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 $\int_{\mathscr{X}} f(x)p(x)\mathrm{d}x$ where $\mathscr{X} \subseteq \mathbb{R}^d$ 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 $\mathscr{X}$ and $f : \mathscr{X} \to \mathbb{R}$ is a *test-function*.

We will use the common approximation

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

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**:

In the **Frank-Wolfe Bayesian Quadrature**, we have

- ☞ $\{w_i\}$ which appear naturally in the Bayesian Quadrature by taking the expectation of a posterior distribution (described in section 2),
- ☞ $\{x_i\}$ 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 $\mathscr{X} \subseteq \mathbb{R}^d$ be a measurable space, μ a measure on $\mathscr{X}$ such that $p = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ where λ denotes the Lebesgue measure on $\mathscr{X}$, $\mathscr{H} \subset L^2(\mathscr{X}, \mathbb{R}; \mu)$ be an RKHS with a reproducing kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, Φ its canonical feature map associated. We denote respectively by $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ and $\|\cdot\|_{\mathscr{H}}$ the bigps product and norm induced on $\mathscr{H}$.

Recall that the following relations hold:

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

Let's denote as [1]:

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

We will use the *maximum mean discrepancy* (MMD) as our main metric to measure the accuracy of the approximation $p[f] \approx \hat{p}[f]$ 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 $f = \frac{1}{\left\|\mu_p - \mu_{\hat{p}}\right\|_{\mathcal{H}}}\left(\mu_p - \mu_{\hat{p}}\right)$.

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

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

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

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

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 :

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

where

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

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

$$\begin{aligned}
\Sigma_{22} &= \left(\mathrm{Cov}\left[f_i, f_j \right] \right)_{1 \le i,j \le n} \\
&= \left(\mathrm{Cov}\left[f(x_i), f(x_j) \right] \right)_{1 \le i,j \le n} \\
&= \left(k(x_i, x_j) \right)_{1 \le i,j \le n} \\
&= K
\end{aligned}$$

$$\mu = \begin{pmatrix} \mathrm{Cov}\left[p[f], f_1 \right] \\ \dots \\ \mathrm{Cov}\left[p[f], f_n \right] \end{pmatrix} K^{-1} \mathbf{f}$$

Let's rewrite the vector from the left:

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

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

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

The variance is then straightforward :

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

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

[1] formula 4 in [1]

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

which leads us to the desired result:

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

3 Frank-Wolfe algorithm

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: `http://arxiv.org/abs/1506.02681v3 [stat.ML]`.

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