SMC samplers

nicolas.chopin@ensae.fr
(based on a previous PG course with O. Papaspiliopoulos)

Summary

- Motivating problems: sequential (or non-sequential) inference and simulation outside SSMs (including normalising constant calculation)
- Feynman-Kac formalisation of such problems
- Specific algorithms: IBIS, tempering SMC, SMC-ABC
- An overarching framework: SMC samplers

Outline

Sequential Bayesian learning

$\mathbb{P}_t(\mathrm{d}\theta)$ posterior distribution of parameters θ, given observations $y_{0:t}$, where $\theta \in \Theta$; typically:

$$\mathbb{P}_t(\mathrm{d}\theta) = \frac{1}{p_t(y_{0:t})} p_t^\theta(y_{0:t}) \mathbb{P}_{-1}(\mathrm{d}\theta)$$

with $\mathbb{P}_{-1}(\mathrm{d}\theta)$ the prior distribution, $p_t^\theta(y_{0:t})$ likelihood and $p_t(y_{0:t})$ marginal likelihood.
Note that

$$\frac{\mathbb{P}_t(\mathrm{d}\theta)}{\mathbb{P}_{t-1}(\mathrm{d}\theta)} = \frac{1}{p_t(y_t|y_{0:t-1})} p_t^\theta(y_t|y_{0:t-1}) \propto p_t^\theta(y_t|y_{0:t-1}).$$

Practical motivations

- sequential learning
- Detection of outliers and structural changes
- Sequential model choice/composition
- 'Big' data
- Data tempering effect

Tempering

Suppose we wish to either sample from, or compute the normalising constant of

$$\mathbb{P}(\mathrm{d}\theta) = \frac{1}{L}\exp\{-V(\theta)\}\mu(\mathrm{d}\theta).$$

Idea: introduce for any $a \in [0, 1]$,

$$\mathbb{P}^a(\mathrm{d}\theta) = \frac{1}{L_a}\exp\{-aV(\theta)\}\mu(\mathrm{d}\theta).$$

Note that

$$\frac{\mathbb{P}^b(\mathrm{d}\theta)}{\mathbb{P}^a(\mathrm{d}\theta)} = \frac{L_a}{L_b}\exp\{(a-b)V(\theta)\} \propto \exp\{(a-b)V(\theta)\}$$

Estimating ratios of normalising constants

Interestingly, we have two identities to compute L_1/L_0:

- Bridge sampling:

$$\frac{L_1}{L_0} = \prod_{i=1}^{n} \frac{L_{a_i}}{L_{a_{i-1}}}$$

for $0 = a_0 < \ldots < a_n = 1$, where

$$\frac{L_{a_i}}{L_{a_{i-1}}} = \int_{\Theta} \exp\{(a_{i-1} - a_i)V(\theta)\}\mathbb{P}^{a_{i-1}}(\mathrm{d}\theta)$$

- Thermodynamic integration:

$$\log(L_1/L_0) = \int_0^1 \mathbb{P}^a \left[V(\Theta)\right] \mathrm{d}a$$

Rare events

Suppose we wish to either sample from, or compute the normalising constant of

$$\mathbb{P}(\mathrm{d}\theta) = \frac{1}{L}\mathbb{1}_E(\theta)\mu(\mathrm{d}\theta).$$

for some set E.

As for tempering, we could introduce a sequence of sets $\Theta = E_0 \supset \ldots \supset E_n = E$, and the corresponding sequence of distributions.

Outline

Statement

Sequence of probability distributions on a common space $(\Theta, \mathcal{B}(\Theta))$, $\mathbb{P}_0(\mathrm{d}\theta), \ldots, \mathbb{P}_T(\mathrm{d}\theta)$. In certain applications interest only in $\mathbb{P}_T$, in others for all $\mathbb{P}_t$, in others mainly interested in normalising constants.

IBIS

Input of IBIS[1]

- A sequence of distributions $\mathbb{P}_t(\mathrm{d}\theta)$ on $(\Theta, \mathcal{B}(\Theta))$
- Weights $G_t(\theta) = \ell_t^{-1}\mathbb{P}_t(\mathrm{d}\theta)/\mathbb{P}_{t-1}(\mathrm{d}\theta)$, with $\mathbb{P}_{-1} := \mathbb{M}_0$, and ℓ_t normalising constant
- MCMC kernels, $M_t(\theta, \mathrm{d}\theta')$ invariant wrt $\mathbb{P}_{t-1}$
- The number of particles N

[1] Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89:539–552

Algorithm 1 IBIS pt1

All operations to be performed for all $n \in 1:N$.

At time 0:

(a) Generate $\Theta_0^n \sim \mathbb{M}_0(\mathrm{d}x_0)$.

(b) Compute $w_0^n = G_0(\Theta_0^n)$, $W_0^n = w_0^n / \sum_{m=1}^N w_0^m$, and $l_0^N = N^{-1} \sum_{n=1}^N w_0^n$.

Algorithm 2 IBIS pt2

Recursively, for $t = 1, \dots, T$:

If degeneracy criterion not fulfilled:

(a) Set $\Theta_t^n = \Theta_{t-1}^n$.

(b) Compute $w_t^n = w_{t-1}^n G_t(\Theta_t^n)$, $W_t^n = w_t^n / \sum_{m=1}^N w_t^m$, and $l_t^N = N^{-1} \sum_{n=1}^N w_t^n$.

If degeneracy criterion fulfilled:

(a) Generate ancestor variables $A_t^n \in 1:N$ independently from $\mathcal{M}(W_{t-1}^{1:N})$.

(b) Generate $\Theta_t^n \sim M_t(\Theta_{t-1}^{A_t^n}, \mathrm{d}\theta)$.

(c) Set $w_t^n = G_t(\Theta_t^n)$. $W_t^n = w_t^n / \sum_{m=1}^N w_t^m$, and $l_t^N = N^{-1} \sum_{n=1}^N w_t^n$.

Output of IBIS

$$\mathbb{P}_t^N(\mathrm{d}\theta) = \sum_{n=1}^{N} W_t^n \delta_{\Theta_t^n}(\mathrm{d}\theta) \quad \text{approximates } \mathbb{P}_t(\mathrm{d}\theta)$$

$$\ell_t^N \quad \text{approximates } \ell_t$$

Main tools in IBIS

- Extension, invariance & particle approx

$$
\begin{aligned}
\mathbb{P}_t(\mathrm{d}\theta') &= \frac{\mathbb{P}_t(\mathrm{d}\theta')}{\mathbb{P}_{t-1}(\mathrm{d}\theta')}\mathbb{P}_{t-1}(\mathrm{d}\theta') \\
&= G_t(\theta') \int_\Theta M_t(\theta, \mathrm{d}\theta')\mathbb{P}_{t-1}(\mathrm{d}\theta) \\
&\approx G_t(\theta') \int_\Theta M_t(\theta, \mathrm{d}\theta')\mathbb{P}^N_{t-1}(\mathrm{d}\theta) \\
&= G_t(\theta') \sum_n M_t(\Theta^n_{t-1}, \mathrm{d}\theta') W^n_{t-1}
\end{aligned}
$$

- Use of two types of invariant transition kernels
- Adaptation

Example: sequential Bayesian learning

$$\frac{\mathbb{P}_t(\mathrm{d}\theta)}{\mathbb{P}_{t-1}(\mathrm{d}\theta)} = \frac{1}{p_t(y_t|y_{0:t-1})} p_t^\theta(y_t|y_{0:t-1})$$

$$\ell_t = p_t(y_t|y_{0:t-1})$$

Typically, $\mathbb{M}_0 = \mathbb{P}_{-1}$; improper priors; analyse in batches

Auto-calibration

A standard choice for MCMC kernel M_t is a Gaussian random walk Metropolis. Then we can calibrate the random walk variance on the empirical variance of the resampled particles.
It is also possible to automatically choose when to do resampling+MCMC:

- for sequential inference, trigger resampling+MCMC when ESS is below (say) $N/2$.
- for tempering SMC, one may choose recursively $\delta_i = a_i - a_{i-1}$ by solving numerically $\mathrm{ESS} = N/2$ (say).

SMC samplers

A principled framework for building FK models for a broad category of problems, outside the rehearsed state-space models: Del Moral et al[2] (2006).

Includes many of the previous ideas as special cases

[2]Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers.
J. R. Stat. Soc. Ser. B Stat. Methodol., 68(3):411–436

SMC samplers

Input of SMC sampler

- A sequence of distributions $\mathbb{P}_t(\mathrm{d}\theta)$ on $(\Theta, \mathcal{B}(\Theta))$
- (Forward) kernels, $M_t(\theta, \mathrm{d}\theta')$, $t = 1 : T$ and backward kernels $\overleftarrow{K}_{t-1}(\theta, \mathrm{d}\theta')$, $t = T - 1 : 0$
- Weights, for $t = 1, \ldots, T$

$$G_t(\theta', \theta) = c_t^{-1} \frac{\mathbb{P}_t(\mathrm{d}\theta) \overleftarrow{K}_{t-1}(\theta, \mathrm{d}\theta')}{\mathbb{P}_{t-1}(\mathrm{d}\theta') M_t(\theta', \mathrm{d}\theta)}$$

and $G_0(\theta) = c_0^{-1} \mathbb{P}_0(\mathrm{d}\theta) / \mathbb{M}_0(\mathrm{d}\theta)$,
- The number of particles N

With these ingredients, define Feynman-Kac models

$$\mathbb{Q}_t(\mathrm{d}\theta_{0:t}) = \frac{1}{L_t}\mathbb{M}_t(\mathrm{d}\theta_{0:t}) G_0(\theta_0) \prod_{s=1}^{t} G_s(\theta_{s-1}, \theta_s)$$

A direct calculation shows that for each t,

$$\mathbb{Q}_t(\mathrm{d}\theta_{0:t}) = \mathbb{P}_t(\mathrm{d}\theta_t) \prod_{s=1}^{t} \overleftarrow{K}_{s-1}(\theta_s, \mathrm{d}\theta_{s-1})$$

The main idea

This is simply based on two ideas we have already developed:

- Feynman-Kac model as a Markov measure
- backward kernel of a MC
- Then, easy to verify by telescoping that

$$\frac{\mathbb{Q}_t(\mathrm{d}\theta_{0:t})}{\mathbb{M}_t(\mathrm{d}\theta_{0:t})} \propto G_0(\theta_0) \prod_{s=1}^{t} G_s(\theta_{s-1}, \theta_s)$$

Special cases

- for trivial dynamics $M_t(\theta, \mathrm{d}\theta') = \delta_\theta(\mathrm{d}\theta')$, we can set $\overleftarrow{K}_{t-1}(\theta, \mathrm{d}\theta') = \delta_\theta(\mathrm{d}\theta')$ (this is a trivial special case of the above)
- If M_t is invariant wrt $\mathbb{P}_t$, then

$$\overleftarrow{K}_{t-1}(\theta, \mathrm{d}\theta') = \mathbb{P}_t(\mathrm{d}\theta') \frac{M_t(\theta', \mathrm{d}\theta)}{\mathbb{P}_t(\mathrm{d}\theta)} \quad G_t(\theta', \theta) = \frac{\mathbb{P}_t(\mathrm{d}\theta')}{\mathbb{P}_{t-1}(\mathrm{d}\theta')}$$

Optimal choice of backward kernels

Fix horizon t; then if we choose

$$\overleftarrow{K}_{s-1}(\theta_s, \mathrm{d}\theta_{s-1}) = \overleftarrow{M}_{s-1}(\theta_s, \mathrm{d}\theta_{s-1}) = \mathbb{M}_t(\mathrm{d}\theta_{s-1}) \frac{M_s(\theta_{s-1}, \mathrm{d}\theta_s)}{\mathbb{M}_t(\mathrm{d}\theta_s)}$$

then

$$\frac{\mathbb{Q}_t(\mathrm{d}\theta_{0:t})}{\mathbb{M}_t(\mathrm{d}\theta_{0:t})} = \left.\frac{\mathbb{P}_t(\mathrm{d}\theta)}{\mathbb{M}_t(\mathrm{d}\theta)}\right|_{\theta=\theta_t}$$

which is clearly optimal (given the fixed forward kernels), but typically intractable; this is of course the case of SIS.