A Computable Measure of Suboptimality for Entropy-Regularised Variational Objectives

Clémentine Chazal, Heishiro Kanagawa, Zheyang Shen, Anna Korba, Chris. J. Oates

October 22th, 2025

Introduction

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

$$\mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0) \tag{1}$$

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

Introduction

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

$$\mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0) \tag{1}$$

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

The principal issue is that P is not tractable,

Introduction

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

$$\mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0) \tag{1}$$

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

The principal issue is that P is not tractable,

- We do not have access to the unormalized density of P (except if $\mathcal{L}$ is linear: if $\mathcal{L} = \int v(x)\mathrm{d}Q(x)$, then $\mathcal{J}(Q) = \mathrm{KLD}(Q||e^{-v}Q_0)$ and $P \propto e^{-v}Q_0$).

Introduction

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

$$\mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0) \tag{1}$$

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

The principal issue is that P is not tractable,

- We do not have access to the unormalized density of P (except if $\mathcal{L}$ is linear: if $\mathcal{L} = \int v(x)\mathrm{d}Q(x)$, then $\mathcal{J}(Q) = \mathrm{KLD}(Q||e^{-v}Q_0)$ and $P \propto e^{-v}Q_0$).
- $\mathcal{J}$ cannot be computed on discrete distributions as $\hat{Q} = \sum_{i=1}^{n} \delta_{x_i}$ because of the Kullback Leibler term.

Introduction

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

$$\mathcal{J}(Q) := \mathcal{L}(Q) + \mathrm{KLD}(Q||Q_0) \tag{1}$$

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

The principal issue is that P is not tractable,

- We do not have access to the unormalized density of P (except if $\mathcal{L}$ is linear: if $\mathcal{L} = \int v(x)\mathrm{d}Q(x)$, then $\mathcal{J}(Q) = \mathrm{KLD}(Q||e^{-v}Q_0)$ and $P \propto e^{-v}Q_0$).
- $\mathcal{J}$ cannot be computed on discrete distributions as $\hat{Q} = \sum_{i=1}^{n} \delta_{x_i}$ because of the Kullback Leibler term.

Intuition : Instead of minimizing $\mathcal{J}$, minimizing the 'size of the gradient of $\mathcal{J}$', $\|\nabla_{\mathrm{V}}\mathcal{J}(Q)\| = \sup_{\|v\|\leqslant 1} \langle \nabla_{\mathrm{V}}\mathcal{J}(Q), v\rangle_{L^2(Q)}$.

Kernel Gradient Discrepancy

Gradient Discrepancy

If Q and Q_0 admit density function respectively q and q_0,

$$\nabla_{\mathrm{V}} \mathcal{J}(Q)(x) = \nabla_{\mathrm{V}} \mathcal{L}(Q)(x) + \nabla \log \frac{q(x)}{q_0(x)}.$$

Gradient Discrepancy

If Q and Q_0 admit density function respectively q and q_0,

$$\nabla_{\mathrm{V}}\mathcal{J}(Q)(x) = \nabla_{\mathrm{V}}\mathcal{L}(Q)(x) + \nabla \log \frac{q(x)}{q_0(x)}.$$

Projecting the $\nabla_{\mathrm{V}}\mathcal{J}(Q)$ on the vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ gives

$$\begin{aligned}
&\int \nabla_{\mathrm{V}}\mathcal{J}(Q)(x) \cdot v(x)\, \mathrm{d}Q(x) \\
&= \int [\nabla_{\mathrm{V}}\mathcal{L}(Q)(x) - (\nabla \log q_0)(x)] \cdot v(x)\, \mathrm{d}Q(x) + \int (\nabla \log q)(x) \cdot v(x)\, \mathrm{d}Q(x) \\
&= \int [\nabla_{\mathrm{V}}\mathcal{L}(Q)(x) - (\nabla \log q_0)(x)] \cdot v(x)\, \mathrm{d}Q(x) - \int (\nabla \cdot v)(x)\, \mathrm{d}Q(x) \\
&= -\int \mathcal{T}_Q v(x)\, \mathrm{d}Q(x), \qquad \mathcal{T}_Q v(x) := [(\nabla \log q_0)(x) - \nabla_{\mathrm{V}}\mathcal{L}(Q)(x)] \cdot v(x) + (\nabla \cdot v)(x).
\end{aligned}$$

Gradient Discrepancy

If Q and Q_0 admit density function respectively q and q_0,

$$\nabla_{\mathrm{V}} \mathcal{J}(Q)(x) = \nabla_{\mathrm{V}} \mathcal{L}(Q)(x) + \nabla \log \frac{q(x)}{q_0(x)}.$$

Projecting the $\nabla_{\mathrm{V}} \mathcal{J}(Q)$ on the vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ gives

$$\begin{aligned}
&\int \nabla_{\mathrm{V}} \mathcal{J}(Q)(x) \cdot v(x) \, \mathrm{d}Q(x) \\
&= \int [\nabla_{\mathrm{V}} \mathcal{L}(Q)(x) - (\nabla \log q_0)(x)] \cdot v(x) \, \mathrm{d}Q(x) + \int (\nabla \log q)(x) \cdot v(x) \, \mathrm{d}Q(x) \\
&= \int [\nabla_{\mathrm{V}} \mathcal{L}(Q)(x) - (\nabla \log q_0)(x)] \cdot v(x) \, \mathrm{d}Q(x) - \int (\nabla \cdot v)(x) \, \mathrm{d}Q(x) \\
&= -\int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x), \qquad \mathcal{T}_Q v(x) \coloneqq [(\nabla \log q_0)(x) - \nabla_{\mathrm{V}} \mathcal{L}(Q)(x)] \cdot v(x) + (\nabla \cdot v)(x).
\end{aligned}$$

Then, one can define the **Gradient Discrepancy**,

$$\mathrm{GD}(Q) \coloneqq \sup_{\substack{v \in \mathcal{V} \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x) \right|.$$

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

$$\mathrm{KGD}_K(Q) := \sup_{\substack{v \in \mathcal{B}_K \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x) \right|$$

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

$$\mathrm{KGD}_K(Q) := \sup_{\substack{v \in \mathcal{B}_K \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x) \right|$$

Remarking that $\mathrm{KGD}_K(Q) = \sup_{\substack{v \in \mathcal{B}_K \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left\langle \int k_K^Q(x, .) \mathrm{d}Q(x), v \right\rangle$ where

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

and $\rho_Q(x) := q_0(x) \exp(-\mathcal{L}'(Q)(x))$,

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

$$\mathrm{KGD}_K(Q) := \sup_{\substack{v\in\mathcal{B}_K \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left| \int \mathcal{T}_Q v(x)\, \mathrm{d}Q(x) \right|$$

Remarking that $\mathrm{KGD}_K(Q) = \sup_{\substack{v\in\mathcal{B}_K \text{ s.t.} \\ (\mathcal{T}_Q v)_- \in \mathcal{L}^1(Q)}} \left\langle \int k_K^Q(x,.)\mathrm{d}Q(x), v \right\rangle$ where

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

and $\rho_Q(x) := q_0(x) \exp(-\mathcal{L}'(Q)(x))$, we finnaly get

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

Experiments

Experiments : MFNN

How to sample from P ? A popular algorithm: **MFLD** (Mean Field Langevin Dynamics algorithm),

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

Experiments : MFNN

How to sample from P ? A popular algorithm: **MFLD** (Mean Field Langevin Dynamics algorithm),

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

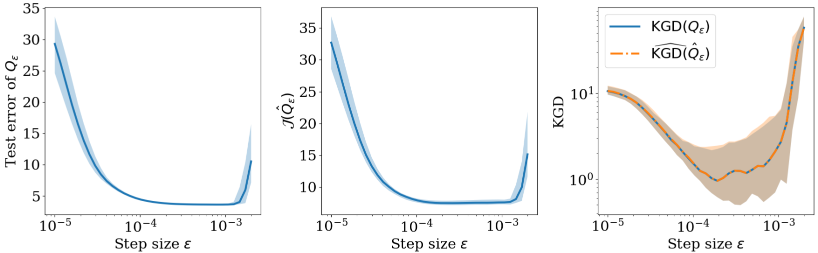
MFNN : Consider independant observations $(z_1, y_n), \ldots, (z_N, y_N)$ linked by $y_i = f(z_i) + \xi_i$, $\xi_i \sim \mathcal{N}(0, \sigma^2)$ where f is a target function. We take $\mathcal{L}$ to be the loss of a regression problem

$$\mathcal{L}(Q) = \frac{\lambda}{N}\sum_{i=1}^{N} \ell(y_i, \mathbb{E}_{X\sim Q}[\Phi(z_i, X)]), \tag{2}$$

where Φ is a Neural Network with parameter X. We want $f \approx \mathbb{E}_{X\sim Q}[\Phi(z_i, X)]$.

MFNN : Stepsize selection with KGD

We propose KGD as a measure to evaluate the best step size for MFLD.



MFNN : Novel sampling algorithms

For this example, we have implemented two new methods whose purpose is to optimise KGD:

- **Variational Inference**: Consider $Q_\theta = T^\theta_\# \mu_0$ for a reference distribution μ_0, we solve

  $$\theta_\star \in \operatorname*{arg\,min}_{\theta \in \Theta} \ \mathrm{KGD}_K(Q_\theta)$$

  by doing a gradient descent on θ.

MFNN : Novel sampling algorithms

For this example, we have implemented two new methods whose purpose is to optimise KGD:

- **Variational Inference**: Consider $Q_\theta = T^\theta_\# \mu_0$ for a reference distribution μ_0, we solve

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

  by doing a gradient descent on θ.

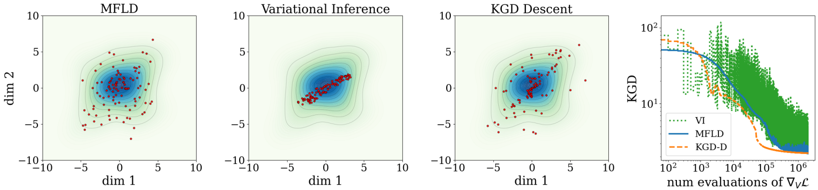
- **KGD Descent**: Let's take the discrete distribution $\hat{Q}_n = \frac{1}{n} \sum_{j=1}^n \delta_{x_j}$, we solve

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

  with gradient descent : $x_i^{t+1} = x_i^t - \varepsilon \nabla_{\mathrm{V}} \mathrm{KGD}^2_K(Q^t_n)(x^t_i)$.

MFNN : Comparison of all the methods

Here is plot the final distribution of the parameters for all the methods.



Predictively Oriented Posteriors

Another example of application is **Predictively Oriented Posteriors**. Let $p(.|x)$ a parametric statistical model for independant data $\{y_i\}_{i=1,...,N}$. Let's take

$$\mathcal{L}(Q) = \frac{1}{2\lambda_N}\mathrm{MMD}^2\left(\int p(\cdot|x)\,\mathrm{d}Q(x), \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}\right) \tag{3}$$

Predictively Oriented Posteriors

Another example of application is **Predictively Oriented Posteriors**. Let $p(.|x)$ a parametric statistical model for independant data $\{y_i\}_{i=1,...,N}$. Let's take

$$\mathcal{L}(Q) = \frac{1}{2\lambda_N} \mathrm{MMD}^2 \left(\int p(\cdot|x) \, \mathrm{d}Q(x), \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i} \right) \tag{3}$$

Algoritms used for this example:

- **Extensible Sampling:** Start from $x_0 \in \mathbb{R}^d$ and then apply the iterative algorithm:

  $$x_n \in \underset{x \in \mathbb{R}^d}{\arg\min} \ \mathrm{KGD}_K \left(\frac{1}{n} \delta_x + \frac{1}{n} \sum_{i=1}^{n-1} \delta_{x_i} \right) \tag{4}$$

  where the minimum is searched on a grid in $\mathbb{R}^d$.

Predictively Oriented Posteriors: Variational Gradient Descent

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

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

Predictively Oriented Posteriors: Variational Gradient Descent

- **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 is proportional to $\int k_K^Q(x, .)\mathrm{d}Q(x)$ which is

$$v_Q(\cdot) \propto \int \{k(x,\cdot)(\nabla \log q_0 - \nabla_{\mathrm{V}} \mathcal{L}(Q))(x) + \nabla_1 k(x,\cdot)\} \, \mathrm{d}Q(x),$$

Predictively Oriented Posteriors: Variational Gradient Descent

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

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

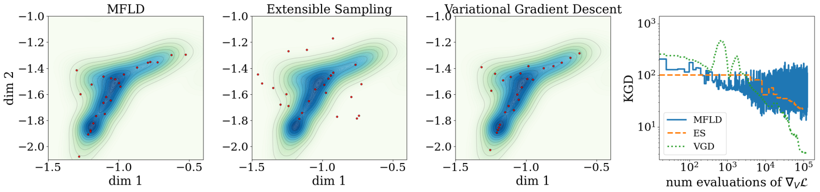
Then the optimal direction is proportional to $\int k_K^Q(x, .)\mathrm{d}Q(x)$ which is

$$v_Q(\cdot) \propto \int \{k(x,\cdot)(\nabla \log q_0 - \nabla_{\mathrm{V}}\mathcal{L}(Q))(x) + \nabla_1 k(x,\cdot)\}\,\mathrm{d}Q(x),$$

And then, we deduce the sampling algorithm sampling algorithm :

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

Predictively Oriented Posteriors : Comparison of the methods



Thank you !