A Computable Measure of Suboptimality for Entropy-Regularised Variational Objectives

Clémentine Chazal[1], Heishiro Kanagawa[2], Zheyang Shen[2], Anna Korba[1], Chris. J. Oates[2,3],

[1]CREST, ENSAE, IP Paris, France

[2]Newcastle University, UK [3]The Alan Turing Institute, UK

Contributions

- Introduction of a new discrepancy, Kernel Gradient Discrepancy (KGD) that captures the closeness to the minimiser of entropy regularized variational objective. This discrepancy can be explicitly computed.
- Study of the convergence properties of KGD.
- Construction of new algorithms.

Introduction and motivations

Consider $P \in \mathcal{P}(\mathbb{R}^d)$ as the minimizer of

$$\mathcal{J}(Q) \coloneqq \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0)$$

where Q_0 is a reference distribution and $\mathcal{L}$ a loss on $\mathcal{P}(\mathbb{R}^d)$.

Problem: P is not tractable:

- We do not have access to unormalized density of P.
- $\mathcal{J}$ cannot be computed on discrete distributions as $\hat{Q} = \sum_{i=1}^n \delta_{x_i}$.

Intuition : If $J : \mathbb{R}^d \to \mathbb{R}$ with assumptions on J, minimisers of $J =$ minimisers of $\|\nabla J\|$.

$\rightarrow$ Instead of minimising $\mathcal{J}$, let's minimise $\|\nabla_{\mathrm{V}} \mathcal{J}(Q)\|$.

Definition: If for any function $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$, the expansion

$$\mathcal{J}((I_d + \varepsilon h)_{\#Q}) = \mathcal{J}(Q) + \varepsilon \langle \nabla_V \mathcal{J}(Q), h \rangle_{L_2} + o(\varepsilon),$$

holds, then $\nabla_V \mathcal{J}(Q) : \mathbb{R}^d \to \mathbb{R}^d$ is the Wasserstein gradient of $\mathcal{J}$.

Kernel Gradient Discrepancy (KGD)

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be a matrix-valued kernel. Let $\mathcal{B}_K = \{v \in \mathcal{H}_K : \|v\|_{\mathcal{H}_K} \leq 1\}$. The Kernel Gradient Discrepancy (KGD) is defined as

$$\mathrm{KGD}_K(Q) \coloneqq \sup_{v \in \mathcal{B}_K} \left| \int \nabla_V \mathcal{J}(Q) \cdot v(x) \, \mathrm{d}Q(x) \right|$$

KGD admits a closed form:

$$\mathrm{KGD}_K(Q) = \left(\iint k_K^Q(x, x') \, \mathrm{d}Q(x) \mathrm{d}Q(x') \right)^{1/2}$$

where

$$k_K^Q(x, x') \coloneqq \sum_{i=1}^d \sum_{j=1}^d \frac{1}{\rho_Q(x)\rho_Q(x')} \partial_{x'_j} \partial_{x_i} \left(\rho_Q(x) K_{i,j}(x, x') \rho_Q(x') \right)$$

is a reproducing kernel and $\rho_Q(x) \coloneqq q_0(x) \exp(-\mathcal{L}'(Q)(x))$.

Theorical Garanties of KGD

We say that $Q_n \xrightarrow{\alpha} Q$, if $\int h \, \mathrm{d}Q_n \to \int h \, \mathrm{d}Q$ for every continuous $h : \mathbb{R}^d \to [0, \infty)$ s.t. $h(x) \lesssim 1 + \|x\|^\alpha$. Under some conditions on k and $\mathcal{L}$,

- **Characterisation of stationnarity:** $\mathrm{KGD}_K(Q) = 0$ if and only if Q is a stationary point of $\mathcal{J}$
- **Continuity:** $\mathrm{KGD}_K(Q_n) \to \mathrm{KGD}_K(Q)$ whenever $Q_n \xrightarrow{\alpha} Q$.
- **Convergence control:** $\mathrm{KGD}_K(Q_n) \to 0$ implies $Q_n \xrightarrow{\alpha} P \in \mathcal{P}_\alpha(\mathbb{R}^d)$.

Experiments: Mean Field Neural Network

Goal: Learning an optimal distribution of the parameters of a neural network Φ with parameter X. Given $(z_1, y_1), ..., (z_N, y_N)$ linked by $y_i = f(z_i) + \xi_i$, $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

$$\mathcal{L}(Q) = \frac{\lambda}{N} \sum_{i=1}^N \ell(y_i, \mathbb{E}_{X\sim Q}[\Phi(z_i, X)]).$$

MFLD (Mean Field Langevin Dynamics algorithm):

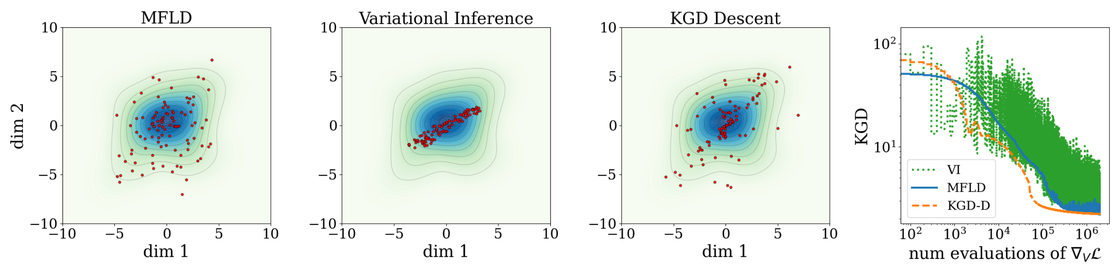
$$X_i^{t+1} = X_i^t + \epsilon[(\nabla \log q_0) - \nabla_{\mathrm{V}} \mathcal{L}(Q_n^t)](X_i^t) + \sqrt{2\epsilon} Z_t^i, \quad Z_t^i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1).$$

Variational Inference: Consider $Q_\theta = T_\#^\theta \mu_0$ for a reference distribution μ_0, solve

$$\theta_\star \in \arg\min_{\theta \in \Theta} \ \mathrm{KGD}_K(Q_\theta).$$

KGD Descent: Let's take the discrete distribution $\hat{Q}_n = \frac{1}{n} \sum_{j=1}^n \delta_{x_j}$, we solve

$$\{x_1, ..., x_n\} \in \arg\min \mathrm{KGD}_K(\hat{Q}_n).$$



Experiments: Predictively Oriented Posteriors

Let $p(.|x)$ a parametric statistical model for independant data $\{y_i\}_{i=1,\dots,N}$.

$$\mathcal{L}(Q) = \frac{1}{2\lambda_N} \mathrm{MMD}^2 \left(\int p(\cdot|x) \, \mathrm{d}Q(x), \frac{1}{N} \sum_{i=1}^N \delta_{y_i} \right) \tag{1}$$

Extensible Sampling: Start with $x_0 \in \mathbb{R}^d$ and then apply:

$$x_n \in \arg\min_{x \in \mathbb{R}^d} \ \mathrm{KGD}_K \left(\frac{1}{n} \delta_x + \frac{1}{n} \sum_{i=1}^{n-1} \delta_{x_i} \right) \tag{2}$$

Variational Gradient Descent: This algorithm is a generalised version of SVGD. Let $v : \mathbb{R}^d \to \mathbb{R}^d$ and $\varepsilon > 0$

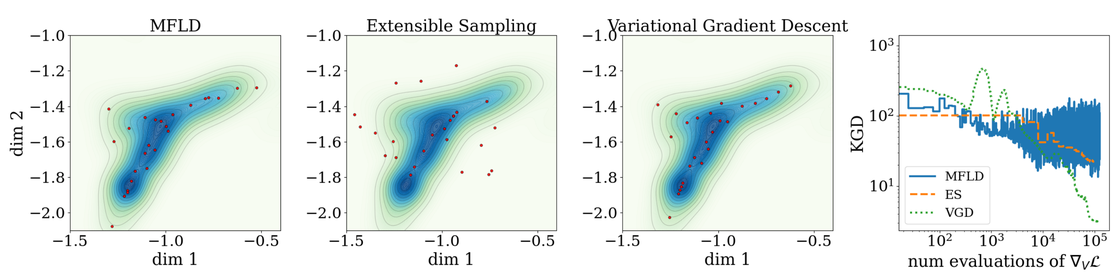
$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathcal{J}((\mathrm{I}_d + \epsilon v)_\# Q) \Big|_{\epsilon=0} = - \int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x).$$

Then the optimal direction in $\mathcal{H}_K$ is

$$v_Q(\cdot) \propto \int k_K^Q(x, .) \mathrm{d}Q(x),$$

we deduce the sampling algorithm :

$$x_i^{t+1} = x_i^t + \frac{1}{n} \sum_{j=1}^n k(x_i^t, x_j^t)(\nabla \log q_0 - \nabla_{\mathrm{V}} \mathcal{L}(Q_n^t))(x_j^t) + \nabla_1 k(x_j^t, x_i^t),$$



[1] H. Kanagawa, A. Barp, A. Gretton, and L. Mackey. Controlling moments with kernel Stein discrepancies, 2025.

[2] L. Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. Transactions on Machine Learning Research, 2022.