# 1 Bayesian statistics and Markov chain Monte Carlo

In the following section we will briefly discuss the main features of a Bayesian statistical approach, comparing it with the classical inferential statistics. Then, we will present a class of techniques used to practically perform Bayesian inference, the Markov chain Monte Carlo (commonly denoted with the acronym MCMC), discussing some of the possible implementations and properties of these methods.

## 1.1 Bayes' formula

The basis of Bayesian statistics can be found in the simple Bayes' formula. Let us consider an event space $\Omega$, a sigma-algebra $\mathcal{A}$, a probability measure $P$ and the probability space $(\Omega, \mathcal{A}, P)$. If $A$ and $B$ are two events in $\Omega$, the probability of the intersection of $A$ and $B$ is given by

$$\begin{aligned} P(A, B) &= P(A|B)P(B) \\ &= P(B|A)P(A). \end{aligned}$$

The equivalence between the two formulations leads to Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

In Bayes' formula, $P(A|B)$ is a probability distribution, therefore its integral has to be equal to one. Therefore, one can rewrite Bayes' rule disregarding the value of $P(B)$ as

$$P(A|B) \propto P(A)P(B|A).$$

The exact value of the probability distribution can be therefore obtained by normalization as

$$P(A|B) = \frac{P(A)P(B|A)}{\int_\Omega P(A)P(B|A)}.$$

The quantities appearing in (1) are commonly referred to as

- posterior distribution $P(A|B)$,
- prior distribution $P(A)$,
- likelihood $P(B|A)$.

The probability distribution of $A$ is often the object of Bayesian inference and the event $B$ is an observable quantity related to $A$. Then the likelihood $P(B|A)$ is not a probability distribution but the likelihood the observations of $B$ have with respect to $A$. Therefore, in order to avoid misinterpretations, we will denote in the following the likelihood by $\mathcal{L}(B|A)$. Moreover, we will adopt in the following sections the notation $\mathcal{Q}$ for the prior and $\pi$ for the posterior distributions, thus obtaining

$$\pi(A|B) \propto \mathcal{Q}(A)\mathcal{L}(B|A).$$

## 1.2 Parametrized models

Bayes' formula opens a new perspective to statistical modeling with respect to the classical inferential standards. In particular, parametrized models are particularly suited to a Bayesian approach. Let us consider a parametrized model for predicting the outcome of an experiment. Let us denote by $\theta$ the parameter driving the experiment and by $X$ a random variable representing its outcome. Let us consider for simplicity $\theta$ as a vector of $\mathbb{R}^{N_p}$, where $N_p$ is the dimension of the parameter space. We will then denote by $\theta_i$ the $i$-th component of the parameter, with $i = 1, \dots, N_p$. For instance, we could consider the toss of a coin and estimate its probability to fall on one of the two side as $\theta$, or more complicated physical models influenced by an intractable source of noise.

In the classical statistic approach we would state an hypothetical distribution for $X$ depending on the parameter (e.g., $X \sim \mathcal{N}(\theta_1, \theta_2)$). Then, let us suppose that a set of observation $\mathcal{Y}_i =$

$\{y_0, y_1, \ldots, y_i\}$, $i = 1, \ldots, N_d$, of the outcome of the experiment is available. We can consider these observations to be produced by a random variable $Y$ representing a quantity connected to the random variable $X$ by a law, that we denote by $f$, and biased by a measurement error, that we denote by $\varepsilon$. For instance, we could consider the following additive observational model

$$Y \sim f(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Gamma). \tag{2}$$

Given this background, there are many techniques available in the classical approach to compute an estimator $\hat{\theta}$ of the true parameter and to state a measure of the uncertainty the statistical model has on the estimator. In particular, one can compute analytically quantities such as the mean square error (MSE) of the estimator or give confidence intervals on $\theta$ such that its true value falls in the interval up to a threshold probability.

In the Bayesian frame the estimation of $\theta$ follows a completely different philosophy. The outcome of the Bayesian inference is neither a value of the parameter nor a set of values in which it is likely to be included, but it is a *probability distribution.* Maintaining the notation introduced above, Bayes' formula in this frame reads

$$\pi(\theta|\mathcal{Y}_i) \propto \mathcal{Q}(\theta)\mathcal{L}(\mathcal{Y}_i|\theta).$$

Let us analyze separately the two terms of this equation.

- Given a set of observations $y_i$, $i = 1, \ldots, N_d$, and an observational model the likelihood $\mathcal{L}(Y|\theta)$ can be evaluated. For example, in the Gaussian case introduced in (2) analytical formulas for the likelihood are available.
- The prior distribution $\mathcal{Q}(\theta)$ has to be established before the observation are obtained. This is a crucial part of the process of Bayesian inference, since in practice if the prior distribution is wrong or inadmissible, the obtained posterior may be negatively affected by this choice.

The two approaches give both equally valid results but in a completely different spirit. While in classical statistics the model driving an experiment is predetermined and its parameters are computed using observations, in the Bayesian frame the object of study is the model behind the parameter itself, which is revealed by the observations.

### 1.2.1 An example: parametrized differential equations

In this paragraph we present a simple example that is useful to understand Bayesian inference of parameters in general and the scope of this work in particular. Let us consider the probability space $(\Omega, \mathcal{F}, P)$, a one-dimensional standard Wiener process $\{W(t)\}_{t\geq 0}$ and a filtration $\{\mathcal{F}(t)\}_{t\geq 0}$ such that $W(t)$ is $\mathcal{F}(t)$ measurable. Moreover, let us consider the following one-dimensional stochastic differential equation (SDE)

$$\begin{aligned} dX(t) &= \lambda X(t)\mathrm{d}t + \mu X(t) dW(t), && 0 < t < T, \\ X(0) &= X_0, && X_0 \in \mathbb{R}, \end{aligned} \tag{3}$$

where $\lambda, \mu$ are real parameters and we consider $X_0$ is a random variable. It is known that under the hypotheses of Itô calculus the solution of (3) is given by

$$X(t) = X_0 \exp\left(\left(\lambda - \frac{1}{2}\mu^2\right) t + \mu W(t)\right),$$

which is a stochastic process often referred to as *geometric Brownian motion.* This equation and its solution have extensively been studied in numerous applications. For example, it used as a simple financial tool in order to model option or stock pricing, with the parameter $\lambda$ which is often referred to as the *drift* and the diffusion coefficient $\mu$ as the *volatility.* Given the model described by (3), we may be interested in inferring the value of one, or more, of its parameters.

Let us consider the following assumptions

- $X_0$ is a known real value,

- the drift coefficient $\lambda$ is known a priori,
- the diffusion coefficient $\mu$ is unknown but a prior distribution $\mathcal{Q}(\mu)$ has been stated,
- the value of the solution $X(t)$ is observable at a set of times $t_i$, $i = 1, \dots, N_d$, such that $t_{N_d} = T$, with a zero-mean additive Gaussian measurement noise $\varepsilon$, i.e., the observations $y_i$, are given by

  $$y_i = x(t_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad \sigma \in \mathbb{R}, \quad i = 1, \dots, N_d,$$

  where we denote by $x(t_i)$ a realization of $X$ evaluated at time $t_i$.

Let us denote as $\mathcal{Y}_i$ the set of all the observations $y_i$ up to time $t_i$. We are interested in estimating the value of the parameter $\mu$. In a Bayesian frame, this corresponds to providing a distribution $\pi$ conditional to the observation following Bayes' rule, i.e.,

$$\pi(\mu|\mathcal{Y}_i) \propto \mathcal{Q}(\mu)\mathcal{L}(\mathcal{Y}_i|\mu).$$

In this simple frame, the knowledge of the analytical form of the solution and of the measurement error gives us an exact notion of the model connecting the parameter and the observations. Hence, for each choice of the value of $\mu$ it is possible to evaluate the likelihood function $\mathcal{L}$ as follows

$$\mathcal{L}(\mathcal{Y}_{N_d}|\mu) = (2\pi\sigma^2)^{-N_d/2} \prod_{k=1}^{N_d} \mathbb{E}\left[\exp\left(-\frac{\sigma^2}{2}(X(t_k) - y_k)^2\right)\right],$$

where we omitted the implicit dependence of the process $X$ on $\mu$. Furthermore, if the prior distribution $\mathcal{Q}$ admits a density in closed form, it is possible to evaluate it on any choice of $\mu$. Therefore, it is possible to compute for each value of $\mu$ the value of the posterior distribution associated with the available set of measurements.

In this simple example the analytical form of any of the quantities of Bayes' formula and the small dimension of the parameter space imply that with a low effort it is possible to determine the value of the posterior distribution. In general this is not true, and as we will show in the next sections fine Monte Carlo techniques have been proposed to generate samples from any distribution.

## 1.3 Markov chain Monte Carlo methods

Markov chain Monte Carlo methods (MCMC) are a class of techniques used to perform Bayesian analyses [7, 8]. In the following we will present the main idea behind the method as well as some examples of their implementation.

Let us consider a model which has a random variable $X$ as its outcome parametrized by a parameter $\theta$ and a set of observations $\mathcal{Y}_i = \{y_1, y_2, \dots, y_i\}$, $i = 1, \dots, N_d$, providing information regarding $X$. Then, thanks to Bayes' rule, we can construct the posterior distribution of $\theta$ by Bayes' rule

$$\pi(\theta|\mathcal{Y}_i) \propto \mathcal{Q}(\theta)\mathcal{L}(\mathcal{Y}_i|\theta).$$

As in the previous paragraphs, let us assume that $\theta$ is a real-valued parameter of dimension $N_p$. If the parameter space has a high dimension, it is computationally expensive exploring all the possible values in order to build the posterior distribution, especially if evaluating the model connecting $\theta$ and the random variable $X$ is non-trivial. If we are interested in knowing the expectation of some measurable function $g \colon \mathbb{R}^{N_p} \to \mathbb{R}$ of $\theta$ we can proceed by the following Monte Carlo evaluation

$$\mathbb{E}\left[g(\theta)\right] = \int_{\mathbb{R}^{N_p}} g(\theta)\pi(\mathrm{d}\theta|\mathcal{Y}_i) \approx \frac{1}{N}\sum_{k=1}^{N} g(\theta^{(k)}), \tag{4}$$

where $\theta^{(k)}$, $k = 1, \dots, N$, is a set of realizations of $\theta$. While the equality in the equation follows from the definition of expectation, there is no guarantee that the Monte Carlo estimator will be a good approximation of the expectation regardless of the samples. MCMC techniques consist in generating samples such that the Monte Carlo approximation is valid without exploring the

**Algorithm 1:** Metropolis-Hastings.

**Data**: $\theta^{(0)} \in \mathbb{R}^{N_p}, N \in \mathbb{N}_0$.
Compute $\pi(\theta_0)$ ;
**for** $i = 0, \dots, N$ **do**
  Draw $\vartheta$ from $q(\theta^{(i)}, \cdot)$ ;
  Compute the acceptance probability $\alpha(\theta^{(i)}, \vartheta)$ as in (5) ;
  Draw $u$ from $\mathcal{U}(0, 1)$ ;
  **if** $\alpha > u$ **then**
    Accept $\vartheta$, set $\theta_{i+1} = \vartheta$ ;
  **else**
    Set $\theta^{(i+1)} = \theta^{(i)}$;
  **end**
**end**

whole parameter space, which would lead to an unaffordable computational time on any modern computer. As the name of the methods suggests, given an initial guess $\theta^{(0)}$, MCMC builds a discrete Markov chain $\{\theta^{(i)}\}_{i\geq 0}$ such that the Monte Carlo approximation in (4) is valid. Formally, this is achieved considering a *transition kernel* $P$ which given the current element of the chain $\theta^{(i)}$ produces the next guess $\theta^{(i-1)}$. Under a set of assumptions on $P$ [8], we have the theoretical guarantee that the samples $\theta^{(i)}$ are drawn from the same *stationary distribution* for $i$ large enough. We can build many transition kernels having this property, and any valid choice of $P$ leads to a different MCMC method. In the following, we will present the widely-used *Metropolis-Hastings* algorithm, as well as two of its variants that were necessary for our work.

### 1.3.1 Metropolis-Hastings algorithm

In this paragraph we will introduce one of the most successful MCMC methods, the Metropolis-Hastings method (MH). In MH, the samples forming the Markov chain are generated following a *proposal distribution* $q\colon \mathbb{R}^{N_p} \times \mathbb{R}^{N_p} \to \mathbb{R}^+$ which satisfies the condition

$$\int_{\mathbb{R}^{N_p}} q(x, y)\mathrm{d}y = 1,$$

thus $q$ is a probability distribution in its second argument. Given the current guess $\theta^{(i)}$, MH proposes the new element of the Markov Chain drawing a value $\vartheta$ from $q(\theta^{(i)}, \cdot)$. The new guess is not automatically accepted as the new element $\theta^{(i+1)}$ of the Markov chain, but it is accepted with a probability, that we denote by $\alpha(\theta^{(i)}, \theta^{(i+1)})$. Formally, the transition kernel $P_{\mathrm{MH}}$ representing the move made by MH from $\theta^{(i)}$ to $\theta^{(i+1)}$ is given by [10]

$$P_{\mathrm{MH}}(\theta^{(i)}, \theta^{(i+1)}) = \alpha(\theta^{(i)}, \theta^{(i+1)})q(\theta^{(i)}, \theta^{(i+1)}) + \delta_{\theta^{(i)}}(\theta^{(i+1)})\rho(\theta^{(i)}),$$

where $\delta_x$ is the Dirac delta centered in $x$ and $\rho$ is defined as

$$\rho(\theta^{(i)}) := 1 - \int_{\mathbb{R}^{N_p}} \alpha(\theta^{(i)}, x)q(\theta^{(i)}, x)\mathrm{d}x.$$

In words, the expression of the transition kernel $P_{\mathrm{MH}}$ is equivalent to stating that the new guess $\vartheta$ generated from the proposal distribution is accepted with probability $\alpha$ and rejected with probability $1-\alpha$. Imposing that $P_{\mathrm{MH}}$ satisfies the hypotheses that guarantee the convergence of MCMC, we can get the expression of the acceptance probability in closed form as

$$\alpha(\theta^{(i)}, \vartheta) = \min\left\{\frac{\pi(\vartheta)q(\vartheta, \theta^{(i)})}{\pi(\theta^{(i)})q(\theta^{(i)}, \vartheta)}, 1\right\}. \tag{5}$$

As it is possible to remark from its pseudo-code, given in Algorithm 1, MH is extremely simple to implement on a computer in any programming language. In fact, the only choice left to the

**Algorithm 2:** Robust adaptive Metropolis.

**Data**: $\theta^{(0)} \in \mathbb{R}^{N_p}$, $N \in \mathbb{N}_0$, $S_0 \in \mathbb{R}^{N_p \times N_p}$, $\alpha^* \in (0,1)$.
Compute $\pi(\theta_0)$ ;
**for** $i = 0, \dots, N$ **do**
    Draw $z$ from $Z \sim \mathcal{N}(0, I)$ ;
    $\vartheta = \theta^{(i)} + S_i z$ ;
    Compute the acceptance probability $\alpha(\theta^{(i)}, \vartheta)$ as in (5) ;
    Draw $u$ from $\mathcal{U}(0,1)$ ;
    **if** $\alpha > u$ **then**
        Accept $\vartheta$, set $\theta_{i+1} = \vartheta$ ;
    **else**
        Set $\theta^{(i+1)} = \theta^{(i)}$;
    **end**
    Compute $S_{i+1}$ as in (8) ;
**end**

user of a MH algorithm is the proposal distribution $q(x, y)$. Unfortunately, this choice could impact negatively the behavior of MH, slowing dramatically its convergence towards the stationary distribution of the Markov chain. Let us first remark that if the proposal distribution is a symmetric function in its two arguments, i.e., $q(x, y) = q(y, x)$, the expression of the acceptance probability simplifies to

$$\alpha(\theta^{(i)}, \vartheta) = \min\left\{\frac{\pi(\vartheta)}{\pi(\theta^{(i)})}, 1\right\}. \tag{6}$$

For example, a Gaussian proposal distribution centered in $\theta^{(i)}$ with covariance matrix $\Sigma$ in $\mathbb{R}^{N_p \times N_p}$ is a common choice for $q(x, y)$. In this case, the proposal distribution is given up to a normalization constant by

$$q(x, y) \propto \exp\left(-\frac{1}{2}(x-y)^T \Sigma^{-1} (x-y)\right). \tag{7}$$

In this work, we mainly used a Gaussian proposal distribution, therefore the acceptance probability will be of the form (6).

Two main issues have to be taken into account before moving on to the practical applications of MH we considered for this work.

1. What is a good choice for the proposal function $q(x, \cdot)$?
2. How can we modify MH in case it is not possible, or not practical, to evaluate the posterior distribution $\pi(\theta)$?

In the following paragraphs we will present two approaches to modify MH targeting these two questions.

### 1.3.2 An adaptive approach

In the frame of MH algorithms, it is important to have a control on the *acceptance ratio*, i.e., the ratio of new proposed values $\vartheta$ that are included in the Markov chain $\{\theta^{(i)}\}_{i \geq 0}$. In the MH frame, the acceptance ratio depends on the chosen proposal distribution, as if the new guess produced via the proposal distribution have a low probability of being accepted, a low value of acceptance ratio will result from the algorithm. If the initial proposal distribution does not provide with acceptable values $\vartheta$, it may be necessary to tune it during the advancement of MH. An algorithm which targets this issue is the robust adaptive Metropolis (RAM) [12].

Let us consider the case a Gaussian proposal distribution $q(x, y)$ as in (7). At the $n$-th step of MH the new guess $\vartheta$ of the parameter is given by

$$\vartheta = \theta^{(n)} + z, \quad Z \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma$ is the covariance matrix. It is possible to build a sequence of matrices such that the convergence properties of MH are not spoiled and the acceptance rate is asymptotically equal to a given value $\alpha^*$ [12]. This is obtained through the following update

$$\vartheta = \theta_k + S_n z_n, \quad Z_n \sim \mathcal{N}(0, I),$$

with $S_n$ a lower triangular positive definite matrix and $I$ the identity matrix. Given an initial choice $S_0$, the matrix $S_n$ is updated at each iteration with a lower triangular matrix $S_{n+1}$ satisfying

$$S_{n+1}S_{n+1}^T = S_n \left( I + \eta_n \left( \alpha(\theta^{(n)}, \vartheta) - \alpha^* \right) \frac{z_n z_n^T}{z_n^T z_n} \right) S_n^T. \tag{8}$$

Hence, we can compute $S_{n+1}$ as the Cholesky factorization of the right hand side. Let us remark that this update has to be performed at each iteration of RAM, both in case $\vartheta$ is accepted and rejected. The sequence $\{\eta_n\}_{n\geq 1}$ can be any sequence decaying to zero with $n$. In this work, we consider

$$\eta_n = n^{-\gamma}, \quad 0.5 < \gamma \leq 1.$$

Often the computational cost needed for the evaluation of the posterior distribution is high with respect to the dimension $N_p$ of the parameter space. In Algorithm 2 we give the pseudo-code for the RAM update. Therefore, performing a Cholesky factorization at each iteration, which has a complexity of $\mathcal{O}(N_p^3)$, does not spoil the performances of RAM with respect to a standard MH.

Let us consider a two-dimensional real random variable $X$ whose distribution has the following density

$$\pi(X) \propto \exp\bigl(-10(X_1^2 - X_2)^2 - (X_1 - 0.25)^4\bigr), \tag{9}$$

where we denoted by $X_i$, $i = 1, 2$ the two components of $X$ and we omitted the normalization constant. We then consider a real value $\sigma$ in the set $\{0.01, 0.5, 2.0\}$ and target the distribution defined by (9) either using a standard MH with the proposal distribution given by a zero-centered normal distribution with covariance $\Sigma = \sigma^2 I$, or using RAM with the same choice of covariance structure as an initial guess and $\alpha^* = 0.4$. We run $N = 5000$ iterations of both algorithms and register all the guesses they produce as well as the final acceptance ratio. Results (Figure 1) show that for the $\sigma = 0.01$ and $\sigma = 2.0$ standard MH fails to properly describe the posterior distribution, either accepting too many guesses and partially describing the posterior, or refusing almost all guesses therefore obtaining an insufficient number of samples. On the other hand, RAM adapts the step and for any choice of $\sigma$ the samples we obtain are equally good, with an acceptance ratio near to $\alpha^*$ (Table 1).

| MCMC | $\sigma = 0.01$ | $\sigma = 0.5$ | $\sigma = 2.0$ |
|---|---|---|---|
| MH | 0.96 | 0.35 | 0.06 |
| RAM | 0.43 | 0.40 | 0.38 |

**Table 1:** Acceptance ratios for MH and RAM with posterior distribution (9)

.

### 1.3.3 Pseudo-marginal Metropolis-Hastings

In this paragraph we discuss the second issue presented above. Let us consider the case in which it is not possible to evaluate the posterior distribution $\pi(\theta)$, or it is too computational expensive. For instance, in the example we provided in Section 1.2.1 the analytical solution of the SDE is computable. If we have a general equation which does not admit a closed-form solution, it is not possible to evaluate the likelihood function. Therefore, the standard MH algorithm and its adaptive version RAM are not applicable.

An algorithm that has been proposed to overcome this issue is the so-called *pseudo-marginal* MCMC [4], which is also known as particle Markov chain Monte Carlo (PMCMC) [1]. The main idea of the proposed pseudo-marginal algorithms is modifying the target of the algorithm to a

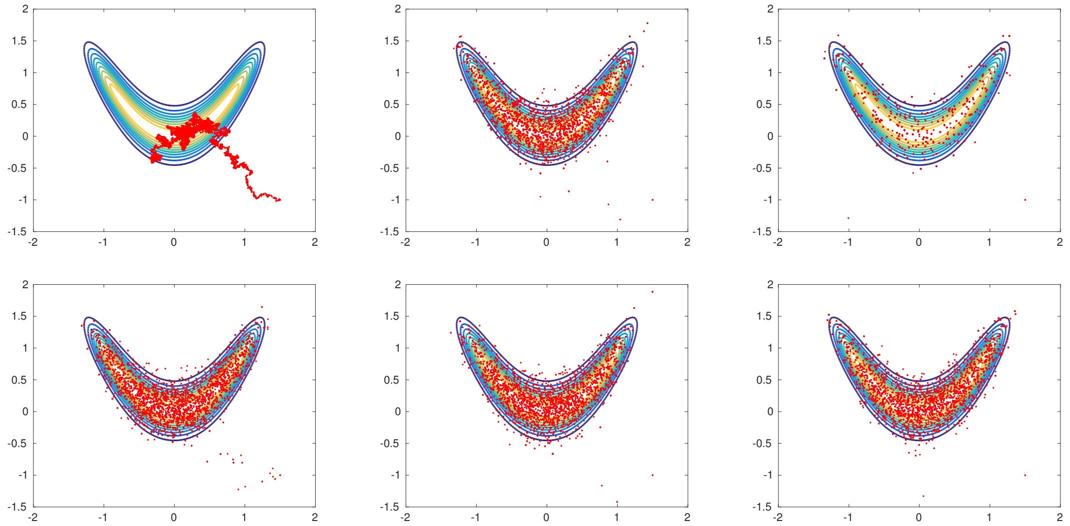



**Figure 1:** Samples produced by MH and RAM for the distribution (9). The contour lines of the density function are plotted for all the sets of results. In the first row we show the results obtained with MH for a normal update with covariance $\Sigma = \sigma^2 I$ with $\sigma = \{0.01, 0.5, 2.0\}$ from left to right. In the second row we show the results obtained with RAM with the same values of $\Sigma$ as an initial guess of the covariance structure.

distribution $\pi(\theta, \xi)$ that admits $\pi(\theta)$ as a marginal distribution and that is easier than $\pi(\theta)$ to evaluate. Then, we can compute an unbiased Monte Carlo approximation $\pi_M(\theta)$ of the marginal distribution as

$$\pi_M(\theta) = \frac{1}{M} \sum_{i=1}^{M} \pi(\theta, \xi^{(i)}), \tag{10}$$

where the values $\xi^{(i)}$ are realizations of the random variable $\xi$. The acceptance probability $\alpha_M$ has then the same form of $\alpha$ in the standard MH, with $\pi_M(\theta)$ instead of the true marginal distribution, i.e.,

$$\alpha_M(\theta^{(i)}, \vartheta) = \min\left\{ \frac{\pi_M(\vartheta) q(\vartheta, \theta^{(i)})}{\pi_M(\theta^{(i)}) q(\theta^{(i)}, \vartheta)}, 1 \right\}. \tag{11}$$

The pseudo-code of the resulting algorithm is shown in Algorithm 3. Let us remark that if the estimator $\pi_M(\theta^{(i)})$ at the $i$-th iteration of MCMC is computed at each iteration and not recycled from the previous iterations, the resulting algorithm is often referred to as Monte Carlo within Metropolis (MCWM) [2] or noisy pseudo-marginal Metropolis [10]. Even though recomputing the estimator may be computationally expensive, the resulting Markov chain has an higher acceptance ratio, i.e., it explores the relevant values of the parameter $\theta$ faster, therefore defining better the posterior distribution. The main issue that has been addressed by the research on this kind of pseudo-marginal algorithms is whether the invariant distribution of the Markov chain converges to the marginal posterior distribution of the random variable $\theta$. It has been shown [2, 10] that under appropriate assumptions the following properties are valid

1. the transition kernel $P_M$ given by (11) converges to an invariant distribution $\pi_M$ with the number of iterations $N$ of MCMC if the number of Monte Carlo draws $M$ is large enough [2, Theorem 9],

2. the invariant distribution $\pi_M$ obtained with MCWM converges to the true marginal distribution $\pi$ if $M$ tends to infinity [10, Theorem 4.1],

3. under stronger assumptions, it is possible to obtain convergence rates of $\pi_M$ to $\pi$ with respect to $M$ [10, Theorem 4.2 and Proposition 4.1].

**Algorithm 3:** Monte Carlo within Metropolis.

**Data**: $\theta^{(0)} \in \mathbb{R}^{N_p}$, $N \in \mathbb{N}_0$.

Compute $\pi(\theta_0)$ ;
**for** $i = 0, \ldots, N$ **do**
  Draw $\vartheta$ from $q(\theta^{(i)}, \cdot)$ ;
  Compute the estimators $\pi_M(\theta^{(i)}, \xi)$ and $\pi_M(\vartheta, \xi)$ as in (10) ;
  Compute the acceptance probability $\alpha_M(\theta^{(i)}, \vartheta)$ as in (11);
  Draw $u$ from $\mathcal{U}(0,1)$ ;
  **if** $\alpha > u$ **then**
    Accept $\vartheta$, set $\theta_{i+1} = \vartheta$ ;
  **else**
    Set $\theta^{(i+1)} = \theta^{(i)}$;
  **end**
**end**

Let us consider the example provided in Section 1.2.1. If we choose an SDE which does not admit a closed-form solution, it is impossible to evaluate the posterior distribution, as the likelihood function does not admit an analytical expression. On the other hand, there exists a large variety of numerical methods [9] that we can apply together with a Monte Carlo approximation to compute an estimator of the likelihood, thus obtaining a value $\pi_M$ as in (10). Hence, while it is impossible in this case to get the exact value of the posterior distribution, we can approximate it through an auxiliary simulation. Therefore, it is possible to apply a MCWM algorithm and obtain an approximation of $\pi(\theta)$ in this case as well.

### 1.3.4 How to deal with inadmissible parameter values

Let us consider without loss of generality a one-dimensional real parameter $\theta$ that can assume values only on a subset of $\mathbb{R}$. For instance, let us consider as the parameter space the interval $I = [a, b]$. If a Gaussian proposal function $q(x, y)$ is adopted in the implementation of MH, the unboundedness of the support of the proposal distribution results in a new guess $\vartheta$ which takes values outside $I$ with a non-zero probability. In this case, we choose to adopt as proposal function a *truncated Gaussian distribution*. The new guess $\vartheta$ is generated by $q(\theta^{(i)}, \cdot)$, which is a truncated Gaussian distribution of mean $\theta^{(i)}$ and fixed variance $\sigma$. The analytical expression of $q$ in this case is given by

$$q(x, y; a, b, \sigma) = \frac{1}{\sigma} \frac{\varphi\left((y - x)/\sigma\right)}{\Phi\left((b - x)/\sigma\right) - \Phi\left((a - x)/\sigma\right)}, \tag{12}$$

where we explicitly added the dependence on $a$, $b$ and $\sigma$. In (12) the function $\varphi$ is defined as

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

and $\Phi$ is the standard Gaussian cumulative distribution function, where we assume that if $b = \infty$ then $\Phi((b - x)/\sigma)$ equals one, and if $a = -\infty$ then $\Phi((a - x)/\sigma)$ equals one. Let us remark that this proposal distribution is not symmetric, therefore $\alpha$ in MH has to take into account the ratio between the proposal distribution evaluated in the old and the new guesses of the parameter. Hence, in this case the acceptance probability is given by

$$\alpha(\theta^{(i)}, \vartheta) = \min\left\{ \frac{\pi(\vartheta)\left(\Phi\left((b - \theta^{(i)})/\sigma\right) - \Phi\left((a - \theta^{(i)})/\sigma\right)\right)}{\pi(\theta^{(i)})\left(\Phi\left((b - \vartheta)/\sigma\right) - \Phi\left((a - \vartheta)/\sigma\right)\right)}, 1 \right\}.$$

Let us consider the example of a non-negative random variable $\theta$. In this case, thanks to the symmetry properties of the function $\Phi$, the acceptance probability $\alpha$ simplifies to

$$\alpha(\theta^{(i)}, \vartheta) = \min\left\{ \frac{\pi(\vartheta)\Phi\left(\theta^{(i)}/\sigma\right)}{\pi(\theta^{(i)})\Phi\left(\vartheta/\sigma\right)}, 1 \right\}.$$

**Algorithm 4:** Monitoring convergence.

**Data**: $K \in \mathbb{N}$; $\theta^{(0),i} \in \mathbb{R}^{N_p}, i = 1, \ldots, K$; $N_0 \in \mathbb{N}$; $\bar{\rho} > 1$.

count $= 0$ ;
**while** $\exists\, i \in \{1, 2, \ldots, N_p\} : \rho_i < \bar{\rho}$ **do**
    Initialize $K$ parallel MCMC with starting value $\theta^{(\text{count}),i}$ for $i = 1, \ldots, K$ ;
    Execute $N_0$ iterations of each MCMC algorithm ;
    Get the chains $\Theta_i$ for each MCMC and assemble the mixed chain $\Theta_{\text{mix}}$ ;
    **foreach** *component* $j$ *of* $\theta$ *in* $\mathbb{R}^{N_p}$ **do**
        **foreach** *chain* $\Theta_i$ **do**
            Compute the variance $V_i^j$ of the component $j$ in $\Theta_i$ ;
        **end**
        Compute the mean within chains $V^j_{\text{mean}}$ averaging the variances $V_i^j$ ;
        Compute the variance $V^j_{\text{mix}}$ of the component $j$ in $\Theta_{\text{mix}}$ ;
        Compute the potential scale reduction factor as in (13) ;
    **end**
    count = count + $N_0$ ;
**end**

As far as the practical implementation is concerned, modern programming languages often provide with generators of pseudo-random Gaussian numbers. In order to obtain a truncated Gaussian distribution, a practical procedure could be generating random numbers until a number in the acceptable range is generated.

### 1.3.5 Monitoring convergence

It is unclear from the discussion of all the variants of MCMC discussed above how to choose an optimal number of samples $N$. In other words, no convergence criteria for the Markov chain has been presented. An interesting approach in order to monitor the convergence of MCMC consists in mixing several Markov chains [5]. The main idea consists in starting $K$ parallel MCMC with different initial guesses and check the properties of the single chains with respect to the chain resulting from the mixing of all the single chains. Let us denote by $\Theta_i$, with $i = 1, \ldots, K$ the single chains, and by $\Theta_{\text{mix}}$ the mixed chain, i.e.

$$\Theta_{\text{mix}} = \bigcup_{i=1}^{K} \Theta_i.$$

Let us consider the estimation of the distribution of a random variable $\theta$ with values in $\mathbb{R}^{N_p}$ with components $\theta_1, \ldots, \theta_{N_p}$. Let us moreover assume that all the MCMC algorithms have reached the iteration $N_0$. Then, we compute separately the population variance within each chain for each component $j$ of the random variable and denote it by $V_i^j$, for $i = 1, \ldots, K$ and $j = 1, \ldots, N_p$, i.e.

$$V_i^j = \frac{1}{N_0} \sum_{k=0}^{N_0} \left( \theta_j - \frac{1}{N_0} \sum_{h=0}^{N_0} \theta_j \right)^2 .$$

Then, we average these variances and denote the result as $V^j_{\text{mean}}$. Finally, we compute the population variance for each component of $\theta$ of the mixed chain $\Theta_{\text{mix}}$ and we denote it $V^j_{\text{mix}}$. We now define the *potential scale reduction factor* $\rho_j$ of the $j$-th component of $\theta$ as [5]

$$\rho_j := \sqrt{V^j_{\text{mix}} / V^j_{\text{mean}}}. \tag{13}$$

Let us remark that $\rho_j$ is greater than one for any sample, and a value close to one implies that all the single chains have in averaged mixed as well as the mixed chain. We therefore check whether all the $\rho_j$ are smaller than a certain threshold $\bar{\rho}$ (e.g., $\bar{\rho} = 1.05$) every $N_0$ iterations, and stop the MCMC algorithms if the condition is verified (Algorithm 4), thus keeping as the output of the mixed chain $\Theta_{\text{mix}}$.

## 2 Probabilistic Methods

Several methods have been developed to integrate an Ordinary Differential Equation (ODE) numerically (...). If a deterministic solver is employed, the error introduced by the numerical method is however difficult to quantify (...). In the frame of statistical analyses, quantifying the impact on the uncertainty of the solution due to the numerical approximation is of the utmost importance (...). Therefore, a new class of probabilistic numerical methods have been recently proposed [3] (...).
Let us consider $f\colon \mathbb{R}^d \to \mathbb{R}^d$ and the following ODE

$$\begin{aligned} \frac{\mathrm{d}u(t)}{\mathrm{d}t} &= f(u), \quad t \in (0, T], \\ u(0) &= u_0, \qquad u_0 \in \mathbb{R}^d. \end{aligned} \tag{14}$$

Integrating numerically (14) with a deterministic method on a time discretization $t_k = kh, k = 0, \ldots, N, T = Nh$ gives a numerical solution $U_k, k = 0, \ldots, N$, defined by

$$\begin{aligned} U_{k+1} &= \Psi(U_k), \quad k = 0, \ldots, N-1, \\ U_0 &= u_0, \end{aligned}$$

where $\Psi$ defines one step of a deterministic numerical method to integrate (14).
The idea behind probabilistic methods is adding at each step of the numerical integration a noise component, *i.e.*,

$$\begin{aligned} U_{k+1} &= \Psi(U_k) + \xi_k(h), \\ U_0 &= u_0, \end{aligned} \tag{15}$$

where $\xi_k(h)$ are i.i.d. Gaussian random variables.

### 2.1 Deterministic methods

- Introduction on RK methods in general with definitions
- Implicit and Explicit methods
- Stabilized explicit methods

### 2.2 Motivation

The numerical solution of ODE's has been investigated deeply in the last decades. In particular, the theoretical and numerical analysis of Runge-Kutta methods has involved mathematicians since their discovery.

- Why are punctual solutions not good?

#### 2.2.1 Numerical example

One of the most famous examples of chaotic ODE's is the Lorenz system. This ODE is defined by

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y - x), & x(0) &= -10, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x(\rho - z) - y, & y(0) &= -1, \\ \S\frac{\mathrm{d}z}{\mathrm{d}t} &= xy - \beta z, & z(0) &= 40. \end{aligned} \tag{16}$$

Edward Lorenz discovered in 1963 that for certain values of the parameters $\rho$, $\sigma$ and $\beta$ this equation has a chaotic behavior, i.e., small variations of the initial conditions result in completely different solutions of the system. A sample trajectory of the solution is depicted in Figure **??**. When solving numerically (16), the perturbation induced by the numerical approximation on the true solution imply that the obtained value could be completely different from the exact solution. Hence, in this

case introducing a probabilistic solution can give a clearer indication of the value and uncertainty of the numerical solution.
Let us consider $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. For this value, the solution has indeed a chaotic behavior. We solve the equation with a time step $h = 0.01$ using the probabilistic solver with $\Psi$ the Gauss collocation method on two stages. In Figure **??** we show 10 realizations of the numerical solution given by the probabilistic solver and the numerical solution without any added perturbation from initial time $t = 0$ to final time $T = 20$. We can remark that all the trajectories given by the probabilistic solver coincide with the punctual solution up to a time $\bar{t}$, approximately equal to 12, where the random perturbation force the numerical solution on another trajectory.

## 2.3 Method properties

- Introduce the analysis

### 2.3.1 Strong convergence

In this section we prove a result about strong convergence of the method defined in (15). The following discrete Gronwall lemma is needed in the proof.

**Proposition 2.1** (Discrete Gronwall Lemma)**.** *Let $y_n$ be a nonnegative sequence and $C_1, C_2$ positive constants. If*

$$y_n \leq C_1 + C_2 \sum_{k=0}^{n-1} y_k,$$

*then*

$$y_n \leq C_1 \exp(nC_2).$$

Two assumptions are necessary to prove the strong convergence result. The first assumption is on the noise model.

**Assumption 2.1.**

$$\mathbb{E}^h |\xi_k(t)\xi_k(t)^T|_F^2 \leq Kt^{2p+1}.$$

*Furthermore, there exists a matrix $Q$ independent of $h$ such that*

$$\mathbb{E}^h[\xi_k(h)\xi_h(h)^T] = Qh^{2p+1},$$

*where $p \geq 1$.*

Let us remark that if $Q = \sigma I$, with $I$ the identity matrix in $\mathbb{R}^{d\times d}$ and $\sigma > 0$, the method (15) can be simulated by

$$U_{k+1} = \Psi_h(U_k) + \sqrt{\sigma} h^{p+\frac{1}{2}} Z_k,$$

where $Z_k$ is a Gaussian random vector with independent entries $Z_{k,i} \sim \mathcal{N}(0, 1), i = 1, \dots, d$.
An assumption on the numerical method is needed.

**Assumption 2.2.** *The function $f$ and a sufficient number of its derivatives are bounded uniformly in $\mathbb{R}^n$ in order to ensure that $f$ is globally Lipschitz and that the numerical flow map $\Psi_h$ has uniform local truncation error of order $q + 1$*

$$\sup_{u\in\mathbb{R}^n} |\Psi_t(u) - \Phi_t(u)| \leq Kt^{q+1}.$$

The following result hold.

**Proposition 2.2** (Strong Convergence)**.** *Under assumptions 2.1 and 2.2 it follows that there is $K > 0$ such that*

$$\sup_{0<kh<T} \mathbb{E}^h |u_k - U_K|^2 \leq Kh^{2\min\{p,q\}}.$$

*Furthermore*

$$\sup_{0\leq t\leq T} \mathbb{E}^h |u(t) - U(t)| \leq Kh^{\min\{p,q\}}.$$

This result implies that a reasonable choice in $p$ of Assumption 2.1 is $p = q$.

*Proof.* Given the method in (15) and writing the exact solution of (14) as

$$u_{k+1} = \Phi_h(u_k),$$

one can compute the truncation error $\epsilon_k = \Psi_h(U_k) - \Phi_h(U_k)$, so that

$$U_{k+1} = \Phi_h(U_k) + \epsilon_k + \xi_k(h).$$

Therefore

$$\begin{aligned} e_{k+1} &= u_k - U_k \\ &= \Phi_h(u_k) - \Phi_h(u_k - e_k) - \epsilon_k - \xi_k(h). \end{aligned}$$

Taking the expectation and under Assumption 2.1

$$\mathbb{E}^h|e_{k+1}|^2 = \mathbb{E}^h|\Phi_h(u_k) - \Phi_h(u_k - e_k) - \epsilon_k|^2 + \mathcal{O}(h^{2p+1}).$$

Developing the square and since $\Phi_h$ is Lipschitz continuous with constant $(1+Lh)$ and $\epsilon_k = \mathcal{O}(h^{q+1})$ thanks to Assumption 2.2

$$\begin{aligned} \mathbb{E}^h|e_{k+1}|^2 \leq &(1 + Lh)^2\mathbb{E}^h|e_k|^2 + \mathbb{E}^h \left| \left( h^{\frac{1}{2}}(\Phi_h(u_k) - \Phi_h(u_k - e_k), h^{-\frac{1}{2}}\epsilon_k \right) \right| \\ &+ \mathcal{O}\left(h^{2q+2}\right) + \mathcal{O}\left(h^{2p+1}\right). \end{aligned}$$

Then, using Cauchy-Schwarz on the inner product

$$\begin{aligned} \mathbb{E}^h|e_{k+1}|^2 &\leq (1 + \mathcal{O}(h))\, \mathbb{E}^h|e_k|^2 + \mathcal{O}\left(h^{2q+1}\right) + \mathcal{O}\left(h^{2p+1}\right) \\ &\leq C_1 h \mathbb{E}^h|e_k|^2 + \mathbb{E}|e_k|^2 + \mathcal{O}\left(h^{2q+1}\right) + \mathcal{O}\left(h^{2p+1}\right) \\ &\leq C_1 h \sum_{i=0}^{k} \mathbb{E}^h|e_i|^2 + \mathcal{O}\left(h^{-1}\right)\left(\mathcal{O}\left(h^{2q+1}\right) + \mathcal{O}\left(h^{2p+1}\right)\right) \\ &\leq C_1 h \sum_{i=0}^{k} \mathbb{E}^h|e_i|^2 + \mathcal{O}\left(h^{2q}\right) + \mathcal{O}\left(h^{2p}\right). \end{aligned}$$

Therefore by Proposition 2.1

$$\begin{aligned} \mathbb{E}^h|e_k|^2 &\leq C_2 h^{2\min\{p,q\}} \exp(C_1 kh) \\ &\leq C_2 h^{2\min\{p,q\}} \exp(C_1 T) \\ &\leq C h^{2\min\{p,q\}}. \end{aligned}$$

□

### 2.3.2 Weak convergence

A result of weak convergence can be proved using a technique of *backward error analysis*. The main idea behind this technique is finding a *modified equation* that the numerical method solves exactly or with a higher accuracy than the original equation.
Let us consider (14) and the numerical method (15). Using the Lie derivative notation, it is possible to find the differential operators $\mathcal{L}$ and $\mathcal{L}^h$ such that for all $\varphi \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$

$$\begin{aligned} \varphi(\Phi_h(u)) &= \left(e^{h\mathcal{L}}\varphi\right)(u), \\ \mathbb{E}\varphi(U_1|U_0 = u) &= \left(e^{h\mathcal{L}^h}\varphi\right)(u). \end{aligned} \tag{17}$$

In particular, $\mathcal{L} = f \cdot \nabla$ and the explicit definition of $\mathcal{L}^h$ is not needed in this scope.
We now introduced a modified ODE

$$\frac{\mathrm{d}\hat{u}}{\mathrm{d}t} = f^h(\hat{u}),$$

and a modified SDE

$$\mathrm{d}\tilde{u} = f^h \tilde{u}\,\mathrm{d}t + \sqrt{h^{2p} Q}\,\mathrm{d}W\,, \tag{18}$$

where $p$ has been introduced in Assumption 2.1. We rewrite the solution of these equations in terms of Lie derivatives as for (17) introducing the differential operators $\hat{\mathcal{L}}$ and $\tilde{\mathcal{L}}$, *i.e.*,

$$\varphi\left(\hat{u}(h)|\hat{u}(0) = u\right) = \left(e^{h\hat{\mathcal{L}}^h}\varphi\right)(u),$$
$$\varphi\left(\tilde{u}(h)|\tilde{u}(0) = u\right) = \left(e^{h\tilde{\mathcal{L}}^h}\varphi\right)(u).$$

Therefore,

$$\hat{\mathcal{L}}^h = f^h \cdot \nabla,$$
$$\tilde{\mathcal{L}}^h = f^h \cdot \nabla + \frac{1}{2} h^{2p} Q : \nabla^2,$$

where $\tilde{\mathcal{L}}^h$ is the *generator* of (18). (... all the passages to get to (28) in [3] ...).

**Assumption 2.3.** *The function $f$ in* (14) *is in $\mathcal{C}^\infty$ and all its derivatives are uniformly bounded in $\mathbb{R}^n$. Furthermore, $f$ is such that for all functions $\varphi$ in $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$*

$$\sup_{u \in \mathbb{R}^n} \left| e^{h\mathcal{L}} \varphi(u) \right| \leq (1 + Lh) \sup_{u \in \mathbb{R}^n} |\varphi(u)|\,,$$
$$\sup_{u \in \mathbb{R}^n} \left| e^{h\mathcal{L}^h} \varphi(u) \right| \leq (1 + Lh) \sup_{u \in \mathbb{R}^n} |\varphi(u)|\,,$$

*for some $L > 0$.*

We can now state the following result about weak convergence.

**Proposition 2.3.** *Consider the numerical method* (15) *and Assumptions 2.1, 2.2 and 2.3. Then for any function $\varphi$ in $\mathcal{C}^\infty$ endowed with the properties of Assumption 2.3,*

$$\left|\varphi(u(T)) - \mathbb{E}^h\left(\varphi(U_k)\right)\right| \leq K h^{\min\{2p,q\}}, \quad kh = T,$$

*and*

$$\left|\mathbb{E}\varphi(\tilde{u}(T)) - \mathbb{E}^h\left(\varphi(U_k)\right)\right| \leq K h^{2p+1}, \quad kh = T,$$

*with $u$ and $\tilde{u}$ solutions of* (14) *and* (18).

*Proof.* proof in [3]. ☐

### 2.3.3 Numerical verification of weak order

Let us consider for (14) the FitzHugh-Nagumo model, defined by

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= c\left(x_1 - \frac{x_1^3}{3} + x_2\right), \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= -\frac{1}{c}(x_1 - a + bx_2), \end{aligned} \tag{19}$$

where $a, b, c \in \mathbb{R}$. In particular, we choose $a = 0.2, b = 0.2, c = 3$. We provide the system with the initial condition $x_1(0) = -1, x_2(0) = 1$. We integrate numerically this equation with (15), using RK4 as $\Psi$. Therefore, Assumption 2.2 holds with $q = 4$. Moreover, we consider $Q$ in Assumption 2.1 to be $Q = \sigma I$ with $\sigma = 0.1$. We approximate the solution up to time $T = 10$ with $p$ in Assumption 2.1 equal to $1, 1.5, 2, 2.5$ and $h$ vary in the range $0.25/(2^i), i = 0, \ldots, 3$. We approximate $\mathbb{E}^h\left(\varphi(U_k)\right)$ using a Monte Carlo simulation over 50000 trajectories and compare it with the solution computed on a fine grid to obtain an estimation of the weak error. Results (Figure 2) show that the predicted order $\min\{2p, q\}$ applies in this example. In particular, since $q = 4$, it is possible to notice that no difference in order is detected between the cases $p = 2$ and $p = 2.5$.

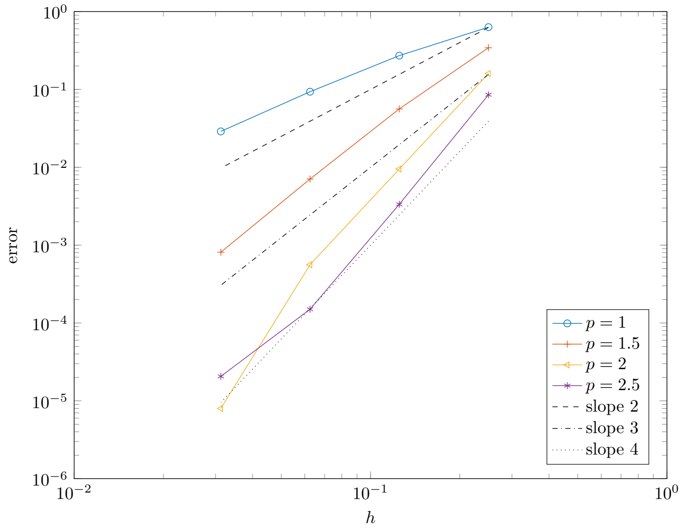



**Figure 2:** Weak order of convergence of (15) applied to (19).

### 2.3.4 Monte Carlo approximation

Let us consider the numerical method introduced in (15) and the Monte Carlo approximation

$$\hat{Z} = \frac{1}{M} \sum_{i=1}^{M} \varphi\left(U_N^{(i)}\right). \tag{20}$$

The mean square error (MSE) of $\hat{Z}$ is given by

$$\begin{aligned} \mathrm{MSE}(\hat{Z}) &= \mathbb{E}\left[\left(\hat{Z} - \varphi\left(u(T)\right)\right)^2\right] \\ &= \mathrm{Var}(\hat{Z}) + \mathbb{E}\left[\hat{Z} - \varphi\left(u(T)\right)\right]^2 \\ &\leq \mathrm{Var}(\hat{Z}) + Ch^{2\min\{2p,q\}}, \end{aligned}$$

where the second term is bounded thanks to proposition 2.3 with $C > 0$. In the standard theory of SDE's, the first term is bounded by $CM^{-1}$, where $C$ is a positive constant. In the probabilistic solver we consider in this work, it is possible to bound the first term with a function of the time step $h$. Intuitively, this favorable property comes from the fact that the noise scale is of the same order of magnitude of the time step.

**Lemma 2.1.** *Consider the numerical method* (15) *applied to a one-dimensional ODE with* $\Psi$ *any explicit Runge-Kutta method on* $s$ *stages and Assumption 2.1. Then the numerical solution* $U_k$ *at time* $t_k = kh$ *satisfies*

$$\mathrm{Var}(U_k) \leq C_1 \mathrm{Var}(U_0) + C_2 \sigma Q h^{2p}, \quad k = 1, \ldots, N,$$

*with* $C_1$, $C_2$ *positive constants.*

*Proof.* Let us consider as the numerical integrator $\Psi$ the Explicit Euler method and $p = 1$, coherently with the common choice that $p$ is equal to the order or the deterministic integrator. Then, thanks to Assumption 2.1 we can write the numerical solution $U_{k+1}$ as

$$U_{k+1} = U_k + hf(U_k) + h^{3/2}(\sigma Q)^{1/2} Z_k,$$

with $Z_k$ a zero-mean random variable independent of $U_k$ such that $\mathrm{Var}(Z_k) = 1$. Then

$$\begin{aligned} \mathrm{Var}(U_{k+1}) &= \mathrm{Var}(U_k + hf(U_k)) + h^3 \sigma Q \\ &\leq 2\,\mathrm{Var}(U_k) + 2h^2\,\mathrm{Var}(f(U_k)) + h^3 \sigma Q, \end{aligned} \tag{21}$$

where we exploited that for any random variables $X$, $Y$,

$$\operatorname{Var}(X+Y) \leq 2\operatorname{Var}(X) + 2\operatorname{Var}(Y). \tag{22}$$

Since $f$ is Lipschitz continuous with constant $C_L$, we can bound the second term in the sum above as

$$\begin{aligned}
\operatorname{Var}(f(U_k)) &= \operatorname{Var}(f(U_k) - f(\mathbb{E}[U_k])) \\
&\leq \mathbb{E}\left[(f(U_k) - f(\mathbb{E}[U_k]))^2\right] \\
&\leq C_L^2 \mathbb{E}\left[(U_k - \mathbb{E}(U_k))^2\right] \\
&= C_L^2 \operatorname{Var}(U_k).
\end{aligned} \tag{23}$$

Hence, we find

$$\begin{aligned}
\frac{1}{2}\operatorname{Var}(U_{k+1}) &\leq (1 + C_L^2 h^2)\operatorname{Var}(U_k) + \frac{1}{2}\sigma Q h^3 \\
&\leq (1 + C_L^2 h^2)^2 \operatorname{Var}(U_{k-1}) + \frac{1}{2}\sigma Q h^3 (1 + C_L^2 h^2) + \frac{1}{2}\sigma Q h^3 \\
&\leq (1 + C_L^2 h^2)^k \operatorname{Var}(U_0) + \frac{1}{2}\sigma Q h^3 \sum_{i=0}^{k-1} (1 + C_L^2 h^2)^i \\
&\leq \exp\big(C_L^2 T^2\big)\operatorname{Var}(U_0) + \frac{1}{2}\sigma Q h^3 \sum_{i=0}^{k-1} (1 + C_L^2 h^2)^i \\
&= C_1 \operatorname{Var}(U_0) + \frac{1}{2}\sigma Q h^3 \sum_{i=0}^{k-1} (1 + C_L^2 h^2)^i
\end{aligned}$$

We then bound the second term as follows

$$\begin{aligned}
\sigma Q h^3 \sum_{i=0}^{k-1} (1 + C_L^2 h^2)^i &= \frac{(1 + C_L^2 h^2)^k - 1}{h^2 C_L^2} \sigma Q h^3 \\
&\leq \frac{\exp\big(k C_L^2 h^2\big) - 1}{C_L^2} \sigma Q h \\
&\leq \frac{\exp\big(T C_L^2 h\big) - 1}{C_L^2} \sigma Q h \\
&= \frac{\sigma Q h}{C_L^2} \sum_{i=1}^{\infty} \frac{(T C_L^2 h)^i}{i!} \\
&\leq \left( \frac{1}{C_L^2 T} \sum_{i=1}^{\infty} \frac{(T^2 C_L^2)^i}{i!} \right) \sigma Q h^2 \\
&= \frac{\exp\big(T^2 C_L^2\big) - 1}{C_L^2 T} \sigma Q h^2 \\
&= C_2 \sigma Q h^2.
\end{aligned} \tag{24}$$

Thus, the result is proved for Explicit Euler. Let us consider now any explicit Runge-Kutta method $\Psi$, and let us rewrite (15) as

$$U_{k+1} = U_k + h\tilde{\Psi}(U_k) + h^{p+1/2}(\sigma Q)^{1/2} Z_k,$$

where $\tilde{\Psi}(x) \coloneqq h^{-1}(\Psi(x) - x)$ is given by

$$\tilde{\Psi}(U_k) = \sum_{i=1}^{s} b_i K_i,$$

and $K_i$, $i = 1, \dots, s$, are the stages of the Runge-Kutta method. Then, proceeding as above

$$\operatorname{Var}(U_{k+1}) \leq 2\operatorname{Var}(U_k) + 2h^2 \operatorname{Var}(\tilde{\Psi}(U_k)) + \sigma Q h^{2p+1}.$$

Let us consider the second term. A direct bound, following from a generalization on $s$ terms of (22) is

$$\operatorname{Var}(\tilde{\Psi}(U_k)) \leq s \sum_{i=1}^{s} b_i^2 \operatorname{Var}(K_i). \tag{25}$$

Hence, we can consider the variance of each stage singularly. Since we are only considering explicit Runge-Kutta method, it is possible to estimate the single variances recursively

$$\begin{aligned}
\operatorname{Var}(K_1) &= \operatorname{Var}(f(U_k)) \leq C_L^2 \operatorname{Var}(U_k), \\
\operatorname{Var}(K_2) &= \operatorname{Var}(f(U_k + ha_{21}K_1)) \leq C_L^2 \operatorname{Var}(U_k + ha_{21}K_1) \\
&\leq 2C_L^2(\operatorname{Var}(U_k) + h^2 a_{21}^2 \operatorname{Var}(K_1)) \\
&\leq 2C_L^2(1 + T^2 a_{21}^2 C_L^2) \operatorname{Var}(U_k) \leq C \operatorname{Var}(U_k) \\
\implies \operatorname{Var}(K_i) &\leq \operatorname{Var}(f(U_k + h \textstyle\sum_{j=1}^{i-1} a_{ij}K_j)) \leq C \operatorname{Var}(U_k), \ \forall i = 2, \dots, s,
\end{aligned}$$

where $C$ is a positive varying from one line to another depending on $C_L$, $T$ and the coefficients of the Runge-Kutta method. We then substitute in (25) and get

$$\operatorname{Var}(U_{k+1}) \leq 2(1 + Csh^2 \textstyle\sum_{i=1}^{s} b_i^2) \operatorname{Var}(U_k) + \sigma Q h^{2p+1}.$$

Finally, we can proceed as explained above in detail in the case of Explicit Euler and recur over $k$ to obtain the desired bound. □

We now consider the same estimation for an implicit Runge-Kutta method

**Lemma 2.2.** *Consider the numerical method* (15) *applied to a one-dimensional ODE with* $\Psi$ *any explicit or implicit Runge-Kutta method on* $s$ *stages and Assumption 2.1. Then, if* $h$ *is small enough, the numerical solution* $U_k$ *at time* $t_k = kh$ *satisfies*

$$\operatorname{Var}(U_k) \leq C_1 \operatorname{Var}(U_0) + C_2 \sigma Q h^{2p}, \quad k = 1, \dots, N,$$

*with* $C_1$, $C_2$ *positive constants.*

*Proof.* Let us consider as $\Psi$ the Implicit Euler method and $p = 1$. Then, we can write one step of the probabilistic method as

$$U_k = U_{k-1} + hf(U_k) + (\sigma Q)^{1/2} h^{3/2} Z_k.$$

Applying (21) and (23) we get

$$\operatorname{Var}(U_k) \leq 2 \operatorname{Var}(U_{k-1}) + 2h^2 C_L^2 \operatorname{Var}(U_k) + \sigma Q h^3.$$

Hence, defining the coefficient $\beta > 0$ as

$$\beta = \frac{1}{1 - 2h^2 C_L^2},$$

and if the time step $h$ is bounded by

$$h < \frac{1}{\sqrt{2} C_L},$$

then $\beta^{-1} > 0$ and we can deduce

$$\begin{aligned}
\frac{1}{2} \operatorname{Var}(U_k) &\leq \beta(\operatorname{Var}(U_{k-1}) + \frac{1}{2} \sigma Q h^3) \\
&\leq \beta^k \operatorname{Var}(U_0) + \frac{1}{2} \left( \sum_{i=1}^{k} \beta^i \right) \sigma Q h^3 \\
&\leq \beta^N \operatorname{Var}(U_0) + \frac{1}{2} T \beta^N \sigma Q h^2,
\end{aligned}$$

which proves the result for the Implicit Euler method. For any implicit or explicit Runge-Kutta method we can write one step of the probabilistic method as

$$U_k = U_{k-1} + h\sum_{i=1}^{s} b_i K_i + (\sigma Q)^{1/2} h^{p+1/2} Z_k.$$

Then thanks to (22) and (23) we obtain

$$\operatorname{Var}(U_k) \leq 2\operatorname{Var}(U_{k-1}) + 2h^2 \operatorname{Var}(\textstyle\sum_{i=1}^{s} b_i K_i) + \sigma Q h^{2p+1}. \tag{26}$$

Let us consider the second term in the bound above. Thanks to the generalization on $s$ terms of (22) we get

$$\operatorname{Var}(\textstyle\sum_{i=1}^{s} b_i K_i) \leq s\sum_{i=1}^{s} b_i^2 \operatorname{Var}(K_i).$$

Considering now the variance of all single stages of the Runge-Kutta scheme, we can exploit (23) and (22) to get

$$\begin{aligned}
\operatorname{Var}(K_i) &= \operatorname{Var}(f(U_{k-1} + h\textstyle\sum_{j=1}^{s} a_{ij}K_j)) \\
&\leq C_L^2 \operatorname{Var}(U_{k-1} + h\textstyle\sum_{j=1}^{s} a_{ij}K_j) \\
&\leq 2C_L^2 \operatorname{Var}(U_{k-1}) + 2C_L^2 h^2 \operatorname{Var}(\textstyle\sum_{j=1}^{s} a_{ij}K_j) \\
&\leq 2C_L^2 \operatorname{Var}(U_{k-1}) + 2C_L^2 h^2 s \max_{i,j=1,\dots,s} a_{ij}^2 \textstyle\sum_{j=1}^{s} \operatorname{Var}(K_j).
\end{aligned}$$

Let us define the constant $\alpha > 0$ as

$$\alpha = 2C_L^2 h^2 s \max_{i,j=1,\dots,s} a_{ij}^2.$$

Then, if the time step $h$ satisfies

$$h < \frac{1}{C_L}\left(\frac{1}{2s\max_{i,j=1,\dots,s} a_{ij}^2}\right)^{1/2},$$

we have that $1-\alpha$ is positive and therefore we can bound the variance of the $i$-th Runge-Kutta stage as

$$\operatorname{Var}(K_i) \leq \frac{2C_L^2}{1-\alpha}\operatorname{Var}(U_{k-1}) + \frac{\alpha}{1-\alpha}\textstyle\sum_{j=1, j\neq i}^{s} \operatorname{Var}(K_j).$$

If for each $i$ we consider a numbering of the Runge-Kutta stages such that $i = s$, we can rewrite the inequality above as

$$\operatorname{Var}(K_s) \leq \frac{2C_L^2}{1-\alpha}\operatorname{Var}(U_{k-1}) + \frac{\alpha}{1-\alpha}\textstyle\sum_{j=1}^{s-1} \operatorname{Var}(K_j).$$

Therefore we can apply the discrete Gronwall inequality (Proposition 2.1) and get

$$\operatorname{Var}(K_i) \leq \frac{2C_L^2}{1-\alpha}\operatorname{Var}(U_{k-1})\exp\left(\frac{\alpha(s-1)}{1-\alpha}\right).$$

Substituting this inequality in (26) we get

$$\begin{aligned}
\frac{1}{2}\operatorname{Var}(U_k) &\leq \left(1 + h^2 s\frac{2C_L^2}{1-\alpha}\exp\left(\frac{\alpha(s-1)}{1-\alpha}\right)\textstyle\sum_{i=1}^{s} b_i^2\right)\operatorname{Var}(U_{k-1}) + \frac{1}{2}\sigma Q h^{2p+1} \\
&\leq \left(1 + h^2 s\frac{2C_L^2}{1-\alpha}\exp\left(\frac{\alpha(s-1)}{1-\alpha}\right)\max_{i=1,\dots,s} b_i^2\right)\operatorname{Var}(U_{k-1}) + \frac{1}{2}\sigma Q h^{2p+1}.
\end{aligned}$$

If we define the constant $\hat{C} > 0$ as

$$\hat{C} := s\frac{2C_L^2}{1-\alpha}\exp\left(\frac{\alpha(s-1)}{1-\alpha}\right)\max_{i=1,\dots,s} b_i^2,$$

we get

$$\frac{1}{2}\operatorname{Var}(U_k) \leq (1+\hat{C}h^2)^k \operatorname{Var}(U_0) + \frac{1}{2}\sigma Q h^{2p+1} \textstyle\sum_{i=0}^{k-1}(1+\hat{C}h^2)^i.$$

For the second term we proceed as in (24) and get for a constant $\tilde{C} > 0$

$$\begin{aligned}\frac{1}{2}\operatorname{Var}(U_k) &\leq (1+\hat{C}h^2)^k \operatorname{Var}(U_0) + \tilde{C}\sigma Q h^{2p} \\ &\leq \exp\left(\hat{C}T^2\right)\operatorname{Var}(U_0) + \tilde{C}\sigma Q h^{2p},\end{aligned}$$

thus obtaining the desired result. □

*Remark* 2.1. Let us remark that in the limit for $h$ going to zero, the coefficient $\alpha$ defined for the Implicit Euler method tends to one. Therefore, asymptotically the variance of the numerical solution is bounded independently of the Lipschitz constant defining the ODE. Conversely, for any explicit Runge-Kutta method the constant depends on $C_L$ for any value of $h$.

*Remark* 2.2. The requirement on the time step $h$ of Lemma 2.2 is reasonable because it is required by the numerical method for its well-posedness. Let us denote by $F$ the function defining one step of the probabilistic Implicit Euler, i.e.,

$$F(X) = U_{k-1} + hf(X) + (\sigma Q)^{1/2} h^{3/2} Z_k.$$

In order to apply Banach's fixed point theorem and therefore admit the existence of a fixed point $U_k$, $F$ has to be a contraption. Therefore, evaluating $F$ on two points $X$ and $Y$, we get

$$\|F(X) - F(Y)\| = h\|f(X) - f(Y)\| \leq hC_L\|X - Y\|.$$

Hence, we have to impose $h < 1/C_L$, which is the same requirement of Lemma 2.2.

We can now consider the MSE of the estimator $\hat{Z}$ introduced in (20).

**Proposition 2.4.** *Under the assumptions of Lemma 2.1 and if $\varphi$ is Lipschitz continuous with constant $C_L$, the following bound for the MSE of $\hat{Z}$ is valid*

$$\operatorname{MSE}(\hat{Z}) \leq C_1 h^{2\min\{2p,q\}} + \frac{C_2}{M}(\operatorname{Var}(U_0) + h^{2p}).$$

*Proof.* The samples $U_N^{(i)}$ are independent and identically distributed as $U_N$, hence

$$\begin{aligned}\operatorname{Var}(\hat{Z}) &= \operatorname{Var}\left(\frac{1}{M}\sum_{i=1}^{M}\varphi\left(U_N^{(i)}\right)\right) \\ &= \frac{1}{M^2}\sum_{i=1}^{M}\operatorname{Var}\left(\varphi\left(U_N\right)\right) \\ &= \frac{1}{M}\operatorname{Var}\left(\varphi\left(U_N\right)\right)\end{aligned}$$

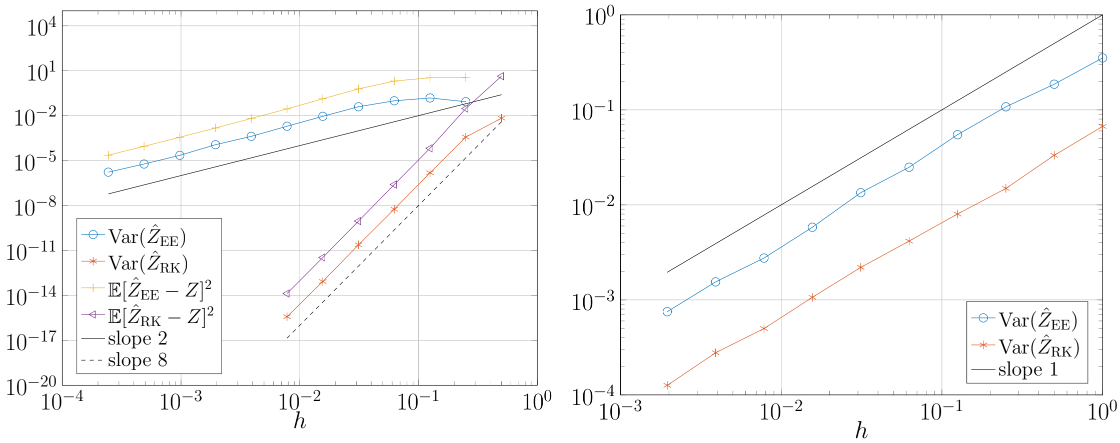
Since the function $\varphi$ is Lipschitz continuous we can use (23) and Lemma 2.1 and get

$$\operatorname{Var}(\hat{Z}) \leq \frac{C}{M}\operatorname{Var}(U_N) \leq \frac{C}{M}(\operatorname{Var}(U_0) + h^{2p})$$

thus obtaining the following bound for the MSE of $\hat{Z}$

$$\operatorname{MSE}(\hat{Z}) \leq C_1 h^{2\min\{2p,q\}} + \frac{C_2}{M}(\operatorname{Var}(U_0) + h^{2p})$$

□




**(a)** Variation of the time step.

**(b)** Variation of the number of trajectories.

**Figure 3:** Variance and squared bias of the Monte Carlo estimator $\hat{Z}$ with Explicit Euler and RK4 applied to (19). The two components of the MSE have the same order of convergence with respect to the time step $h$. Conversely, the order of convergence with respect to the number of trajectories $M$ with fixed $h$ of the variance of $\hat{Z}$ is equal to one for both methods

*Remark* 2.3. Let us remark that in case the initial condition $U_0$ is a known deterministic value, i.e., $\mathrm{Var}(U_0)$ is equal to zero, and the noise scale $p$ is chosen equal to the order of the Runge-Kutta integrator $q$, the bound of the MSE can be rewritten simply as

$$\mathrm{MSE}(\hat{Z}) \leq C_1 h^{2q} + C_2 \frac{h^{2q}}{M}.$$

*Remark* 2.4. In [3] the authors argue that a reasonable choice for the noise scale $p$ is the order of the deterministic solver $q$, thus for a deterministic initial condition the result above is valid. This result is extremely favourable from the point of view of computational cost. Let us assume that the a tolerance $\varepsilon$ and that the numerical error is measured by means of the square root of the MSE. Then, in order to attain this tolerance we have to impose

$$h = \mathcal{O}(\varepsilon^{1/q}),$$

without any condition on the number of trajectories $M$, which we can consider to be $\mathcal{O}(1)$. Hence, the computational cost is

$$\mathrm{cost} = \frac{M}{h} = \mathcal{O}(\varepsilon^{-1/q}).$$

### 2.3.5 Numerical experiment

We consider the FitzHug-Nagumo problem introduced in (19) with the same initial conditions and parameter values and integrate it up to the final time $T = 10$ with the probabilistic integrator. We choose the function $\varphi$ to be given by $\varphi(X) = X^T X$ and generate a reference solution $Z$ with RK4 computed on a fine time step. We choose as deterministic integrator Explicit Euler and RK4 and the noise scale $p$ equal to $q$, i.e., one and four respectively. We choose $M = 10$ and the time step $h = 0.5/2^i$ with $i = 0, 1, \ldots, 11$. Then we compute 300 times the estimator $\hat{Z}$ for all the values of the time step, thus estimating its variance and bias. Numerical results (Figure 4) confirm the theoretical bound presented in Lemma 2.1, as the order of convergence of the variance of $\hat{Z}$ to zero is of order 2 and 8 with respect to $h$ for Explicit Euler and RK4 respectively independently of $M$. We perform another experiment fixing the value of $h$ to 0.5 and varying the number of trajectories in the values $M = 2^i$ with $i = 0, 1, \ldots, 9$. As in the first experiment, we compute 300 times $\hat{Z}$ in order to estimate its variance. Results (Figure 3b) show that the variance has an order equal to 1 for both the methods with respect to $M^{-1}$, thus confirming the theoretical result.

## 2.4 Stability analysis

In the previous sections we analyzed the behavior of the probabilistic integrator for ODE's in terms of its convergence with respect to the time step. Moreover, we analyzed the convergence of a Monte Carlo approximation of the probabilistic solution towards the exact solution. Another key feature of numerical methods is their stability. We recall that the one step of the numerical method can be written, under Assumption 2.1, as

$$U_{k+1} = \Psi(U_k, U_{k+1}) + \sqrt{\sigma Q} h^{p+1/2} Z_k,$$

with $Z_k$, $k = 0, \dots, N$ i.i.d. zero-mean normal random variables with unitary variance. We can write therefore the numerical method as

$$U_{k+1} = \Psi(U_k, U_{k+1}) + \sqrt{\sigma Q} h^p \Delta W_k,$$

where the random variables $\Delta W_k$, $k = 0, \dots, N$, are standard Wiener increments. This is the stochastic Runge-Kutta method applied to the SDE

$$\mathrm{d}U(t) = f(U(t))\mathrm{d}t + \sqrt{\sigma Q} h^p \mathrm{d}W(t).$$

## 2.5 To do at Christmas

- Repeat the computations for MLMC in case $\mathrm{Var}(U_0) = \mathcal{O}(1)$.

# 3 Bayesian inference of the parameters of an ODE

We now consider the probabilistic solver introduced in the previous section and study its behavior when used in the context of Bayesian inference. Let us consider $\theta$ a parameter in $\mathbb{R}^{N_p}$, a function $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ depending on the parameter, an element $u_0$ of $\mathbb{R}^d$ and the following ODE

$$\begin{aligned} \frac{\mathrm{d}u_\theta}{\mathrm{d}t}(t) &= f_\theta(u_\theta(t)), \\ u_\theta(0) &= u_0, \end{aligned} \tag{27}$$

where we write explicitly the dependence of the solution $u$ on the parameter. In the following, we consider $\theta$ to be unknown a priori, and we denote by $\bar{\theta}$ its true value. Moreover, let us consider a set of observed data $\mathcal{Y}_i$ defined as

$$\mathcal{Y}_i = \{y_1, y_2, \dots, y_i\}, \quad i = 1, \dots, D,$$

where $D$ is the total number of observations. We consider the data to be a Gaussian linear function of the exact solution of (27) computed at the true value of $\theta$ at a discrete set of times $t_i$, i.e.

$$y_i = u_{\bar{\theta}}(t_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma), \quad i = 1, \dots, D.$$

If the solution of (27) is computable analytically, then thanks to Bayes theorem we know that once a prior distribution $\mathcal{Q}(\theta)$ is specified, the posterior distribution of $\theta$ is given by Bayes' formula and can be expressed as

$$\pi(\theta|\mathcal{Y}) \propto \mathcal{Q}(\theta)\mathcal{L}(\mathcal{Y}|u_\theta(t)). \tag{28}$$

Under the hypothesis that the observational error is normally distributed, the likelihood function is easy to compute and is given by

$$\mathcal{L}(\mathcal{Y}|u_\theta(t)) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{D} (u_\theta(t_i) - y_i)^T \Gamma^{-1} (u_\theta(t_i) - y_i)\right).$$

In most of applications, the exact solution of (27) is not computable in closed form, and therefore Bayes rule cannot be applied directly as in (28). Therefore, a MCMC technique has to be applied to obtain an estimation of the parameter.

## 3.1 Approximation of the likelihood

An unbiased estimator of the likelihood has to be obtained at each step of the MCMC algorithm in order to compute the acceptance probability. In particular, we approximate the likelihood using the probabilistic solver with time step $h$, thus obtaining the approximation a Monte Carlo estimation obtained with $M$ simulated trajectories with time step $h$, i.e., for each value of $\theta$ we consider

$$\mathcal{L}(\mathcal{Y}|\theta) \approx \mathcal{L}^h(\mathcal{Y}|\theta).$$

Then, we approximate the value of $\mathcal{L}^h$ with a Monte Carlo simulation thus obtaining the following unbiased estimator

$$\mathcal{L}^h_M(\mathcal{Y}|\theta) \approx \mathcal{L}^h(\mathcal{Y}|\theta).$$

We then use this value for computing the acceptance probability in a MWCM algorithm (see Section 1.3.3) to perform Bayesian inference on the value of the parameter $\theta$. A question which often arises in literature [1, 4, 11] is how many samples $M$ it would be advisable to choose in order to consider the obtained posterior distribution a good approximation of the true posterior. For each value of $\theta$, we can apply Proposition 2.4, thus obtaining

$$\mathrm{MSE}(\mathcal{L}^h_M(\mathcal{Y}|\theta)) \le C_1 h^{2\min\{2p,q\}} + \frac{C_2}{M} h^{2p}.$$

Therefore, at each step of the MCMC algorithm we can control with the time step the goodness of the likelihood estimator.

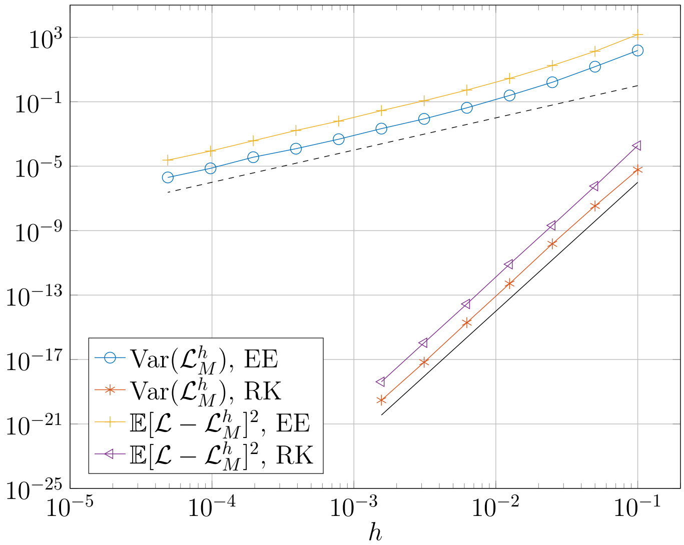



**Figure 4:** Approximation of the likelihood at a fixed value $\bar{\theta}$.

### 3.1.1 Numerical example

We consider the FitzHug-Nagumo problem defined in (19) and the parameters fixed to the true value $\bar{\theta} = (0.2, 0.2, 3.0)$. We generate ten equispaced observations from initial time $t = 0$ to final time $T = 1$ adding a zero-mean Gaussian perturbation with variance equal to $10^{-2}$ to the two components of a numerical solution computed with a small time step. Then we generate a reference solution using the same small time step without noise in order to have an approximation of $\mathcal{L}(\mathcal{Y}|\bar{\theta})$ with negligible error. Then we compute 300 realizations of $\mathcal{L}_h^M(\mathcal{Y}|\bar{\theta})$ using Euler Forward or RK4 as a deterministic integrator with time steps $h$ in the set $h = 0.1 \cdot 2^{-i}$ with $i = 0, \ldots, 11$, for Euler Forward and $h = 0.1 \cdot 2^{-i}$ for $i = 0, \ldots, 6$, for RK4. Hence, we estimate the bias of $\mathcal{L}(\mathcal{Y}|\bar{\theta})$ with respect to the true value of the likelihood and its variance.

## 3.2 Numerical example

Let us consider the two-dimensional FitzHug-Nagumo ODE defined in (19) and the problem of determining the values of the parameters $\theta = (a, b, c)^T$ in $\mathbb{R}^3$. We consider as the true value of $\theta$ the vector $\bar{\theta} = (0.2, 0.2, 3)$. We produce a set of synthetic observations $\mathcal{Y}_{10}$ from a numerical solution $\tilde{u}$ computed using $\bar{\theta}$ and a small time step at times $t_i = 1, 2, ..., 10$, with an additive independent Gaussian noise, i.e.,

$$y_i = \tilde{u}_{\bar{\theta}}(t_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10^{-2} I), \quad i = 1, \ldots, 10,$$

where $I$ is the identity matrix in $\mathbb{R}^{2\times 2}$. Therefore, we consider a diagonal noise with independent normal components having all variance $10^{-2}$. We approximate the posterior distribution $\pi(\theta|\mathcal{Y})$ with both the deterministic and the probabilistic solvers using time step $h = 0.1$. We use the RAM algorithm for the deterministic case and the RAM algorithm applied to MCWM for the probabilistic integrator. In both cases, the proposal distribution $q(x, y)$ is a Gaussian with variance adapted by RAM, and the prior distribution $\mathcal{Q}(\theta)$ is normal with unitary variance and mean $\bar{\theta}$. We consider 50000 iterations of MCMC in both the deterministic and the probabilistic case, with the first 10% of guesses considered as a burn-in. Results (Figure 5) show that the posterior distribution obtained using the deterministic solver is concentrated and biased with respect to the true value of the parameter. On the other side, the probabilistic solver provides with a posterior distribution having a wider support which contains the true value of the parameter. This confirms the claim that the probabilistic solver allows to identify the source of error given by the numerical integration, while in the deterministic case this uncertainty does not result from the obtained distributions.

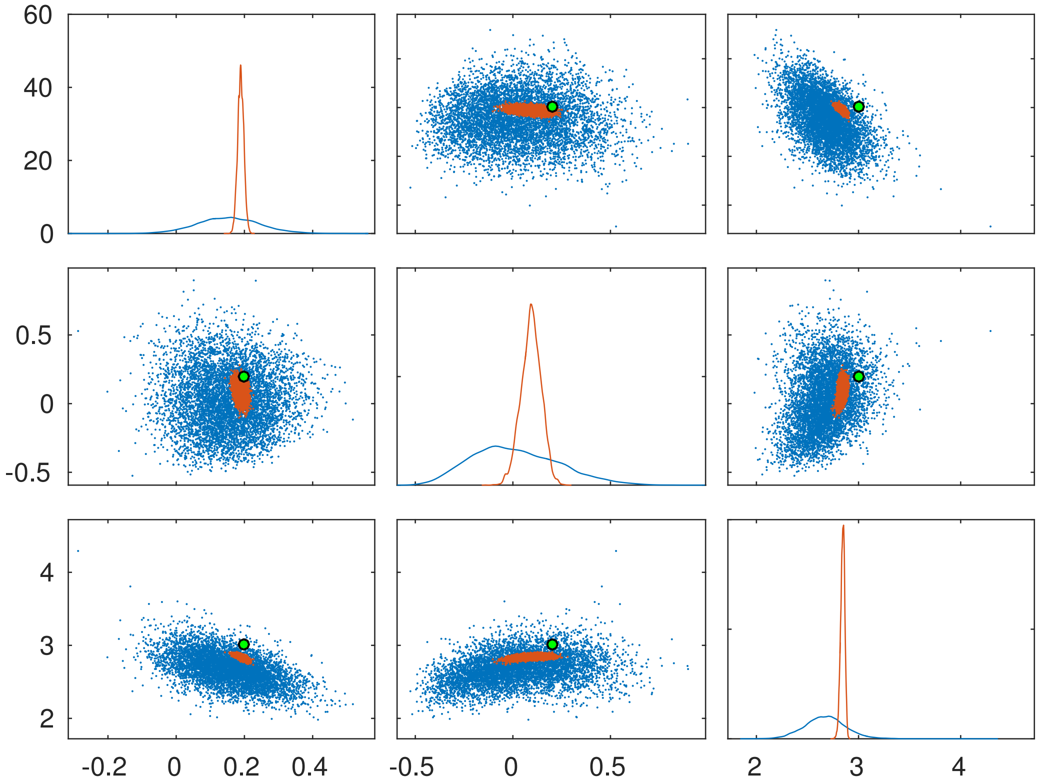



**Figure 5:** Posterior distribution for the parameter $\theta$ defining the FitzHug-Nagumo model. The posterior distributions given by the probabilistic and the deterministic solvers are displayed in blue and red respectively. The true value of the parameters is displayed in thick green dots.

### 3.3 Convergence of the posterior distribution

We wish to study the convergence of the posterior distribution obtained using the probabilistic method with respect to the true posterior distribution. In order to study the convergence, it is necessary to introduce a notion of distance between two probability measures. A standard measure is the *total variation distance*, defined in the following.

**Definition 3.1.** *Given two probability measures $\nu$ and $\mu$ on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, the total variation distance between $\nu$ and $\mu$ is defined as*

$$d_{\mathrm{TV}}(\nu, \mu) := \sup_{A \in \mathcal{B}(\mathcal{X})} |\nu(A) - \mu(A)| .$$

*Moreover, if $\nu$ and $\mu$ admit a densities $f$ and $g$ respectively with respect to a dominating measure $\lambda$, then the total variation distance can be expressed as*

$$d_{\mathrm{TV}}(\nu, \mu) := \frac{1}{2} \int_{\mathcal{X}} (f(x) - g(x)) \mathrm{d}\lambda(x).$$

Other notions of distance can be employed when the total variation distance is not practical to compute, such as the Hellinger distance, which is defined as follows [6].

**Definition 3.2.** *If $f, g$ are densities of the measures $\mu$ and $\nu$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with respect to the Lebesgue measure, the Hellinger distance between $\nu$ and $\mu$ is defined as*

$$d^2_{\mathrm{Hell}}(\mu, \nu) := \frac{1}{2} \int_{\mathbb{R}^n} \left( \sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mathrm{d}x$$

The Hellinger distance allows us to estimate the total variation distance as the following inequalities hold [6]

$$\frac{d^2_{\mathrm{Hell}}(\mu, \nu)}{2} \leq d_{\mathrm{TV}}(\mu, \nu) \leq d_{\mathrm{Hell}}(\mu, \nu). \tag{29}$$

Let us consider the posterior distribution given by MCMC with approximated likelihood. We can compute the second moment of the Hellinger distance as

$$\begin{aligned}
2\mathbb{E}^{\xi} \left[ d^2_{\mathrm{Hell}}(\pi(\theta|\mathcal{Y}), \pi_h^M(\theta|\mathcal{Y})) \right] &= \mathbb{E}^{\xi} \left[ \int \left( \sqrt{\pi(\theta|\mathcal{Y})} - \sqrt{\pi_h^M(\theta|\mathcal{Y})} \right)^2 \mathrm{d}\theta \right] \\
&= \mathbb{E}^{\xi} \left[ \int \left( \sqrt{\mathcal{Q}(\theta)\mathcal{L}(\mathcal{Y}|\theta)} - \sqrt{\mathcal{Q}(\theta)\mathcal{L}_h^M(\mathcal{Y}|\theta)} \right)^2 \mathrm{d}\theta \right] \\
&= \int \mathbb{E}^{\xi} \left[ \left( \sqrt{\mathcal{L}(\mathcal{Y}|\theta)} - \sqrt{\mathcal{L}_h^M(\mathcal{Y}|\theta)} \right)^2 \right] \mathrm{d}\mathcal{Q}(\theta) \\
&= \int \mathrm{MSE} \left( \sqrt{\mathcal{L}_h^M(\mathcal{Y}|\theta)} \right) \mathrm{d}\mathcal{Q}(\theta) \\
&\leq \int C(\theta) h^{2q} \mathrm{d}\mathcal{Q}(\theta) \\
&= h^{2q} \int C(\theta) \mathrm{d}\mathcal{Q}(\theta),
\end{aligned}$$

where $C(\theta)$ is the constant appearing in Proposition 2.4. Let us remark that the constant depends on the Lipschitz constant of the function defining the ODE, which depends non-trivially on $\theta$. Finally, defining

$$\tilde{C} := \sqrt{\frac{1}{2} \int C(\theta) \mathrm{d}\mathcal{Q}(\theta)},$$

we get the following bound on the second moment of the Hellinger distance between the approximated posterior and the posterior obtained with the exact solution

$$\mathbb{E}^{\xi} \left[ d^2_{\mathrm{Hell}}(\pi(\theta|\mathcal{Y}), \pi_h^M(\theta|\mathcal{Y})) \right] \leq \tilde{C}^2 h^{2q}.$$

Then, thanks to Jensen's inequality

$$\mathbb{E}^{\xi}\left[d_{\mathrm{Hell}}(\pi(\theta|\mathcal{Y}), \pi_h^M(\theta|\mathcal{Y}))\right] \leq \mathbb{E}^{\xi}\left[d_{\mathrm{Hell}}^2(\pi(\theta|\mathcal{Y}), \pi_h^M(\theta|\mathcal{Y}))\right]^{1/2}$$
$$\leq \tilde{C}h^q.$$

Let us remark that thanks to (29), this bound is equally true for the total variation distance.

### 3.4 Convergence of the Monte Carlo estimation

We are now interested in the convergence of the expectation of the parameter $\theta$ inferred by MCMC. Let us consider a function $g\colon \mathbb{R}^{N_p} \to \mathbb{R}$ such that $g \in L^{\infty}(\mathbb{R}^{N_p})$. Then, we wish to bound the distance between the expectation of $g(\theta)$ computed with respect to the true measure and to the measure targeted by MCMC implemented with the probabilistc solver and time step $h$. Thanks to the previous result on the total variation distance we get

$$\begin{aligned}
\mathbb{E}^{\xi}\left|\mathbb{E}^{\pi}\left[g(\theta)\right] - \mathbb{E}^{\pi_h^M}\left[g(\theta)\right]\right| &= \mathbb{E}^{\xi}\left|\int g(\theta)(\pi(\theta|\mathcal{Y}) - \pi_h^M(\theta|\mathcal{Y}))\mathrm{d}\theta\right| \\
&\leq \|g\|_{\infty}\mathbb{E}^{\xi}\left[\int \left|\pi(\theta|\mathcal{Y}) - \pi_h^M(\theta|\mathcal{Y})\right|\mathrm{d}\theta\right] \\
&= 2\|g\|_{\infty}\mathbb{E}^{\xi}\left[d_{\mathrm{TV}}(\pi, \pi_h^M)\right] \\
&\leq 2\|g\|_{\infty}Ch^q.
\end{aligned}$$

We can now compute the variance of the expectation of $g(\theta)$ computed MCMC and the probabilistic integrator as

$$\begin{aligned}
\mathrm{Var}^{\xi}(\mathbb{E}^{\pi_h^M}\left[g(\theta)\right]) &= \mathrm{Var}^{\xi}\left(\int g(\theta)\pi_h^M(\theta)\mathrm{d}\theta\right) \\
&= \mathrm{Var}^{\xi}\left(\int g(\theta)\mathcal{L}_h^M(\mathcal{Y}|\theta)\mathrm{d}\mathcal{Q}(\theta)\right) \\
&= \int g(\theta)\,\mathrm{Var}^{\xi}(\mathcal{L}_h^M(\mathcal{Y}|\theta))\mathrm{d}\mathcal{Q}(\theta) \\
&\leq Ch^{2q}\int g(\theta)\mathrm{d}\mathcal{Q}(\theta) \\
&= \hat{C}h^{2q},
\end{aligned}$$

where we applied Proposition 2.4. Hence, the MSE of this estimation is bounded quadratically with respect to the order of the employed Runge-Kutta method, i.e.,

$$\begin{aligned}
\mathrm{MSE}(\mathbb{E}^{\pi_h^M}\left[g(\theta)\right]) &= \mathbb{E}^{\xi}\left[\mathbb{E}^{\pi_h^M}\left[g(\theta)\right] - \mathbb{E}^{\pi}\left[g(\theta)\right]\right]^2 + \mathrm{Var}^{\xi}(\mathbb{E}^{\pi_h^M}\left[g(\theta)\right]) \\
&\leq Ch^{2q}.
\end{aligned}$$

#### 3.4.1 Numerical experiment

We consider the FitzHug-Nagumo model (19) and produce observations $\mathcal{Y}$ at times $t_i = i$ for $i = 1, \ldots, 10$ from a reference solution with additive noise with variance $10^{-2}$. We produce a reference posterior distribution using the result of a MCMC algorithm obtained with a small time step $h$. We then vary $h$ in order to observe the convergence of the posterior distribution towards the reference, as well as the convergence of the Monte Carlo estimation. We consider 100000 iterations of RAM applied to MCWM. As discussed above, the number of trajectories $M$ used to approximate the numerical likelihood does not have an influence on the convergence rate of the posterior distribution to the true posterior. Therefore, we just fix $M$ to be equal to one. We consider as the function $g$ of the parameter the Euclidean norm, therefore we consider the approximation

$$\|\theta\| \approx 10^{-6}\sum_{i=1}^{10^6}\left\|\theta^{(i)}\right\|.$$

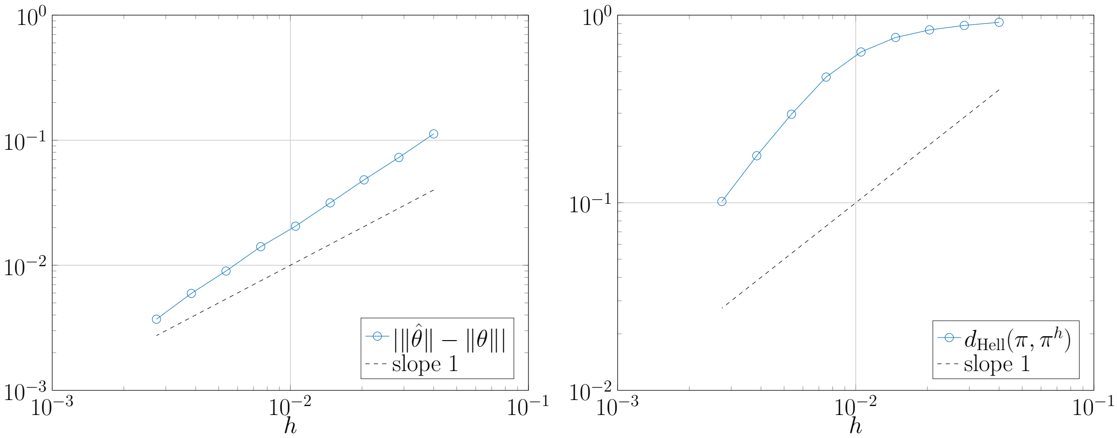



**Figure 6:** Convergence of the parameter to its stationary value and of the Hellinger distance of the probability distributions.

Results (Figure 6) show the convergence obtained averaging 10 realizations of the entire MCMC chain used in order to simulate the expectation with respect to the random variable $\sigma$.

## References


[1] C. Andrieu, A. Doucet, and R. Holenstein, *Particle Markov chain Monte Carlo methods*, J. R. Stat. Soc. Ser. B. Stat. Methodol., (2010), pp. 269 – 342.

[2] C. Andrieu and G. O. Roberts, *The pseudo-marginal approach for efficient Monte Carlo computations*, Ann. Statist., 37 (2009), pp. 697–725.

[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] A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn, *Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator*, Biometrika, (2015), pp. 1 – 19.

[5] A. Gelman and K. Shirley, *Inference from simulations and monitoring convergence*, in Handbook of Markov chain Monte Carlo, S. Brooks, A. Gelman, G. J. Jones, and X.-L. Meng, eds., CRC press, 2011, ch. 6, pp. 163–174.

[6] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Int. Stat. Rev., 70 (2002), pp. 419–435.

[7] W. R. Gilks, *Markov chain Monte Carlo*, Encyclopedia of Biostatistics, 4 (2005).

[8] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences, 160, Springer, 2005.

[9] P. E. Kloeden, E. Platen, and H. Schurz, *Numerical solution of SDE through computer experiments*, Universitext, Springer-Verlag, Berlin, 1994.

[10] F. J. Medina-Aguayo, A. Lee, and G. O. Roberts, *Stability of noisy Metropolis–Hastings*, Stat. Comput., 26 (2016), pp. 1187–1211.

[11] M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn, *On some properties of Markov chain Monte Carlo simulation methods based on the particle filter*, J. Econometrics, 171 (2012), pp. 134–151.

[12] M. Vihola, *Robust adaptive Metropolis algorithm with coerced acceptance rate*, Stat. Comput., 22 (2012), pp. 997–1008.