On the Convergence of Gradient Descent in GANs: MMD GAN As a Gradient Flow

Youssef Mroueh*
IBM Research

Truyen Nguyen*
University of Akron

Abstract

We consider the maximum mean discrepancy (MMD) GAN problem and propose a parametric kernelized gradient flow that mimics the min-max game in gradient regularized MMD GAN. We show that this flow provides a descent direction minimizing the MMD on a statistical manifold of probability distributions. We then derive an explicit condition which ensures that gradient descent on the parameter space of the generator in gradient regularized MMD GAN is globally convergent to the target distribution. Under this condition, we give non asymptotic convergence results of gradient descent in MMD GAN. Another contribution of this paper is the introduction of a dynamic formulation of a regularization of MMD and demonstrating that the parametric kernelized descent for MMD is the gradient flow of this functional with respect to the new Riemannian structure. Our obtained theoretical result allows ones to treat gradient flows for quite general functionals and thus has potential applications to other types of variational inferences on a statistical manifold beyond GANs. Finally, numerical experiments suggest that our parametric kernelized gradient flow stabilizes GAN training and guarantees convergence.

1 Introduction

Generative Adversarial Networks (GANs) were introduced in [1] and have attracted growing attention in the machine learning community. Implicit Generative models such as GANs can be seen as learning a distribution via optimizing a functional defined on a statistical manifold. The statistical manifold refers to the parametrization of the *Generator*. There are a plethora of works on functionals that are optimized in GANs for example the Jensen-Shanon divergence in the original work [1]; general ϕ divergences in [2]; the neural net distance in [3]; integral probability metrics such as the Wasserstein order 1 distance considered in [4]; the maximum mean discrepancy [5] considered in [6–10]. Despite their striking empirical success, the rigorous understanding of the convergence in *distributional sense* of gradient descent in GANs remains less understood. Much of the theoretical analysis has been dedicated to the stability of the min-max game via the introduction of gradient regularizers [11–17]. Min-max convergence rates for a large class of GANs was studied in [18, 19], however these bounds are not specific to gradient descent in GANs. In this work we aim at understanding the distributional convergence properties of gradient descent in the context of MMD GANs. The work closest to ours is [20] that establishes global convergence of the generator using the Wasserstein 1 distance. Nevertheless, gradient descent is not explicitly considered in [20].

We summarize our main contributions in this work as follows:

• We introduce in Section 3 a new gradient regularizer for the MMD. This regularizer has the form of a parametric energy, where the gradient is taken with respect to the generator parameters, instead of the input space as usually considered in previous works. We call the new proposed discrepancy $\text{MMD}_{\alpha,\beta}$.

• We consider the MMD GAN problem in Section 4 and propose a new descent direction in terms of the witness function of the parametric regularized $\text{MMD}_{\alpha,\beta}$. We give in this section detailed descriptions and properties of the corresponding continuous flow.

• We analyze in Section 5 the non-asymptotic distributional convergence properties of gradient descent in MMD GAN when using the $\text{MMD}_{\alpha,\beta}$ witness functions to drive the generator updates.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

• We derive in Section 6 a dynamic formulation of the MMD on the statistical manifold of probability distributions and use it to propose a novel regularization of the MMD, which we call $d_{\alpha,\beta}$. We show that $d_{\alpha,\beta}$ admits a Riemannian metric tensor and investigate gradient flows for general functionals w.r.t. this structure. Intriguingly, we show that gradient descent in MMD GANs driven by the witness function of $\mathrm{MMD}_{\alpha,\beta}$ coincides with the gradient flow of the MMD w.r.t. this new geometric structure.

• Finally, we discuss related works in Section 7, and validate experimentally our theoretical findings in Section 8.

2 Preliminaries

Let Ω be an open region in $\mathbb{R}^d$ and let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) generated by a kernel $k(x,y)$ on $\Omega\times\Omega$. Let $\mathcal{Z}\subset\mathbb{R}^m$ be an open region on a lower dimensional space endowed with a probability distribution ν on $\mathcal{Z}$. Let Θ be a parameter space in $\mathbb{R}^p$, and $(\theta,z)\to G_\theta(z)=\left(G_\theta^1(z),...,G_\theta^d(z)\right)\in\Omega$ be a generator function defined on Θ. We assume that k is bounded, G_θ is differentiable in θ, and

$$\begin{aligned}\|k(x,.)-k(y,.)\|_{\mathcal{H}}&\leq L\,\|x-y\|\,,\\ \|G_\theta(z)-G_{\theta'}(z)\|&\leq M(z)\,\|\theta-\theta'\|\end{aligned}\tag{1}$$

for some constant $L>0$ and function $M:\mathcal{Z}\to[0,\infty)$ with $\mathbb{E}_\nu[M(z)^2]<\infty$. In (1) and throughout the paper, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the standard Euclidean inner product and norm, while $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the inner product and norm on $\mathcal{H}$. We use ∇ and ∇_w to respectively denote the standard gradient and the standard gradient with respect to the variable w. For two vectors $\mathbf{a}$ and $\mathbf{b}$, by $\mathbf{a}\otimes\mathbf{b}$ we mean the matrix whose (i,j) entry is $\mathbf{a}_i\mathbf{b}_j$. For a probability distribution ρ on Ω, let $\mu_\rho(x):=\int k(x,y)\rho(dy)$ denote its kernel mean embedding. As k is bounded, we have $\mu_\rho\in\mathcal{H}$ and $\int f(x)\rho(dx)=\langle f,\mu_\rho\rangle_{\mathcal{H}}$ for every $f\in\mathcal{H}$ (see [21]). Note that $\mu_{\rho_1-\rho_2}=\mu_{\rho_1}-\mu_{\rho_2}$ by linearity.

We consider the following statistical manifold of probability distributions:

$$\mathcal{P}_\Theta:=\{q_\theta=(G_\theta)_\#\nu,\ \ \theta\in\Theta\}.$$

The main objective functions considered in this paper are parametric energy regularizations of $\mathrm{MMD}(p,q_\theta):=\|\mu_p-\mu_{q_\theta}\|_{\mathcal{H}}$. The introduction of the following operators plays an important role in understanding these objective functions and associated gradient flows.

Definition 2.1 (Matrix Mass)**.** *Let $J_\theta G_\theta(z)=\left(\frac{\partial G^j_\theta(z)}{\partial\theta_i}\right)_{ij}\in\mathbb{R}^{p\times d}$ denote the Jacobian of G_θ with respect to θ. Then for $\theta\in\Theta$, we define a matrix valued kernel Γ_θ on $\mathcal{Z}\times\mathcal{Z}$ as follows, for $(z,z')\in\mathcal{Z}\times\mathcal{Z}$:*

$$\Gamma_\theta(z,z'):=J_\theta G_\theta(z)^\top J_\theta G_\theta(z')\in\mathbb{R}^{d\times d}.$$

Definition 2.2 (Parametric Grammian - Mass Corrected Grammian of Derivatives)**.** *For $\theta\in\Theta$, let $L_\theta:\mathcal{H}\to\mathbb{R}^p$ be the operator given by*

$$L_\theta(f):=\int J_\theta G_\theta(z)\nabla f(G_\theta(z))\nu(dz)$$

and $L_\theta^\top:\mathbb{R}^p\to\mathcal{H}$ be the operator given by

$$\begin{aligned}L_\theta^\top(v)&:=\int\langle\nabla_\theta[k(G_\theta(z)],.),v\rangle\,\nu(dz)\\&=\int\langle\nabla_x k(G_\theta(z),.),J_\theta G_\theta(z)^\top v\rangle\,\nu(dz).\end{aligned}$$

Then the parametric Grammian $D_\theta:\mathcal{H}\to\mathcal{H}$ is defined by $D_\theta:=L_\theta^\top L_\theta$.

The main properties of these operators are summarized as follows. Proofs are given in the Appendix.

Proposition 2.3. *For each $\theta\in\Theta$, we have*

i) $L_\theta^\top$ is the adjoint operator of L_θ, i.e., $\langle L_\theta f,v\rangle=\langle f,L_\theta^\top v\rangle_{\mathcal{H}}$ for $f\in\mathcal{H}$ and $v\in\mathbb{R}^p$.

ii) D_θ is symmetric, i.e., $\langle D_\theta f,g\rangle_{\mathcal{H}}=\langle f,D_\theta g\rangle_{\mathcal{H}}$ for $f,g\in\mathcal{H}$.

iii) $\langle f,D_\theta f\rangle_{\mathcal{H}}=\left\|\nabla_\theta\int f(x)q_\theta(dx)\right\|^2=\|\nabla_\theta[\langle f,\mu_{q_\theta}\rangle_{\mathcal{H}}]\|^2\geq0$. In particular, D_θ is a positive operator and hence its spectrum is contained in $[0,\infty)$.

iv) For $f\in\mathcal{H}$, we have $(D_\theta f)(x)=\langle D(x,\cdot),f\rangle_{\mathcal{H}}$ with $D(y,y'):=\iint\langle\partial_\theta k(G_\theta(z),y),\partial_\theta k(G_\theta(z'),y')\rangle\nu(dz)\nu(dz')$, where: $\partial_\theta k(G_\theta(z),y)=J_\theta G_\theta(z)\nabla_x k(G_\theta(z),y)$. Equivalently $D(y,y')=\iint Trace(\nabla_x k(G_\theta(z),y)\otimes\Gamma_\theta(z,z')\nabla_x k(G_\theta(z'),y'))\nu(dz)\nu(dz')$.

3 A Novel Parametric Energy Regularization of MMD

Let us introduce a parametric energy regularization of MMD and this notion of discrepancy will play a central role in this paper. For parameters α, $\beta\geq0$, the regularized discrepancy between a given probability distribution p on Ω and a parametric distribution $q_\theta\in\mathcal{P}_\Theta$ is defined by

$$\mathrm{MMD}_{\alpha,\beta}(p,q_\theta):=\sup_{f\in E_{\alpha,\beta}}\int f(x)\,p(dx)-\int f(x)\,q_\theta(dx)\tag{2}$$

with $E_{\alpha,\beta} := \Big\{ f \in \mathcal{H} : \alpha \Big\| \nabla_\theta \int f(x) q_\theta(dx) \Big\|^2_{\mathbb{R}^p} + \beta \|f\|^2_{\mathcal{H}} \leq \frac{1}{2} \Big\}$. The case $\alpha = 0$ and $\beta = 1/2$ corresponds to the MMD [5], while the case $\alpha = 1/2$ and $\beta = 0$ shares some similarity with the usual kernelized Sobolev discrepancy [22]. The main difference in this definition with the Sobolev discrepancy is that the parametric energy $\|\nabla_\theta[\langle f, \ _q \rangle_{\mathcal{H}}]\|^2_{\mathbb{R}^p} = \langle f, D_\theta f\rangle_{\mathcal{H}}$ is used in place of the standard energy $\|\nabla f\|^2_{L^2_q}$.

Remark 3.1. *Note that our regularization for* MMD *while it shares similarities with WGAN-GP [11], it is different since our gradient penalty is with respect to the generator parameter whereas it is with respect to the input in WGAN-GP.*

Hereafter, $I : \mathcal{H} \to \mathcal{H}$ denotes the identity operator. Then it follows from property iii) in Proposition 2.3 that $\alpha D_\theta + \beta I$ is invertible whenever $\alpha \geq 0$ and $\beta > 0$. Proposition 2.3 also allows us to express the constraint $E_{\alpha,\beta}$ in (2) as

$$E_{\alpha,\beta} = \Big\{ f \in \mathcal{H} : 2 \langle f, (\alpha D_\theta + \beta I) f\rangle_{\mathcal{H}} \leq 1 \Big\}. \quad (3)$$

This constraint can be interpreted as a regularization through the following unconstrained formulation.

Proposition 3.2. *Let* $\Delta_\theta(f) := \int f(x)\, p(dx) - \int f(x)\, q_\theta(dx)$. *For* $\alpha \geq 0$ *and* $\beta > 0$, *we have*

$$\begin{aligned}
&\mathrm{MMD}_{\alpha,\beta}(p, q_\theta)^2 \\
&= \sup_{f \in \mathcal{H}} \Big\{ \Delta_\theta(f) - \frac{\alpha}{2} \|\nabla_\theta[\langle f, \ _q \rangle_{\mathcal{H}}]\|^2_{\mathbb{R}^p} - \frac{\beta}{2} \|f\|^2_{\mathcal{H}} \Big\} \\
&= \frac{1}{2} \Big\langle \ _{p-q}, (\alpha D_\theta + \beta I)^{-1} \ _{p-q} \Big\rangle_{\mathcal{H}}.
\end{aligned}$$

Moreover, the witness function f^* *realizing the above supremum is given by:*

$$(\alpha D_\theta + \beta I) f^* = \ _{p-q}. \quad (4)$$

The next result shows that the regularized $\mathrm{MMD}_{\alpha,\beta}$ is upper bounded by the MMD, and gives a characterization on distributions for which the two discrepancies are the same.

Corollary 3.3. *For* $\alpha \geq 0$ *and* $\beta > 0$, *we have*

$$\sqrt{2\beta}\,\mathrm{MMD}_{\alpha,\beta}(p, q_\theta) \leq \mathrm{MMD}(p, q_\theta).$$

In case $\alpha > 0$, *the equality happens if and only if* $D_\theta \ _{p-q} = 0$.

4 Generative Adversarial Networks via Parametric Regularized Flows

Let p be the target distribution which is a probability measure on Ω. Consider the functional $\mathcal{F}(q_\theta) = \frac{1}{2}\mathrm{MMD}^2(p, q_\theta) = \frac{1}{2}\| \ _{p-q} \|^2_{\mathcal{H}}$. We now focus on the MMD GAN problem:

$$\begin{aligned}
\min_{q \in \mathcal{P}} \mathcal{F}(q) &= \min_{\theta \in} \mathcal{F}(q_\theta) \\
&= \min_{\theta \in} \frac{1}{2} \|\mathbb{E}_p k(x, .) - \mathbb{E}_{z \sim \nu} k(G_\theta(z), .)\|^2_{\mathcal{H}}.
\end{aligned}$$

This problem has been investigated in several works [6–8, 22, 23]. In this section we propose a new descent direction in the parameter space of the generator.

Continuous Descent. We consider the following dynamic for any sequence of functions $f_t \in \mathcal{H}$, $t \geq 0$:

$$\frac{d\theta_t}{dt} = L_{\theta_t}(f_t) = \int J_{\theta_t} G_{\theta_t}(z') \nabla_x f_t(G_{\theta_t}(z'))\nu(dz'). \quad (5)$$

For a given $z \in \mathcal{Z}$, the dynamic of the generator is as follows

$$\begin{aligned}
\frac{dG_{\theta_t}(z)}{dt} &= J_{\theta_t} G_{\theta_t}(z)^\top \frac{d\theta_t}{dt} \quad (6)\\
&= \int J_{\theta_t} G_{\theta_t}(z)^\top J G_{\theta_t}(z') \nabla_x f_t(G_{\theta_t}(z'))\nu(dz') \\
&= \int \Gamma_{\theta_t}(z, z') \nabla_x f_t(G_{\theta_t}(z'))\nu(dz'). \quad (7)
\end{aligned}$$

While the dynamic of particles is usually given by velocities defined at each particle, the generator's dynamic at a given z is the average of the mass corrected velocities of all other samples from the generator. The mass correction is driven by the matrix valued kernel Γ_{θ_t} that defines a similarity in the hidden space $\mathcal{Z}$. For example in particles descent such as Sobolev Descent [22] the dynamic of particles X_t is given by:

$$\frac{dX_t}{dt} = \nabla_x f_t(X_t),$$

and this simple advection dynamic is to be contrasted with the generator dynamic (7).

Using the dynamic of the generator in (7) and item iv) in Proposition 2.3, we also have the following dynamic of the mean embedding:

$$\begin{aligned}
&\frac{d}{dt} \ _{q_t} = \frac{d}{dt} \int k(G_{\theta_t}(z), .)\nu(dz) \\
&= \int \Big\langle \nabla_x k(G_{\theta_t}(z), .), \frac{dG_{\theta_t}(z)}{dt} \Big\rangle \nu(dz) \\
&= \int\int \langle \nabla_x k(G_{\theta_t}(z), .), \Gamma_{\theta_t}(z, z') \nabla_x f_t(G_{\theta_t}(z'))\rangle\, \nu(dz')\nu(dz) \\
&= D_{\theta_t} f_t. \quad (8)
\end{aligned}$$

Thanks to (8), it is easy to derive the dynamic of the MMD distance:

$$\frac{d\mathcal{F}(q_{\theta_t})}{dt} = \Big\langle \ _{p-q_t}, -\frac{d}{dt} \ _{q_t} \Big\rangle_{\mathcal{H}} = -\Big\langle \ _{p-q_t}, D_{\theta_t} f_t \Big\rangle_{\mathcal{H}}. \quad (9)$$

Let us consider the following choices for the sequence f_t:

• **Witness functions of** MMD**.** In that case, we set $f_t = \ _{p-q_{\theta_t}}$. Using the generator updates given in (5), we have therefore:

$$\frac{d\mathcal{F}(q_{\theta_t})}{dt} = - \langle \ _{p-q_{\theta_t}}, D_{\theta_t} \ _{p-q_{\theta_t}} \rangle_{\mathcal{H}} \leq 0.$$

This is a valid descent direction, and is similar in spirit to the MMD flows of [24]. Nevertheless as shown for the particles case in [24], it does not lead to convergence. In the discrete case, [24] introduced a noising scheme that has convergence guarantees.

• **Witness functions of** $\mathrm{MMD}_{1,0}$**.** In that case, let us assume that solutions f_t of $D_{\theta_t} f_t = \ _{p-q_{\theta_t}}$ exist. Then by using $d\theta_t = L_{\theta_t}(f_t)dt$ we obtain:

$$\frac{d\mathcal{F}(q_{\theta_t})}{dt} = - \langle \ _{p-q_{\theta_t}}, \ _{p-q_{\theta_t}} \rangle = -2\mathcal{F}(q_{\theta_t}).$$

While this seems to be the ideal choice as it gives us an exponential convergence, it comes with the caveat that D_{θ_t} may be singular and hence we have either no solution or infinitely many solutions for f_t. These derivations and the singularity issue of operator D_θ motivated the introduction of $\mathrm{MMD}_{\alpha,\beta}$, and we consider hereafter its flows.

• **Witness functions of** $\mathrm{MMD}_{\alpha,\beta}$**.** Let f_t be the unique witness function of $\mathrm{MMD}_{\alpha,\beta}$ between p and $q_{\theta_t} = (G_{\theta_t})_{\#}\nu$ given by

$$(\alpha D_{\theta_t} + \beta I) f_t = \ _{p-q_{\theta_t}}. \tag{10}$$

Theorem 4.1 below gives the dynamic of the MMD when the generator parameters are updated according to Equation (5) with f_t being the $\mathrm{MMD}_{\alpha,\beta}$ witness functions given in Equation (10).

Theorem 4.1 (Parametric Regularized Flows Decrease the MMD Distance)**.** *Assume that α, $\beta > 0$. Then the dynamic (5)–(10) defined by the witness function of the parametric regularized* MMD *decreases the functional $\mathcal{F}(q_\theta)$:*

$$\frac{d\mathcal{F}(q_{\theta_t})}{dt} = -\frac{2}{\alpha}\Big[\mathcal{F}(q_{\theta_t}) - \beta\, \mathrm{MMD}_{\alpha,\beta}(p, q_{\theta_t})^2\Big] \leq 0. \tag{11}$$

Moreover, we have $\frac{d\mathcal{F}(q_{\theta_t})}{dt} < 0$ if and only if $D_{\theta_t} \ _{p-q_{\theta_t}} \neq 0$.

We see from Theorem 4.1 that $\mathrm{MMD}_{\alpha,\beta}$ witness functions alleviate the singularity issue of D_{θ_t}, but slows down the convergence by introducing a damping term that is proportional to $\mathrm{MMD}^2_{\alpha,\beta}$.

5 Non-Asymptotic Convergence Of Gradient Descent In MMD GAN

In Section 4, we showed that $\mathrm{MMD}_{\alpha,\beta}$ witness functions provide descent directions for continuous MMD GAN. In this section we turn to (discrete) gradient descent in the parameter space of the generator, and give non asymptotic convergence results for gradient descent in regularized MMD GANs.

Discrete Descent Directions. We would like to identify directions of θ along which the functional $\mathcal{F}(q_\theta)$ decreases its value. For this, let us compute the rate $\frac{d}{d\varepsilon}\big|_{\varepsilon=0}\mathcal{F}(q_{\theta+\varepsilon v})$ for each vector $v \in \mathbb{R}^p$.

Lemma 5.1. *Let $\theta \in \Theta$. Then for any vector $v \in \mathbb{R}^p$, we have*

$$\frac{d}{d\varepsilon}\mathcal{F}(q_{\theta+\varepsilon v}) = -\langle \ _{p-q_{\theta+\varepsilon v}}, L^\top_{\theta+\varepsilon v} v\rangle_{\mathcal{H}} \ \textit{for every } \varepsilon \geq 0.$$

Lemma 5.1 implies that v is a descent direction of $\mathcal{F}(q_\theta)$ if and only if $\langle \ _{p-q_\theta}, L^\top_\theta v\rangle_{\mathcal{H}} > 0$. The next result gives one such direction.

Proposition 5.2. *Let $\theta \in \Theta$, and assume that $D_\theta \ _{p-q_\theta} \neq 0$. Let $v^* = L_\theta f$ with $f \in \mathcal{H}$ being the solution of $(\alpha D_\theta + \beta I) f = \ _{p-q_\theta}$. Then v^* is a descent direction of $\mathcal{F}(q_\theta)$. Precisely, we have*

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}(q_{\theta+\varepsilon v^*}) = -\Big[\alpha\|D_\theta f\|^2_{\mathcal{H}} + \beta\langle f, D_\theta f\rangle_{\mathcal{H}}\Big] < 0.$$

In particular, $\mathcal{F}(q_{\theta+\varepsilon v^*}) < \mathcal{F}(q_\theta)$ if $\varepsilon > 0$ is small enough.

Discrete Time Descent for MMD GAN. Hereafter, we denote $\|h\|_{\mathcal{H}} := \big(\sum_{i=1}^d \|h_i\|^2_{\mathcal{H}}\big)^{\frac{1}{2}}$ for $h = (h_1, ..., h_d) \in \mathcal{H}^d$ and $\|A\| := (\sum_{i,j} a_{ij}^2)^{\frac{1}{2}}$ for a matrix $A = (a_{ij})$. The next result holds under the following extra conditions for k and G:

$$\|\nabla_x k(z_1, .) - \nabla_x k(z_2, .)\|_{\mathcal{H}} \leq \tilde{L}\, \|z_1 - z_2\|, \tag{12}$$

$$\|J_\theta G_\theta(z) - J_{\theta'} G_{\theta'}(z)\| \leq \tilde{M}(z)\, \|\theta - \theta'\| \tag{13}$$

with $\tilde{L} > 0$ being a constant and $\mathbb{E}_\nu[\tilde{M}(z)] < \infty$.

In Theorem 5.3 below we find conditions under which we can achieve global convergence in MMD of gradient descent in MMD GAN. When the kernel is characteristic, this is equivalent to the global weak convergence of gradient descent in MMD GAN to the target distribution. Gradient descent for MMD GAN is given by the following updates: for all $\ell \geq 1$, the witness functions between p and $q_{\theta_\ell} = (G_{\theta_\ell})_{\#}\nu$ update:

$$f_\ell := (\alpha_\ell D_{\theta_\ell} + \beta_\ell I)^{-1} \ _{p-q_{\theta_\ell}}, \tag{14}$$

and the generator update:

$$\theta_{\ell+1} := \theta_\ell + \varepsilon_\ell L_{\theta_\ell}(f_\ell), \tag{15}$$

where α_ℓ, β_ℓ are sequences of regularization parameters, and ε_ℓ is a sequence of learning rates.

Theorem 5.3. *Assume that k and G satisfy (1) and (12)–(13). Let $\lambda_i(\theta) > 0$ be the smallest non-zero eigenvalue of D_θ, and define $a(\theta, f) := 1 - \frac{||P_{\text{Null}(D_\theta)} f||_{\mathcal{H}}^2}{||f||_{\mathcal{H}}^2}$ with $P_{\text{Null}(D_\theta)}$ being the projection to the null space of D_θ. Let $\mathcal{F}(q_\theta) = \frac{1}{2}\text{MMD}(p, q_\theta)^2$. Consider the gradient descent updates given in (14) and (15). Let $0 < \tau < 1$ and $\theta_1 \in \Theta$ be the starting point chosen such that:*

$$a(\theta_1, f_1) > \tau. \tag{16}$$

The sequence ε_ℓ is chosen so that the following two conditions are satisfied for each ℓ:

$$a(\theta_{\ell+1}, f_{\ell+1}) > \tau \tag{17}$$

and

$$2C(2\beta_\ell)^{-1}\left(1 + \sqrt{\mathcal{F}(q_{\theta_1})}\right) \leq \varepsilon_\ell^{-1} \tag{18}$$

with $C > 0$ depending only on the constants C_1, C_2, C_3, C_4 given in Lemma B.1.

Under the condition (17) on ε_ℓ we have $\tau < a(\theta_\ell, f_\ell) \leq 1$. Let $\chi_j := \frac{\lambda_i(\theta_j) a(\theta_j, f_j)}{\alpha_j \lambda_i(\theta_j) a(\theta_j, f_j) + \beta_j} > 0$. Then we have:

$$\mathcal{F}(q_{\theta_{\ell+1}}) \leq \mathcal{F}(q_{\theta_1}) \exp\Big(-\sum_{j=1}^{\ell} \varepsilon_j \chi_j\Big) \quad \forall \ell \geq 1.$$

In particular for $\alpha_\ell \leq \frac{\tau}{2}$ and $\beta_\ell = \alpha_\ell \lambda_i(\theta_\ell)$, we obtain $\chi_\ell \geq 1$ and it follows that:

$$\mathcal{F}(q_{\theta_{\ell+1}}) \leq \mathcal{F}(q_{\theta_1}) \exp\Big(-\sum_{j=1}^{\ell} \varepsilon_j\Big).$$

Consequently if $\sum_{j=1}^{\infty} \varepsilon_j = +\infty$ (meaning that ε_j decays as $1/\sqrt{j}$ for e.g) and conditions (17) and (18) hold, then we obtain $\mathcal{F}(\theta_\ell) \to 0$ as $\ell \to \infty$. As we see from Theorem 5.3, not all learning rates are admissible. At a given iteration ℓ, we select a learning rate ε_ℓ so that $f_{\ell+1} \notin \text{Null}(D_{\theta_{\ell+1}})$ (ensured by condition (17)) and so that (18) holds as well. To understand condition (17), recall from Theorem 4.1 that having f_ℓ not in the null space of D_{θ_ℓ} means that we have a strict descent. Notice that $f_\ell \notin \text{Null}(D_{\theta_\ell})$ if and only if $\delta_{p-q_{\theta_\ell}} \notin \text{Null}(D_{\theta_\ell})$.

6 Parametric Kernelized Flows for a General Functional

The flow of the MMD functional (i.e. $\frac{1}{2}\text{MMD}(p, q_\theta)^2$) analyzed in the previous sections is driven by the gradient of the witness function between p and q_θ of the discrepancy $\text{MMD}_{\alpha,\beta}$. In this section we discover a Riemannian structure on the statistical manifold of probability distributions and show that the continuous gradient descent in MMD GANs described in Section 4 coincides with the gradient flow of the functional with respect to this new geometric structure. We also develop a rigorous theory for treating gradient flows of general functionals and thus open a way for other types of variational inferences beyond GANs.

6.1 Dynamic Formulation For MMD on a Statistical Manifold

The following result gives a dynamic formulation of MMD and allows us to discover a Riemannian structure associated to MMD. This is analogous to Benamou–Brenier dynamic formulation of the Wasserstein of order 2 [25]:

$$W_2^2(p, q) = \inf_{(q_t, f_t)} \int_0^1 \int \|\nabla_x f_t(x)\|^2 q_t(dx) dt$$
$$\text{s.t } \frac{\partial q_t(x)}{\partial t} = -div(q_t \nabla_x f_t(x)), \ q_0 = q, \ q_1 = p. \tag{19}$$

The main difference is that their flows are with respect to the standard L_q^2 energy, while ours as explained in Section 6.2, are driven by the parametric energy $||\nabla_\theta[\langle f, \delta_q \rangle_{\mathcal{H}}]||^2$.

Theorem 6.1 (Dynamic MMD on a Statistical Manifold)**.** *Assume that for any $\theta_0, \theta_1 \in \Theta$, there exists a path $(\theta_t, f_t)_{t\in[0,1]}$ such that $\theta_{t=0} = \theta_0$, $\theta_{t=1} = \theta_1$, $f_t \in \mathcal{H}$, and*

$$\partial_t \theta_t = L_{\theta_t} f_t \ \text{and} \ D_{\theta_t} f_t = 2\delta_{q_{\theta_1} - q_{\theta_t}} \quad \forall t \in [0, 1]. \tag{20}$$

Let q_{θ_0} and q_{θ_1} be two probability measures in $\mathcal{P}_\Theta$. Then we have the following dynamic form of MMD between distributions defined on the statistical manifold $\mathcal{P}_\Theta$:

$$\text{MMD}^2(q_{\theta_0}, q_{\theta_1}) = \min_{(\theta_t, f_t)} \left\{ \int_0^1 \|D_{\theta_t} f_t\|_{\mathcal{H}}^2 dt \ , \right.$$
$$\left. \partial_t \theta_t = L_{\theta_t} f_t, \ f_t \in \mathcal{H}, \ \theta_{t=0} = \theta_0, \ \theta_{t=1} = \theta_1. \right\}$$

Theorem 6.1 is proven under Assumption 20 that guarantees the existence of a solution. This assumption is not realistic since D_{θ_t} can be singular. Nevertheless, we state this theorem to motivate the introduction in the next section of the dynamic form of a regularized MMD that alleviates this singularity issue.

6.2 Regularized MMD and Gradient Flows on a Statistical Manifold

Motivated by the result in Theorem 6.1 and to alleviate the singularity issue in Assumption (20), we define the following regularized version of MMD:

Definition 6.2 (Regularized MMD on a Statistical Manifold)**.** *Let $\alpha, \beta > 0$. Define*

$$d_{\alpha,\beta}(q_{\theta_0}, q_{\theta_1})^2 = \min_{(\theta_t, f_t)} \int_0^1 \alpha \|D_{\theta_t} f_t\|^2_{\mathcal{H}} + \beta \langle f_t, D_{\theta_t} f_t\rangle_{\mathcal{H}} \, dt,$$

$$\partial_t \theta_t = L_{\theta_t} f_t, \; f_t \in \mathcal{H}, \; \theta_{t=0} = \theta_0, \; \theta_{t=1} = \theta_1.$$

Note that the regularization we introduced here is the parametric energy $\|\nabla_\theta[\langle f_t, \mu_q\rangle_{\mathcal{H}}]|_{\theta=\theta_t}\|^2 = \langle f_t, D_{\theta_t} f_t\rangle_{\mathcal{H}}$, which plays a similar role as the kinetic energy in Benamou-Brenier's formula. The evolution of θ_t in our form is analogous to the continuity equation in the $\mathcal{W}_2$ dynamic form [25]. Conditions on the kernel and the generator family under which we can guarantee existence of the solution for this problem are out of the scope of this work, since our interest in this heuristic Riemannian structure is solely in order to define an appropriate tangent space and Riemannian metric tensor. We leave the analysis of $d_{\alpha,\beta}$ to a future work.

This dynamic formulation gives rise to the following Riemannian metric tensor on the tangent space of Θ: for $\theta \in \Theta$, let

$$\begin{aligned} g_\theta(\xi_1, \xi_2) &:= \alpha \langle D_\theta \varphi_1, D_\theta \varphi_2\rangle_{\mathcal{H}} + \beta \langle \varphi_1, D_\theta \varphi_2\rangle_{\mathcal{H}} \\ &= \langle (\alpha D_\theta + \beta I)\varphi_1, D_\theta \varphi_2\rangle_{\mathcal{H}} \end{aligned}$$

where $\xi_i = L_\theta(\varphi_i) = \nabla_\theta[\langle \varphi_i, \mu_q\rangle_{\mathcal{H}}] \in \mathbb{R}^p$ with $\varphi_i \in \mathcal{H}$ $(i = 1, 2)$. We note that $g_\theta(\xi_1, \xi_2) = g_\theta(\xi_2, \xi_1)$ due to the symmetry of D_θ (see property ii) of Proposition 2.3). Then it follows that

$$d_{\alpha,\beta}(q_{\theta_0}, q_{\theta_1})^2 = \min_{(\theta_t, f_t)} \left\{ \int_0^1 g_\theta(\partial_t \theta_t, \partial_t \theta_t) dt \right\},$$

$$\partial_t \theta_t = L_{\theta_t} f_t, \; f_t \in \mathcal{H}, \; \theta_{t=0} = \theta_0, \; \theta_{t=1} = \theta_1.$$

Let us assume that $\alpha, \beta > 0$ from now on. For a functional $\mathcal{F} : \mathcal{P}_\Theta \to \mathbb{R}$, let $\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_\theta)$ denote the gradient of $\mathcal{F}$ with respect to the metric $d_{\alpha,\beta}$. That is, $\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_\theta)$ is a vector in $\mathbb{R}^p$ satisfying

$$\left.\frac{d\mathcal{F}(q_{\theta_t})}{dt}\right|_{t=0} = g_\theta(\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_\theta), \xi) \tag{21}$$

for every differentiable curve $t \mapsto \theta_t \in \Theta$ with $\theta_{t=0} = \theta$ and $\partial_t \theta_t|_{t=0} = \xi = L_\theta \varphi$ for some $\varphi \in \mathcal{H}$. The next theorem shows us how to compute such gradient.

Theorem 6.3. *Let $\mathcal{F}(q_\theta)$ be a functional depending only on the kernel mean embedding of q_θ. Precisely, assume that $\mathcal{F}(q_\theta) = H(\mu_q)$ for some functional H with the chain rule property*

$$\partial_{\theta_i}[H(\mu_q)] = \langle h_\theta, \partial_{\theta_i}[\mu_q]\rangle_{\mathcal{H}} \tag{22}$$

for some function $h_\theta \in \mathcal{H}$ and for all $\theta \in \Theta$. Then the gradient of $\mathcal{F}$ w.r.t. the discrepancy $d_{\alpha,\beta}$ is given by

$$\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_\theta) = L_\theta u,$$

where

$$(\alpha D_\theta + \beta I)u = h_\theta. \tag{23}$$

Let $\mathcal{F}(q_\theta)$ be the functional as in Theorem 6.3, and consider the gradient flow of $\mathcal{F}(q_\theta)$ with respect to $d_{\alpha,\beta}$. We note that this is a gradient regularized flow. According to Theorem 6.3, the equation of this flow is given by

$$\partial_t \theta_t = -\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_{\theta_t}) = -L_{\theta_t} u_t, \tag{24}$$

where

$$(\alpha D_{\theta_t} + \beta I)u_t = h_{\theta_t}. \tag{25}$$

The following proposition shows that these gradient flows are indeed descent directions of the functional defined on the statistical manifold $\mathcal{P}_\Theta$:

Proposition 6.4. *Along the gradient flow (24)–(25) of $\mathcal{F}(q_\theta)$, we have*

$$\begin{aligned} \frac{d}{dt}\mathcal{F}(q_{\theta_t}) &= -\left(\alpha\|D_{\theta_t} u_t\|^2_{\mathcal{H}} + \beta\|L_{\theta_t} u_t\|^2\right) \\ &= -\frac{1}{\alpha}\left[\|h_{\theta_t}\|^2_{\mathcal{H}} - \beta\left\langle h_{\theta_t}, (\alpha D_{\theta_t} + \beta I)^{-1} h_{\theta_t}\right\rangle_{\mathcal{H}}\right] \\ &\le 0, \end{aligned}$$

where h_θ is defined by (22). Moreover, we have $\frac{d\mathcal{F}(q_{\theta_t})}{dt} < 0$ if and only if $D_{\theta_t} h_{\theta_t} \neq 0$.

Intuition on the role of D_θ in the gradient flow. The operator D_θ plays a central role in our framework and we give here an intuitive interpretation of its role from the gradient flow lens. Let $\{(\lambda_j(\theta), v_j(\theta))\}_{j=1}^{\infty}$ be the eigenvalues and eigenfunctions of the operator D_θ. Then as we have $u_t = \sum_{j=0}^{\infty} \frac{1}{\alpha\lambda_j(\theta_t)+\beta} \langle h_{\theta_t}, v_j(\theta_t)\rangle_{\mathcal{H}} v_j(\theta_t)$, the flow equation can be written as follows: $\partial_t \theta_t = -\mathrm{grad}_{d_{\alpha,\beta}} \mathcal{F}(q_{\theta_t}) = -\sum_{j=0}^{\infty} \frac{1}{\alpha\lambda_j(\theta_t)+\beta} \langle h_{\theta_t}, v_j(\theta_t)\rangle_{\mathcal{H}} L_{\theta_t}(v_j(\theta_t))$. The eigenfunctions of D_{θ_t} provide the descent directions $L_{\theta_t}(v_j(\theta_t))$, that are linearly combined according to the similarity of $v_j(\theta_t)$ and h_{θ_t} and weighted by a factor $1/(\alpha\lambda_j(\theta_t)+\beta)$. Small eigenvalues are noisy directions and spectral filtering them via α and β favors descent directions with larger eigenvalues.

6.3 Gradient Flows of Particular Functionals: MMD GAN as Gradient Flow

MMD **GAN as a Gradient Flow w.r.t. $d_{\alpha,\beta}$.** The next result shows that the flow (5)–(10) of the parametric gradient regularized MMD GAN coincides with the gradient flow of $\mathcal{F}(q_\theta) = \frac{1}{2}\text{MMD}(p, q_\theta)^2$ with respect to $d_{\alpha,\beta}$.

Corollary 6.5. *For $\mathcal{F}(q_\theta) = \frac{1}{2}\text{MMD}(p, q_\theta)^2$, we have $\text{grad}_{d\ ;}\mathcal{F}(q_\theta) = L_\theta u$ with $(\alpha D_\theta + \beta I)u = - \ _{p-q}$. Consequently, the parametric regularized flow (5)–(10) on the statistical manifold is the gradient flow of $\mathcal{F}$ with respect to $d_{\alpha,\beta}$*

Proof. This follows immediately from Theorem 6.3 noting that $\partial_{\theta_i}[\mathcal{F}(q_\theta)] = -\langle \ _{p-q} , \partial_{\theta_i}[\ _q]\rangle_{\mathcal{H}}$. ☐

Gradient flows of Generic Functionals on the Statistical Manifold. Our framework is not limited to the MMD functional. In the corollary below we exhibit additional examples of functional $\mathcal{F}(q_\theta)$ defined on the statistical manifold along with their gradient flows w.r.t. $d_{\alpha,\beta}$.

Corollary 6.6.

- *For the (potential) energy functional $\mathcal{F}_1(q_\theta) = \int V(x)q_\theta(dx)$ with $V \in \mathcal{H}$, we have $\text{grad}_{d\ ;}\mathcal{F}_1(q_\theta) = L_\theta u$, where $(\alpha D_\theta + \beta I)u = V$.*
- *For the (entropy) functional $\mathcal{F}_1(q_\theta) = \int f(\mu_q\ (x))dx$ with $f : \mathbb{R} \to \mathbb{R}$ being continuously differentiable, we have $\text{grad}_{d\ ;}\mathcal{F}_1(q_\theta) = L_\theta u$, where $(\alpha D_\theta + \beta I)u = \int f'(\ _q\ (x))k(x, .)dx$.*
- *For the (interaction) functional $\mathcal{F}_3(q_\theta) = \int f(x)g(y)q_\theta(dx)q_\theta(dy)$ with $f, g \in \mathcal{H}$, we have $\text{grad}_{d\ ;}\mathcal{F}_3(q_\theta) = L_\theta u$, where $(\alpha D_\theta + \beta I)u = \langle f, \ _q \rangle_{\mathcal{H}}\, g + \langle g, \ _q \rangle_{\mathcal{H}}\, f$.*

Continuous Time Descent In this section we analyze the convergence properties of the kernelized parametric gradient flows of functionals defined on the statistical manifold w.r.t. to $d_{\alpha,\beta}$. The following proposition studies the convergence behavior of the parametric flows.

Proposition 6.7 (Convergence up to a barrier)**.** *Let $\mathcal{F}(q_\theta)$ be the functional as in Theorem 6.3. Assume that there exists a continuous function $\gamma : \Theta \to [0, \infty)$ such that*

$$\|h_\theta\|^2_{\mathcal{H}} \geq \gamma(\theta)\,\mathcal{F}(q_\theta) \quad \forall\theta \in \Theta. \tag{26}$$

Consider the dynamic $t \in [0, +\infty) \to \theta_t \in \Theta$ of the gradient flow (24)–(25). Then we have for every $t \geq 0$:

$$\mathcal{F}(q_{\theta_t}) \leq \mathcal{F}(q_{\theta_0})e^{-\int_0^t \frac{\ _i(\ _s)a(\ _s;u_s)\ (\ _s)}{\ _i(\ _s)a(\ _s;u_s)+\ }ds}, \tag{27}$$

where $\lambda_i(\theta)$ and $a(\theta, u)$ are defined in Theorem 5.3.

Remark 6.8. *Note that condition (26) is satisfied by the MMD for $\gamma(\theta) = 2$, since $h_\theta = - \ _{p-q}$, and hence $\|h_\theta\|^2_{\mathcal{H}} = 2\,\mathcal{F}(q_\theta)$.*

7 Related Work

MMD GAN. Since their introduction in [1], many cost functions have been proposed for training GANs. Related to our work is MMD GAN introduced in [6, 7] and later improved in [8].

Gradient Regularized GANs. GAN's training is notoriously known to be unstable. Wasserstein GAN [4] tackled that issue by considering $\mathcal{W}_1$ as a loss function for training. Imposing lipchitizity in practice is challenging and a gradient penalty in the input space was introduced in WGAN-GP [11] as means to impose lipchitizity. Sobolev GAN [12] connected this gradient penalty to advection and semi-Sobolev norms. The work [16] considered this input gradient regularizer in MMD GANs.

Closely related to our study are the gradient penalties on the parameter space of the generator and the discriminator considered in [13]. It was demonstrated in [14, 15] that these gradient regularizers ensure stability and convergence of the min/max game. [17] showed that GANs with a variant of a gradient penalty on the parameter space of the discriminator is locally stable. Even though these works studied the stability of the min/max game and its convergence to a saddle point, they did not show the global convergence of the discrete gradient flow in the distributional sense which is what we prove in this paper thanks to the newly proposed Riemannian structure.

Gradient Flows. Wasserstein $\mathcal{W}_2$ flows for minimizing functionals over probabilities (see the excellent introduction [26]) are related to our work. These flows are non parametric and are given by the continuity equation. Recently Wasserstein flows were extended to statistical manifold [27], and a kernelized Wasserstein natural gradient flow was introduced in [24]. Kernelized particle flows such as Stein Descent [28–30], Sobolev Descent [22] and MMD flows [23] are not defined on a statistical manifold and only act on the particle level. [29, 30] introduce similar Riemannian structures to ours for the Stein geometry, nevertheless they are not defined on a statistical manifold. It would be interesting to derive a dynamical form of the Stein metric on a manifold of parametric explicit densities. [31] studied minimum stein estimators and introduced a natural Stein gradient, and it would be interesting to connect their study to a definition of an appropriate dynamic Stein structure on a parametric manifold.

8 Numerical Experiments

We consider a finite dimensional RKHS defined by a random feature map [32, 33] $\Phi(x) : \Omega \to \mathbb{R}^m$. Φ

(a) Trajectories of Flows for = 0 : Euclidean Gradients Flows

(b) MMD Loss

(c) Trajectories of Flows for = 100 : Kernelized Gradients Flows w.r.t. $d_{;}$

(d) MMD Loss

Figure 1: Trajectories of Kernelized Gradient ows of the MMD functional for = 0 (no gradient regularization) and = 100. It is clear that the Riemannian structure induced by $d_{;}$, > 0 guarantees the convergence, while GAN su er from cycles and mode collapse for = 0 .

in our case is a 4 layers Relu Network with hidden dimension $m = 512$, with weights sampled from standard Gaussian and then xed [34]. The space $H = f f(x) = hw; (x)i ; w 2 R^m g$.

MMD GAN as a Kernelized Gradient Flow. In parametric gradient regularized MMD GANs, we nd the MMD ; (p; q) witness function f between p and q , and then update the generator according to: $_{+1}$ = + "L f . We showed that this coincides exactly with the gradient ow of the MMD functional w.r.t. the geometric structure $d_{;}$ introduced in this paper, and we proved its global convergence under mild assumptions. We show here an example, where the target distribution is a two dimensional mixture of Gaussians with 8 modes. In this example, Z is 256 dimensional space and is standard Gaussian on Z . The generator G is a 2 layer Relu Network (hidden size 128) and two dimensional output. We x the mini-batch size in the training to 512, the learning rate of the witness function and generator is set to 1e ^{4}. We train the witness function f of MMD ; with stochastic gradient (RmsProp [35]) for 5 iterations. Note that in this case only the last linear layer is updated in the witness function, and is kept xed. We then update the parameter with gradient descent on E_q f . We use also RmsProp to ensure that the learning rates are adaptive. Note that given the witness function f , the discrete ow update given in Eq. (15) is exactly the gradient descent update on E_q f .

We give in Figure 1 the trajectories of the ows for = 100 and = 0 for L = 100K iterations. Since in our implementation (Appendix E) we use stochastic gradient for learning the witness function f , the desired regularization e ect of is ensured by early stopping and SGD [36, 37]. The case = 100 corresponds to the gradient ow w.r.t. Riemannian structure $d_{;}$, the case = 0 corresponds to unregularized MMD GAN that uses Euclidean gradients on the objective MMD (p; q) [6 8]. We see from Figure 1 that the kernelized ows induced by $d_{;}$ for > 0 ensure graceful convergence of the generator to the target distribution as predicted by Theorem 5.3, while GAN su ers from cycles and mode collapse for = 0 .

In Appendix F, we give trajectories for neural witness functions (i.e. is learned as well), and we see similar behavior to the xed kernel. We leave for future work the extension of $d_{;}$ and its gradient ows to neural function spaces as in [3, 38, 39]. Analysing the dynamic of the min-max optimization of parametric gradient regularized MMD GAN descent in the neural case will be interesting to conduct in the spirit of [40].

9 Conclusion

We propose an energy regularized gradient descent to MMD GAN training and derive a condition guaranteeing its global convergence. We demonstrate that the resulting ow coincides with the gradient ow of MMD with respect to a newly proposed Riemannian structure on the statistical manifold of probability distributions. Our investigation deepens the role of gradient regularization in GANs. Future directions include more understanding about the relationship between the static discrepancy and the dynamic discrepancy introduced in this paper, and investigation of other variational problems on a statistical manifold using the proposed Riemannian structure.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[2] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In NIPS, 2016.

[3] Tengyu Ma. Generalization and equilibrium in generative adversarial nets (gans) (invited talk). Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing - STOC 2018, 2018.

[4] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. ICML , 2017.

[5] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. JMLR , 2012.

[6] Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In ICML , 2015.

[7] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In UAI , 2015.

[8] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. Advances in Neural Information Processing Systems, 2017.

[9] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. Mcgan: Mean and covariance feature matching gan. ICML , 2017.

[10] Cicero Nogueira dos Santos, Youssef Mroueh, Inkit Padhi, and Pierre Dognin. Learning implicit generative models by matching perceptual features. In The IEEE International Conference on Computer Vision (ICCV) , October 2019.

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. Advances in Neural Information Processing Systems, 2017.

[12] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. ICLR , 2018.

[13] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. Advances in neural information processing systems, 2017.

[14] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In Advances in neural information processing systems, pages 2018 2028, 2017.

[15] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? Proceedings of the 35th International Conference on Machine Learning 2018.

[16] Michael Arbel, Dougal J. Sutherland, Mikolaj Binkowski, and Arthur Gretton. On gradient regularizers for mmd gans. Advances in Neural Information Processing Systems, 2018.

[17] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In Advances in neural information processing systems, pages 5585 5595, 2017.

[18] Ananya Uppal, Shashank Singh, and Barnabas Poczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In Advances in Neural Information Processing Systems, pages 9086 9097, 2019.

[19] Tengyuan Liang. On how well generative adversarial networks learn distributions. Preprint arXiv:1811.03179, 2020.

[20] Leon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. In Braverman Readings in Machine Learning, pages 229 268, 2017.

[21] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyonds. Foundations and Trends in Machine Learning, pages 1 141, 2017.

[22] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In Aistats, Proceedings of Machine Learning Research, 2019.

[23] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient ow. Advances in Neural Information Processing Systems, 2019.

[24] M Arbel, A Gretton, W Li, and G Montufar. Kernelized wasserstein natural gradient. In International Conference on Learning Representations, 2020.

[25] Jean-David Benamou and Yann Brenier. A computational uid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik, 2000.

[26] Filippo Santambrogio. f Euclidean, metric, and Wasserstein g gradient ows: an overview. Bulletin of Mathematical Sciences, 7(1):87 154, 2017.

[27] Yifan Chen and Wuchen Li. Optimal transport natural gradient for statistical manifolds with continuous sample space. Information Geometry, 2018.

[28] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Advances in Neural Information Processing Systems 29, 2016.

[29] Qiang Liu. Stein variational descent as a gradient ow. Advances in neural information processing systems, 2017.

[30] A. Duncan, N. Nuesken, and L. Szpruch. On the geometry of stein variational gradient descent. Preprint arXiv:1912.00894, 2019.

[31] Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators. In Advances in Neural Information Processing Systems, pages 12964 12976, 2019.

[32] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. Advances in neural information processing systems, 2007.

[33] Francis Bach. Breaking the curse of dimensionality with convex neural networks. The Journal of Machine Learning Research, 18(1):629 681, 2017.

[34] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. Advances in Neural Information Processing Systems 22, pages 342 350, 2009.

[35] T. Tieleman and G. Hinton. Lecture 6.5 RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[36] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. Constructive Approximation, 2007.

[37] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM Rev., 60(2), 2018.

[38] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In Advances in neural information processing systems, pages 3036 3046, 2018.

[39] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Advances in Neural Information Processing Systems 31, pages 8571 8580, 2018.

[40] Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotsko , and Joan Bruna. A mean- eld analysis of two-player zero-sum games. Preprint arXiv:2002.06277, 2020.

[41] David G. Luenberger. Optimization by Vector Space Methods. John Wiley & Sons, Inc., USA, 1st edition, 1997.

On the Convergence of Gradient Descent in GANs: MMD GAN As a Gradient Flow Supplementary Material

A Proofs

The appendix parallels the paper in terms of sections and contains proofs of results presented within each Section.

A.1 Preliminaries

Proof of Proposition 2.3. The proof follows from noting that $q_\theta = (G_\theta)_\# \nu$ and $L_\theta(f) = \nabla_\theta \int f(x) q_\theta(dx) = \nabla_\theta[\langle f, \mu_{q_\theta}\rangle_{\mathcal{H}}]$.

[i)] L_θ and $L_\theta^\top$ are adjoint operators:

$$\begin{aligned}
\langle L_\theta(f), v\rangle &= \left\langle \int J G_\theta(z)\nabla f(G_\theta(z))\, \nu(dz), v\right\rangle \\
&= \int v^\top J G_\theta(z)\nabla f(G_\theta(z))\, \nu(dz) \\
&= \int \left\langle \nabla f(G_\theta(z)), J G_\theta(z)^\top v\right\rangle \nu(dz) \\
&= \left\langle f, \int \langle \nabla_x k(G_\theta(z), .), J G_\theta(z)^\top v\rangle \nu(dz)\right\rangle_{\mathcal{H}} \\
&= \left\langle f, L_\theta^\top v\right\rangle_{\mathcal{H}}.
\end{aligned}$$

[ii)] D_θ is symmetric, i.e., $\langle D_\theta f, g\rangle_{\mathcal{H}} = \langle f, D_\theta g\rangle_{\mathcal{H}}$ for $f, g \in \mathcal{H}$.

$$\begin{aligned}
\langle D_\theta f, g\rangle_{\mathcal{H}} &= \langle L_\theta^\top L_\theta f, g\rangle_{\mathcal{H}} \\
&= \langle L_\theta f, L_\theta g\rangle \text{ (applying (i))} \\
&= \langle f, L_\theta^\top L_\theta g\rangle_{\mathcal{H}} \text{ (applying (i) again)} \\
&= \langle f, D_\theta g\rangle_{\mathcal{H}}
\end{aligned}$$

[iii)] $\langle f, D_\theta f\rangle_{\mathcal{H}} = \|\nabla_\theta[\langle f, \mu_{q_\theta}\rangle_{\mathcal{H}}]\|^2 \geq 0$. In particular, D_θ is a positive operator and hence its spectrum is contained in $[0, \infty)$.

$$\begin{aligned}
\langle f, D_\theta f\rangle_{\mathcal{H}} = \langle L_\theta f, L_\theta f\rangle = ||L_\theta f||^2 &= \left\|\int J G_\theta(z)\nabla f(G_\theta(z))\, \nu(dz)\right\|^2 \\
&= \left\|\nabla_\theta \int f(G_\theta(z))\, \nu(dz)\right\|^2 \\
&= \left\|\left\langle f, \nabla_\theta \int k(G_\theta(z), .)\, \nu(dz)\right\rangle_{\mathcal{H}}\right\|^2 \\
&= \left\|\nabla_\theta \left\langle f, \int k(G_\theta(z), .)\, \nu(dz)\right\rangle_{\mathcal{H}}\right\|^2 \text{ (by regularity of the kernel and linearity)} \\
&= \left\|\nabla_\theta \langle f, \mu_{q_\theta}\rangle_{\mathcal{H}}\right\|^2
\end{aligned}$$

[iv)] For $f \in \mathcal{H}$, we have $(D_\theta f)(x) = \langle D_\theta(x, \cdot), f\rangle_{\mathcal{H}}$ with $D_\theta(y, y') := \iint \langle \partial_\theta k(G_\theta(z), y), \partial_\theta k(G_\theta(z'), y')\rangle \nu(dz)\nu(dz')$, where: $\partial_\theta k(G_\theta(z), y) = J G_\theta(z)\nabla_x k(G_\theta(z), y)$. Equivalently $D_\theta(y, y') = \iint \mathrm{Trace}(\nabla_x k(G_\theta(z), y) \Gamma_\theta(z, z')\nabla_x k(G_\theta(z'), y'))\, \nu(dz)\nu(dz')$:

$$\begin{aligned}
\langle f, D_\nu g\rangle_{\mathcal H} &= \langle L_\nu f, L_\nu g\rangle\\
&= \left\langle \int J_\theta G_\theta(z)\nabla f(G_\theta(z))\,\mu(dz), \int J_\theta G_\theta(z')\nabla g(G_\theta(z'))\,\mu(dz')\right\rangle\\
&= \int\int \langle J_\theta G_\theta(z)\nabla f(G_\theta(z)), J_\theta G_\theta(z')\nabla g(G_\theta(z'))\rangle\,\mu(dz)\,\mu(dz')\\
&= \int\int \left\langle \nabla f(G_\theta(z)), J_\theta G_\theta(z)^\top J_\theta G_\theta(z')\nabla g(G_\theta(z'))\right\rangle\,\mu(dz)\,\mu(dz')\\
&= \int\int \langle \nabla f(G_\theta(z)), \Gamma(z,z')\nabla g(G_\theta(z'))\rangle\,\mu(dz)\,\mu(dz')\\
&= \left\langle f, \int\int \mathrm{Trace}\left(\nabla_x k(G_\theta(z),\cdot)\,\Gamma(z,z')\nabla_x k(G_\theta(z'),\cdot)\right)\mu(dz)\,\mu(dz')\right\rangle_{\mathcal H};
\end{aligned}$$

hence we identify the operator D_ν. □

A.2 Parametric Energy Regularization of MMD

Proof of Proposition 3.2. It follows from the fact $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal H}$ that $\int f(x)p(dx) = \langle f, \mu_p\rangle_{\mathcal H}$ for $f \in \mathcal H$. This and definition (2) gives $\mathrm{MMD}_{\lambda,\beta}(p,q) = \sup_{f\in E_{\lambda,\beta}} \langle f, \mu_p - \mu_q\rangle_{\mathcal H}$. Then as the quadratic form in (3) is convex in $f \in \mathcal H$ and Slater's condition holds due to $2\langle f, (\lambda D + \beta I)f\rangle_{\mathcal H} - 1 < 0$ when $f = 0$, we can apply standard duality result [41, Thm. 8.7.1] to obtain

$$\mathrm{MMD}_{\lambda,\beta}(p,q) = \min_{\gamma\ge 0}\sup_{f\in\mathcal H}\left\{\langle f, \mu_p-\mu_q\rangle_{\mathcal H} + \gamma\left(1 - 2\langle f,(\lambda D+\beta I)f\rangle_{\mathcal H}\right)\right\}.$$

Since the above supremum is attained at the funtion f satisfying $\mu_p - \mu_q = 4\gamma(\lambda D + \beta I)f$, we further get

$$\begin{aligned}
\mathrm{MMD}_{\lambda,\beta}(p,q) &= \min_{\gamma\ge 0}\left\{\frac{1}{8\gamma}\left\langle \mu_p-\mu_q, (\lambda D+\beta I)^{-1}(\mu_p-\mu_q)\right\rangle_{\mathcal H} + \gamma\right\}\\
&= \sqrt{\frac{1}{2}\langle \mu_p-\mu_q, (\lambda D+\beta I)^{-1}(\mu_p-\mu_q)\rangle_{\mathcal H}}. \qquad (28)
\end{aligned}$$

By the same reasoning, we also have

$$\begin{aligned}
&\sup_{f\in\mathcal H}\left\{\int f(x)\,p(dx) - \int f(x)\,q(dx) - \frac{\lambda}{2}\|\nabla_\theta[\langle f, \mu_{q_\theta}\rangle_{\mathcal H}]\|^2 - \frac{\beta}{2}\|f\|^2_{\mathcal H}\right\}\\
&= \sup_{f\in\mathcal H}\langle f, \mu_p-\mu_q\rangle_{\mathcal H} - \frac{1}{2}\langle f,(\lambda D+\beta I)f\rangle_{\mathcal H} = \frac{1}{2}\left\langle \mu_p-\mu_q, (\lambda D+\beta I)^{-1}(\mu_p-\mu_q)\right\rangle_{\mathcal H}
\end{aligned}$$

and the witness function realizing this supremum is given by:

$$(\lambda D + \beta I)f^* = \mu_p - \mu_q.$$

This together with (28) gives the conclusion of the proposition. □

Proof of Corollary 3.3. By Proposition 3.2, we have

$$\mathrm{MMD}(p,q)^2 - 2\beta\,\mathrm{MMD}_{\lambda,\beta}(p,q)^2 = \|\mu_p-\mu_q\|^2_{\mathcal H} - \beta\langle \mu_p-\mu_q, f^*\rangle_{\mathcal H} = \langle \mu_p-\mu_q, \mu_p-\mu_q - \beta f^*\rangle_{\mathcal H}.$$

But equation (4) for f^* implies that $\mu_p - \mu_q - \beta f^* = \lambda D f^*$. Therefore, we obtain

$$\begin{aligned}
\mathrm{MMD}(p,q)^2 - 2\beta\,\mathrm{MMD}_{\lambda,\beta}(p,q)^2 &= \lambda\langle \mu_p-\mu_q, Df^*\rangle_{\mathcal H}\\
&= \lambda\langle(\lambda D+\beta I)f^*, Df^*\rangle_{\mathcal H} = \lambda\left[\lambda\|Df^*\|^2_{\mathcal H} + \beta\langle f^*, Df^*\rangle_{\mathcal H}\right] \ge 0. \qquad (29)
\end{aligned}$$

This in particular gives the first conclusion of the corollary. When $\lambda > 0$, it also shows that the equality happens if and only if $Df^* = 0$. We next claim that $Df^* = 0$ is equivalent to $D(\mu_p - \mu_q) = 0$. Indeed, if $Df^* = 0$ then it follows from (4) that $\mu_p - \mu_q = \beta f^*$ and hence $D(\mu_p - \mu_q) = \beta Df^* = 0$. Conversely, if $D(\mu_p - \mu_q) = 0$ then we obtain from (4) that $\lambda D(Df^*) + \beta Df^* = 0$ yielding $\lambda\langle D(Df^*), Df^*\rangle_{\mathcal H} + \beta\langle Df^*, Df^*\rangle_{\mathcal H} = 0$. As the two terms are nonnegative, this only happens if $Df^* = 0$. Thus we have proved the claim and the proof is complete. □

A.3 Generative Adversarial Networks via Parametric Regularized Flows

Proof of Theorem 4.1. From the definition of F and by using (8), we have

$$\frac{dF(q_t)}{dt} = \left\langle \mu_p - \mu_{q_t}, \frac{d}{dt}\mu_{q_t}\right\rangle_{\mathcal{H}} = \left\langle \mu_p - \mu_{q_t}, D_{q_t} f_t\right\rangle_{\mathcal{H}}.$$

But from the identity (29) we obtain

$$\left\langle \mu_p - \mu_{q_t}, D_{q_t} f_t\right\rangle_{\mathcal{H}} = \frac{2}{\lambda}\left[F(q_t) - \mathrm{MMD}_{\mathcal{H};\lambda}(p;q_t)^2\right].$$

Therefore, we conclude that the equality in (11) holds. On the other hand, the nonpositivity in (11) and the strict inequality assertion are just a consequence of Corollary 3.3. □

B Non-Asymptotic Convergence Of Gradient Descent In MMD GAN

Proof of Lemma 5.1. It is easy to see that

$$\mathrm{MMD}(p;q)^2 = \langle \mu_q - \mu_p, \mu_q - \mu_p\rangle_{\mathcal{H}} = \iint k(x;y)\,q(dx)\,q(dy) - 2\int \mu_p(x)\,q(dx) + \|\mu_p\|_{\mathcal{H}}^2.$$

It follows that

$$\begin{aligned}
\frac{d}{d\varepsilon}F(q_{\theta+\varepsilon v}) &= \frac{1}{2}\frac{d}{d\varepsilon}\iint k(G_{\theta+\varepsilon v}(z);G_{\theta+\varepsilon v}(z'))\,\nu(dz)\,\nu(dz') - \frac{d}{d\varepsilon}\int \mu_p(G_{\theta+\varepsilon v}(z))\,\nu(dz)\\
&= \frac{1}{2}\Big[\iint \langle \nabla_x k(G_{\theta+\varepsilon v}(z);G_{\theta+\varepsilon v}(z')), J_{\theta+\varepsilon v}G_{\theta+\varepsilon v}(z)^\top v\rangle\,\nu(dz)\,\nu(dz')\\
&\quad + \iint \langle \nabla_y k(G_{\theta+\varepsilon v}(z);G_{\theta+\varepsilon v}(z')), J_{\theta+\varepsilon v}G_{\theta+\varepsilon v}(z')^\top v\rangle\,\nu(dz)\,\nu(dz')\Big]\\
&\quad - \int \langle \nabla \mu_p(G_{\theta+\varepsilon v}(z)), J_{\theta+\varepsilon v}G_{\theta+\varepsilon v}(z)^\top v\rangle\,\nu(dz).
\end{aligned}$$

Since $\nabla_x k(z;z') = \nabla_y k(z';z)$ due to the symmetry of $k(x;y)$, we deduce that

$$\begin{aligned}
\frac{d}{d\varepsilon}F(q_{\theta+\varepsilon v}) &= \iint \langle \nabla_x k(G_{\theta+\varepsilon v}(z);G_{\theta+\varepsilon v}(z')), J_{\theta+\varepsilon v}G_{\theta+\varepsilon v}(z)^\top v\rangle\,\nu(dz)\,\nu(dz')\\
&\quad - \int \langle \nabla \mu_p(G_{\theta+\varepsilon v}(z)), J_{\theta+\varepsilon v}G_{\theta+\varepsilon v}(z)^\top v\rangle\,\nu(dz)\\
&= \int (L^\top_{\theta+\varepsilon v} v)(G_{\theta+\varepsilon v}(z'))\,\nu(dz') - \langle L_{\theta+\varepsilon v}\mu_p, v\rangle.
\end{aligned}$$

Observe that as $L^\top_\theta v \in \mathcal{H}$, we have $(L^\top_\theta v)(G_\theta(z')) = \langle k(G_\theta(z');\cdot), L^\top_\theta v\rangle_{\mathcal{H}}$ which implies that $\int (L^\top_\theta v)(G_\theta(z'))\,\nu(dz') = \left\langle \int k(G_\theta(z');\cdot)\,\nu(dz'), L^\top_\theta v\right\rangle_{\mathcal{H}} = \langle \mu_{q_\theta}, L^\top_\theta v\rangle_{\mathcal{H}}$. Thus we can rewrite the above expression as

$$\frac{d}{d\varepsilon}F(q_{\theta+\varepsilon v}) = \langle \mu_{q_{\theta+\varepsilon v}}, L^\top_{\theta+\varepsilon v} v\rangle_{\mathcal{H}} - \langle L_{\theta+\varepsilon v}\mu_p, v\rangle = \langle \mu_p - \mu_{q_{\theta+\varepsilon v}}, L^\top_{\theta+\varepsilon v} v\rangle_{\mathcal{H}}.$$

□

Proof of Proposition 5.2. From Lemma 5.1 and by our choice of v_θ we get

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} F(q_{\theta+\varepsilon v_\theta}) = -\langle \mu_p - \mu_q, L^\top_\theta L_\theta f_\theta\rangle_{\mathcal{H}} = -\langle \mu_p - \mu_q, D_\theta f_\theta\rangle_{\mathcal{H}} = -\left[\|D_\theta f_\theta\|_{\mathcal{H}}^2 + \lambda\langle f_\theta, D_\theta f_\theta\rangle_{\mathcal{H}}\right].$$

Since $D_\theta(\mu_p - \mu_q) \neq 0$, we have from the proof of Corollary 3.3 that $D_\theta f_\theta \neq 0$. Therefore, the above expression is negative, i.e., v_θ is a descent direction. □

B.1 Discrete time descent for MMD GAN

In order to prove Theorem 5.3 we need two auxiliary lemmas. The first one is:

Lemma B.1. Assume that k and G satisfy (1) and (12)–(13). Then we have for any $\theta \in \Theta$, $\varepsilon \geq 0$, and $v \in \mathbb{R}^p$ that

$$\|\mu_{q_\theta} - \mu_{q_{\theta+\varepsilon v}}\|_{\mathcal{H}} \leq C_1 \varepsilon \|v\|; \qquad \|L^\top_{\theta+\varepsilon v} v\|_{\mathcal{H}} \leq C_2 \|v\|;$$
$$\|(L^\top_{\theta+\varepsilon v} - L^\top_\theta) v\|_{\mathcal{H}} \leq (C_3 + C_4)\varepsilon \|v\|^2;$$

where $C_1 := L\,\mathbb{E}_\nu[M(z)]$, $C_2 := \sqrt{dp}\,L\,\mathbb{E}_\nu[M(z)]$, $C_3 := \sqrt{p}\,\tilde{L}\,\mathbb{E}_\nu[M(z)^2]$, and $C_4 := \sqrt{d}\,L\,\mathbb{E}_\nu[\tilde{M}(z)]$.

Proof. We have

$$\|\mu_{q_\theta} - \mu_{q_{\theta+\varepsilon v}}\|_{\mathcal{H}} = \left\| \int \big(k(G_\theta(z), \cdot) - k(G_{\theta+\varepsilon v}(z), \cdot) \big)\, \nu(dz) \right\|_{\mathcal{H}} \leq \int \big\| k(G_\theta(z), \cdot) - k(G_{\theta+\varepsilon v}(z), \cdot) \big\|_{\mathcal{H}}\, \nu(dz).$$

Hence by using Lipschitz condition (1), we get

$$\|\mu_{q_\theta} - \mu_{q_{\theta+\varepsilon v}}\|_{\mathcal{H}} \leq L\varepsilon\|v\| \int M(z)\, \nu(dz) =: C_1 \varepsilon \|v\|.$$

In the following calculations, we use the observation that if $h = (h_1, \ldots, h_d) \in \mathcal{H}^d$ and $b \in \mathbb{R}^d$ then

$$\|\langle h, b\rangle\|_{\mathcal{H}} = \left\| \sum_{i=1}^d h_i b_i \right\|_{\mathcal{H}} \leq \sum_{i=1}^d \|h_i\|_{\mathcal{H}} |b_i| \leq \Big(\sum_{i=1}^d \|h_i\|^2_{\mathcal{H}} \Big)^{\frac{1}{2}} \Big(\sum_{i=1}^d b_i^2 \Big)^{\frac{1}{2}} = \|h\|_{\mathcal{H}} \|b\|.$$

Also note that condition (1) implies that

$$\|\nabla_x k(x, \cdot)\|_{\mathcal{H}} \leq \sqrt{d}\,L \quad \text{and} \quad \|J_\theta G_\theta(z)\| \leq \sqrt{p}\,M(z).$$

We next have

$$\begin{aligned}
\|L^\top_{\theta+\varepsilon v} v\|_{\mathcal{H}} &= \left\| \int \langle \nabla_x k(G_{\theta+\varepsilon v}(z), \cdot), J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top v\rangle\, \nu(dz) \right\|_{\mathcal{H}} \\
&\leq \int \big\| \langle \nabla_x k(G_{\theta+\varepsilon v}(z), \cdot), J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top v\rangle \big\|_{\mathcal{H}}\, \nu(dz) \\
&\leq \|v\| \int \|\nabla_x k(G_{\theta+\varepsilon v}(z), \cdot)\|_{\mathcal{H}} \|J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top\|\, \nu(dz) \\
&\leq \sqrt{dp}\,L\|v\| \int M(z)\, \nu(dz) =: C_2 \|v\|.
\end{aligned}$$

Finally, observe that

$$\|(L^\top_{\theta+\varepsilon v} - L^\top_\theta) v\|_{\mathcal{H}} = \left\| \int A(z;\theta)\, \nu(dz) \right\|_{\mathcal{H}} \leq \int \|A(z;\theta)\|_{\mathcal{H}}\, \nu(dz);$$

where

$$\begin{aligned}
A(z;\theta) :=\ & \langle \nabla_x k(G_{\theta+\varepsilon v}(z), \cdot), J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top v\rangle - \langle \nabla_x k(G_\theta(z), \cdot), J_\theta G_\theta(z)^\top v\rangle \\
=\ & \big\langle \nabla_x k(G_{\theta+\varepsilon v}(z), \cdot) - \nabla_x k(G_\theta(z), \cdot), J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top v \big\rangle \\
&+ \big\langle \nabla_x k(G_\theta(z), \cdot), \big(J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top - J_\theta G_\theta(z)^\top \big) v \big\rangle.
\end{aligned}$$

It follows that $\|(L^\top_{\theta+\varepsilon v} - L^\top_\theta) v\|_{\mathcal{H}} \leq I_1 + I_2$ with

$$\begin{aligned}
I_1 &\leq \|v\| \int \|\nabla_x k(G_{\theta+\varepsilon v}(z), \cdot) - \nabla_x k(G_\theta(z), \cdot)\|_{\mathcal{H}} \|J_{\theta+\varepsilon v} G_{\theta+\varepsilon v}(z)^\top\|\, \nu(dz) \\
&\leq \sqrt{p}\,\tilde{L}\|v\| \int \|G_{\theta+\varepsilon v}(z) - G_\theta(z)\| M(z)\, \nu(dz) \leq \sqrt{p}\,\tilde{L}\varepsilon\|v\|^2 \int M(z)^2\, \nu(dz) =: C_3 \varepsilon \|v\|^2
\end{aligned}$$

and

$$I_2 \le \|v\| \int \|\nabla_x k(G_\gamma(z);\cdot)\|_{\mathcal{H}} \, \big\| J_{\gamma+\varepsilon v} G_{\gamma+\varepsilon v}(z)^\top - J_\gamma G_\gamma(z)^\top \big\| \, \mu(dz)$$
$$\le \sqrt{d} L \varepsilon \|v\|^2 \int M(z)\, \mu(dz) =: C_4 \varepsilon \|v\|^2.$$

Notice that we have used condition (12) to estimate I_1 and condition (13) to estimate I_2. By combining the above estimates, we obtain the desired result. □

The second one is:

Lemma B.2. Under the assumptions of Theorem 5.3, we have

$$F(q_{\gamma_{\ell+1}}) - F(q_{\gamma_\ell}) \le -\frac{1}{2}\varepsilon_\ell \langle \sqrt{q_{\gamma_\ell}}; D_{\gamma_\ell} f_\ell \rangle_{\mathcal{H}} \quad \text{for all } \ell \ge 1. \tag{30}$$

Proof. Let $\ell \ge 1$ be arbitrary. For convenience, let us write $\gamma := \gamma_\ell$, $v := L_{\gamma_\ell} f_\ell$, $\gamma' := \gamma_\ell + \varepsilon_\ell v$, and $f := f_\ell$. Then by Lemma 5.1, we have

$$\frac{d}{d\varepsilon} F(q_{\gamma+\varepsilon v}) = -\langle \sqrt{q_{\gamma+\varepsilon v}}; L^\top_{\gamma+\varepsilon v} v \rangle_{\mathcal{H}} \quad \forall \varepsilon \in [0, \varepsilon_\ell].$$

Therefore,

$$\begin{aligned} F(q_{\gamma'}) - F(q_\gamma) &= \int_0^{\varepsilon_\ell} \frac{d}{d\varepsilon} F(q_{\gamma+\varepsilon v})\, d\varepsilon = -\int_0^{\varepsilon_\ell} \langle \sqrt{q_{\gamma+\varepsilon v}}; L^\top_{\gamma+\varepsilon v} v\rangle_{\mathcal{H}}\, d\varepsilon \\ &= -\varepsilon_\ell \langle \sqrt{q_\gamma}; L^\top_\gamma v\rangle_{\mathcal{H}} - \int_0^{\varepsilon_\ell} \Big[\langle \sqrt{q_{\gamma+\varepsilon v}}; L^\top_{\gamma+\varepsilon v} v\rangle_{\mathcal{H}} - \langle \sqrt{q_\gamma}; L^\top_\gamma v\rangle_{\mathcal{H}} \Big] d\varepsilon. \end{aligned} \tag{31}$$

To estimate the integral term, observe that

$$\begin{aligned} &\big|\langle \sqrt{q_{\gamma+\varepsilon v}}; L^\top_{\gamma+\varepsilon v} v\rangle_{\mathcal{H}} - \langle \sqrt{q_\gamma}; L^\top_\gamma v\rangle_{\mathcal{H}}\big| \\ &\quad = \big|\langle \sqrt{q_{\gamma+\varepsilon v}} - \sqrt{q_\gamma}; L^\top_{\gamma+\varepsilon v} v\rangle_{\mathcal{H}} + \langle \sqrt{q_\gamma}; (L^\top_{\gamma+\varepsilon v} - L^\top_\gamma) v\rangle_{\mathcal{H}}\big| \\ &\quad \le \|\sqrt{q_{\gamma+\varepsilon v}} - \sqrt{q_\gamma}\|_{\mathcal{H}} \|L^\top_{\gamma+\varepsilon v} v\|_{\mathcal{H}} + \|\sqrt{q_\gamma}\|_{\mathcal{H}} \|(L^\top_{\gamma+\varepsilon v} - L^\top_\gamma) v\|_{\mathcal{H}}. \end{aligned}$$

This together with Lemma B.1 and the definition of $F(q_\gamma)$ gives

$$\big|\langle \sqrt{q_{\gamma+\varepsilon v}}; L^\top_{\gamma+\varepsilon v} v\rangle_{\mathcal{H}} - \langle \sqrt{q_\gamma}; L^\top_\gamma v\rangle_{\mathcal{H}}\big| \le C\varepsilon \Big(1 + \sqrt{F(q_\gamma)}\Big) \|v\|^2$$

with $C > 0$ depending only on the constants C_1, C_2, C_3, C_4 given in Lemma B.1. Then by combining with (31) and $L^\top_\gamma v = D_\gamma f$ we arrive at

$$F(q_{\gamma'}) - F(q_\gamma) \le -\varepsilon_\ell \langle \sqrt{q_\gamma}; D_\gamma f\rangle_{\mathcal{H}} + \frac{C}{2}\varepsilon_\ell^2 \Big(1 + \sqrt{F(q_\gamma)}\Big)\|v\|^2.$$

Since $\langle \sqrt{q_\gamma}; D_\gamma f\rangle_{\mathcal{H}} = \langle (D_\gamma + \lambda I) f; D_\gamma f\rangle_{\mathcal{H}} = \|D_\gamma f\|^2_{\mathcal{H}} + \lambda\langle f; D_\gamma f\rangle_{\mathcal{H}}$ and $\|v\|^2 = \langle f; D_\gamma f\rangle_{\mathcal{H}}$ by iii) of Proposition 2.3, we also have $\|v\|^2 \le \frac{1}{\lambda}\langle \sqrt{q_\gamma}; D_\gamma f_\ell\rangle_{\mathcal{H}}$. Thus we infer further that

$$F(q_{\gamma_{\ell+1}}) - F(q_{\gamma_\ell}) \le -\Big[1 - C(2\lambda)^{-1}\varepsilon_\ell \Big(1 + \sqrt{F(q_{\gamma_\ell})}\Big)\Big] \varepsilon_\ell \langle \sqrt{q_{\gamma_\ell}}; D_{\gamma_\ell} f_\ell\rangle_{\mathcal{H}} \quad \forall \ell \ge 1, \tag{32}$$

where we have returned to the original notation. We claim that (32) implies in particular that $F(q_{\gamma_\ell}) \le F(q_{\gamma_1})$ for all $\ell \ge 1$. Indeed, from (32) for $\ell = 1$ and condition (18) we get $F(q_{\gamma_2}) - F(q_{\gamma_1}) \le 0$. Now if $F(q_{\gamma_\ell}) \le F(q_{\gamma_1})$ for some $\ell \ge 1$, then again (32) and condition (18) give

$$F(q_{\gamma_{\ell+1}}) - F(q_{\gamma_\ell}) \le -\Big[1 - C(2\lambda)^{-1}\varepsilon_\ell \Big(1 + \sqrt{F(q_{\gamma_1})}\Big)\Big] \varepsilon_\ell \langle \sqrt{q_{\gamma_\ell}}; D_{\gamma_\ell} f_\ell\rangle_{\mathcal{H}} \le 0.$$

Thus, $F(q_{\ell+1}) \;\; F(q_\ell) \;\; F(q_1)$. Therefore, it follows by induction that the claim holds true. Owing to this claim and condition (18), we deduce from (32) that

$$F(q_{\ell+1}) \;\; F(q_\ell) \qquad \left[1 \quad C(2\;) \;{}^{1}\varepsilon_\ell \;\; 1 + \sqrt{\overline{F(q_1)}} \right] \varepsilon_\ell \langle\; p \;\; q_\ell ; D_\ell f_\ell \rangle_H$$
$$\frac{1}{2}\varepsilon_\ell \langle\; p \;\; q \;; D_\ell f_\ell \rangle_H$$

for every $\ell \;\; 1$. □

Proof of Theorem 5.3. In order to not to clutter notations we note hereafter $\;:=\;\ell$, and $\;:=\;\ell$. We have $\langle\; p \;\; q_\ell ; D_\ell f_\ell \rangle_H = 2F(q_\ell) \;\; 2 \;\; \mathrm{MMD}_{;}\;\; (p; q_\ell)^2$ from (29). Thus by combining with Lemma B.2 we obtain

$$\frac{F(q_{\ell+1}) \;\; F(q_\ell)}{\varepsilon_\ell} \qquad {}^{1}F(q_\ell) + \qquad {}^{1}\mathrm{MMD}_{;}\;\; (p; q_\ell)^2 \quad 8\ell \quad 1:$$

Let $f\; _j; d_j g_{j=0 \ldots 1}$ be the eigensystem of D_ℓ, the eigenvalues are given in decreasing order. Let $_i(\;)$ be the smallest non zero eigenvalue.

$$\begin{aligned}
2\mathrm{MMD}_{;}\;\; (p; q_\ell)^2 &= \Big\langle\; p \;\; q_\ell ; f_\ell \Big\rangle_H \;=\; \langle (D_t + \;I\;) f_\ell ; f_\ell \rangle_H \\
&= \sum_{j=0}^{1} \;_j \langle f_\ell ; d_j \rangle^2 + \;\; \|f_\ell\|_H^2 \\
&= \sum_{j=0}^{i} \;_j \langle f_\ell ; d_j \rangle^2 + \;\; \|f_\ell\|_H^2 \\
&\quad\; _i(\;)\|f_\ell\|_H^2 \left(1 \quad \frac{\|P_{\mathrm{Null}(D_\ell)} f_\ell\|_H^2}{\|f_\ell\|_H^2} \right) + \;\; \|f_\ell\|_H^2 ;
\end{aligned}$$

where the projection on the null space of D_ℓ:

$$P_{\mathrm{Null}(D_\ell)} f_\ell = \sum_{j>i} \langle d_j ; f_\ell \rangle d_j :$$

Let $a(\ell; f_\ell) = 1 \quad \frac{\|P_{\mathrm{Null}(D_\ell)} f_\ell\|_H^2}{\|f_\ell\|_H^2}$. By our choice of $_1$ and $\varepsilon_\ell \;\; _1$, we have $f_\ell \notin \mathrm{Null}(D_\ell)$, since we choose $\varepsilon_\ell \;\; _1$ so that:

$$a(\ell; f_\ell) > \;\; > \;\; 0$$

We conclude that:

$$\Big\langle\; p \;\; q_\ell ; f_\ell \Big\rangle_H \quad \|f_\ell\|_H^2 \left(\;_i(\;) a(\ell; f_\ell) + \;\; \right)$$

Using this we obtain

$$\Big\langle\; p \;\; q_\ell ; f_\ell \Big\rangle_H \quad \|\; p \;\; q_\ell\|_H \|f_\ell\|_H \quad \|\; p \;\; q_\ell\|_H \sqrt{\frac{\Big\langle\; p \;\; q_\ell ; f_\ell \Big\rangle_H}{_i(\;) a(\ell; f_\ell) + \;}}$$

yielding

$$\Big\langle\; p \;\; q_\ell ; f_\ell \Big\rangle_H \quad \frac{1}{_i(\;) a(\ell; f_\ell) + \;} \|\; p \;\; q_\ell\|_H^2 :$$

Hence it follows that:

$$\mathrm{MMD}_{;}\;\; (p; q_\ell)^2 \quad \frac{F(q_\ell)}{_i(\;) a(\ell; f_\ell) + \;} :$$

Thus

$$\frac{F(q_{\ell+1}) \;\; F(q_\ell)}{\varepsilon_\ell} \qquad {}^{1}\; 1 \quad \frac{}{_i(\;) a(\ell; f_\ell) + \;} \quad F(q_\ell) = \frac{_i(\;) a(\ell; f_\ell)}{_i(\;) a(\ell; f_\ell) + \;} F(q_\ell)$$

which implies

$$F(q_{\ell+1}) \quad \Big[1 \quad \frac{"_\ell \quad _i(\ell)a(\ell;f_\ell)}{_i(\ell)a(\ell;f_\ell)+} \Big] F(q_\ell)$$

for every $\ell = 1;2;:::$ By iterating this estimate, we arrive at

$$F(q_{\ell+1}) \quad F(q_1) \prod_{j=1} \Big[1 \quad \frac{"_j \quad _i(_j)a(_j;f_j)}{_j \quad _i(_j)a(_j;f_j)+ \quad _j} \Big] \quad 8\ell \quad 1:$$

$$= F(q_1) \prod_{j=1} \Big[1 \quad "_j \quad _j \Big]$$

$$F(q_1)\exp(\sum_{j=1}^{\ell} "_j \quad _j):$$

In particular we have for for $\ell \quad \frac{}{2}$, and $\ell = \quad \ell \quad _i(\ell)$, we have noting that for all j, we have: $< a(_j;f_j) \quad 1$:

$$_j = \frac{_i(_j)a(_j;f_j)}{_j \quad _i(_j)a(_j;f_j)+ \quad _j} \quad \frac{_i(\;)}{2 \quad _j \quad _i(\;)} = \frac{}{2 \quad _j} \quad 1;$$

and hence we have:

$$F(q_{\ell+1}) \quad F(q_1)\exp(\sum_{j=1}^{\ell} "_j):$$

□

C Parametric Kernelized Flows for a General Functional

C.1 Dynamic Formulation For MMD on a Statistical Manifold

Proof of Theorem 6.1. By using the dynamic of q_t in (8), we have

$$\mathrm{MMD}(q_0;q_1) = k \quad q_1 \quad q_0 k_H = \int_0^1 \frac{d}{dt} \quad q_t \, dt \quad _H = \int_0^1 D \quad _t f_t \, dt \quad _H$$

which together with Jensen inequality yields

$$\mathrm{MMD}^2(q_0;q_1) = \int_0^1 D \quad _t f_t \, dt \quad ^2_H \quad \int_0^1 k D \quad _t f_t k_H^2 \, dt:$$

Thus it is enough to prove that there exists an admissible path$(_t;f_t)$ satisfying

$$\mathrm{MMD}^2(q_0;q_1) = \int_0^1 D \quad _t f_t \quad ^2_H \, dt: \tag{33}$$

Let $(_t;f_t)$ be the path given by assumption (20). In particular, we have

$$D \quad _t f_t = 2 \quad q_1 \quad q_t:$$

This together with a calculation in the proof of Theorem 4.1 gives

$$\frac{d}{dt}\mathrm{MMD}^2(q_t;q_1) = \quad 2 \Big\langle q_1 \quad q_t ; D \quad _t f_t \Big\rangle_H = \quad D \quad _t f_t \quad ^2_H :$$

It follows that this path satis es optimal property (33) since

$$\mathrm{MMD}^2(q_0;q_1) = \int_0^1 \frac{d}{dt}\mathrm{MMD}^2(q_t;q_1)dt = \int_0^1 D \quad _t f_t \quad ^2_H \, dt:$$

□

C.2 Regularized λ MMD and gradient flows on a Statistical Manifold

Proof of Theorem 6.3. Let $t \mapsto \theta_t \in \Theta$ be a differentiable curve passing through θ at $t = 0$ and with tangent vector

$$\partial_t \theta_t \big|_{t=0} = \xi = L_\theta \varphi \quad \text{for } \varphi \in \mathcal{H}. \tag{34}$$

Then by using the chain rule we obtain

$$\begin{aligned}\frac{dF(q_{\theta_t})}{dt}\Big|_{t=0} &= \Big\langle \nabla_\theta [F(q_{\theta_t})], \frac{d\theta_t}{dt}\Big\rangle_{\mathbb{R}^p}\Big|_{t=0} = \Big\langle \langle h_{\theta_t}, \nabla_\theta [\rho_{q_{\theta_t}}]\rangle_{\mathcal{H}}, \frac{d\theta_t}{dt}\Big\rangle_{\mathbb{R}^p}\Big|_{t=0} \\ &= \Big\langle h_{\theta_t}, \nabla_\theta [\rho_{q_{\theta_t}}]\frac{d\theta_t}{dt}\Big\rangle_{\mathcal{H}}\Big|_{t=0} = \Big\langle h_{\theta_t}, \frac{d}{dt}[\rho_{q_{\theta_t}}]\Big\rangle_{\mathcal{H}}\Big|_{t=0}.\end{aligned}$$

Hence it follows from (34), (8), and (23) that

$$\frac{dF(q_{\theta_t})}{dt}\Big|_{t=0} = \langle h_\theta, D_\theta \varphi\rangle_{\mathcal{H}} = \langle (D_\theta + \lambda I)u, D_\theta \varphi\rangle = g_\theta(L_\theta u, \xi)_{\mathcal{H}}.$$

Therefore, we conclude from definition (21) that $\mathrm{grad}_{d_{\lambda,\theta}} F(q_\theta) = L_\theta u$. □

Gradient flows of Generic Functionals on the Statistical Manifold. Our framework is not limited to the MMD functional. In the following corollary we exhibit additional examples of functional $F(q_\theta)$ defined on the statistical manifold along with their gradient flows w.r.t. $d_{\lambda,\theta}$.

Corollary C.1.
- For the (potential) energy functional $F_1(q_\theta) = \int V(x) q_\theta(dx)$ with $V \in \mathcal{H}$, we have $\mathrm{grad}_{d_{\lambda,\theta}} F_1(q_\theta) = L_\theta u$; where $(D_\theta + \lambda I)u = V$.
- For the (entropy) functional $F_1(q_\theta) = \int f(\rho_{q_\theta}(x))dx$ with $f : \mathbb{R} \to \mathbb{R}$ being continuously differentiable, we have $\mathrm{grad}_{d_{\lambda,\theta}} F_1(q_\theta) = L_\theta u$; where $(D_\theta + \lambda I)u = \int f'(\rho_{q_\theta}(x))k(x,\cdot)dx$.
- For the (interaction) functional $F_3(q_\theta) = \int f(x)g(y)q_\theta(dx)q_\theta(dy)$ with $f, g \in \mathcal{H}$, we have $\mathrm{grad}_{d_{\lambda,\theta}} F_3(q_\theta) = L_\theta u$; where $(D_\theta + \lambda I)u = \langle f, \rho_{q_\theta}\rangle_{\mathcal{H}}\, g + \langle g, \rho_{q_\theta}\rangle_{\mathcal{H}}\, f$.

Proof of Corollary C.1. This is a consequence of Theorem 6.3 observing that $\partial_{\theta_i}[F_1(q_\theta)] = \langle V, \partial_{\theta_i}[\rho_{q_\theta}]\rangle_{\mathcal{H}}$ since $F_1(q_\theta) = \langle V, \rho_{q_\theta}\rangle_{\mathcal{H}}$ and $\partial_{\theta_i} F_2(q_\theta) = \langle \int f'(\rho_{q_\theta}(x))k(x,\cdot)dx, \partial_{\theta_i}\rho_{q_\theta}\rangle_{\mathcal{H}}$ since $F_2(q_\theta) = \int f(\langle \rho_{q_\theta}, k(x,\cdot)\rangle_{\mathcal{H}})dx$. Note also that as $F_3(q_\theta) = \langle f, \rho_{q_\theta}\rangle_{\mathcal{H}}\langle g, \rho_{q_\theta}\rangle_{\mathcal{H}}$, we have

$$\partial_{\theta_i}[F_3(q_\theta)] = \big\langle \langle f, \rho_{q_\theta}\rangle_{\mathcal{H}}\, g + \langle g, \rho_{q_\theta}\rangle_{\mathcal{H}}\, f, \partial_{\theta_i}[\rho_{q_\theta}]\big\rangle_{\mathcal{H}}.$$

□

Proof of Proposition 6.4. This general fact can be seen from the proof of Theorem 6.3. Indeed,

$$\begin{aligned}\frac{d}{dt}F(q_{\theta_t}) &= g_{\theta_t}(\mathrm{grad}_{d_{\lambda,\theta}} F(q_{\theta_t}), \partial_t \theta_t) = -g_{\theta_t}(L_{\theta_t}u_t, L_{\theta_t}u_t) \\ &= -\langle (D_{\theta_t} + \lambda I)u_t, D_{\theta_t}u_t\rangle_{\mathcal{H}} = -\langle h_{\theta_t}, D_{\theta_t}u_t\rangle_{\mathcal{H}}.\end{aligned} \tag{35}$$

On the other hand, we obtain from equation (25) for u_t that

$$\langle h_{\theta_t}, D_{\theta_t}u_t\rangle_{\mathcal{H}} + \lambda\langle h_{\theta_t}, u_t\rangle_{\mathcal{H}} = \|h_{\theta_t}\|^2_{\mathcal{H}}$$

which gives

$$\langle h_{\theta_t}, D_{\theta_t}u_t\rangle_{\mathcal{H}} = \frac{1}{\lambda}\Big[\|h_{\theta_t}\|^2_{\mathcal{H}} - \lambda\langle h_{\theta_t}, u_t\rangle_{\mathcal{H}}\Big] = \frac{1}{\lambda}\Big[\|h_{\theta_t}\|^2_{\mathcal{H}} - \lambda\langle h_{\theta_t}, (D_{\theta_t} + \lambda I)^{-1}h_{\theta_t}\rangle_{\mathcal{H}}\Big].$$

Therefore, we deduce the first conclusion of the proposition. From (35), we also see that $\frac{d}{dt}F(q_{\theta_t}) = 0$ if and only if $\|D_{\theta_t}u_t\|^2_{\mathcal{H}} + \lambda\langle u_t, D_{\theta_t}u_t\rangle_{\mathcal{H}} = 0$. That is, $\frac{d}{dt}F(q_{\theta_t}) = 0$ if and only if $D_{\theta_t}u_t = 0$. But by exactly the same reason as in the proof of Corollary 3.3, we have $D_{\theta_t}u_t = 0$ is equivalent to $D_{\theta_t}h_{\theta_t} = 0$. Thus the last conclusion of the proposition follows. □

D Convergence of Flows

Proof of Proposition 6.7. From Proposition 6.4, we have

$$\frac{d}{dt}F(q_t) = \frac{1}{\ }\Big[\ \|h_t\|_H^2 \qquad \langle h_t, u_t\rangle_H\ \Big]. \tag{36}$$

By the proof of Theorem 5.3 we have:

$$\langle h_t, u_t\rangle_H \qquad \frac{1}{{}_i(\ _t)a(\ _t, u_t) + \ }\|h_t\|_H^2.$$

This together with (36) and assumption (26) gives

$$\frac{d}{dt}F(q_t) \qquad \frac{1}{\ }\ 1 \qquad \frac{\ }{{}_i(\ _t)a(\ _t, u_t) + \ }\ \|h_t\|_H^2 \qquad \frac{{}_i(\ _t)a(\ _t, u_t)\ (\ _t)}{{}_i(\ _t)a(\ _t, u_t) + \ }F(q_t).$$

It follows that

$$\log\ \frac{F(q_t)}{F(q_0)} = \int_0^t \frac{d}{ds}\ \log(F(q_t))\ ds \qquad \int_0^t \frac{{}_i(\ _s)a(\ _s, u_s)\ (\ _s)}{{}_i(\ _s)a(\ _s, u_s) + \ }ds,$$

This gives decay estimate (27). □

E Algorithm

Algorithm 1 MMD GAN with Discrete Parametric Kernelized Flows w.r.t. $d_{;}$

Input: > 0 gradient penalty weight, Learning rate, n_c number of iterations for training the critic, N batch size

Initialize witness function $f(x) = \langle v; (x) \rangle$, Generator G, initialize $v; ;$. Note by p parameters of $v;$.

repeat

{Update of the witness function of $MMD^2_{;}$ $(p; G_{;\#})$ }

for $j = 1$ to n_c do

Sample a minibatch $x_i; i = 1 \ldots N; x_i$ p

Sample a minibatch $z_i; i = 1 \ldots N; z_i$

Compute $\widehat{MMD}^2_{;} (p; G_{;\#}( )) = M(p;)$

$M(p;) = \frac{1}{N}\sum_{i=1}^{N} f(G( )(z_i)) \quad \frac{1}{N}\sum_{i=1}^{N} f(x_i) + \quad r \quad \frac{1}{N}\sum_{i=1}^{N} f(G(z_i))^{2}$

{We omit the regularizer since SGD achieves similar e et}

{A better estimate of $\|r \quad E\, f(G(z))\|^2$ can be obtained by using N^2 independent samples or by using a variance estimate, but we found this estimate to be enough in practice }

if Fixed Kernel then

v v RmsProp($r_v M(p;)$)

else if Learned Kernel then

p p RmsProp($r_p M(p;)$)

end if

end for

{Generator update: Parametric Discrete Kernel Flow w.r.t. $d_{;}$ on the statistical Manifold }

Sample $z_i; i = 1 \ldots N; z_i$

$d \quad r \quad \frac{1}{N}\sum_{i=1}^{N} f(G(z_i))$

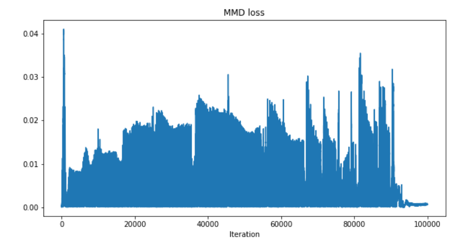
RmsProp(d)

until converges

F Additional Experiments and Plots

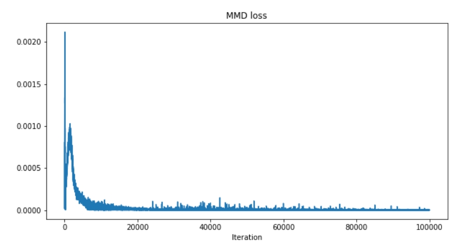
F.1 MMD GAN when the witness function is a multilayer Neural Network: i.e. Learned Kernel

(a) Fixed Kernel: Trajectories of Flows for = 0: Euclidean Gradients Flows



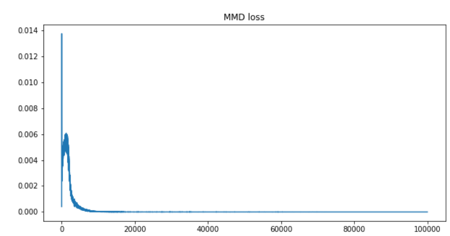
(b) MMD Loss

(c) Fixed Kernel: Trajectories of Flows for = 100: Kernelized Gradients Flows w.r.t. d ;



(d) MMD Loss

(e) Learned Kernel: Trajectories of Flows for = 100: Kernelized Gradients Flows w.r.t. d ;



(f) MMD Loss

Figure 2: Trajectories of Kernelized Gradient flows of the MMD functional for $\alpha = 0$ (no gradient regularization) and $\alpha = 100$. It is clear that the Riemannian structure induced by $d_{\alpha,\beta}$, $\alpha > 0$ guarantees the convergence, while we suffer from cycles and mode collapse for $\alpha = 0$. When the kernel is learned , i.e. the witness function is a multilayer neural network, the gradient flow exhibits qualitatively similar behavior to the fixed kernel one (compare Fig 1c) of the fixed kernel to Fig 2e of the learned kernel). This suggests possibly a Neural Tangent Kernel Regime that needs further analysis.