Gradient Flows on Kernel Divergence Measures

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Measure-theoretic Approaches
and Optimal Transportation in Statistics, 2022

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, consistency
- Noise injection for improved convergence

KALE and KALE flow

- Introduction to KALE as a variational lower bound on the KL divergence
- Wasserstein-2 gradient flow on KALE
- Properties in relation to MMD

Arbel, Korba, Salim, G., Maximum Mean Discrepancy Gradient Flow (NeurIPS 2019)

Glaser, Arbel, G., KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support (NeurIPS 2021)

Motivation

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

Gradient flow on MMD

- MMD (and related IPMs) are GAN critics
- Understand dynamics of GAN training
- Neural network training dynamics

Gradient flow on KALE

- The KALE (and other lower bounds on ϕ-divergences) are GAN critics
- Understand dynamics of GAN training

Source and target might have disjoint support: KL undefined!

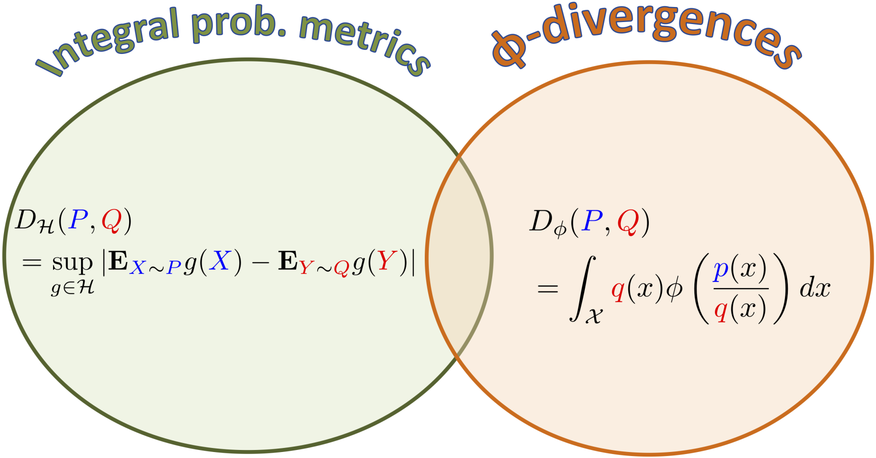
Binkowski, Sutherland, Arbel, G., Demystifying MMD GANs (ICLR 2018)

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

Arbel, Zhou, G. Generalized Energy-Based Models, (ICLR 2021)

Nowozin, Cseke, Tomioka, NeurIPS (2016)

Divergences



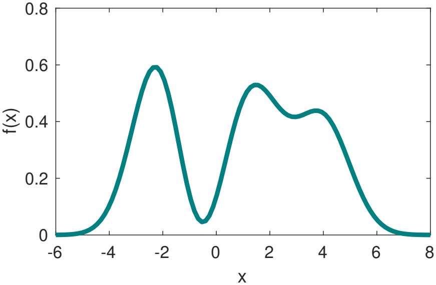
The MMD, and MMD flow

All of kernel methods

"The kernel trick"

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x)$$

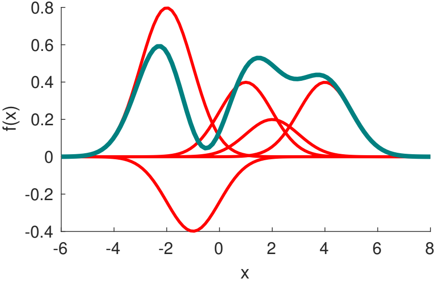
$$= \sum_{i=1}^{m} \alpha_i \underbrace{k(x_i, x)}_{\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{F}}}$$



All of kernel methods

"The kernel trick"

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \sum_{i=1}^{m} \alpha_i \underbrace{k(x_i, x)}_{\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{F}}}$$



$$f_\ell := \sum_{i=1}^{m} \alpha_i \varphi_\ell(x_i)$$

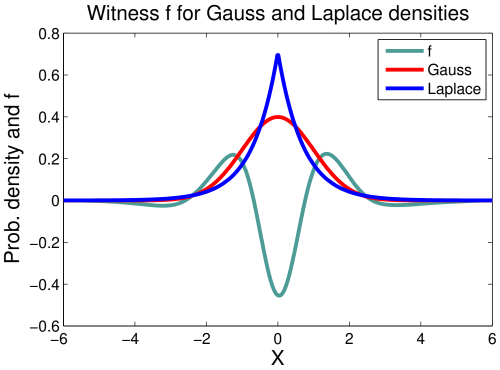
Function of infinitely many features expressed using m coefficients.

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

$$(F = \text{unit ball in RKHS } \mathcal{F})$$



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

For characteristic RKHS $\mathcal{F}$, $MMD(P, Q; F) = 0$ 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]

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

$$(F = \text{unit ball in RKHS } \mathcal{F})$$

A result for the proof on the next slide:

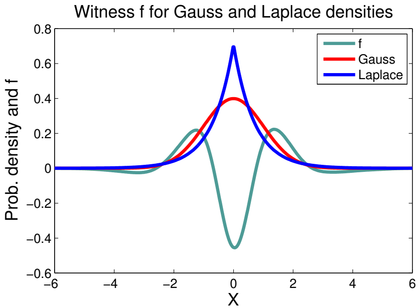
$$\mathrm{E}_P(f(X)) = \mathrm{E}_P \left\langle f, \varphi(X) \right\rangle_{\mathcal{F}} = \left\langle f, \mathrm{E}_P \varphi(X) \right\rangle_{\mathcal{F}} = \left\langle f, \mu_P \right\rangle_{\mathcal{F}}$$

(always true if kernel is bounded)

Integral prob. metric vs feature difference

The MMD:

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



Integral prob. metric vs feature difference

The MMD:

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

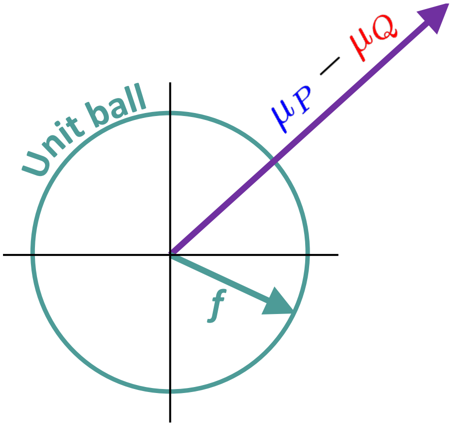
use

$$\mathrm{E}_P f(X) = \langle \mu_P, f \rangle_{\mathcal{F}}$$

Integral prob. metric vs feature difference

The MMD:

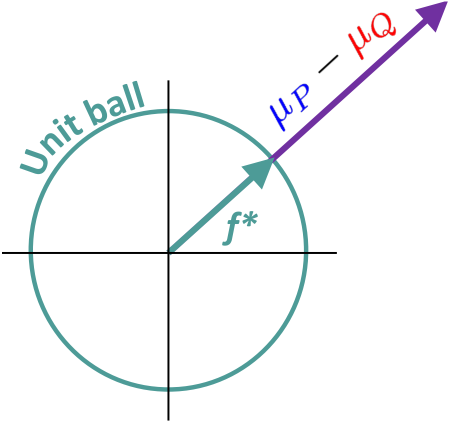
$$MMD(P, Q; F)$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathrm{E}_P f(X) - \mathrm{E}_Q f(Y) \right]$$
$$= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature difference

The MMD:

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



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

Integral prob. metric vs feature difference

The MMD:

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

$$f^*(x) \propto \mu_P(x) - \mu_Q(x) = \mathrm{E}_P k(X, x) - \mathrm{E}_Q k(Y, x)$$

Function view and feature view equivalent

Computing the MMD

The maximum mean discrepancy is the distance between feature means:

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

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

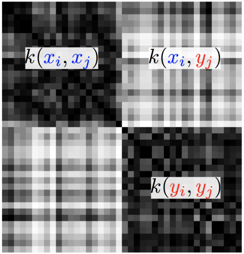
Computing the MMD

The maximum mean discrepancy is the distance between feature means:

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

Empirical estimate:

$$
\begin{aligned}
\widehat{MMD}^2 = &\frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) \\
&+ \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) \\
&- \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)
\end{aligned}
$$



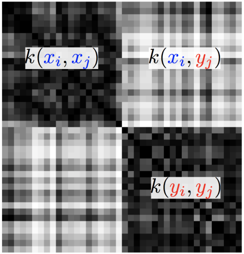
Computing the MMD

The maximum mean discrepancy is the distance between feature means:

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

Empirical witness:

$$\hat{f}_{\nu^\star,\nu_t}(z) \propto \sum_j k(z, x_j) - \sum_j k(z, y_j)$$

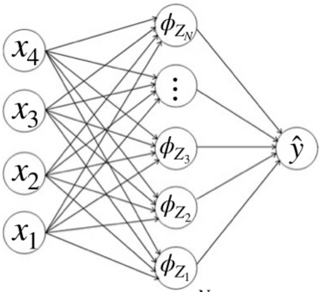


MMD Flow



Motivation: Neural Net training

$(x, y) \sim data$



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

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

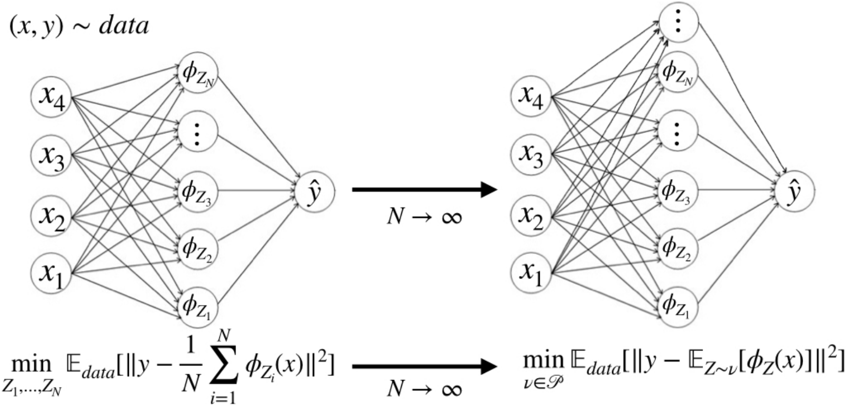
Optimization using gradient descent:

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

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

Motivation: Neural Net training

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



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:

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

Want to prove global convergence of GD when $n \to \infty$ and

$$\phi_Z(x) = w\, g_\theta(x), \qquad Z = (w, \theta)$$

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:

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

Want to prove global convergence of GD when $n \to \infty$ and

$$\phi_Z(x) = w\, g_\theta(x), \qquad Z = (w, \theta)$$

Connection to the MMD:

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

Preliminaries: Wasserstein gradient flow on MMD

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)

Preliminaries: Wasserstein gradient flow on MMD

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:

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

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

Wasserstein gradient flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ 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

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

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

Wasserstein gradient flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ 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

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

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 flows

Tangent space of $\left(\mathcal{P}_2(\mathbb{R}^d), W_2\right)$ 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

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

Under reasonable assumptions [A. Theorem 10.4.13]

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

where first variation in direction ξ:

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

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 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:

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

McKean-Vlasof 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 (1)

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[1] convex).

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

Consistency (2)

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow [A, Proposition 2]

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assume the dissipation rate is controlled (path-dependent Lojasiewicz inequality)

$$\mathcal{F}(\nu_t) \leq C\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- From above, [A, Proposition 7]:

$$\mathcal{F}(\nu_t) \leq \frac{1}{\mathcal{F}(\nu_0)^{-1} + 2C^{-1}t}$$

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

Consistency (2)

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow [A, Proposition 2]

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assume the dissipation rate is controlled (path-dependent Lojasiewicz inequality)

$$\mathcal{F}(\nu_t) \leq C\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- From above, [A, Proposition 7]:

$$\mathcal{F}(\nu_t) \leq \frac{1}{\mathcal{F}(\nu_0)^{-1} + 2C^{-1}t}$$

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

Consistency (2)

Dissipation inequalities:

- Rate by which $\mathcal{F}$ decreases along the gradient flow [A, Proposition 2]

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- Assume the dissipation rate is controlled (path-dependent Lojasiewicz inequality)

$$\mathcal{F}(\nu_t) \leq C\mathbb{E}_{\nu_t}[\|\nabla f_{\nu^\star,\nu_t}\|^2]$$

- From above, [A, Proposition 7]:

$$\mathcal{F}(\nu_t) \leq \frac{1}{\mathcal{F}(\nu_0)^{-1} + 2C^{-1}t}$$

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

Consistency (2)

Check: Lojasiewicz inequality for MMD?

- Does there exist $C > 0$ such that

$$\mathcal{F}(\nu_t) \leq C\mathbb{E}_{\nu_t}[||\nabla f_{\nu^\star,\nu_t}||^2]$$

- By Cauchy-Schwarz in the RKHS,[A, eq. 16]

$$\mathcal{F}(\nu_t) =: \frac{1}{2}MMD^2(\nu_t, \nu^\star) \leq S(\nu^\star|\nu_t)\mathbb{E}_{\nu_t}[||\nabla f_{\nu^\star,\nu_t}||^2]$$

where $S(\nu^\star|\nu_t)$ is the Negative Sobolev Distance[2]

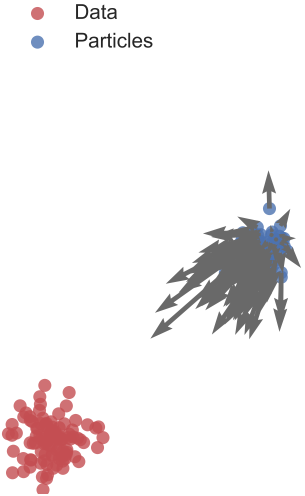
- Require $S(\nu^\star|\nu_t) < C$ for entire sequence ν_t: hard to check in theory, fails in practice.

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

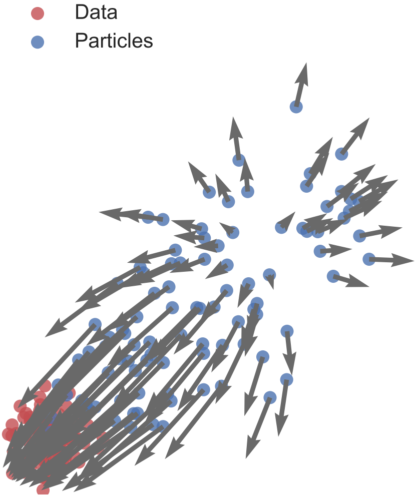
[2] $S(\nu^\star|\nu_t) = \sup_{g,\mathbb{E}_{Z\sim\nu_t}[||\nabla g(Z)||^2]\leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^\star}[g(U)]|$

MMD flow in practice

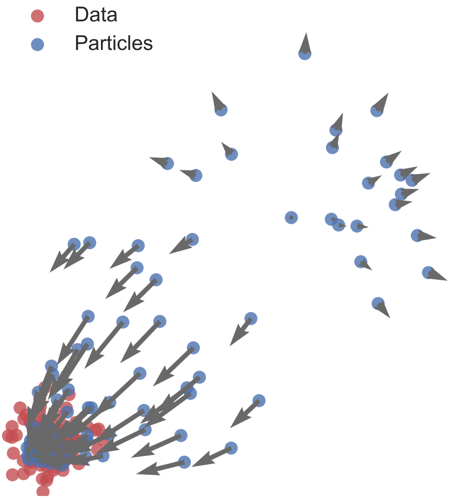
MMD flow in practice



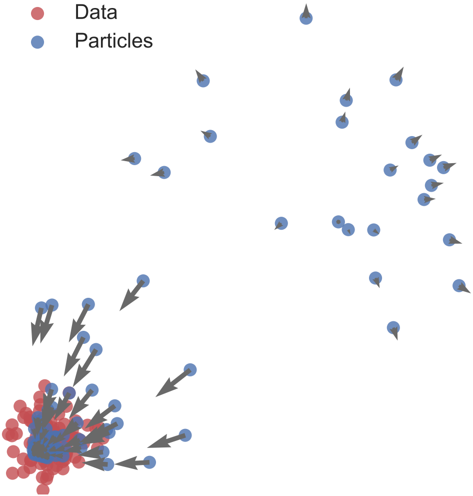
MMD flow in practice



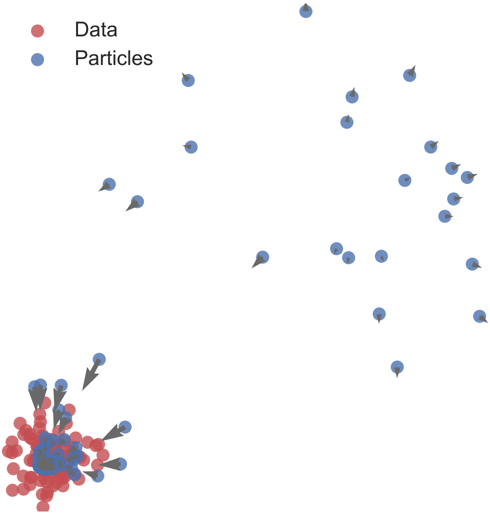
MMD flow in practice



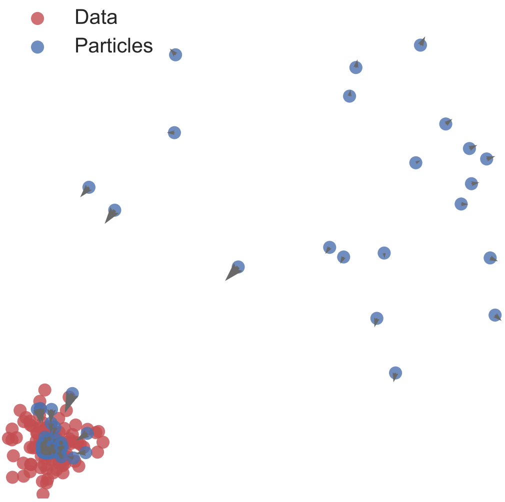
MMD flow in practice



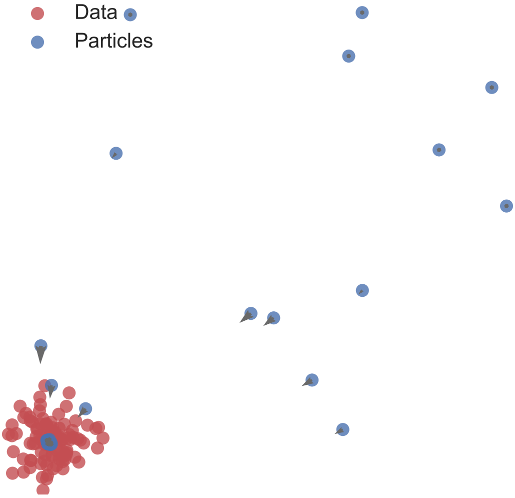
MMD flow in practice



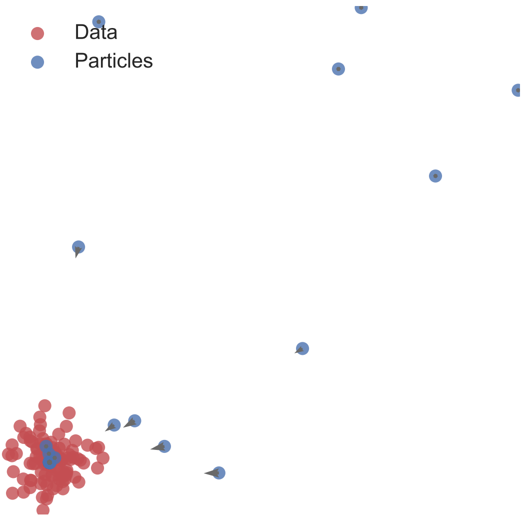
MMD flow in practice



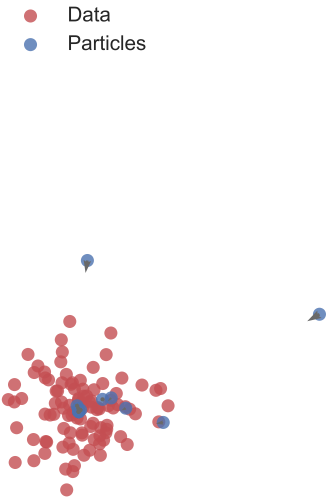
MMD flow in practice



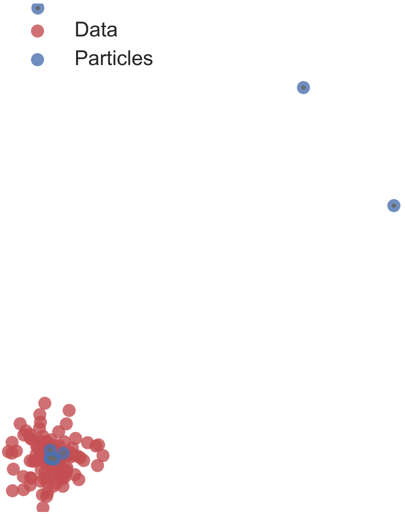
MMD flow in practice



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.

Can these observations be used to improve convergence?

Noise injection to improve 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,[3] but extended to interacting particles.
- Different from entropic regularization:

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

[3] Chaudhari, Oberman, Osher, Soatto, Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. Research in the Mathematical Sciences (2017)
Hazan, Levy, Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. ICML (2016).

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: Back to the failure mode: $\nabla f_{\nu^\star,\nu_t}(Z_t + \beta_t u_t) \simeq 0$

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

Noise injection: consistency

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

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: Back to the failure mode: $\nabla f_{\nu^\star, \nu_t}(Z_t + \beta_t u_t) \simeq 0$

Need β_t such that:

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

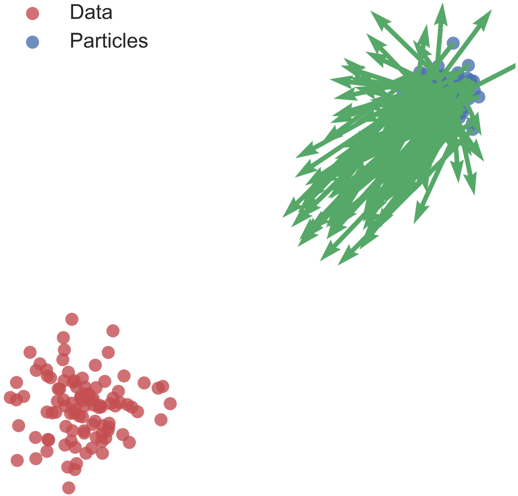
$$\sum_i^t \beta_i^2 \underset{t \to \infty}{\rightarrow} \infty$$

Then [A, Proposition 8]

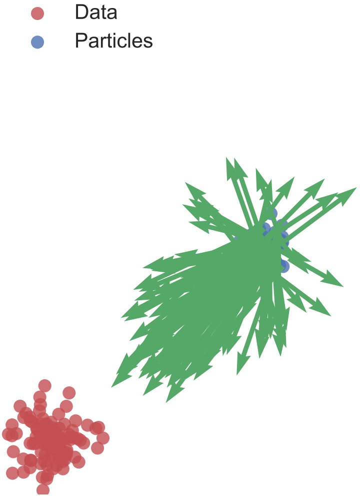
$$\mathcal{F}(\nu_t) \leq \mathcal{F}(\nu_0) e^{-C\gamma \sum_i^t \beta_i^2}.$$

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

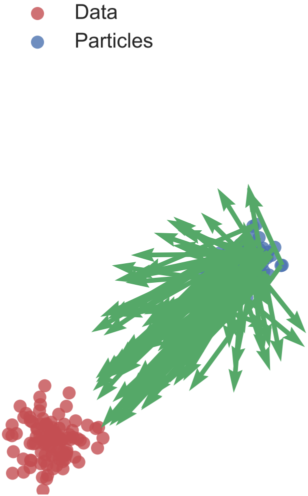
Noise injected MMD flow in practice



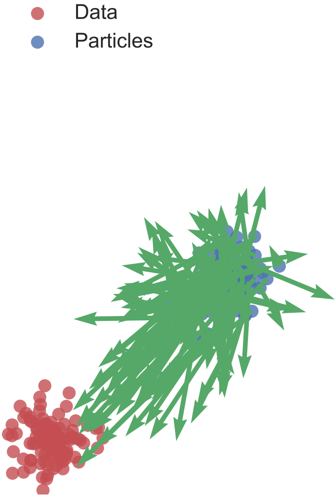
Noise injected MMD flow in practice



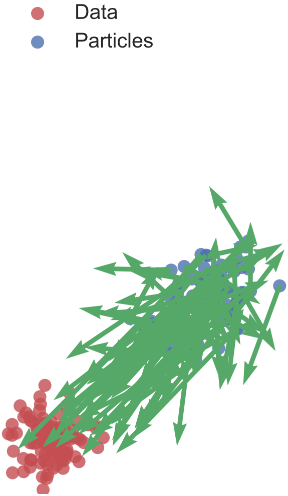
Noise injected MMD flow in practice



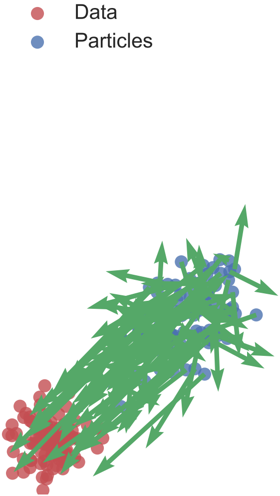
Noise injected MMD flow in practice



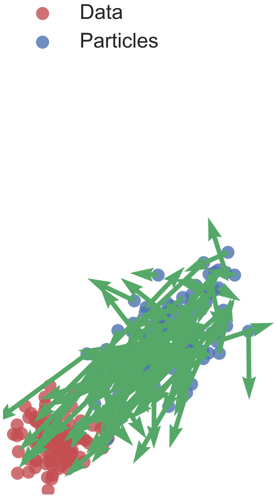
Noise injected MMD flow in practice



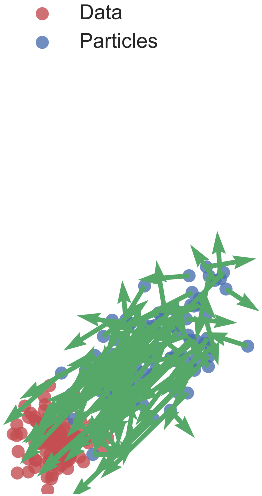
Noise injected MMD flow in practice



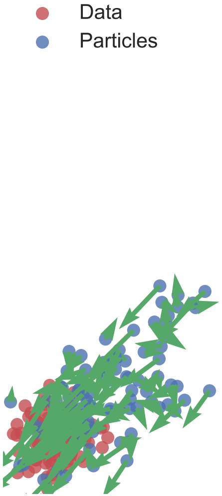
Noise injected MMD flow in practice



Noise injected MMD flow in practice

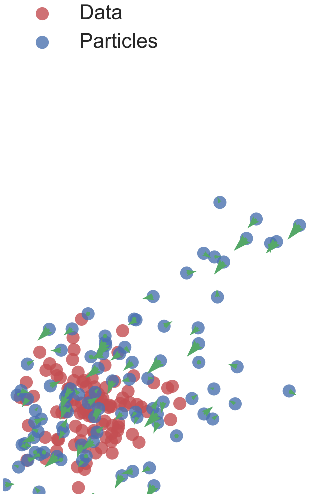


Noise injected MMD flow in practice



Noise injected MMD flow in practice

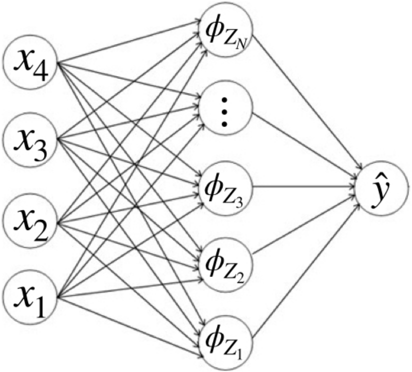
Noise injected MMD flow in practice



Noise injected MMD flow in practice

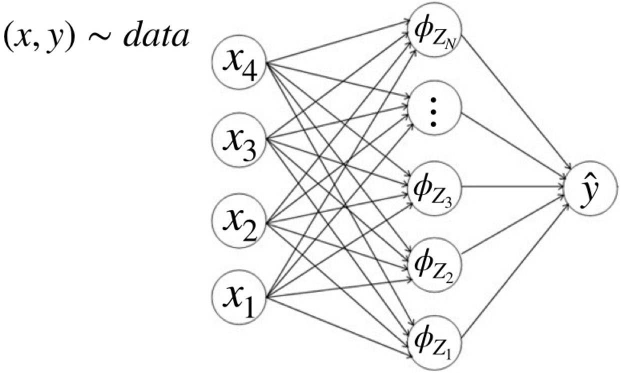
Noise injection: neural net setting

$(x, y) \sim data$



$$\min_{Z_1,\ldots,Z_N} \mathbb{E}_{data}[\| \frac{1}{M} \sum_{m}^{M} \phi_{U^m}(x) - \frac{1}{N} \sum_{n=1}^{N} \phi_{Z^n}(x) \|^2]$$

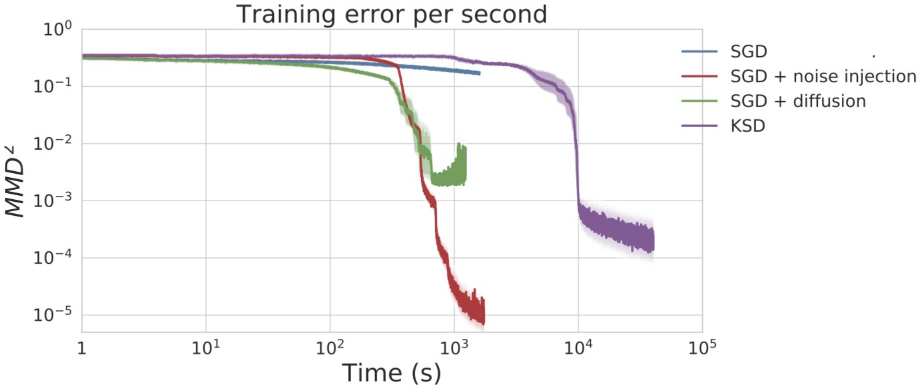
Noise injection: neural net setting



$$\min_{Z_1,\ldots,Z_N} MMD^2(\nu^*, \frac{1}{N}\sum_{n=1}^{N} \delta_{Z^n})$$

$$k(Z, Z') = \mathbb{E}_{data}[\phi_Z(x)\phi_{Z'}(x)]$$

Noise injection: neural net setting



The KALE, and KALE flow

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{p(z)}{q(z)}\right) q(z)dz$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- Example: $\phi(u) = u\log(u)$ gives KL divergence,

$$\begin{aligned} D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)dz \\ &= \int \left(\frac{p(z)}{q(z)}\right) \log\left(\frac{p(z)}{q(z)}\right) q(z)dz \end{aligned}$$

The ϕ-divergences

Define the ϕ-divergence(f-divergence):

$$D_\phi(P, Q) = \int \phi\left(\frac{p(z)}{q(z)}\right) q(z)dz$$

where ϕ is convex, lower-semicontinuous, $\phi(1) = 0$.

- Example: $\phi(u) = u\log(u)$ gives KL divergence,

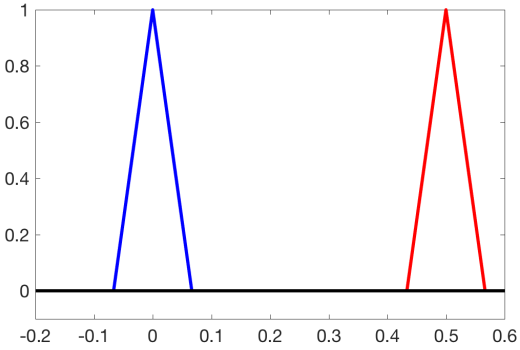
$$\begin{aligned} D_{KL}(P, Q) &= \int \log\left(\frac{p(z)}{q(z)}\right) p(z)dz \\ &= \int \left(\frac{p(z)}{q(z)}\right) \log\left(\frac{p(z)}{q(z)}\right) q(z)dz \end{aligned}$$

The challenge of disjoint support

Simple example: disjoint support.

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

$$D_{KL}(P, Q) = \infty \qquad D_{JS}(P, Q) = \log 2$$

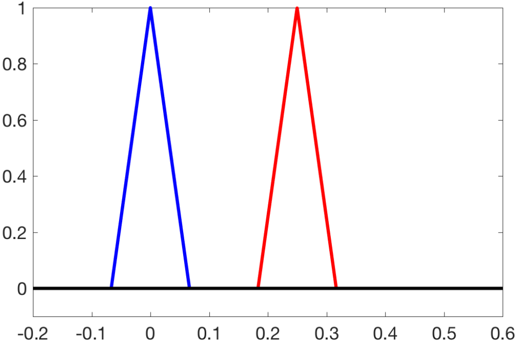


The challenge of disjoint support

Simple example: disjoint support.

Goodfellow et al. (NeurIPS 2014), Arjovsky and Bottou [ICLR 2017]

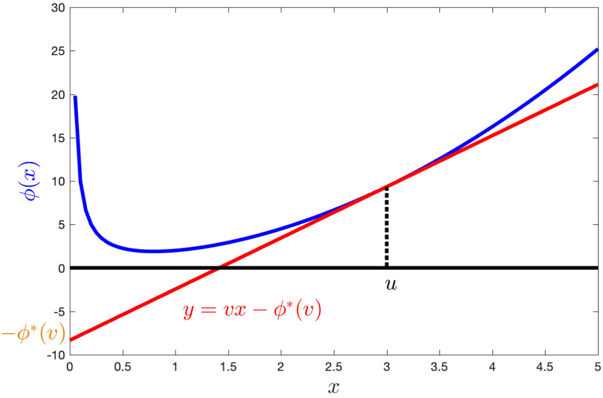
$$D_{KL}(P, Q) = \infty \qquad D_{JS}(P, Q) = \log 2$$



ϕ-divergences in practice

Notation: the conjugate (Fenchel) dual

$$\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.$$



- $\phi^*(v)$ is negative intercept of tangent to ϕ with slope v

ϕ-divergences in practice

Notation: the conjugate (Fenchel) dual

$$\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.$$

- For a convex l.s.c. ϕ we have

$$\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}$$

ϕ-divergences in practice

Notation: the conjugate (Fenchel) dual

$$\phi^*(v) = \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.$$

- For a convex l.s.c. ϕ we have

$$\phi^{**}(x) = \phi(x) = \sup_{v \in \mathbb{R}} \{xv - \phi^*(v)\}$$

- KL divergence:

$$\phi(x) = x \log(x) \qquad \phi^*(v) = \exp(v - 1)$$

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$D_\phi(P, Q) = \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz$$

$$= \int q(z) \underbrace{\sup_{f_z}\left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right)}_{\phi\left(\frac{p(z)}{q(z)}\right)}$$

$\phi^*(v)$ is dual of $\phi(x)$.

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$
\begin{aligned}
D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\
&= \int q(z) \sup_{f_z} \left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right) \\
&\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) - \mathrm{E}_Q \phi^*(f(Y))
\end{aligned}
$$

(restrict the function class)

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

A variational lower bound

A lower-bound ϕ-divergence approximation:

$$\begin{aligned} D_\phi(P, Q) &= \int q(z)\phi\left(\frac{p(z)}{q(z)}\right) dz \\ &= \int q(z) \sup_{f_z} \left(\frac{p(z)}{q(z)} f_z - \phi^*(f_z)\right) \\ &\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) - \mathrm{E}_Q \phi^*(f(Y)) \end{aligned}$$

(restrict the function class)

Bound tight when:

$$f^\diamond(z) = \partial\phi\left(\frac{p(z)}{q(z)}\right)$$

if ratio defined.

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \underbrace{\exp\left(f(Y)\right)}_{\phi^*(f(Y)+1)}$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

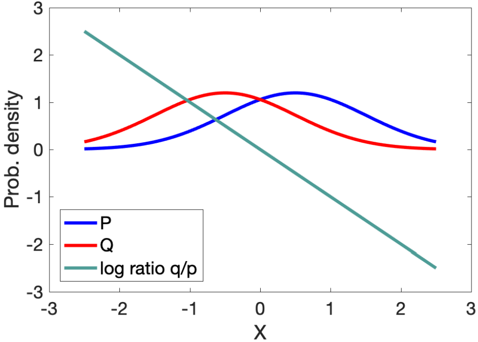
$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp(f(Y))$$

Bound tight when:

$$f^{\diamond}(z) = \log \frac{q(z)}{p(z)}$$

if ratio defined.



Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\, dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(f(Y)\right)$$

$$\approx \sup_{f \in \mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(f(y_i))\right] + 1$$

$$x_i \overset{\text{i.i.d.}}{\sim} P$$

$$y_i \overset{\text{i.i.d.}}{\sim} Q$$

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(f(Y)\right)$$

$$\approx \sup_{f \in \mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(f(y_i))\right] + 1$$

This is a

KL

Approximate

Lower-bound

Estimator.

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\,dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(f(Y)\right)$$

$$\approx \sup_{f \in \mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(f(y_i))\right] + 1$$

This is a

K

A

L

E

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010); Nowozin, Cseke, Tomioka, NeurIPS (2016)

Case of the KL

$$D_{KL}(P, Q) = \int \log\left(\frac{p(z)}{q(z)}\right) p(z)\, dz$$

$$\geq \sup_{f \in \mathcal{H}} \mathrm{E}_P f(X) + 1 - \mathrm{E}_Q \exp\left(f(Y)\right)$$

$$\approx \sup_{f \in \mathcal{H}} \left[\frac{1}{n}\sum_{j=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} \exp(f(y_i))\right] + 1$$

The KALE divergence

Nguyen, Wainwright, Jordan, IEEE Transactions on Information Theory (2010);
Nowozin, Cseke, Tomioka, NeurIPS (2016)

Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} E_P f(X) - E_Q \exp\left(f(Y)\right) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$$\|w\|^2_{\mathcal{H}} \quad \text{penalized}$$

Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

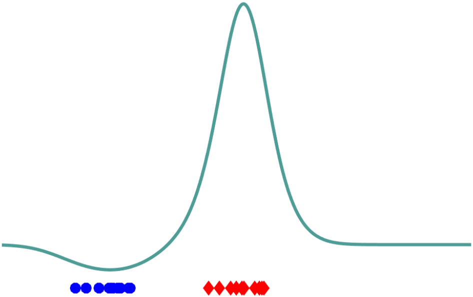
Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} E_P f(X) - E_Q \exp(f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$\|w\|_{\mathcal{H}}^2$ penalized

$KALE(Q, P; \mathcal{H})$=0.18



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

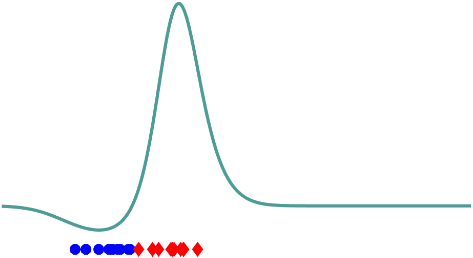
Empirical properties of KALE

$$KALE(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}} E_P f(X) - E_Q \exp(f(Y)) + 1$$

$$f = \langle w, \phi(x) \rangle_{\mathcal{H}} \qquad \mathcal{H} \text{ an RKHS}$$

$\|w\|^2_{\mathcal{H}}$ penalized

$$KALE(Q, P; \mathcal{H}) = 0.12$$



Glaser, Arbel, G. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," (NeurIPS 2021, Section 2)

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\max}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

Zhang, Liu, Zhou, Xu, and He. "On the Discrimination-Generalization Tradeoff in GANs" (ICLR 2018, Corollary 2.4; Theorem B.1)
Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (1)

Key requirements on $\mathcal{H}$ and $\mathcal{X}$:

- Compact domain $\mathcal{X}$,
- $\mathcal{H}$ dense in the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ wrt $\|\cdot\|_\infty$.
- If $f \in \mathcal{H}$ then $-f \in \mathcal{H}$ and $cf \in \mathcal{H}$ for $0 \leq c \leq C_{\max}$.

Theorem: $KALE(P, Q; \mathcal{H}) \geq 0$ and $KALE(P, Q; \mathcal{H}) = 0$ iff $P = Q$.

$\mathcal{H}$ dense in $C(\mathcal{X})$ for $\mathcal{X} \subset \mathbb{R}^d$ when:

$$\mathcal{H} = \operatorname{span}\{\sigma(w\top x + b) : [w, b] \in \Theta\}$$

$\sigma(u) = \max\{u, 0\}^\alpha$, $\alpha \in \mathbb{N}$, and $\{\lambda\theta : \lambda \geq 0, \theta \in \Theta\} = \mathbb{R}^{d+1}$.

Zhang, Liu, Zhou, Xu, and He. “On the Discrimination-Generalization Tradeoff in GANs”

(ICLR 2018, Corollary 2.4; Theorem B.1)
Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $\text{KALE}(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)

Topological properties of KALE (2)

Additional requirement: all functions in $\mathcal{H}$ Lipschitz in their inputs with constant L

Theorem: $\mathrm{KALE}(P, Q^n; \mathcal{H}) \to 0$ iff $Q^n \to P$ under the weak topology.

Partial proof idea:

$$\begin{aligned}
\mathrm{KALE}(P, Q; \mathcal{H}) &= \int f \, dP - \int \exp(f) \, dQ + 1 \\
&= -\int f(x) \, dQ(x) + f(x') \, dP(x') \\
&\quad - \int \underbrace{(\exp(f) - f - 1)}_{\geq 0} \, dQ \\
&\leq \int f(x') \, dP(x') - \int f(x) \, dQ(x) \leq L W_1(P, Q)
\end{aligned}$$

Liu, Bousquet, Chaudhuri. "Approximation and Convergence Properties of Generative Adversarial Learning" (NeurIPS 2017); Arbel, Liang, G. (ICLR 2021, Proposition 1)

KALE vs KL vs MMD

A scaled KALE (non-degenerate for $\lambda = 0$ or $\lambda \to \infty$):

$$\mathrm{KALE}_\lambda(P, Q; \mathcal{H}) = (1+\lambda) \sup_{f \in \mathcal{H}} \left[E_P f(X) - E_Q \exp(f(Y)) + 1 - \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \right]$$

MMD limit:

$$\lim_{\lambda \to +\infty} \mathrm{KALE}_\lambda(P, Q; \mathcal{H}) = \frac{1}{2} \mathrm{MMD}^2(P, Q).$$

KL limit (assuming $\log \frac{\mathrm{d}P}{\mathrm{d}Q} \in \mathcal{H}$):

$$\lim_{\lambda \to 0} \mathrm{KALE}_\lambda(P, Q; \mathcal{H}) = \mathrm{KL}(P, Q).$$

Glaser, Arbel, G. (NeurIPS 2021, Proposition 1)

Wasserstein gradient flow on KALE

First variation of the $KALE_\lambda(\nu, \nu^\star)$

$$\frac{\partial \mathrm{KALE}_\lambda}{\partial \nu}(\nu)(z) := (1+\lambda)\, f_{\nu,\nu^\star}(z)$$

where $f_{\nu,\nu^\star}$ is the solution of

$$f_{\nu,\nu^\star} = \arg\max_{f \in \mathcal{H}} \{\mathcal{K}(f, \nu)\},$$

where

$$\mathcal{K}(f, \nu) := E_\nu f(X) - E_{\nu^*} \exp(f(Y)) + 1 - \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$$

Wasserstein gradient flow on KALE

First variation of the $KALE_\lambda(\nu, \nu^\star)$

$$\frac{\partial \mathrm{KALE}_\lambda}{\partial \nu}(\nu)(z) := (1+\lambda)\, f_{\nu,\nu^\star}(z)$$

where $f_{\nu,\nu^\star}$ is the solution of

$$f_{\nu,\nu^\star} = \arg\max_{f \in \mathcal{H}} \{\mathcal{K}(f, \nu)\},$$

where

$$\mathcal{K}(f,\nu) := E_\nu f(X) - E_{\nu^*} \exp(f(Y)) + 1 - \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$$

Proof (idea):

$$\frac{\partial \mathrm{KALE}_\lambda}{\partial \nu} = \frac{\partial \mathcal{K}(f_{\nu,\nu^\star}, \nu)}{\partial \nu} + \underbrace{\frac{\partial \mathcal{K}(f,\nu)}{\partial f}\bigg|_{f=f_{\nu,\nu^\star}}}_{=0} \frac{\partial f_{\nu,\nu^\star}}{\partial \nu}$$

as long as $\frac{\partial f_{\nu,\nu^\star}}{\partial \nu}$ exists (via implicit function theorem)

Wasserstein gradient flow on KALE

The W_2 gradient flow of the KALE:

$$\partial_t \nu_t = -(1+\lambda)\mathrm{div}(\nu_t \nabla f_{\nu_t,\nu^\star}), \qquad \nu_0 = P_0$$

where

$$f_{\nu,\nu^\star} = \arg\max_f \mathcal{K}(f,\nu)$$

Glaser, Arbel, G. (NeurIPS 2021, Lemma 3)

Consistency (2)

Again, under the (strong!) assumption

$$S(\nu^\star|\nu_t) := \sup_{g, \mathbb{E}_{Z\sim\nu_t}[||\nabla g(Z)||^2]\leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^\star}[g(U)]|$$
$$\leq C$$

we have

$$\text{KALE}(\nu_t) \leq \frac{1}{\text{KALE}(\nu_0)^{-1} + C^{-1}t}$$

Once again, noise injection can be used (similar result to MMD flow).

Glaser, Arbel, G. (NeurIPS 2021, Proposition 3)

Consistency (2)

Again, under the (strong!) assumption

$$S(\nu^\star|\nu_t) := \sup_{g, \mathbb{E}_{Z\sim\nu_t}[\|\nabla g(Z)\|^2]\leq 1} |\mathbb{E}_{Z\sim\nu_t}[g(Z)] - \mathbb{E}_{U\sim\nu^\star}[g(U)]|$$
$$\leq C$$

we have

$$\mathrm{KALE}(\nu_t) \leq \frac{1}{\mathrm{KALE}(\nu_0)^{-1} + C^{-1}t}$$

Once again, noise injection can be used (similar result to MMD flow).

Compare with linear rate for Wasserstein-2 flow on KL when $\nu^\star$ satisfies log-Sobolev inequality with constant ρ:

$$\frac{d}{dt}KL(\nu_t, \nu^\star) \leq -2\rho KL(\nu_t, \nu^\star)$$

Glaser, Arbel, G. (NeurIPS 2021, Proposition 3)

KALE flow vs MMD flow in practice

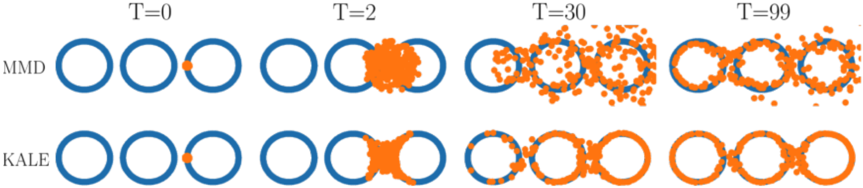


Figure 1: MMD and KALE flow trajectories for “three rings” target

Glaser, Arbel, G. (NeurIPS 2021)

Summary

- Gradient flows based on kernel dependence measures:
  - MMD flow is simpler, KALE flow is mode-seeking
  - Noise injection can improve convergence
- NeurIPS 2019, NeurIPS 2021

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

Summary

- Gradient flows based on kernel dependence measures:
  - MMD flow is simpler, KALE flow is mode-seeking
  - Noise injection can improve convergence
- NeurIPS 2019, NeurIPS 2021

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

KALE as GAN critic:

ICLR 2021:

arXiv.org > stat > arXiv:2003.05033

Statistics > Machine Learning

[Submitted on 10 Mar 2020 (v1), last revised 24 Jun 2020 (this version, v3)]

Generalized Energy Based Models

Michael Arbel, Liang Zhou, Arthur Gretton

NeurIPS 2020:

arXiv.org > cs > arXiv:2003.06060

Search...

Help | Advanc

Computer Science > Machine Learning

[Submitted on 12 Mar 2020 (v1), last revised 24 Mar 2020 (this version, v2)]

Your GAN is Secretly an Energy-based Model and You Should use Discriminator Driven Latent Sampling

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, Yoshua Bengio

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT