Asymptotic analysis of genealogies induced by sequential Monte Carlo algorithms

Suzie Brown

March 12, 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

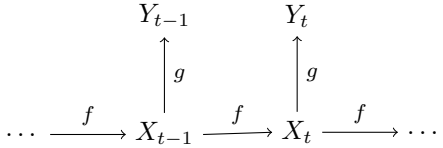
References for this section are Doucet et al. (2001), Del Moral et al. (2006), and Doucet and Johansen (2009).

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

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

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

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 $p(x_t|y_{0:t})$ 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

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

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

for $i = 1, \ldots, N$ **do**
- $x_0^{(i)} \sim \mu(\cdot)$ ▷ initialise
- $\tilde{w}_0^{(i)} \leftarrow g(y_0|x_0^{(i)})$
- $w_0^{(i)} \leftarrow \tilde{w}_0^{(i)} / \sum \tilde{w}_0^{(j)}$

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
  - $x_t^{(i)} \sim f(\cdot|\tilde{x}_t^{(i)})$ ▷ propagate particles
  - $\tilde{w}_t^{(i)} \leftarrow g(y_t|x_t^{(i)})$ ▷ calculate weights
  - $w_t^{(i)} \leftarrow \tilde{w}_t^{(i)} / \sum \tilde{w}_t^{(j)}$ ▷ normalise weights
- **end for**

end for

If only the latest filtering distribution is required, we can marginalise out $\mathbf{x}_{0:t-1}$ at each step by simply throwing away the particle histories and keeping only the particle approximation $\mathbf{x}_t$ to the filtering distribution at the current time t. The algorithm progresses in a Markovian fashion, only ever referring to the particles at the immediately previous step, so filtering distributions can be approximated with minimal memory usage. If, say, the mean and variance of $X_t \mid y_{0:t}$ at each time t are required, we can store just these summary statistics, plus the two most recent generations of particles, and throw away all other information about the particles at previous time steps. This is vital if one wishes to carry out filtering in an on-line fashion, as it prevents the memory requirements accumulating more than necessary.

The form of the RESAMPLE function in Algorithm 1 is discussed in Section 5.

2.4 Ancestral degeneracy

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

3 SMC as a 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

3.1 Kingman's coalescent

Imagine we have a population with fixed size N over discrete generations, where each individual is descended from one individual of the previous generation. Then for each individual in the present generation, we can trace their *lineage* back through the generations. If we trace two lineages back in time, at some generation they may descend from the same individual, at which point we say they have *coalesced*. Once two lineages have coalesced they will stay together going backwards in time. The combined lineages of $n \leq N$ of the present individuals therefore forms a tree, or several non-overlapping trees, the entirety of which we refer to as the *ancestry* or *genealogy* of those n individuals.

Kingman's *n-coalescent* provides a model for such genealogies. Kingman showed in (Kingman, 1982*a*,*b*,*c*) that the n-coalescent is the limiting process for samples from a wide class of population models as $N \to \infty$.

The defining feature of the model is that each pair of lineages merges with unit rate. This means that many coalescences occur while there are many distinct lineages present. In particular, the n-coalescent can be formulated as a Poisson process where pairs of lineages coalesce independently at rate 1, with the pair to coalesce being chosen uniformly at random (Wakeley, 2009, Section 3.2).

In the notation of Wakeley (2009), let T_i; $i = 2, \ldots, n$ be the i^{th} coalescence time, that is, the length of time for which there are exactly i branches in the sample genealogy. The n-coalescent is the process in which these times are distributed as independent Exponentials with rate $\binom{i}{2}$.

Möhle (1998) writes the same process in terms of the infinitesimal generator Q of a Markov process on the set of equivalence relations on n elements, having entries

$$q_{\xi\eta} = \begin{cases} -\binom{b}{2} & \text{if } \xi = \eta \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases}$$

where b is the number of equivalence classes of ξ, and $\xi \prec \eta$ means that η is a state with exactly one more pair of lineages coalesced compared to ξ.

The *Kingman coalescent* is the process on the whole population of size $N \to \infty$, such that the genealogy of any sample of size $n < N$ individuals from the present generation is an n-coalescent.

4 Conditional SMC

- motivation: particle MCMC, need for multiple lineages
- result: coalescence rate etc in terms of standard multinomial one; verification of assumptions of KJJS theorem (but exile horrible calculations to appendix)

Conditional SMC differs from the standard algorithm in that one predetermined trajectory (that is, a sequence of particle positions and the corresponding ancestral line) is conditioned to survive all of the propagation and resampling steps. We will refer to this sequence as the *immortal trajectory*, following the terminology used for conditioned Galton-Watson processes, and the *immortal particle* will refer to the particle in a particular generation that is part of the immortal trajectory.

The conditional SMC algorithm was proposed by Andrieu et al. (2010) for use in the *particle Gibbs* sampler, which they introduce as part of a more general class of particle MCMC methods. In the particle Gibbs sampler, the standard SMC algorithm does not admit the desired target distribution, so this conditional version must be used instead.

When used as a component of the particle Gibbs algorithm, the immortal trajectory $x^*_{0:T}$ for each SMC run is sampled from the trajectories output from the previous run (Andrieu et al., 2010, Section 2.4.3). However, for our purposes we just consider a single SMC run for which the immortal trajectory is fixed.

A conditional SMC algorithm employing multinomial resampling is described in Algorithm 2.

5 Alternative resampling schemes

- overview of the main variance-reducing schemes
- results: theorem for residual resampling (hopefully)
- maybe results for other schemes

Algorithm 2 Conditional SMC with multinomial resampling

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

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

There is a great deal of flexibility in the function referred to as RESAMPLE in Algorithm 1. The most straightforward choice is multinomial resampling (Efron and Tibshirani, 1994), which is also the easiest to analyse. However, multinomial resampling is well known to be sub-optimal in terms of the resulting Monte Carlo variance, and is rarely used in practice. For instance, Douc and Cappé (2005) proves that both residual resampling and stratified resampling yield lower variance. In this section we will present some resampling schemes that claim to perform better than multinomial resampling.

6 Discussion

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

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010), 'Particle markov chain monte carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 269–342.

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.

Douc, R. and Cappé, O. (2005), Comparison of resampling schemes for particle filtering, *in* 'Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on', IEEE, pp. 64–69.

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.

Efron, B. and Tibshirani, R. J. (1994), *An introduction to the bootstrap*, CRC press.

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

Kingman, J. (1982*a*), 'The coalescent', *Stochastic processes and their applications* **13**(3), 235–248.

Kingman, J. (1982*b*), Exchangeability and the evolution of large populations, *in* 'Proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th-9th April, 1981, in Honour of Professor Bruno de Finetti', North-Holland, Amsterdam.

Kingman, J. (1982*c*), 'On the genealogy of large populations', *Journal of Applied Probability* **19**(A), 27–43.

Möhle, M. (1998), 'Robustness results for the coalescent', *Journal of applied probability* **35**(2), 438–447.

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

Wakeley, J. (2009), *Coalescent theory: an introduction*, number 575: 519.2 WAK.