UPPSALA UNIVERSITET

Sequential Monte Carlo methods

Lecture 8 – Path space view of the particle filter

Thomas Schön, Uppsala University

2017-08-25

Outline – Lecture 8

Aim: Introduce the path space view of the particle filter, explain the path degeneracy problem and briefly mention the low-variance resampling methods.

Outline:

1. Path space view of the particle filter
2. Path degeneracy
3. Mitigating the path degeneracy problem
   a. Effective samples size (ESS)
   b. Low variance resampling
4. Parameter inference in SSMs

Reminder – the bootstrap particle filter

Algorithm 1 Bootstrap particle filter (for $i = 1, \dots, N$)

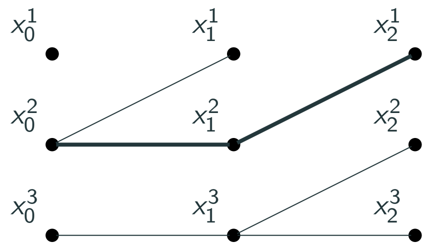
1. **Initialization ($t = 0$):**
   (a) Sample $x_0^i \sim p(x_0)$.
   (b) Set initial weights: $w_0^i = 1/N$.
2. **for** $t = 1$ **to** T **do**
   (a) **Resample:** sample ancestor indices $a_t^i \sim \mathcal{C}(\{w_{t-1}^j\}_{j=1}^N)$.
   (b) **Propagate:** sample $x_t^i \sim p(x_t \mid x_{t-1}^{a_t^i})$. $x_{0:t}^i = \{x_{0:t-1}^{a_t^i}, x_t^i\}$.
   (c) **Weight:** compute $\widetilde{w}_t^i = p(y_t \mid x_t^i)$ and normalize $w_t^i = \widetilde{w}_t^i / \sum_{j=1}^N \widetilde{w}_t^j$.

The **ancestor indices** $\{a_t^i\}_{i=1}^N$ allow us to keep track of exactly what happens in each resampling step.

Note the bookkeeping added to the propagation step 2b.

Bookkeeping – ancestral path

Example evolution of three particles for $t = 0, 1, 2$.



The **ancestral path** of x_2^1, i.e. $x_{0:2}^1$, is shown as the thick line.

Bookkeeping – ancestor indices

At time $t = 1$, particle x_0^2 is resampled twice and particle x_0^3 is resampled once (whereas particle x_0^1 is not resampled). Hence, at time $t = 1$, the **ancestor indices** are

$$a_1^1 = 2, \; a_1^2 = 2 \text{ and } a_1^3 = 3.$$

Similarly, at time $t = 2$, the **ancestor indices** are given by

$$a_2^1 = 2, \; a_2^2 = 3 \text{ and } a_2^3 = 3.$$

The **ancestral path** of x_2^1, i.e. $x_{0:2}^1$, is shown as a thick line. It is defined recursively from the ancestor indices

$$x_{0:2}^1 = (x_0^{a_1^{a_2^1}}, x_1^{a_2^1}, x_2^1) = (x_0^{a_1^2}, x_1^2, x_2^1) = (x_0^2, x_1^2, x_2^1).$$

Bootstrap PF targeting the joint filtering PDF

Algorithm 2 joint filtering bootstrap PF (for $i = 1, \dots, N$)

1. **Initialization ($t = 0$):**
   (a) Sample $x_0^i \sim p(x_0)$.
   (b) Set initial weights: $w_0^i = 1/N$.
2. **for** $t = 1$ **to** T **do**
   (a) **Resample:** sample ancestor indices $a_t^i \sim \mathcal{C}(\{w_{t-1}^j\}_{j=1}^N)$.
   (b) **Propagate:** sample $x_t^i \sim p(x_t \mid x_{t-1}^{a_t^i})$. $x_{0:t}^i = \{x_{0:t-1}^{a_t^i}, x_t^i\}$.
   (c) **Weight:** compute $\widetilde{w}_t^i = p(y_t \mid x_t^i)$ and normalize $w_t^i = \widetilde{w}_t^i / \sum_{j=1}^N \widetilde{w}_t^j$.

Bootstrap PF targeting the joint filtering PDF

It can be shown that Algorithm 2 targets the joint filtering pdf

$$p(x_{0:t} \mid y_{1:t}) = p(x_{0:t-1} \mid y_{1:t-1}) \frac{p(x_t \mid x_{t-1}) p(y_t \mid x_t)}{p(y_t) \mid y_{1:t-1})}.$$

It resamples entire trajectories $x_{0:t}^i$, not just individual states x_t^i.

Resulting approximation of the joint filtering PDF

$$\widehat{p}^N(x_{0:t} \mid y_{1:t}) = \sum_{i=1}^N w_t^i \delta_{x_{0:t}^i}(x_{0:t}).$$

Problem: While it can actually be shown that the estimate $\widehat{p}^N(x_{0:t} \mid y_{1:t})$ produced by Algorithm 2 converge asymptotically as $N \to \infty$ it is still **not** a good approximation of $p(x_{0:t} \mid y_{1:t})$!

Why?

Path degeneracy

ex) Path degeneracy

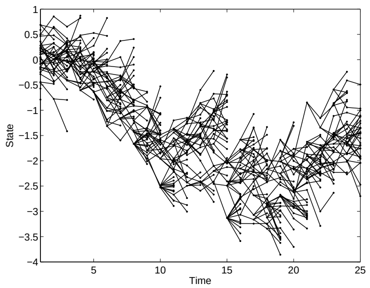
1D Gaussian random walk, measured in Gaussian noise, $T = 25$.

Target the joint filtering density using a bootstrap PF (Alg. 2) with $N = 30$ particles.

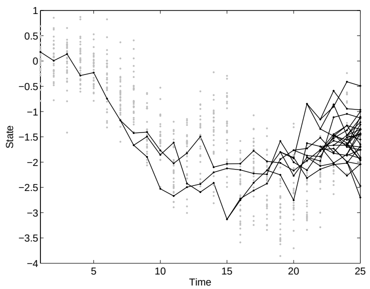
$$\widehat{p}(x_{0:25} \mid y_{1:25}) = \sum_{i=1}^{30} w_{25}^{i} \delta_{x_{0:25}^{i}}(x_{0:25}).$$

ex) Path degeneracy

ex) Path degeneracy



At each point in time all particles are plotted using a black dot and each particle is connected with its ancestor using a black line.



The grey dots represents $\widehat{p}(x_t \mid y_{1:t})$ at each point in time.

The black lines represents $\widehat{p}(x_{0:25} \mid y_{1:t})$.

ex) Path degeneracy

Note that all ancestral paths $\{x_{0:25}^{i}\}_{i=1}^{N}$ share a common ancestor at time $t = 6$ (and consequently for all times $t < 6$ as well).

Let us use the resulting particle system $\{w_{25}^{i}, x_{0:25}^{i}\}_{i=1}^{N}$ to compute a Monte Carlo estimate of $\mathbb{E}[x_3 \mid y_{1:25}]$,

$$\mathbb{E}[x_3 \mid y_{1:25}] \approx \sum_{i=1}^{30} w_{25}^{i} x_3^{i}$$

Boils down to an estimate using a **single** sample, since x_3^{i} is **identical** for all $i = 1, \ldots, 30$.

Path degeneracy

Path degeneracy follows as a direct consequence of resampling.

The resampling step will by construction result in that for any time s there exists a time $t > s$ such that the PF approximation $\widehat{p}^N(x_{0:t} \mid y_{1:t})$ consists of a single particle at time s.

In the above example this happened for $s = 6$ and $t = 25$.

Mitigating path degeneracy

Mitigating the path degeneracy problem

The impact of the path degeneracy problem can be reduced:

1. Do not resample at each iteration, when?
2. Better resampling algorithms
3. ...

Effective sample size (ESS)

The effective sample size (ESS) N_{eff} is a diagnostics tool that tells us when our weights are problematic in the sense that they are close to being degenerate.

$$N_{\text{eff}} = \frac{N}{\mathbb{E}_q[\omega^2(x^i)]} \leq N.$$

We cannot evaluate N_{eff} exactly, but we can compute an estimate

$$\widehat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^N (w^i)^2}.$$

"ESS-adaptive resampling": When $\widehat{N}_{\text{eff}}$ falls below some threshold N_{thres} we resample the particles, otherwise we continue without resampling.

Ex) Effective sample size (ESS)

Ex. 1) Let $w^i = 1/N$ for all $i = 1, \dots, N$ (independent samples),

$$\widehat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w^i)^2} = \frac{1}{N \times 1/N^2} = N.$$

Ex. 2) Let $w^i = 0$ for $i = 1, \dots, N-1$ and $w^N = 1$ (completely degenerate),

$$\widehat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w^i)^2} = 1.$$

Bootstrap PF with ESS-adaptive resampling

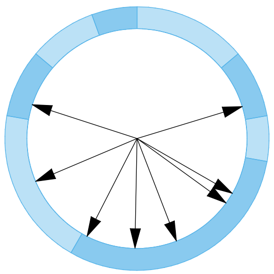
Algorithm 3 joint filtering bootstrap PF (for $i = 1, \dots, N$)

1. **Initialization ($t = 0$):**
   (a) Sample $x_0^i \sim p(x_0)$.
   (b) Set initial weights: $w_0^i = 1/N$.
2. **for** $t = 1$ **to** T **do**
   (a) Compute $\widehat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w_{t-1}^i)^2}$.
   (b) **ESS-adapted resample:** If $\widehat{N}_{\text{eff}} < N_{\text{thres}}$ sample ancestor indices $a_t^i \sim \mathcal{C}(\{w_{t-1}^j\}_{j=1}^N)$ and set $w_{t-1}^i = 1/N$.
   If $\widehat{N}_{\text{eff}} \geq N_{\text{thres}}$ set $a_t^i = i$.
   (c) **Propagate:** sample $x_t^i \sim p(x_t \mid x_{t-1}^{a_t^i})$. $x_{0:t}^i = \{x_{0:t-1}^{a_t^i}, x_t^i\}$.
   (d) **Weight:** compute $\widetilde{w}_t^i = p(y_t \mid x_t^i) w_{t-1}$ and normalize $w_t^i = \widetilde{w}_t^i / \sum_{j=1}^N \widetilde{w}_t^j$.

Multinomial resampling

Multinomial resampling introduced during lecture 4

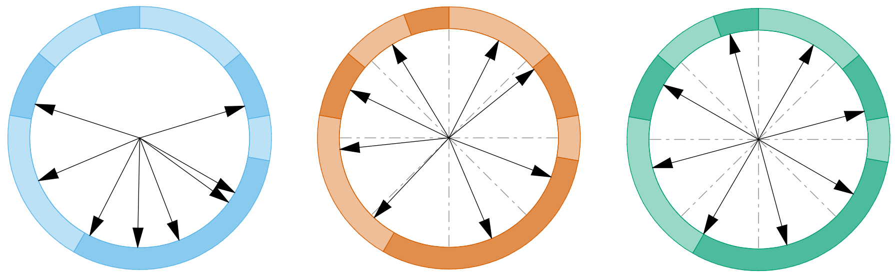
$$a^i \sim \mathcal{C}(\{w^j\}_{j=1}^N), \qquad \mathbb{P}(a^i = j) = w^j.$$



Blue circular disc – weights $\{w^i\}_{i=1}^8$.

Solid arrows – selected particles $\{x^i\}_{i=1}^8$.

Alternative implementations of resampling



Multinomial Stratified Systematic

Divide the circle into strata (grey dashed lines).

Stratified resampling randomly selects 1 sample from each strata.

Systematic resampling randomly generates 1 offset and then it picks one sample from each strata using this offset.

Figures borrowed from the paper L.M. Murray, A. Lee and P.E. Jacob (2016). **Parallel resampling in the particle filter**. *Journal of Computational and Graphical Statistics*. 25(3):789–805, 2016.

Removing the path degeneracy problem

The impact of the path degeneracy problem can sometimes be completely removed by **backward simulation** (results in particle **smoothers**).

Fredrik Lindsten and Thomas B. Schön. **Backward simulation methods for Monte Carlo statistical inference.** *Foundations and Trends in Machine Learning*, 6(1):1-143, 2013.

Fixed-lag smoother

In estimating the fixed-lag smoothing density $p(x_{t-l+1:t} \mid y_{1:t})$ for some small $l > 1$ we can make use of

$$\widehat{p}(x_{t-l+1:t} \mid y_{1:t}) = \sum_{i=1}^{N} w_t^i \delta_{x_{t-l+1:t}^i}(x_{t-l+1:t}),$$

where the particle system comes from a particle filter targeting the joint filtering density.

If l is taken too large we activate the path degeneracy problem to such a degree that it will not work.

The particle MCMC (particle MH and particle Gibbs) algorithms provide good solutions to state smoothing problems.

Parameter inference in SSMs

Nonlinear state space model

$$\begin{aligned} X_t &= f(X_{t-1}, \theta) + V_t, & X_t \mid (X_{t-1} = x_{t-1}, \theta = \theta) &\sim p(x_t \mid x_{t-1}, \theta), \\ Y_t &= g(X_t, \theta) + E_t, & Y_t \mid (X_t = x_t, \theta = \theta) &\sim p(y_t \mid x_t, \theta), \\ X_0 &\sim p(x_0 \mid \theta). & X_0 &\sim p(x_0 \mid \theta). \end{aligned}$$

Two different parameter inference formulations differing in the way the unknown parameters θ are modelled:

- **Maximum likelihood:** θ modelled as **deterministic**.
- **Bayesian:** θ modelled as **stochastic**.

Central object – data distribution/likelihood

The data distribution can be computed by marginalizing

$$p(x_{0:T}, y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid x_t, \theta) \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \theta) p(x_0 \mid \theta)$$

w.r.t. the state trajectory $x_{0:T}$

$$p(y_{1:T} \mid \theta) = \int p(x_{0:T}, y_{1:T} \mid \theta) \mathrm{d}x_{0:T}.$$

Average over all possible values for the state trajectory $x_{0:T}$.

Alternative way of performing the averaging:

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1}, \theta) = \prod_{t=1}^{T} \int p(y_t \mid x_t, \theta) \underbrace{p(x_t \mid y_{1:t-1}, \theta)}_{\text{approx. by PF}} \mathrm{d}x_t$$

A few concepts to summarize lecture 8

Ancestral path: By starting from a particle x_t^i at time t and tracing its ancestors backwards in time via the ancestor indices we obtain $x_{0:t}^i$, which is the ancestral path for particle x_t^i.

Path degeneracy: The resampling step will by construction result in that for any time s there exists a time $t > s$ such that the PF approximation $\widehat{p}(x_{0:t} \mid y_{1:t})$ consists of a single particle at time s.

Effective sample size (ESS): An importance sampling diagnostics tool that tells us when our weights are problematic in the sense that they are close to being degenerate, i.e. it provides a way of gauging the extent of the weight degeneracy.

Backward simulation: Generates samples backwards in time. When backward sampling can be implemented it removes the path degeneracy problem (only possible in off-line situations).

Likelihood function: Deterministic function of θ obtained by inserting the available measurements into the data distribution.