Particle rejuvenation of Rao-Blackwellized Sequential Monte Carlo smoothers for Conditionally Linear and Gaussian models

Ngoc Minh Nguyen*, Sylvain Le Corff† and Eric Moulines‡

The authors are grateful to the associate editor and the anonymous referees for their comments and remarks to improve the manuscript. These suggestions have been taken into consideration to propose a revision of the original paper. It has also been carefully proof read to remove remaining typos and minor mistakes. We provide below detailed answers for all comments.

Associate editor

The paper surveys recent Rao-Blackwellized particle smoothers and presents a particle rejuvenation approach. The particle rejuvenation is interesting and seems useful, but it also shares some resemblance with the MCMC-based rejuvenation strategy proposed by [LBS+16, Section V.B] in the context of Rao-Blackwellized FFBS and with the strategy proposed by [FWT10] in the context of two-filter smoothing. Please relate the proposed method to these prior works. The detailed comments by the reviewers need to be addressed as well.

It is true that our rejuvenation step shares some resemblance with the algorithms of [FWT10] and [LBS+16, Section V.B]. As noted in [FWT10, Secion 2.6], two-filter smoothers are prone to suffer from degeneracy issues when the algorithm associates forward particles at time $i-1$ with backward particles at time i. The authors illustrate this issue in the case where the hidden state is an AR(2) process. To overcome the weakness of such standard two-filter approaches, [FWT10] samples new particles at time i conditional on forward particles up to time $i-1$ and on backward particles from time $i+1$ which are appropriately weighted. This allows to produce new particles at time i and to obtain a SMC approximation whose support is not restricted to the backward particles at time $i+1$ as in the original two-filter algorithm. The particle rejuvenation proposed in our paper exploits this idea in the specific case of linear and Gaussian models where explicit computations allow to produce an approximation using $(a^k_{1:i-1})_{1\leq k\leq N}$ and $(\tilde{a}^k_{i+1:n})_{1\leq k\leq N}$ with support $\{1,\dots,J\}$. In addition, it produces a direct approximation of the target distribution based on $(a^k_{1:i-1})_{1\leq k\leq N}$ and $(\tilde{a}^k_{i+1:n})_{1\leq k\leq N}$ without any additional importance sampling steps.

On the other hand, another modification of the FFBS algorithm based on a Markov chain Monte Carlo sampling step was introduced in [LBS+16]. Instead of sampling from explicitly from (12), [LBS+16] proposed to draw a forward path $a_{1:i-1}$ in $(a^k_{1:i-1})_{1\leq k\leq N}$ and a sate a_i in $\{1,\dots,J\}$

*LTCI, CNRS and Télécom ParisTech, 46 rue Barrault 75634 Paris Cedex 13, France.
†Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.
‡Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France.

according to:

$$\widetilde{q}(a_{1:i}|a_{i+1:n}, y_{1:n}) = \sum_{k=1}^{N} \widetilde{\vartheta}_{i-1}^{k} \widetilde{q}(a_i|a_{1:i-1}^{k}, a_{i+1:n}, y_{1:n}) \delta_{a_{1:i-1}^{k}}(a_{1:i-1}) \,,$$

where $(\widetilde{\vartheta}_{i-1}^{k})_{1\leq k\leq N}$ *are adjustment multipliers and* $\widetilde{q}(a_i|a_{1:i-1}^{k}, a_{i+1:n}, y_{1:n})$ *is a proposal kernel chosen by the user. This means that an ancestral path* $a_{1:i-1}^{\star}$ *is sampled in* $(a_{1:i-1}^{k})_{1\leq k\leq N}$ *with weights* $(\widetilde{\vartheta}_{i-1}^{k})_{1\leq k\leq N}$ *and* $a_i^{\star}$ *is sampled from* $\widetilde{q}(\cdot|a_{1:i-1}^{\star}, a_{i+1:n}, y_{1:n})$*. Then, the proposed sequence* $a_{1:i}^{\star}$ *is accepted or rejected using the usual Metropolis-Hastings acceptance ratio. The choice of MCMC rejuvenation has interesting practical consequences as the computation of the acceptance ratio only requires to compute the posterior probability (11) for the proposed sequence* $a_{1:i}^{\star}$ *while our technique is based on the computation of (11) for all combinations of sequences* $(a_{1:i-1}^{k})_{1\leq k\leq N}$ *and states* $a_i \in \{1, \ldots, J\}$*. The method proposed in [LBS+16] is computationally less costly as its complexity depends mainly on the number of MCMC samples produced by the algorithm. However it also requires a fair amount of tuning to choose efficient adjustment multipliers and proposal kernels. Our proposed method is computationally more intensive but allows to sample explicitly from (12) without additional tuning parameters. Note that both methods provide similar results for the same computational costs. These remarks were added at the end of Section 2.2 and Section 2.3.1.*

Referee 1

A clear presentation of the state-of-the-art algorithms, e.g. what posterior pdf incl. linear states each algorithm (RB-FFBS and RB Two-filter smoother) tries to approximate, as well as a full description of these algorithms. In the current manuscript, the authors focus entirely on the SMC approximation part of the RB filters, and the modifications they intend to do (it is not really mentioned that the linear Gaussian part of the problem is not going to be presented, apologies if I missed it)

The overall presentation of the paper has been modified to take this comment into account. First, the paper is not presented as a survey paper anymore as we want to focus on fully Rao-Blackwellized SMC approximations only and to highlight the improvements that can be made with simple rejuvenation steps. Therefore, we only mention the algorithm based on Kim's approximation in the introduction but detail only two-filter and FFBS based algorithms. In addition the Rao-Blackwellized two-filter of [BDM10] has been much more detailed in Section 2.3.1 before introducing the particle rejuvenation step in Section 2.3.2. For both algorithms a specific section is devoted to the original method and another section to the rejuvenation step. In addition, these rejuvenation steps are compared to other two-filter and FFBS improvements proposed in the literature.

Some (more) intuition, about concepts and why certain steps are performed. Especially in the presentation of the RBTF in Section 2.4 I had problems to follow the overall idea. E.g. where do we integrate out the linear states in the backward pass of the original formulation presented by [BDM10]. I would expect this description before the modification of particle rejuvenation is presented. Also, some intuition about the explicit marginalization proposed in (9) would be helpful. Why is it chosen, what is the benefit compared to the original formulation? Currently, one gets easily lost in the derivations without maintaining the overall picture.

This comment helped to greatly improved the readability of the paper. As explained in the previous

paragraph, Section 2.3.1 is now devoted to the original Rao-Blackwellized two-filter method and Section 2.3.2 to the rejuvenation step. The idea supporting two-filter based algorithms is clearly presented in Section 2.3.1 and most of the computations have also been postponed to the appendix in order to highlight the overall picture and avoid technicalities. Some additional motivations to introduce rejuvenation are given at the end of Sections 2.2 and 2.3.2 (see the answer to the associate editor). Two-filter smoothers are prone to suffer from degeneracy issues when the algorithm associates forward particles at time $i-1$ with backward particles at time i. The particle rejuvenation proposed in our paper exploits the idea introduced in [FWT10] in the specific case of linear and Gaussian models where explicit computations allows to produce an approximation using $(a^k_{1:i-1})_{1\leq k\leq N}$ and $(\tilde{a}^k_{i+1:n})_{1\leq k\leq N}$ with support $\{1,\dots,J\}$ and without any additional sampling steps.

Some comparisons in terms of performance, complexity, detailing the superiority (or not) of one method over the other. E.g. what are the pros and cons of RB-FFBS and RBTF. Also some numerical comparisons in the simulation section are expected.

Even is the paper is not a survey of all Rao-Blackwellized methods for CLGM anymore, some additional numerical experiments were added in Section 3. The simulation study which was previously given in Section 3 was devoted either to FFBS or two-filter based methods. Now, both algorithms are used and compared to their counterpart with rejuvenation. Our main objective is then to show the improvement of the particle rejuvenation steps for both methods. This also confirms that FFBS algorithm performs better even with a sharp choice for the γ_i's but also that rejuvenation steps improve both the accuracy and the variance of each method.

The minor comments were taken into consideration and typos were corrected. Some specific details are given below.

The authors talk about two FFBS algorithms that they compare to the FFBS with particle rejuvenation. But which two? Kim's and Lindsten's? Some more details would be helpful, also which forward filter has been used.

The two FFBS algorithms are the original version of [LBS+16] and the algorithm with rejuvenation. A first (costly) run of the FFBS algorithm with rejuvenation with a great number of particles ($N = 5000$) is used as a benchmark value. Then the posterior distributions are estimated by the original methods and the methods with rejuvenation with fewer particles and compared to this benchmark run.

Figure 1 shall be Table 1? In the text it is 2500 iterations, in the Table caption it is 3500? How is Std. Dev obtained? I mean is this simulated data or real data where the true parameters θ are usually unknown?

The estimated values were indeed obtained after 2500 iterations of the algorithm. The standard deviations are obtained with 50 independent Monte Carlo runs of the estimation algorithm (and the estimated value is set as the mean over all runs). This is now stated clearly in the paper.

p.7, L.51: Is Ω_i^{-1}, Λ_i correct? In the paper by Lindsten it is Ω_i, Λ_i^{-1}.

The equation is correct. The difference comes from the fact that $\|z\|_A = z'Az$ in the paper by

Linddsten and $\|z\|_A = z'A^{-1}z$ *in our paper.*

Referee 2

It is well known that SMC algorithms, in their most basic form, are numerically unstable when used for approximating distributions defined on path spaces of increasing dimension. This is due to the particle path degeneracy phenomenon caused by the resampling mechanism of SMC algorithms. The degeneracy appears to be a potential problem for the estimators discussed in the paper, since the approximation of the backward kernel (on, e.g., p. 8, line 52) involves quantities such as the mean and the covariance matrix of the conditional distribution of z_i given the states $a_{1:i}$ and the observations, evaluated along the particle genealogical paths. I miss a discussion on this. If the effect exists, does the rejuvenation proposed by the authors counteract the same? (I do not think so).

As explained to the associate editor and to Referee 1, it is true that particle smoothers are prone to degenerate when they are based on the genealogy of the forward or backward particles. In addition, in the specific case of two-filter based smoothers, [FWT10, Section 2.6] noted that these algorithms are prone to suffer from degeneracy issues when the algorithm associates forward particles at time $i-1$ *with backward particles at time* i*. The particle rejuvenation proposed in our paper produces an approximation using* $(a^k_{1:i-1})_{1\leq k\leq N}$ *and* $(\tilde{a}^k_{i+1:n})_{1\leq k\leq N}$ *with support* $\{1,\ldots,J\}$*. In the specific case of linear and Gaussian models, Kalman filtering techniques allow to compute explicitly the approximation of the target distribution at time* i *based on* $(a^k_{1:i-1})_{1\leq k\leq N}$ *and* $(\tilde{a}^k_{i+1:n})_{1\leq k\leq N}$*.*

Therefore, the support of the approximated distribution is not restricted to $(\tilde{a}^k_i)_{1\leq k\leq N}$ *to overcome particle depletion. Marginalization over the linear states builds new trajectories of the form* $(a^k_{1:i-1}, j, \tilde{a}^\ell_{i+1:n})_{1\leq k,\ell\leq N}$ *for all* $1\leq j\leq J$ *which have not been sampled by the forward and backward filters. Although this does not overcome the overall degeneracy issue, numerical experiments highlight that this improves the performance of all Rao-Blackwellized SMC smoothers. Additional comments were added at the end of Section 2.2 and Section 2.3.1.*

The rejuvenation approach proposed in Section 2.3 makes sense in that it allows the production of backward draws that are not restricted to the discrete support formed by the particles generated in the forward filtering pass. On the other hand, it is not obvious that there is a lot of variance to be gained in doing this (even though the simulation study in Section 3 indicates some improvement). Thus, I miss a heuristic motivation for the rejuvenation step in Section 2.3. Are there specific models for which the rejuvenation step is supposed improve significantly the FFBS algorithm? As far as I understand, there is a price to pay for the rejuvenation in terms of some additional computational work, as the Kalman filter needs to be run for all the possible extensions of each particle path. So, is the particle rejuvenation worthy of this additional effort? I guess it depends, e.g., on the number J of possible regimes?

The intuition supporting particle rejuvenation is explained in the answers to the associate editor and to Referee 1. There is no theoretical result to prove that this approach provides a better approximation of the target distributions. However, our numerical experiments hint that this actually improves the accuracy and variance of the proposed estimators which would be even more interesting with additional regimes or additional rejuvenation steps. [FWT10, Section 2.6] illustrated the degeneracy issue of two-filter based smoothers with the example of an AR(2) process. Our rejuvenation exploits the idea introduced in [FWT10, Section 2.6] to overcome degeneracy issues in

the context of CLGM. In addition, the additional computational cost is not prohibitive as Kalman filtering steps are much faster than the combination of the forward and backward particles whose complexity grows with N^2 (but who is stil necessary when no rejuvenation step is introduced).

It took a while before I understood the backward filter algorithm in Section 2.4. More specifically, the expression of the importance weights on p. 11 seems to be based on the identity

$$\tilde{p}_i(a_{i:n}|y_{i:n}) \propto \left\{ \prod_{u=i}^{n-1} Q(a_u, a_{u+1}) \right\} \int \gamma_i(a_i, z_i) p(y_{i:n}|z_i, a_{i:n}) \mathrm{d}z_i \,.$$

which never appears explicitly in the description of the backward filter. I would recommend the authors to include a derivation of the backward filter on the basis of this identity.

The identity was added in Section 2.3.1 in a more detailed derivation of the backward filter. We thank Referee 2 for this help to improve the presentation of the backward filter.

Finally, I miss a discussion on the last sampling operation of the Rao-Blackwellized two-filter smoother, i.e., the sampling of the mixture obtained when plugging the approximations into (9). This should yield a mixture with N^2 components, implying a not innocent computational complexity when N is large.

The additional computational cost of the rejuvenation step comes from the Kalman filtering steps needed to extend all backward trajectories (built with particles). These steps are much faster than the combination of the forward and backward particles whose complexity grows with N^2. However, this combination is also necessary in the original two-filter algorithm of [BDM10] with no rejuvenation. A comment was added in the numerical section.

All minor comments were taken into consideration and typos were corrected.