On gradient regularizers for MMD GANs

Michael Arbel
Gatsby Computational Neuroscience Unit
University College London
michael.n.arbel@gmail.com

Dougal J. Sutherland
Gatsby Computational Neuroscience Unit
University College London
dougal@gmail.com

Mikołaj Bińkowski
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

Abstract

We propose a principled method for gradient-based regularization of the critic of GAN-like models trained by adversarially optimizing the kernel of a Maximum Mean Discrepancy (MMD). Our method is based on studying the behavior of the optimized MMD, and constrains the gradient based on analytical results rather than an optimization penalty. Experimental results show that the proposed regularization leads to stable training and outperforms state-of-the art methods on image generation, including on 160×160 CelebA and 64×64 ImageNet.

1 Introduction

There has been an explosion of interest in *implicit generative models* (IGMs) over the last few years, beginning with the introduction of generative adversarial networks (GANs) [14]. These models allow approximate samples from a complex high-dimensional target distribution $\mathbb{P}$, using a model distribution $\mathbb{Q}_\theta$, where estimation of likelihoods, exact inference, and so on are not tractable. GAN-type IGMs have yielded very impressive empirical results, particularly for image generation, far beyond the quality of samples seen from most earlier generative models [e.g. 16, 19–21, 35].

These excellent results, however, have depended on adding a variety of methods of regularization and other tricks to stabilize the notoriously difficult optimization problem of GANs [35, 39]. Some of this difficulty is perhaps because when a GAN is viewed as minimizing a discrepancy $\mathcal{D}_{\mathrm{GAN}}(\mathbb{P}, \mathbb{Q}_\theta)$, its gradient $\nabla_\theta \mathcal{D}_{\mathrm{GAN}}(\mathbb{P}, \mathbb{Q}_\theta)$ does not provide useful signal to the generator if the target and model distributions are not absolutely continuous, as is nearly always the case [1].

An alternative set of losses are the integral probability metrics (IPMs) [33], many of which give credit to generator distributions $\mathbb{Q}_\theta$ "near" to the target distribution $\mathbb{P}$ [2, 7, Section 4 of 13]. IPMs are defined in terms of a *critic function*: a "well behaved" function with large amplitude where the mass of $\mathbb{P}$ and $\mathbb{Q}_\theta$ differs most. The IPM is the difference in the expected critic under $\mathbb{P}$ and $\mathbb{Q}_\theta$, and is zero when the distributions agree. The Wasserstein IPMs, whose critics are made smooth via a Lipschitz constraint, have been particularly successful as losses for IGMs [2, 12, 16]. But the Lipschitz constraint needs to hold uniformly, which can be hard to enforce. A popular approach has been to apply a gradient constraint only in expectation [16]: the gradient of the critic is constrained to be near norm 1 on points chosen uniformly along lines between samples from $\mathbb{P}$ and $\mathbb{Q}$.

Another class of integral probability metrics used as IGM losses are the Maximum Mean Discrepancies (MMDs) [15], as proposed by [11, 25]. In these cases, the critic function is a member of a reproducing kernel Hilbert space. Far better performance can be obtained, however, when the MMD is not defined directly on the generator and target samples, but on learned features of these samples.

Preprint. Work in progress.

Since the MMD is an IPM, the gradient regularisation approaches for Wasserstein GANs can apply when learning these features: [24] use weight clipping [2], and [4, 6] use the gradient penalty of [16].

The recent Sobolev GAN [30] uses a similar constraint on the expected gradient norm, but phrases it as estimating a Sobolev IPM rather than loosely approximating Wasserstein. This expectation can be taken over the same distribution as [16], but other measures are also proposed, such as $(\mathbb{P} + \mathbb{Q}_\theta)/2$. A second recent approach, the spectrally normalized GAN [29], controls the Lipschitz constant of the critic by enforcing the spectral norms of the weight matrices to be 1. Gradient penalties also benefit GANs based on f-divergences [34]: for instance, the spectral normalization technique of [29] can be applied to the critic network of an f-GAN. Alternatively, a gradient penalty can be defined to approximate the effect of blurring $\mathbb{P}$ and $\mathbb{Q}_\theta$ with noise [37], which addresses the problem of non-overlapping support [1]. This approach has recently been shown to yield locally convergent optimization in some cases with non-continuous distributions, where the original GAN does not [27].

In this paper, we introduce a novel regularization for the MMD GAN critic of [4, 6, 24], that *directly targets generator performance*, rather than adopting regularization methods intended to approximate Wasserstein distances [2, 16]. The new MMD regularizer derives from an approach widely used in semi-supervised learning [9, Section 2], where the aim is to define a classification function f which is positive on $\mathbb{P}$ (the positive class) and negative on $\mathbb{Q}_\theta$ (negative class), in the absence of labels on many of the samples. The decision boundary between the classes is assumed to be in a region of low density for both $\mathbb{P}$ and $\mathbb{Q}_\theta$: f should therefore be flat where $\mathbb{P}$ and $\mathbb{Q}_\theta$ have support (areas with constant label), and have a larger slope in regions of low density. Bousquet et al. [9] propose as their regularizer on f a sum of the variance and a density-weighted gradient norm.

We adopt a related penalty on the MMD critic, with the difference that we only apply the penalty on $\mathbb{P}$: thus, the critic is flatter where $\mathbb{P}$ has high mass, but does not vanish on the generator samples from $\mathbb{Q}_\theta$ (which we optimize). In excluding $\mathbb{Q}_\theta$ from the critic function constraint, we also avoid the concern raised by [29] that a critic depending on $\mathbb{Q}_\theta$ will change with the current minibatch – potentially leading to less stable learning. The resulting discrepancy is no longer an integral probability metric: it is asymmetric, and the critic function class depends on the target $\mathbb{P}$ being approximated.

We first discuss in Section 2 how MMD-based losses can be used to learn implicit generative models, and how a naive approach could fail. This motivates our new discrepancies, introduced in Section 3. Section 4 demonstrates that these losses outperform state-of-the-art models for image generation.

2 Learning implicit generative models with MMD-based losses

An IGM is a model $\mathbb{Q}_\theta$ which aims to approximate a target distribution $\mathbb{P}$ over a space $\mathcal{X} \subseteq \mathbb{R}^d$. We will define $\mathbb{Q}_\theta$ by a *generator* function $G_\theta : \mathcal{Z} \to \mathcal{X}$, implemented as a deep network with parameters θ, where $\mathcal{Z}$ is a space of latent codes, say $\mathbb{R}^{128}$. We assume a simple fixed distribution on $\mathcal{Z}$, say $Z \sim \mathcal{N}(0, I)$, and call $\mathbb{Q}_\theta$ the distribution of $G_\theta(Z)$. We will consider learning by minimizing a discrepancy $\mathcal{D}$ between distributions, satisfying $\mathcal{D}(\mathbb{P}, \mathbb{Q}_\theta) \geq 0$ and $\mathcal{D}(\mathbb{P}, \mathbb{P}) = 0$, which we call our *loss*. We aim to minimize $\mathcal{D}(\mathbb{P}, \mathbb{Q}_\theta)$ by using stochastic gradient descent on an estimator of $\mathcal{D}$.

In the present work, we will build losses $\mathcal{D}$ based on the Maximum Mean Discrepancy,

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{f : \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)], \tag{1}$$

an integral probability metric where the critic class is the unit ball within $\mathcal{H}_k$, the reproducing kernel Hilbert space with a kernel k. The optimization in (1) admits a simple closed-form optimal critic, $f^*(t) \propto \mathbb{E}_{X \sim \mathbb{P}}[k(X, t)] - \mathbb{E}_{Y \sim \mathbb{Q}}[k(Y, t)]$. There is also a simple unbiased, closed-form estimator of the MMD with appealing statistical properties [15] – in particular, the estimator's sample complexity is *independent* of the dimension of $\mathcal{X}$, compared to the exponential dependence of Wasserstein [46].

The MMD is *weak* for any bounded kernel with Lipschitz embeddings [43, Theorem 3.2(b)], in the sense that if $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$ then $\mathrm{MMD}(\mathbb{P}_n, \mathbb{P}) \to 0$, where $\xrightarrow{D}$ denotes convergence in distribution. (The same is true for the Wasserstein distance.) This is desirable in that it means the loss can provide better signal to the generator as $\mathbb{Q}_\theta$ approaches $\mathbb{P}$, as opposed to e.g. Jensen-Shannon where the loss could be constant and then jump to 0 [e.g. 2, Example 1]. The MMD is *strict*, zero iff $\mathbb{P} = \mathbb{Q}_\theta$, for any *characteristic* kernel [42]. The Gaussian RBF kernel, for example, yields an MMD which satisfies both properties. Thus in principle, one need not conduct any alternating optimization in an IGM at all, but merely minimize the MMD over generator parameters θ.

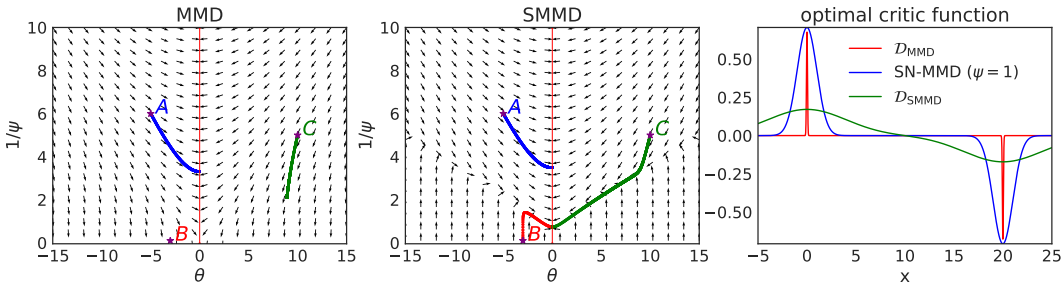


Figure 1: The problem of Section 2: $\mathbb{P} = \delta_0$, $\mathbb{Q} = \delta_\theta$, $\phi_\psi(x) = \psi x$. Left: gradient vector fields for the MMD. Center: for the SMMD (Section 3.2). Right: optimal critics for $\theta = 20$.

Despite these appealing properties, using simple pixel-level kernels leads to poor generator samples [7, 11, 25, 44]. More recent MMD GANs [4, 6, 24] achieve better results by parameterizing a family of kernels $\{k_\psi\}_{\psi\in\Psi}$ in a loss we will call the Optimized MMD, previously studied by [41, 43]:

$$\mathcal{D}_{\text{MMD}}(\mathbb{P}, \mathbb{Q}) := \sup_{\psi\in\Psi} \text{MMD}_{k_\psi}(\mathbb{P}, \mathbb{Q}). \tag{2}$$

We primarily consider kernels defined by some fixed kernel K on top of a learned representation $\phi_\psi : \mathcal{X} \to \mathbb{R}^s$, i.e. $k_\psi(x, y) = K(\phi_\psi(x), \phi_\psi(y))$, denoted $k_\psi = K \circ \phi_\psi$. This is sufficient for reasonable choices of K and ϕ; we need not try to ensure each k_ψ is characteristic, as did [24].

Proposition 1. *Suppose K is characteristic, and $\{\phi_\psi\}_{\psi\in\Psi}$ is rich enough that for any $\mathbb{P} \neq \mathbb{Q}$, there is some $\psi \in \Psi$ such that $\phi_\psi(\mathbb{P}) \neq \phi_\psi(\mathbb{Q})$. Then $\mathcal{D}_{\text{MMD}}$ is strict: if $\mathbb{P} \neq \mathbb{Q}$, then $\mathcal{D}_{\text{MMD}}(\mathbb{P}, \mathbb{Q}) > 0$.*

Proof. Let $\hat{\psi} \in \Psi$ be such that $\phi_{\hat{\psi}}(\mathbb{P}) \neq \phi_{\hat{\psi}}(\mathbb{Q})$. Then, since K is characteristic,

$$\mathcal{D}_{\text{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi\in\Psi} \text{MMD}_K(\phi_\psi(\mathbb{P}), \phi_\psi(\mathbb{Q})) \geq \text{MMD}_K(\phi_{\hat{\psi}}(\mathbb{P}), \phi_{\hat{\psi}}(\mathbb{Q})) > 0. \qquad \square$$

To estimate $\mathcal{D}_{\text{MMD}}$, one can conduct alternating optimization to estimate a $\hat{\psi}$ and then update the generator according to $\text{MMD}_{k_{\hat{\psi}}}$, similar to the scheme used in GANs and WGANs.[1] Unlike $\mathcal{D}_{\text{GAN}}$ or $\mathcal{W}$, fixing a $\hat{\psi}$ and optimizing the generator still yields a sensible distance $\text{MMD}_{k_{\hat{\psi}}}$.

Early attempts at minimizing $\mathcal{D}_{\text{MMD}}$ in an IGM, though, were unsuccessful [44, footnote 7]. This could be because for some kernel classes, $\mathcal{D}_{\text{MMD}}$ is stronger than Wasserstein or MMD. Consider the following example in $\mathbb{R}$ [27]: we wish to model a point mass at the origin, $\mathbb{P} = \delta_0$, with any possible point mass, $\mathbb{Q}_\theta = \delta_\theta$ for $\theta \in \mathbb{R}$. Taking $\phi_\psi(x) = \psi x$ for $\psi \in \mathbb{R}$ and $K(a, b) = \exp\left(-\frac{1}{2}(a - b)^2\right)$,

$$\text{MMD}^2_{k_\psi}(\delta_0, \delta_\theta) = 2\left(1 - \exp\left(-\tfrac{1}{2}\psi^2\theta^2\right)\right), \qquad \mathcal{D}_{\text{MMD}}(\delta_0, \delta_\theta) = \begin{cases} \sqrt{2} & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}. \tag{3}$$

Considering $\mathcal{D}_{\text{MMD}}(\delta_0, \delta_{1/n}) = \sqrt{2} \not\to 0$, though $\delta_{1/n} \xrightarrow{D} \delta_0$, shows that this $\mathcal{D}_{\text{MMD}}$ is not weak. This also causes optimization issues. Figure 1 (left) shows gradient vector fields in parameter space, $v(\theta, \psi) \propto \left(-\nabla_\theta \text{MMD}^2_{k_\psi}(\delta_0, \delta_\theta), \nabla_\psi \text{MMD}^2_{k_\psi}(\delta_0, \delta_\theta)\right)$. Some sequences following v converge to an optimal solution $(0, \psi)$, but others are stuck because there is essentially no gradient (B). Figure 1 (right, red) shows that the optimal $\mathcal{D}_{\text{MMD}}$ critic is very sharp near $\mathbb{P}$ and $\mathbb{Q}$; this is less true for cases where the algorithm converged. We can avoid these issues if we ensure a bounded Lipschitz critic:[2]

Proposition 2. *Assume the critics $f_\psi(x) = (\mathbb{E}_{X\sim\mathbb{P}} k_\psi(X, x) - \mathbb{E}_{Y\sim\mathbb{Q}} k_\psi(Y, x))/\text{MMD}_{k_\psi}(\mathbb{P}, \mathbb{Q})$ are uniformly bounded and have a common Lipschitz constant: $\sup_{x\in\mathcal{X},\psi\in\Psi}|f_\psi(x)| < \infty$ and $\sup_{\psi\in\Psi}\|f_\psi\|_L < \infty$. In particular, this holds when $k_\psi = K \circ \phi_\psi$ and*

$$\sup_{a\in\mathbb{R}^s} K(a, a) < \infty, \quad \|K(a, \cdot) - K(b, \cdot)\|_{\mathcal{H}_K} \leq L_K\|a - b\|_{\mathbb{R}^s}, \quad \sup_{\psi\in\Psi}\|\phi_\psi\|_L \leq L_\phi < \infty.$$

Then $\mathcal{D}_{\text{MMD}}$ is weak: if $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$, then $\mathcal{D}_{\text{MMD}}(\mathbb{P}_n, \mathbb{P}) \to 0$.

[1] This form of estimator is justified by an envelope theorem [28], but note it is invariably biased [6].

[2] [24, Theorem 4] makes a similar claim to Proposition 2, but its proof was incorrect: it tries to uniformly bound $\text{MMD}_{k_\psi} \leq \mathcal{W}^2$, but the bound used is for a Wasserstein in terms of $\|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_{k_\psi}}$.

Proof. The main result is [10, Corollary 11.3.4]. To show the claim for $k_\psi = K \circ \phi_\psi$, note that $|f_\psi(x) - f_\psi(y)| \le \|f_\psi\|_{\mathcal{H}_{k_\psi}} \|k_\psi(x,\cdot) - k_\psi(y,\cdot)\|_{\mathcal{H}_{k_\psi}}$, which since $\|f_\psi\|_{\mathcal{H}_{k_\psi}} = 1$ is

$$\|K(\phi_\psi(x),\cdot) - K(\phi_\psi(y),\cdot)\|_{\mathcal{H}_K} \le L_K \|\phi_\psi(x) - \phi_\psi(y)\|_{\mathbb{R}^s} \le L_K L_\phi \|x - y\|_{\mathbb{R}^d}. \qquad \square$$

Indeed, if we put a box constraint on ψ [24] or especially if we regularize the gradient of the critic function [6], the resulting MMD GAN generally matches or outperforms WGAN-based models.

Unfortunately, though, an additive gradient penalty doesn't substantially change the vector field of Figure 1. The distance we will propose in Section 3.2 has much nicer convergence behavior.

3 New discrepancies for learning implicit generative models

Our aim here is to introduce a discrepancy that can provide useful gradient information when used as an IGM loss. All proofs of results in this section can be found in Appendix A.

3.1 Gradient-Constrained Maximum Mean Discrepancy

Letting $\mathcal{H}$ be an RKHS with kernel k and norm $\|\cdot\|_{\mathcal{H}}$, we define the Gradient-Constrained MMD as

$$S_{k,\mu,\lambda}(\mathbb{P},\mathbb{Q}) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{S(\mu),\lambda} \le 1}} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)], \tag{4}$$

$$\text{where } \|f\|^2_{S(\mu),\lambda} := \|f\|^2_{L^2(\mu)} + \|\nabla f\|^2_{L^2(\mu)} + \lambda \|f\|^2_{\mathcal{H}}, \tag{5}$$

$\|\cdot\|_{L^2(\mu)}$ is the L^2 norm under some measure μ, and $\lambda > 0$. The second term $\|\nabla f\|^2_{L^2(\mu)}$ encourages the function f to be flat where μ has mass. In in this work, we will focus on $\mu = \mathbb{P}$, flattening the critic in the vicinity of the target sample. We add the first term following [9]: in one dimension and with μ uniform, $\|\cdot\|_{S(\mu),0}$ is then an RKHS norm with the Laplace kernel $\kappa(x,y) = \exp(-\|x-y\|)$, which is also a Sobolev space. The correspondence to a Sobolev norm is lost in higher dimensions [see e.g. 47, Ch. 10], but we also found it beneficial in practice to retain this form.

We can obtain a more tractable expression for $S_{k,\mu,\lambda}$ after making a few mild assumptions (in Appendix A); we assume $\mathbb{P}$ and $\mathbb{Q}$ have integrable first moments, and the others are satisfied e.g. by a Gaussian kernel and differentiable ϕ_ψ with $\mu = \mathbb{P}$. Let $\eta := \mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)] - \mathbb{E}_{Y\sim\mathbb{Q}}[k(Y,\cdot)] \in \mathcal{H}$ be the difference in kernel mean embeddings; recall $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \|\eta\|_{\mathcal{H}}$. We also need the regularized covariance-type operator D_λ, which is a linear operator on functions in $\mathcal{H}$ given by

$$\langle f, D_\lambda g\rangle_{\mathcal{H}} := \int f(x)\, g(x)\, \mu(\mathrm{d}x) + \sum_{i=1}^{d} \int \partial_i f(x)\, \partial_i g(x)\, \mu(\mathrm{d}x) + \lambda \langle f, g\rangle_{\mathcal{H}}. \tag{6}$$

Proposition 3. *Under Assumptions (A) to (D), the Gradient-Constrained MMD is given by*

$$S_{k,\mu,\lambda}(\mathbb{P},\mathbb{Q}) = \sqrt{\langle \eta, D_\lambda^{-1}\eta\rangle_{\mathcal{H}}}. \tag{7}$$

Computing (7) is in general intractable, as it involves inverting the infinite-dimensional operator D_λ. Proposition 4 moves towards a computable estimator by specializing to empirical measures for μ.

Proposition 4. *Let $\hat{\mu} = \sum_{m=1}^{M} \delta_{X_m}$ be an empirical measure of M points. Let $\eta(X) \in \mathbb{R}^M$ have mth entry $\eta(X_m)$, and $\nabla\eta(X) \in \mathbb{R}^{Md}$ have (m,i)th entry[3] $\partial_i \eta(X_m)$. Then under Assumptions (A) to (D), the Gradient-Constrained MMD is*

$$S^2_{k,\hat{\mu},\lambda}(\mathbb{P},\mathbb{Q}) = \frac{1}{\lambda}\left(\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) - \bar{P}(\nabla)\right) \tag{8}$$

$$\bar{P}(\eta) = \begin{bmatrix}\eta(X) \\ \nabla\eta(X)\end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix}K & G^{\mathsf{T}} \\ G & H\end{bmatrix} + M\lambda I_{M+Md}\right)^{-1} \begin{bmatrix}\eta(X) \\ \nabla\eta(X)\end{bmatrix}, \tag{9}$$

where K is the kernel matrix $K_{m,m'} = k(X_m, X_{m'})$, G is the matrix of left derivatives[4] $G_{(m,i),m'} = \partial_i k(X_m, X_{m'})$, and H that of derivatives of both arguments $H_{(m,i),(m',j)} = \partial_i \partial_{j+d} k(X_m, X_{m'})$.

[3] We use (m,i) to denote $(m-1)d + i$; thus $\nabla\eta(X)$ stacks $\nabla\eta(X_1), \ldots, \nabla\eta(X_M)$ into one vector.

[4] We use $\partial_i k(x,y)$ to denote the partial derivative with respect to x_i, and $\partial_{i+d} k(x,y)$ that for y_i.

Removing the kernel matrices in (9) would yield a penalty term similar to that of the critic gradient penalty: $\bar{P}(\eta) = \frac{1}{\lambda M}\sum_{m=1}^{M} \eta(X_m)^2 + \|\nabla\eta(X_m)\|^2$.

We estimate $S_{k,\hat{\mu},\lambda}$ from samples by plugging in an estimator $\hat{\eta}$ for η, which is a difference of sample means.This discrepancy indeed works well in practice: we demonstrate in Appendix D.2 that optimizing this estimator with an adversarially trained kernel network yields a good generative model on MNIST. But the linear system of size $M(1+d)$ in (9) is impractical: even on the 28×28 MNIST dataset and using a low-rank approximation, the model took days to converge. We therefore describe a faster approximation in the next section.

The criterion (4) relates to some divergences previously used in IGM training. The Fisher GAN [31] imposes only the variance constraint $\|f\|^2_{L^2((\mathbb{P}+\mathbb{Q}_\theta)/2)} \leq 1$. The Sobolev GAN [30] constrains $\|\nabla f\|^2_{L^2(\mu)} \leq 1$, along with a vanishing boundary condition on f to ensure a well-defined solution (although this was not used in the implementation, and can cause very unintuitive critic behavior; see Appendix B). The authors considered several choices of μ, including the WGAN-GP measure [16] and mixtures $(\mathbb{P}+\mathbb{Q}_\theta)/2$. Rather than enforcing their constraints in closed form, though, these models used additive regularization. We will compare to the Sobolev GAN in experiments.

3.2 Scaled Maximum Mean Discrepancy

We will now derive a lower bound on the Gradient-Constrained MMD which can be estimated in time linear in dimension d, and is therefore more suited for learning high-dimensional distributions, while retaining many of its attractive qualities.

Proposition 5. *Under Assumptions (A) to (D), the following inequality holds for all f in $\mathcal{H}$:*

$$\|f\|_{S(\mu),\lambda} \leq \sigma^{-1}_{k,\mu,\lambda}\|f\|_{\mathcal{H}}, \tag{10}$$

where $\sigma_{k,\mu,\lambda} = 1/\sqrt{\lambda + \int k(x,x)\mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \partial_i\partial_{i+d}k(x,x)\mu(\mathrm{d}x)}$.

We then define the Scaled Maximum Mean Discrepancy based on the bound of Proposition 5:

$$\mathrm{SMMD}_{k,\mu,\lambda}(\mathbb{P},\mathbb{Q}) := \sup_{\substack{f\in\mathcal{H}\\ \sigma^{-1}_{k,\mu,\lambda}\|f\|_{\mathcal{H}}\leq 1}} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)] = \sigma_{k,\mu,\lambda}\,\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}). \tag{11}$$

Because the constraint in the optimization of (11) is tighter than in that of (4), we have that $\mathrm{SMMD}_{k,\mu,\lambda}(\mathbb{P},\mathbb{Q}) \leq S_{k,\mu,\lambda}(\mathbb{P},\mathbb{Q})$. The Sobolev norm, and a fortiori the gradient norm under μ, is thus also controlled for the SMMD critic. We also show in Appendix D.1 that $\mathrm{SMMD}_{k,\mu,\lambda}$ behaves similarly to $S_{k,\mu,\lambda}$ on Gaussians.

We will need to estimate the scaling factor $\sigma_{k,\mu,\lambda}$. If μ is a probability distribution, an estimator based on $k(x,x)$ and $\partial_i\partial_{i+d}k(x,x)$ from samples is straightforward. But if $k_\psi = K\circ\phi_\psi$ and K is translation-invariant, $K(a,b) = g(-\|a-b\|^2)$, then $\sigma^{-2}_{k,\mu,\lambda} = \lambda + g(0) + 2g'(0)\,\mathbb{E}_\mu\left[\|\nabla\phi_\psi(X)\|^2\right]$. Or if K is linear, $K(a,b) = a^\mathsf{T}b$, then $\sigma^{-2}_{k,\mu,\lambda} = \lambda + \mathbb{E}_\mu\left[\|\phi_\psi(X)\|^2 + \|\nabla\phi_\psi(X)\|^2\right]$.

Of course, if μ and k are fixed, the SMMD is simply a constant times the MMD, and so behaves in essentially the same way as the MMD. But optimizing SMMD over a kernel family behaves quite differently from $\mathcal{D}_{\mathrm{MMD}}$ (2) for a given kernel family. Define the Optimized SMMD as

$$\mathcal{D}_{\mathrm{SMMD}}(\mathbb{P},\mathbb{Q}) := \sup_{\psi\in\Psi}\mathrm{SMMD}_{k_\psi,\mu,\lambda}(\mathbb{P},\mathbb{Q}). \tag{12}$$

Figure 1, right, shows the vector field for this loss in the simple example introduced in Section 2. The optimization surface is far more amenable: in particular the location B, which formerly had an extremely small gradient that made learning effectively impossible, now converges in very few steps.

Uniform bounds vs bounds in expectation Controlling $\|\nabla f_\psi\|^2_{L^2(\mu)} = \mathbb{E}_\mu\|\nabla f_\psi(X)\|^2$ does not necessarily imply a bound on $\|f\|_L = \sup_{x\in\mathcal{X}}\|\nabla f_\psi(X)\|$. Thus Proposition 2 doesn't necessarily apply to $\mathcal{D}_{\mathrm{SMMD}}$ with arbitrary kernel families. But it is far easier to have a tight control of $\|\nabla f_\psi\|^2_{L^2(\mu)}$ than of $\|f\|_L$, which is crucial in practice. If we bound $\|f\|_L$ with e.g. spectral normalization (SN) [29], we can significantly reduce the expressiveness of the parametric family. In

the example of Section 2, constraining $\|\phi_\psi\|_{\text{op}} = 1$ limits us to only $\phi(x) = x$. Thus $\mathcal{D}_{\text{MMD}}$ becomes simply the MMD with an RBF kernel of bandwidth 1, which has poor gradients when θ is far from 0 (Figure 1, right, blue). But for this kernel, where $\sigma_{k,\mu,\lambda}$ doesn't depend on μ, controlling the upper bound on $\|\nabla f_\psi\|^2_{L^2(\mu)}$ in (10) implies a bound on $\|f\|_L$, one tighter than obtained by constraining $\|\phi\|_L$. In general, the Cauchy-Schwartz bound (10) allows jointly adjusting the smoothness of k_ψ and the critic f, while SN would control the two smoothnesses independently. Likewise, limiting $\|\phi\|_L$ by limiting the Lipschitz norm of each layer could substantially reduce capacity, while $\|\nabla f_\psi\|_{L^2(\mu)}$ need not be decomposed by layer. Another advantage is that using e.g. $\mu = \mathbb{P}$ provides a data-dependent measure of complexity: we do not needlessly prevent ourselves from using critics whose poor behavior is only far away from any of the data points.

Spectral parametrization When the generator is near a local optimum, the critic might identify only one direction on which $\mathbb{Q}_\theta$ and $\mathbb{P}$ differ. If the generator parameterization is such that there is no local way for the generator to correct it, the critic may begin to single-mindedly focus on this difference, choosing redundant convolutional filters. If this occurs, the generator will be motivated to fix this single direction while ignoring all other aspects of the distributions. We can help avoid this collapse by using a critic parameterization amenable to diverse filters with higher-rank weight matrices. Miyato et al. [29] propose to parameterize the weight matrices as $W = \gamma \bar{W}/\|\bar{W}\|_{\text{op}}$, where $\|\bar{W}\|_{op}$ is the spectral norm of $\bar{W}$. This parametrization works well, particularly with $\mathcal{D}_{\text{SMMD}}$; Figure 2b shows the singular values of the second layer of a critic's network (and Figure 7, in the appendix, shows more layers). The conditioning of the weight matrix remains stable throughout training for the spectral parametrization, while it worsens through training in the default case.

4 Experiments

We evaluated unsupervised image generation on three datasets: CIFAR-10 [23] (60 000 images, 32×32), CelebA [26] (202 599 face images, resized and cropped to 160×160 as in [6]), and the more challenging ILSVRC2012 (ImageNet) dataset [38] (1 281 167 images, resized to 64×64). Code for all of these experiments is available at `github.com/MichaelArbel/Scaled-MMD-GAN`.

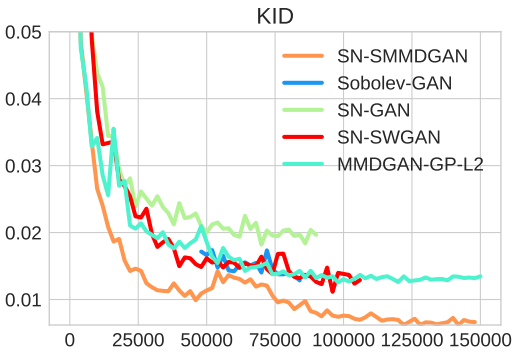
Losses All models are based on a scalar-output critic network $\phi_\psi : \mathcal{X} \to \mathbb{R}$. The WGAN and Sobolev GAN use a critic $f = \phi_\psi$, while the GAN uses a discriminator $D_\psi(x) = 1/(1 + \exp(-\phi_\psi(x)))$. The MMD-based methods use a kernel $k_\psi(x, y) = \exp(-(\phi_\psi(x) - \phi_\psi(y))^2/2)$. We also consider SMMD with a linear top-level kernel, $k(x, y) = \phi_\psi(x)\phi_\psi(y)$; because this becomes essentially identical to a WGAN (Appendix C), we refer to this method as SWGAN. SMMD, SWGAN, and Sobolev GAN use $\mu = \mathbb{P}$. We choose $\lambda = 0.1$ and scale the loss to obtain:

$$\text{SMMD: } \frac{\hat{\text{MMD}}^2_{k_\psi}(\mathbb{P}, \mathbb{Q}_\theta)}{1 + 10\,\mathbb{E}_{\hat{\mathbb{P}}}\left[\|\nabla\phi_\psi(X)\|^2\right]}, \quad \text{SWGAN: } \frac{\mathbb{E}_{\hat{\mathbb{P}}}\left[\phi_\psi(X)\right] - \mathbb{E}_{\hat{\mathbb{Q}}_\theta}\left[\phi_\psi(X)\right]}{\sqrt{1 + 10\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[|\phi_\psi(X)|^2\right] + \mathbb{E}_{\hat{\mathbb{P}}}\left[\|\nabla\phi_\psi(X)\|^2\right]\right)}}.$$

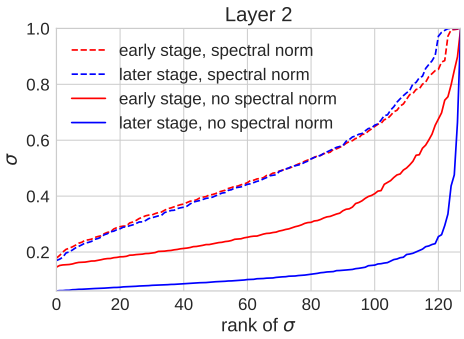
Architecture For CIFAR-10, we used the standard CNN architecture proposed by [29] with a 7-layer critic and a 4-layer generator. For CelebA, we used a 5-layer DCGAN discriminator and a 10-layer ResNet generator as in [6]. For ImageNet, we used a 10-layer ResNet for both the generator and discriminator. In all experiments we used 64 filters for the smallest convolutional layer, and double it at each layer.[5] The input codes for the generator are drawn from a 128 dimensional uniform. We consider two parameterizations for each critic: a standard one where the parameters can take any real value, and a spectral parametrization (denoted SN-) as above [29]. Models without explicit gradient control (SN-GAN, SN-MMDGAN, SN-MMGAN-L2, SN-WGAN) fix $\gamma = 1$; others learn γ.

Training All models were trained for 150 000 generator updates on a single GPU, except for ImageNet where the model was trained on 3 GPUs simultaneously. To avoid communication overheads we averaged the MMD estimate on each GPU, giving the block MMD estimator [48]. We always used 64 samples per GPU from each of $\mathbb{P}$ and $\mathbb{Q}$, and 5 critic updates per generator step. We used initial learning rates of 0.0001 for CIFAR-10 and CelebA, 0.0002 for ImageNet, and decayed these rates using the KID adaptive scheme of [6]. Every 2 000 steps, generator samples are compared to those from 20 000 steps ago; if the relative KID test [8] fails to show an improvement three consecutive times, the learning rate is decayed by 0.8. We used the Adam optimizer [22] with $\beta_1 = 0.5$, $\beta_2 = 0.9$.

[5] The number of filters in the 7-layer critic used for CIFAR10 was doubled every two layers, as in [29].



(a) Scores for CelebA through the training process.



(b) Singular values for two SMMD models.

Figure 2: The training process on CelebA.

Evaluation To compare the sample quality of different GAN models, we considered three different scores, all based on the Inception network [45] trained for classification on ImageNet dataset. The *Inception Score* (IS) [39] is based on the entropy of predicted labels; higher values are better. Though standard, this metric has many issues, particularly on datasets other than ImageNet [3, 6, 17]. The *FID* [17] instead measures the similarity between samples from the generator and the target as the Frechét distance between Gaussians fit to intermediate representations of samples in the Inception network. It is much more sensible than the Inception and becoming standard, but its estimator is strongly biased [6]. The *KID* [6] is similar to FID, but by using a polynomial-kernel MMD its estimates enjoy better statistical properties and are easier to compare. (A similar score was recommended by [18].) We use default parameters in the implementation of [6] for all scores.

Table 1: Mean (standard deviation) of score estimates, based on 50 000 samples from each model.

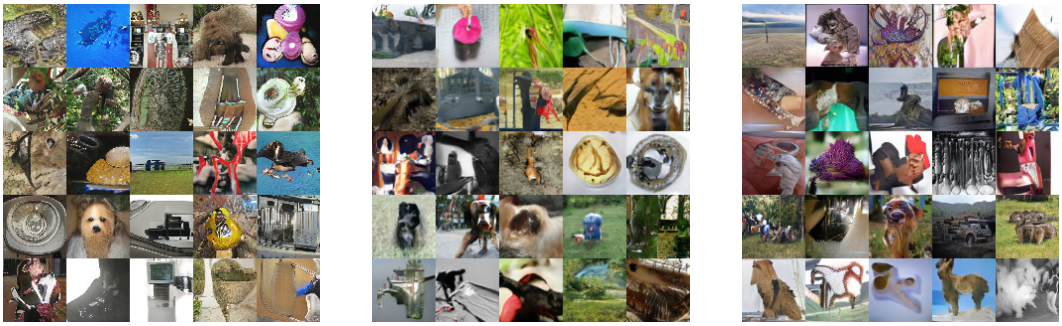
(a) CIFAR-10 and CelebA.

	CIFAR-10			CelebA		
Method	IS	FID	KID	IS	FID	KID
MMDGAN-GP-L2	7.55(0.12)	23.6(0.1)	0.014(0.0)	2.61(0.01)	20.55(0.25)	0.013(0.001)
WGAN-GP	6.68(0.06)	40.2(0.1)	0.024(0.001)	2.72(0.01)	29.24(0.22)	0.022(0.001)
SN-GAN	7.29(0.12)	26.1(0.1)	0.015(0.001)	2.64(0.03)	27.7(0.5)	0.018(0.001)
Sobolev-GAN	3.21(0.03)	132.2(0.1)	0.113(0.001)	2.84(0.05)	20.4(0.1)	0.014(0.001)
SMMDGAN	7.48(0.1)	25.1(0.3)	0.015(0.001)	2.83(0.04)	25.9(0.1)	0.019(0.001)
SN-SMMDGAN	7.43(0.11)	24.5(0.2)	0.015(0.001)	2.91(0.04)	12.2(0.1)	0.006(0.0)
SN-SWGAN	7.34(0.13)	27.9(0.4)	0.016(0.001)	2.86(0.05)	18.6(0.1)	0.011(0.0)

(b) ImageNet.

Method	IS	FID	KID
BGAN	10.70(0.46)	43.9(0.3)	0.047(0.0)
SN-GAN	11.16(0.13)	47.5(0.1)	0.044(0.0)
SN-SMMDGAN	10.91(0.15)	36.6(0.2)	0.035(0.0)

Results Table 1a presents the scores for models trained on both CIFAR-10 and CelebA datasets. On CIFAR-10, SMMD-GAN and SN-SMMD-GAN performed comparably to MMD GAN-GP and SN-GAN. But on CelebA, SN-SMMD GAN dramatically outperformed the other methods with the same architecture in all three metrics. It also trained faster (Figure 2a). It is also worth noting that SN-SWGAN far outperformed WGAN-GP on both datasets. Interestingly, Sobolev GAN did quite well on CelebA but terribly on CIFAR-10. Table 1b presents the scores for SN-SMMD GAN trained



(a) Scaled MMD GAN with SN (b) SN-GAN (c) Boundary Seeking GAN

(d) Scaled MMD GAN with SN (e) Scaled WGAN with SN (f) MMD GAN with GP+L2

Figure 3: Samples from various models. Top: 64×64 ImageNet; bottom: 160×160 CelebA.

on ImageNet, and the scores of pre-trained models using BGAN [5] and SN-GAN [29].[6] Our model substantially outperformed both methods in terms of FID and KID scores, and is similar to SN-GAN in Inception score. Figure 3 shows samples on ImageNet and CelebA; Appendix D.4 has more.

Spectrally normalized WGANs / MMDGANs To control for the contribution of the spectral parametrization to the performance, we evaluated (in Table 2, Appendix D.3) variants of MMDGANs and WGANs using spectral normalization. WGAN and MMDGAN with spectral normalization, where the parameter γ is fixed to 1, led to unstable training and didn't converge at all (Figure 8) despite many attempts (in Table 2, Appendix D.3) . The gradient control due to SN is thus probably too loose for MMD GAN and WGAN. We also considered variants of these models with a learned γ while also adding a gradient penalty and an L_2 penalty on critic activations [6, footnote 19]. These also generally gave poor results – except for SN-MMDGAN-GP-L2, which was excellent, but not substantially improved over MMDGAN-GP-L2. We ran the same experiments on CelebA, but aborted the runs early when it became clear that training was not successful (including for SN-MMDGAN-GP-L2).

Rank collapse We occasionally observed the failure mode for SMMD where the critic becomes low-rank, discussed in Section 3.2, especially on CelebA; this failure was obvious even in the training objective. Figure 2b (later stage) is one of these examples. Spectral parametrization seemed to prevent this behavior. We also found one could avoid collapse by reverting to an earlier checkpoint and increasing the RKHS regularization parameter λ. We did not do this for any of the experiments here, but did use $1 + 0.1\,\mathbb{E}_{\mathbb{P}}\|\nabla\phi_\psi(X)\|^2$ in the SMMDGAN loss on CelebA to avoid collapse.

5 Conclusion

We studied gradient regularization for MMD-based critics in implicit generative models, clarifying how previous techniques relate to the $\mathcal{D}_{\mathrm{MMD}}$ loss. Based on these insights, we proposed the Gradient-Constrained MMD and its approximation the Scaled MMD, a new loss function for IGMs that controls gradient behavior in a principled way and obtains excellent performance in practice.

[6] These models are courtesy of the respective authors and also trained at 64×64 resolution. SN-GAN used the same architecture as our model, but trained for 250 000 generator iterations; BS-GAN used a similar 5-layer ResNet architecture and trained for 74 epochs, comparable to SN-GAN.

One interesting area of future study for these distances is their behavior when used to diffuse particles distributed as $\mathbb{Q}$ towards particles distributed as $\mathbb{P}$. Mroueh et al. [30, Appendix A.1] began such a study for the Sobolev GAN loss; [32] proved convergence and studied discrete-time approximations.

Another area to explore is the geometry of these losses, as studied by Bottou et al. [7], who showed potential advantages of the Wasserstein geometry over the MMD. Their results, though, do not address $\mathcal{D}_{\mathrm{MMD}}$; $S_{k,\mu,\lambda}$ and $\mathcal{D}_{\mathrm{SMMD}}$'s geometry is potentially even more different.

References

[1] M. Arjovsky and L. Bottou. "Towards Principled Methods for Training Generative Adversarial Networks." In: *ICLR*. 2017. arXiv: 1701.04862.

[2] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks." In: *ICML*. 2017. arXiv: 1701.07875.

[3] S. Barratt and R. Sharma. *A Note on the Inception Score*. 2018. arXiv: 1801.01973.

[4] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. *The Cramer Distance as a Solution to Biased Wasserstein Gradients*. 2017. arXiv: 1705.10743.

[5] D. Berthelot, T. Schumm, and L. Metz. *BEGAN: Boundary Equilibrium Generative Adversarial Networks*. 2017. arXiv: 1703.10717.

[6] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. "Demystifying MMD GANs." In: *ICLR*. 2018. arXiv: 1801.01401.

[7] L. Bottou, M. Arjovsky, D. Lopez-Paz, and M. Oquab. *Geometrical Insights for Implicit Generative Modeling*. 2017. arXiv: 1712.07822.

[8] W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. "A Test of Relative Similarity For Model Selection in Generative Models." In: *ICLR*. 2016. arXiv: 1511.04581.

[9] O. Bousquet, O. Chapelle, and M. Hein. "Measure Based Regularization." In: *NIPS*. 2004.

[10] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge University Press, 2002.

[11] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. "Training generative neural networks via Maximum Mean Discrepancy optimization." In: *UAI*. 2015. arXiv: 1505.03906.

[12] A. Genevay, G. Peyré, and M. Cuturi. "Learning Generative Models with Sinkhorn Divergences." In: *AISTATS*. 2018. arXiv: 1706.00292.

[13] T. Gneiting and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation." In: *JASA* 102.477 (2007), pp. 359–378.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets." In: *NIPS*. 2014. arXiv: 1406.2661.

[15] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. "A Kernel Two-Sample Test." In: *JMLR* 13 (2012).

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. "Improved Training of Wasserstein GANs." In: *NIPS*. 2017. arXiv: 1704.00028.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium." In: *NIPS*. 2017. arXiv: 1706.08500.

[18] G. Huang, Y. Yuan, Q. Xu, C. Guo, Y. Sun, F. Wu, and K. Weinberger. *An empirical study on evaluation metrics of generative adversarial networks*. 2018. URL: https://openreview.net/forum?id=Sy1f0e-R-.

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. *Multimodal Unsupervised Image-to-Image Translation*. 2018. arXiv: 1804.04732.

[20] Y. Jin, K. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang. *Towards the Automatic Anime Characters Creation with Generative Adversarial Networks*. 2017. arXiv: 1708.05509.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." In: *ICLR*. 2018. arXiv: 1710.10196.

[22] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *ICLR*. 2015. arXiv: 1412.6980.

[23] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. 2009.

[24] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. "MMD GAN: Towards Deeper Understanding of Moment Matching Network." In: *NIPS*. 2017. arXiv: 1705.08584.

[25] Y. Li, K. Swersky, and R. Zemel. "Generative Moment Matching Networks." In: *ICML*. 2015. arXiv: 1502.02761.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep learning face attributes in the wild." In: *ICCV*. 2015.

[27] L. Mescheder, A. Geiger, and S. Nowozin. *Which Training Methods for GANs do actually Converge?* 2018. arXiv: 1801.04406.

[28] P. Milgrom and I. Segal. "Envelope theorems for arbitrary choice sets." In: *Econometrica* 70.2 (2002), pp. 583–601.

[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks." In: *ICLR*. 2018. arXiv: 1802.05927.

[30] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. "Sobolev GAN." In: *ICLR*. 2018.

[31] Y. Mroueh and T. Sercu. "Fisher GAN." In: *NIPS*. 2017. arXiv: 1705.09675.

[32] Y. Mroueh, T. Sercu, and A. Raj. *Sobolev Descent: Variational Transport of Distributions via Advection*. Private communication. Apr. 2018.

[33] A. Müller. "Integral Probability Metrics and their Generating Classes of Functions." In: *Advances in Applied Probability* 29.2 (1997), pp. 429–443.

[34] S. Nowozin, B. Cseke, and R. Tomioka. "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization." In: *NIPS*. 2016. arXiv: 1606.00709.

[35] A. Radford, L. Metz, and S. Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." In: *ICLR*. 2016. arXiv: 1511.06434.

[36] J. R. Retherford. "Review: J. Diestel and J. J. Uhl, Jr., Vector measures." In: *Bull. Amer. Math. Soc.* 84.4 (July 1978), pp. 681–685.

[37] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. "Stabilizing Training of Generative Adversarial Networks through Regularization." In: *NIPS*. 2017. arXiv: 1705.09367.

[38] O. Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. 2014. arXiv: 1409.0575.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. "Improved Techniques for Training GANs." In: *NIPS*. 2016. arXiv: 1606.03498.

[40] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[41] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. "Kernel choice and classifiability for RKHS embeddings of probability distributions." In: *NIPS*. 2009.

[42] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." In: *JMLR* 12 (2011), pp. 2389–2410. arXiv: 1003.0887.

[43] B. Sriperumbudur. "On the optimal estimation of probability mesaures in weak and strong topologies." In: *Bernoulli* 22.3 (2016), pp. 1839–1893. arXiv: 1310.8240.

[44] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy." In: *ICLR*. 2017. arXiv: 1611.04488.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the Inception Architecture for Computer Vision." In: *CVPR*. 2016. arXiv: 1512.00567.

[46] J. Weed and F. Bach. *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*. 2017. arXiv: 1707.00087.

[47] H. Wendland. *Scattered Data Approximation*. Cambridge, UK: Cambridge University Press, 2005.

[48] W. Zaremba, A. Gretton, and M. B. Blaschko. "B-tests: Low Variance Kernel Two-Sample Tests." In: *NIPS*. 2013. arXiv: 1307.1954.

A Proofs

We first review some basic properties of the Reproducing Kernel Hilbert Spaces. We consider here a separable RKHS $\mathcal{H}$ with basis $(e_i)_{i\in I}$, where I is either finite if $\mathcal{H}$ is finite-dimensional, or $I = \mathbb{N}$ otherwise. We also assume that the reproducing kernel k is continuously twice differentiable. Then the following reproducing properties hold for any given function f in $\mathcal{H}$:

$$f(x) = \langle f, k(x, .)\rangle_{\mathcal{H}} \tag{13}$$
$$\partial_i f(x) = \langle f, \partial_i k(x, .)\rangle_{\mathcal{H}} \tag{14}$$

We say that an operator $A : \mathcal{H} \mapsto \mathcal{H}$ is Hilbert-Schmidt if $\|A\|_{HS}^2 = \sum_{i\in I} \|Ae_i\|_{\mathcal{H}}^2$ is finite. $\|A\|_{\mathrm{HS}}$ is called the Hilbert-Schmidt norm of A. The space of Hilbert-Schmidt operators itself a Hilbert space with the inner product $\langle A, B\rangle_{HS} = \sum_{i\in I}\langle Ae_i, Be_i\rangle_{\mathcal{H}}$. Moreover, we say that an operator A is trace-class if its trace norm is finite, i.e. $\|A\|_1 = \sum_{i\in I}\langle e_i, (A^*A)^{\frac{1}{2}} e_i\rangle_{\mathcal{H}} < \infty$. The outer product $f \otimes g$ for $f, g \in \mathcal{H}$ gives an $\mathcal{H} \to \mathcal{H}$ operator such that $(f \otimes g)v = \langle g, v\rangle_{\mathcal{H}} f$ for all v in $\mathcal{H}$.

Given two vectors f and g in $\mathcal{H}$ and a Hilbert-Schmidt operator A we have the following properties:

1. The outer product $f \otimes g$ is a Hilbert-Schmidt operator with Hilbert-Schmidt norm given by: $\|f \otimes g\|_{\mathrm{HS}} = \|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}$.
2. The inner product between two rank-one operators $f \otimes g$ and $u \otimes v$ is $\langle f \otimes g, u \otimes v\rangle_{\mathrm{HS}} = \langle f, u\rangle_{\mathcal{H}}\langle g, v\rangle_{\mathcal{H}}$.
3. The following identity holds: $\langle f, Ag\rangle_{\mathcal{H}} = \langle f \otimes g, A\rangle_{\mathrm{HS}}$.

We will need the following assumptions about the distributions $\mathbb{P}$ and $\mathbb{Q}$, the measure μ, and the kernel k:

- **(A)** $\mathbb{P}$ and $\mathbb{Q}$ have integrable first moments.
- **(B)** $\sqrt{k(x, x)}$ grows at most linearly in x: for all x in $\mathcal{X}$, $\sqrt{k(x, x)} \leq C(\|x\| + 1)$
- **(C)** The kernel k is twice continuously differentiable.
- **(D)** The functions $x \mapsto k(x, x)$ and $x \mapsto \partial_i\partial_{i+d}k(x, x)$ for $1 \leq i \leq d$ are μ-integrable.

We are now ready to begin proving things.

Proof of Proposition 3. Let f be a function in $\mathcal{H}$. We will first express the squared λ-regularized Sobolev norm of f (5) as a quadratic form in $\mathcal{H}$. Recalling the reproducing properties of (13) and (14), we have:

$$\|f\|_{S(\mu),\lambda}^2 = \int \langle f, k(x, \cdot)\rangle_{\mathcal{H}}^2 \, \mu(\mathrm{d}x) + \sum_{i=1}^{d} \int \langle f, \partial_i k(x, \cdot)\rangle_{\mathcal{H}}^2 \, \mu(\mathrm{d}x) + \lambda\|f\|_{\mathcal{H}}^2.$$

Now introduce the operator $D_x = k(x, \cdot) \otimes k(x, \cdot) + \sum_{i=1}^{d} \partial_i k(x, \cdot) \otimes \partial_i k(x, \cdot)$; note that the operator D (6) is $D = \mathbb{E}_\mu \, D_X$. Using 2, one further gets

$$\|f\|_{S(\mu),\lambda}^2 = \int \langle f \otimes f, D_x\rangle_{\mathrm{HS}} \, \mu(\mathrm{d}x) + \lambda\|f\|_{\mathcal{H}}^2.$$

Under Assumption (D), and using Lemma 1, one can take the integral inside the inner product, which leads to $\|f\|_{S(\mu),\lambda}^2 = \langle f \otimes f, D\rangle_{\mathrm{HS}} + \lambda\|f\|_{\mathcal{H}}^2$. Finally, using 3 it follows that

$$\|f\|_{S(\mu),\lambda}^2 = \langle f, D_\lambda f\rangle_{\mathcal{H}}.$$

Under Assumptions (A) and (B), 1 applies, and it follows that $k(x, \cdot)$ is also Bochner integrable under $\mathbb{P}$ and $\mathbb{Q}$. Thus

$$\mathbb{E}_{\mathbb{P}}\left[\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right] - \mathbb{E}_{\mathbb{Q}}\left[\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right] = \langle f, \mathbb{E}_{\mathbb{P}}\left[k(x, \cdot)\right] - \mathbb{E}_{\mathbb{P}}\left[k(x, \cdot)\right]\rangle_{\mathcal{H}} = \langle f, \eta\rangle_{\mathcal{H}},$$

where η is defined as this difference in mean embeddings.

Since D_λ is symmetric positive definite, its square-root $D_\lambda^{\frac{1}{2}}$ is well-defined and is also invertible. For any $f \in \mathcal{H}$, let $g = D_\lambda^{\frac{1}{2}} f$, so that $\langle f, D_\lambda f \rangle_\mathcal{H} = \|g\|_\mathcal{H}^2$. Note that for any $g \in \mathcal{H}$, there is a corresponding $f = D_\lambda^{-\frac{1}{2}} g$. Thus we can re-express the maximization problem in (4) in terms of g:

$$\begin{aligned} S_{k,\mu,\lambda}(\mathbb{P}, \mathbb{Q}) &:= \sup_{\substack{f \in \mathcal{H} \\ \langle f, D_\lambda f \rangle_\mathcal{H} \leq 1}} \langle f, \eta \rangle_\mathcal{H} = \sup_{\substack{g \in \mathcal{H} \\ \|g\|_\mathcal{H} \leq 1}} \langle D_\lambda^{-\frac{1}{2}} g, \eta \rangle_\mathcal{H} \\ &= \sup_{\substack{g \in \mathcal{H} \\ \|g\|_\mathcal{H} \leq 1}} \langle g, D_\lambda^{-\frac{1}{2}} \eta \rangle_\mathcal{H} = \|D_\lambda^{-\frac{1}{2}} \eta\|_\mathcal{H} = \sqrt{\langle \eta, D_\lambda^{-1} \eta \rangle_\mathcal{H}}. \end{aligned}$$ □

Proof of Proposition 4. Let $g \in \mathcal{H}$ be the solution to the regression problem $D_\lambda g = \eta$:

$$\begin{aligned} \frac{1}{M} \sum_{m=1}^{M} \left[g(X_m) k(X_m, \cdot) + \sum_{i=1}^{d} \partial_i g(X_m) \partial_i k(X_m, \cdot) \right] + \lambda g &= \eta \\ g = \frac{1}{\lambda} \eta - \frac{1}{\lambda M} \sum_{m=1}^{M} \left[g(X_m) k(X_m, \cdot) + \sum_{i=1}^{d} \partial_i g(X_m) \partial_i k(X_m, \cdot) \right]. \end{aligned} \tag{15}$$

Taking the inner product of both sides of (15) with $k(X_{m'}, \cdot)$ for each $1 \leq m' \leq M$ yields the following M equations:

$$g(X_{m'}) = \frac{1}{\lambda} \eta(X_{m'}) - \frac{1}{\lambda M} \sum_{m=1}^{M} \left[g(X_m) K_{m,m'} + \sum_{i=1}^{d} \partial_i g(X_m)\, G_{(m,i),m'} \right]. \tag{16}$$

Doing the same with $\partial_j k(X_{m'}, \cdot)$ gives Md equations:

$$\partial_j g(X_{m'}) = \frac{1}{\lambda} \partial_j \eta(X_{m'}) - \frac{1}{\lambda M} \sum_{m=1}^{M} \left[g(X_m) G_{(m',j),m} + \sum_{i=1}^{d} \partial_i g(X_m) H_{(m,i),(m',j)} \right]. \tag{17}$$

From (15), it is clear that g is a linear combination of the form:

$$g(x) = \frac{1}{\lambda} \eta(x) - \frac{1}{\lambda M} \sum_{m=1}^{M} \left[\alpha_m k(X_m, x) + \sum_{i=1}^{d} \beta_{m,i} \partial_i k(X_m, x) \right],$$

where the coefficients $\alpha := (\alpha_m = g(X_m))_{1 \leq m \leq M}$ and $\beta := (\beta_{m,i} = \partial_i g(X_m))_{\substack{1 \leq m \leq M \\ 1 \leq i \leq d}}$ satisfy the system of equations (16) and (17). We can rewrite this system as

$$\begin{bmatrix} K + M\lambda I_M & G^\intercal \\ G & H + M\lambda I_{Md} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = M \begin{bmatrix} \eta(X) \\ \nabla \eta(X) \end{bmatrix},$$

where I_M, I_{Md} are the identity matrices of dimension M, Md. Since K and H must be positive semidefinite, an inverse exists. We conclude by noticing that

$$S_{k,\hat{\mu},\lambda}(\mathbb{P}, \mathbb{Q})^2 = \langle \eta, g \rangle_\mathcal{H} = \frac{1}{\lambda} \|\eta\|_\mathcal{H}^2 - \frac{1}{\lambda M} \sum_{m=1}^{M} \left[\alpha_m \eta(X_m) + \sum_{i=1}^{d} \beta_{m,i} \partial_i \eta(X_m) \right].$$ □

As an aside, the following form is perhaps also of interest. Let $T : \mathcal{H} \to \mathbb{R}^{M+Md}$ be the linear operator

$$T = \sum_{m=1}^{M} \bar{e}_m \otimes k(X_m, \cdot) + \sum_{m=1}^{M} \sum_{i=1}^{d} \bar{e}_{m+(m,i)} \otimes \partial_i k(X_m, \cdot),$$

where $\bar{e}_i$ is the ith standard basis vector for $\mathbb{R}^{M+Md}$. Then $\begin{bmatrix} \eta(X) \\ \nabla \eta(X) \end{bmatrix} = T\eta$, and $\begin{bmatrix} K & G^\intercal \\ G & H \end{bmatrix} = TT^*$. Thus (8) can be written

$$S_{k,\hat{\mu},\lambda}^2 = \frac{1}{\lambda} \left\langle \eta, \left(I - T^* (TT^* + M\lambda I)^{-1} T \right) \eta \right\rangle_\mathcal{H}.$$

Proof of Proposition 5. The key idea here is to use the Cauchy-Schwarz inequality for the Hilbert-Schmidt inner product. Letting $f \in \mathcal{H}$, $\|f\|^2_{S(\mu),\lambda}$ is

$$\int f(x)^2\mu(dx) + \int \|\nabla f(x)\|^2\mu(dx) + \lambda\|f\|^2_{\mathcal{H}}$$

$$\overset{(a)}{=} \int \langle f, k(x,\cdot)\otimes k(x,\cdot)f\rangle_{\mathcal{H}}\mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \langle f, \partial_i k(x,\cdot)\otimes \partial_i k(x,\cdot)f\rangle_{\mathcal{H}}\mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}}$$

$$\overset{(b)}{=} \int \langle f\otimes f, k(x,\cdot)\otimes k(x,\cdot)\rangle_{\mathrm{HS}}\,\mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \langle f\otimes f, \partial_i k(x,\cdot)\otimes \partial_i k(x,\cdot)\rangle_{\mathrm{HS}}\,\mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}}$$

$$\overset{(c)}{\leq} \|f\|^2_{\mathcal{H}}\left[\int k(x,x)\mu(dx) + \sum_{i=1}^{d}\int \partial_i\partial_{i+d}k(x,x)\mu(dx) + \lambda\right].$$

(a) follows from the reproducing properties (13) and (14) and relation 2. (b) is obtained using 3, while (c) follows from the Cauchy-Schwarz inequality and 1. □

Lemma 1. *Under Assumption (D), D_x is Bochner integrable and its integral D is a trace-class symmetric positive semi-definite operator with $D_\lambda = D + \lambda I$ invertible for any positive λ. Moreover, for any Hilbert-Schmidt operator A we have: $\langle A, D\rangle_{HS} = \int \langle A, D_x\rangle_{HS}\mu(dx)$.*

Under Assumptions (A) and (B), $k(x,\cdot)$ is Bochner integrable with respect to any probability distribution $\mathbb{P}$ with finite first moment and the following relation holds: $\langle f, \mathbb{E}_{\mathbb{P}}[k(x,\cdot)]\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[\langle f, k(x,\cdot)\rangle_{\mathcal{H}}]$ for all f in $\mathcal{H}$.

Proof. The operator D_x is positive self-adjoint. It is also trace-class, as by the triangle inequality

$$\|D_x\|_1 \leq \|k(x,\cdot)\otimes k(x,\cdot)\|_1 + \sum_{i=1}^{d}\|\partial_i k(x,\cdot)\otimes\partial_i k(x,\cdot)\|_1$$

$$= \|k(x,\cdot)\|^2_{\mathcal{H}} + \sum_{i=1}^{d}\|\partial_i k(x,\cdot)\|^2_{\mathcal{H}} < \infty.$$

By Assumption (D), we have that $\int \|D_x\|_1\mu(dx) < \infty$ which implies that D_x is μ-integrable in the Bochner sense [36, Definition 1 and Theorem 2]. Its integral D is trace-class and satisfies $\|D\|_1 \leq \int \|D_x\|_1\mu(\mathrm{d}x)$. This allows to have $\langle A, D\rangle_{HS} = int\langle A, D_x\rangle_{HS}\mu(dx)$ for all Hilbert-Schmidt operator A. Moreover, the integral preserves the symmetry and positivity. It follows that $\mathcal{D}_\lambda$ is invertible. The Bochner integrability of $k(x,.)$ under a distribution $\mathbb{P}$ with finite moment follows directly from Assumptions (A) and (B), since $\int \|k(x,.)\|\mathbb{P}(dx) \leq C\int(\|x\|+1)\mathbb{P}(dx) < \infty$. This allows to write $\langle f, \mathbb{E}_{\mathbb{P}}[k(x,.)]\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[\langle f, k(x,.)\rangle_{\mathcal{H}}]$. □

B Vector fields of Gradient-Constrained MMD and Sobolev GAN critics

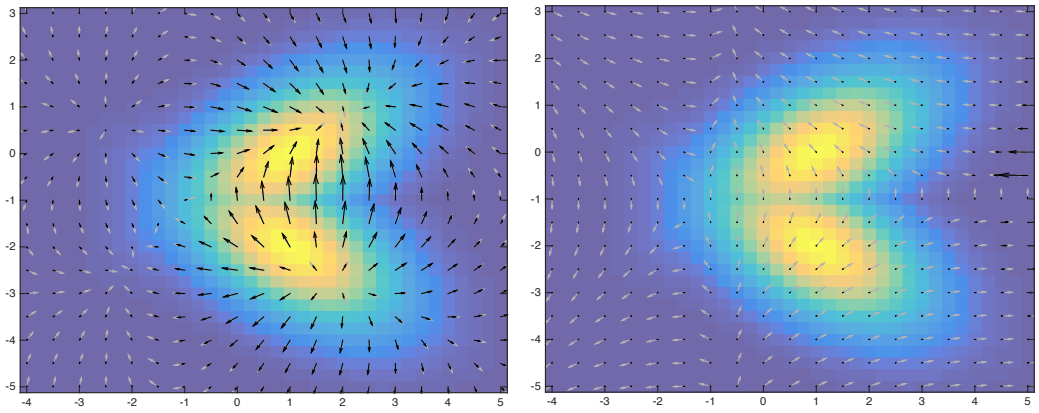
Mroueh et al. [30] argue that *the gradient of the critic (...) defines a transportation plan for moving the distribution mass* (from generated to reference distribution) and present the solution of Sobolev PDE for 2-dimensional Gaussians. We observed that in this simple example the gradient of the Sobolev critic can be very high outside of the areas of high density, which is not the case with the Gradient-Constrained MMD. Figure 4 presents critic gradients in both cases, using $\mu = (\mathbb{P}+\mathbb{Q})/2$ for both.

This unintuitive behavior is most likely related to the vanishing boundary condition, assumed by Sobolev GAN. Solving the actual Sobolev PDE, we found that the Sobolev critic has very high gradients close to the boundary in order to match the condition; moreover, these gradeints point to the direction opposite direction to the reference distribution.

C Near-equivalence of WGAN and linear-kernel MMD GANs

For an MMD GAN-GP with kernel $k(x,y) = \phi(x)\phi(y)$, we have that

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}\,\phi(x) - \mathbb{E}_{\mathbb{Q}}\,\phi(Y)|$$



(a) Gradient-Constrained MMD critic gradient. (b) Sobolev IPM critic gradient.

Figure 4: Vector fields of critic gradients between two Gaussians. The grey arrows show normalized gradients, i.e. gradient directions, while the black ones are the actual gradients. Note that for the Sobolev critic, gradients are magnitudes higher on the right hand side of the plot than in the areas of high density of the given distributions.

and that the critic function is

$$\frac{\eta(t)}{\|\eta\|_{\mathcal{H}}} = \frac{\mathbb{E}_{X\sim\mathbb{P}}\,\phi(X)\phi(t) - \mathbb{E}_{Y\sim\mathbb{Q}}\,\phi(Y)\phi(t)}{|\mathbb{E}_{\mathbb{P}}\,\phi(X) - \mathbb{E}_{\mathbb{Q}}\,\phi(Y)|} = \operatorname{sign}\left(\mathbb{E}_{X\sim\mathbb{P}}\,\phi(X) - \mathbb{E}_{Y\sim\mathbb{Q}}\,\phi(Y)\right)\phi(t).$$

Thus if we assume $\mathbb{E}_{X\sim\mathbb{P}}\,\phi(X) > \mathbb{E}_{Y\sim\mathbb{Q}}\,\phi(Y)$, as that is the goal of our critic training, we see that the MMD becomes identical to the WGAN loss, and the gradient penalty is applied to the same function.

(MMD GANs, however, would typically train on the unbiased estimator of MMD^2, giving a very slightly different loss function. [6] also applied the gradient penalty to η rather than the true critic $\eta/\|\eta\|$.)

The SMMD with a linear kernel is thus analogous to applying the scaling operator to a WGAN; hence the name SWGAN.

D Additional experiments

D.1 Comparison of Gradient-Constrained MMD to Scaled MMD

Figure 5 shows the behavior of the MMD, the Gradient-Constrained SMMD, and the Scaled MMD when comparing Gaussian distributions. We can see that $\mathrm{MMD} \propto \mathrm{SMMD}$ and the Gradient-Constrained MMD behave similarly in this case, and that optimizing the SMMD and the Gradient-Constrained MMD is also similar. Optimizing the MMD would yield essentially one everywhere.

D.2 IGMs with Optimized Gradient-Constrained MMD loss

We implemented the estimator of Proposition 4 using the empirical mean estimator of η, and sharing samples for $\mu = \mathbb{P}$. To handle the large but approximately low-rank matrix system in (9), we used an incomplete Cholesky decomposition [40, Algorithm 5.12] to obtain $R \in \mathbb{R}^{\ell\times M(1+d)}$ such that $\begin{bmatrix} K & G^\mathsf{T} \\ G & H \end{bmatrix} \approx R^\mathsf{T} R$. Then the Woodbury matrix identity allows an efficient evaluation of (9):

$$\left(R^\mathsf{T}R + M\lambda I\right)^{-1} = \frac{1}{M\lambda}\left(I - R(RR^\mathsf{T} + M\lambda I)^{-1}R\right).$$

Even though only a small ℓ is required for a good approximation, and the full matrices K, G, and H need never be constructed, backpropagation through this procedure is slow and not especially GPU-friendly; training on CPU was faster. Thus we were only able to run the estimator on MNIST, and even that took days to conduct the optimization on powerful workstations.

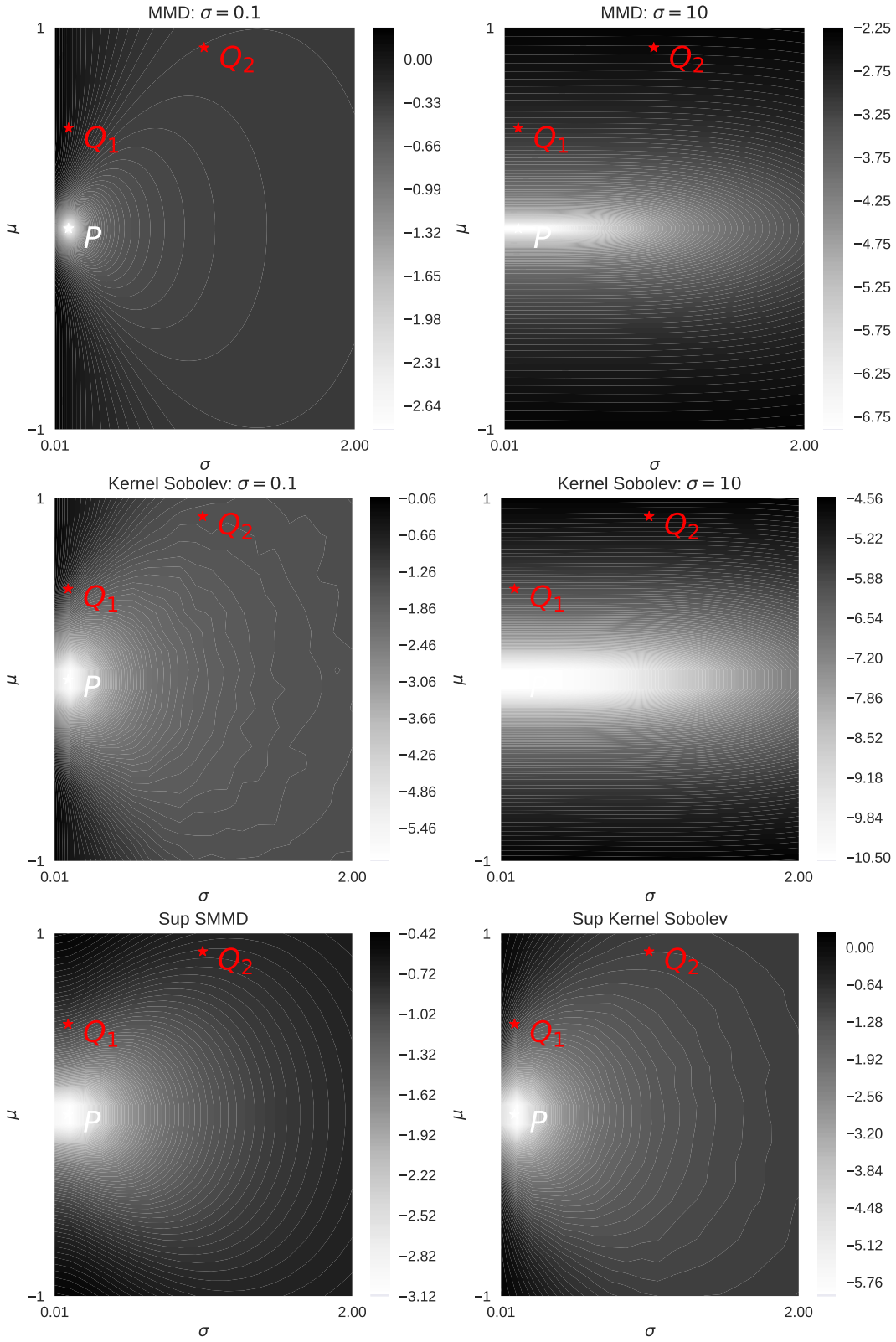
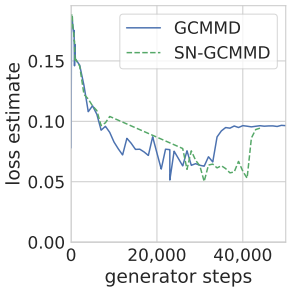


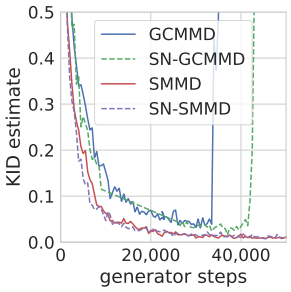
Figure 5: Plots of various distances between one dimensional Gaussians, where $P = \mathcal{N}(0, 0.1^2)$, and the colors show $\log \mathcal{D}(P, \mathcal{N}(\mu, \sigma^2))$. All distances use $\lambda = 1$. Top left: MMD with a Gaussian kernel of bandwidth $\psi = 0.1$. Top right: MMD with bandwidth $\psi = 10$. Middle left: Gradient-Constrained MMD with bandwidth $\psi = 0.1$. Middle right: Gradient-Constrained MMD with bandwidth $\psi = 10$. Bottom left: Optimized SSMD, allowing any $\psi \in \mathbb{R}$. Bottom right: Optimized Gradient-Constrained MMD.

(a) Without spectral normalization; 32 000 generator iterations.

(b) With spectral normalization; 41 000 generator iterations.



(c) Generator losses.



(d) KID scores.

Figure 6: The MNIST models with Optimized Gradient-Constrained MMD loss.

The learned models, however, were reasonable. Using a DCGAN architecture, batches of size 64, and a procedure that otherwise agreed with the setup of Section 4, samples with and without spectral normalization are shown in Figures 6a and 6b. After the points in training shown, however, the same rank collapse as discussed in Section 4 occurred. Here it seems that spectral normalization may have delayed the collapse, but not prevented it. Figure 6c shows generator loss estimates through training, including the obvious peak at collapse; Figure 6d shows KID scores based on the MNIST-trained convnet representation [6], including comparable SMMD models for context. The fact that SMMD models converged somewhat faster than Gradient-Constrained MMD models here may be more related to properties of the estimator (8) rather than the distances; more work would be needed to fully compare the behavior of the two distances.

D.3 Spectral normalization and Scaled MMD

Figure 7 shows the distribution of critic weight singular values, like Figure 2b, at more layers. Figure 8 and Table 2 show results for the spectral normalization variants considered in the experiments. MMDGAN, with neither spectral normalization nor a gradient penalty, did surprisingly well in this case, though it fails badly in other situations.

Figure 7 compares the decay of singular values for layer of the critic's network at both early and later stages of training in two cases: with or without the spectral parametrization. The model was trained on CelebA using SMMD. Figure 8 shows the evolution per iteration of Inception score, FID and KID for Sobolev-GAN, MMDGAN and variants of MMDGAN and WGAN using spectral normalization. It is often the case that this parametrization alone is not enough to achieve good results.

Table 2: Mean (standard deviation) of score evaluations on CIFAR-10 for different methods using Spectral Normalization.

Method	IS	FID	KID
MMDGAN	7.01(0.12)	30.9(0.3)	0.02(0.001)
SN-MMDGAN	5.23(0.06)	69.0(0.2)	0.046(0.002)
SN-WGAN	2.95(0.03)	162.3(0.2)	0.125(0.002)
SN-MMDGAN-GP	5.02(0.06)	137.0(0.1)	0.084(0.002)
SN-WGAN-GP	4.49(0.1)	73.1(0.1)	0.058(0.001)
SN-MMDGAN-GP-L2	7.43(0.06)	23.4(0.0)	0.015(0.001)
SN-MMDGAN-L2	4.46(0.07)	67.0(0.0)	0.053(0.002)

D.4 Additional samples

Figures 9 and 10 give extra samples from the models.

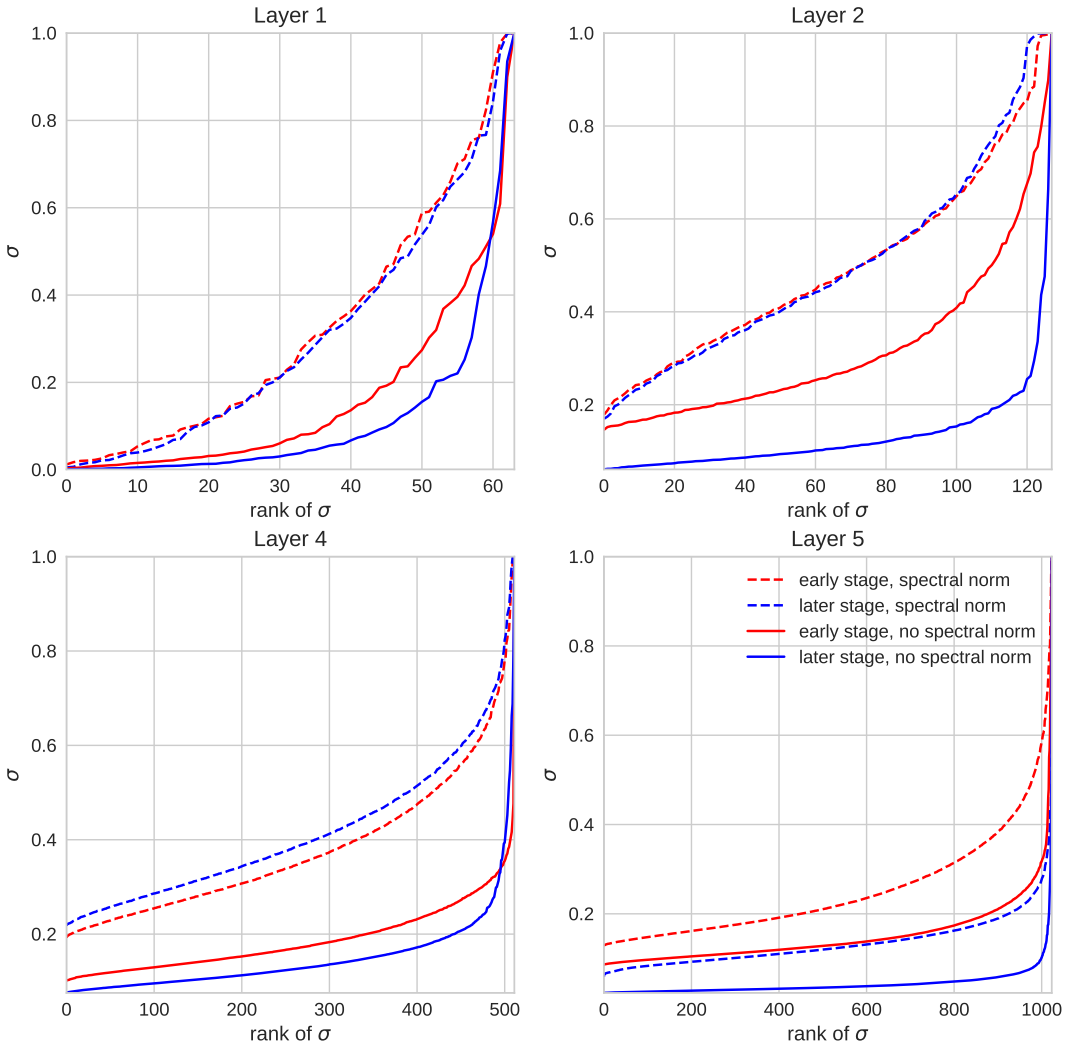


Figure 7: Singular values at different layers, for the same setup as Figure 2b.

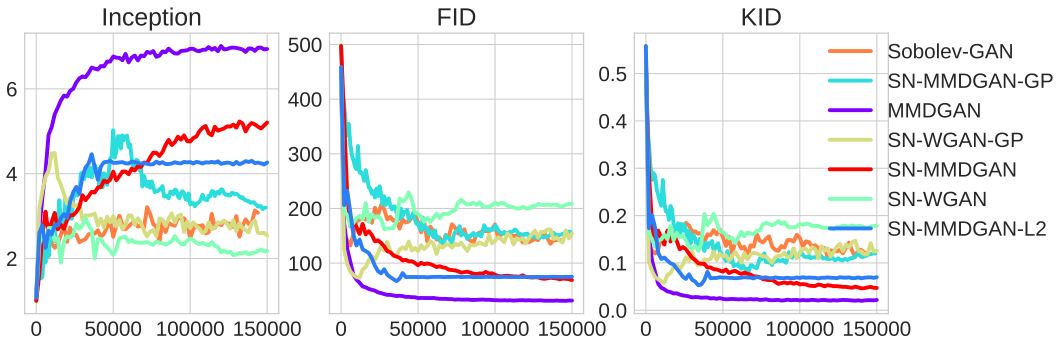


Figure 8: Evolution per iteration of different scores for variants of methods, mostly using spectral normalization, on CIFAR-10.

Figure 9: Samples from a generator trained on ImageNet dataset using Scaled MMD with Spectral Normalization: SN-SMMDGAN.

(a) SNGAN (b) SobolevGAN

(c) MMDGAN-GP-L2 (d) SN-SMMD GAN

(e) SN SWGAN (f) SMMD GAN

Figure 10: Comparison of samples from different models trained on CelebA with 160×160 resolution.