Statistical and Geometrical properties of the regularized kernel Kullback Leibler divergence

Clémentine Chazal[1] Anna Korba[1] Francis Bach[2]

CREST/ENSAE, IP Paris[1] , INRIA, Paris[2]

Abstract

In this paper [2] we study the properties of the kernel Kullback-Leibler divergence (KKL), introduced in [1], with the aim of performing sampling by using the divergence as the objective of an optimisation problem. Our contributions are to propose a regularized version of the KKL, which is consistent for empirical measures and to derive a Wasserstein gradient of the KKL which has enabled to implement a sampling algorithm.

Introduction and motivations

Problem: To approximate a target distribution q on $\mathbb{R}^d$, we solve the optimization problem

$$\boxed{\min_{p \in \mathcal{P}(\mathbb{R}^d)} \mathcal{F}(p)}$$

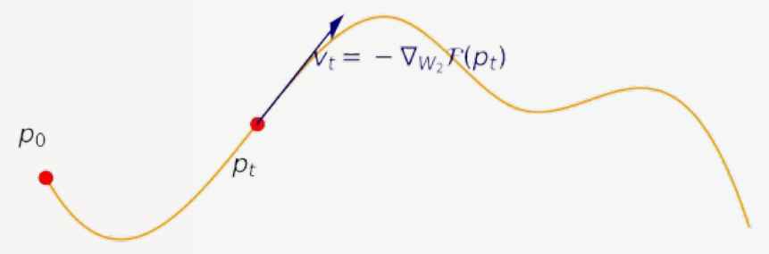
where $\mathcal{F}(p) = D(p||q)$ with D a divergence or a distance.

Wasserstein gradient flow:

- If for any function $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$, the expansion
$$\mathcal{F}((I_d + \varepsilon h)_{\#p}) = \mathcal{F}(p) + \varepsilon \langle \nabla_{W_2} \mathcal{F}(p), h \rangle_p + o(\varepsilon),$$
holds, then $\nabla_{W_2}\mathcal{F}(p) : \mathbb{R}^d \to \mathbb{R}^d$ is the Wasserstein gradient of $\mathcal{F}$.
- Analogy between gradient flow and Wasserstein gradient flow

$$\begin{cases} \text{Gradient Flow} \\ x(0) = x_0, \\ x'(t) = -\nabla f(x(t)). \end{cases}$$

$$\begin{cases} \text{Wasserstein Gradient Flow} \\ p(0) = p_0, \\ \partial_t p(t) = -\nabla_{W_2}\mathcal{F}(p(t)). \end{cases}$$



The choice of D dictates the overall dynamics. In this project we selected the regularized Kernel Kullback Leibler Divergence.

Kernel Kullback Leibler divergence (KKL)

Kernel Kullback Leibler divergence (KKL): Given $\mathcal{H}$ a RKHS with reproducing kernel k. For $p \ll q$, the KKL divergence is

$$\mathrm{KKL}(p||q) := \mathrm{Tr}[\Sigma_p(\log \Sigma_p - \log \Sigma_q)]$$

where

$$\Sigma_p = \int k(.,x)k(.,x)^* dp(x).$$

If k^2 and $\forall x \in \mathbb{R}^d$, $k(x,x) = 1$ then

$$\mathrm{KKL}(p||q) = 0 \Leftrightarrow p = q.$$

Regularized KKL **:** To handle cases where $p \not\ll q$, the regularized KKL is defined for $\alpha \in]0,1[$ as

$$\boxed{\mathrm{KKL}_\alpha(p \parallel q) := \mathrm{KKL}(p \parallel (1-\alpha)q + \alpha p)}$$

Closed form for regularized KKL on empirical distributions

Regularized KKL **for empirical distributions:** Let $x_1, \dots, x_n \sim p$, $y_1, \dots, y_m$ and note $\hat{p} = \frac{1}{n}\Sigma_{i=1}^n \delta_{x_i}$ and $\hat{q} = \frac{1}{m}\Sigma_{j=1}^m \delta_{y_j}$. Regularized KKL admits a closed form expression

$$\mathrm{KKL}_\alpha(\hat{p}||\hat{q}) = \mathrm{Tr}\left(\frac{1}{n}K_{\hat{p}} \log \frac{1}{n}K_{\hat{p}}\right) - \mathrm{Tr}\left(I_\alpha K \log(K)\right),$$

$$I_\alpha = \begin{pmatrix} \frac{1}{\alpha}I & 0 \\ 0 & 0 \end{pmatrix} \text{ and } K = \begin{pmatrix} \frac{\alpha}{n}K_{\hat{p}} & \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{\hat{p},\hat{q}} \\ \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{\hat{q},\hat{p}} & \frac{1-\alpha}{m}K_{\hat{q}} \end{pmatrix}$$

and $K_{\hat{p}} = (k(x_i, x_j))_{i,j=1}^n$, $K_{\hat{q}} = (k(y_i, y_j))_{i,j=1}^m$, $K_{\hat{p},\hat{q}} = (k(x_i, y_j))_{i,j=1}^{n,m}$.

Wasserstein gradient for empirical measures:

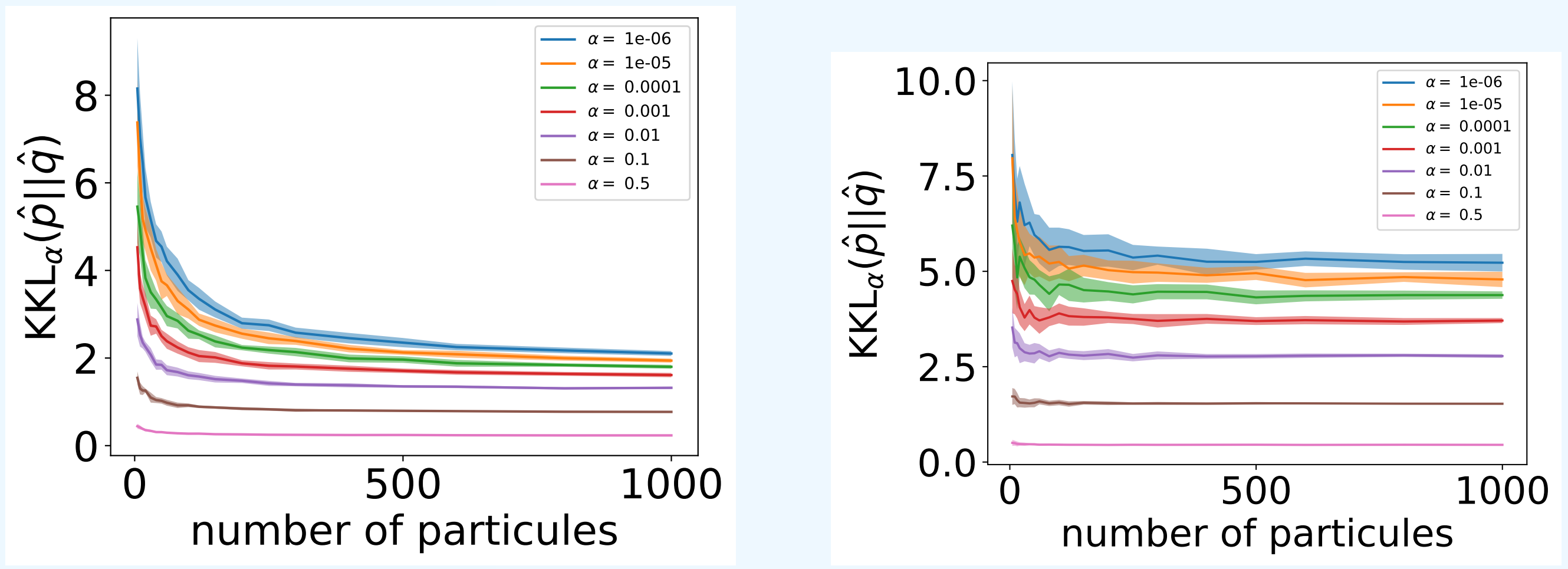
$$\nabla_{W_2}\mathcal{F}(\hat{p})(x) = \nabla_x \left(S(x)^T g(K_{\hat{p}})S(x) - T(x)^T g(K)T(x) - T(x)^T AT(x)\right)$$

where $S(x) = (\frac{1}{\sqrt{n}}k(x, x_i))_i$, $T(x) = ((\sqrt{\frac{\alpha}{n}}k(x, x_i))_i, (\sqrt{\frac{1-\alpha}{m}}k(x, y_j))_j)$ and A is a matrix depending on the eigenvalues and eigenvectors of K.

Theorical properties of the regularized KKL

- The regularized KKL is consistant to the true KKL for $p \ll q$ when $\alpha \to 0$:
$$\mathrm{KKL}_\alpha(p||q) \underset{\alpha \to 0}{\to} \mathrm{KKL}(p||q).$$
- $\alpha \to \mathrm{KKL}_\alpha(p||q)$ is decreasing.
- Consistency of the regularized KKL for empirical measures:
$$\mathbb{E}|\mathrm{KKL}_\alpha(\hat{p}||\hat{q}) - \mathrm{KKL}_\alpha(p||q)| \leqslant C_{p,\alpha}\frac{\log n}{\sqrt{m \wedge n}} + C'_{p,\alpha}\frac{\log^2 n}{m \wedge n}.$$

The following experiments illustrate the previous theorical results.



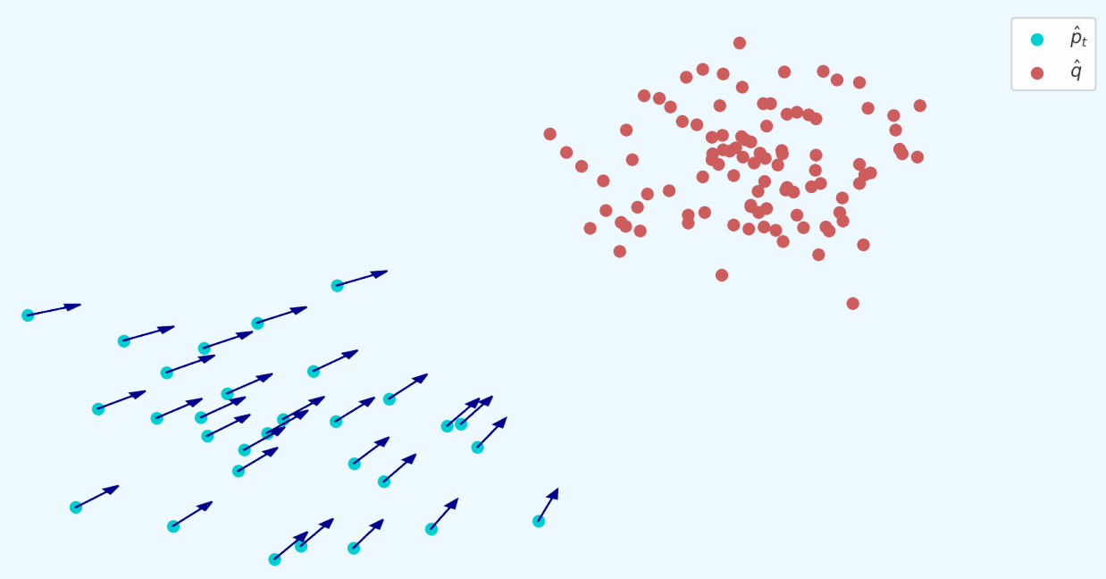
$d = 10$ $d = 2$

Sampling experiments

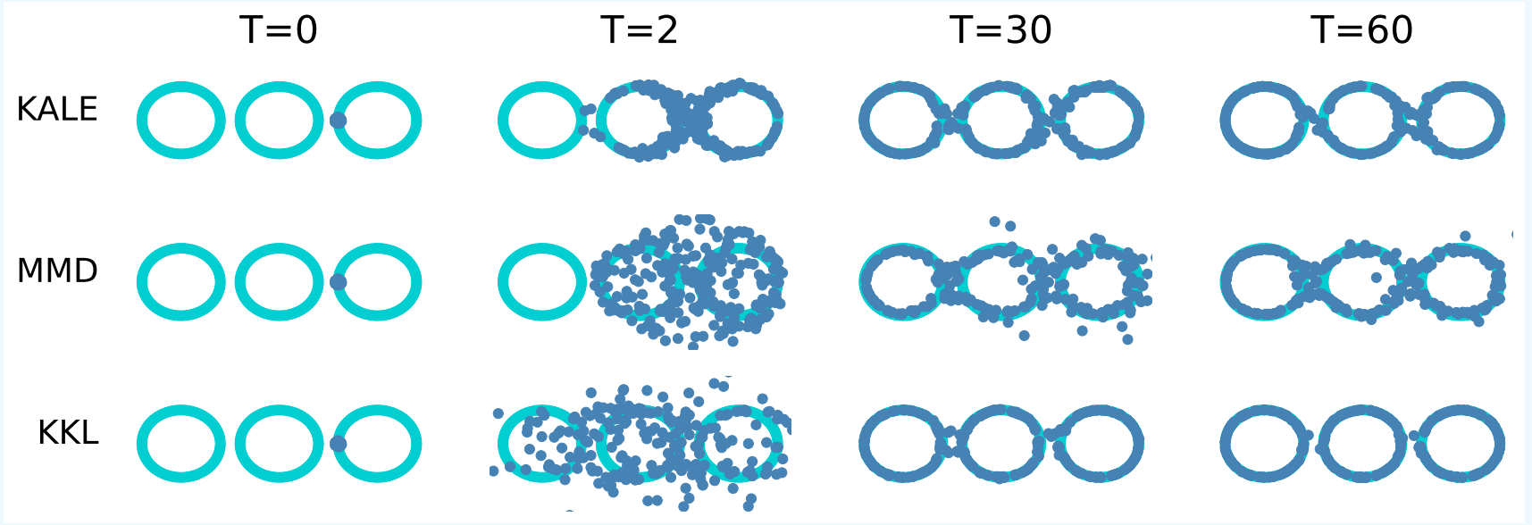
Now we fix $\hat{q}$, we optimize $\hat{p}$ by a discretisation of the Wasserstein gradient flow of the regularized KKL.

Descent scheme: Let $\hat{p}_t = \frac{1}{n}\Sigma_{i=1}^n \delta_{x_i^t}$, $\gamma > 0$, $t = 1, \dots, T$.

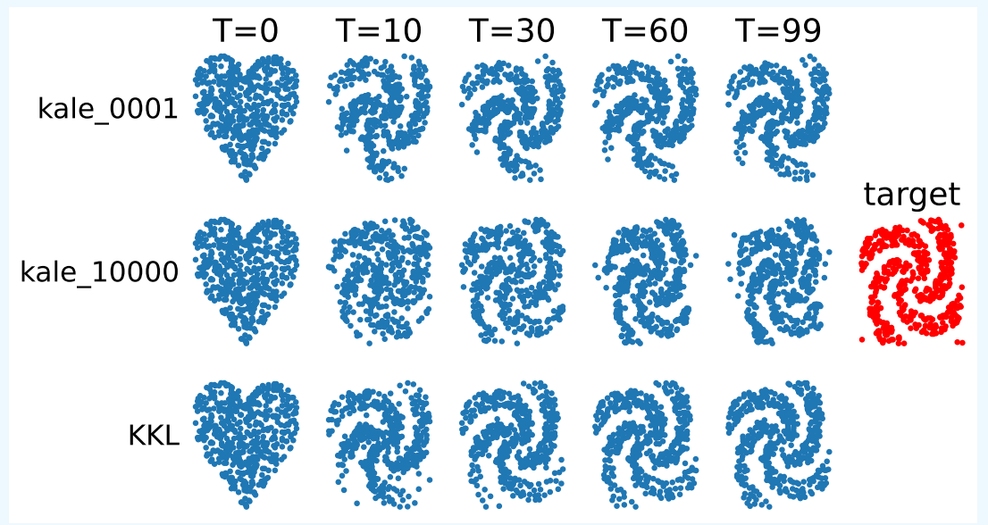
- $x_{t+1}^i = x_t^i - \gamma \nabla_{W_2}\mathcal{F}(\hat{p}_t)(x_t^i)$
- $\hat{p}_{t+1} = (I_d - \gamma \nabla_{W_2}\mathcal{F}(\hat{p}_t))_{\#\hat{p}}$



Experiments:



MMD, KALE and KKL flow for 3 rings target.



Shape transfer

Reference

[1] Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2022.

[2] Clémentine Chazal, Anna Korba, and Francis Bach. Statistical and geometrical properties of regularized kernel kullback-leibler divergence. *NeurIPS*, 2024.