Gradient Flows on the Maximum Mean Discrepancy

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Google Deepmind

First International Conference on Probabilistic Numerics (Probnum 2025)

Outline

MMD and MMD flow

- Introduction to MMD as an integral probability metric
- Connection with neural net training
- Wasserstein-2 Gradient Flow on the MMD
- Convergence: adaptive kernel
  - Neural Net implementation
  - Interpolation to χ^2

Arbel, Korba, Salim, G., Maximum Mean Discrepancy Gradient Flow (NeurIPS 2019)
Galashov, De Bortoli, G., Deep MMD Gradient Flow without adversarial training (ICLR 2025)
Chen, Mustafi, Glaser, Korba. G, Sriperumbudur (De)-regularized Maximum Mean Discrepancy Gradient Flow (submitted JMLR)

Outline

MMD and MMD flow

- Introduction to MMD as an integral probability metric
- Connection with neural net training
- Wasserstein-2 Gradient Flow on the MMD
- Convergence: adaptive kernel
  - Neural Net implementation
  - Interpolation to χ^2

Main motivation: gradient flow when the target distribution represented by samples

- A different kind of particle flow to diffusion models
- Neural network training dynamics

Arbel, Korba, Salim, G., Maximum Mean Discrepancy Gradient Flow (NeurIPS 2019)
Galashov, De Bortoli, G., Deep MMD Gradient Flow without adversarial training (ICLR 2025)
Chen, Mustafi, Glaser, Korba. G, Sriperumbudur (De)-regularized Maximum Mean Discrepancy Gradient Flow (submitted JMLR)

The MMD, and MMD flow

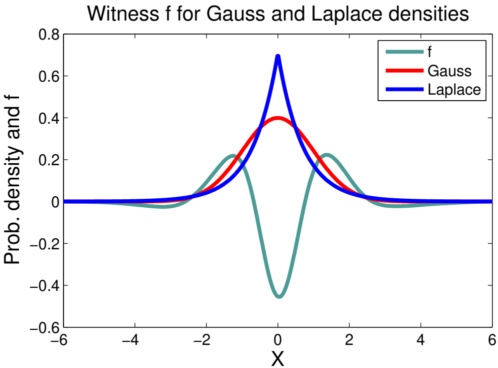
The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}}$$

$$\langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} = k(x, x')$$



The MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

$$MMD(P, Q; F) := \sup_{\|f\| \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{H}}$$

$$\langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} = k(x, x')$$

For characteristic RKHS $\mathcal{H}$, $MMD(P, Q) = 0$ iff $P = Q$

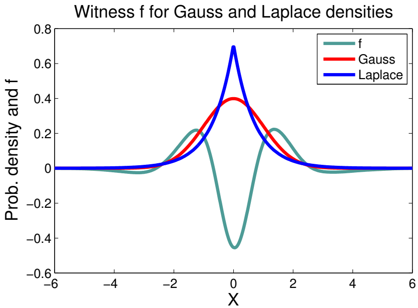
Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded varation 1 (Kolmogorov metric) [Müller, 1997]
- Bounded Lipschitz (Wasserstein distances) [Dudley, 2002]

The MMD and witness in closed form

The MMD:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{H}} \le 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$



The MMD and witness in closed form

The MMD:

$$MMD(P, Q)$$
$$= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$
$$= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{H}}$$

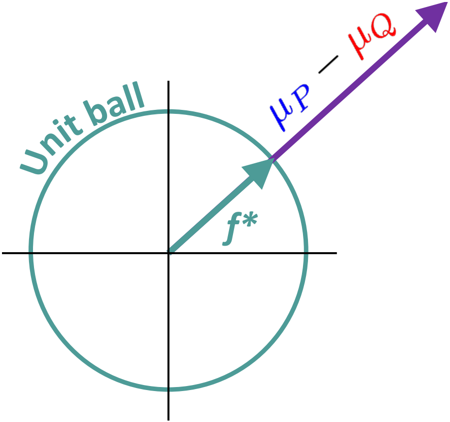
use

$$\begin{aligned} \mathrm{E}_P f(X) &= \mathrm{E}_P \langle \varphi(X), f \rangle_{\mathcal{H}} \\ &= \langle \mathrm{E}_P [\varphi(X)], f \rangle_{\mathcal{H}} \\ &= \langle \mu_P, f \rangle_{\mathcal{H}} \end{aligned}$$

The MMD and witness in closed form

The MMD:

$$MMD(P, Q) = \sup_{\|f\|_{\mathcal{H}} \le 1} [\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y)] = \sup_{\|f\|_{\mathcal{H}} \le 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{H}} = \|\mu_P - \mu_Q\|_{\mathcal{H}}$$



$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

The MMD and witness in closed form

The MMD:

$$
\begin{aligned}
& MMD(P, Q) \\
&= \sup_{\|f\|_{\mathcal{H}} \le 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right] \\
&= \sup_{\|f\|_{\mathcal{H}} \le 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{H}} \\
&= \|\mu_P - \mu_Q\|_{\mathcal{H}}
\end{aligned}
$$

$$
\begin{aligned}
f^*(x) &\propto \langle \mu_P - \mu_Q, \varphi(x) \rangle_H \\
&= \mathrm{E}_P k(X, x) - \mathrm{E}_Q k(Y, x)
\end{aligned}
$$

The MMD and witness in closed form

The MMD:

$$
\begin{aligned}
& MMD(P, Q) \\
& = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right] \\
& = \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{H}} \\
& = \|\mu_P - \mu_Q\|_{\mathcal{H}}
\end{aligned}
$$

In terms of kernels:

$$
\begin{aligned}
MMD^2(P, Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{H}} \\
&= \underbrace{\mathrm{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathrm{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathrm{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity.

MMD Flow (NeurIPS 19)

arXiv > stat > arXiv:1906.04370

Statistics > Machine Learning

[Submitted on 11 Jun 2019 (v1), last revised 3 Dec 2019 (this version, v2)]

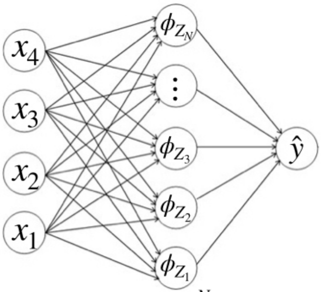
Maximum Mean Discrepancy Gradient Flow

Michael Arbel, Anna Korba, Adil Salim, Arthur Gretton



Motivation: Neural Net training

$(x, y) \sim data$



$$\min_{Z_1,\dots,Z_N} \mathbb{E}_{data}[\|y - \frac{1}{N}\sum_{i=1}^{N} \phi_{Z_i}(x)\|^2]$$

$$\min_{Z_1,\dots,Z_N \in \mathcal{Z}} \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}\right)$$

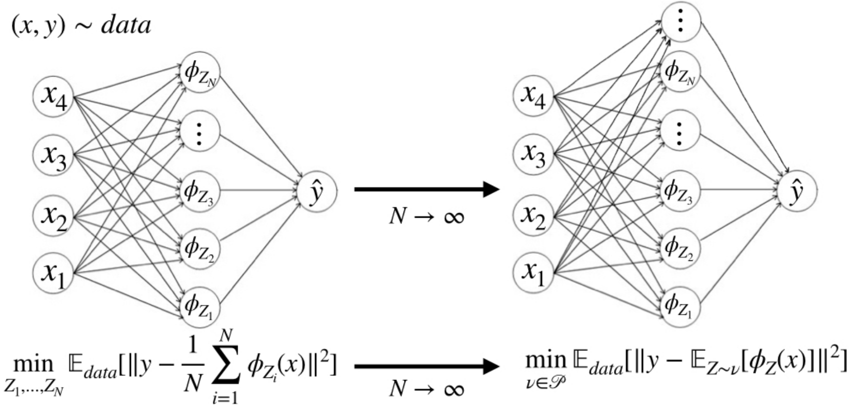
Optimization using gradient descent:

$$Z_i^{t+1} = Z_i^t - \gamma \nabla_{Z_i} \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i^t}\right)$$

Chizat, Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport", NeurIPS (2018)

Motivation: Neural Net training

$$\min_{Z_1,\ldots,Z_n \in \mathcal{Z}} \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}\right) \quad \xrightarrow[n\to\infty]{} \quad \min_{\nu\in\mathcal{P}} \mathcal{L}(\nu)$$



Chizat, Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport", NeurIPS (2018)

Motivation: Neural Net training

From previous slide:

$$\min_{\nu \in \mathcal{P}} \mathcal{L}(\nu) := \mathbb{E}_{(x,y)}[\|y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]\|^2]$$

Chizat, Bach. “On the global convergence of gradient descent for over-parameterized models using optimal transport”, NeurIPS (2018)

Motivation: Neural Net training

From previous slide:

$$\min_{\nu \in \mathcal{P}} \mathcal{L}(\nu) := \mathbb{E}_{(x,y)}[||y - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]||^2]$$

Connection to the MMD:

- Assume well-specified setting, $y(x) = \mathbb{E}_{U\sim\nu^\star}[\phi_U(x)]$
- Random feature formulation,

$$\mathcal{L}(\nu) = \mathbb{E}_x\left[||\mathbb{E}_{U\sim\nu^\star}[\phi_U(x)] - \mathbb{E}_{Z\sim\nu}[\phi_Z(x)]||^2\right] = MMD^2(\nu, \nu^\star)$$

- The kernel is: $k(U, Z) = \mathbb{E}_x[\phi_U(x)^\top \phi_Z(x)]$.

Chizat, Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport", NeurIPS (2018)

Intuition: MMD as "force field" on ν

Assume henceforth

$$\nu, \nu^* \in \mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \ : \ \int \|x\|^2 d\mu(x) < \infty \right\}.$$

MMD as free energy: target ν^*, current distribution ν

$$\mathcal{F}(\nu) := \frac{1}{2} MMD^2(\nu^*, \nu) = \frac{1}{2} \underbrace{\mathrm{E}_{\nu} k(x, x')}_{\text{interaction}} + \frac{1}{2} \underbrace{\mathrm{E}_{\nu^*} k(y, y')}_{\text{constant}} - \underbrace{\mathrm{E}_{\nu, \nu^*} k(x, y)}_{\text{confinement}}$$

[A] Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008)

Intuition: MMD as "force field" on ν

Assume henceforth

$$\nu, \nu^* \in \mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \ : \ \int \|x\|^2 d\mu(x) < \infty \right\}.$$

MMD as free energy: target ν^*, current distribution ν

$$\mathcal{F}(\nu) := \frac{1}{2} MMD^2(\nu^*, \nu) = \frac{1}{2} \underbrace{\mathrm{E}_{\nu} k(x, x')}_{\text{interaction}} + \frac{1}{2} \underbrace{\mathrm{E}_{\nu^*} k(y, y')}_{\text{constant}} - \underbrace{\mathrm{E}_{\nu, \nu^*} k(x, y)}_{\text{confinement}}$$

Consider $\{\mathrm{y}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu^*$ and $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu$.
Force on a particle z:

$$-\sum_j \nabla_z k(z, x_j) + \sum_j \nabla_z k(z, \mathrm{y}_j) = -\nabla_z \hat{f}_{\nu^*, \nu_t}(z)$$

Can we formalize this?

[A] Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008)

Wasserstein gradient flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ at μ is $h \in L^2(\mu)$ where $h : \mathbb{R}^d \to \mathbb{R}^d$.
Define $\nabla_{W_2}\mathcal{F}(\mu)$ of $\mathcal{F}$ at μ using Taylor expansion

$$\mathcal{F}((\mathrm{Id} + \epsilon h)_{\#\mu}) = \mathcal{F}(\mu) + \epsilon \left\langle \nabla_{W_2}\mathcal{F}(\mu), h \right\rangle_{L^2(\mu)} + o(\epsilon) \tag{1}$$

[A] Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008)

Wasserstein gradient flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ at μ is $h \in L^2(\mu)$ where $h : \mathbb{R}^d \to \mathbb{R}^d$.
Define $\nabla_{W_2}\mathcal{F}(\mu)$ of $\mathcal{F}$ at μ using Taylor expansion

$$\mathcal{F}((\mathrm{Id} + \epsilon h)_{\#\mu}) = \mathcal{F}(\mu) + \epsilon \left\langle \nabla_{W_2}\mathcal{F}(\mu), h \right\rangle_{L^2(\mu)} + o(\epsilon) \qquad (1)$$

The gradient flow is then:

$$\partial_t \nu_t = \mathrm{div}(\nu_t \nabla_{W_2}\mathcal{F}(\nu_t))$$

[A] Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008)

Wasserstein gradient flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ at μ is $h \in L^2(\mu)$ where $h : \mathbb{R}^d \to \mathbb{R}^d$.
Define $\nabla_{W_2}\mathcal{F}(\mu)$ of $\mathcal{F}$ at μ using Taylor expansion

$$\mathcal{F}((\mathrm{Id} + \epsilon h)_{\#\mu}) = \mathcal{F}(\mu) + \epsilon \left\langle \nabla_{W_2}\mathcal{F}(\mu), h \right\rangle_{L^2(\mu)} + o(\epsilon) \qquad (1)$$

The gradient flow is then:

$$\partial_t \nu_t = \mathrm{div}(\nu_t \nabla_{W_2}\mathcal{F}(\nu_t))$$

Under reasonable assumptions [A. Theorem 10.4.13]

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla \mathcal{F}'(\mu).$$

where first variation in direction ξ:

$$\mathcal{F}(\mu + \epsilon\xi) = \mathcal{F}(\mu) + \epsilon \int \mathcal{F}'(\mu)(x)\, d\xi(x) + o(\epsilon) \qquad \mu + \epsilon\xi \in \mathcal{P}_2(\mathbb{R}^d) \quad (2)$$

[A] Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008)

Wasserstein gradient flow on MMD

First variation of $\frac{1}{2}MMD^2(\nu^\star, \nu) =: \mathcal{F}(\nu)$

$$\mathcal{F}'(\nu)(z) := f_{\nu^\star,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^\star}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

The W_2 gradient flow of the MMD:

$$\partial_t \nu_t = \mathrm{div}(\nu_t \nabla_{W_2}\mathcal{F}(\nu_t)) = \mathrm{div}(\nu_t \nabla f_{\nu^\star,\nu_t})$$

Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008, Ch. 10)

Mroueh. Sercu, and Raj. Sobolev Descent. (AISTATS, 2019)

Arbel, Korba, Salim, G. (NeurIPS 2019)

Wasserstein gradient flow on MMD

First variation of $\frac{1}{2}MMD^2(\nu^\star, \nu) =: \mathcal{F}(\nu)$

$$\mathcal{F}'(\nu)(z) := f_{\nu^\star,\nu}(z) = 2\left(\mathbb{E}_{U\sim\nu^\star}[k(U,z)] - \mathbb{E}_{U\sim\nu}[k(U,z)]\right)$$

The W_2 gradient flow of the MMD:

$$\partial_t \nu_t = \mathrm{div}(\nu_t \nabla_{W_2}\mathcal{F}(\nu_t)) = \mathrm{div}(\nu_t \nabla f_{\nu^\star,\nu_t})$$

McKean-Vlasov dynamics for particles (existence and uniqueness under Assumption A):

$$dZ_t = -\nabla_{Z_t} f_{\nu^\star,\nu_t}(Z_t)dt, \qquad Z_0 \sim \nu_0$$

Assumption A: $k(x,x) \leq K$, for all $x \in \mathbb{R}^d$, $\sum_{i=1}^d \|\partial_i k(x,\cdot)\|^2 \leq K_{1d}$ and $\sum_{i,j=1}^d \|\partial_i \partial_j k(x,\cdot)\|^2 \leq K_{2d}$, d indicates scaling with dimension.

Ambrosio, Gigli, and Savaré. Gradient flows: in metric spaces and in the space of probability measures. (2008, Ch. 10)

Mroueh. Sercu, and Raj. Sobolev Descent. (AISTATS, 2019)

Arbel, Korba, Salim, G. (NeurIPS 2019)

Wasserstein gradient flow on the MMD

Forward Euler scheme [A, Section 2.2]:

$$\nu_{n+1} = (I - \gamma \nabla f_{\nu^\star, \nu_t})_{\#} \nu_n$$

$$Z_{n+1} = Z_n - \gamma \nabla_{Z_n} f_{\nu^\star, \nu_n}(Z_n), \qquad Z_0 \sim \nu_0,\ Z_n \sim \nu_n$$

Under Assumption A, ν_n approaches ν_t as $\gamma \to 0$

[A] Arbel, Korba, Salim, G. (NeurIPS 2019)

Wasserstein gradient flow on the MMD

Forward Euler scheme [A, Section 2.2]:

$$\nu_{n+1} = (I - \gamma \nabla f_{\nu^\star,\nu_t})_\# \nu_n$$
$$Z_{n+1} = Z_n - \gamma \nabla_{Z_n} f_{\nu^\star,\nu_n}(Z_n), \qquad Z_0 \sim \nu_0,\ Z_n \sim \nu_n$$

Under Assumption A, ν_n approaches ν_t as $\gamma \to 0$

Consistency? Does ν_t converge to $\nu^\star$ as $t \to \infty$?

[A] Arbel, Korba, Salim, G. (NeurIPS 2019)

Consistency

Can we use geodesic (displacement) convexity?

- A geodesic ρ_t between ν_1 and ν_2 is given by the transport map $T_{\nu_1}^{\nu_2} : \mathbb{R}^d \to \mathbb{R}^d$:

$$\rho_t = \left((1-t)\mathrm{Id} + tT_{\nu_1}^{\nu_2}\right)_{\#\nu_1}$$

Consistency

Can we use geodesic (displacement) convexity?

- A geodesic ρ_t between ν_1 and ν_2 is given by the transport map $T_{\nu_1}^{\nu_2} : \mathbb{R}^d \to \mathbb{R}^d$:

$$\rho_t = \left((1-t)\mathrm{Id} + tT_{\nu_1}^{\nu_2}\right)_{\#\nu_1}$$

- A functional $\mathcal{F}$ is displacement convex if:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_1) + t\mathcal{F}(\nu_2)$$

Consistency

Can we use geodesic (displacement) convexity?

- A geodesic ρ_t between ν_1 and ν_2 is given by the transport map $T_{\nu_1}^{\nu_2} : \mathbb{R}^d \to \mathbb{R}^d$:

$$\rho_t = \left((1-t)\mathrm{Id} + tT_{\nu_1}^{\nu_2}\right)_{\#\nu_1}$$

- A functional $\mathcal{F}$ is displacement convex if:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_1) + t\mathcal{F}(\nu_2)$$

MMD is not displacement convex in general (it is always mixture convex[1]).

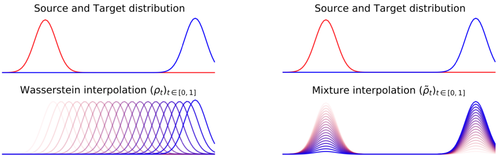
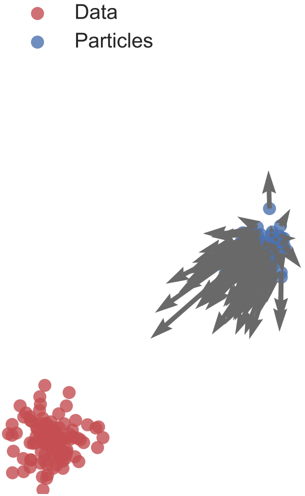


Figure from Korba, Salim, ICML 2022 Tutorial, "Sampling as First-Order Optimization over a space of probability measures"

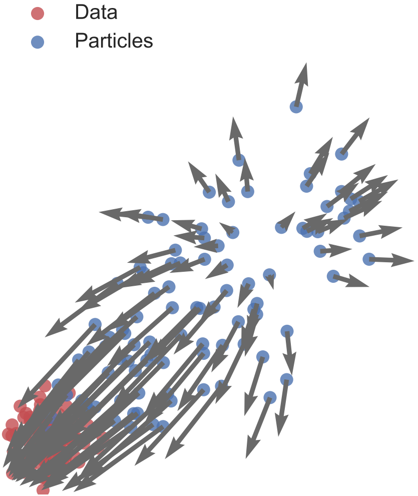
1. $\mathcal{F}(t\nu_1 + (1-t)\nu_2) \leq t\mathcal{F}(\nu_1) + (1-t)\mathcal{F}(\nu_2) \quad \forall t \in [0,1]$).

MMD flow in practice

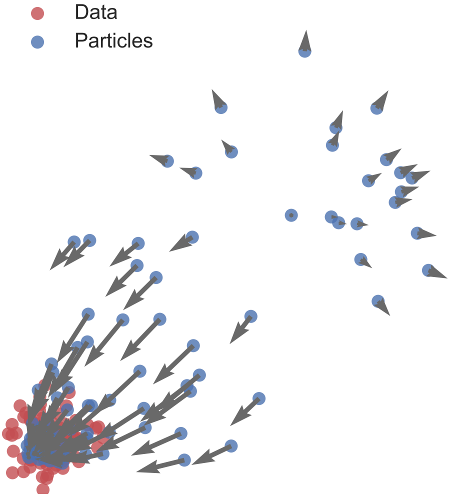
MMD flow in practice



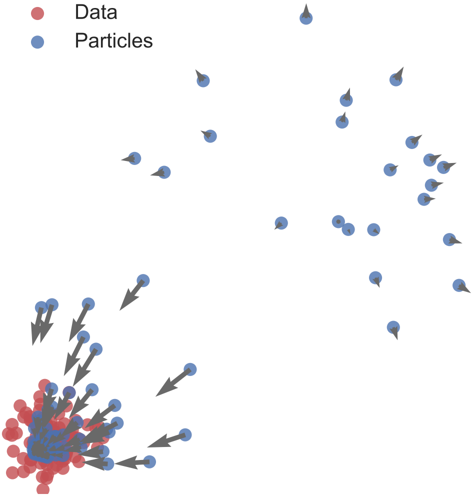
MMD flow in practice



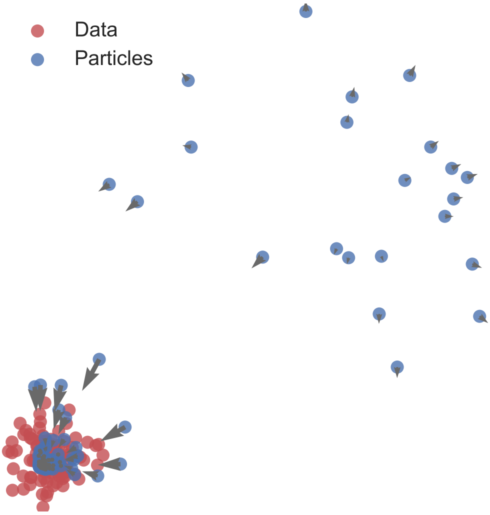
MMD flow in practice



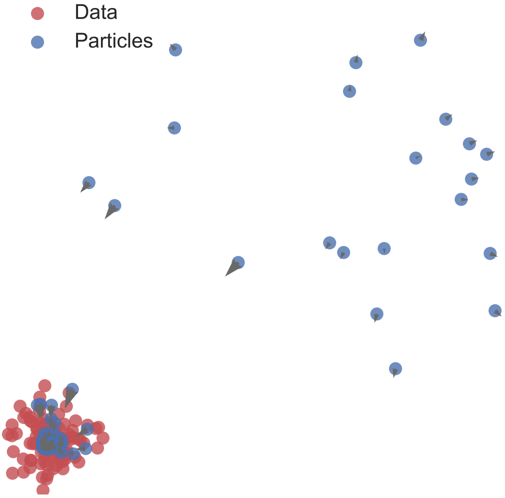
MMD flow in practice



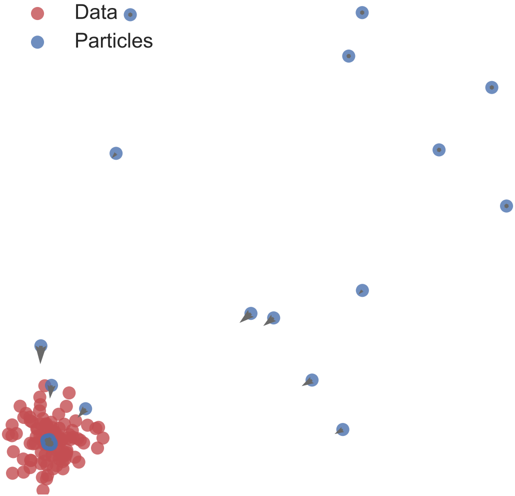
MMD flow in practice



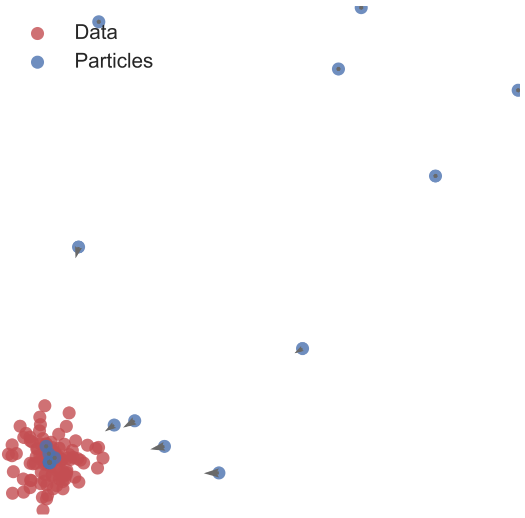
MMD flow in practice



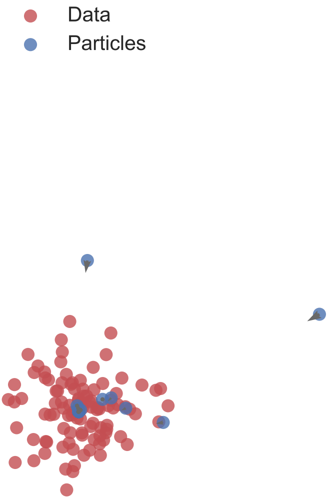
MMD flow in practice



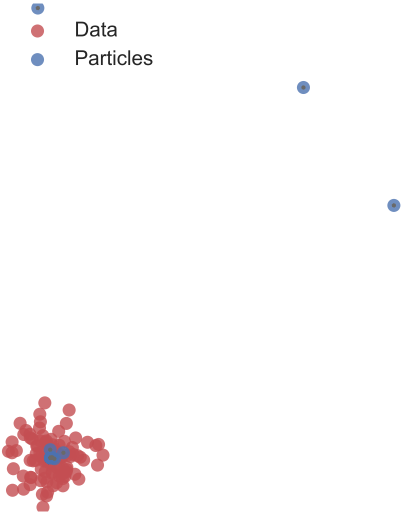
MMD flow in practice



MMD flow in practice



MMD flow in practice



Empirical observations

Some observations:

- Almost all particles tend to collapse at the center of mass m of the target $\nu^\star$, i.e.: $(\nu_t \simeq \delta_m)$
  - However, the loss stops decreasing: $\nabla f_{\nu^\star,\nu_t}(z) \simeq 0$ for z on the support of ν_t (and is small when far from $\nu^\star$)...
  - ...and in general, $\nabla f_{\nu^\star,\nu_t}(z) \neq 0$ outside the support of ν_t.

Empirical observations

Some observations:

- Almost all particles tend to collapse at the center of mass m of the target $\nu^\star$, i.e.: $(\nu_t \simeq \delta_m)$
  - However, the loss stops decreasing: $\nabla f_{\nu^\star,\nu_t}(z) \simeq 0$ for z on the support of ν_t (and is small when far from $\nu^\star$)...
  - ...and in general, $\nabla f_{\nu^\star,\nu_t}(z) \neq 0$ outside the support of ν_t.

Idea: Adapt the kernel according to distance of ν_t to $\nu^\star$.

- "Broad" kernel when distributions far apart,
- "narrow" kernel when they are close.

Noise injection in NeurIPS 2019 was a first attempt.

Noise injection for convergence

Noise injection: Evaluate $\nabla f_{\nu^\star,\nu_t}$ outside of the support of ν_t to get a better signal!

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to continuation methods,[1] but extended to interacting particles.
- Different from entropic regularization:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t) + \beta_t u_t$$

- Blur RKHS kernel with t-dependent Gaussian noise

Noise injection for convergence

Noise injection: Evaluate $\nabla f_{\nu^\star,\nu_t}$ outside of the support of ν_t to get a better signal!

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to continuation methods,[1] but extended to interacting particles.
- Different from entropic regularization:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t) + \beta_t u_t$$

- Blur RKHS kernel with t-dependent Gaussian noise

Noise injection for convergence

Noise injection: Evaluate $\nabla f_{\nu^\star,\nu_t}$ outside of the support of ν_t to get a better signal!

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to continuation methods,[1] but extended to interacting particles.
- Different from entropic regularization:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t) + \beta_t u_t$$

- Blur RKHS kernel with t-dependent Gaussian noise

Noise injection for convergence

Noise injection: Evaluate $\nabla f_{\nu^\star,\nu_t}$ outside of the support of ν_t to get a better signal!

- Sample $u_t \sim \mathcal{N}(0,1)$ and β_t is the noise level:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$$

- Similar to continuation methods,[1] but extended to interacting particles.
- Different from entropic regularization:

$$Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t) + \beta_t u_t$$

- Blur RKHS kernel with t-dependent Gaussian noise

Noise injection: consistency

Recall: $Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star, \nu_t}(Z_t + \beta_t u_t); \qquad Z_t \sim \nu_t$

Tradeoff for β_t

- Large β_t: $\nu_{t+1} - \nu_t$ not a descent direction any more: $\mathcal{F}(\nu_{t+1}) > \mathcal{F}(\nu_t)$
- Small β_t: does not converge

[A] Arbel, Korba, Salim, G. (NeurIPS 2019)

Noise injection: consistency

Recall: $Z_{t+1} = Z_t - \gamma \nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t)$; $\quad Z_t \sim \nu_t$

Tradeoff for β_t

- Large β_t: $\nu_{t+1} - \nu_t$ not a descent direction any more: $\mathcal{F}(\nu_{t+1}) > \mathcal{F}(\nu_t)$
- Small β_t: does not converge

Need β_t such that:

$$\mathcal{F}(\nu_{t+1}) - \mathcal{F}(\nu_t) \leq -C\gamma \mathbb{E}_{\substack{X_t \sim \nu_t \\ u_t \sim \mathcal{N}(0,1)}} [\|\nabla f_{\nu^\star,\nu_t}(X_t + \beta_t u_t)\|^2]$$

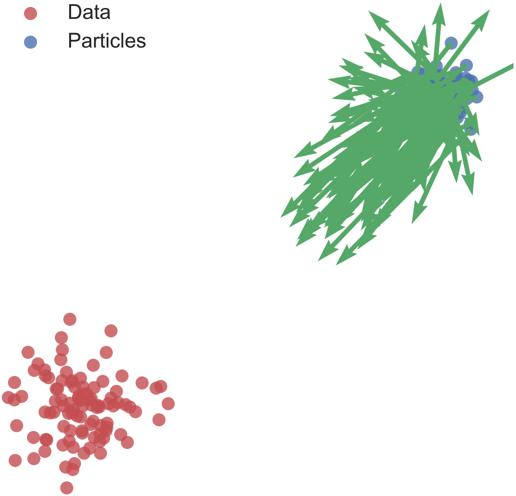
$$\sum_i^t \beta_i^2 \underset{t\to\infty}{\rightarrow} \infty$$

Then [A, Proposition 8]

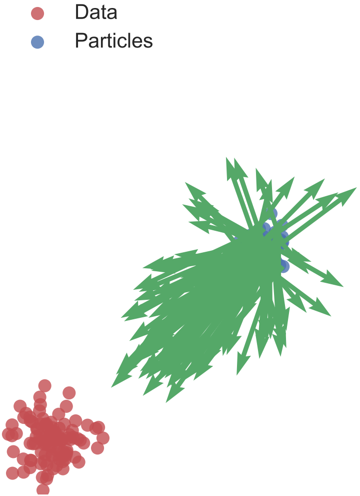
$$\mathcal{F}(\nu_t) \leq \mathcal{F}(\nu_0) e^{-C\gamma \sum_i^t \beta_i^2}.$$

[A] Arbel, Korba, Salim, G. (NeurIPS 2019)

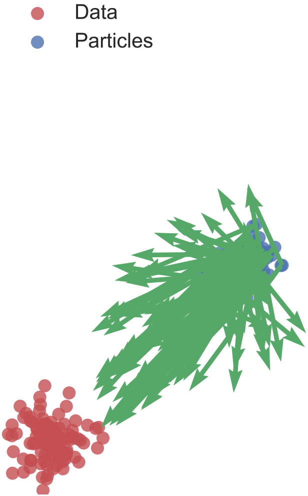
Noise injected MMD flow in practice



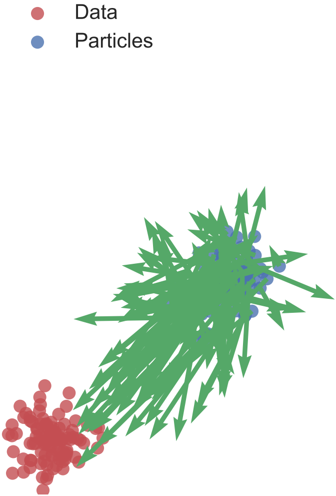
Noise injected MMD flow in practice



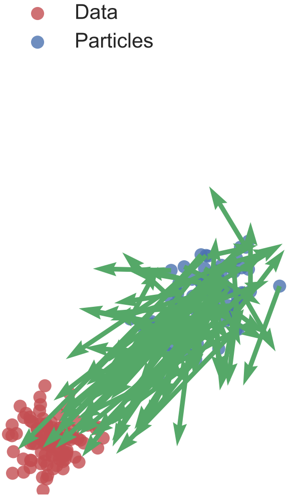
Noise injected MMD flow in practice



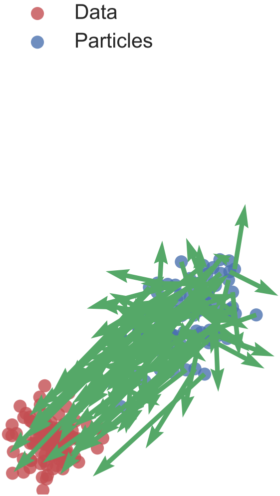
Noise injected MMD flow in practice



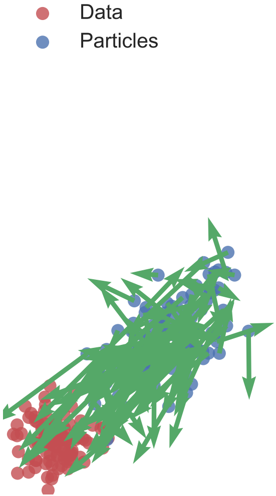
Noise injected MMD flow in practice



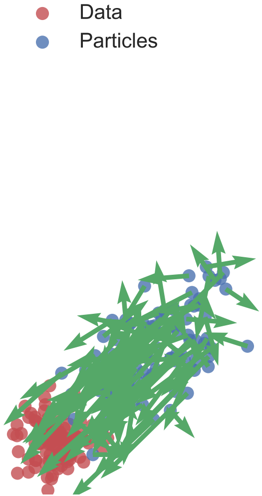
Noise injected MMD flow in practice



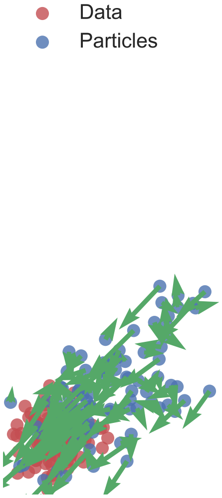
Noise injected MMD flow in practice



Noise injected MMD flow in practice

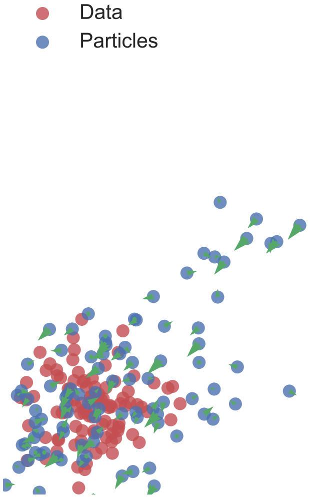


Noise injected MMD flow in practice



Noise injected MMD flow in practice

Noise injected MMD flow in practice



Noise injected MMD flow in practice

Adaptive MMD Flow (ICLR 25)

arXiv > cs > arXiv:2405.06780

Computer Science > Machine Learning

[Submitted on 10 May 2024]

Deep MMD Gradient Flow without adversarial training

Alexandre Galashov, Valentin de Bortoli, Arthur Gretton

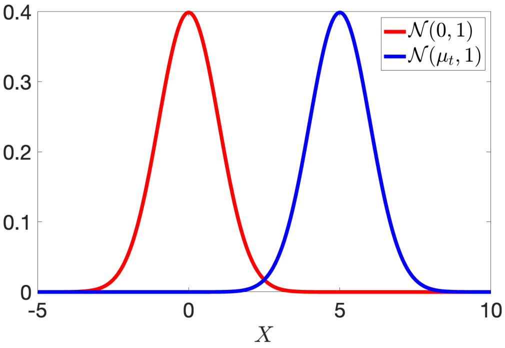
Will an adaptive kernel help?

Define the two measures:

$$\nu^\star := \mathcal{N}(0, \sigma^2 \mathrm{Id}) \qquad \nu_t := \mathcal{N}(\mu_t, \sigma^2 \mathrm{Id}).$$

Consider the family of MMDs:

$$\mathrm{MMD}_\alpha^2(\nu^\star, \nu_t) \qquad \text{with} \qquad k_\alpha(x, y) = \alpha^{-d} \exp[-\|x - y\|^2/(2\alpha^2)]$$



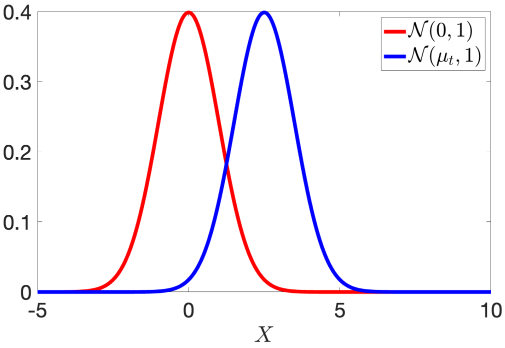
Will an adaptive kernel help?

Define the two measures:

$$\nu^\star := \mathcal{N}(0, \sigma^2 \mathrm{Id}) \qquad \nu_t := \mathcal{N}(\mu_t, \sigma^2 \mathrm{Id}).$$

Consider the family of MMDs:

$$\mathrm{MMD}^2_\alpha(\nu^\star, \nu_t) \qquad \text{with} \qquad k_\alpha(x, y) = \alpha^{-d} \exp[-\|x - y\|^2 / (2\alpha^2)]$$



Will an adaptive kernel help?

Choose kernel such that:

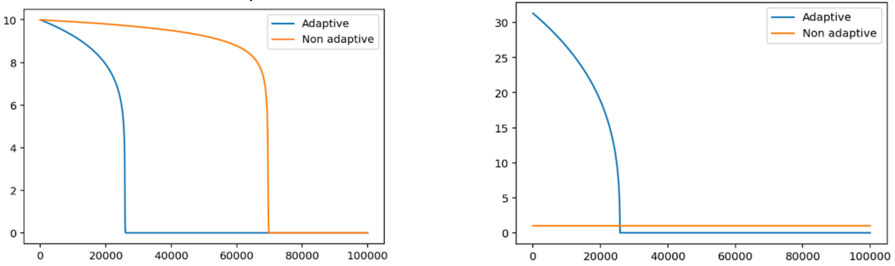
$$\alpha^\star = \text{argmax}_{\alpha \geq 0} \|\nabla_{\mu_t} \text{MMD}^2_\alpha(\nu^\star, \nu_t)\|.$$

Then

$$\alpha^\star = \text{ReLU}(\|\mu_t\|^2/(d+2) - 2\sigma^2)^{1/2}.$$

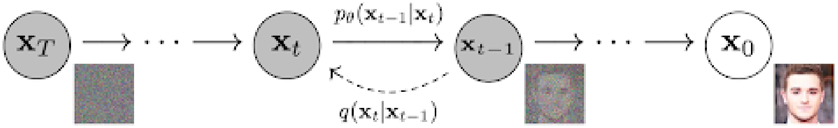
Value of mean μ_t

Kernel width α



How to train an adaptive MMD (1)

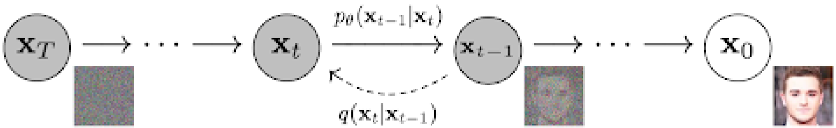
Diffusion:



Generate forward path $\tilde{\nu}_t$, $t \in [0,1]$, such that $\tilde{\nu}_0 = \nu^\star$, and $\tilde{\nu}_1 = \mathrm{N}(0, \mathrm{Id})$ is a Gaussian noise.

How to train an adaptive MMD (1)

Diffusion:



Generate forward path $\tilde{\nu}_t$, $t \in [0,1]$, such that $\tilde{\nu}_0 = \nu^\star$, and $\tilde{\nu}_1 = \mathrm{N}(0, \mathrm{Id})$ is a Gaussian noise.

Given samples $\tilde{x}_0 \sim \tilde{\nu}_0$, the samples $\tilde{x}_t | \tilde{x}_0$ are given by

$$\tilde{x}_t = \alpha_t \tilde{x}_0 + \beta_t \epsilon, \quad \epsilon \in \mathrm{N}(0, \mathrm{Id}),$$

with $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$.

- low t: $\tilde{x}_t$ close to the original data $\tilde{x}_0$,
- high t: $\tilde{x}_t$ close to a unit Gaussian

Schedule (α_t, β_t) is the variance-preserving one of Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole. Score-based generative modeling through stochastic differential equations (ICLR 2021)

How to train an adaptive MMD (2)

Time-dependent MMD training loss:

$$\mathcal{F}(\theta, t) := \frac{1}{2}\mathrm{E}_{\tilde{\nu}_t} k_{\theta,t}(\tilde{x}_t, \tilde{x}_t') + \mathrm{E}_{\tilde{\nu}_t,\nu^*} k_{\theta,t}(\tilde{x}_t, y)$$

with kernel

$$k_{\theta,t}(x, y) = \phi(x; t, \theta)^\top \phi(y; t, \theta)$$

and witness $f^{(\theta,t)}_{\nu^\star,\tilde{\nu}_t}$.

Gulrajani, Ahmed, Arjovsky, Dumoulin, Courville, Improved Training of Wasserstein GANs (NeurIPS 2017)

Binkowski, Sutherland, Arbel, G. (NeurIPS 2018)

How to train an adaptive MMD (2)

Time-dependent MMD training loss:

$$\mathcal{F}(\theta, t) := \frac{1}{2}\mathrm{E}_{\tilde{\nu}_t} k_{\theta,t}(\tilde{x}_t, \tilde{x}_t') + \mathrm{E}_{\tilde{\nu}_t,\nu^*} k_{\theta,t}(\tilde{x}_t, y)$$

with kernel

$$k_{\theta,t}(x, y) = \phi(x; t, \theta)^\top \phi(y; t, \theta)$$

and witness $f^{(\theta,t)}_{\nu^\star,\tilde{\nu}_t}$.

Train θ by minimizing noise-conditional loss on forward path:

$$\mathcal{F}_{\mathrm{tot}}(\theta, t) = \mathcal{F}(\theta, t) + \lambda_{\ell_2}\mathcal{F}_{\ell_2}(\theta, t) + \lambda_\nabla \mathcal{F}_\nabla(\theta, t),$$

$$\mathcal{F}_{\mathrm{tot}}(\theta) = \mathbb{E}_{t\sim U[0,1]}\left[\mathcal{F}_{\mathrm{tot}}(\theta, t)\right]$$

where

- $\mathcal{F}_{\ell_2}(\theta, t)$ is a "variance"-style penalty
- $\mathcal{F}_\nabla(\theta, t) = \frac{1}{N}\sum_{i=1}^N (\|\nabla f^{(\theta,t)}_{\nu^\star,\tilde{\nu}_t}(\tilde{x}_{t,i})\|_2 - 1)^2$, is a gradient penalty

Gulrajani, Ahmed, Arjovsky, Dumoulin, Courville, Improved Training of Wasserstein GANs (NeurIPS 2017)

Binkowski, Sutherland, Arbel, G. (NeurIPS 2018)

Sample generation

Algorithm Noise-adaptive MMD gradient flow

Sample initial particles $Z \sim \mathrm{N}(0, \mathrm{Id})$
Set $\Delta t = (t_{\max} - t_{\min})/T$
for $i = T$ to 0 **do**
  Set the noise level $t = i\Delta t$
  Set $Z_t^0 = Z$
  for $n = 0$ to $N_s - 1$ **do**
    $Z_t^{n+1} = Z_t^n - \eta \nabla f_{\nu^\star, \nu_t}^{(\theta^\star, t)}(Z_t^n)$
  end for
  Set $Z = Z_t^N$
end for
Output Z

Results

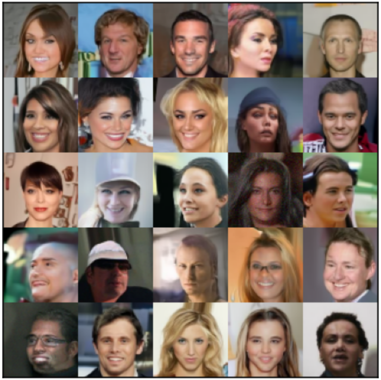
Table: Unconditional generation, CIFAR-10. MMD GAN (orig.), used mixed-RQ kernel. "Orig." – original paper, "impl." – our implementation.

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

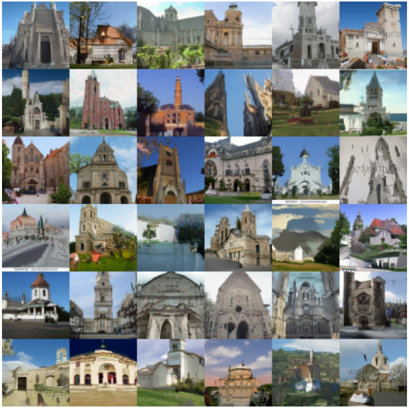
DDPM from (Ho et al., 2020). Discriminator flows include two KL gradient flows trained adversarially: JKO-Flow (Fan et al., 2022) and Deep Generative Wasserstein Gradient Flows (DGGF-KL) (Heng et al., 2023). GS-MMD-RK is Generative Sliced MMD Flows with Riesz Kernels (Hertrich et al., 2024)

Images

CELEB-A (64x64)



LSUN Church (64x64)



Summary

- Gradient flows based on kernel dependence measures
- NeurIPS 2019, NeurIPS 2021, ICLR 2025, JMLR (submitted)

NeurIPS 2019:

arXiv > stat > arXiv:1906.04370

Statistics > Machine Learning

[Submitted on 11 Jun 2019 (v1), last revised 3 Dec 2019 (this version, v2)]

Maximum Mean Discrepancy Gradient Flow

Michael Arbel, Anna Korba, Adil Salim, Arthur Gretton

NeurIPS 2021:

arXiv > stat > arXiv:2106.08929

Statistics > Machine Learning

[Submitted on 16 Jun 2021 (v1), last revised 29 Oct 2021 (this version, v2)]

KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support

Pierre Glaser, Michael Arbel, Arthur Gretton

Adaptive MMD (ICLR 25):

arXiv > cs > arXiv:2405.06780

Computer Science > Machine Learning

[Submitted on 10 May 2024]

Deep MMD Gradient Flow without adversarial training

Alexandre Galashov, Valentin de Bortoli, Arthur Gretton

(De)regularized MMD (JMLR, submitted):

arXiv > stat > arXiv:2409.14980

Search..

Help | Ac

Statistics > Machine Learning

[Submitted on 23 Sep 2024]

(De)-regularized Maximum Mean Discrepancy Gradient Flow

Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, Bharath K. Sriperumbudur

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT