiple noise levels. As noted above, our work borrows from the training of diffusion models, as we train a discriminator on multiple noise levels of the forward diffusion process (Ho et al., 2020). This gives better control of the training samples for the (noise adapted) discriminator than using an incompletely trained GAN generator.

7 Experiments

Understanding DMMD **behavior in 2-D.** Our aim is to get an understanding of the behavior of DMMD described in Section 4. We expect DMMD to mimic GAN discriminator training via noise conditional discriminator learning. To see whether this manifests in practice, we design an experiment with Radial Basis Function (RBF) kernel for MMD, $k_t(x, y) = \exp[-\|x - y\|^2/(2\sigma^2(t; \theta))]$, where the noise dependent kernel width function $\sigma(\cdot; \theta) : [0, 1] \to [0, +\infty)$ is parameterized by $\theta \in \mathbb{R}^K$. This parameter controls the coarseness of the MMD discriminator. We consider 2-D checkerboard dataset, see Figure 1, left. We learn noise-conditional kernel widths $\sigma(t; \theta)$ using a neural network. As baselines, we train MMD-GAN where distriminator learns σ, as well as MMD gradient flow with fixed values of σ and with a manually selected noise-dependent $\sigma(t) = 0.1(1 - t) + 0.5t$ called *linear interpolation*. All experimental details are provided in Appendix C.

We report the learned RBF kernel widths for DMMD in Figure 2, left. As expected, as noise level goes from high to low, the kernel width $\sigma(t)$ decreases. In Figure 2, center, we show the learned MMD-GAN kernel width parameter σ as a function of training iterations. When the training progresses, this parameter decreases, since the corresponding generator produces samples, close to the target distribution. The behaviors of DMMD and MMD-GAN are quite similar and so as the range of values for the kernel widths is also similar. This highlights our point that DMMD mimics the training of a GAN discriminator. The exact dynamics for $\sigma(t)$ in DMMD depends on the parameters of the forward diffusion process (6). The sharp phase transition is consistent with the phase transition highlighted in Section 3. In addition, we report $\mathrm{MMD}^2(P_t, P; t)$ for different methods in Figure 2, right. We see that DMMD behaves similarly to *linear interpolation*, but is more nuanced for higher noise levels. The samples are reported in Figure 1. DMMD produces samples which are visually better than the other baselines. For RBF kernel, we noticed the presence of outliers. The amount of outliers generally depends on the kernel, see Appendix of (Hertrich et al., 2023) for more details.

Image generation We study the performance of DMMD on unconditional image generation of CIFAR10 (Krizhevsky et al., 2009). We use the same forward diffusion process as in (Ho et al., 2020) to produce noisy images. We use a U-Net (Ronneberger et al., 2015) backbone for discriminator feature network $\phi(x, t; \theta)$, with a slightly different architecture from the one used in (Ho et al., 2020), see Appendix F. For all the image-based experiments, we use linear base kernel (11). We explored using other kernels such as RBF and Rational Quadratic (RQ), but did not find an improvement in performance. We use FID (Heusel et al., 2018) and Inception Score (Salimans et al., 2016) for evaluation, see Appendix F. Unless specified otherwise, we use the number $N_{\mathrm{p}} = 200$ of particles for Algorithm 2. We provide ablation over the number of particles in Appenidx F.3. The total number of iterations for DMMD equals to $T \times N_{\mathrm{s}}$, where T is the number of noise levels and N_{s} is the number of steps per noise level. For consistency with diffusion models, we call this *number of function evaluations* (NFE). For DMMD, we show performance with different NFEs. As we show in Appendix G (see Table 6), there is an improvement on FID as we increase NFEs, but only up to a point (NFE=250).

Table 1: **Unconditional image generation on CIFAR-10**. For MMD GAN (orig.), we used mixed-RQ kerned (see (Bińkowski et al., 2021)). "Orig." – original paper, "impl." – our implementation. For JKO-Flow (Fan et al., 2022), the NFE is taken from their Figure 12.

Method	FID	IS	NFE
MMD GAN (orig.)	39.90	6.51	-
MMD GAN (impl.)	13.62	8.93	-
DDPM (orig.)	3.17	9.46	1000
DDPM (impl.)	5.19	8.90	100
Discriminator flow baselines			
DGGF-KL	28.80	-	110
JKO-Flow	23.10	7.48	~ 150
MMD flow baselines			
MMD-GAN-Flow	450	1.21	100
GS-MMD-RK	55.00	-	86
DMMD (ours)	**8.31**	**9.09**	100
DMMD (ours)	**7.74**	**9.12**	250

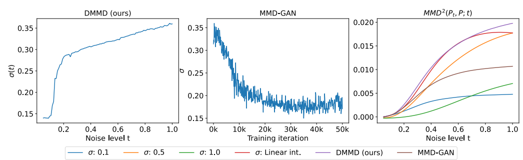

Figure 2: **Toy experiment**. *Left*, learned RBF kernel widths $\sigma(t)$ for DMMD. *Center*, σ for MMD-GAN as function of training iterations. *Right*, $\mathrm{MMD}^2(P_t, P; t)$ for different methods.

Table 2: **Approximate sampling** performance on CIFAR10. IS stands for Inception score

Method	FID	IS	NFE
DMMD	8.31	9.09	100
DMMD-e	**8.21**	8.99	102
a-DMMD	24.86	9.10	50
a-DMMD-e	**9.185**	8.70	52
a-DMMD-a	11.22	9.00	52

As baselines we consider our implementation of MMD-GAN (Bińkowski et al., 2021) with linear base kernel and DDPM (Ho et al., 2020) using the same neural network backbones as for DMMD. We also report results from the original papers. On top of that, we consider baselines based on *discriminator flows*. JKO-Flow (Fan et al., 2022), which uses JKO (Jordan et al., 1998) scheme for the KL gradient flow. Deep Generative Wasserstein Gradient Flows (DGGF-KL) (Heng et al., 2023), which uses particle-based approach (similar to (3)) for the KL gradient flow. These approaches use adversarial training to train discriminators, see Section 6 for more details. On top of that, we consider Generative Sliced MMD Flows with Riesz Kernels (GS-MMD-RK) (Hertrich et al., 2023) which uses similar particle based approach to DGGF-KL to construct MMD flow, but uses fixed (kernel) discriminator. On top of that, we report results using a discriminator flow defined on a trained MMD-GAN discriminator which we call MMD-GAN-Flow. More details on experiments are given in Appendix F. The results are provided in Table 1.

We see that DMMD achieves better performance than the MMD GAN. As expected, MMD-GAN-Flow does not work at all. This is because the MMD-GAN discriminator at convergence was trained on samples close to the target distribution. Making a parallel with RBF kernel experiment from, this means that the gradient of MMD will be very small on samples far away from the target distribution. This highlights the benefit of adaptive MMD discriminators. Moreover, we also see that DMMD performs better than GS-MMD-RK, which uses fixed kernel. This highlights the advantage of learning discriminator features in DMMD. DMMD achieves superior performance compared to other discriminator flow baselines. We believe that one of the reasons of the under-performance of these methods is adversarial training, which makes the hyperparameters choice tricky. DMMD on the other hand, relies on a simple non-adversarial training procedure from Algortihm 1. Finally, we see that DDPM performs better than DMMD. This is not surprising, since both, U-Net architecture and forward diffusion process (6) were optimized for DDPM performance. Nevertheless, DMMD demonstrates strong empirical performance as a discriminator flow method trained without adversarial training. The samples from our method are provided in Appendix I.1. We provide results on CELEB-A, LSUN Church and MNIST below.

Approximate sampling. We run approximate MMD gradient flow (see Section 5) with the same discriminator as for DMMD. We call this variant a-DMMD, where a stands for *approximate*. On top of that, we use denoising procedure described in Section 4. Starting from the samples given by a-DMMD, we do 2 gradient flow steps with higher learning rate using either approximate gradient flow, which we call a-DMMD-a, or exact gradient flow (9) applied to a single particle, which we call a-DMMD-e, e stands for *exact*. On top of that, we apply the denoising to DMMD, which we call DMMD-e. Results are provided in Table 2. We observe that a-DMMD performs worse than DMMD, which is as expected. Applying a denoising step improves performance of a-DMMD, bringing it closer to DMMD. This suggests that the approximation (13) moves the particles close to the target distribution; but once close to the target, a more refined procedure is required. By contrast, we see that denoising helps DMMD only marginally. This suggests that the *exact* noise-conditional witness function (10) accurately captures fine detais close to the target distribution.

Results on MNIST, CELEB-A (64x64) and LSUN-Church (64x64) Besides CIFAR-10, we study the performance of DMMD on MNIST (Lecun et al., 1998), CELEB-A (64x64 (Liu et al., 2015) and LSUN-Church (64x64) (Yu et al., 2016). For MNIST and CELEB-A, we consider the same splits and evaluation regime as in (Franceschi et al., 2023). For LSUN Church, the splits and the evaluation regime are taken from (Ho et al., 2020). For more details, see Appendix F.1. As baselines, we consider our implementations of DDPM (Ho et al., 2020), MMD-GAN (Bińkowski et al., 2021). In addition to DMMD, we report the performance of *Discriminator flow* baseline from (Franceschi et al., 2023) with numbers taken from the corresponding paper. This baseline uses adversarial training together with MMD gradient flow to produce samples. The results are provided in Table 3. We see that DMMD performance is better compared to the discriminator flow and MMD-GAN, which is consistent with our findings on CIFAR-10. It also underperforms compared to DDPM. The corresponding samples are provided in Appendix I.2.

Table 3: **Unconditional image generation on additional datasets**. The metric used is FID. The number of gradient flow steps for DMMD is 100.

Dataset	MMD-GAN	DDPM	DMMD	Disc. flow
MNIST	7.0	1.94	3.0	4.0
CELEB-A	12.1	6.72	8.3	41.0
LSUN	8.4	3.84	6.1	-

8 Conclusion

In this paper we have presented a method to train a noise conditional discriminator without adversarial training, using a forward diffusion process. We use this noise conditional discriminator to generate samples using a noise adaptive MMD gradient flow. We provide theoretical insight into why an adaptive gradient flow can provide faster convergence than the non-adaptive variant. We demonstrate strong empirical performance of our method on uncoditional image generation of CIFAR10, as well as on additional, similar image datasets. We propose a scalable approximation of our approach which has close to the original empirical performance.

A number of questions remain open for future work. The empirical performance of DMMD will be of interest in regimes where diffusion models could be ill-behaved, such as in generative modeling on Riemannian manifolds; as well as on larger datasets such as ImageNet. DMMD provides a way of training a discriminator, which may be applicable in other areas where a domain-adaptive discriminator might be required. Finally, it will be of interest to establish theoretical foundations for DMMD in general settings, and to derive convergence results for the associated flow.

References

Aiello, E., Valsesia, D., and Magli, E. Fast inference in denoising diffusion models via mmd finetuning, 2023.

Altekrüger, F., Hertrich, J., and Steidl, G. Neural wasserstein gradient flows for maximum mean discrepancies with riesz kernels, 2023.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, 2. ed edition, 2008. ISBN 978-3-7643-8722-8 978-3-7643-8721-1. OCLC: 254181287.

Ansari, A. F., Ang, M. L., and Soh, H. Refining deep generative models via discriminator gradient flow, 2021.

Arbel, M., Sutherland, D. J., Bińkowski, M., and Gretton, A. On gradient regularizers for mmd gans. *Advances in neural information processing systems*, 31, 2018.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow, 2019.

Arbel, M., Zhou, L., and Gretton, A. Generalized energy based models, 2021.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans), 2017.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans, 2021.

Bortoli, V. D. Convergence of denoising diffusion models under the manifold hypothesis, 2023.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis, 2019.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. The union of manifolds hypothesis and its implications for deep generative modelling. *arXiv preprint arXiv:2207.02862*, 2022.

Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y., and Bengio, Y. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling, 2021.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Fan, J., Zhang, Q., Taghvaei, A., and Chen, Y. Variational wasserstein gradient flow, 2022.

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

Franceschi, J.-Y., de Bézenac, E., Ayed, I., Chen, M., Lamprier, S., and Gallinari, P. A neural tangent kernel perspective of gans, 2022.

Franceschi, J.-Y., Gartrell, M., Santos, L. D., Issenhuth, T., de Bézenac, E., Chen, M., and Rakotomamonjy, A. Unifying gans and score-based diffusion as generative particle models, 2023.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 2018.

Glaser, P., Arbel, M., and Gretton, A. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support, 2021.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans, 2017.

Hagemann, P., Hertrich, J., Altekrüger, F., Beinert, R., Chemseddine, J., and Steidl, G. Posterior sampling based on gradient flows of the mmd with negative distance kernel, 2023.

Heng, A., Ansari, A. F., and Soh, H. Deep generative wasserstein gradient flows, 2023. URL https://openreview.net/forum?id=zjSeBTEdXp1.

Hertrich, J., Wald, C., Altekrüger, F., and Hagemann, P. Generative sliced mmd flows with riesz kernels, 2023.

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, 2018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks, 2020.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data, 2020a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan, 2020b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of gans, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. Mmd gan: Towards deeper understanding of moment matching network, 2017.

Li, Y. and van der Schaar, M. On error propagation of diffusion models, 2024.

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., and Mallya, A. Generative adversarial networks for image and video synthesis: Algorithms and applications, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild, 2015.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Muller, A. Integral probability metrics and their generating classes of functions. volume 29, pp. 429–443. Advances in Applied Probability, 1997.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization, 2016.

Pidstrigach, J. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans, 2016.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 06 2018. ISBN 9780262256933. doi: 10.7551/mitpress/4175.001.0001. URL https://doi.org/10.7551/mitpress/4175.001.0001.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.co.uk/books?id=hV8o5R7_5tkC.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pp. 2022–12, 2022.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion gans, 2022.

Xu, Y., Zhao, Y., Xiao, Z., and Hou, T. Ufogen: You forward once large scale text-to-image generation via diffusion gans. *arXiv preprint arXiv:2311.09257*, 2023.

Yang, Z., Feng, R., Zhang, H., Shen, Y., Zhu, K., Huang, L., Zhang, Y., Liu, Y., Zhao, D., Zhou, J., and Cheng, F. Eliminating lipschitz singularities in diffusion models, 2023.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. On the discrimination-generalization tradeoff in gans. In *6th International Conference on Learning Representations*, 2018.