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

Youssef Mroueh* and Truyen Nguyen*

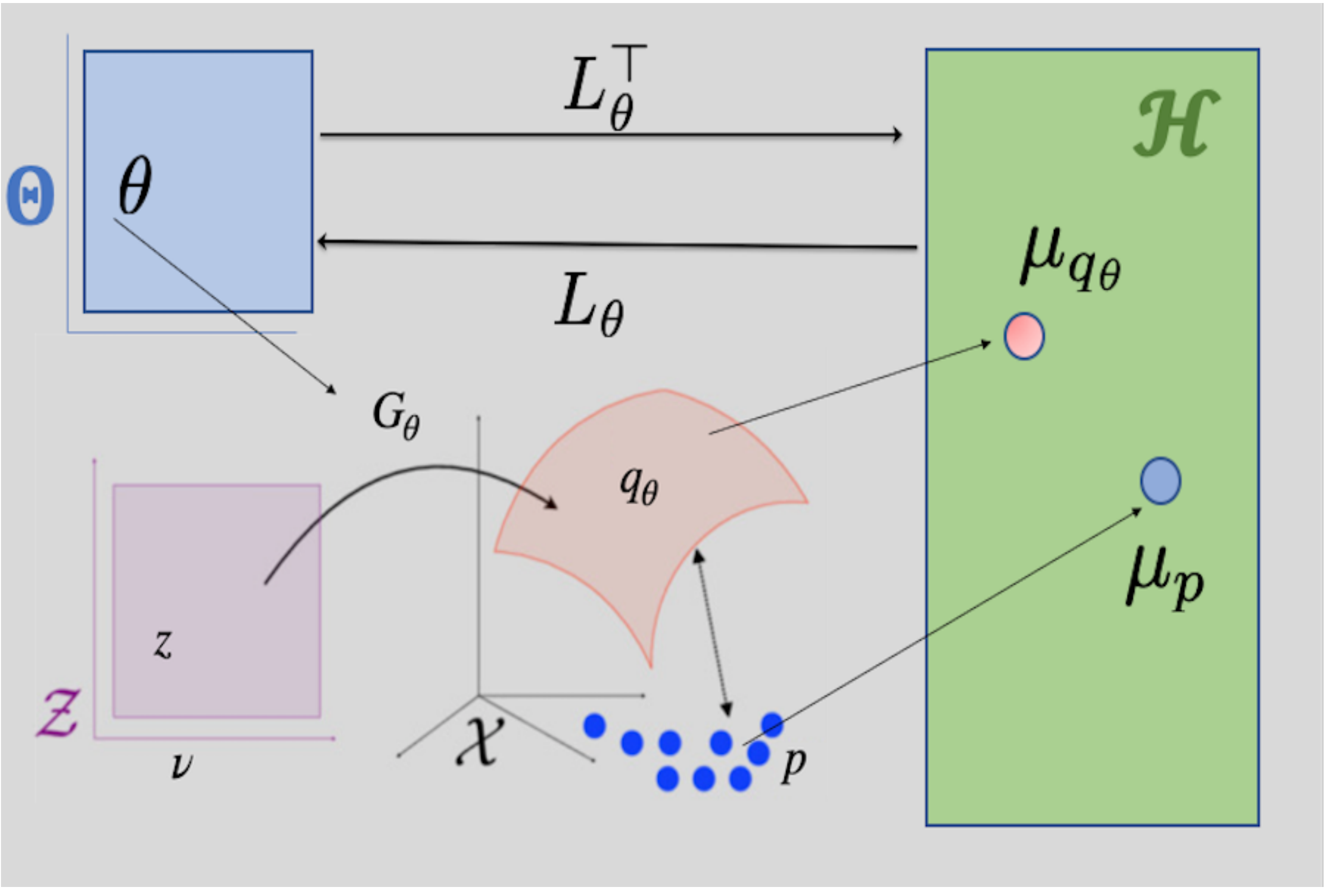
IBM Research & Akron University

https://arxiv.org/abs/2011.02402

Outline

- Parametric Gradient Regularization for MMD
- Assumptions for the convergence of MMD GAN both in the continuous and discrete case)
- New Riemannian structure and its gradient flow on the parametric statistical manifold
- Parametric Regularized MMD GAN as a gradient flow of the MMD functional w.r.t to the new Riemannian structure

Parametric Regularized MMD



Parametric Regularized MMD

$$\mathrm{MMD}_{\alpha,\beta}(p,q_\theta)^2 = \sup_{f\in\mathscr{H}} \left\{ \mathbb{E}_p f(x) - \mathbb{E}_{q_\theta} f(x) - \frac{\alpha}{2} \left\| \int \nabla_\theta f(G_\theta(z)) d\nu(z) \right\|^2_{\mathbb{R}^p} - \frac{\beta}{2} \|f\|^2_{\mathscr{H}} \right\}$$

$$d\theta_t = L_{\theta_t}(f_t)dt$$

choice of f_t $\quad (\alpha D(\theta_t) + \beta I) f_t = \mu_p - \mu_{q_{\theta_t}}$ Witness function of MMD_{α, β}

$$q_{\theta_t} = (G_{\theta_t})_{\#}(\nu)$$

Parametric Regularized MMD GAN

$$d\theta_t = L_{\theta_t}(f_t)dt$$

choice of f_t $$(\alpha D(\theta_t) + \beta I) f_t = \mu_p - \mu_{q_{\theta_t}}$$

Witness function of MMD_{α, β}

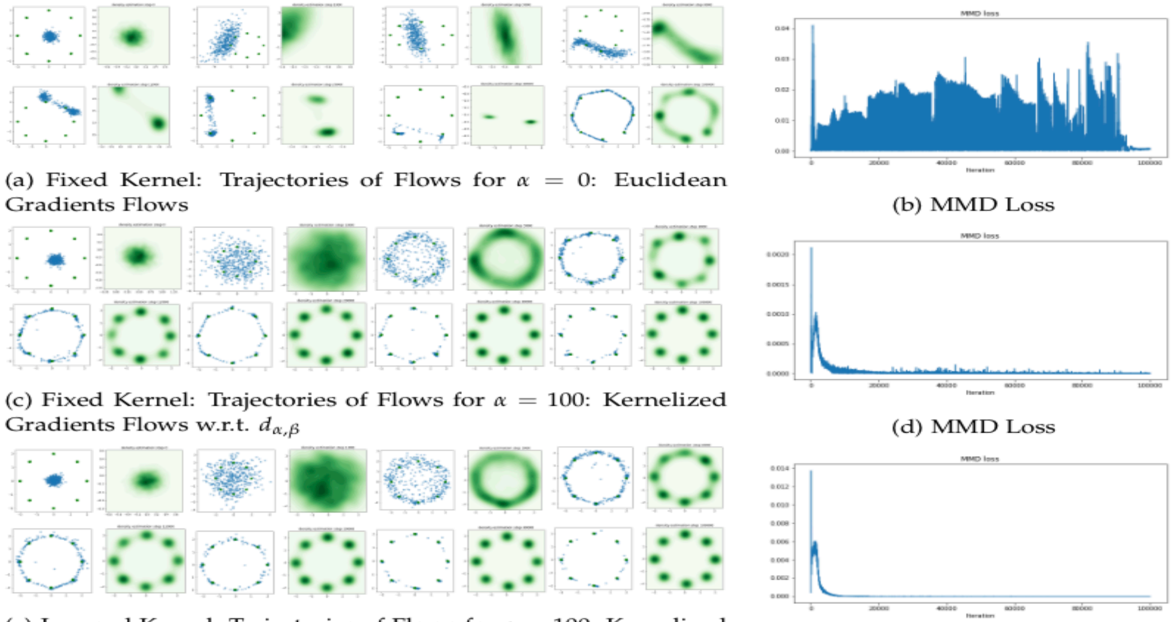
$$q_{\theta_t} = (G_{\theta_t})_{\#}(\nu)$$

Theorem 1 (Parametric Regularized Flows Decrease the MMD Distance). *Assume that $\alpha, \beta > 0$. Then the dynamic defined by the witness function of the parametric regularized MMD decreases the functional $\mathscr{F}(q_\theta)$:*

$$\frac{d\mathscr{F}(q_{\theta_t})}{dt} = -\frac{2}{\alpha}\left[\mathscr{F}(q_{\theta_t}) - \beta\, MMD_{\alpha,\beta}(p, q_{\theta_t})^2\right] \leq 0. \quad (1)$$

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

Experiments



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

(b) MMD Loss

(c) Fixed Kernel: Trajectories of Flows for $\alpha = 100$: Kernelized Gradients Flows w.r.t. $d_{\alpha,\beta}$

(d) MMD Loss

(e) Learned Kernel: Trajectories of Flows for $\alpha = 100$: Kernelized Gradients Flows w.r.t. $d_{\alpha,\beta}$

(f) MMD Loss

- Target Distribution is a mixture of Gaussian
- (a) Unregularized MMD GAN goes through cycles , does not converge
- (c) MMD GAN regularized with Parametric Gradients with fixed Kernel: convergent
- (e) MMD GAN regularized with Parametric Gradients with Learned Kernel: convergent

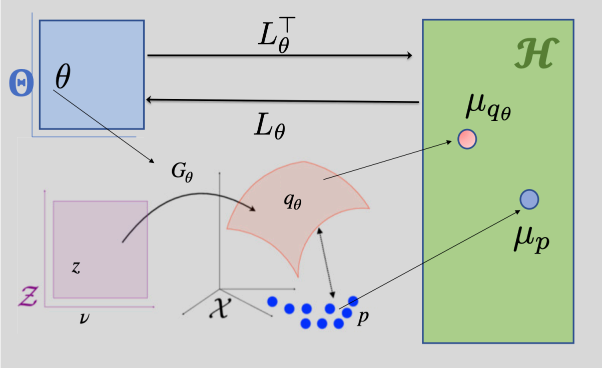
On the Convergence of Gradient Descent in GANs: MMD GAN As a Gradient Flow

Youssef Mroueh* and Truyen Nguyen*
IBM Research & Akron University

Outline

- Parametric Gradient Regularization for MMD
- Assumptions for the convergence of MMD GAN both in the continuous and discrete case)
- New Riemannian structure and its gradient flow on the parametric statistical manifold
- Parametric Regularized MMD GAN as a gradient flow of the MMD functional w.r.t to the new Riemannian structure

Parametric Regularized MMD I



$$\mathscr{F}(q_\theta) = \|\mu_p - \mu_{q_\theta}\|_{\mathscr{H}}$$

$$L_\theta : \mathscr{H} \to \mathbb{R}^p \quad L_\theta(f) := \int J_\theta G_\theta(z) \nabla f(G_\theta(z))\nu(dz)$$

$$f \to L_\theta(f)$$

$$L_\theta^\top : \mathbb{R}^p \to \mathscr{H}, \quad L_\theta^\top(v) := \int \langle \nabla_\theta [k(G_\theta(z)], .), v\rangle \, \nu(dz)$$

$$v \to L_\theta^\top(v)$$

Parametric Grammian $D_\theta : \mathscr{H} \to \mathscr{H}$ is defined by $D_\theta := L_\theta^\top L_\theta$.

Image Credit : Edited from original figure by Gabriel Peyré

Parametric Regularized MMD II

$$\mathrm{MMD}_{\alpha,\beta}(p, q_\theta)^2 = \sup_{f \in \mathscr{H}} \left\{ \mathbb{E}_p f(x) - \mathbb{E}_{q_\theta} f(x) - \frac{\alpha}{2} \left\| \int \nabla_\theta f(G_\theta(z)) d\nu(z) \right\|^2_{\mathbb{R}^p} - \frac{\beta}{2} \|f\|^2_{\mathscr{H}} \right\}$$

$$\mathrm{MMD}_{\alpha,\beta}(p, q_\theta)^2 = \sup_{f \in \mathscr{H}} \left\{ \langle f, \mu_p - \mu_{q_\theta} \rangle_{\mathscr{H}} - \frac{1}{2} \langle f, (\alpha D(\theta) + \beta I) f \rangle_{\mathscr{H}} \right\}$$

$$\mathrm{MMD}^2_{\alpha,\beta}(p, q_\theta) = \frac{1}{2} \left\langle \mu_{p-q_\theta}, (\alpha D_\theta + \beta I)^{-1} \mu_{p-q_\theta} \right\rangle_{\mathscr{H}}$$

witness function: $\quad (\alpha D_\theta + \beta I) f = \mu_p - \mu_{q_\theta}$

Parametric Regularized MMD GAN

$$d\theta_t = L_{\theta_t}(f_t) dt$$

choice of f_t $\quad (\alpha D(\theta_t) + \beta I) f_t = \mu_p - \mu_{q_{\theta_t}}$

$$q_{\theta_t} = (G_{\theta_t})_\#(\nu)$$

Theorem 1 (Parametric Regularized Flows Decrease the MMD Distance). *Assume that $\alpha, \beta > 0$. Then the dynamic defined by the witness function of the parametric regularized MMD decreases the functional $\mathscr{F}(q_\theta)$:*

$$\frac{d\mathscr{F}(q_{\theta_t})}{dt} = -\frac{2}{\alpha} \left[\mathscr{F}(q_{\theta_t}) - \beta \, MMD_{\alpha,\beta}(p, q_{\theta_t})^2 \right] \leq 0. \quad (1)$$

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

Riemannian Structure and Gradient Flow

Definition 1 (Regularized MMD on a Statistical Manifold). *Let $\alpha, \beta > 0$. Define*

$$d_{\alpha,\beta}(q_{\theta_0}, q_{\theta_1})^2 = \min_{\theta_t, f_t} \int_0^1 \left(\alpha \|D_{\theta_t} f_t\|^2_{\mathscr{H}} + \beta \langle f_t, D_{\theta_t} f_t \rangle_{\mathscr{H}} \right) dt,$$
$$\partial_t \theta_t = L_{\theta_t} f_t, \; f_t \in \mathscr{H}, \; \theta_{t=0} = \theta_0, \; \theta_{t=1} = \theta_1.$$

$$\mathscr{F}(q_\theta) = H(\mu_{q_\theta}) \qquad \partial_{\theta_t}[H(\mu_{q_\theta})] = \langle h_\theta, \partial_{\theta_t}[\mu_{q_\theta}] \rangle_{\mathscr{H}}$$

Consider

$$\partial_t \theta_t = -\underset{d_{\alpha,\beta}}{grad} \mathscr{F}(q_{\theta_t}) = -L_{\theta_t} u_t, \quad (1)$$

where

$$(\alpha D_{\theta_t} + \beta I) u_t = h_{\theta_t}. \quad (2)$$

MMD GAN as A Gradient Flow w.r.t $d_{\{\alpha\beta\}}$

- Parametric Gradient Regularized MMD GAN update

$$d\theta_t = L_{\theta_t}(f_t) dt$$

choice of f_t $\quad (\alpha D(\theta_t) + \beta I) f_t = \mu_p - \mu_{q_{\theta_t}}$

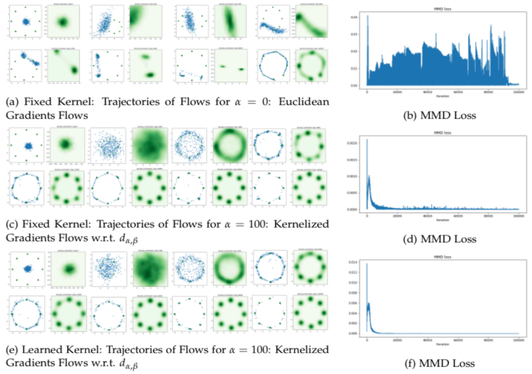
$$q_{\theta_t} = (G_{\theta_t})_\#(\nu)$$

$$\Updownarrow$$

$$\partial_t \theta_t = -\underset{d_{\alpha,\beta}}{\mathrm{grad}} \frac{1}{2} \mathrm{MMD}^2(p, q_{\theta_t})$$

- Parametric Gradient Flow of MMD w.r.t $d_{\{\alpha\beta\}}$

Experiments



(a) Fixed Kernel: Trajectories of Flows for $\alpha = 0$: Euclidean Gradients Flows
(b) MMD Loss
(c) Fixed Kernel: Trajectories of Flows for $\alpha = 100$: Kernelized Gradients Flows w.r.t. $d_{\alpha,\beta}$
(d) MMD Loss
(e) Learned Kernel: Trajectories of Flows for $\alpha = 100$: Kernelized Gradients Flows w.r.t. $d_{\alpha,\beta}$
(f) MMD Loss

- Target Distribution is a mixture of Gaussian
- (a) Unregularized MMD GAN goes through cycles , does not converge
- (c) MMD GAN regularized with Parametric Gradients with fixed Kernel: convergent
- (d) MMD GAN regularized with Parametric Gradients with Learned Kernel: convergent

Questions ,

mroueh@us.ibm.com
tn8@uakron.edu