Summer School - Probabilistic Numerics

Giacomo Garegnani

Dobbiaco - June 2017

This brief report summarizes the topics highlighted in the Dobbiaco Summer School 2017, which took place from the 19 June to the 23 June 2017. Speakers were Philipp Hennig (Max Planck Institute for Intelligent Systems), and Chris Oates (Newcastle university and Alan Turing institute).

1 A machine learning approach to Probabilistic Numerics

In this first part, we summarize results of probabilistic numerics in which prior and posterior distributions are Gaussians. We show how considering the Gaussian distribution allows translating a probabilistic problem into linear algebra computations, thus showing an application on the computation of approximate integrals and on the solution of ordinary differential equations (ODEs).

1.1 Motivation: Gaussian distribution and linear algebra

Gaussian random variables offer a solid framework for Bayesian inference and therefore the implementation of efficient probabilistic numerical methods. Due to their exponential structure, Gaussian random variables are a natural choice in order to treat inference with the instruments of linear algebra. In particular, if we denote by $p(x) = \mathcal{N}(x; \mu, \Sigma)$ a Gaussian distribution with mean x and covariance Σ, it is true that

- products of Gaussians are Gaussians

$$\mathcal{N}(x; a, A)\mathcal{N}(x; b, B) = \mathcal{N}(x; c, C)\mathcal{N}(a; b, A + B),$$

- marginals of Gaussians are Gaussians

$$\int \mathcal{N}\left[\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right] \mathrm{d}y = \mathcal{N}(x; \mu_x, \Sigma_{xx}), \tag{1}$$

- conditionals of Gaussians are Gaussians

$$p(x \mid y) = \mathcal{N}\big(x; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\big),$$

- linear projections of Gaussians are Gaussians

$$p(z) = \mathcal{N}(z; \mu, \Sigma) \implies p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^T).$$

These properties show that the family of Gaussian distributions is closed with respect to all the operations needed for Bayesian inference, and computation of resulting distributions is only done in terms of linear algebra of means and covariance structures.

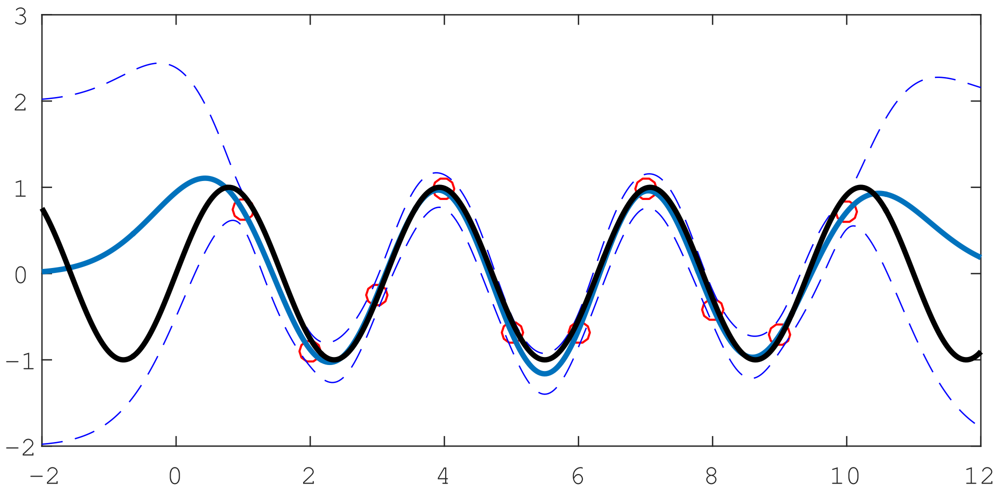

Figure 1: Posterior distribution. Parameters $g = sin(2x)$ (black line), observation noise $\sigma = 0.1$ (red circles), exponential kernel $k(x, y) = \exp(-(x-y)^2/2)$. Mean of the posterior (thick blue) and standard deviations (dashed blue).

1.2 Gaussian processes

The fundamental instrument in order to build probabilistic numerics methods based on Gaussian distributions are Gaussian processes. First of all, the notion of a Mercer kernel is needed.

Definition 1.1 (Mercer Kernel)**.** A function $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a Mercer kernel if, for any finite collection $X = [x_1, \ldots, x_N]$, the matrix $k_{XX} \in \mathbb{R}^{N \times N}$, with $(k_{XX})_{ij} = k(x_i, x_j)$, is positive semidefinite.

Provided with a definition of Mercer kernel, a Gaussian process is directly defined.

Definition 1.2. Let $\mu\colon \mathbb{X} \to \mathbb{R}$ be any function and $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ a Mercer kernel. A Gaussian process $p(f) = \mathcal{GP}(f; \mu, k)$ is a probability distribution over the function $f\colon \mathbb{X} \to \mathbb{R}$, such that every finite restriction to function values $f_X := [f(x_1), \ldots, f(x_N)]$ is a Gaussian distribution $p(f_X) = \mathcal{N}(f_X; \mu_X, k_{XX})$.

The most common example of Gaussian process is the (one-dimensional) Wiener process W, whose distribution can be seen as $p(w) = \mathcal{GP}(w; 0, k_w)$, where $k_w(x, x') = \min\{x, x'\}$. The first application of Gaussian processes is learning in a Bayesian fashion a function from data. In fact, given a valid kernel k and a function μ, we can set the prior distribution over a space of function $f\colon \mathbb{X} \to \mathbb{R}$ to be

$$p(f) = \mathcal{GP}(f; \mu, k). \tag{2}$$

Let us consider then a set of data $\{(x_i, G_i)\}_{i=1}^N$, where $x_i \in \mathbb{X}$ and $G_i \in \mathbb{R}$. We furthermore consider $G_i = g(x_i) + \varepsilon$, where ε is a random disturbance such that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The posterior distribution, i.e., $p(f \mid G)$, is then Gaussian. In fact, if we consider a set of points $\{X_i\}_{i=1}^M$ where we wish to know the value of the mean and variance of the posterior distribution, we can simply obtain that by

$$\mu(X) = k_{xX}(k_{xx} + \sigma^2 I)^{-1} G, \tag{3}$$

$$\Sigma(X) = k_{XX} - k_{xX}(k_{xx} + \sigma^2 I)^{-1} k_{Xx}, \tag{4}$$

where x and X are the vectors of the observations $\{x_i\}_{i=1}^N$ and $\{X_i\}_{i=1}^M$.

1.3 Approximate integrals with $\mathcal{GP}$

Given $F \in \mathbb{R}^{d\times d}$ and $L \in \mathbb{R}^d$, let us consider the d-dimensional SDE

$$\mathrm{d}x(t) = Fx(t)\mathrm{d}t + L\mathrm{d}W_t. \tag{5}$$

This equation, with $x(t_0) = x_0$, locally describes the behavior of a Gaussian process $p(x)$ defined as

$$p(x) = \mathcal{GP}\big(x; e^{F(t-t_0)}, k\big), \quad k(t,t') = \int_{t_0}^{t\wedge t'} e^{F(t-\tau)}LL^T e^{F^T(t'-\tau)}\mathrm{d}\tau. \tag{6}$$

A discrete sampling of the Gaussian process leads to a Markov chain whose transition probability is given by $p(x_{t_i+1} \mid x_{t_i}) = \mathcal{N}(x_{t_{i+1}}; A_{t_i}x_{t_i}, Q_{t_i})$, where

$$A_{t_i} = e^{F(t_{i+1}-t_i)}, \quad Q_{t_i} = \int_0^{t_{i+1}-t_i} e^{F\tau}LL^T e^{F\tau}\mathrm{d}\tau. \tag{7}$$

Consider the problem of integrating a known function $f\colon \mathbb{X} \to \mathbb{R}$ in a probabilistic manner. At the continuous time level, we consider the following SDE

$$\mathrm{d}z(x) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} z(x)\mathrm{d}x + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\mathrm{d}W_t, \tag{8}$$

where the vector $z(x)$ encodes $z(x) = \big(\int_{x_0}^x f(\tilde{x})\mathrm{d}\tilde{x}, f(x)\big)^T$. The aim is constructing a Gaussian process $p(z \mid f) = \mathcal{GP}(z; m, P)$, i.e., finding the evolution in space of the mean m and variance P. If we consider an evenly spaced grid with spacing h and following the procedure above for a general SDE, we get the matrices

$$A(h) = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}, \quad Q(h) = \begin{pmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{pmatrix}. \tag{9}$$

These matrices are computed via (7) and give the update formula for the Markov process defined by (8). Moreover, we have to consider the observations, which are given by the value of the function f on the grid points. These are accounted for in a Kalman filter setting [5] via the matrix $H = (0, 1)$ and the scalar $R = 0$ (noiseless observations). For each point in the grid, there are two phases

(i) Update mean and covariance of z (*prediction step*) via the Markov chain generator by

$$m^- \leftarrow Am, \quad P^- \leftarrow A(h)PA(h)^T + Q. \tag{10}$$

(ii) Consider the observations $y = f(x) - Hm^-$, and update mean and covariance accordingly (*update step*) via

$$m \leftarrow m^- + Ky, \quad P \leftarrow (I - KH)P^-, \tag{11}$$

where $K = P^-H^T/(HP^-H^T)$ is known as the *Kalman gain*.

It is possible to find that following these updates, the mean of the integral (i.e., m_1) is given by the classical trapezoidal rule. The covariance structure P is adding information about the uncertainty on the computation of the integral.

1.4 An ODE solver based on $\mathcal{GP}$

The simple linear algebra structure descending from Gaussian processes enables to formulate ODE solvers actively updating a posterior distribution over the numerical solution with a slight increase in computational cost with respect to classical solvers. The first example of a Bayesian solver for ODEs has been prematurely proposed by J. Skilling [7], and is an active area of research (e.g., H. Kersting and P. Hennig [4] and M. Schober, D.K. Duvenaud, P. Hennig [6]).

Let us consider $f\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ and the ODE

$$y'(t) = f(y(t), t), \quad y(0) = y_0 \in \mathbb{R}^d. \tag{12}$$

The main idea behind Bayesian solvers of ODEs descends practically from the probabilistic interpretation of approximate integrals, and their algorithms exploit the typical updates of Kalman filters. In particular, we consider the SDE

$$\mathrm{d}z(x) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} z(x)\mathrm{d}x + \begin{pmatrix} 0 \\ \vartheta \end{pmatrix} \mathrm{d}W_t, \tag{13}$$

which is the same as (8), with a relaxation parameter ϑ. Computing the matrices with (7), one gets

$$A(h) = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}, \quad Q(h) = \vartheta^2 \begin{pmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{pmatrix}. \tag{14}$$

Conversely to the computation of integrals, in this case the points in which evaluating the function f are unknown a priori. Hence, after the *prediction step* and before the *update step*, observations are generated as $y_i = f(Hm_i^-, t_i)$. It is possible to show [6] that considering these matrices the prediction step is equivalent to a step of the Explicit Euler method, while after the update step the value of m is equal to that of the trapezoidal rule. Overall, the mean of the method converges in order two to the true solution of the ODE.

Kersting and Hennig [4] show that it possible to build higher order methods just enlarging the dimension of the state space of the SDE (8). In particular, considering matrices

$$F = \begin{pmatrix} 0 & 1 & 0 & & \cdots & 0 \\ & 0 & 1 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & 0 & 1 & 0 \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{q+1 \times q+1}, \quad L = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vartheta \end{pmatrix} \in \mathbb{R}^{q+1}, \tag{15}$$

it is possible to get a method whose mean is converging with order q to the true solution. In particular, it has been shown (Hennig, Osborne, Girolami, unpublished) that for $q = 1, 2, 3, 4$ and a careful choice of the evaluation times the method descending from this choice of matrices is locally equivalent to Runge-Kutta methods of order q.

2 A general framework of Bayesian numerical methods

In this section, we consider a general definition and application of Bayesian numerical methods, which has been first formalized by A. Stuart [8].

2.1 A definition of Bayesian numerical methods

Many probabilistic methods have been proposed for the solution of classical problems of numerical analysis. Not all of them though can be classified as being Bayesian. Any numerical method can be written in terms of

- An unobservable quantity $x \in \mathcal{X}$,
- A quantity of interest $Q(x) \in \mathcal{Q}$,
- An information operator $x \mapsto A(x) \in \mathcal{A}$, where $\dim \mathcal{A} = n < \infty$.

The goal of deterministic method is to find $Q(x)$ using the information given by $A(x)$. For example, we can consider the problem of integrating a function $x(t)$. Then our information $A(x)$ will be a set of points $\{x(t_i)\}_{i=1}^n$, and our quantity of interest will be $Q(x) = \int x(t)\mathrm{d}t$. Given $a \in \mathcal{A}$, the numerical method can be encoded by some function $b\colon \mathcal{A} \to \mathcal{Q}$ such that the output of the method is $b(a)$. In the Bayesian framework, a prior distribution on the unobservable quantity x has to be specified, and we will denote it by P_x, and the result is a posterior distribution $P_{x|a}$ which is used to approximate $Q(x)$.

Definition 2.1. A numerical method $B(P_x, a)$ is *Bayesian* if and only if

$$B(P_x, a) = Q_{\#}P_{x|a}, \tag{16}$$

where $Q_{\#}P_{x|a}$ is the *push forward measure* associated to Q and $P_{x|a}$, i.e., for any operator T, set S and measure μ, the measure such that

$$(T_{\#}\mu)(S) = \mu\big(T^{-1}(S)\big), \tag{17}$$

where T^{-1} denotes the counter-image of S through T.

Remark 2.1. If we denote the set of measures on a set Ω as $\mathcal{P}_\Omega$, then $P_x \in \mathcal{P}_\mathcal{X}$ and $B(P_x, a) \in \mathcal{P}_\mathcal{Q}$.

This rigorous and restrictive definition of a Bayesian probabilistic numerical method has been proposed by J. Cockayne et al. [2]. In this work, it is possible to find an interesting classification of several probabilistic solvers present in literature.

2.2 Well-posedeness of Bayesian numerical methods

From the definition above it is unclear how the posterior distribution $P_{x|a} \in \mathcal{P}_\mathcal{X}$ over the unobservable quantity x is defined. In the finite-dimensional case (i.e., $\dim \mathcal{X} < \infty$), it is possible to work with the Lebesgue measure and hence Bayes' rule states

$$p(x \mid a) = \frac{p(a \mid x)p(x)}{p(a)}. \tag{18}$$

In our setting, the unobservable quantity x is in an infinite dimensional space $\mathcal{X}$ (e.g., if x is the solution of a differential equation $\mathcal{X}$ is a functional space). Hence, it is necessary to work with Radon-Nikodym derivatives, i.e.,

$$\frac{\mathrm{d}P_{x|a}}{\mathrm{d}P_x} = \frac{p(a \mid x)}{p(a)}. \tag{19}$$

Let us remark that the quantity on the right of the equality is well defined as $\dim \mathcal{A} = n < \infty$. Depending on the prior P_x, even Radon-Nikodym derivatives could be undefined, and it is necessary to give more structure to the spaces $\mathcal{X}$, $\mathcal{Q}$ and $\mathcal{A}$ in order to define the posterior distribution $P_{x|a}$. A detailed treatment of these conditions, including the theory of *disintegration*, is given by Cockayne et al. [2]. In this work, an overview of the methods available to sample from $P_{x|a}$ is given. In particular, if P_x is Gaussian, Monte Carlo methods are available to fulfill this purpose efficiently.

2.3 Probabilistic solution of PDEs

Consider a domain $\Omega \subset \mathbb{R}^d$, two operators $\mathcal{D}$ and $\mathcal{B}$ and the partial differential equation

$$\begin{aligned} \mathcal{D}x(t) &= g(t), \quad t \in \Omega, \\ \mathcal{B}x(t) &= b(t), \quad t \in \partial\Omega. \end{aligned} \tag{20}$$

A deterministic method to approximate the solution $x(t)$ with a function $\hat{x}(t)$ is the *symmetric collocation.* Given a function $k\colon \Omega^2 \to \mathbb{R}$, the numerical solution is given by

$$\hat{x}(t) = \sum_{i=1}^{N} w_i \bar{\mathcal{D}} k(t, t_i), \tag{21}$$

where $\{w_i\}_{i=1}^N$ are real weights and $\bar{\mathcal{D}}$ is the adjoint of $\mathcal{D}$. if $T = (t_1, t_2, \dots, t_N)^T$ and $\mathbf{g} = (g(t_1), g(t_2), \dots, g(t_N))^T$, then the weights are given by

$$\mathbf{w} = [\mathcal{D}\bar{\mathcal{D}} K(T, T)]^{-1} \mathbf{g}, \tag{22}$$

where $K(T, T')$ is the matrix with elements $[K(T, T')]_{ij} = k(t_i, t'_j)$. Then the numerical solution is

$$\hat{x}(t) = \bar{\mathcal{D}} K(t, T) [\mathcal{D}\bar{\mathcal{D}} K(T, T)]^{-1} \mathbf{g}. \tag{23}$$

This numerical method can be extended naturally to a Bayesian probabilistic method. Let us consider a Gaussian prior $P_x\colon x \sim \mathcal{GP}(0, k)$ and the information operator

$$A(x) = \begin{pmatrix} \mathcal{D}x(t_1) \\ \vdots \\ \mathcal{D}x(t_N) \end{pmatrix} = \begin{pmatrix} g(t_1) \\ \vdots \\ g(t_N) \end{pmatrix}, \tag{24}$$

with furthermore $Q(x) = x$. The posterior $P_{x|a}$ is then a $\mathcal{GP}$ (see Section 1), and has parameters m and Σ given by

$$\begin{aligned} m(t) &= \bar{\mathcal{D}} K(t, T) [\mathcal{D}\bar{\mathcal{D}} K(T, T)]^{-1} \mathbf{g}, \implies \text{deterministic method} \\ \Sigma(t, t') &= k(t, t') - \bar{\mathcal{D}} K(t, T)) [\mathcal{D}\bar{\mathcal{D}} K(T, T)]^{-1} \mathcal{D} K(T, t'). \end{aligned} \tag{25}$$

For this method, if h is the fill distance (i.e., the maximum distance between two points t_i and t_j in the domain Ω) it is possible to prove that under the posterior

$$P_{x|a}\{x' : \|x' - x\|_{L^2(\Omega)} < \varepsilon\} = 1 - \mathcal{O}(\varepsilon^{-1} h^{2\beta - 2\rho - d}), \tag{26}$$

where β is related to the choice of the kernel k, ρ is the order of $\mathcal{D}$ and d is the dimensionality of the problem.

Defining a posterior distribution over the solution x can be exploited in inverse problems. Let us consider $\mathcal{D} = \mathcal{D}_\vartheta$ for an unknown parameter ϑ. In classical Bayesian inverse problems, it is possible to reconstruct its value via MCMC algorithms, where for a set of data $\mathcal{Y}$, the likelihood function $\mathcal{L}(\mathcal{Y} \mid \vartheta)$ is approximated via a deterministic algorithm. This can lead to posterior distributions over the parameter ϑ which are concentrated far from its true value. This considerations have been presented both in works strictly related with Bayesian forward problems (Cockayne et al, [1]), and where the forward problem is solved with a non-Bayesian probabilistic integrator (Conrad et al, [3]). If a posterior distribution $P_{x|a}$ is given, the likelihood can be computed via marginalization as

$$\begin{aligned} \mathcal{L}(\mathcal{Y} \mid \vartheta) &\propto \int p(y \mid \vartheta, x) \mathrm{d}P_{x|a} \\ &\implies \mathcal{Y} \mid \vartheta \sim \mathcal{N}(m, \Gamma + \Sigma), \end{aligned} \tag{27}$$

where Γ is the variance structure of the observational noise, and m, Σ are given in (25). Such an approximation of the likelihood can be exploited in a Monte Carlo algorithm (MCMC) for determining the posterior distribution $P_{\vartheta|\mathcal{Y}}$.

3 Final considerations

The methods proposed in the first section of this report were presented by P. Hennig, while the second part is related to the work of C. Oates.

In the first part, the proposed methods have a clear practical interest in applications of machine learning and artificial intelligence. In fact, the main goal of these methods is maintaining a low computational cost, which translates in low computational times and efficiency, while providing a complete probabilistic interpretation of the numerical solution. Although indisputably interesting, these methods lack rigorous classic numerical analysis and seem too tailored for specific purposes.

In the second part, a complete theoretical analysis of Bayesian methods in the statistics sense is developed. It is undeniable that such a classification effort is relevant. However, in this case the issue is the opposite with respect to the first part, as no attention is devoted to the development and analysis of concrete numerical schemes. In any case, the complete theoretical framework for forward and inverse Bayesian problems, given especially the work of A. Stuart [8] and Cockayne et al. [2], can be relevant for the analysis of particular methods. Finally, even after a discussion with C. Oates it is unclear why a Bayesian approach, intended in the sense of Definition 2.1, should be preferred to any probabilistic method as, e.g., the one of the probabilistic methods presented in [3] or of our RTS-RK method.

References

[1] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami, *Probabilistic meshless methods for partial differential equations and bayesian inverse problems*, 2016.

[2] ——, *Bayesian probabilistic numerical methods*, 2017.

[3] P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput., (2016).

[4] H. Kersting and P. Hennig, *Active uncertainty calibration in Bayesian ODE solvers*, in Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), AUAI Press, 2016, pp. 309–318.

[5] S. Särkkä, *Bayesian filtering and smoothing*, vol. 3, Cambridge University Press, 2013.

[6] M. Schober, D. K. Duvenaud, and P. Hennig, *Probabilistic ODE solvers with Runge-Kutta means*, in Advances in neural information processing systems, 2014, pp. 739–747.

[7] J. Skilling, *Bayesian Solution of Ordinary Differential Equations*, Springer Netherlands, 1992, pp. 23–37.

[8] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.