Automatic differentiation via off-policy particle filters

October 10, 2023

Abstract

Kevin has checked numerically that the measurement off policy (MOP) particle filter has a fixed-seed derivative matching Scibior & Wood (SW). Here, we present the MOP algorithm in order to check that it is properly weighted. The argument for this remains somewhat heuristic, reasoning about properly weighted samples in a similar way to Jun Liu's 2001 book. We also introduce a MOP-α algorithm whose fixed-seed derivative interpolates between the basic ignore-resampling-in-derivative (IRID) PF ($\alpha = 0$) and MOP ($\alpha = 1$).

1 The Measurement off policy (MOP) particle filter

Algorithm 1: Measurement off policy (MOP) SMC:

Initialize filter particles: simulate $X_{0,j}^{F,\theta} \sim f_{X_0}(\,\cdot\,;\theta)$ for j in $1\!:\!J$
Initialize relative weights: $w_{0,j}^{F,\theta} = 1$ for j in $1\!:\!J$
for n *in* $1\!:\!N$ **do**
    Simulate for prediction: $X_{n,j}^{P,\theta} \sim f_{X_n|X_{n-1}}\big(\,\cdot\,|X_{n-1,j}^{F};\theta\big)$ for j in $1\!:\!J$
    Prediction weights: $w_{n,j}^{P,\theta} = w_{n-1,j}^{F,\theta}$ for j in $1\!:\!J$
    Evaluate measurement density: $g_{n,j}^{\theta} = f_{Y_n|X_n}(y_n^*|X_{n,j}^{P,\theta};\theta)$ for j in $1\!:\!J$
    Before-resampling conditional likelihood: $L_n^{B,\theta} = \dfrac{\sum_{j=1}^J g_{n,j}^{\theta} w_{n,j}^{P,\theta}}{\sum_{j=1}^J w_{n,j}^{P,\theta}}$
    Conditional likelihood under ϕ: $L_n^{\phi} = \frac{1}{J}\sum_{m=1}^J g_{n,m}^{\phi}$
    Normalize weights: $\tilde{g}_{n,j}^{\phi} = \dfrac{g_{n,j}^{\phi}}{J L_n^{\phi}}$ for j in $1\!:\!J$
    Apply systematic resampling to select indices $k_{1:J}$ with $\text{Prob}\big(k_j = m\big) = \tilde{g}_{n,m}^{\phi}$
    Resample particles: $X_{n,j}^{F,\theta} = X_{n,k_j}^{P,\theta}$
    Filter weights corrected for resampling: $w_{n,j}^{FC,\theta} = w_{n,j}^{P,\theta} \times \dfrac{g_{n,j}^{\theta}}{g_{n,j}^{\phi}}$ for j in $1\!:\!J$
    Resample filter weights: $w_{n,j}^{F,\theta} = w_{n,k_j}^{FC,\theta}$ for j in $1\!:\!J$
    After-resampling conditional likelihood: $L_n^{A,\theta} = L_n^{\phi}\dfrac{\sum_{j=1}^J w_{n,j}^{F,\theta}}{\sum_{j=1}^J w_{n,j}^{P,\theta}}$
end

- MOP-SMC requires that the algorithm is first run at ϕ, for which it is a vanilla particle filter. $g^{\phi}_{n,j}$ and $\tilde{g}^{\phi}_{n,j}$ are computed at this first pass. Then it is run at θ, with the seed kept fixed.

- Here, the resampling rule for particle j depends on j (and therefore ϕ and $X^{P,\phi}_{n,j}$) but not on θ or $X^{P,\theta}_{n,j}$. This does not have the standard form for a weighted particle filter, for which we can have general reweighting rules but they are usually a funciton of the location of the particle and the model parameter, θ. That may be why it is harder to see that this is a properly weighted filter. Nevertheless, inspection of the algorithm shows that each particle is properly reweighted to account for its resampling probability, in step 12, so it should be properly weighted.

- The final estimate of the likelihood is either based on the after-resampling conditional likelihood estimate
$$L^A(\theta) = \left(\frac{1}{J}\sum_{j=1}^{J} w^{F,\theta}_{N,j}\right)\prod_{n=1}^{N} L^{\phi}_n \tag{1}$$
or the before-resampling estimate,
$$L^B(\theta) = \prod_{n=1}^{N} \frac{\sum_{j=1}^{J} w^{P,\theta}_{N,j}\, g^{\theta}_{N,j}}{\sum_{j=1}^{J} w^{P,\theta}_{N,j}}. \tag{2}$$
with $L^B(\theta)$ presumably having slightly lower variance.

- Weighted samples representing the filter distribution, $f_{X_n|Y_{1:n}}(x_n \,|\, y^*_{1:n}\,;\theta)$ are either $\big\{\big(X^{F,\theta}_{n,j}, w^{F,\theta}_{n,j}\big), j \text{ in } 1\!:\!J\big\}$ or $\big\{\big(X^{P,\theta}_{n,j}, g^{\theta}_{n,j}\, w^{P,\theta}_{n,j}\big), j \text{ in } 1\!:\!J\big\}$, meaning that an expectation over $f_{X_n|Y_{1:n}}(x_n \,|\, y^*_{1:n}\,;\theta)$ is consistently estimated by a corresponding weighted average of the filter or prediction particles.

- As long as `rprocess` is a continuously differentiable function of θ for fixed seed, and `dmeasure` is a continuously differentiable function of θ, and $g^{\phi}_{n,j} \neq 0$, we see that MOP is a continuously differentiable function of θ for fixed seed. Since it also provided an unbiased estimate of the likelihood, this justifies exchanging the order of differentiation and integration to ensure that its derivative is an unbiased estimate of the deriative of the likelihood.

- Taking the derivative with respect to θ at $\theta = \phi$, step 12 looks very much like the stop gradient approach.

Here's an outline of an argument explaining why MOP is properly weighted.

Lemma 1. *Suppose that $\{(X_j, u_j), j = 1, \ldots, J\}$ is properly weighted for f_X, meaning that, if $X \sim f_X$,*
$$E[h(X)] \approx \sum_{j=1}^{J} \frac{w_j}{\sum_{k=1}^{J} u_k} h(X_j),$$
for some appropriate formalization of $\approx$. Now, let $\{(Y_j, v_j), j = 1, \ldots, J\}$ be a sample drawn from $\{(X_j, u_j)\}$ where (X_j, u_j) is represented, on average, $\pi_j J$ times. This could amount to binomial resampling having J draws each with probability p_j, or systematic resampling. Suppose
$$(Y_j, v_j) = \big(X_{a(j)}, u_{a(j)}/\pi_{a(j)}\big),$$
where a is called the ancestor function. Then, $\{(Y_j, v_j), j = 1, \ldots, J\}$ is properly weighted for f_X.

Notably, Lemma 1 permits $\pi_{1:J}$ to depend on $\{(X_j, u_j)\}$ as long as the resampling is carried out independently of $\{(X_j, u_j)\}$ conditional on $\pi_{1:J}$.

Lemma 2. *Suppose that $Z_j \sim f_{Z|X}(\cdot|X_j)$ where $f_{Z|X}$ is a conditional probability density function corresponding to a joint density $f_{X,Z}$ with marginal densities f_X and f_Z. Then, $\{(Z_j, u_j)\}$ is properly weighted for f_Z.*

Lemma 3. *Suppose that $(X'_j, u'_j) = \big(X_j, u_j\, f_{Z|X}(z^*|X_j)\big)$. Then, $\{(X'_j, u'_j)\}$ is properly weighted for $f_{X|Z}(\cdot|z^*)$.*

Recursively applying Lemmas 1, 2 and 3, we obtain that the MOP filter is properly weighted. Specifically, suppose inductively that $\big\{\big(X^{F,\theta}_{n-1,j}, w^{F,\theta}_{n-1,j}\big)\big\}$ is properly weighted for $f_{X_{n-1}|Y_{1:n-1}}(x_{n-1}|y^*_{1:n-1};\theta)$. Then, Lemma 2 tells us that $\big\{\big(X^{P,\theta}_{n,j}, w^{P,\theta}_{n,j}\big)\big\}$ is properly weighted for $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)$. Lemma 3 tells us that $\big\{\big(X^{P,\theta}_{n,j}, w^{P,\theta}_{n,j} g^{\theta}_{n,j}\big)\big\}$ is therefore properly weighted for $f_{X_n|Y_{1:n}}(x_n|y^*_{1:n};\theta)$. Lemma 1 guarantees that the resampling rule, given by

$$\big(X^{F,\theta}_{n,j}, w^{F,\theta}_{n,j}\big) = \big(X^{P,\theta}_{n,a(j)}, w^{P,\theta}_{n,j} g^{\theta}_{n,j} / g^{\phi}_{n,j}\big),$$

with resampling weights proportional to $g^{\phi}_{n,j}$, is therefore also properly weighted for $f_{X_n|Y_{1:n}}(x_n|y^*_{1:n};\theta)$.

This has addressed filtering, but not quite yet the likelihood evaluation. For this we use the following lemma.

Lemma 4. *$f_{Y_n|Y_{1:n-1}}(y^*_n|y^*_{1_n-1};\theta)$ is properly estimated by either the before-resampling estimate,*

$$L^{B,\theta}_n = \frac{\sum_{j=1}^J g^{\theta}_{n,j} w^{P,\theta}_{n,j}}{\sum_{j=1}^J w^{P,\theta}_{n,j}}, \tag{3}$$

or by the after-resampling estimate,

$$L^{A,\theta}_n = L^{\phi}_n \frac{\sum_{j=1}^J w^{F,\theta}_{n,j}}{\sum_{j=1}^J w^{P,\theta}_{n,j}}. \tag{4}$$

where L^{ϕ}_n comes from step 8.

Here, (3) is a direct consequence of our earlier result that $\{\big(X^{P,\theta}_{n,j}, w^{P,\theta}_{n,j}\big)\}$ is properly weighted for $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)$. To see (4), we write the numerator of (3) as

$$L^{\phi}_n \sum_{j=1}^J \left[\frac{g^{\theta}_{n,j}}{g^{\phi}_{n,j}} w^{P,\theta}_{n,j}\right] \frac{g^{\phi}_{n,j}}{L^{\phi}_n}$$

Using Lemma 1, we see this is properly estimated by

$$L^{\phi}_n \sum_{j=1}^J w^{F,\theta}_{n,j},$$

from which we obtain (4).

Using Lemma 4, we obtain a likelihood estimate,

$$L^{A,\theta} = \prod_{n=1}^N \left(L^{\phi}_n \frac{\sum_{j=1}^J w^{F,\theta}_{n,j}}{\sum_{j=1}^J w^{P,\theta}_{n,j}} \right).$$

Since $w^{F,\theta}_{n,j} = w^{P,\theta}_{n+1,j}$, this is a telescoping product. The remaining terms are $\sum_{j=1}^J w^{P,\theta}_{0,j} = J$ on the denominator and $\sum_{j=1}^J w^{F,\theta}_{N,j}$ on the numerator. This derives the MOP estimate in (1).

$L^{B,\theta}$ should generally be preferred, since there is no reason to include the extra variability from resampling when calculating the conditional log likelihood, but it lacks the nice telescoping product.

Algorithm 2: Measurement off policy (MOP-α) SMC:

1 Initialize filter particles: simulate $X_{0,j}^{F,\theta} \sim f_{X_0}(\,\cdot\,;\theta)$ for j in $1\!:\!J$
2 Initialize relative weights: $w_{0,j}^{F,\theta} = 1$ for j in $1\!:\!J$
3 **for** n *in* $1\!:\!N$ **do**
4   Simulate for prediction: $X_{n,j}^{P,\theta} \sim f_{X_n|X_{n-1}}\big(\,\cdot\,|X_{n-1,j}^{F};\theta\big)$ for j in $1\!:\!J$
5   Prediction weights with discounting: $w_{n,j}^{P,\theta} = \big(w_{n-1,j}^{F,\theta}\big)^{\alpha}$ for j in $1\!:\!J$
6   Evaluate measurement density: $g_{n,j}^{\theta} = f_{Y_n|X_n}(y_n^*|X_{n,j}^{P,\theta};\theta)$ for j in $1\!:\!J$
7   Before-resampling conditional likelihood: $L_n^{B,\theta,\alpha} = \dfrac{\sum_{j=1}^{J} g_{n,j}^{\theta} w_{n,j}^{P,\theta}}{\sum_{j=1}^{J} w_{n,j}^{P,\theta}}$
8   Conditional likelihood under ϕ: $L_n^{\phi} = \frac{1}{J}\sum_{m=1}^{J} g_{n,m}^{\phi}$
9   Normalize weights: $\tilde{g}_{n,j}^{\phi} = \dfrac{g_{n,j}^{\phi}}{J L_n^{\phi}}$ for j in $1\!:\!J$
10   Apply systematic resampling to select indices $k_{1:J}$ with $\mathrm{Prob}\big(k_j = m\big) = \tilde{g}_{n,m}^{\phi}$
11   Resample particles: $X_{n,j}^{F,\theta} = X_{n,k_j}^{P,\theta}$
12   Filter weights corrected for resampling: $w_{n,j}^{FC,\theta} = w_{n,j}^{P,\theta} \times \dfrac{g_{n,j}^{\theta}}{g_{n,j}^{\phi}}$ for j in $1\!:\!J$
13   Resample filter weights: $w_{n,j}^{F,\theta} = w_{n,k_j}^{FC,\theta}$ for j in $1\!:\!J$
14   After-resampling conditional likelihood: $L_n^{A,\theta,\alpha} = L_n^{\phi}\,\dfrac{\sum_{j=1}^{J} w_{n,j}^{F,\theta}}{\sum_{j=1}^{J} w_{n,j}^{P,\theta}}$
15 **end**

2 MOP-α

- MOP has been rewritten from an earlier draft so that now the only difference between MOP and MOP-α is the discounting exponent α in step 5.
- When $\alpha = 1$ we recover MOP.
- When $\alpha = 0$, $w_{n,j}^{P,\theta} = 1$ and so $L_n^{B,\theta}$ corresponds to an equally weighted conditional likelihood, as in a regular particle filter, but with resampling carried out according to ϕ not θ. When calculating a derivative with respect to θ at $\theta = \phi$, this corresponds to ignoring differentiation through the resampling, i.e., the fixed-seed derivative of the PF algorithm. To see this, we notice that the fixed seed derivative which "ignores" resampling is not really ignoring it; in an infinitesimal neighborhood of ϕ, the resampling is constant and so the derivative is zero. The problem is not that the fixed-seed derivative is missing something, just that we cannot correctly move the derivative through an expectation for a discontinuous function.
- When $\alpha < 1$, the product

$$L^{A,\alpha}(\theta) = \prod_{n=1}^{N} L_n^{A,\theta,\alpha}$$

no longer has a telescoping cancellation.

3 Unresolved questions

- $L_n^{A,\theta}$ and $L_n^{B,\theta}$ are asymptotically equivalent as J increases, but are not identical. However, both are identical for a particle filter (i.e., MOP when $\theta = \phi$). Presumably, this could allow their fixed-seeed derivatives to be the same at $\theta = \phi$, though it is not obvious if this is true. In that case, we would have an alternative explanation of the telescoping identity of Scibior & Wood.
- It seems possible to carry out an extra numerical pass where one recalculates $L^A(\theta)$ taking advantage of the cancellation, which could make the derivative faster and/or more memory efficient, depending on what is the bottleneck. This may not be an immediate priority.
- The cancellation difference between $L_n^{A,\theta}$ and $L_n^{B,\theta}$ is moot for $\alpha < 1$, so there's not much reason not to use $L_n^{B,\theta}$.