Probabilistic solvers for ODE's and Bayesian inference of parametrized models

Master Project - Master in CSE

Giacomo Garegnani
Supervisor: Prof. Assyr Abdulle
Experts: Prof. Fabio Nobile, Dr. Kostas Zygalakis

EPFL

02/02/2017

Outline of the presentation

1. Introduction on Bayesian inference and MCMC
2. Probabilistic solvers for ODE's
3. Bayesian inference inverse problems with differential equations

Consider Ω event space, $\mathcal{A}$ σ-algebra, P probability measure and $(\Omega, \mathcal{A}, P)$. Given A, B in Ω, Bayes' formula reads

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \propto P(B \mid A)P(A).$$

Normalization constant $P(B)$ can be replaced as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\int_\Omega P(B \mid A)P(A)},$$

as $P(A \mid B)$ is a probability distribution.

Problem. Consider two events A, B in Ω and the probability space $(\Omega, \mathcal{A}, P)$. We want to infer the probability distribution of A given B as

$$\underbrace{\pi(A \mid B)}_{\text{posterior}} \propto \overbrace{\mathcal{Q}(A)}^{\text{prior}} \underbrace{\mathcal{L}(B \mid A)}_{\text{likelihood}}$$

In models parametrized by a parameter θ, we deduce the distribution of θ through observations $\mathcal{Y}_n = \{y_1, y_2, \ldots, y_n\}$ as

$$\pi(\theta \mid \mathcal{Y}_n) \propto \mathcal{Q}(\theta)\mathcal{L}(\mathcal{Y}_n \mid \theta).$$

Goal. Approximate the expectation under the distribution $\pi(\theta \mid \mathcal{Y})$ of a functional of the parameter $\theta \in \mathbb{R}^{N_p}$ with a Monte Carlo sum, i.e.,

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

where $\theta^{(k)}$ are realizations of θ.

Problem. How do we generate samples $\theta^{(k)}$, with $k = 1, \ldots, N$ so that the approximation (5) holds?
$\rightsquigarrow$ MCMC [Gilks, 2005, e.g.]

Idea. Generate samples $\theta^{(k)}$, with from a Markov chain with kernel P until the chain reaches its stationary distribution. Different choices of the Markov kernel lead to different MCMC algorithms.

Metropolis-Hastings (MH). Choose a *proposal distribution* $q(x, y)$ such that

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

Then, given current sample $\theta^{(i)}$, the transition kernel P_{MH} giving $\theta^{(i+1)}$ is defined as

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

where $\alpha(\theta^{(i)}, \vartheta)$ is the *acceptance probability* and is given by

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

and ρ is defined as

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

Algorithm 1 Metropolis-Hastings.

Given $\theta^{(0)} \in \mathbb{R}^{N_p}$, $N \in \mathbb{N}_0$.
for $i = 0, \ldots, N$ **do**
  Draw ϑ from $q(\theta^{(i)}, \cdot)$;
  Compute the acceptance probability $\alpha(\theta^{(i)}, \vartheta)$ as

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

  Draw u from $\mathcal{U}(0, 1)$;
  if $\alpha > u$ **then**
    Accept ϑ, set $\theta_{i+1} = \vartheta$;
  else
    set $\theta_{i+1} = \theta^{(i)}$
  end if
end for

Remark. If the proposal distribution is symmetric, i.e., $q(x, y) = q(y, x)$, then

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

For example, Gaussian proposal [Kaipio and Somersalo, 2005]

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

Problems. Two main problems

1. How to choose an efficient proposal distribution?
2. How to modify MH if it is not possible to evaluate the posterior distribution?

Problem. Bad proposal distribution $q(x, y) \implies$ inefficient algorithms. Measure efficiency with *acceptance ratio*.

Idea. Adapt $q(x, y)$ to obtain a chosen acceptance ratio α^*. Choose $q(x, y)$ Gaussian, the new guess ϑ is

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

where $S_n \in \mathbb{R}^{N_p \times N_p}$ is lower triangular definite positive. Then,

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

where $\eta_n \xrightarrow{n \to \infty} 0$.

Two-dimensional distribution π with density [Kaipio and Somersalo, 2005]

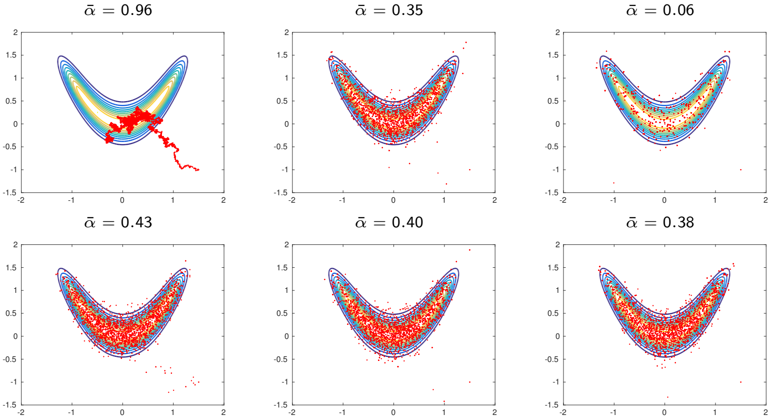
$$\pi(X) \propto \exp\left(-10(X_1^2 - X_2)^2 - (X_1 - 0.25)^4\right),$$

Setting of the experiment. Given $\sigma = \{0.01, 0.5, 2.0\}$, compare

- standard MH with Gaussian proposal with covariance σI,
- RAM with $S_0 = \sigma I$ and $\alpha^* = 0.4$.

Draw $N = 5000$ samples and compute final acceptance ratio $\bar{\alpha}$.

Robust adaptive Metropolis [Vihola, 2012], numerical experiment



Samples produced by MH and RAM for the distribution with standard MH (first row) and RAM (second row).

Problem. Impossible to evaluate $\pi(\theta)$ (no closed form available).

Idea. Find evaluable $\pi(\theta, \xi)$ that admits $\pi(\theta)$ as marginal distribution, then compute

$$\hat{\pi}_M(\theta) = \frac{1}{M} \sum_{i=1}^{M} \pi(\theta, \xi^{(i)}),$$

with $\xi^{(i)}$, $i = 1, \dots, M$ realizations of ξ. Use $\hat{\pi}_M$ for $\alpha(\theta^{(i)}, \vartheta)$.

Remark. The rest of MH is unchanged.

Outline of the presentation

1. Introduction on Bayesian inference and MCMC
2. Probabilistic solvers for ODE's
3. Bayesian inference inverse problems with differential equations

Problem. Given $f \colon \mathbb{R}^d \to \mathbb{R}^d$ and the autonomous ODE

$$u'(t) = f(u), \quad u(0) = u_0,$$

build a probabilistic numerical solution. There exists flow map $\Phi_t(y)$ such that

$$u(t) = \Phi_t(u_0).$$

Idea. Given $h > 0$, the flow map of a Runge-Kutta method $\Psi_h(y)$ is

$$u_{k+1} = \Psi_h(u_k), \quad k = 0, 1, \dots,$$

consider $\xi_k(h)$ i.i.d. random variables in $\mathbb{R}^d$ and compute

$$U_{k+1} = \underbrace{\Psi_h(U_k)}_{\text{deterministic}} + \overbrace{\xi_k(h)}^{\text{random}}, \quad k = 0, 1, \dots,$$

Deterministic component $\rightsquigarrow$ Runge-Kutta methods.

Definition

Given $s \in \mathbb{N}^*$, $(b_i)_{i=1}^s$, $(a_{ij})_{i,j=1}^s$, $h > 0$, one step of an s-stage Runge-Kutta method is

$$
\begin{aligned}
K_i &= f\big(U_0 + h \textstyle\sum_{j=1}^s a_{ij} K_j\big), \quad i = 1, \dots, s, \\
U_1 &= U_0 + h \textstyle\sum_{i=1}^s b_i K_i.
\end{aligned}
$$

The method is of order q if $\exists C > 0$ independent of h such that

$$\|u(h) - U_1\| \leq C h^{q+1}.$$

Monte Carlo computations $\rightsquigarrow$ Explicit methods privileged.

Consider chaotic differential equation, e.g., Lorenz system

$$
\begin{aligned}
x' &= \sigma(y - x), & x(0) &= -10, \\
y' &= x(\rho - z) - y, & y(0) &= -1, \\
z' &= xy - \beta z, & z(0) &= 40.
\end{aligned}
$$

Small perturbation $\implies$ uncontrollable deviation of the solution.
Deterministic solvers not reliable for any time step $h > 0$.
$\rightsquigarrow$ Family of M probabilistic numerical solutions.

Relevant properties to be analyzed

- strong order of convergence,
- weak order of convergence,
- behavior Monte Carlo approximations,
- stability.

Recall. The probabilistic method is defined as

$$U_{k+1} = \Psi_h(U_k) + \xi_k(h), \quad k = 0, 1, \dots,$$

for suitable random variables $\xi_k(h)$.

It is possible to prove a result of strong convergence.

Definition (Strong convergence)

The probabilistic method has strong order r if $\exists C > 0$ independent of h such that for h small enough

$$\sup_{t_k = kh} \mathbb{E}|U_k - u(t_k)| \leq Ch^r.$$

Assumption (Variance of random variables)

The variables $\xi_k(t)$ *satisfy for* $p \geq 1$

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

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

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

Assumption (Order of the deterministic component)

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* Ψ_h *has uniform local truncation error of order* $q+1$*, i.e.,*

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

Theorem (Strong Convergence)

Under the assumptions above, there exists $K > 0$ *such that*

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

Idea of the proof. Compute truncation error between exact and numerical solutions, divide deterministic and probabilistic contribution and derive a recurrence on the error. Apply then discrete Gronwall's lemma to bound the error.

Definition (Weak convergence)

The probabilistic method has weak order r if $\exists C > 0$ independent of h such that for any function φ sufficiently smooth

$$\sup_{t_k = kh} |\mathbb{E}[\varphi(U_k)] - \varphi(u(t_k))| \leq Ch^r,$$

for h small enough.

Idea. Introduce a modified SDE

$$\mathrm{d}\tilde{u} = f^h \tilde{u} \,\mathrm{d}t + \sqrt{h^{2p} Q} \,\mathrm{d}W \,, \tag{1}$$

and study the convergence of U_k and $\tilde{u}$ to u for $h \to 0$.

Theorem (Weak convergence)

For any function φ sufficiently smooth

$$|\varphi(u(T)) - \mathbb{E}[\varphi(U_k)]| \leq Kh^{\min\{2p,q\}}, \quad kh = T,$$

and

$$|\mathbb{E}[\varphi(\tilde{u}(T))] - \mathbb{E}[\varphi(U_k)]| \leq Kh^{2p+1}, \quad kh = T.$$

Idea of the proof. Use techniques of backwards error analysis, finding a modified ODE and SDE such that the numerical error is of higher order.

Problem. Study convergence of Monte Carlo approximations.

Numerical solution $\rightsquigarrow \mathcal{Q}_h$ (inaccessible),
M samples of numerical solution $\rightsquigarrow Q_h^M$ (accessible),
Exact solution $\rightsquigarrow \delta_u$.

Convergence scheme, we expect

$$\mathcal{Q}_h^M \xrightarrow{M \to \infty} \mathcal{Q}_h \xrightarrow{h \to 0} \delta_u .$$

Second convergence already treated, first unclear [Kersting and Hennig, 2016].

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

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

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

References II

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

[Hairer et al., 2002] Hairer, E., Lubich, C., and Wanner, G. (2002).
Geometric numerical integration: structure-preserving algorithms for ordinary differential equations.
Springer-Verlag, Berlin and New York.

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

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

[Medina-Aguayo et al., 2016] Medina-Aguayo, F. J., Lee, A., and Roberts, G. O. (2016).
Stability of noisy Metropolis–Hastings.
Stat. Comput., 26(6):1187–1211.

[Vihola, 2012] Vihola, M. (2012).
Robust adaptive Metropolis algorithm with coerced acceptance rate.
Stat. Comput., 22(5):997–1008.