# Particle Filters

Chris Sherlock

November 16, 2005

## 1 Basics

**Notation** $x_t$ = hidden state, $y_t$ =observed data, $\epsilon_t, \nu_t$ = noise

**Assumptions**

- **Markov:** $X_t = f(x_{t-1}, \epsilon_t)$ or equivalently $P(x_t|x_{1:t-1}) = P(x_t|x_{t-1})$ is known.
- **Observed | Hidden:** $Y_t = g(x_t, \nu_t)$ or equivalently $P(y_t|x_t)$ is known.
- **Prior for $X_0$:** $X_0$ has a known prior distribution.

e.g. $X_t$ is the position and velocity of a ship and $y_t$ is its bearing.

**Basic updating formula** Using 'prior' and 'posterior' at time $t$ to refer to the distribution of $X_t$ before and after $Y_t$ is observed we have: Posterior for $X_t \propto$ Likelihood for $Y_t$ given $x_t \times$ Prior for $X_t$

$$P(x_{t+1}|y_{1:t+1}) \propto P(y_{t+1}|x_{t+1}) \; P(x_{t+1}|y_{1:t}) = P(y_{t+1}|x_{t+1}) \int dx_t \; P(x_{t+1}|x_t) P(x_t|y_{1:t}) \qquad (1)$$

Equation (1) presents the posterior at time $t$ in terms of the posterior at time $t-1$. It is the basis for the *Kalman filter* in which it is assumed that

$$\begin{aligned} \mathbf{X}_t &= \mathbf{F}\mathbf{X}_{t-1} + \boldsymbol{\epsilon}_t \\ \mathbf{Y}_t &= \mathbf{G}\mathbf{X}_t + \boldsymbol{\nu}_t \end{aligned}$$

with known matrices $\mathbf{F}, \mathbf{G}$ and Gaussian $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\nu}_t$ with known variance. It is also the basis for the *forward-backward* algorithm (e.g. Baum *et al.*, 1970) where $X_t$ may take only a finite number of values. In more general situations a particle filter approach is used. The following introduction to particle filters is based on Gordon *et al.* (1993) (the SIR filter) and Pitt and Shepherd (1999) (the ASIR filter). It is not intended to be exhaustive!

## 2 The SIR filter

The aim of the SIR (sampling, importance-resampling) filter is to produce for each time $t$ a discrete set of values and associated weights $(x_t^{(i)}, w_t^{(i)})$ such that any posterior expectation may be approximated as

$$E_{X_t|y_{1:t}}[h(X_t)] \approx \sum_{i=1}^{N} h\left(x_t^{(i)}\right) w_t^{(i)}$$

For simplicity of exposition conditioning on '**old data**' $y_{1:t-1}$ (prior knowledge) is **implicit** in probability expressions throughout the remainder of this summary, conditioning on 'new data' $y_t$ or $y_{t+1}$ (posterior knowledge) implies conditioning on all previous $y$'s. Now

$$\begin{aligned} E_{X_t|y_{1:t}}\left[h(X_t)\right] &= \int dx_t\ h(x_t)P\left(x_t|y_t\right) \\ &= \int dx_t\ h(x_t)\frac{P\left(x_t|y_t\right)}{P\left(x_t\right)}P\left(x_t\right) \\ &= \int dx_t\ h(x_t)\frac{P\left(y_t|x_t\right)}{P\left(y_t\right)}P\left(x_t\right) \end{aligned}$$

So

$$E_{X_t|y_{1:t}}\left[h(X_t)\right] = \frac{1}{P\left(y_t\right)}E_{X_t|y_{1:t-1}}\left[h(X_t)P\left(y_t|x_t\right)\right] \approx \frac{1}{N}\frac{1}{P\left(y_t\right)}\sum_{i=1}^{N} h\left(x_t^{(i)}\right) P\left(y_t|x_t^{(i)}\right) \tag{2}$$

Where $x_t^{(1)}, \ldots, x_t^{(N)}$ are a sample from the prior at time $t$ i.e. from the prior for $X_t$. Also

$$P\left(y_t\right) = \int dx_t P\left(x_t\right) P\left(y_t|x_t\right) = E_{X_t|y_{1:t-1}}\left[P\left(y_t|x_t\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N} P\left(y_t|x_t^{(i)}\right) \tag{3}$$

Combining (2) and (3) we obtain

$$E_{X_t|y_{1:t}}\left[h(X_t)\right] \approx \sum_{i=1}^{N} h\left(x_t^{(i)}\right) w_t^{(i)} \tag{4}$$

where

$$w_t^{(i)} = \frac{P\left(y_t|x_t^{(i)}\right)}{\sum_{j=1}^{N} P\left(y_t|x_t^{(j)}\right)} \tag{5}$$

We now consider how an approximate sample from the prior at time $t$ may be used to produce an approximate sample from the prior at time $t+1$.

Using (4) we may obtain a continuous approximation to the prior at time $t+1$:

$$\begin{aligned} P\left(x_{t+1}|y_t\right) &= \int dx_t\ P\left(x_{t+1}|x_t\right) P\left(x_t|y_t\right) \\ &= E_{X_t|y_{1:t}}\left[P\left(x_{t+1}|X_t\right)\right] \\ &\approx \sum_{i=1}^{N} P\left(x_{t+1}|x_t^{(i)}\right) w_t^{(i)} \\ &:= \hat{P}\left(x_{t+1}|y_t\right) \end{aligned}$$

where $x_t^{(1)}, \ldots, x_t^{(N)}$ is a sample from the prior at time $t$ and the $w_t^{(i)}$ are as defined in (5).

To sample from this mixture distribution, first sample the mixture element $i$ from the discrete distribution with weights $w_t^{(i)}$ and then sample from the conditional distribution given the mixture element $P\left(x_{t+1}|x_t^{(i)}\right)$.

Repeat this $N$ times to produce a new sample: we have constructed $N$ particles from the (approximate) prior for $X_{t+1}$ using $N$ particles from the prior for $X_t$.

Since our desired output is the set of values and associated weights, the SIR algorithm is best considered as

1. **Re-sampling:** sample $N$ new particles from the $N$ old particles with discrete distribution $(w_t^{(1)}, \ldots, w_t^{(N)})$
2. **Propogation:** for each resampled particle (from the values at time $t$) sample a new particle value from density $P\left(x_{t+1}|x_t^{(i)}\right)$.
3. **Filtering/Re-weighting:** assign to each particle a weight $w_t^{(i)} \propto P\left(y_t|x_t^{(i)}\right)$

Steps (1) and (2) are equivalent to sampling N times from our approximation to the prior, $\hat{P}(.)$.

### 2.1 Weaknesses

Since stage (1) samples with replacement, the number of particles with distinct ancestors in the original prior monotonically decreases with the number of iterations.

This problem is exacerbated if at any stage the prior is much flatter than the likelihood. In this case particles near the narrow likelihood peak will be given a much greater weight than any others so that most of the other particles will not be resampled. An alternative way of viewing the problem is that since the posterior will closely resemble the likelihood in shape and position, most of the support of the prior, from which we have sampled, plays a minor role in the support of the posterior, and the corresponding particles are relatively unimportant.

Heuristically the prior will be much flatter than the likelihood if either $Var[\epsilon_t]$ is large or $Var[\nu_t]$ is small.

Adding random noise to each particle after the resampling stage (1) (*jittering*) helps to alleviate this problem. It is equivalent to using kernel smoothing after stage (1) to obtain a sample from a continuous distribution.

A further weakness is that we are approximating our prior (and posterior) distributions by mixtures - so these approximations will be poor in their tails (not discussed further here).

## 3 Adapted filters

Given a sample $x_t^{(i)}$ and weights $w_t^{(i)}$ (1) allows us to approximate the posterior density at time $t+1$

$$\hat{P}(x_{t+1}|y_{t+1}) = P(y_{t+1}|x_{t+1}) \sum_{i=1}^{N} P\left(x_{t+1}|x_t^{(i)}\right) w_t^{(i)}$$

In the SIR filter the $x_t^{(i)}$ are draws from the prior for $x_t$ and the weights are proportional to the likelihood $P\left(y_t|x_t^{(i)}\right)$. As already noted, some or many of the weights may be negligible and so each corresponding $x_t^{(i)}$ is probably wasted as it is very unlikely to be sampled from; equivalently the number of points $x_t^{(i)}$ in the main mass of the likelihood may be small. However *we already know* $y_t$ before sampling the $x_t^{(i)}$, so a more efficient set of draws should be possible; such filters are termed *adapted*.

### 3.1 A first attempt

We may extend the derivation of (4) to allow for a sample $\{x_t^{(i)}\}$ from any function $g(x_t|y_t)$:

$$\begin{aligned} E_{X_t|y_{1:t}}\left[h(X_t)\right] &= \int dx_t \; h(x_t) P\left(x_t|y_t\right) \\ &= \int dx_t \; h(x_t) \frac{P\left(x_t|y_t\right)}{P\left(x_t\right)} P\left(x_t\right) \\ &= \int dx_t \; h(x_t) \frac{P\left(y_t|x_t\right) P\left(x_t\right)}{P\left(y_t\right) g(x_t|y_t)} g(x_t|y_t) \\ &= \frac{1}{P\left(y_t\right)} E_{g\left(x_t|y_t\right)}\left[h(x_t) \frac{P\left(y_t|x_t\right) P\left(x_t\right)}{g(x_t|y_t)}\right] \end{aligned}$$

So that (4) still applies but with new weights

$$w_t^{(i)} \propto \frac{P\left(y_t|x_t^{(i)}\right) P\left(x_t^{(i)}\right)}{g(x_t^{(i)}|y_t)} \tag{6}$$

subject to $\sum w_t^{(j)} = 1$.

Our continuous approximation to the prior is again

$$\hat{P}\left(x_{t+1}|y_t\right) = \sum_{i=1}^{N} P\left(x_{t+1}|x_t^{(i)}\right) w_t^{(i)}$$

but with the $x_t^{(i)}$ now sampled from $g(x_t|y_t)$ and the $w_t^{(i)}$ as defined in (6).

Note that if $g(x_t|y_t) = \hat{P}\left(x_t^{(i)}\right)$ then we reduce back to the SIR filter.

The algorithm is now

- **Re-sampling and Propogation:** sample $N$ times from $g(x_t|y_t)$.
- **Reweighting:** calculate a new set of weights from (6).

The third stage described in Pitt and Shepherd (1999), resampling again with the weights as probabilities, is unnecessary and only introduces noise.

If we can choose a $g(x_t|y_t)$ that closely follows the posterior density of $x_t$, $P\left(x_t|y_t\right)$ then the weights will be relatively similar and the particles will be propogated relatively evenly. However to calculate each $w_t^{(i)}$ we must evaluate

$$P\left(x_t^{(i)}\right) \approx \sum_{j=1}^{N} P\left(x_t^{(i)}|x_{t-1}^{(j)}\right) w_{t-1}^{(j)}$$

and so we need $O(N^2)$ evaluations of the Markov probability at each time point, which renders the algorithm too inefficient for practical use.

### 3.2 The ASIR filter

For any sample and associated weights $(x_t^{(1)}, w_t^{(1)}), \ldots, (x_t^{(1)}, w_t^{(1)})$ our continuous approximation to the posterior at time $t+1$ is

$$\begin{aligned}\hat{P}(x_{t+1}|y_{t+1}) &\propto P(y_{t+1}|x_{t+1}) \sum_{i=1}^{N} P\left(x_{t+1}|x_t^{(i)}\right) w_t^{(i)} \\ &= \sum_{i=1}^{N} w_t^{(i)} P\left(y_{t+1}, x_{t+1}|x_t^{(i)}\right) \\ &= \sum_{i=1}^{N} w_t^{(i)} P\left(y_{t+1}|x_t^{(i)}\right) P\left(x_{t+1}|y_{t+1}, x_t^{(i)}\right)\end{aligned}$$

which is a mixture distribution with weights proportional to

$$w_t^{(i)} P\left(y_{t+1}|x_t^{(i)}\right)$$

Consider the above at time $t$ instead of $t+1$: our approximation to the joint posterior distribution of $X_t$ and the mixture element $i_t$ is therefore

$$\hat{P}(x_t, i_t|y_t) \propto P(y_t|x_t) P\left(x_t|x_{t-1}^{(i_t)}\right) w_{t-1}^{(i_t)} = P\left(y_t|x_{t-1}^{(i_t)}\right) P\left(x_t|y_t, x_{t-1}^{(i_t)}\right) w_{t-1}^{(i_t)} \tag{7}$$

Sampling from this and then discarding the mixture label $i_t$ produces a sample from the approximate posterior distribution for $X_t$.

In general this is not possible so let us approximate the joint posterior instead by

$$g(x_t, i_t|y_t) = q\left(x_t|y_t, x_{t-1}^{(i_t)}\right) \beta_{t-1}^{(i_t)}$$

with

$$\sum \beta_{t-1}^{(i)} = 1 \quad \text{and} \quad \int q\left(x_t|y_t, x_{t-1}^{(i_t)}\right) = 1$$

Then provided $\beta_{t-1}^{(i)}$ does not depend on $x_t$, the marginal probability for the $i^{th}$ mixture element is

$$\int dx_t \ \beta_{t-1}^{(i)} q\left(x_t|y_t, x_{t-1}^{(i)}\right) = \beta_{t-1}^{(i)}$$

So we may sample from $g(x_t, i_t|y_t)$ by choosing the mixture element with probability $\beta_{t-1}^{(i_t)}$ and then simulating $x_t$ from the transition density given the mixture element and the latest observation $q\left(x_t|y_t, x_{t-1}^{(i_t)}\right)$. Discarding the label we obtain a sample from $g(x_t|y_t)$ rather than $P(x_t|y_t)$ so any expectation will need to be reweighted. Consider an approximate posterior expectation at time $t$.

$$\begin{aligned}E_{X_t|y_{1:t}}[h(X_t)] &= \int dx_t \ h(x_t) \sum_{i=1}^{N} P(x_t, i|y_t) \\ &= \sum_{i=1}^{N} \int dx_t \ h(x_t) \frac{P(x_t, i|y_t)}{g(x_t, i|y_t)} g(x_t, i|y_t)\end{aligned}$$

$$
\begin{aligned}
&\propto E_{g(x_t,i|y_t)}\left[h(x_t)\frac{P\left(y_t|x_t\right)P\left(x_t|x_{t-1}^{(i)}\right)w_{t-1}^{(i)}}{q\left(x_t|y_t,x_{t-1}^{(i)}\right)\beta_{t-1}^{(i)}}\right]\\
&\approx \sum_{j=1}^{N} h(x_t^{(j)})\frac{P\left(y_t|x_t^{(j)}\right)P\left(x_t^{(j)}|x_{t-1}^{\left(i_t^{(j)}\right)}\right)w_{t-1}^{\left(i_t^{(j)}\right)}}{q\left(x_t^{(j)}|y_t,x_{t-1}^{\left(i_t^{(j)}\right)}\right)\beta_{t-1}^{\left(i_t^{(j)}\right)}}\\
&= \sum_{j=1}^{N} h(x_t^{(j)})\, w_t^{(j)}
\end{aligned}
$$

where $\left(x_t^{(1)},i_t^{(1)}\right),\ldots,\left(x_t^{(N)},i_t^{(N)}\right)$ are a sample from $g(x_t,i_t|y_t)$ and

$$
w_t^{(j)} \propto \frac{P\left(y_t|x_t^{(j)}\right)P\left(x_t^{(j)}|x_{t-1}^{\left(i_t^{(j)}\right)}\right)w_{t-1}^{\left(i_t^{(j)}\right)}}{q\left(x_t^{(j)}|y_t,x_{t-1}^{\left(i^{(j)}\right)}\right)\beta_{t-1}^{\left(i_t^{(j)}\right)}} \tag{8}
$$

Note that the weight now depends on where the particle was sampled from and therefore potentially on the whole history of the particle, through $w_{t-1}^{\left(i_t^{(j)}\right)}$ .

The ASIR algorithm is therefore

- **Re-sampling and Propogation:** sample $N$ times from $g(x_t,i_t|y_t)$ . For each pair $\left(X_t^{(j)},i_t^{(j)}\right)$ first sample $i_t$ from the discrete distribution with probabilities $\left(\beta_{t-1}^{(1)},\ldots,\beta_{t-1}^{(N)}\right)$ and then sample $X_t$ from $q\left(x_t|y_t,x_{t-1}^{(i_t)}\right)\beta_{t-1}^{(i_t)}$.
- **Reweighting:** calculate a new set of weights from (8).

We now consider specific instances of the above algorithm. Set

$$
\beta_{t-1}^{(i)} \propto w_{t-1}^{(i)}P\left(y_t|x_{t-1}^{(i)}\right) \tag{9}
$$

$$
q\left(x_t|y_t,x_{t-1}^{(i)}\right) = P\left(x_t|y_t,x_{t-1}^{(i)}\right) \tag{10}
$$

then by (7) all the $w_t^{(j)}$ are equal. This is the optimal solution, however it is only possible to calculate the above probabilities exactly in specific cases such as the non-linear Gaussian measurement model:

If $X_t|x_{t-1} \sim N\left(\mu(x_{t-1}),\sigma^2(x_{t-1})\right)$ and $Y_t|x_t \sim N(x_t,1)$ then clearly

$$
Y_t|x_{t-1} \sim N\left(\mu(x_{t-1}),1+\sigma^2(x_{t-1})\right)
$$

Also

$$
P\left(x_t|y_t,x_{t-1}^{(i)}\right) \propto P\left(x_t,y_t|x_{t-1}^{(i)}\right) = P\left(y_t|x_t\right)P\left(x_t|x_{t-1}\right)
$$

which is Gaussian. So the optimal weights may be calculated exactly. Similarly if $P\left(x_t|x_{t-1}\right)$ is log-concave then it may be approximated by a Gaussian and a sample with near-optimal weights obtained.

Alternatively, and more generically, substitute

$$\begin{aligned} \beta_{t-1}^{(i)} &\propto w_{t-1}^{(i)} P\left(y_t|\mu_t^{(i)}\right) \\ q\left(x_t|y_t, x_{t-1}^{(i)}\right) &= P\left(x_t|x_{t-1}^{(i)}\right) \end{aligned}$$

where $\mu_t^{(i)}$ is 'the mean, mode, a draw, or some other likely value' associated with $P\left(x_t|x_{t-1}^{(i)}\right)$. Our new weights are

$$w_t^{(j)} \propto \frac{P\left(y_t|x_t^{(j)}\right) P\left(x_t^{(j)}|x_{t-1}^{\left(i_t^{(j)}\right)}\right) w_{t-1}^{\left(i_t^{(j)}\right)}}{q\left(x_t^{(j)}|y_t, x_{t-1}^{\left(i_t^{(j)}\right)}\right) \beta_{t-1}^{\left(i_t^{(j)}\right)}} = \frac{P\left(y_t|x_t^{(j)}\right)}{P\left(y_t|\mu_t^{\left(i_t^{(j)}\right)}\right)}$$

# References


Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximisation technique occuring in the statistical analysis of probababilstic functions of Markov chains. *The Annals of Mathematical Statistics* **41**, 164–171.

Gordon, N., Salmond, D. and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE proceedings-F* **140**, 107–113.

Pitt, M. K. and Shepherd, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* **94**, 590–599.