Asymptotic analysis of genealogies induced by sequential Monte Carlo algorithms

Suzie Brown

March 11, 2019

1 Introduction

- organisation of the report

Sequential Monte Carlo has become a popular tool, particularly in applications such as object tracking, where there is a natural sequential component and we wish to infer underlying states from noisy observations. While particle methods can be very effective for filtering, it is more difficult to apply them to smoothing because they typically suffer very badly from ancestral degeneracy in the particle genealogies.

When attempting to mitigate this problem, one often encounters a trade-off between ancestral degeneracy (arising from resampling) and weight degeneracy (arising from sequential importance sampling). However, while weight degeneracy is a reasonably well-quantified problem, there exists little in the way of tools for quantifying ancestral degeneracy a priori. There have been some simulation studies attempting to cast light on the magnitude of this problem, but analytical findings remain elusive, since the complexity of the most commonly used particle methods makes it difficult to obtain any rigorous results. Consequently, there is a wealth of pertinent open questions in this area. This work attempts to extend a first result for a standard class of SMC algorithms to the more sophisticated algorithms which are typically used in practice.

Throughout this document we will use the compact notation $X_{m:n}$ as shorthand for $X_m, X_{m+1}, \ldots, X_n$, as well as $X_{-n} := X_0, \ldots, X_{n-1}, X_{n+1}, \ldots X_N$.

2 Foo

- the problem with smoothing: ancestral vs. weight degeneracy
- motivating plots

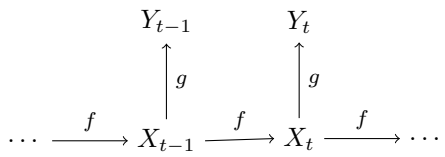
References for this section are Doucet et al. (2001), Del Moral et al. (2006), and Doucet and Johansen (2009).

2.1 Class of models

Although sequential Monte Carlo (SMC) methods can be applied in a much more general setting, they are particularly easy to motivate in the setting of state space models, where the "sequential" nature follows naturally from the discrete time steps present in the model. For the purposes of presenting the algorithm, let us consider a time-homogeneous state space model consisting of an unobservable discrete-time Markov process $X_{0:T}$ and observables $Y_{0:T}$, satisfying the conditional independence structure

$$X_{t+1} \perp X_{0:t-1}, X_{t+2:T} \mid X_t$$
$$Y_t \perp Y_{-t}, X_{-t} \mid X_t$$

for all $t \in \{0, 1, \ldots, T\}$, as represented by the graphical model below.



We assume for notational convenience that $x_0, \ldots, x_T$ take values in a common state space $\mathcal{X}$, and $y_0, \ldots, y_T$ in a common state space $\mathcal{Y}$, but these assumptions can be dropped.

Suppose we have the following model:

$$
\begin{aligned}
&X_0 \sim \mu(\cdot) \\
&X_{t+1} \mid (X_t = x_t) \sim f(\cdot|x_t) \qquad t = 0, \ldots, T-1 \\
&Y_t \mid (X_t = x_t) \sim g(\cdot|x_t) \qquad t = 0, \ldots, T
\end{aligned}
$$

where $(X_t)_{t=0}^T$ is an unobservable discrete-time Markov process and the observables $(Y_t)_{t=0}^T$ satisfy $Y_t \perp \{Y_{-t}, X_{-t}\} \mid X_t$.

We assume that the *transition* and *emission* kernels have densities which are denoted by f and g respectively, but this is not necessary in general. We only require that we can sample from $\mu(\cdot)$ and $f(\cdot|x)$, and calculate *unnormalised* potentials $g(y|x)$, for all x, y.

2.2 Inference in state space models

Suppose we are in a Bayesian setting, where μ is our prior distribution at time 0, observations y_t arrive sequentially, and we want to infer information about the hidden states (either on- or off-line). The three main inference problems are:

Filtering (where is it now?) $p(x_t|y_{0:t})$

Prediction (where will it go next?) $p(x_{t+1}|y_{0:t})$

Smoothing (where has it been?) $p(x_{0:t}|y_{0:t})$

In the on-line setting, we take as our prior the posterior distribution from the previous time step $t-1$, and update it using the new observation y_t. The inference must be fast enough to keep up with the rate of arrival of observations, so in particular the complexity of the update must not increase with T. In the off-line setting, we take μ as the prior distribution, and infer the set of posteriors once all $T+1$ observations have arrived.

Prediction and filtering are essentially equivalent, because given a filtering distribution, the corresponding predictive distribution can be obtained by applying the transition kernel f. Smoothing is considered a harder task because it requires us to infer many more parameters from the same amount of information; indeed the dimension of the problem increases linearly with T.

In the case of linear Gaussian state space models (i.e. where f and g are Gaussian densities that depend only linearly on their arguments), the posterior distributions of interest are available analytically, by way of the Kalman filter (Kalman, 1960) and Rauch-Tung-Striebel (RTS) smoother recursions (Rauch et al., 1965). The other analytic case occurs if the state space of $(X_t)_{t=0}^\infty$ is finite, in which case the forward-backward algorithm (Baum, 1972) yields the exact posteriors.

2.3 Sequential Monte Carlo

In more complex models such techniques are not feasible, and we are forced to resort to Monte Carlo methods. For state space models, Markov chain Monte Carlo methods are not very effective due to the high dimension of the parameter space. But we can exploit the sequential nature of the underlying dynamics to decompose the problem into a sequence of inferences of more manageable dimension. This is the motivation behind sequential Monte Carlo (SMC) methods.

The conditional independence structure in the model implies that the (joint) marginal distribution of the hidden states $X_{0:t}$ is given by

$$p(x_{0:t}) = \mu(x_0) \prod_{i=1}^{t} f(x_i \mid x_{i-1})$$

and that the likelihood of the observations $y_{0:t}$ given the underlying states $x_{0:t}$ takes the form

$$p(y_{0:t} \mid x_{0:t}) = \prod_{i=0}^{t} g(y_i \mid x_i).$$

The smoothing distribution $p(x_t \mid y_{0:T})$ is obtained from $p(x_{0:T} \mid y_{0:T})$ by marginalising. Using the conditional independence structure, we can write

$$p(x_{0:t} \mid y_{0:t}) \propto g(y_t \mid x_t) f(x_t \mid x_{t-1}) p(x_{0:t-1} \mid y_{0:t-1}) \tag{1}$$

$$\propto \mu(x_0) g(y_0 \mid x_0) \prod_{i=1}^{t} f(x_i \mid x_{i-1}) g(y_i \mid x_i) \tag{2}$$

for $t = 0, \ldots, M$, where the one-step recursion (1) is obtained using Bayes rule, and (2) is obtained by applying (1) t times. The filtering distribution $p(x_t \mid y_{0:t})$ can be obtained from (1) by marginalising out $x_{0:t-1}$, which is straightforward if Monte Carlo samples are available. The predictive distributions can also be derived from the smoothing distributions using

$$p(x_{t+1} \mid y_{0:t}) = g(x_{t+1} \mid x_t) p(x_{0:t} \mid y_{0:t}).$$

SMC provides a method to approximate to (1), given a model specification and a sequence of observations. Like the underlying process, the algorithm proceeds sequentially, returning its approximation to the smoothing distribution at each time step. A generic SMC algorithm is presented below.

Algorithm 1 Standard SMC

Inputs: $\mu : \mathcal{X} \to [0,1]$; $f : \mathcal{X} \times \mathcal{X} \to [0,1]$; $g : \mathcal{Y} \times \mathcal{X} \to [0,1]$; $y_{0:M} \in \mathcal{Y}^T$; $N \in \mathbb{N}$

for $i = 1, \ldots, N$ **do**

- $x_0^{(i)} \sim \mu(\cdot)$ ▷ initialise
- $\tilde{w}_0^{(i)} \leftarrow g(y_0 \mid x_0^{(i)})$
- $w_0^{(i)} \leftarrow \tilde{w}_0^{(i)} / \sum \tilde{w}_0^{(j)}$

end for

for $t = 1, \ldots, T$ **do**

- **for** $i = 1, \ldots, N$ **do**
  - $\tilde{x}_t^{(i)} \leftarrow \text{RESAMPLE}(\mathbf{x}_{t-1}, \mathbf{w}_{t-1})$ ▷ resample particles
  - $x_t^{(i)} \sim f(\cdot \mid \tilde{x}_t^{(i)})$ ▷ propagate particles
  - $\tilde{w}_t^{(i)} \leftarrow g(y_t \mid x_t^{(i)})$ ▷ calculate weights
  - $w_t^{(i)} \leftarrow \tilde{w}_t^{(i)} / \sum \tilde{w}_t^{(j)}$ ▷ normalise weights
- **end for**

end for

3 SMC as a coalescent

- k-coalescent & Kingman coalescent
- pop gen literature about large population cts time limits of various models
- resampling viewed backwards in time: branching process → coalescent process
- asymptotic properties of SMC lit review: CLT, path storage, coalescence etc.
- the gap in knowledge that we aim to fill
- (remark: although SMC has other problems in high dimension, the coalescence rate doesn't depend on the dimension...)

4 Conditional SMC

- motivation: particle MCMC, need for multiple lineages
- conditional (multinomial) SMC algorithm and its context within particle Gibbs
- result: coalescence rate etc in terms of standard multinomial one; verification of assumptions of KJJS theorem (but exile horrible calculations to appendix)
- simulations & conclusions thence

Algorithm 2 Conditional SMC with multinomial resampling

Require: $N, T, \mu, \{K_t\}, \{g_t\}, x^*_{0:T}$

for $i \in \{1, \ldots, N\}$ **do**
  Sample $X_0^{(i)} \sim \mu$
end for
Sample $a_0^* \sim \text{Uniform}(\{1, \ldots, N\})$
$X_0^{(a_0^*)} \leftarrow x_0^*$
for $i \in \{1, \ldots, N\}$ **do**
  $w_0^{(i)} \leftarrow \frac{g_0(X_0^{(i)})}{\sum_{j=1}^N g_0(X_0^{(j)})}$
end for
for $t \in \{0, \ldots, T-1\}$ **do**
  Sample $a_t^{(1:N)} \sim \text{Categorical}(\{1, \ldots, N\}, w_t^{(1:N)})$
  Sample $a_{t+1}^* \sim \text{Uniform}(\{1, \ldots, N\})$
  $a_t^{(a_{t+1}^*)} \leftarrow a_t^*$
  for $i \in \{1, \ldots, N\}$ **do**
    Sample $X_{t+1}^{(i)} \sim K_{t+1}(X_t^{(a_t^{(i)})}, \cdot)$
  end for
  $X_{t+1}^{(a_{t+1}^*)} \leftarrow X_{t+1}^*$
  for $i \in \{1, \ldots, N\}$ **do**
    $w_{t+1}^{(i)} \leftarrow \frac{g_{t+1}(X_t^{(a_t^{(i)})}, X_{t+1}^{(i)})}{\sum_{j=1}^N g_{t+1}(X_t^{(a_t^{(j)})}, X_{t+1}^{(j)})}$
  end for
end for

5 Alternative resampling schemes

- multinomial not really used in practice, but other schemes hard to analyse
- overview of the main variance-reducing schemes
- results: theorem for residual resampling (hopefully)
- maybe results for other schemes
- simulation comparing all of them, and conclusions thence

6 Discussion

- results so far
- impact of this work: to practitioners, to enriching the SMC literature, interpretation within pop gen.
- future directions

References

Baum, L. (1972), 'An inequality and associated maximization technique occurring in the statistical analysis of probabilistic functions of markov chains', *Inequalities,* **3**, 1–8.

Del Moral, P., Doucet, A. and Jasra, A. (2006), 'Sequential monte carlo samplers', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3), 411–436.

Doucet, A., De Freitas, N. and Gordon, N. (2001), An introduction to sequential monte carlo methods, *in* 'Sequential Monte Carlo methods in practice', Springer, pp. 3–14.

Doucet, A. and Johansen, A. M. (2009), 'A tutorial on particle filtering and smoothing: Fifteen years later', *Handbook of nonlinear filtering* **12**(656-704), 3.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Journal of basic Engineering* **82**(1), 35–45.

Rauch, H. E., Striebel, C. and Tung, F. (1965), 'Maximum likelihood estimates of linear dynamic systems', *AIAA journal* **3**(8), 1445–1450.