A Computable Measure of Suboptimality for Entropy-Regularised Variational Objectives

Clémentine Chazal

CREST, ENSAE Paris

Model-Oriented Data Analysis and Optimum Design (mODa14)

June 19th, 2026

Joint work with




Heishiro Kanagawa
Newcastle University, UK

Zheyang Shen
Newcastle University, UK

Anna Korba
CREST/ Ensae, Paris, France

Chris J. Oates
Newcastle University/The Alan Turing Institute, UK

Introduction

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

$$\mathcal{J}(Q) \coloneqq \underbrace{\mathcal{L}(Q)}_{\text{loss}} + \underbrace{\mathrm{KL}(Q||Q_0)}_{\textit{regularisation}} \tag{1}$$

where $Q_0 = \mathcal{N}(0, I_d)$, $\mathrm{KL}(Q||Q_0) = \int \log(Q/Q_0)dQ$ if $Q \ll Q_0$, $+\infty$ otherwise.

[1] [McLatchie et al., 2025]

Introduction

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

$$\mathcal{J}(Q) := \underbrace{\mathcal{L}(Q)}_{\text{loss}} + \underbrace{\mathrm{KL}(Q||Q_0)}_{\textit{regularisation}} \tag{1}$$

where $Q_0 = \mathcal{N}(0, I_d)$, $\mathrm{KL}(Q||Q_0) = \int \log(Q/Q_0)dQ$ if $Q \ll Q_0$, $+\infty$ otherwise.

Examples of $\mathcal{L}$:

- **Bayesian statistics**: (linear) $\mathcal{L}(Q) = -\int \log p(\mathcal{D}|x)\, \mathrm{d}Q(x)$ for data $\mathcal{D}$ and parametric model $p(.|x)$.

[1] [McLatchie et al., 2025]

Introduction

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

$$\mathcal{J}(Q) \coloneqq \underbrace{\mathcal{L}(Q)}_{\text{loss}} + \underbrace{\mathrm{KL}(Q||Q_0)}_{\textit{regularisation}} \tag{1}$$

where $Q_0 = \mathcal{N}(0, I_d)$, $\mathrm{KL}(Q||Q_0) = \int \log(Q/Q_0) dQ$ if $Q \ll Q_0$, $+\infty$ otherwise.

Examples of $\mathcal{L}$:

- **Bayesian statistics**: (linear) $\mathcal{L}(Q) = -\int \log p(\mathcal{D}|x)\, \mathrm{d}Q(x)$ for data $\mathcal{D}$ and parametric model $p(.|x)$.
- **Mean-Field Neural Network (MFNN)**: (quadratic) $\mathcal{L}(Q) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbb{E}_{x\sim Q}[\Phi(z_i, x)])^2$ for data $(z_i, y_i)_{i=1}^{N}$ and neural network Φ.

[1] [McLatchie et al., 2025]

Introduction

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

$$\mathcal{J}(Q) \coloneqq \underbrace{\mathcal{L}(Q)}_{\text{loss}} + \underbrace{\mathrm{KL}(Q||Q_0)}_{\textit{regularisation}} \tag{1}$$

where $Q_0 = \mathcal{N}(0, I_d)$, $\mathrm{KL}(Q||Q_0) = \int \log(Q/Q_0)dQ$ if $Q \ll Q_0$, $+\infty$ otherwise.

Examples of $\mathcal{L}$:

- **Bayesian statistics**: (linear) $\mathcal{L}(Q) = -\int \log p(\mathcal{D}|x)\, \mathrm{d}Q(x)$ for data $\mathcal{D}$ and parametric model $p(.|x)$.
- **Mean-Field Neural Network (MFNN)**: (quadratic) $\mathcal{L}(Q) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbb{E}_{x\sim Q}[\Phi(z_i, x)])^2$ for data $(z_i, y_i)_{i=1}^N$ and neural network Φ.
- **Predictively oriented posterior**[1]: $\mathcal{L}(Q) = -\frac{1}{\lambda_N}\sum_{i=1}^{N} S\left(\int p(\cdot|x)\, \mathrm{d}Q(x), \delta_{y_i}\right)$, $(y_i)_{i=1,..,n}$ observations.

[1] [McLatchie et al., 2025]

Problem and Objective

The only thing we know about P is that it satisfies

$$\frac{\mathrm{d}P}{\mathrm{d}Q_0} \propto \exp(-\mathcal{L}'(P))$$

where $\mathcal{L}'(Q) : \mathbb{R}^d \to \mathbb{R}$ is the first variation[2].

[2] $\lim_{\epsilon \to 0} \frac{1}{\epsilon}\{\mathcal{L}(Q + \epsilon\chi) - \mathcal{L}(Q)\} = \int \mathcal{L}'(Q)\,\mathrm{d}\chi,\ \forall \chi = R - Q,\ R \in \mathcal{P}(\mathbb{R}^d)$.

Problem and Objective

The only thing we know about P is that it satisfies

$$\frac{\mathrm{d}P}{\mathrm{d}Q_0} \propto \exp(-\mathcal{L}'(P))$$

where $\mathcal{L}'(Q) : \mathbb{R}^d \to \mathbb{R}$ is the first variation[2].

- **Goal:** Approximate P with a discrete distribution $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$:

[2] $\lim_{\epsilon \to 0} \frac{1}{\epsilon}\{\mathcal{L}(Q + \epsilon\chi) - \mathcal{L}(Q)\} = \int \mathcal{L}'(Q)\,\mathrm{d}\chi,\ \forall \chi = R - Q,\ R \in \mathcal{P}(\mathbb{R}^d)$.

Problem and Objective

The only thing we know about P is that it satisfies

$$\frac{\mathrm{d}P}{\mathrm{d}Q_0} \propto \exp(-\mathcal{L}'(P))$$

where $\mathcal{L}'(Q) : \mathbb{R}^d \to \mathbb{R}$ is the first variation[2].

- **Goal:** Approximate P with a discrete distribution $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$:

Problem:

- On discrete distributions, $\hat{Q} = \sum_{i=1}^{n} \delta_{x_i}$, $\mathrm{KL}(\hat{Q}||Q_0) = +\infty$ and so $\mathcal{J}(\hat{Q}) = +\infty$.

[2] $\lim_{\epsilon \to 0} \frac{1}{\epsilon}\{\mathcal{L}(Q + \epsilon\chi) - \mathcal{L}(Q)\} = \int \mathcal{L}'(Q)\,\mathrm{d}\chi$, $\forall \chi = R - Q$, $R \in \mathcal{P}(\mathbb{R}^d)$.

Problem and Objective

The only thing we know about P is that it satisfies

$$\frac{\mathrm{d}P}{\mathrm{d}Q_0} \propto \exp(-\mathcal{L}'(P))$$

where $\mathcal{L}'(Q) : \mathbb{R}^d \to \mathbb{R}$ is the first variation[2].

- **Goal:** Approximate P with a discrete distribution $\hat{Q}=\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$:

Problem:

- On discrete distributions, $\hat{Q} = \sum_{i=1}^{n}\delta_{x_i}$, $\mathrm{KL}(\hat{Q}||Q_0) = +\infty$ and so $\mathcal{J}(\hat{Q}) = +\infty$.

$\rightarrow$ Find a substitute loss for $\mathcal{J}$ that does not need the condition $Q \ll Q_0$.

[2] $\lim_{\epsilon\to 0}\frac{1}{\epsilon}\{\mathcal{L}(Q+\epsilon\chi)-\mathcal{L}(Q)\} = \int \mathcal{L}'(Q)\,\mathrm{d}\chi$, $\forall\chi = R - Q,\ R \in \mathcal{P}(\mathbb{R}^d)$.

Motivation from $\mathbb{R}^d$

Consider $J : \mathbb{R}^d \to \mathbb{R}$ and the minimisation problem

$$x \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} J(x).$$

If x^* is a minimiser of J then $\nabla J(x^*) = 0$.

Motivation from $\mathbb{R}^d$

Consider $J : \mathbb{R}^d \to \mathbb{R}$ and the minimisation problem

$$x \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} J(x).$$

If x^* is a minimiser of J then $\nabla J(x^*) = 0$.

• **Idea:** instead of looking for minimiser directly let's look for stationary points: solve the minimisation

$$x \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \|\nabla J(x)\|^2.$$

Motivation from $\mathbb{R}^d$

Consider $J : \mathbb{R}^d \to \mathbb{R}$ and the minimisation problem

$$x \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} J(x).$$

If x^* is a minimiser of J then $\nabla J(x^*) = 0$.

• **Idea:** instead of looking for minimiser directly let's look for stationary points: solve the minimisation

$$x \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \|\nabla J(x)\|^2.$$

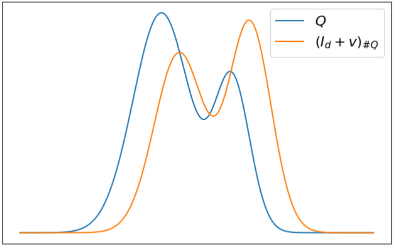
$\rightarrow$ We want to generalise this idea for $\mathcal{J}$ in $\mathcal{P}(\mathbb{R}^d)$.

Gradient in $\mathcal{P}(\mathbb{R}^d)$: Wasserstein gradient

Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$, consider the push forward distribution

$$(I_d + \varepsilon v)_{\#Q}$$

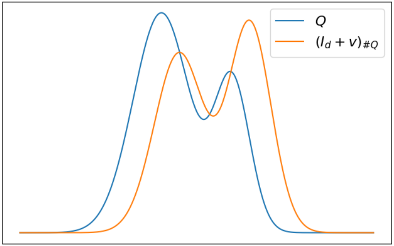
Compare $\mathcal{J}((I_d + \varepsilon v)_{\#Q})$ and $\mathcal{J}(Q)$.



Gradient in $\mathcal{P}(\mathbb{R}^d)$: Wasserstein gradient

Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$, consider the push forward distribution

$$(I_d + \varepsilon v)_{\#Q}$$



Compare $\mathcal{J}((I_d + \varepsilon v)_{\#Q})$ and $\mathcal{J}(Q)$.

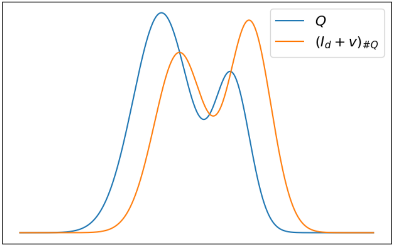
$\rightarrow$ $\nabla_V \mathcal{J}(Q) : \mathbb{R}^d \to \mathbb{R}^d$ is the direction with steepest slope for $\mathcal{J}$.

Gradient in $\mathcal{P}(\mathbb{R}^d)$: Wasserstein gradient

Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$, consider the push forward distribution

$$(I_d + \varepsilon v)_{\#Q}$$

Compare $\mathcal{J}((I_d + \varepsilon v)_{\#Q})$ and $\mathcal{J}(Q)$.



$\rightarrow \nabla_V \mathcal{J}(Q) : \mathbb{R}^d \to \mathbb{R}^d$ is the direction with steepest slope for $\mathcal{J}$.

Definition

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

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

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

The Wasserstein gradient can be computed by the formula:

$$\nabla_{\mathrm{V}} \mathcal{J}(Q)(x) := \nabla_x \mathcal{J}'(Q)(x)$$

for each $x \in \mathbb{R}^d$.

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,

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

To measure the 'size' of the gradient let us project the $\nabla_{\mathrm{V}}\mathcal{J}(Q)$ on the vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ gives

$$\begin{aligned}\langle \nabla_{\mathrm{V}}\mathcal{J}(Q), v\rangle_{L^2(Q)} &= \int \nabla_{\mathrm{V}}\mathcal{J}(Q)(x) \cdot v(x)\,\mathrm{d}Q(x) \\ &= -\int \mathcal{T}_Q v(x)\,\mathrm{d}Q(x)\end{aligned}$$

with $\mathcal{T}_Q v(x) \coloneqq [(\nabla \log q_0)(x) - \nabla_{\mathrm{V}}\mathcal{L}(Q)(x)] \cdot v(x) + (\nabla \cdot v)(x).$

Gradient Discrepancy

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

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

To measure the 'size' of the gradient let us project the $\nabla_{\mathrm{V}}\mathcal{J}(Q)$ on the vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ gives

$$\begin{aligned}\langle \nabla_{\mathrm{V}}\mathcal{J}(Q), v\rangle_{L^2(Q)} &= \int \nabla_{\mathrm{V}}\mathcal{J}(Q)(x) \cdot v(x)\, \mathrm{d}Q(x) \\ &= -\int \mathcal{T}_Q v(x)\, \mathrm{d}Q(x)\end{aligned}$$

with $\mathcal{T}_Q v(x) \coloneqq [(\nabla \log q_0)(x) - \nabla_{\mathrm{V}}\mathcal{L}(Q)(x)] \cdot v(x) + (\nabla \cdot v)(x)$.

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

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

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

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

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

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

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

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

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

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

and $\rho_Q(x) \coloneqq q_0(x)\exp(-\mathcal{L}'(Q)(x))$, we finaly get

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

→ KGD is defined on discrete distribution.

Theorical guarantees on KGD

1. **Separability:** $\mathrm{KGD}_K(Q) = 0$ if and only if Q is a stationary point of $\mathcal{J}$. RKHS with sufficiently many vector fields

[3] We extend the results of [Barp et al., 2024] obtained for KSD.

Theorical guarantees on KGD

1. **Separability:** $\mathrm{KGD}_K(Q) = 0$ if and only if Q is a stationary point of $\mathcal{J}$. RKHS with sufficiently many vector fields

Definition (α-convergence)

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. $h(x) \lesssim 1 + \|x\|^\alpha$.

2. **Continuity:** $\mathrm{KGD}_K(Q_n) \to \mathrm{KGD}_K(Q)$ whenever $Q_n \xrightarrow{\alpha} Q$. Growth and continuity of kernel.

[3] We extend the results of [Barp et al., 2024] obtained for KSD.

Theorical guarantees on KGD

1. **Separability:** $\mathrm{KGD}_K(Q) = 0$ if and only if Q is a stationary point of $\mathcal{J}$. RKHS with sufficiently many vector fields

Definition (α-convergence)

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. $h(x) \lesssim 1 + \|x\|^\alpha$.

2. **Continuity:** $\mathrm{KGD}_K(Q_n) \to \mathrm{KGD}_K(Q)$ whenever $Q_n \xrightarrow{\alpha} Q$. Growth and continuity of kernel.
3. **Convergence control**[3]**:** $\mathrm{KGD}_K(Q_n) \to 0$ implies $Q_n \xrightarrow{\alpha} P \in \mathcal{P}_\alpha(\mathbb{R}^d)$. Separability/Continuity + Uniform integrability (Dissipative $\nabla \log q0$).

[3] We extend the results of [Barp et al., 2024] obtained for KSD.

Existence and uniqueness of stationary points

In which case $\mathcal{J}$ admits a unique minimiser ?

Assume that

1. *(convexity)* $\mathcal{L}$ is convex and lower-bounded;
2. *(regularity)* $(Q, x) \mapsto \mathcal{L}'(Q)(x)$ is (weakly) continuous in Q and bounded in (Q, x).

Then $\mathcal{J}$ admits a unique minimiser P.

Experiments

Predictively Oriented Posteriors

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

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

Predictively Oriented Posteriors

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

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

- If $p(.|x)$ is well specified for a true parameter x^*, then $P = \delta_{x^*}$.

Predictively Oriented Posteriors

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

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

- If $p(.|x)$ is well specified for a true parameter x^*, then $P = \delta_{x^*}$.
- If $p(.|x)$ is misspecified, the set of possible distribution to approximate the data is larger.

Algorithms

- **MFLD:** (Mean Field Langevin Dynamics algorithm[4]),

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

  Algoritms used for this example:

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

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

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

[4] [Del Moral, 2013]

Predictively Oriented Posteriors: Variational Gradient Descent

- **Variational Gradient Descent:** This algorithm is a generalised version of SVGD[5]. Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $v \in \mathcal{H}_K$ and $\varepsilon > 0$

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

[5] introduced in [Liu and Wang, 2016], and generalised in [Wang and Liu, 2019].

Predictively Oriented Posteriors: Variational Gradient Descent

- **Variational Gradient Descent:** This algorithm is a generalised version of SVGD[5]. Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $v \in \mathcal{H}_K$ 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 proportional to $\int k_K^Q(x,.)\mathrm{d}Q(x)$ which is

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

[5] introduced in [Liu and Wang, 2016], and generalised in [Wang and Liu, 2019].

Predictively Oriented Posteriors: Variational Gradient Descent

- **Variational Gradient Descent:** This algorithm is a generalised version of SVGD[5]. Let $v : \mathbb{R}^d \to \mathbb{R}^d$, $v \in \mathcal{H}_K$ 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 proportional to $\int k_K^Q(x,.)\mathrm{d}Q(x)$ which is

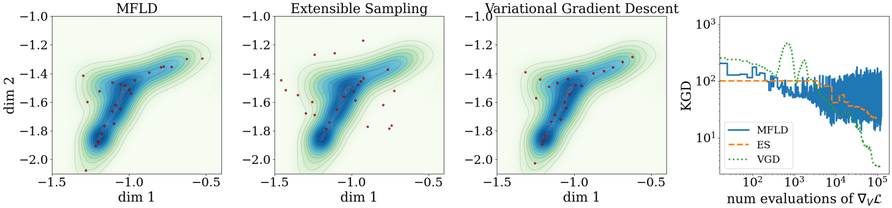
$$v_Q(\cdot) \propto \int \{k(x,\cdot)(\nabla \log q_0 - \nabla_{\mathrm{V}}\mathcal{L}(Q))(x) + \nabla_1 k(x,\cdot)\}\,\mathrm{d}Q(x).$$

And then, we deduce the sampling algorithm:

$$\begin{aligned} x_i^{t+1} &= x_i^t + v_{Q_n}(x_i^t), \quad \forall i = 1,..,n,\ t = 1,..,T \\ &= x_i^t + \frac{1}{n}\sum_{j=1}^n k(x_i^t, x_j^t)(\nabla \log q_0 - \nabla_{\mathrm{V}}\mathcal{L}(Q_n^t))(x_j^t) + \nabla_1 k(x_j^t, x_i^t). \end{aligned}$$

[5] introduced in [Liu and Wang, 2016], and generalised in [Wang and Liu, 2019].

Predictively Oriented Posteriors : Comparison of the methods



- Contours are plotted running for $t = 10^5$, stepsize $\eta = 10^{-6}$. kernel density estimation is used to compute the density.
- •: final particles distribution.
- 4th graph : Evolutions of KGD with iterations: confirms that the values used for KGD are consistent with the results obtained.

Mean Field Neural Network

MFNN : Consider independant observations $(z_1, y_n), ..., (z_N, y_N)$ 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

$$\mathcal{L}(Q) = \frac{\lambda}{N} \sum_{i=1}^{N} \ell(y_i, \mathbb{E}_{X \sim Q}[\Phi(z_i, X)]), \tag{3}$$

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

Mean Field Neural Network

MFNN : Consider independant observations $(z_1, y_n), ..., (z_N, y_N)$ 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

$$\mathcal{L}(Q) = \frac{\lambda}{N} \sum_{i=1}^{N} \ell(y_i, \mathbb{E}_{X \sim Q}[\Phi(z_i, X)]), \tag{3}$$

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

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

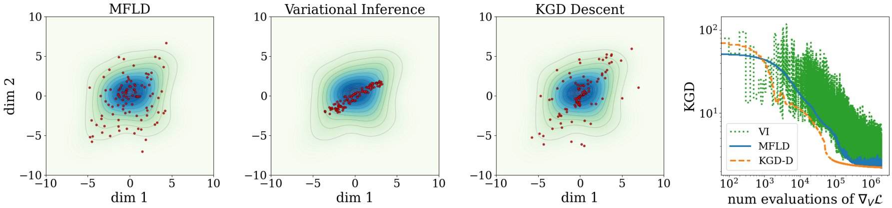
  by doing a gradient descent on θ.

- **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 : $x_i^{t+1} = x_i^t - \varepsilon \nabla_{\mathrm{V}} \mathrm{KGD}^2_K(Q_n^t)(x_i^t)$.

Mean Field Neural Network: Comparison of the methods



- Same setting as previous figure.
- Problem: The new methods require to differentiate KGD.

References

Barp, A., Simon-Gabriel, C.-J., Girolami, M., and Mackey, L. (2024).
Targeted separation and convergence with kernel discrepancies.
Journal of Machine Learning Research, 25(378):1–50.

Del Moral, P. (2013).
Mean field simulation for Monte Carlo integration.
Monographs on Statistics and Applied Probability, 126(26):6.

Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose Bayesian inference algorithm.
Advances in Neural Information Processing Systems, 29.

McLatchie, Y., Cherief-Abdellatif, B.-E., Frazier, D. T., and Knoblauch, J. (2025).
Predictively oriented posteriors.
arXiv preprint arXiv:2510.01915.

Wang, D. and Liu, Q. (2019).
Nonlinear Stein variational gradient descent for learning diversified mixture models.
In *International Conference on Machine Learning*, pages 6576–6585. PMLR.

Link between KGD and KSD in Linear case

- For a target distribution P with density p, the **Kernel Stein Discrepancy (KSD)** is defined as:

$$\mathrm{KSD}^2(Q|P) := \sup_{v \in \mathcal{B}_K} \left| \int \mathcal{T}_Q v(x) \, \mathrm{d}Q(x) \right| = \iint k_P(x, y) dQ(x) dQ(y)$$

where $k_P(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{1}{p(x)p(x')} \partial_{x'_j} \partial_{x_i} \left(p(x) K_{i,j}(x, x') p(x') \right)$.

- If $\mathcal{L}(Q) = \int V(x) \mathrm{d}Q(x)$, $V : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{L}'(Q)(x) = V(x)$. Then

$$\mathcal{J}(Q) = \mathrm{KLD}(Q || e^{-V} Q_0)$$

and so $P \propto e^{-V} Q_0$ and $\rho_Q(x) = q_0(x) e^{-\mathcal{L}'(Q)(x)} = q_0(x) e^{-V(x)} = p$. We have

$$\mathrm{KGD}_K(Q) = \mathrm{KSD}(Q || P)$$