Statistical and Geometrical properties of regularized Kernel Kullback-Leibler divergence

Clémentine Chazal
joint work with Anna Korba (ENSAE) and Francis Bach (INRIA)

CREST, ENSAE

Abstract. In this work we study the properties of the kernel Kullback Leibler divergence (KKL) introduced in [1], in order to do sampling with Wasserstein Gradient flows. Our approach involves optimization in the space of probability measures, using the KKL divergence as the objective function. This divergence measures a kind of distance between two probability distributions: the target distribution, which is only partially known, and a set of particles intended to represent as well as possible the target distribution. By performing an adapted gradient descent on the particle distribution, we can transport the particles from an initial position to a good representation of the target. Our contributions are to propose a regularized version of the KKL divergence which is defined for any distribution, to derive a statistical upper bound for the convergence of the KKL on empirical distribution to the KKL on continuous distribution and to propose a closed form expression for the KKL and its Wasserstein gradient which enables to implement a sampling algorithm implying gradient descent.

Keywords: Sampling · Kernel methods · Optimal transport · Wasserstein gradient flows.

1 Introduction

Sampling in Machine Learning consists in approaching as close as possible a target probability distribution q for which we only know partial information. For example, in Bayesian inference one can derive the posterior distribution of the parameters of some neural network. In this case the density of q is know up to a normalisation constant. There are various ways of solving this problem, including MCMC methods [2] or variational inference methods [3]. In generative modelling, only a set of observation from q is available and the goal is to generate data whose distribution is similar to the training set distribution. To solve the problem in this case, it is usual to minimise a divergence or distance D between a discrete probability distribution $\hat{p}$ that we construct iteratively and $\hat{q}$ the empirical distribution from the observations of q that we possess. This is the approach we study in this article. It can be formulated as an optimization problem

$$\min_{p \in \mathcal{P}(\mathcal{X})} D(p||q)$$

where $\mathcal{P}(\mathcal{X})$ is the space of probability distributions on $\mathcal{X}$. The choice for the divergence D is quite important, indeed depending on its geometry, the minimization can be more or less efficient. For example, in [4] D is chosen to be Maximum Mean Discrepancy (MMD) divergence which has its weaknesses when used on its own because the MMD has many local minima. We chose the Kernel Kullback Leibler divergence (KKL) introduced in [1] which is defined in the following section. Then, Wasserstein's gradient descent principle is close to the classical gradient descent. For a fixed target q, let's note $\mathcal{F}(p) = D(p||q)$. Let $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$ and consider the push forward distribution $(I_d + \varepsilon h)_{\#p}$. If we can develop

$$\mathcal{F}((I_d + \varepsilon h)_{\#p}) = \mathcal{F}(p) + \varepsilon \langle \nabla_{W2}\mathcal{F}(p), h\rangle_p + o(\varepsilon),$$

then $\nabla_{W2}\mathcal{F}(p) : \mathbb{R}^d \to \mathbb{R}^d$ is called the Wasserstein gradient of $\mathcal{F}$. Then if $\hat{p}$ is supported on a finite number of points we can build the following gradient descent algorithm. Let $x_t^1, ...x_t^n = \hat{p}^t$.

$$x_{t+1}^i = x_t^i - \gamma \nabla_{W2}\mathcal{F}(\hat{p}_t)(x_t^i). \tag{1}$$

1.1 Definition of the regularized Kernel Kullback Leibler divergence

The Kernel Kullback Leibler (KKL) divergence is defined as follows. Let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel, $\mathcal{H}$ its Hilbert space with reproducing kernel (RKHS) and φ its transformation function. For a probability distribution $p \in \mathcal{P}(\mathcal{X})$, the covariance operator $\Sigma_p : \mathcal{H} \to \mathcal{H}$ is given by :

$$\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$$

where $*$ represents the transposition in $\mathcal{H}$. For $p, q \in \mathcal{P}(\mathcal{X})$, the kernel Kullback-Leibler divergence (KKL) is defined in [1] as:

$$\mathrm{KKL}(p||q) := Tr(\Sigma_p \log \Sigma_p) - Tr(\Sigma_p \log \Sigma_q) = \sum_{(\lambda,\gamma)\in \Lambda_p \times \Lambda_q} \lambda \log\left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma\rangle^2_{\mathcal{H}}. \tag{2}$$

According to [1], if $\mathcal{X}$ is compact, if k is a positive definite continuous kernel with $k(x, x) = 1, \quad \forall x \in \mathcal{X}$ and if k^2 is universal, then $D(\Sigma_p||\Sigma_q) = 0 \Leftrightarrow p = q$. This makes the Kullback-Leibler (KKL) kernel divergence an interesting candidate for optimisation on $\mathcal{P}(\mathcal{X})$. In practice , we choose to use for k the gaussian kernel which satisfies these conditions.

A major issue with KKL is that it is finite only if the support of p is included in the support of q. In order to have KKL defined for any discrete distributions $\hat{p}$ and $\hat{q}$ we defined a regularized version of KKL which is for $\alpha \in]0, 1[$,

$$\mathrm{KKL}_\alpha(p||q) := \mathrm{KKL}(p||(1-\alpha)q+\alpha p) = Tr(\Sigma_p \log \Sigma_p) - Tr(\Sigma_p \log((1-\alpha)\Sigma_q + \alpha\Sigma_p)). \tag{3}$$

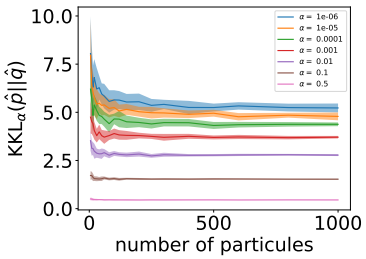
2 Contributions

2.1 Statistical bound

In this work, we use the regularized KKL on empirical distributions. A main question is then to know if it is a good approximation of the KKL on continuous distributions and if it converges to its value in population when the number of particles goes to infinity. We derived an upper bound on the difference between $\mathrm{KKL}_\alpha(\hat{p}||\hat{q})$ and $\mathrm{KKL}_\alpha(p||q)$ for p and q supported respectively on n and m points. Under some hypothesis, the bound is

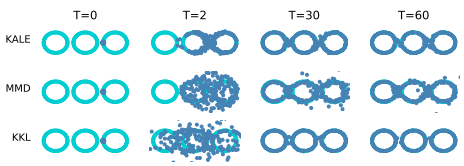
$$|\mathrm{KKL}_\alpha(\hat{p}||\hat{q}) - \mathrm{KKL}_\alpha(p||q)| \leqslant C_1 \frac{\log n}{\sqrt{n \wedge m}} + C_2 \frac{(\log n)^2}{m \wedge n}$$

where C_1 and C_2 depend on α and on the hypothesis. We plot the evolution of $\mathrm{KKL}_\alpha(\hat{p}||\hat{q})$ for an increasing number of particle n for $\hat{p}$ and $\hat{q}$ for different values of the parameter α. We observe that in each case it seems to converge.



2.2 Sampling experiment

We derived a closed form solution for the Wasserstein gradient of the KKL divergence which enabled us to implement the gradient descent algorithm as in 1. We decided to repeat an experiment carried out in [5] and compare our results with those obtained. The experiment consists in taking as target distribution q a uniform distribution over three rings, and initializing the $\hat{p}$ distribution by a set of points concentrated at one point on one of the rings (dark blue points). The efficiency of our method is clear to see.



References

1. Bach, F. (2022). Information theory with kernel methods. IEEE Transactions on Information Theory, 69(2), 752-775.
2. Roberts, G. O., Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms.
3. Blei, D. M., Kucukelbir, A., McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518), 859-877.
4. Arbel, M., Korba, A., Salim, A., Gretton, A. (2019). Maximum mean discrepancy gradient flow. Advances in Neural Information Processing Systems, 32.
5. Glaser, P., Arbel, M., Gretton, A. (2021). KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. Advances in Neural Information Processing Systems, 34, 8018-8031.