BAYESIAN STATISTICS PROJECT

FRANK-WOLFE BAYESIAN QUADRATURE: PROBABILISTIC INTEGRATION WITH THEORETICAL GUARANTEES

AUTHORS

LYN BOUSSENGUI, ANTOINE BROWAEYS
ACHRAF BZILI, NICOLAS CLOAREC

Notations

We start with some notations we will use along this report.

- For any function, $g(\cdot)$ denotes the function $g : x \mapsto g(x)$.
- In integrals, $\mathrm{d}x$ denote $\mathrm{d}\lambda(x)$, *i.e.* with respect to the Lebesgue measure.

Introduction

The goal of the article [1] is to compute efficiently the integrals of the form $\int_{\mathcal{X}} f(x)p(x)\mathrm{d}x$ where $\mathcal{X} \subseteq \mathbb{R}^d$ is a measurable space, $d \geq 1$ integer representing the dimension of the problem, p a probability density with respect to the Lebesgue measure on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ is a *test-function*.

We will use the common approximation

$$\int_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i) \tag{1}$$

but of course the real challenge lies in the choice of sequences $\{x_i\}$ and $\{w_i\}$:

- **Monte Carlo** : $w_i = \frac{1}{n}$ and x_i realization of multivariate random variable $X_i \overset{iid}{\sim} X$ where X has $p(\cdot)$ as probability distribution.
- **Kernel herding** :
- **Quasi-Monte Carlo** :
- **Frank-Wolfe Bayesian Quadrature** :

- $\{w_i\}$ appear naturally in the Bayesian Quadrature by taking the expectation of a posterior distribution (described in section 2),
- $\{x_i\}$ are selected by the Frank-Wolfe algorithm in order to minimize a posterior variance (described in section 3).

The main interest of the method developed in [1] is the super fast *exponential* convergence to the true value of the integral compared to the other methods mentioned above.

Through this report, we will detail every results from [1] with the goal to clarify and explain details that could have been omitted intentionally or not and which, in our view, make the Briol's and al. approach more natural, intuitive and easier to understand.

1 Background

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable space, μ a measure on $\mathcal{X}$ such that $p = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ where λ denotes the Lebesgue measure on $\mathcal{X}$, $\mathcal{H} \subset L^2(\mathcal{X}, \mathbb{R}; \mu)$ be an RKHS with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, Φ its canonical feature map associated. We denote respectively by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ the bigps product and norm induced on $\mathcal{H}$.

Recall that the following relations hold:

$$\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H} \tag{2}$$
$$\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \tag{3}$$
$$\forall (x, y) \in \mathcal{X}^2 \quad k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \tag{4}$$

Let's denote as [1]:

$$p[f] := \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x$$
$$\hat{p}[f] := \sum_{i=1}^{n} w_i f(x_i).$$

We will use the *maximum mean discrepancy* (MMD) as our main metric to measure the accuracy of the approxi-

mation $p[f] \approx \hat{p}[f]$ in the worst case scenario and which is defined as

$$\mathrm{MMD}\left(\{x_i, w_i\}_{i=1}^n\right) := \sup_{f \in \mathcal{H}:\|f\|_{\mathcal{H}}=1} |p[f] - \hat{p}[f]|.$$

Let's show (formula 3. in [1]) that MMD can be rewrite as

$$\mathrm{MMD}\left(\{x_i, w_i\}_{i=1}^n\right) = \left\|\mu_p - \mu_{\hat{p}}\right\|_{\mathcal{H}} \tag{5}$$

where $\mu_p(\cdot) = p[\Phi(\cdot)]$ and $\mu_{\hat{p}}(\cdot) = \hat{p}[\Phi(\cdot)]$.

- **For all f in $\mathcal{H}$, we have $p[f] = \langle f, \mu_p \rangle_{\mathcal{H}}$.** By using the dirac delta function, the continuity of the inner product and viewing integral as a limit of a sum, we get

$$\begin{aligned} p[f] &= \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \delta_x[f]\,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \langle f, \Phi(x)\rangle_{\mathcal{H}}\,\mathrm{d}\mu(x) \\ &= \left\langle f, \int_{\mathcal{X}} \Phi(x)\,\mathrm{d}\mu(x)\right\rangle_{\mathcal{H}} \\ &= \langle f, \mu_p\rangle_{\mathcal{H}} \end{aligned}$$

- **For all f in $\mathcal{H}$, we have $\hat{p}[f] = \langle f, \mu_{\hat{p}}\rangle_{\mathcal{H}}$.**

$$\begin{aligned} \hat{p}[f] &= \sum_{i=1}^n w_i f(x_i) \\ &= \sum_{i=1}^n w_i \delta_{x_i}[f] \\ &= \sum_{i=1}^n w_i \langle f, \Phi(x_i)\rangle_{\mathcal{H}} \\ &= \left\langle f, \sum_{i=1}^n w_i \Phi(x_i)\right\rangle_{\mathcal{H}} \\ &= \langle f, \mu_{\hat{p}}\rangle_{\mathcal{H}} \end{aligned}$$

- By using previous results and the Cauchy-schwartz inequality, we get :

$$\begin{aligned} \mathrm{MMD}\left(\{x_i, w_i\}_{i=1}^n\right) &= \sup_{f \in \mathcal{H}:\|f\|_{\mathcal{H}}=1} \left|\langle f, \mu_p - \mu_{\hat{p}}\rangle_{\mathcal{H}}\right| \\ &\leq \sup_{f \in \mathcal{H}:\|f\|_{\mathcal{H}}=1} \|f\|_{\mathcal{H}} \left\|\mu_p - \mu_{\hat{p}}\right\|_{\mathcal{H}} \\ &= \left\|\mu_p - \mu_{\hat{p}}\right\|_{\mathcal{H}} \end{aligned}$$

with equality if and only if f and $\mu_p - \mu_{\hat{p}}$ are linearly dependent. We deduce the desired result by taking $f = \dfrac{1}{\left\|\mu_p - \mu_{\hat{p}}\right\|_{\mathcal{H}}}\left(\mu_p - \mu_{\hat{p}}\right)$.

2 Bayesian Quadrature

Let's place a functional prior on the integrand f and denote by $(\Omega, \mathcal{F}, \mathbb{P})$ its probability space associated. We will assume that f to be a *centered* **gaussian process** with the kernel k as its covariance function, *i.e.*

$$\begin{aligned} &\forall x \in \mathcal{H}, \quad \mathbb{E}\, f(x) = 0 \\ &\forall x, y \in \mathcal{H}, \quad \mathrm{Cov}\left[f(x), f(y)\right] = k(x, y) \end{aligned}$$

A useful property is that $p[f]$ is a gaussian variable and then completely defined by its second-order statistics:

$$\mathbb{E}\, p[f] = 0 \tag{6}$$

$$\mathbb{V}\, p[f] = \int_{\mathcal{X}^2} k(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \tag{7}$$

By switching integrals using Fubini's theorem, we get

$$\begin{aligned} \mathbb{E}\, p[f] &= \int_{\Omega} p[f](w)\,\mathrm{d}\mathbb{P}(w) \\ &= \int_{\Omega}\int_{\mathcal{X}} f(x, w)\,\mathrm{d}\mu(x)\,\mathrm{d}\mathbb{P}(w) \\ &= \int_{\mathcal{X}} \underbrace{\int_{\Omega} f(x, w)\,\mathrm{d}\mathbb{P}(w)}_{\mathbb{E} f(x) = 0}\,\mathrm{d}\mu(x) = 0 \end{aligned}$$

$$
\begin{aligned}
\mathbb{V}\,p[f] = \mathbb{E}\,p[f]^2 &= \int_\Omega p[f](w)^2 \,\mathrm{d}\mathbb{P}(w) \\
&= \int_\Omega \left(\int_{\mathcal{X}} f(x,w)\,\mathrm{d}\mu(x) \right)^2 \mathrm{d}\mathbb{P}(w) \\
&= \int_\Omega \int_{\mathcal{X}^2} f(x,w) f(y,w)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\,\mathrm{d}\mathbb{P}(w) \\
&= \int_{\mathcal{X}^2} \underbrace{\int_\Omega f(x,w) f(y,w)\,\mathrm{d}\mathbb{P}(w)}_{=\mathrm{Cov}[f(x),f(y)]=k(x,y)} \mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\
&= \int_{\mathcal{X}^2} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)
\end{aligned}
$$

Assume that samples $\{x_i\}$ and $\{f_i\} := \{f(x_i)\}$ are given for $i = 1$ to n and denote by $K := \big(k(x_i,x_j)\big)_{1\leq i,j\leq n}$. A natural question arises: how to update the weights $\{w_i\}_{i=1}^n$?

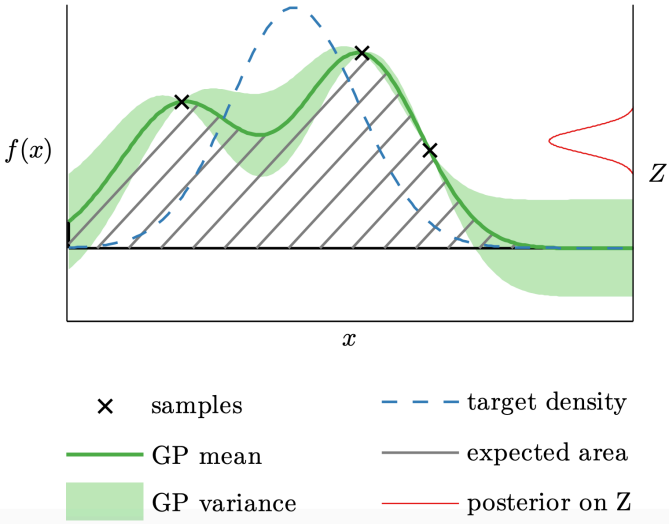


Figure 1: **An illustration of Bayesian Quadrature.** *Source:* [3]

First of all, let's determine the conditional distribution $p[f]|\,\mathbf{f}$ where $\mathbf{f} = (f_1,\dots,f_n)^T$. Since both $p[f]$ and $\mathbf{f}$ are gaussian, we can use the conditional gaussian rule:

By denoting, $y_1 := p[f]$, $y_2 = \mathbf{f}$, $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{1,1}, \Sigma_{1,2} \\ \Sigma_{2,1}, \Sigma_{2,2} \end{pmatrix}$ its covariance matrix by blocks, we have :

$$p[f]|\,\mathbf{f} \sim \mathcal{N}(\mu, \widetilde{\Sigma})$$

where

$$
\begin{cases} \mu &= \Sigma_{12}\,\Sigma_{22}^{-1}\,\mathbf{f} \\ \widetilde{\Sigma} &= \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}. \end{cases}
$$

Let's determine what μ and $\widetilde{\Sigma}$ look like in our context.

$$
\begin{aligned}
\Sigma_{22} &= \big(\mathrm{Cov}\left[f_i, f_j\right]\big)_{1\leq i,j\leq n} \\
&= \big(\mathrm{Cov}\left[f(x_i), f(x_j)\right]\big)_{1\leq i,j\leq n} \\
&= \big(k(x_i,x_j)\big)_{1\leq i,j\leq n} \\
&= K \\
\mu &= \begin{pmatrix} \mathrm{Cov}\left[p[f], f_1\right] \\ \dots \\ \mathrm{Cov}\left[p[f], f_n\right] \end{pmatrix} K^{-1}\mathbf{f}
\end{aligned}
$$

Let's rewrite the vector from the left:

$$
\begin{aligned}
\mathrm{Cov}\left[p[f], f_i\right] &= \int_\Omega p[f](w) f(x_i,w)\,\mathrm{d}\mathbb{P}(w) \\
&= \int_\Omega \int_{\mathcal{X}} f(x,w)\,\mathrm{d}\mu(x) f(x_i,w)\,\mathrm{d}\mathbb{P}(w) \\
&= \int_{\mathcal{X}} \underbrace{\int_\Omega f(x,w) f(x_i,w)\,\mathrm{d}\mathbb{P}(w)}_{=\mathrm{Cov}[f(x),f(x_i)]=k(x,x_i)} \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} k(x,x_i)\,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \Phi(x_i)(x)\,\mathrm{d}\mu(x) \\
&= p[\Phi(x_i)] \\
&= \mu_p(x_i)
\end{aligned}
$$

By denoting $z := (z_i)_{i=1}^n = \big(\mu_p(x_i)\big)_{i=1}^n \in \mathbb{R}^n$, we get the desired formula[1] for the expectation:

$$\mu = z^T K^{-1}\mathbf{f} \tag{8}$$

The variance is then straightforward :

[1] formula 4 in [1]

$$\widetilde{\Sigma} = \mathbb{V}\,p[f] - z^T K^{-1} z$$
$$= \int_{\mathcal{X}^2} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - z^T K^{-1} z$$

We need to simplify the integral to obtain the desired formula for the variance:

$$\begin{aligned}\int_{\mathcal{X}^2} k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) &= \int_{\mathcal{X}}\int_{\mathcal{X}} \Phi(x)(y)\,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} p[\Phi(x)]\,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} \mu_p(x)\,\mathrm{d}\mu(x) \\ &= p[\mu_p]\end{aligned}$$

which leads us to the desired result:

$$\widetilde{\Sigma} = p[\mu_p] - z^T K^{-1} z$$

Both the conditional expectation and variance are essentials in the FWBQ[2] algorithm:

- $\mu = z^T K^{-1}\mathbf{f}$ which can be written as $\mu = \sum_{i=1}^{n} w_i^{BQ} f_i$ where $w^{BQ} := \left(K^{-1}\right)^T z$. μ appears to be the **most natural estimation of our integral** $p[f]$. It also gives us the **new weights to update**.
- $\widetilde{\Sigma} = p[\mu_p] - z^T K^{-1} z$ which is also equal to $\mathrm{MMD}\left(\{x_i, w_i\}_{i=1}^n\right)^2$ according to [3] and can be interpreted as our **uncertainty**. Therefore we need to **choose** $\{x_i\}$ **which minimize this quantity**. This will be achieved by using the Frank Wolfe algorithm described in the next section.

3 Frank-Wolfe algorithm

Let's J be a convex differentiable real-valued function on a domain $\mathcal{G} \subset \mathcal{H}$ which is supposed to be a compact and convex.

The Frank-Wolfe algorithm [4] proposes a method to solve the following optimization problem :

$$\begin{aligned}&\text{Minimize } \{J(\mathbf{x})\} \\ &\text{subject to } \mathbf{x} \in \mathcal{G}\end{aligned} \tag{9}$$

Let's describe here both the Frank-Wolfe algorithm and its variant with the Line Search :

- **Initialization** $g_1 = \overline{g}_1 \in \mathcal{G}$ and step-size sequence $\{\rho_i\}_{i=1}^n$ (not required in the FWLS algorithm).
- **Iterations** For $i = 2$ to n :
- **Search direction** we replace in (9) J by its first-order Taylor expansion around g_k to solve the following subproblem

$$\begin{aligned}&\text{minimize } \left\{J(g_{i-1}) + \nabla J(g_{i-1})^{\mathrm{T}}(g - g_{i-1})\right\} \\ &\text{subject to } g \in \mathcal{G}\end{aligned} \tag{10}$$

Let's denote as [1]

$$\overline{g}_i := \arg\min_{g\in\mathcal{G}} \left\langle g, \nabla J(g_{i-1})\right\rangle_{\mathcal{H}} \tag{11}$$

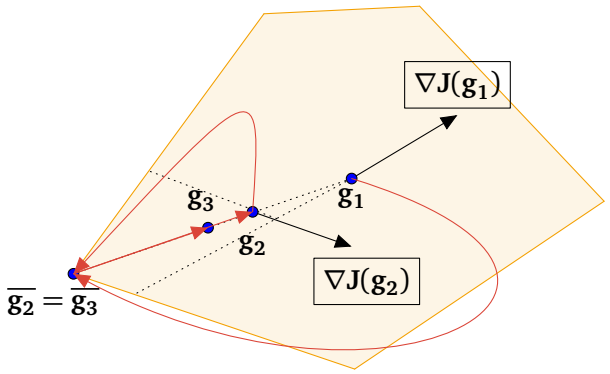
where we have removed in (10) terms independent of the optimization variable.

- **New iteration point** : we choose g_i as a *convex*-combination of g_{i-1} and $\overline{g}_i$.

$$g_i = (1-\rho_i)\,g_{i-1} + \rho_i \overline{g}_i \tag{12}$$

- **FW** ρ_i is determined by the sequence given at the initialization
- **FWLS**

$$\rho_i := \mathrm{argmin}_{\rho\in[0,1]} J\left((1-\rho)g_{i-1} + \rho\overline{g}_i\right)$$



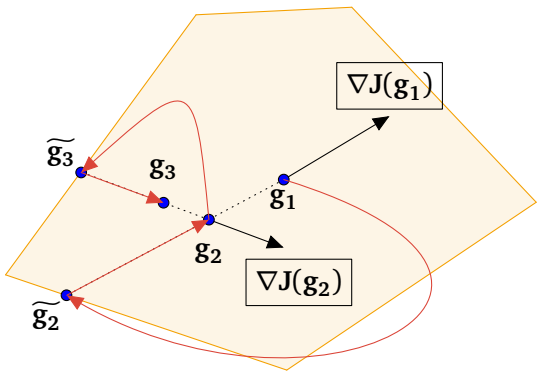
An illustration of the Frank Wolfe algorithm.

In equation (11), we don't have necessarily $\overline{g}_i \in \mathrm{span}\{\nabla g_{i-1}\}$. In fact, we can clearly see in the illustration above that the point

$$\begin{aligned}\widetilde{g}_i &:= \arg\min_{g\in\mathcal{G}\cap\mathrm{span}\{\nabla g_{i-1}\}} \left\langle g, \nabla J(g_{i-1})\right\rangle_{\mathcal{H}} \\ &= \arg\min_{\alpha\geq 0} \{-\alpha\nabla g_{i-1} \mid \alpha\nabla g_{i-1} \in \mathcal{G}\}\end{aligned}$$

would not be optimal as $\langle \overline{g}_i, \nabla g_{i-1}\rangle_{\mathcal{H}} < \langle \widetilde{g}_i, \nabla g_{i-1}\rangle_{\mathcal{H}}$.

[2] Frank-Wolfe Bayesian Quadrature



Comparison with a descent algorithm.

4 Consistency and Contraction

5 Experimental Results

5.1 Overview of the code

We have implemented the following algorithms : – FW and FWBQ (when Line-Search mode is disabled) – FWLS and FWLSBQ (when Line-Search mode is enabled). Two notebooks are available. The first notebook applies the different algorithms to a simulation study. The second notebook is an example of how the algorithms can be used to solve a problem in a specific area.

5.2 Simulation Study

We also used an exponentiated-quadratic (EQ) kernel $k(x, x') := \lambda^2 exp(-\frac{1}{2}\sigma^2\|x - x'\|_2^2)$. EQ kernel is a relevant choice when p is a mixture of gaussians. Moreover, the mean element μ_p has a closed-form expression. In order to replicate the simulation study of the paper, we took p as a mixture of 20 two-dimensional gaussians. Two dimensions for each gaussians allow us to plot selected points, the function f, the approximation of the mean element g_n and the mean element μ_p so that we can come up with a visual representation of the algorithms.

5.3 An example of application

Resources

[1] François-Xavier Briol et al. “Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees”. In: *Advances in Neural Information Processing Systems 28, 1162–1170, 2015* (June 8, 2015). arXiv: https://arxiv.org/pdf/1506.02681.pdf [stat.ML].

[2] David Duvenaud. “Bayesian Quadrature: Model-based Approximate Integration”. University Lecture. URL: https://www.cs.toronto.edu/~duvenaud/talks/intro_bq.pdf.

[3] Ferenc Huszár and David Duvenaud. “Optimally-Weighted Herding is Bayesian Quadrature”. In: (Apr. 7, 2012). arXiv: https://arxiv.org/pdf/1204.1664.pdf [stat.ML].

[4] Michael Patriksson. “The FrankWolfe algorithm”. University Lecture. URL: http://www.math.chalmers.se/Math/Grundutb/CTH/tma946/0203/fw_eng.pdf.