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 [3] 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 [3] 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 [3] 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 [3]:

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

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 [3]) 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 $\{w_i\}_{i=1}^n$?

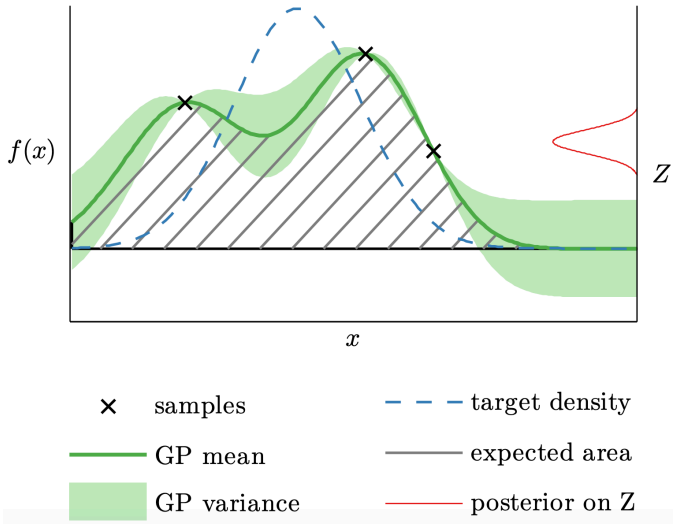


Figure 1: **An illustration of Bayesian Quadrature.** *Source:* [6]

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 [3]

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

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 [6] 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 [7] 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 [3]

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

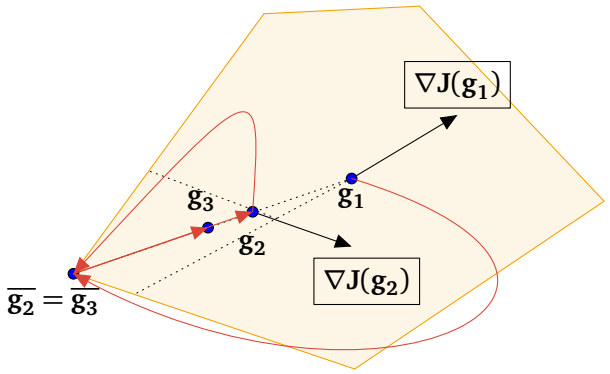
where we have removed in (10) terms independent of the optimization variable.

- **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.

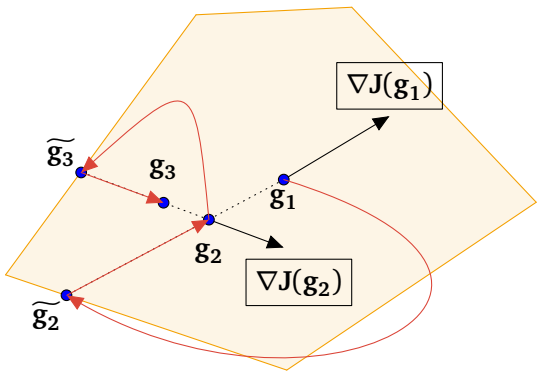
In equation (11), we don't have necessarily $\overline{g}_i \in \mathrm{span}\{\nabla J(g_{i-1})\}$. In fact, we can clearly see in the illustration above that the point

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

would not be optimal as

$$\langle \overline{g}_i, \nabla J(g_{i-1})\rangle_{\mathcal{H}} < \langle \widetilde{g}_i, \nabla J(g_{i-1})\rangle_{\mathcal{H}}.$$

[2] Frank-Wolfe Bayesian Quadrature



Comparison with a descent algorithm.

Another important fact of the Frank Wolfe algorithm is that the output g_n can easily be expressed as a linear combination of atoms $\overline{g_i}$.

In fact we have the following formula, which corresponds to the equation (7) in [3]:

For all $n \geq 2$:

$$g_n = \sum_{k=1}^{n} \underbrace{\rho_k \left\{ \prod_{k<i\leq n} (1-\rho_i) \right\}}_{:=w_k^{FW}} \cdot \overline{g_k} = \sum_{k=1}^{n} w_k^{FW} \cdot \overline{g_k} \tag{13}$$

with $\rho_1 = 1$.

The formula in [3] is slightly different :

$$g_n = \sum_{i=1}^{n} \rho_{i-1} \left\{ \prod_{j=i+1}^{n} (1-\rho_{j-1}) \right\} \cdot \overline{g}_i$$

with $\rho_0 = 1$ but clearly doesn't work. In fact according to the algorithm, by taking $n = 2$ we should have :

$$g_2 = (1-\rho_2) \cdot \overline{g_1} + \rho_2 \cdot \overline{g_2}$$

while their formula gives

$$\begin{aligned} g_2 &= \sum_{i=1}^{2} \rho_{i-1} \left\{ \prod_{j=i+1}^{2} (1-\rho_{j-1}) \right\} \overline{g}_i \\ &= (1-\rho_1) \cdot \overline{g_1} + \rho_1 \cdot \overline{g_2}. \end{aligned}$$

Let's prove (13) by induction.

• **Base case** : $n = 2$. Our formula gives :

$$\begin{aligned} g_2 &= \sum_{k=1}^{2} \rho_k \left\{ \prod_{k<i\leq 2} (1-\rho_i) \right\} \cdot \overline{g_k} \\ &= (1-\rho_2) \cdot \overline{g_1} + \rho_2 \cdot \overline{g_2} \end{aligned}$$

as we have set $\rho_1 = 1$, as expected.

• **Step case** : Assume the formula holds for $n-1$ and let's prove it for n. Using the recurrence relation of (12) and our assumption, we have :

$$\begin{aligned} g_n &= (1-\rho_n) \cdot g_{n-1} + \rho_n \cdot \overline{g}_n \\ &= (1-\rho_n) \cdot \sum_{k=1}^{n} \rho_k \left\{ \prod_{k<i\leq n-1} (1-\rho_i) \right\} \cdot \overline{g_k} + \rho_n \cdot \overline{g}_n \\ &= \sum_{k=1}^{n} \rho_k \left\{ \prod_{k<i\leq n} (1-\rho_i) \right\} \cdot \overline{g_k} + \rho_n \cdot \overline{g}_n \\ &= \sum_{k=1}^{n} \rho_k \left\{ \prod_{k<i\leq n} (1-\rho_i) \right\} \cdot \overline{g_k} \end{aligned}$$

which ends the proof.

It is common to take harmonic coefficients in the standard FW algorithm[3] as a choice for $\{\rho_i\} := \left\{\frac{1}{i}\right\}_{i=1}^{n}$, which gives us uniform weights as showed bellow :

$$w_k^{FW} = \frac{1}{k} \left\{ \prod_{k<i\leq n} \left(1 - \frac{1}{i}\right) \right\} = \frac{1}{k} \cdot \frac{k}{n} = \frac{1}{n}.$$

Let's set here the framework of our optimization problem.

We want to approximate the mean element μ_p with a linear combination of elements of the form :

$$\Phi(x) = k(\cdot\,,\, x)\,.$$

In fact, let's say we have:

$$\mu_p \approx \hat{\mu_p} := \sum_{k=1}^{n} w_k \Phi(x_k).$$

Using the reproducing property, we will then have :

$$\hat{p}[f] = \langle f, \mu_p \rangle_{\mathcal{H}} = \sum_{k=1}^{n} w_k f(x_k),$$

which is a quadrature rule.
In order to do this, let's assume that we optimize the following convex function :

$$J(g) := \frac{1}{2} \left\| g - \mu_p \right\|_{\mathcal{H}}^2 \tag{14}$$

on the domain $\mathcal{G} \subseteq \mathcal{H}$ where $\mathcal{G}$ refers to the closure of the convex hull of $\Phi(\mathcal{X})$, which is assumed to be uniformly bounded, *i.e.*

$$\exists R > 0 : \forall x \in \mathcal{X}, \|\Phi(x)\|_{\mathcal{H}} \leq R$$

But a question rises : Why atoms g_i should be of the form $\Phi(x_i)$?

In fact, let's g be a point in the domain. Because of its definition, g can simply be expressed as a convex combination of elements in the feature space $\Phi(\mathcal{X})$, *i.e.* $g = \sum_k \alpha_k \Phi(x_k)$ with $\sum_k \alpha_k = 1$ and thus we have :

[3] Let's remember that in the FWLS algorithm, $\{\rho_i\}$ and thus $\left\{w_k^{FWLS}\right\}$ are determind by the line search step.

$$\begin{aligned}\left\langle g, \nabla J(g_{i-1})\right\rangle_{\mathscr{H}} &= \sum_k \alpha_k \cdot \left\langle \Phi(x_k), \nabla J(g_{i-1})\right\rangle_{\mathscr{H}} \\ &\geq \left(\sum_k \alpha_k\right) \cdot \min_k \left\langle \Phi(x_k), \nabla J(g_{i-1})\right\rangle_{\mathscr{H}} \\ &= \left\langle \Phi(x_{k_0}), \nabla J(g_{i-1})\right\rangle_{\mathscr{H}}\end{aligned}$$

for k_0 such that $k_0 \in \arg\min \left\langle \Phi(x_k), \nabla J(g_{i-1})\right\rangle_{\mathscr{H}}$.

We now understand how can the Frank Wolfe algorithm be useful here : because the minimization of (11) can be restricted at extreme points of the domain, selecting atoms of the form $\Phi(x_k)$ allow us to select interesting points $\{x_k\}$ which are needed for the bayesian quadrature which are solution of the following optimization problem :

$$x_k \in \arg\min_{x \in \mathscr{X}} \left\langle \Phi(x), g_{k-1} - \mu_p\right\rangle_{\mathscr{H}} \tag{15}$$

where $g \in \mathscr{G}$ in the left term of the inner product has been replaced by $\Phi(x) \in \mathscr{G}$ with $x \in \mathscr{X}$.

Therefore, if we denote by $\{w_l\}_{l=1}^{i-1}$ the coefficients in front of $\{\Phi(x_l)\}_{l=1}^{i-1}$ in g_{i-1} and using the reproducing kernel property, we have :

$$\begin{aligned}\left\langle \Phi(x), g_{i-1} - \mu_p\right\rangle_{\mathscr{H}} &= \left\langle \Phi(x), \sum_{l=1}^{i-1} w_l^{(i-1)} \Phi(x_l) - \mu_p \right\rangle_{\mathscr{H}} \\ &= \sum_{l=1}^{i-1} w_l^{(i-1)} k(x, x_l) - \mu_p(x)\end{aligned}$$

which gives an explicit formula for the optimization problem[4] with known quantities. In fact, in the simulations that we have been able to reproduce successfully, the choice of a gaussian kernel k allow us to compute $\mu_p(\cdot)$ easily. See [3] *Appendix C* for more details.

4 The Frank-Wolfe Bayesian Quadrature algorithm

We describe here the final algorithm which is a combination of the two previous ones. At each iteration `i`:

- **Selecting a new x_i** through the Frank Wolfe algorithm with g_{i-1} depending on the BQ weights of the iteration `i-1` and FW design points of all previous iteration `j` with $1 \leq j < i$.

- **Selecting new weights** $\left\{w_k^{BQ}\right\}_{k=1}^{i}$ with the bayesian quadrature.

It's important to note that weights at each iteration `i` are used only for the next iteration `i+1` while the design points x_i are used for all iteration `j` with $i < j \leq n$.

5 Consistency

We will establish here the main result of the article [3] and referred as Theorem 1:

The posterior mean $\hat{p}_{\text{FWBQ}}[f]$ converges to the true integral $p[f]$ at the following rates:

$$\begin{aligned}\left|p[f] - \hat{p}_{\text{FWBQ}}[f]\right| &\leq MMD\left(\{x_i, w_i\}_{i=1}^n\right) \\ &\leq \begin{cases} \frac{2D^2}{R} n^{-1} & \text{for FWBQ} \\ \sqrt{2} D \exp\left(-\frac{R^2}{2D^2} n\right) & \text{for FWLSBQ} \end{cases}\end{aligned} \tag{16}$$

where the FWBQ uses step-size $\rho_i = 1/(i+1)$, $D \in (0, \infty)$ is the diameter of the marginal polytope $\mathscr{G}$ and $R \in (0, \infty)$ gives the radius of the **largest**[a] ball of center μ_p included in $\mathscr{G}$.

[a] There was a typo here in [3] as the authors mentioned the radius of the smallest ball which is clearly zero. In fact, greater is the radius and better is the inequality, so of course we are more interested in large radius rather than smaller ones.

It was a bit disappointing to see that the authors of [3] did not prove the most interesting part of the theorem, which is the introduction of the diameter of the marginal polytope (D) and the radius of the largest ball of center μ_p (R). Instead, it came out of nowhere quoting [1], which is of course deeply unsatisfactory from a scientific standpoint as it gives no clear understanding about the intuition behind this result.

We were able to find more details about this result in [2] and [4]. Let's start assuming as [1] that μ_p is in the relative interior of $\mathscr{G}$:

$$\exists r > 0 \text{ such that } B(\mu_p, r) \subset \mathscr{G}. \tag{17}$$

- **Step 1** We start with a proposition[5] :

We denote by R the radius of the largest ball of center μ_p. Using the same notations as previously and the assumption above, we have at iteration `i`:

$$\left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i}\right\rangle_{\mathscr{H}} + R\left\|\mu_p - g_{i-1}\right\|_{\mathscr{H}} \leq 0 \tag{18}$$

By denoting $d = \mu_p - g_{i-1}$ and using the definition of R we have :

$$g = \mu_p + R \cdot \frac{d}{\|d\|_{\mathscr{H}}} \in \mathscr{G}.$$

[4] We remark that if the optimization problem on $g \in \mathscr{G}$ was a classic convex problem, this is not the case anymore when we perform the optimization on $x \in \mathscr{X}$.

[5] Proposition 3.1 in [2].

Moreover by using the definition of $\overline{g_i}$ we get:

$$\begin{aligned}\left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}} &= \left\langle -\nabla J(g_{i-1}), \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}} \\ &\leq \left\langle -\nabla J(g_{i-1}), \mu_p - g \right\rangle_{\mathscr{H}} \\ &\leq -R \cdot \left\langle \mu_p - g_{i-1}, \frac{d}{\|d\|_{\mathscr{H}}} \right\rangle_{\mathscr{H}} \\ &\leq -R\,\|d\|_{\mathscr{H}}\end{aligned}$$

which is the desired result.

• **Step 2** There is an explicit formula for ρ_i at the i^{th} iteration.

If we denote by $\rho^\star$ the optimum in the line search at the i^{th} iteration, the following formula holds :

$$\rho^\star = \frac{\left\langle g_{i-1} - \mu_p, g_{i-1} - \overline{g_i} \right\rangle_{\mathscr{H}}}{\|g_{i-1} - \overline{g_i}\|^2_{\mathscr{H}}} \tag{19}$$

We will show here the proof from [2] instead of the one in [3] as it requires less computation and more intuition[a]. Let

$$f : \left| \begin{array}{ccc} [0;1] & \longrightarrow & \mathbb{R} \\ \rho & \longmapsto & J((1-\rho)\, g_{i-1} + \rho \overline{g_i}) \end{array} \right.$$

Because J is quadratic so is f and we can replace f by its quadratic approximation :

$$\begin{aligned} f(\rho) = J(g_{i-1}) &+ \rho \langle \overline{g_i} - g_{i-1}, \nabla J(g_{i-1}) \rangle \\ &+ \frac{1}{2}\rho^2 \|g_{i-1} - \overline{g_i}\|^2 \end{aligned}$$

Taking the derivative of this expression, gives :

$$\rho^\star = \frac{\left\langle g_{i-1} - \mu_p, g_{i-1} - \overline{g_i} \right\rangle_{\mathscr{H}}}{\|g_{i-1} - \overline{g_i}\|^2_{\mathscr{H}}}.$$

which is the desired result but we need to ensure that $\rho^\star$ is between 0 and 1.
By definition of $\overline{g_i}$:

$$\begin{aligned} \overline{g_i} &= \arg\min_{g \in \mathscr{G}} \langle g, \nabla J(g_{i-1}) \rangle_{\mathscr{H}} \\ &\arg\min_{g \in \mathscr{G}} \langle g - g_{i-1}, \nabla J(g_{i-1}) \rangle_{\mathscr{H}} \end{aligned}$$

Because $g_{i-1} \in \mathscr{G}$[b], we have

$$\begin{aligned} \left\langle \underbrace{g_{i-1} - \mu_p}_{=\nabla J(g_{i-1})}, g_{i-1} - \overline{g_i} \right\rangle_{\mathscr{H}} &\geq \langle \nabla J(g_{i-1}), g_{i-1} - g_{i-1} \rangle_{\mathscr{H}} \\ &\geq 0 \end{aligned}$$

so $\rho^\star \geq 0$. Using (18), we have

$$\left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}} \leq 0. \tag{20}$$

Thus

$$\begin{aligned} &\left\langle g_{i-1} - \mu_p, g_{i-1} - \overline{g_i} \right\rangle_{\mathscr{H}} \\ &= \left\langle \mu_p - g_{i-1}, (\mu_p - g_{i-1}) - (\mu_p - \overline{g_i}) \right\rangle_{\mathscr{H}} \\ &= \left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} - \left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}} \\ &\leq \left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} - \left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}} \\ &\quad + \left(\left\|\mu_p - \overline{g_i}\right\|^2_{\mathscr{H}} - \underbrace{\left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle_{\mathscr{H}}}_{\leq 0} \right) \\ &= \left\|(\mu_p - g_{i-1}) - (\mu_p - \overline{g_i})\right\|^2_{\mathscr{H}} \\ &= \|g_{i-1} - \overline{g_i}\|^2_{\mathscr{H}} \end{aligned}$$

which gives us $\rho^\star \leq 1$.

[a] the authors of [3] do not give a full proof of this result the most important (showing that $\rho^\star \in [0;1]$).
[b] Remember that g_{i-1} is a convex combination of elements of $\Phi(\mathscr{X})$.

• **Step 3** We show here the linear convergence rate of the algorithm :

$$\left\|\mu_p - g_i\right\|^2_{\mathscr{H}} \leq \left(1 - q^2\right) \cdot \left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} \tag{21}$$

where $q = \frac{R}{D} > 0$.

Using (12) and the fact that J is equal to its quadratic approximation, we have :

$$\begin{aligned} \left\|\mu_p - g_i\right\|^2_{\mathscr{H}} &= (\rho^\star)^2 \left\|\overline{g_i} - g_{i-1}\right\|^2_{\mathscr{H}} \\ &\quad + 2\rho^\star \left\langle \mu_p - g_{i-1}, g_{i-1} - \overline{g_i} \right\rangle_{\mathscr{H}} \\ &\quad + \left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} \end{aligned}$$

Substituting the value of $\rho^\star$ from step 2 yields to :

$$\begin{aligned} &\left\|\mu_p - g_i\right\|^2_{\mathscr{H}} \\ &= \frac{\left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} \left\|\mu_p - \overline{g_i}\right\|^2_{\mathscr{H}} - \left\langle \mu_p - g_i, \mu_p - \overline{g_i} \right\rangle^2_{\mathscr{H}}}{\left\|\overline{g_i} - g_{i-1}\right\|^2_{\mathscr{H}}} \end{aligned}$$

We will need the two following inequalities in this third step:

$$\left\|\mu_p - \overline{g_i}\right\|^2_{\mathscr{H}} \leq \left\|\overline{g_i} - g_{i-1}\right\|^2_{\mathscr{H}} \tag{22}$$

$$R^2 \left\|\mu_p - g_{i-1}\right\|^2_{\mathscr{H}} \leq \left\langle \mu_p - g_{i-1}, \mu_p - \overline{g_i} \right\rangle^2_{\mathscr{H}} \tag{23}$$

The first one can be showed introducing μ_p in the right term and using (20), while the second one comes from (18) and the fact that the square function is decreasing on $\mathbb{R}_-$.

We are now able to prove the result :

$$
\begin{aligned}
&\left\|\mu_p - g_i\right\|_{\mathcal{H}}^2 \\
&\quad = \frac{\left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2 \left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}}^2 - \left\langle \mu_p - g_i, \mu_p - \overline{g_i} \right\rangle_{\mathcal{H}}^2}{\left\|\overline{g_i} - g_{i-1}\right\|_{\mathcal{H}}^2} \\
&\quad \leq \frac{\left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2 \left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}}^2 - R^2 \left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2}{\left\|\overline{g_i} - g_{i-1}\right\|_{\mathcal{H}}^2} \\
&\quad \leq \frac{\left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2 \left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}}^2 - R^2 \left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2}{\left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}}^2} \\
&\quad \leq \left(1 - \frac{R^2}{\left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}}^2}\right) \cdot \left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2 \\
&\quad \leq \left(1 - \frac{R^2}{D^2}\right) \cdot \left\|\mu_p - g_{i-1}\right\|_{\mathcal{H}}^2
\end{aligned}
$$

where the last inequality comes from

$$\left\|\mu_p - \overline{g_i}\right\|_{\mathcal{H}} \leq \sup_{g_1, g_2 \in \mathcal{G}} \|g_1 - g_2\|_{\mathcal{H}} := D$$

6 Contraction

7 Experimental Results

7.1 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.

7.2 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.

7.3 An example of application

Resources

[1] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. "On the Equivalence between Herding and Conditional Gradient Algorithms". In: *ICML 2012 International Conference on Machine Learning, Edimburgh : Royaume-Uni (2012)* (Mar. 20, 2012). arXiv: http://arxiv.org/abs/1203.4523v2 [cs.LG].

[2] Amir Beck and Marc Teboulle. "A conditional gradient method with linear rate of convergence for solving convex linear systems". In: *Mathematical Methods of Operations Research* 59 (Jan. 2004), pp. 235–247. DOI: 10.1007/s001860300327.

[3] 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].

[4] Yutian Chen, Max Welling, and Alex Smola. "Super-Samples from Kernel Herding". In: (Mar. 15, 2012). arXiv: http://arxiv.org/abs/1203.3472v1 [cs.LG].

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

[6] 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].

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