BML Project: Optimally-Weighted Herding is Bayesian Quadrature

Clément Bonet[*]
Department of Mathematics
ENS Paris-Saclay
94230 Cachan
`clement.bonet@telecom-paris.fr`

Christophe Vuong
Department of Mathematics
ENS Paris-Saclay
94230 Cachan
`christophe.vuong@telecom-paris.fr`

Abstract

Kernel herding is a deterministic method of choosing samples summarising a probability distribution. A related task is choosing samples for estimating integrals using Bayesian quadrature. The authors show that the criterion minimised when selecting samples in kernel herding is equivalent to the posterior variance in Bayesian quadrature. They then show that sequential Bayesian quadrature can be viewed as an optimal weighted version of herding, and demonstrate empirically a rate of convergence faster than $\mathcal{O}\left(\frac{1}{n}\right)$.

1 Introduction

In the article (Huszár and Duvenaud, 2012), the authors deal with the problem of computing integrals of the form $Z_{f,p} = \int f(x)p(x)dx$ where the distribution p is known analytically. To approximate such integrals, there exists several methods which are more or less efficient.

For example, we can use Monte Carlo methods which consist to evaluate random samples of p $i.i.d$ in order to approximate $Z_{f,p}$ with the law of large numbers. But this method converges only in $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ and is therefore pretty slow. Another classical method is to use Markov Chain Monte Carlo methods (MCMC) such as Metropolis-Hastings for example. These methods can deal with the case when we can't directly sample from p. The convergence rates of these methods depend on several factors.

In this paper, (Huszár and Duvenaud, 2012) present two quasi Monte Carlo methods, which consist of designing pseudo samples in a deterministic way: Kernel Herding and Bayesian Quadrature. Then they empirically show that the convergence rate is bigger than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

2 Herding

2.1 Maximum Mean Discrepancy

For selecting pseudosamples, herding relies on an objective based on the maximum mean discrepancy (MMD; Sriperumbudur et al., 2009). MMD measures the divergence between two distributions, p and q with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$\mathrm{MMD}_{\mathcal{F}}\left(p,q\right) = \sup_{f\in\mathcal{F}} \left| \int f_x p(x)dx - \int f_x q(x)dx \right| \tag{1}$$

[*] `https://github.com/clbonet/MVA-BML-Project-Optimally-Weighted-Herding-is-Bayesian-Quadrature` for the code of the project

Preprint. Under review.

Intuitively, if two distributions are close in the MMD sense, then no matter which function f we choose from $\mathcal{F}$, the difference in its integral over p or q should be small. A particularly interesting case is when the function class $\mathcal{F}$ is functions of unit norm from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. In this case, the MMD between two distributions can be conveniently expressed using expectations of the associated kernel $k(x, x')$ only (Sriperumbudur et al., 2009):

$$\begin{aligned} MMD^2_{\mathcal{H}}(p, q) = & \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}} \left| \int f_x p(x) dx - \int f_x q(x) dx \right|^2 \\ & = \|\mu_p - \mu_q\|^2_{\mathcal{H}} \\ & = \iint k(x, y) p(x) p(y) dx dy - 2 \iint k(x, y) p(x) q(y) dx dy \\ & + \iint k(x, y) q(x) q(y) dx dy, \end{aligned} \tag{2}$$

2.2 Criterion

Herding uses maximum mean discrepancy to evaluate of how well the sample set $\{x_1, \ldots, x_N\}$ represents the target distribution p:

$$\begin{aligned} \epsilon^2_{herding}\left(\{x_1, \ldots, x_N\}\right) = & \operatorname{MMD}^2_{\mathcal{H}}\left(p, \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n}\right) \\ = & \iint k(x, y) p(x) p(y) dx dy \\ & - \frac{2}{N} \sum_{n=1}^{N} \int k(x, x_n) p(x) dx + \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_n, x_m) \end{aligned} \tag{3}$$

For selecting sequentially samples, we minimized the following objective function at each iteration:

$$\begin{aligned} x_{n+1} \longleftarrow & \operatorname*{argmin}_{x \in \mathcal{X}} \epsilon^2_{herding}\left(\{x_1, \ldots, x_n, x\}\right) \\ & = \operatorname*{argmax}_{x \in \mathcal{X}} 2\mathbb{E}_{x' \sim p}\left[k(x, x')\right] - \frac{1}{n+1} \sum_{m=1}^{n} k(x, x_m) \end{aligned}$$

Intuitively the first term encourages the sampling around p and the second is a penalty in order to not sample too close of the others samples.

Then, when we have enough samples, we can approximate $Z_{p,f}$ with $\frac{1}{N} \sum_{n=1}^{N} f(x_n)$.

For the following, we choose the standard choice of Gaussian kernel k and $\mathcal{H}$ its RKHS:

$$k(x, y) = \frac{1}{(2\pi)^{D/2} |\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x - y)^\top \Gamma^{-1}(x - y)\right)$$

2.3 Example of a Gaussian as a distribution (Jebara and Kondor, 2003)

Let the distribution be $p(x) \sim \mathcal{N}(\mu, \Sigma)$

Let $\underline{Q = (\Gamma^{-1} + \Sigma^{-1})^{-1}}$, then $\underline{m(x) = Q(\Gamma^{-1}x + \Sigma^{-1}\mu)}$

We use that result (see the proof in appendix A):

$$\mathbb{E}_{y\sim p}\left[k(x,y)\right] = \frac{|Q|^{1/2}\exp\left[-\frac{1}{2}\left(x^\top\Gamma^{-1}x + \mu^\top\Sigma^{-1}\mu - m(x)^\top Q^{-1}m(x)\right)\right]}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}$$

$$\mathbb{E}_{y\sim p}\left[k(x,y)\right] = \frac{e^{-\frac{1}{2}\left(x^\top(\Gamma^{-1}-\Gamma^{-1}Q\Gamma^{-1})x+\mu^\top(\Sigma^{-1}-\Sigma^{-1}Q\Sigma^{-1})\mu-2\mu^\top\Sigma^{-1}Q\Gamma^{-1}x\right)}\left|\Gamma^{-1}+\Sigma^{-1}\right|^{-1/2}}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}$$

$$\begin{aligned}\Gamma^{-1}-\Gamma^{-1}Q\Gamma^{-1} &= \Gamma^{-1}-\Gamma^{-1}(\Gamma^{-1}+\Sigma^{-1})^{-1}\Gamma^{-1} = \Gamma^{-1}-\Gamma^{-1}(\Gamma^{-1}+\Sigma^{-1})^{-1}(\Gamma^{-1}+\Sigma^{-1})+\Gamma^{-1}Q\Sigma^{-1}\\ &= \Gamma^{-1}-\Gamma^{-1}I_d+\Gamma^{-1}Q\Sigma^{-1} = \Gamma^{-1}Q\Sigma^{-1}\end{aligned}$$

The same goes for $\Sigma^{-1}-\Sigma^{-1}Q\Sigma^{-1} = \Gamma^{-1}Q\Sigma^{-1}$. So,

$$\begin{aligned}\mathbb{E}_{y\sim p}\left[k(x,y)\right] &= \frac{e^{-\frac{1}{2}\left(x^\top(\Gamma^{-1}Q\Sigma^{-1})x+\mu^\top(\Sigma^{-1}Q\Gamma^{-1})\mu-2\mu^\top\Sigma^{-1}Q\Gamma^{-1}x\right)}\left|\Gamma^{-1}+\Sigma^{-1}\right|^{-1/2}}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}\\ &= \frac{\left|\Gamma^{-1}+\Sigma^{-1}\right|^{-1/2}}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}\exp((x-\mu)^\top R^{-1}(x-\mu))\end{aligned}$$

where $\underline{R = \Sigma(\Sigma^{-1}+\Gamma^{-1})\Gamma = \Gamma+\Sigma}$

$$\boxed{\mathbb{E}_{y\sim p}\left[k(x,y)\right] = \frac{1}{(2\pi)^{D/2}\left|R\right|^{1/2}}\exp((x-\mu)^\top R^{-1}(x-\mu))} \tag{4}$$

2.4 Example of mixture of Gaussians as a distribution

With eq. (4), by the linearity of the integral, for:
$p(x) = \sum\limits_{i=1}^{K}\alpha_i\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_i)^\top\Sigma_i^{-1}(x-\mu_i)\right)$ and $\sum\limits_{i=1}^{K}\alpha_i = 1$. Let $R_i = \Gamma+\Sigma_i$,

$$\mathbb{E}_{y\sim p}\left[k(x,y)\right] = \sum_{i=1}^{K}\alpha_i\frac{1}{(2\pi)^{D/2}\left|R_i\right|^{1/2}}\exp\left[-\frac{1}{2}\left\{(x-\mu_i)^\top R_i^{-1}(x-\mu_i)\right\}\right] \tag{5}$$

Integrating over it boils down to do the same computations as in eq. (5), replacing Γ by R_i and x by μ_i.

Then, we have:

$$\mathbb{E}_{x,y\sim p}\left[k(x,y)\right] = \sum_{1\le i,j\le K}\alpha_i\alpha_j\frac{1}{(2\pi)^{D/2}\left|S_{ij}\right|^{1/2}}\exp\left[-\frac{1}{2}\left\{(\mu_i-\mu_j)^\top S_{ij}^{-1}(\mu_i-\mu_j)\right\}\right] \tag{6}$$

with $\underline{S_{ij} = R_i+\Sigma_i = \Gamma+\Sigma_i+\Sigma_j}$

From those equations, we derive the values for $\epsilon^2_{herding}$ and the objective function.

3 Bayesian Quadrature

3.1 Estimator

The Bayesian Quadrature method presented in Huszár and Duvenaud (2012) consists of putting a Gaussian process prior over f with kernel function k and mean 0. Then we can compute the posterior distribution $p(f(x')|f(x_1), ..., f(x_n))$ with $x_1, ..., x_n$ n points where f have been evaluated. Finally, this implies a distribution over $Z_{f,p}$, and we will be able to compute its expectation.

Let's compute the posterior:

Let $f(X) = (f(x_1), ..., f(x_n))$ then by definition of the prior, $f(X) \sim \mathcal{N}(0, K)$ where $K = \begin{pmatrix} k(x_i, x_j) \end{pmatrix}_{1 \leq i,j \leq n}$.

Let $\tilde{f} = \begin{pmatrix} f(x_1) \\ . \\ . \\ . \\ f(x_n) \\ f(x') \end{pmatrix}$ $q = \begin{pmatrix} k(x_1, x') \\ . \\ . \\ . \\ k(x_n, x') \end{pmatrix}$

Then:

$$\begin{aligned} p(f(x')|f(X)) &\propto p(f(x'), f(x_1), ..., f(x_n)) \\ &\propto exp(-\frac{1}{2}\tilde{f}^T \begin{pmatrix} K & q \\ q^T & k(x', x') \end{pmatrix}^{-1} \tilde{f}) \end{aligned}$$

By applying the inversion of partitioned matrix lemma (see in Rasmussen and Williams (2006) for example), we have:

$$\begin{pmatrix} K & q \\ q^T & k(x', x') \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{K} & \tilde{q} \\ \tilde{q}^T & \tilde{k}(x', x') \end{pmatrix}$$

with:

$$\begin{cases} \tilde{K} & = K^{-1} + K^{-1}q(k(x', x') - q^T K^{-1} q)^{-1} q^T K^{-1} \\ \tilde{q} & = -K^{-1}q(k(x', x') - q^T K^{-1} q)^{-1} \\ \tilde{k}(x', x') & = \frac{1}{k(x', x') - q^T K^{-1} q} \end{cases}$$

So, $p(f(x')|f(X)) \propto exp(-\frac{1}{2}\tilde{k}(x', x')f(x')^2 + 2f(X)^T \tilde{q} f(x'))$

We deduce that $\underline{p(f(x')|f(X)) = \mathcal{N}(f(X)^T K^{-1} q, k(x', x') - q^T K^{-1} q)}$.

Then we find:

$$\mathbb{E}_f \left[Z_{f,p}|f(X)\right] = f(X)^T K^{-1} z \tag{7}$$

with $z_n = \mathbb{E}_{x' \sim p}\left[k(x_n, x')\right]$ (see appendix B).

3.2 Criterion

To evaluate the uncertainty of the samples, and therefore select in a sequential way as Herding the new samples, we can use the posterior variance which is given by:

$$Var(Z_{f,p}|f(X)) = \mathbb{E}_{x,x' \sim p}\left[k(x, x')\right] - z^T K^{-1} z$$

Indeed (Rasmussen and Ghahramani, 2002):

$$\begin{aligned}
Var(Z_{f,p}|f(X)) &= \mathbb{E}_f\left[(Z_{f,p} - \mathbb{E}\left[Z_{f,p}|f(X)\right])^2|f(X)\right] \\
&= \mathbb{E}_f\left[\left(\int f(x)p(x)dx - \int \mathbb{E}_f\left[f(x)|f(X)\right]p(x)dx\right)^2\middle|f(X)\right] \\
&= \mathbb{E}_f\left[\left(\int \left(f(x) - \mathbb{E}_f\left[f(x)|f(X)\right]\right)p(x)dx\right)^2\middle|f(X)\right] \\
&= \int\int\int\left\{\Big(f(x) - \mathbb{E}_f\left[f(x)|f(X)\right]\Big)\Big(f(x') - \mathbb{E}_f\left[f(x')|f(X)\right]\Big)\right\} \\
&\quad p(x)p(x')p(f|f(X))dxdx'df \\
&= \int\int Cov_f(f(x), f(x'))p(x)p(x')dxdx' \quad \text{by Fubini} \\
&= \int\int (k(x,x') - q(x)K^{-1}q(x'))p(x)p(x')dxdx' \\
&= \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] - \int q(x)p(x)dx\, K^{-1}\int q(x')p(x')dx' \\
&= \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] - \mathbb{E}_{x\sim p}\left[q(x)\right]^T K^{-1}\mathbb{E}_{x'\sim p}\left[q(x')\right]
\end{aligned}$$

$$Var(Z_{f,p}|f(X)) = \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] - z^T K^{-1} z \tag{8}$$

Moreover, it can be shown using the representer theorem (Huszár and Duvenaud, 2012) that:

Proposition 3.1.

$$Var(Z_{f,p}|f(X)) = MMD^2(p, q_{BQ}) = \|\mu_p - \mu_{q_{BQ}}\|^2$$

where $q_{BQ} = \sum_{n=1}^{N} w_{BQ}^{(n)}\delta_{x_n}$ *with* $w_{BQ}^{(n)} = \sum_m z_m^T K_{mn}^{-1}$

Proof. If we use extend the support of the supremum to measures as in (Kanagawa et al., 2018) for the definition of MMD, with μ_p the kernel mean and $z = (\mu_p(x_i))_{i=1,\dots N}$,

$$\begin{aligned}
\|\mu_p - \mu_{q_{BQ}}\|^2 &= \|\mu_p\|^2 - 2\langle\mu_p, \mu_{q_{BQ}}\rangle + \|\mu_{q_{BQ}}\|^2 \\
&= \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] - 2w_{BQ}^\top z + w_{BQ}^\top K^{-1} w_{BQ} \quad \text{by the reproducing property} \\
&= \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] - z^\top K^{-1} z \quad \text{by the definition of the weights}
\end{aligned}$$

□

We can therefore use the same sequential method as Herding to select new samples:

$$\begin{aligned}
x_{n+1} &\longleftarrow \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \epsilon_{BQ}^2\left(\{x_1,\dots,x_n,x\}\right) \\
&= \underset{x\in\mathcal{X}}{\operatorname{argmax}}\, z^T K^{-1} z
\end{aligned}$$

Furthemore, assuming proposition 3.1, one can derive (see appendix C):

$$\epsilon_{BQ}^2\left(\{x_1,\dots,x_N\}\right) \le \epsilon_{herding}^2\left(\{x_1,\dots,x_N\}\right)$$

It means that ϵ_{BQ}^2 is also a minimax bound on estimation error with respect to an RKHS.

4 Implementation

```
Algorithm 1: Herding Sequential Sampling
Result: samples
samples = {}
$l = 0$
while $l \leq N$ do
    $newSample = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, 2\mathbb{E}_{x' \sim p}\left[k(x, x')\right] - \frac{1}{l+1} \sum_{m=1}^{l} k(x, x_m)$
    $samples = samples \cup \{newSample\}$
    $l \leftarrow l + 1$
end
```

In practice, we find the maximum over a lot of random points over $\mathcal{X}$ for new sample at each step, let say $n_{queries}$ points with respect to the uniform law over the hypercube defined by the extremal values of the function f. There are l additions at step l and we take the maximum over $n_{queries}$ scores. So Herding sequential sampling is in $\mathcal{O}\left(N^2\right)$, even though in for our experiments $n_{queries}$ is bigger than N

```
Algorithm 2: Sequential Bayesian Quadrature
Result: samples
samples = {}
$l = 0$
while $l \leq N$ do
    $newSample = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, z^T K^{-1} z = \sum_{m,n=1}^{l+1} \mathbb{E}_{x \sim p}\left[k(x_n, x)\right] \mathbb{E}_{x \sim p}\left[k(x_m, x)\right] K_{nm}^{-1}$
    $samples = samples \cup \{newSample\}$
    $l \leftarrow l + 1$
end
```

This algorithm is similar to the previous one, although it compute an inverse of Gram matrix of size $l \times l$ at each step in addition to z. This leads to a complexity of $\mathcal{O}\left(N^3\right)$ as we use the inverse of partitioned matrix lemma. Although this cost may seem prohibitive, the approach is justified in some important applications where this cubic computational cost is negligible compared to the cost of one evaluation of the integrand.

5 Results

5.1 Setting

We used a gaussian kernel $k(x,y) = \frac{2}{\pi} e^{-\|x-y\|^2}$, and f following a gaussian mixture of 20 components with the same covariance matrix as k. This ensures an important assumption: f is in the RKHS $\mathcal{H}$ associated to k.

$$f(x) = \sum_{i=1}^{20} \beta_i k(x, c_i) \quad \|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{20} \sum_{j=1}^{20} \beta_i \beta_j k(c_i, c_j) = 1$$

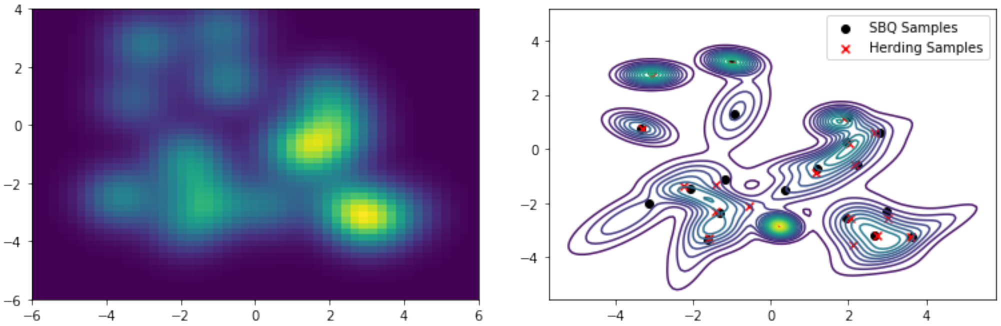
Starting from $\tilde{f} = \sum_{i=1}^{20} \tilde{\beta}_i k(c_i, .)$, where β_i are drawn at random, we define:

$$f = \sum_{i=1}^{20} \frac{\beta_i}{\sqrt{\sum_{i=1}^{20} \sum_{j=1}^{20} \beta_i \beta_j k(c_i, c_j)}} k(x, c_i)$$

Then we have a closed form for the integral as shown in eq. (4) with $R_j = \Gamma + \Sigma_j$:

$$Z_{f,p} = \sum_{i=1}^{20} \sum_{j=1}^{K} \beta_i \alpha_j \frac{1}{(2\pi)^{D/2} \left|R_j\right|^{1/2}} \exp((c_i - \mu_j)^\top R_j^{-1} (c_i - \mu_j))$$

In the following we want to compare the convergence rate. The MMD gives an upper bound of it for those functions f. Let denote N the evaluation budget. Indeed, since both Kernel Herding and SBQ do not need values of f at the query samples, the budget for those methods is only N. We do not consider kernel evaluations and evaluations with respect of the distribution in the budget, as they are fixed.



(a) Density of Gaussian mixture

(b) 20 first samples of SBQ and Herding

5.2 MMD as an upper bound of mean absolute error

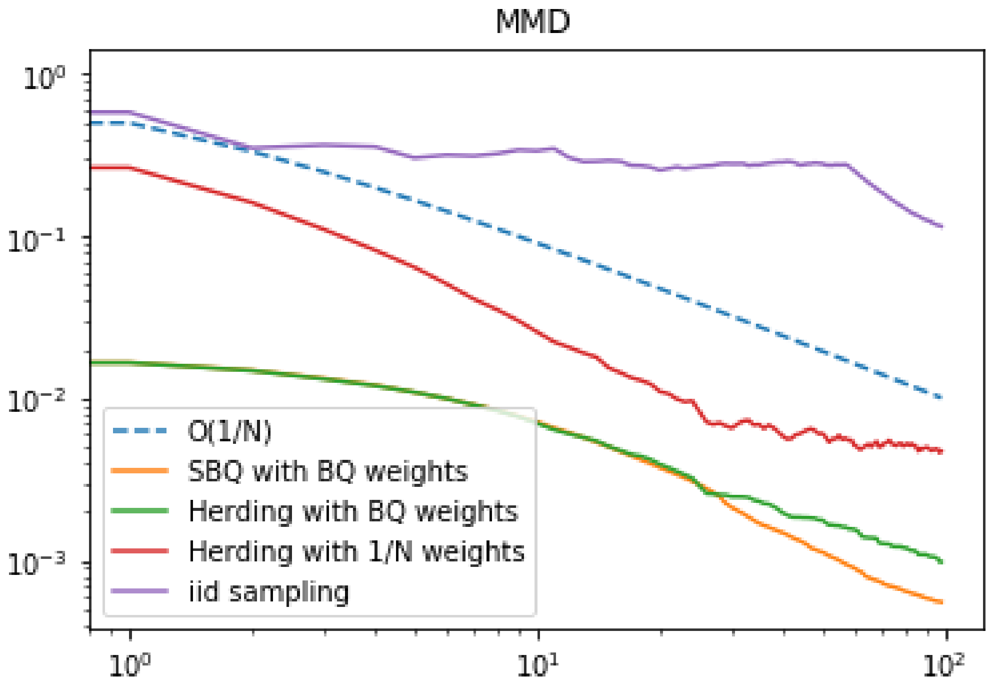
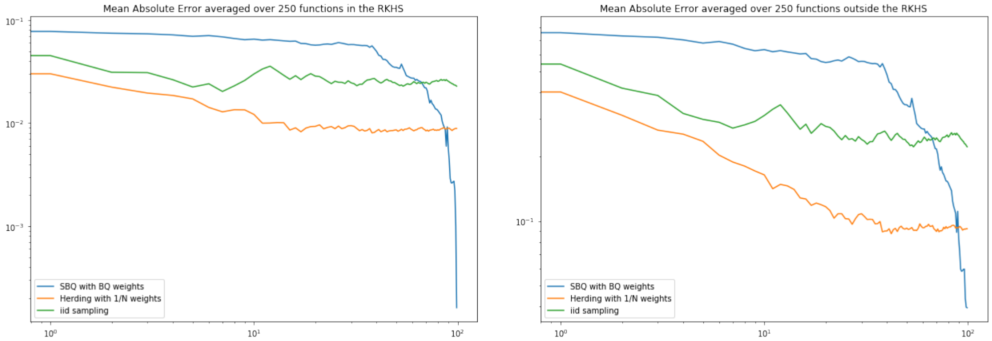


Figure 2: MMD



(a) Mean Absolute Error for 250 random functions within the RKHS

(b) Mean Absolute Error for 250 random functions outside of the RKHS

We also try for functions outside the RKHS $\mathcal{H}$. We use the following lemma:

Lemma 5.1. *Let* Λ *such that* $\Lambda - \frac{\Gamma}{2} \notin \mathcal{S}_D^+$, $q(x) = \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \exp((x-c_i)^\top \Lambda_i^{-1}(x-c))$ *is not in* $\mathcal{H}$.

Proof. Let remind a spectral characterization of $\mathcal{H}$ with reproducing kernel the Gaussian kernel k (Kanagawa et al., 2018) using Bochner's theorem.:

$$\mathcal{H} = \left\{ f \ : \int \frac{|\hat{f}(\omega)|^2}{\hat{\phi}(\omega)} d\omega < +\infty \right\} \quad \text{where } \phi(u) = \frac{1}{(2\pi)^{D/2} \, |\Gamma|^{1/2}} \exp\left(-\frac{1}{2} u^\top \Gamma^{-1} u \right) \tag{9}$$

Using the form of the characteristic function of a Gaussian: $\hat{q}(\omega) = \exp\left(-ic^\top \omega - \frac{1}{2}\omega^\top \Lambda \omega \right)$. Then,

$$\frac{|\hat{q}(\omega)|^2}{\hat{\phi}(\omega)} = \exp\left(-\omega^\top (\Lambda - \frac{\Gamma}{2})\omega \right) \tag{10}$$

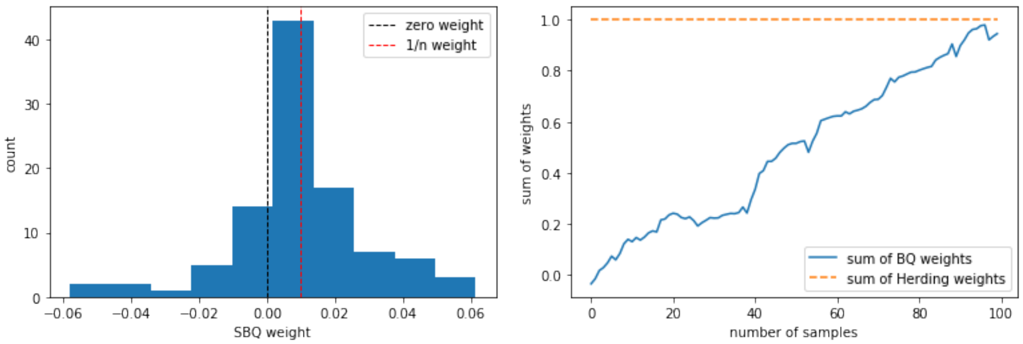
Since $\Lambda - \frac{\Gamma}{2} \notin \mathcal{S}_D^+$, the quadratic form $\omega \longmapsto -\omega^\top (\Lambda - \frac{\Gamma}{2})\omega$ is not upper bounded and continuous, hence the integral of eq. (10) is not finite. Thus with eq. (9), $\underline{q \notin \mathcal{H}}$. □

We generate random mixture of 20 Gaussians whose covariances are different from the one of k. With $\Gamma = \sigma^2 I_D$, it is likely to get the conditions of the lemma lemma 5.1 with a broad distribution of their eigenvalues which can be lower than $\sigma^2/2$.

$$f(x) = \sum_{i=1}^{20} \beta_i q_i(x) \quad q_i(x) = \frac{1}{(2\pi)^{D/2} \, |\Lambda_i|^{1/2}} \exp((x-c_i)^\top \Lambda_i^{-1}(x-c_i))$$

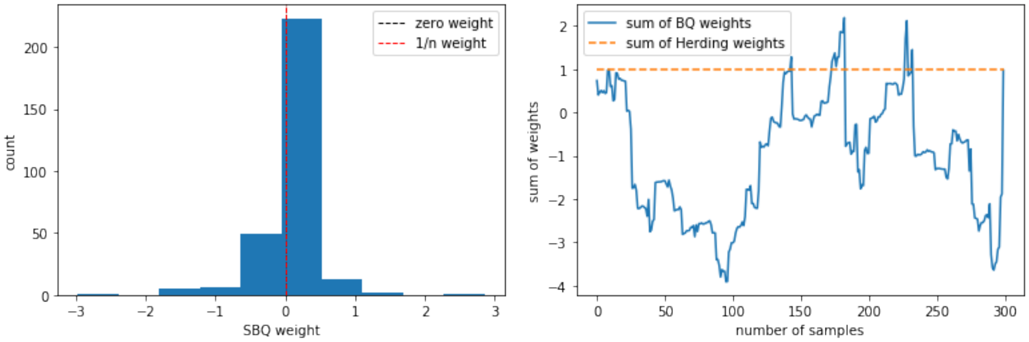
5.3 Discussion over SBQ weights

We plotted the distribution of the BQ weights as well as their sum. We observed that the sum is not always equal to 1 and that some weights are even negative. For 100 samples, the sum of the first weights is very close to 0, and seems to tend towards 1. When we plotted it with 300 samples, we did not really observe the same behaviour as described in the paper. Indeed, the first weights are closer to 1 than before, and the cumulative sum fluctuates much more, because of some weights which are far from the majority, and some which are closer to 0. The proposition 5 in (Karvonen et al., 2018) does give a minimum for the number of positive weights following SBQ rule which is increasing with the number of samples. The trend can be noticed in our graphs. (Karvonen et al., 2018) also give an upper bound for the magnitude weights which is increasing with the number of samples. And we observe indeed that with 300 samples, the magnitude is bigger in the sense that there are weights which are equals to -3 or 3 when for 100 samples, there are all between -0.1 and 0.1.



(a) Distribution of weights for 100 samples

(b) Sum of the weights for 100 samples



(a) Distribution of weights for 300 samples

(b) Sum of the weights for 300 samples

References

F. Huszár and D. Duvenaud. Optimally-weighted herding is bayesian quadrature, 2012.

T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels, 2003.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018.

T. Karvonen, M. Kanagawa, and S. Särkkä. On the positivity and magnitudes of bayesian quadrature weights, 2018.

C. E. Rasmussen and Z. Ghahramani. Bayesian monte carlo, 2002.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures, 2009.

A Computation of the expectation of the Gaussian kernel

Let $\underline{Q = (\Gamma^{-1} + \Sigma^{-1})^{-1}}$, then $\underline{m(x) = Q(\Gamma^{-1}x + \Sigma^{-1}\mu)}$

$$\begin{aligned}
\mathbb{E}_{y\sim p}\left[k(x,y)\right] &= \int \frac{\exp\left(-\frac{1}{2}(x-y)^\top\Gamma^{-1}(x-y)\right)\exp\left(-\frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu)\right)}{(2\pi)^D\left|\Gamma\right|^{-1/2}\left|\Sigma\right|^{-1/2}}dy \\
&= \frac{e^{-\frac{1}{2}\left(x^\top\Gamma^{-1}x+\mu^\top\Sigma^{-1}\mu\right)}}{(2\pi)^D\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}\int e^{-\frac{1}{2}\left(+y^\top(\Gamma^{-1}+\Sigma^{-1})y-2(x^\top\Gamma^{-1}+\mu^\top\Sigma^{-1})y\right)}dy \\
&= \frac{e^{-\frac{1}{2}\left(x^\top\Gamma^{-1}x+\mu^\top\Sigma^{-1}\mu\right)}\left|\Gamma^{-1}+\Sigma^{-1}\right|^{1/2}}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}\int\frac{e^{-\frac{1}{2}\left(y^\top(\Gamma^{-1}+\Sigma^{-1})y-2(x^\top\Gamma^{-1}+\mu^\top\Sigma^{-1})y\right)}}{(2\pi)^{D/2}\left|\Gamma^{-1}+\Sigma^{-1}\right|^{1/2}}dy \\
\mathbb{E}_{y\sim p}\left[k(x,y)\right] &= \frac{\left|Q\right|^{1/2}\exp\left[-\frac{1}{2}\left(x^\top\Gamma^{-1}x+\mu^\top\Sigma^{-1}\mu-m(x)^\top Q^{-1}m(x)\right)\right]}{(2\pi)^{D/2}\left|\Gamma\right|^{1/2}\left|\Sigma\right|^{1/2}}
\end{aligned}$$

B Computation of the BQ estimator

With the notations of section 3.1

$$\begin{aligned}
\mathbb{E}_f\left[Z_{f,p}|f(X)\right] &= \mathbb{E}_f\left[\int f(x)p(x)dx|f(X)\right] = \int\int f(x)p(x)dxp(f|f(X))df \\
&= \int\int f(x)p(f(x)|f(X))dfp(x)dx = \int\mathbb{E}_{f\sim p(.|f(X)}\left[f(x)\right]p(x)dx \\
&= \int f(X)^TK^{-1}q(x)p(x)dx = f(X)^TK^{-1}\mathbb{E}_{x\sim p}\left[q(x)\right]
\end{aligned}$$

C ϵ_{BQ} as a minimax bound

With the notation of section 3.1.

Because of the fact that $\mathbb{E}_f\left[Z_{f,p}|f(X)\right]$ is a Bayes estimator, it has the minimal expected squared error amongst all estimators, so:

$$\begin{aligned}
\epsilon_{BQ}^2 &= Var(Z_{f,p}|f(X)) \\
&= MMD^2(p, q_{BQ}) \\
&= \sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}}\left|\int f_xp(x)dx-\sum_{n=1}^N f_{x_n}w_{BQ}^{(n)}\right|^2 \\
&= \inf_{\hat{Z}}\sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}}\left|Z_{f,p}-\hat{Z}(f_{x_1},...,f_{x_N})\right|^2 \\
&= \inf_{w\in\mathbb{R}^N}\sup_{\substack{f\in\mathcal{H}\\ \|f\|_{\mathcal{H}}=1}}\left|\int f_xp(x)dx-\sum_{n=1}^N f_{x_n}w_n\right|^2 \\
&\leq MMD^2\left(p,\frac{1}{N}\sum_{n=1}^N\delta_{x_n}\right) \\
&= \epsilon_{herding}^2
\end{aligned}$$