

Adaptive methods for sequential importance sampling with application to state space models

Julien Cornebise · Éric Moulines · Jimmy Olsson

Received: 28 February 2008 / Accepted: 11 July 2008 / Published online: 15 August 2008
© Springer Science+Business Media, LLC 2008

Abstract In this paper we discuss new adaptive proposal strategies for sequential Monte Carlo algorithms—also known as particle filters—relying on criteria evaluating the quality of the proposed particles. The choice of the proposal distribution is a major concern and can dramatically influence the quality of the estimates. Thus, we show how the long-used coefficient of variation (suggested by Kong et al. in *J. Am. Stat. Assoc.* 89(278–288):590–599, 1994) of the weights can be used for estimating the chi-square distance between the target and instrumental distributions of the auxiliary particle filter. As a by-product of this analysis we obtain an auxiliary adjustment multiplier weight type for which this chi-square distance is minimal. Moreover, we establish an empirical estimate of linear complexity of the Kullback-Leibler divergence between the involved distributions. Guided by these results, we discuss adaptive designing of the particle filter proposal distribution and illustrate the methods on a numerical example.

Keywords Adaptive Monte Carlo · Auxiliary particle filter · Coefficient of variation · Kullback-Leibler

This work was partly supported by the National Research Agency (ANR) under the program “ANR-05-BLAN-0299”.

J. Cornebise (✉) · É. Moulines
Institut des Télécoms, Télécom ParisTech, 46 Rue Barrault,
75634 Paris Cedex 13, France
e-mail: julien.cornebise@telecom-paristech.fr

É. Moulines
e-mail: eric.moulines@telecom-paristech.fr

J. Olsson
Center of Mathematical Sciences, Lund University, Box 118,
SE-22100 Lund, Sweden
e-mail: jimmy@maths.lth.se

divergence · Cross-entropy method · Sequential Monte Carlo · State space models

1 Introduction

Easing the role of the user by tuning automatically the key parameters of *sequential Monte Carlo (SMC) algorithms* has been a long-standing topic in the community, notably through adaptation of the particle sample size or the way the particles are sampled and weighted. In this paper we focus on the latter issue and develop methods for adjusting adaptively the proposal distribution of the particle filter.

Adaptation of the number of particles has been treated by several authors. In Legland and Oudjane (2006) (and later Hu et al. 2008, Sect. IV) the size of the particle sample is increased until the total weight mass reaches a positive threshold, avoiding a situation where all particles are located in regions of the state space having zero posterior probability. Fearnhead and Liu (2007, Sect. 3.2) adjust the size of the particle cloud in order to control the error introduced by the resampling step. Another approach, suggested by Fox (2003) and refined in Soto (2005) and Straka and Simandl (2006), consists of increasing the sample size until the *Kullback-Leibler divergence (KLD)* between the true and estimated target distributions is below a given threshold.

Unarguably, setting an appropriate sample size is a key ingredient of any statistical estimation procedure, and there are cases where the methods mentioned above may be used for designing satisfactorily this size; however increasing the sample size only is far from being always sufficient for achieving efficient variance reduction. Indeed, as in any algorithm based on importance sampling, a significant discrepancy between the proposal and target distributions may

require an unreasonably large number of samples for decreasing the variance of the estimate under a specified value. For a very simple illustration, consider importance sampling estimation of the mean m of a normal distribution using as importance distribution another normal distribution having zero mean and same variance: in this case, the variance of the estimate grows like $\exp(m^2)/N$, N denoting the number of draws, implying that the sample size required for ensuring a given variance grows exponentially fast with m .

This points to the need for adapting the importance distribution of the particle filter, e.g., by adjusting at each iteration the particle weights and the proposal distributions; see e.g. Doucet et al. (2000), Liu (2004), and Fearnhead (2008) for reviews of various filtering methods. These two quantities are critically important, since the performance of the particle filter is closely related to the ability of proposing particles in state space regions where the posterior is significant. It is well known that sampling using as proposal distribution the mixture composed by the current particle importance weights and the prior kernel (yielding the classical bootstrap particle filter of Gordon et al. 1993) is usually inefficient when the likelihood is highly peaked or located in the tail of the prior.

In the sequential context, the successive distributions to be approximated (e.g. the successive filtering distributions) are the iterates of a nonlinear random mapping, defined on the space of probability measures; this nonlinear mapping may in general be decomposed into two steps: a prediction step which is linear and a nonlinear correction step which amounts to compute a normalisation factor. In this setting, an appealing way to update the current particle approximation consists of sampling new particles from the distribution obtained by propagating the current particle approximation through this mapping; see e.g. Hürzeler and Künsch (1998), Doucet et al. (2000), and Künsch (2005) (and the references therein). This sampling distribution guarantees that the conditional variance of the importance weights is equal to zero. As we shall see below, this proposal distribution enjoys other optimality conditions, and is in the sequel referred to as the *optimal sampling distribution*. However, sampling from the optimal sampling distribution is, except for some specific models, a difficult and time-consuming task (the in general costly auxiliary accept-reject developed and analysed by Künsch (2005) being most often the only available option).

To circumvent this difficulty, several sub-optimal schemes have been proposed. A first type of approaches tries to mimic the behavior of the optimal sampling without suffering the sometimes prohibitive cost of rejection sampling. This typically involves localisation of the modes of the unnormalised optimal sampling distribution by means of some optimisation algorithm, and the fitting of over-dispersed student's t -distributions around these modes; see for example

Shephard and Pitt (1997), Doucet et al. (2001), and Liu (2004) (and the references therein). Except in specific cases, locating the modes involves solving an optimisation problem for every particle, which is quite time-consuming.

A second class of approaches consists of using some classical approximate non-linear filtering tools such as the *extended Kalman filter* (EKF) or the *unscented transform Kalman filter* (UT/UKF); see for example Doucet et al. (2001) and the references therein. These techniques assume implicitly that the conditional distribution of the next state given the current state and the observation has a single mode. In the EKF version of the particle filter, the linearisation of the state and observation equations is carried out for each individual particle. Instead of linearising the state and observation dynamics using Jacobian matrices, the UT/UKF particle filter uses a deterministic sampling strategy to capture the mean and covariance with a small set of carefully selected points (*sigma points*), which is also computed for each particle. Since these computations are most often rather involved, a significant computational overhead is introduced.

A third class of techniques is the so-called *auxiliary particle filter* (APF) suggested by Pitt and Shephard (1999), who proposed it as a way to build data-driven proposal distributions (with the initial aim of robustifying standard SMC methods to the presence of outlying observations); see e.g. Fearnhead (2008). The procedure comprises two stages: in the first-stage, the current particle weights are modified in order to select preferentially those particles being most likely proposed in regions where the posterior is significant. Usually this amounts to multiply the weights with so-called *adjustment multiplier weights*, which may depend on the next observation as well as the current position of the particle and (possibly) the proposal transition kernels. Most often, this adjustment weight is chosen to estimate the predictive likelihood of the next observation given the current particle position, but this choice is not necessarily optimal.

In a second stage, a new particle sample from the target distribution is formed using this proposal distribution and associating the proposed particles with weights proportional to the inverse of the adjustment multiplier weight.¹ APF procedures are known to be rather successful when the first-stage distribution is appropriately chosen, which is not always straightforward. The additional computational cost depends mainly on the way the first-stage proposal is designed. The APF method can be mixed with EKF and UKF leading

¹The original APF proposed by Pitt and Shephard (1999) features a second resampling procedure in order to end-up with an equally weighted particle system. This resampling procedure might however severely reduce the accuracy of the filter: Carpenter et al. (1999) give an example where the accuracy is reduced by a factor of 2; see also Douc et al. (2008) for a theoretical proof.

to powerful but computationally involved particle filter; see, e.g., Andrieu et al. (2003).

None of the suboptimal methods mentioned above minimise any sensible risk-theoretic criterion and, more annoyingly, both theoretical and practical evidences show that choices which seem to be intuitively correct may lead to performances even worse than that of the plain bootstrap filter (see for example Douc et al. 2008 for a striking example). The situation is even more unsatisfactory when the particle filter is driven by a state space dynamic different from that generating the observations, as happens frequently when, e.g., the parameters are not known and need to be estimated or when the model is misspecified.

Instead of trying to guess what a good proposal distribution should be, it seems sensible to follow a more risk-theoretically founded approach. The first step in such a construction consists of choosing a sensible risk criterion, which is not a straightforward task in the SMC context. A natural criterion for SMC would be the variance of the estimate of the posterior mean of a target function (or a set of target functions) of interest, but this approach does not lead to a practical implementation for two reasons. Firstly, in SMC methods, though closed-form expression for the variance at any given current time-step of the posterior mean of any function is available, this variance depends explicitly on all the time steps before the current time. Hence, choosing to minimise the variance at a given time-step would require to optimise all the simulations up to that particular time step, which is of course not practical. Because of the recursive form of the variance, the minimisation of the conditional variance at each iteration of the algorithm does not necessarily lead to satisfactory performance on the long-run. Secondly, as for the standard importance sampling algorithm, this criterion is not *function-free*, meaning that a choice of a proposal can be appropriate for a given function, but inappropriate for another.

We will focus in the sequel on function-free risk criteria. A first criterion, advocated in Kong et al. (1994) and Liu (2004) is the *chi-square distance* (CSD) between the proposal and the target distributions, which coincides with the *coefficient of variation* (CV^2) of the importance weights. In addition, as heuristically discussed in Kong et al. (1994), the CSD is related to the *effective sample size*, which estimates the number of i.i.d. samples equivalent to the weighted particle system.² In practice, the CSD criterion can be estimated, with a complexity that grows linearly with the number of particles, using the empirical CV^2 which can be shown to converge to the CSD as the number of particles tends to infinity. In this paper we show that a similar property

still holds in the SMC context, in the sense that the CV^2 still measures a CSD between two distributions μ^* and π^* , which are associated with the proposal and target distributions of the particle filter (see Theorem 1(ii)). Though this result does not come as a surprise, it provides an additional theoretical footing to an approach which is currently used in practice for triggering resampling steps.

Another function-free risk criterion to assess the performance of importance sampling estimators is the KLD between the proposal and the target distributions; see Cappé et al. (2005, Chap. 7). The KLD shares some of the attractive properties of the CSD; in particular, the KLD may be estimated using the negated empirical *entropy* \mathcal{E} of the importance weights, whose computational complexity is again linear in the number of particles. In the SMC context, it is shown in Theorem 1(i) that \mathcal{E} still converges to the KLD between the same two distributions μ^* and π^* associated with the proposal and the target distributions of the particle filter.

Our methodology to design appropriate proposal distributions is based upon the minimisation of the CSD and KLD between the proposal and the target distributions. Whereas these quantities (and especially the CSD) have been routinely used to detect sample impoverishment and trigger the resampling step (Kong et al. 1994), they have not been used for adapting the simulation parameters in SMC methods.

We focus here on the auxiliary sampling formulation of the particle filter. In this setting, there are two quantities to optimise: the adjustment multiplier weights (also called *first-stage weights*) and the parameters of the proposal kernel; together these quantities define the mixture used as instrumental distribution in the filter. We first establish a closed-form expression for the limiting value of the CSD and KLD of the auxiliary formulation of the proposal and the target distributions. Using these expressions, we identify a type of auxiliary SMC adjustment multiplier weights which minimise the CSD and the KLD for a given proposal kernel (Proposition 2). We then propose several optimisation techniques for adapting the proposal kernels, always driven by the objective of minimising the CSD or the KLD, in coherence with what is done to detect sample impoverishment (see Sect. 5). Finally, in the implementation section (Sect. 6), we use the proposed algorithms for approximating the filtering distributions in several state space models, and show that the proposed optimisation procedure improves the accuracy of the particle estimates and makes them more robust to outlying observations.

2 Informal presentation of the results

2.1 Adaptive importance sampling

Before stating and proving rigorously the main results, we discuss informally our findings and introduce the proposed

²In some situations, the estimated ESS value can be misleading: see the comments of Stephens and Donnelly (2000) for a further discussion of this.

methodology for developing adaptive SMC algorithms. Before entering into the sophistication of sequential methods, we first briefly introduce adaptation of the standard (non-sequential) importance sampling algorithm.

Importance sampling (IS) is a general technique to compute expectations of functions with respect to a target distribution with density $p(x)$ while only having samples generated from a different distribution—referred to as the *proposal distribution*—with density $q(x)$ (implicitly, the dominating measure is taken to be the Lebesgue measure on $\mathcal{E} \triangleq \mathbb{R}^d$). We sample $\{\xi_i\}_{i=1}^N$ from the proposal distribution q and compute the unnormalised importance weights $\omega_i \triangleq W(\xi_i)$, $i = 1, \dots, N$, where $W(x) \triangleq p(x)/q(x)$. For any function f , the self-normalised importance sampling estimator may be expressed as $IS_N(f) \triangleq \Omega_N^{-1} \sum_{i=1}^N \omega_i f(\xi_i)$, where $\Omega_N \triangleq \sum_{i=1}^N \omega_i$. As usual in applications of the IS methodology to Bayesian inference, the target density p is known only up to a normalisation constant; hence we will focus only on a self-normalised version of IS that solely requires the availability of an unnormalised version of p (see Geweke 1989). Throughout the paper, we call a set $\{\xi_i\}_{i=1}^N$ of random variables, referred to as *particles* and taking values in \mathcal{E} , and nonnegative weights $\{\omega_i\}_{i=1}^N$ a *weighted sample* on \mathcal{E} . Here N is a (possibly random) integer, though we will take it fixed in the sequel. It is well known (see again Geweke 1989) that, provided that f is integrable with respect to p , i.e. $\int |f(x)|p(x) dx < \infty$, $IS_N(f)$ converges, as the number of samples tends to infinity, to the target value

$$\mathbb{E}_p[f(X)] \triangleq \int f(x)p(x) dx,$$

for any function $f \in \mathcal{C}$, where \mathcal{C} is the set of functions which are integrable with respect to the target distribution p . Under some additional technical conditions, this estimator is also asymptotically normal at rate \sqrt{N} ; see Geweke (1989).

It is well known that IS estimators are sensitive to the choice of the proposal distribution. A classical approach consists of trying to minimise the asymptotic variance with respect to the proposal distribution q . This optimisation is in closed form and leads (when f is a non-negative function) to the optimal choice $q^*(x) = f(x)p(x) / \int f(x)p(x) dx$, which is, since the normalisation constant is precisely the quantity of interest, rather impractical. Sampling from this distribution can be done by using an accept-reject algorithm, but this does not solve the problem of choosing an appropriate proposal distribution. Note that it is possible to approach this optimal sampling distribution by using the *cross-entropy method*; see Rubinstein and Kroese (2004) and de Boer et al. (2005) and the references therein. We will discuss this point later on.

For reasons that will become clear in the sequel, this type of objective is impractical in the sequential context, since

the expression of the asymptotic variance in this case is recursive and the optimisation of the variance at a given step is impossible. In addition, in most applications, the proposal density is expected to perform well for a range of typical functions of interest rather than for a specific target function f . We are thus looking for *function-free* criteria. The most often used criterion is the CSD between the proposal distribution q and the target distribution p , defined as

$$d_{\chi^2}(p \parallel q) = \int \frac{\{p(x) - q(x)\}^2}{q(x)} dx, \tag{2.1}$$

$$= \int W^2(x)q(x) dx - 1, \tag{2.2}$$

$$= \int W(x)p(x) dx - 1. \tag{2.3}$$

The CSD between p and q may be expressed as the variance of the importance weight function W under the proposal distribution, i.e.

$$d_{\chi^2}(p \parallel q) = \text{Var}_q[W(X)].$$

This quantity can be estimated by computing the squared coefficient of variation of the unnormalized weights (Evans and Swartz 1995, Sect. 4):

$$CV^2(\{\omega_i\}_{i=1}^N) \triangleq N\Omega_N^{-2} \sum_{i=1}^N \omega_i^2 - 1. \tag{2.4}$$

The CV^2 was suggested by Kong et al. (1994) as a means for detecting weight degeneracy. If all the weights are equal, then CV^2 is equal to zero. On the other hand, if all the weights but one are zero, then the coefficient of variation is equal to $N - 1$ which is its maximum value. From this it follows that using the estimated coefficient of variation for assessing accuracy is equivalent to examining the normalised importance weights to determine if any are relatively large.³ Kong et al. (1994) showed that the coefficient of variation of the weights $CV^2(\{\omega_i\}_{i=1}^N)$ is related to the *effective sample size* (ESS), which is used for measuring the overall efficiency of an IS algorithm:

$$N^{-1} \text{ESS}(\{\omega_i\}_{i=1}^N) \triangleq \frac{1}{1 + CV^2(\{\omega_i\}_{i=1}^N)} \rightarrow \{1 + d_{\chi^2}(p \parallel q)\}^{-1}.$$

Heuristically, the ESS measures the number of i.i.d. samples (from p) equivalent to the N weighted samples. The smaller the CSD between the proposal and target distributions is,

³Some care should be taken for small sample sizes N ; the CV^2 can be low because q sample only over a subregion where the integrand is nearly constant, which is not always easy to detect.

the larger is the ESS. This is why the CSD is of particular interest when measuring efficiency of IS algorithms.

Another possible measure of fit of the proposal distribution is the KLD (also called *relative entropy*) between the proposal and target distributions, defined as

$$d_{\text{KL}}(p \parallel q) \triangleq \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx, \quad (2.5)$$

$$= \int p(x) \log W(x) dx, \quad (2.6)$$

$$= \int W(x) \log W(x) q(x) dx. \quad (2.7)$$

This criterion can be estimated from the importance weights using the negative *Shannon entropy* \mathcal{E} of the importance weights:

$$\mathcal{E}(\{\omega_i\}_{i=1}^N) \triangleq \Omega_N^{-1} \sum_{i=1}^N \omega_i \log \left(N \Omega_N^{-1} \omega_i \right). \quad (2.8)$$

The Shannon entropy is maximal when all the weights are equal and minimal when all weights are zero but one. In IS (and especially for the estimation of rare events), the KLD between the proposal and target distributions was thoroughly investigated by Rubinstein and Kroese (2004), and is central in the *cross-entropy* (CE) methodology.

Classically, the proposal is chosen from a family of densities q_θ parameterised by θ . Here θ should be thought of as an element of Θ , which is a subset of \mathbb{R}^k . The most classical example is the family of student's t -distributions parameterised by mean and covariance. More sophisticated parameterisations, like mixture of multi-dimensional Gaussian or Student's t -distributions, have been proposed; see, e.g., Oh and Berger (1992, 1993), Evans and Swartz (1995), Givens and Raftery (1996), Liu (2004, Chap. 2, Sect. 2.6), and, more recently, Cappé et al. (2008) in this issue. In the sequential context, where computational efficiency is a must, we typically use rather simple parameterisations, so that the two criteria above can be (approximatively) solved in a few iterations of a numerical minimisation procedure.

The optimal parameters for the CSD and the KLD are those minimising $\theta \mapsto d_{\chi^2}(p \parallel q_\theta)$ and $\theta \mapsto d_{\text{KL}}(p \parallel q_\theta)$, respectively. In the sequel, we denote by θ_{CSD}^* and θ_{KLD}^* these optimal values. Of course, these quantities cannot be computed in closed form (recall that even the normalisation constant of p is most often unknown; even if it is known, the evaluation of these quantities would involve the evaluation of most often high-dimensional integrals). Nevertheless, it is possible to construct consistent estimators of these optimal parameters. There are two classes of methods, detailed below.

The first uses the fact that the the CSD $d_{\chi^2}(p \parallel q_\theta)$ and the KLD $d_{\text{KL}}(p \parallel q_\theta)$ may be approximated by (2.4) and (2.8),

substituting in these expressions the importance weights by $\omega_i = W_\theta(\xi_i^\theta)$, $i = 1, \dots, N$, where $W_\theta \triangleq p/q_\theta$ and $\{\xi_i^\theta\}_{i=1}^N$ is a sample from q_θ . This optimisation problem formally shares some similarities with the classical minimum chi-square or maximum likelihood estimation, but with the following important difference: the integrations in (2.1) and (2.5) are with respect to the proposal distribution q_θ and not the target distribution p . As a consequence, the particles $\{\xi_i^\theta\}_{i=1}^N$ in the definition of the coefficient of variation (2.4) or the entropy (2.8) of the weights constitute a sample from q_θ and not from the target distribution p . As the estimation progresses, the samples used to approach the limiting CSD or KLD can, in contrast to standard estimation procedures, be updated (these samples could be kept fixed, but this is of course inefficient).

The computational complexity of these optimisation problems depends on the way the proposal is parameterised and how the optimisation procedure is implemented. Though the details of the optimisation procedure is in general strongly model dependent, some common principles for solving this optimisation problem can be outlined. Typically, the optimisation is done recursively, i.e. the algorithm defines a sequence θ_ℓ , $\ell = 0, 1, \dots$, of parameters, where ℓ is the iteration number. At each iteration, the value of θ_ℓ is updated by computing a direction $p_{\ell+1}$ in which to step, a step length $\gamma_{\ell+1}$, and setting

$$\theta_{\ell+1} = \theta_\ell + \gamma_{\ell+1} p_{\ell+1}.$$

The search direction is typically computed using either Monte Carlo approximation of the finite-difference or (when the quantities of interest are sufficiently regular) the gradient of the criterion. These quantities are used later in conjunction with classical optimisation strategies for computing the step size $\gamma_{\ell+1}$ or normalising the search direction. These implementation issues, detailed in Sect. 6, are model dependent. We denote by M_ℓ the number of particles used to obtain such an approximation at iteration ℓ . The number of particles may vary with the iteration index; heuristically there is no need for using a large number of simulations during the initial stage of the optimisation. Even rather crude estimation of the search direction might suffice to drive the parameters towards the region of interest. However, as the iterations go on, the number of simulations should be increased to avoid “zig-zagging” when the algorithm approaches convergence. After L iterations, the total number of generated particles is equal to $N = \sum_{\ell=1}^L M_\ell$. Another solution, which is not considered in this paper, would be to use a stochastic approximation procedure, which consists of fixing $M_\ell = M$ and letting the stepsize γ_ℓ tend to zero. This appealing solution has been successfully used in Arouna (2004).

The computation of the finite difference or the gradient, being defined as expectations of functions depending on θ , can be performed using two different approaches. Starting

from definitions (2.3) and (2.6), and assuming appropriate regularity conditions, the gradient of $\theta \mapsto d_{\chi^2}(p \parallel q_\theta)$ and $\theta \mapsto d_{\text{KL}}(p \parallel q_\theta)$ may be expressed as

$$G_{\text{CSD}}(\theta) \triangleq \nabla_\theta d_{\chi^2}(p \parallel q_\theta) = \int p(x) \nabla_\theta W_\theta(x) dx, \\ = \int q_\theta(x) W_\theta(x) \nabla_\theta W_\theta(x) dx, \tag{2.9}$$

$$G_{\text{KLD}}(\theta) \triangleq \nabla_\theta d_{\text{KL}}(p \parallel q_\theta) = \int p(x) \nabla_\theta \log[W_\theta(x)] dx, \\ = \int q_\theta(x) \nabla_\theta W_\theta(x) dx. \tag{2.10}$$

These expressions lead immediately to the following approximations,

$$\hat{G}_{\text{CSD}}(\theta) = M^{-1} \sum_{i=1}^M W_\theta(\xi_i^\theta) \nabla_\theta W_\theta(\xi_i^\theta), \tag{2.11}$$

$$\hat{G}_{\text{KLD}}(\theta) = M^{-1} \sum_{i=1}^M \nabla_\theta W_{\theta_\ell}(\xi_i^{\theta_\ell}). \tag{2.12}$$

There is another way to compute derivatives, which shares some similarities with *pathwise derivative estimates*. Recall that for any $\theta \in \Theta$, one may choose F_θ so that the random variable $F_\theta(\epsilon)$, where ϵ is a vector of independent uniform random variables on $[0, 1]^d$, is distributed according to q_θ . Therefore, we may express $\theta \mapsto d_{\chi^2}(p \parallel q_\theta)$ and $\theta \mapsto d_{\text{KL}}(p \parallel q_\theta)$ as the following integrals,

$$d_{\chi^2}(p \parallel q_\theta) = \int_{[0,1]^d} w_\theta(x) dx, \\ d_{\text{KL}}(p \parallel q_\theta) = \int_{[0,1]^d} w_\theta(x) \log[w_\theta(x)] dx,$$

where $w_\theta(x) \triangleq W_\theta \circ F_\theta(x)$. Assuming appropriate regularity conditions (i.e. that $\theta \mapsto W_\theta \circ F_\theta(x)$ is differentiable and that we can interchange the integration and the differentiation), the differential of these quantities with respect to θ may be expressed as

$$G_{\text{CSD}}(\theta) = \int_{[0,1]^d} \nabla_\theta w_\theta(x) dx, \\ G_{\text{KLD}}(\theta) = \int_{[0,1]^d} \{\nabla_\theta w_\theta(x) \log[w_\theta(x)] + \nabla_\theta w_\theta(x)\} dx.$$

For any given x , the quantity $\nabla_\theta w_\theta(x)$ is the pathwise derivative of the function $\theta \mapsto w_\theta(x)$. As a practical matter, we usually think of each x as a realization of the output of an ideal random generator. Each $w_\theta(x)$ is then the output of the simulation algorithm at parameter θ for the random number x . Each $\nabla_\theta w_\theta(x)$ is the derivative of the simulation output with respect to θ with the random numbers held fixed.

These two expressions, which of course coincide with (2.9) and (2.10), lead to the following estimators,

$$\tilde{G}_{\text{CSD}}(\theta) = M^{-1} \sum_{i=1}^M \nabla_\theta w_\theta(\epsilon_i), \\ \tilde{G}_{\text{KLD}}(\theta) = M^{-1} \sum_{i=1}^M \{\nabla_\theta w_\theta(\epsilon_i) \log[w_\theta(\epsilon_i)] + \nabla_\theta w_\theta(\epsilon_i)\},$$

where each element of the sequence $\{\epsilon_i\}_{i=1}^M$ is a vector on $[0, 1]^d$ of independent uniform random variables. It is worthwhile to note that if the number $M_\ell = M$ is kept fixed during the iterations and the uniforms $\{\epsilon_i\}_{i=1}^M$ are drawn once and for all (i.e. the same uniforms are used at the different iterations), then the iterative algorithm outlined above solves the following problem:

$$\theta \mapsto \text{CV}^2 \left(\{w_\theta(\epsilon_i)\}_{i=1}^M \right), \tag{2.13}$$

$$\theta \mapsto \mathcal{E} \left(\{w_\theta(\epsilon_i)\}_{i=1}^N \right). \tag{2.14}$$

From a theoretical standpoint, this optimisation problem is very similar to M -estimation, and convergence results for M -estimators can thus be used under rather standard technical assumptions; see for example Van der Vaart (1998). This is the main advantage of fixing the sample $\{\epsilon_i\}_{i=1}^M$. We use this implementation in the simulations.

Under appropriate conditions, the sequence of estimators $\theta_{\ell, \text{CSD}}^*$ or $\theta_{\ell, \text{KLD}}^*$ of these criteria converge, as the number of iterations tends to infinity, to θ_{CSD}^* or θ_{KLD}^* which minimise the criteria $\theta \mapsto d_{\chi^2}(p \parallel q_\theta)$ and $\theta \mapsto d_{\text{KL}}(p \parallel q_\theta)$, respectively; these theoretical issues are considered in a companion paper.

The second class of approaches considered in this paper is used for minimising the KLD (2.14) and is inspired by the cross-entropy method. This algorithm approximates the minimum θ_{KLD}^* of (2.14) by a sequence of pairs of steps, where each step of each pair addresses a simpler optimisation problem. Compared to the previous method, this algorithm is derivative-free and does not require to select a step size. It is in general simpler to implement and avoid most of the common pitfalls of stochastic approximation. Denote by $\theta_0 \in \Theta$ an initial value. We define recursively the sequence $\{\theta_\ell\}_{\ell \geq 0}$ as follows. In a first step, we draw a sample $\{\xi_i^{\theta_\ell}\}_{i=1}^{M_\ell}$ and evaluate the function

$$\theta \mapsto Q_\ell(\theta, \theta_\ell) \triangleq \sum_{i=1}^{M_\ell} W_{\theta_\ell}(\xi_i^{\theta_\ell}) \log q_\theta(\xi_i^{\theta_\ell}). \tag{2.15}$$

In a second step, we choose $\theta_{\ell+1}$ to be the (or any, if there are several) value of $\theta \in \Theta$ that maximises $Q_\ell(\theta, \theta_\ell)$. As above, the number of particles M_ℓ is increased during the successive iterations. This procedure resembles closely the

Monte Carlo EM (Wei and Tanner 1991) for maximum likelihood in incomplete data models. The advantage of this approach is that the solution of the maximisation problem $\theta_{\ell+1} = \operatorname{argmax}_{\theta \in \Theta} \in \mathcal{Q}_\ell(\theta, \theta_\ell)$ is often on closed form. In particular, this happens if the distribution q_θ belongs to an exponential family (EF) or is a mixture of distributions of NEF; see Cappé et al. (2008) for a discussion. The convergence of this algorithm can be established along the same lines as the convergence of the MCEM algorithm; see Fort and Moulines (2003). As the number of iterations ℓ increases, the sequence of estimators θ_ℓ may be shown to converge to θ_{KLD}^* . These theoretical results are established in a companion paper.

2.2 Sequential Monte Carlo methods

In the sequential context, where the problem consists of simulating from a sequence $\{p_k\}$ of probability density function, the situation is more difficult. Let \mathcal{X}_k be denote the state space of distribution p_k and note that this space may vary with k , e.g. in terms of increasing dimensionality. In many applications, these densities are related to each other by a (possibly random) mapping, i.e. $p_k = \Psi_{k-1}(p_{k-1})$. In the sequel we focus on the case where there exists a non-negative function $l_{k-1} : (\xi, \tilde{\xi}) \mapsto l_{k-1}(\xi, \tilde{\xi})$ such that

$$p_k(\tilde{\xi}) = \frac{\int l_{k-1}(\xi, \tilde{\xi}) p_{k-1}(\xi) d\xi}{\int p_{k-1}(\xi) \int l_{k-1}(\xi, \tilde{\xi}) d\tilde{\xi} d\xi} \tag{2.16}$$

As an example, consider the following generic nonlinear dynamic system described in state space form:

– *State (system) model*

$$X_k = a(X_{k-1}, U_k) \leftrightarrow \overbrace{q(X_{k-1}, X_k)}^{\text{Transition Density}}, \tag{2.17}$$

– *Observation (measurement) model*

$$Y_k = b(X_k, V_k) \leftrightarrow \overbrace{g(X_k, Y_k)}^{\text{Observation Density}}. \tag{2.18}$$

By these equations we mean that each hidden state X_k and data Y_k are assumed to be generated by nonlinear functions $a(\cdot)$ and $b(\cdot)$, respectively, of the state and observation noises U_k and V_k . The state and the observation noises $\{U_k\}_{k \geq 0}$ and $\{V_k\}_{k \geq 0}$ are assumed to be mutually independent sequences of i.i.d. random variables. The precise form of the functions and the assumed probability distributions of the state and observation noises U_k and V_k imply, via a change of variables, the transition probability density function $q(x_{k-1}, x_k)$ and the observation probability density function $g(x_k, y_k)$, the latter being referred to as the *likelihood of the observation*. With these definitions, the process

$\{X_k\}_{k \geq 0}$ is Markovian, i.e. the conditional probability density of X_k given the past states $X_{0:k-1} \triangleq (X_0, \dots, X_{k-1})$ depends exclusively on X_{k-1} . This distribution is described by the density $q(x_{k-1}, x_k)$. In addition, the conditional probability density of Y_k given the states $X_{0:k}$ and the past observations $Y_{0:k-1}$ depends exclusively on X_k , and this distribution is captured by the likelihood $g(x_k, y_k)$. We assume further that the initial state X_0 is distributed according to a density function $\pi_0(x_0)$. Such nonlinear dynamic systems arise frequently in many areas of science and engineering such as target tracking, computer vision, terrain referenced navigation, finance, pollution monitoring, communications, audio engineering, to list only a few.

Statistical inference for the general nonlinear dynamic system above involves computing the *posterior distribution* of a collection of state variables $X_{s:s'} \triangleq (X_s, \dots, X_{s'})$ conditioned on a batch $Y_{0:k} = (Y_0, \dots, Y_k)$ of observations. We denote this posterior distribution by $\phi_{s:s'|k}(X_{s:s'}|Y_{0:k})$. Specific problems include *filtering*, corresponding to $s = s' = k$, *fixed lag smoothing*, where $s = s' = k - L$, and *fixed interval smoothing*, with $s = 0$ and $s' = k$. Despite the apparent simplicity of the above problem, the posterior distributions can be computed in closed form only in very specific cases, principally, the linear Gaussian model (where the functions $a(\cdot)$ and $b(\cdot)$ are linear and the state and observation noises $\{U_k\}_{k \geq 0}$ and $\{V_k\}_{k \geq 0}$ are Gaussian) and the discrete *hidden Markov model* (where X_k takes its values in a finite alphabet). In the vast majority of cases, nonlinearity or non-Gaussianity render analytic solutions intractable—see Anderson and Moore (1979), Kailath et al. (2000), Ristic et al. (2004), Cappé et al. (2005).

Starting with the initial, or prior, density function $\pi_0(x_0)$, and observations $Y_{0:k} = y_{0:k}$, the posterior density $\phi_{k|k}(x_k|y_{0:k})$ can be obtained using the following *prediction-correction* recursion (Ho and Lee 1964):

– *Prediction*

$$\begin{aligned} \phi_{k|k-1}(x_k|y_{0:k-1}) \\ = \phi_{k-1|k-1}(x_{k-1}|y_{0:k-1})q(x_{k-1}, x_k), \end{aligned} \tag{2.19}$$

– *Correction*

$$\phi_{k|k}(x_k|y_{0:k}) = \frac{g(x_k, y_k)\phi_{k|k-1}(x_k|y_{0:k-1})}{\ell_{k|k-1}(y_k|y_{0:k-1})}, \tag{2.20}$$

where $\ell_{k|k-1}$ is the predictive distribution of Y_k given the past observations $Y_{0:k-1}$. For a fixed data realisation, this term is a normalising constant (independent of the state) and is thus not necessary to compute in standard implementations of SMC methods.

By setting $p_k = \phi_{k|k}$, $p_{k-1} = \phi_{k-1|k-1}$, and

$$l_{k-1}(x, x') = g(x_k, y_k)q(x_{k-1}, x_k),$$

we conclude that the sequence $\{\phi_{k|k}\}_{k \geq 1}$ of filtering densities can be generated according to (2.16).

The case of fixed interval smoothing works entirely analogously: indeed, since

$$\begin{aligned} \phi_{0:k|k-1}(x_{0:k}|y_{0:k-1}) \\ = \phi_{0:k-1|k-1}(x_{0:k-1}|y_{0:k-1})q(x_{k-1}, x_k) \end{aligned}$$

and

$$\phi_{0:k|k}(x_k|y_{0:k}) = \frac{g(x_k, y_k)\phi_{k|k-1}(x_{0:k}|y_{0:k-1})}{\ell_{k|k-1}(y_k|y_{0:k-1})},$$

the flow $\{\phi_{0:k|k}\}_{k \geq 1}$ of smoothing distributions can be generated according to (2.16) by letting $p_k = \phi_{0:k|k}$, $p_{k-1} = \phi_{0:k-1|k-1}$, and replacing $l_{k-1}(x_{0:k-1}, x'_{0:k}) dx'_{0:k}$ by $g(x'_k, y_k)q(x_{k-1}, x'_k) dx'_k \delta_{x_{0:k-1}}(dx'_{0:k-1})$, where δ_a denotes the Dirac mass located in a . Note that this replacement is done formally since the unnormalised transition kernel in question lacks a density in the smoothing mode; this is due to the fact that the Dirac measure is singular with respect to the Lebesgue measure. This is however handled by the measure theoretic approach in Sect. 4, implying that all theoretical results presented in the following will comprise also fixed interval smoothing.

We now adapt the procedures considered in the previous section to the sampling of densities generated according to (2.16). Here we focus on a single time-step, and drop from the notation the dependence on k which is irrelevant at this stage. Moreover, set $p_k = \mu$, $p_{k-1} = \nu$, $l_k = l$, and assume that we have at hand a weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^N$ targeting ν , i.e., for any ν -integrable function f , $\Omega_N^{-1} \sum_{i=1}^N \omega_{N,i} f(\xi_{N,i})$ approximates the corresponding integral $\int f(\xi)\nu(\xi) d\xi$. A natural strategy for sampling from μ is to replace ν in (2.16) by its particle approximation, yielding

$$\mu_N(\tilde{\xi}) \triangleq \sum_{i=1}^N \frac{\omega_{N,i} \int l(\xi_{N,i}, \tilde{\xi}) d\tilde{\xi}}{\sum_{j=1}^N \omega_{N,j} \int l(\xi_{N,j}, \tilde{\xi}) d\tilde{\xi}} \left[\frac{l(\xi_{N,i}, \tilde{\xi})}{\int l(\xi_{N,i}, \tilde{\xi}) d\tilde{\xi}} \right]$$

as an approximation of μ , and simulate \tilde{M}_N new particles from this distribution; however, in many applications direct simulation from μ_N is infeasible without the application of computationally expensive auxiliary accept-reject techniques introduced by Hürzeler and Künsch (1998) and thoroughly analysed by Künsch (2005). This difficulty can be overcome by simulating new particles $\{\tilde{\xi}_{N,i}\}_{i=1}^{\tilde{M}_N}$ from the instrumental mixture distribution with density

$$\pi_N(\tilde{\xi}) \triangleq \sum_{i=1}^N \frac{\omega_{N,i} \psi_{N,i}}{\sum_{j=1}^N \omega_{N,j} \psi_{N,j}} r(\xi_{N,i}, \tilde{\xi}),$$

where $\{\psi_{N,i}\}_{i=1}^N$ are the so-called *adjustment multiplier weights* and r is a Markovian transition density function,

i.e., $r(\xi, \tilde{\xi})$ is a nonnegative function and, for any $\xi \in \Xi$, $\int r(\xi, \tilde{\xi}) d\tilde{\xi} = 1$. If one can guess, based on the new observation, which particles are most likely to contribute significantly to the posterior, the resampling stage may be anticipated by increasing (or decreasing) the importance weights. This is the purpose of using the multiplier weights $\psi_{N,i}$. We associate these particles with importance weights $\{\mu_N(\tilde{\xi}_{N,i})/\pi_N(\tilde{\xi}_{N,i})\}_{i=1}^{\tilde{M}_N}$. In this setting, a new particle position is simulated from the transition proposal density $r(\xi_{N,i}, \cdot)$ with probability proportional to $\omega_{N,i} \psi_{N,i}$. Happlessly, the importance weight $\mu_N(\tilde{\xi}_{N,i})/\pi_N(\tilde{\xi}_{N,i})$ is expensive to evaluate since this involves summing over N terms.

We thus introduce, as suggested by Pitt and Shephard (1999), an *auxiliary variable* corresponding to the selected particle, and target instead the probability density

$$\begin{aligned} \mu_N^{\text{aux}}(i, \tilde{\xi}) \\ \triangleq \frac{\omega_{N,i} \int l(\xi_{N,i}, \tilde{\xi}) d\tilde{\xi}}{\sum_{j=1}^N \omega_{N,j} \int l(\xi_{N,j}, \tilde{\xi}) d\tilde{\xi}} \left[\frac{l(\xi_{N,i}, \tilde{\xi})}{\int l(\xi_{N,i}, \tilde{\xi}) d\tilde{\xi}} \right] \end{aligned} \quad (2.21)$$

on the product space $\{1, \dots, N\} \times \Xi$. Since μ_N is the marginal distribution of μ_N^{aux} with respect to the particle index i , we may sample from μ_N by simulating instead a set $\{(I_{N,i}, \tilde{\xi}_{N,i})\}_{i=1}^{\tilde{M}_N}$ of indices and particle positions from the instrumental distribution

$$\pi_N^{\text{aux}}(i, x') \triangleq \frac{\omega_{N,i} \psi_{N,i}}{\sum_{j=1}^N \omega_{N,j} \psi_{N,j}} r(\xi_{N,i}, \tilde{\xi}) \quad (2.22)$$

and assigning each draw $(I_{N,i}, \tilde{\xi}_{N,i})$ the weight

$$\tilde{\omega}_{N,i} \triangleq \frac{\mu_N^{\text{aux}}(I_{N,i}, \tilde{\xi}_{N,i})}{\pi_N^{\text{aux}}(I_{N,i}, \tilde{\xi}_{N,i})} = \psi_{N,I_{N,i}}^{-1} \frac{l(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})}{r(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})}. \quad (2.23)$$

Hereafter, we discard the indices and let $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ approximate the target density μ . Note that setting, for all $i \in \{1, \dots, N\}$, $\psi_{N,i} \equiv 1$ yields the standard bootstrap particle filter presented by Gordon et al. (1993). In the sequel, we assume that each adjustment multiplier weight $\psi_{N,i}$ is a function of the particle position $\psi_{N,i} = \Psi(\xi_{N,i})$, $i \in \{1, \dots, N\}$, and define

$$\Phi(\xi, \tilde{\xi}) \triangleq \Psi^{-1}(\xi) \frac{l(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})}, \quad (2.24)$$

so that $\mu_N^{\text{aux}}(i, \tilde{\xi})/\pi_N^{\text{aux}}(i, \tilde{\xi})$ is proportional to $\Phi(\xi_{N,i}, \tilde{\xi})$. We will refer to the function Ψ as the *adjustment multiplier function*.

2.3 Risk minimisation for sequential adaptive importance sampling and resampling

We may expect that the efficiency of the algorithm described above depends highly on the choice of adjustment multiplier weights and proposal kernel.

In the context of state space models, Pitt and Shephard (1999) suggested to use an approximation, defined as the value of the likelihood evaluated at the mean of the prior transition, i.e. $\psi_{N,i} \triangleq g(\int x'q(\xi_{N,i}, x') dx', y_k)$, where y_k is the current observation, of the predictive likelihood as adjustment multiplier weights. Although this choice of the weight outperforms the conventional bootstrap filter in many applications, as pointed out in Andrieu et al. (2003), this approximation of the predictive likelihood could be very poor and lead to performance even worse than that of the conventional approach if the dynamic model $q(x_{k-1}, x_k)$ is quite scattered and the likelihood $g(x_k, y_k)$ varies significantly over the prior $q(x_{k-1}, x_k)$.

The optimisation of the adjustment multiplier weight was also studied by Douc et al. (2008) (see also Olsson et al. 2007) who identified adjustment multiplier weights for which the increase of asymptotic variance at a single iteration of the algorithm is minimal. Note however that this optimisation is done using a *function-specific* criterion, whereas we advocate here the use of *function-free* criteria.

In our risk minimisation setting, this means that both the adjustment weights and the proposal kernels need to be adapted. As we will see below, these two problems are in general intertwined; however, in the following it will be clear that the two criteria CSD and KLD behave differently at this point. Because the criteria are rather involved, it is interesting to study their behaviour as the number of particles N grows to infinity. This is done in Theorem 1, which shows that the CSD $d_{\chi^2}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}})$ and KLD $d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}})$ converges to $d_{\chi^2}(\mu^* \parallel \pi_{\psi}^*)$ and $d_{\text{KL}}(\mu^* \parallel \pi_{\psi}^*)$, respectively, where

$$\begin{aligned} \mu^*(\xi, \tilde{\xi}) &\triangleq \frac{v(\xi)l(\xi, \tilde{\xi})}{\iint v(\xi)l(\xi, \tilde{\xi}) d\xi d\tilde{\xi}}, \\ \pi_{\psi}^*(\xi, \tilde{\xi}) &\triangleq \frac{v(\xi)\Psi(\xi)r(\xi, \tilde{\xi})}{\iint v(\xi)\Psi(\xi)r(\xi, \tilde{\xi}) d\xi d\tilde{\xi}}. \end{aligned} \tag{2.25}$$

The expressions (2.25) of the limiting distributions then allow for deriving the adjustment multiplier weight function Ψ and the proposal density l minimising the corresponding discrepancy measures. In absence of constraints (when Ψ and l can be chosen arbitrarily), the optimal solution for both the CSD and the KLD consists of setting $\Psi = \Psi^*$ and $r = r^*$, where

$$\Psi^*(\xi) \triangleq \int l(\xi, \tilde{\xi}) d\tilde{\xi} = \int \frac{l(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})} r(\xi, \tilde{\xi}) d\tilde{\xi}, \tag{2.26}$$

$$r^*(\xi, \tilde{\xi}) \triangleq l(\xi, \tilde{\xi})/\Psi^*(\xi). \tag{2.27}$$

This choice coincides with the so-called *optimal sampling strategy* proposed by Hürzeler and Künsch (1998) and developed further by Künsch (2005), which turns out to be *optimal* (in absence of constraints) in our risk-minimisation setting.

Remark 1 The limiting distributions μ^* and π_{ψ}^* have nice interpretations within the framework of state space models (see the previous section). In this setting, the limiting distribution μ^* at time k is the joint distribution $\phi_{k:k+1|k+1}$ of the *filtered* couple $X_{k:k+1}$, that is, the distribution of $X_{k:k+1}$ conditionally on the observation record $Y_{0:k+1}$; this can be seen as the asymptotic target distribution of our particle model. Moreover, the limiting distribution π^* at time k is only slightly more intricate: Its first marginal corresponds to the filtering distribution at time k reweighted by the adjustment function Ψ , which is typically used for incorporating information from the new observation Y_{k+1} . The second marginal of π^* is then obtained by propagating this weighted filtering distribution through the Markovian dynamics of the proposal kernel R ; thus, π_{ψ}^* describes completely the asymptotic instrumental distribution of the APF, and the two quantities $d_{\text{KL}}(\mu^* \parallel \pi^*)$ and $d_{\chi^2}(\mu^* \parallel \pi^*)$ reflect the asymptotic discrepancy between the true model and the particle model at the given time step.

In presence of constraints on the choice of Ψ and r , the optimisation of the adjustment weight function and the proposal kernel density is intertwined. By the so-called *chain rule for entropy* (see Cover and Thomas 1991, Theorem 2.2.1), we have

$$d_{\text{KL}}(\mu^* \parallel \pi_{\psi}^*) = \int \frac{v(\xi)}{v(\Psi^*)} \Psi^*(\xi) \log \left(\frac{\Psi^*(\xi)/v(\Psi^*)}{\Psi(\xi)/v(\Psi)} \right) d\xi$$

where $v(f) \triangleq \int v(\xi)f(\xi) d\xi$. Hence, if the optimal adjustment function can be chosen freely, then, whatever the choice of the proposal kernel is, the best choice is still $\Psi_{\text{KL},r}^* = \Psi^*$: the best that we can do is to choose $\Psi_{\text{KL},r}^*$ such that the two marginal distributions $\xi \mapsto \int \mu^*(\xi, \tilde{\xi}) d\tilde{\xi}$ and $\xi \mapsto \int \pi^*(\xi, \tilde{\xi}) d\tilde{\xi}$ are identical. If the choices of the weight adjustment function and the proposal kernels are constrained (if, e.g., the weight should be chosen in a pre-specified family of functions or the proposal kernel belongs to a parametric family), nevertheless, the optimisation of Ψ and r decouple asymptotically. The optimisation for the CSD does not lead to such a nice decoupling of the adjustment function and the proposal transition; nevertheless, an explicit expression for the adjustment multiplier weights can still be found in this case:

$$\Psi_{\chi^2,r}^*(\xi) \triangleq \sqrt{\int \frac{l^2(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})} d\tilde{\xi}}$$

$$= \sqrt{\int \frac{l^2(\xi, \tilde{\xi})}{r^2(\xi, \tilde{\xi})} r(\xi, \tilde{\xi}) d\tilde{\xi}} \tag{2.28}$$

Compared to (2.26), the optimal adjustment function for the CSD is the L^2 (rather than the L^1) norm of $\xi \mapsto l^2(\xi, \tilde{\xi})/r^2(\xi, \tilde{\xi})$. Since $l(\xi, \tilde{\xi}) = \Psi^*(\xi)r^*(\xi, \tilde{\xi})$ (see definitions (2.26) and (2.27)), we obtain, not surprisingly, if we set $r = r^*$, $\Psi_{\chi^2, r}^*(\xi) = \Psi^*(\xi)$.

Using this risk minimisation formulation, it is possible to select the adjustment weight function as well as the proposal kernel by minimising either the CSD or the KLD criteria. Of course, compared to the sophisticated adaptation strategies considered for adaptive importance sampling, we focus on elementary schemes, the computational burden being quickly a limiting factor in the SMC context.

To simplify the presentation, we consider in the sequel the adaptation of the proposal kernel; as shown above, it is of course possible and worthwhile to jointly optimise the adjustment weight and the proposal kernel, but for clarity we prefer to postpone the presentation of such a technique to a future work. The optimisation of the adjustment weight function is in general rather complex: indeed, as mentioned above, the computation of the optimal adjustment weight function requires the computing of an integral. This integral can be evaluated in closed form only for a rather limited number of models; otherwise, a numerical approximation (based on cubature formulae, Monte Carlo etc.) is required, which may therefore incur a quite substantial computational cost. If proper simplifications and approximations are not found (which are, most often, model specific) the gains in efficiency are not necessarily worth the extra cost. In state space (tracking) problems simple and efficient approximations, based either on the EKF or the UKF (see for example Andrieu et al. 2003 or Shen et al. 2004), have been proposed for several models, but the validity of this sort of approximations cannot necessarily be extended to more general models.

In the light of the discussion above, a natural strategy for adaptive design of π_N^{aux} is to minimise the empirical estimate \mathcal{E} (or CV^2) of the KLD (or CSD) over all proposal kernels belonging to some parametric family $\{r_\theta\}_{\theta \in \Theta}$. This can be done using straightforward adaptations of the two methods described in Sect. 2.1. We postpone a more precise description of the algorithms and implementation issues to Sect. 4, where more rigorous measure-theoretic notation is introduced and the main theoretical results are stated.

3 Notation and definitions

To state precisely the results, we will now use measure-theoretic notation. In the following we assume that all random variables are defined on a common probability space

$(\Omega, \mathcal{F}, \mathbb{P})$ and let, for any general state space $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$, $\mathcal{P}(\mathcal{E})$ and $\mathbb{B}(\mathcal{E})$ be the sets of probability measures on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ and measurable functions from \mathcal{E} to \mathbb{R} , respectively.

A kernel K from $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ to some other state space $(\tilde{\mathcal{E}}, \mathcal{B}(\tilde{\mathcal{E}}))$ is called *finite* if $K(\xi, \tilde{\mathcal{E}}) < \infty$ for all $\xi \in \mathcal{E}$ and *Markovian* if $K(\xi, \tilde{\mathcal{E}}) = 1$ for all $\xi \in \mathcal{E}$. Moreover, K induces two operators, one transforming a function $f \in \mathbb{B}(\mathcal{E} \times \tilde{\mathcal{E}})$ satisfying $\int_{\tilde{\mathcal{E}}} |f(\xi, \tilde{\xi})| K(\xi, d\tilde{\xi}) < \infty$ into another function

$$\xi \mapsto K(\xi, f) \triangleq \int_{\tilde{\mathcal{E}}} f(\xi, \tilde{\xi}) K(\xi, d\tilde{\xi})$$

in $\mathbb{B}(\mathcal{E})$; the other transforms a measure $\nu \in \mathcal{P}(\mathcal{E})$ into another measure

$$A \mapsto \nu K(A) \triangleq \int_{\mathcal{E}} K(\xi, A) \nu(d\xi) \tag{3.1}$$

in $\mathcal{P}(\tilde{\mathcal{E}})$. Furthermore, for any probability measure $\mu \in \mathcal{P}(\mathcal{E})$ and function $f \in \mathbb{B}(\mathcal{E})$ satisfying $\int_{\mathcal{E}} |f(\xi)| \mu(d\xi) < \infty$, we write $\mu(f) \triangleq \int_{\mathcal{E}} f(\xi) \mu(d\xi)$.

The *outer product* of the measure γ and the kernel T , denoted by $\gamma \otimes T$, is defined as the measure on the product space $\mathcal{E} \times \tilde{\mathcal{E}}$, equipped with the product σ -algebra $\mathcal{B}(\mathcal{E}) \otimes \mathcal{B}(\tilde{\mathcal{E}})$, satisfying

$$\gamma \otimes T(A) \triangleq \iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \gamma(d\xi) T(\xi, d\tilde{\xi}) \mathbb{1}_A(\xi, \xi') \tag{3.2}$$

for any $A \in \mathcal{B}(\mathcal{E}) \otimes \mathcal{B}(\tilde{\mathcal{E}})$. For a non-negative function $f \in \mathbb{B}(\mathcal{E})$, we define the modulated measure $\gamma[f]$ on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ by

$$\nu[f](A) \triangleq \nu(f \mathbb{1}_A), \tag{3.3}$$

for any $A \in \mathcal{B}(\mathcal{E})$.

In the sequel, we will use the following definition. A set \mathcal{C} of real-valued functions on \mathcal{E} is said to be *proper* if the following conditions hold: (i) \mathcal{C} is a linear space; (ii) if $g \in \mathcal{C}$ and f is measurable with $|f| \leq |g|$, then $f \in \mathcal{C}$; (iii) for all $c \in \mathbb{R}$, the constant function $f \equiv c$ belongs to \mathcal{C} .

Definition 1 A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on \mathcal{E} is said to be *consistent* for \mathcal{C} for the probability measure $\nu \in \mathcal{P}(\mathcal{E})$ and the set \mathcal{C} if, for any $f \in \mathcal{C}$, as $N \rightarrow \infty$,

$$\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \nu(f),$$

$$\Omega_N^{-1} \max_{1 \leq i \leq M_N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0,$$

$$\Omega_N \triangleq \sum_{i=1}^{M_N} \omega_{N,i}.$$

Alternatively, we will sometimes say that the weighted sample in Definition 1 *targets* the measure ν .

Thus, suppose that we are given a weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ targeting $\nu \in \mathcal{P}(\mathcal{E})$. We wish to transform this sample into a new weighted particle sample approximating the probability measure

$$\mu(\cdot) \triangleq \frac{\nu L(\cdot)}{\nu L(\tilde{\mathcal{E}})} = \frac{\int_{\mathcal{E}} L(\xi, \cdot) \nu(d\xi)}{\int_{\mathcal{E}} L(\xi', \tilde{\mathcal{E}}) \nu(d\xi')} \tag{3.4}$$

on some other state space $(\tilde{\mathcal{E}}, \mathcal{B}(\tilde{\mathcal{E}}))$. Here L is a finite transition kernel from $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ to $(\tilde{\mathcal{E}}, \mathcal{B}(\tilde{\mathcal{E}}))$. As suggested by Pitt and Shephard (1999), an auxiliary variable corresponding to the selected stratum, and target the measure

$$\mu_N^{\text{aux}}(\{i\} \times A) \triangleq \frac{\omega_{N,i} L(\xi_{N,i}, \tilde{\mathcal{E}})}{\sum_{j=1}^{M_N} \omega_{N,j} L(\xi_{N,j}, \tilde{\mathcal{E}})} \left[\frac{L(\xi_{N,i}, A)}{L(\xi_{N,i}, \tilde{\mathcal{E}})} \right] \tag{3.5}$$

on the product space $\{1, \dots, M_N\} \times \mathcal{E}$. Since μ_N is the marginal distribution of μ_N^{aux} with respect to the particle position, we may sample from μ_N by simulating instead a set $\{(I_{N,i}, \tilde{\xi}_{N,i})\}_{i=1}^{M_N}$ of indices and particle positions from the instrumental distribution

$$\pi_N^{\text{aux}}(\{i\} \times A) \triangleq \frac{\omega_{N,i} \psi_{N,i}}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}} R(\xi_{N,i}, A) \tag{3.6}$$

and assigning each draw $(I_{N,i}, \tilde{\xi}_{N,i})$ the weight

$$\tilde{\omega}_{N,i} \triangleq \psi_{N,I_{N,i}}^{-1} \frac{dL(\xi_{N,I_{N,i}}, \cdot)}{dR(\xi_{N,I_{N,i}}, \cdot)}(\tilde{\xi}_{N,i})$$

being proportional to $d\mu_N^{\text{aux}}/d\pi_N^{\text{aux}}(I_{N,i}, \tilde{\xi}_{N,i})$ —the formal difference with (2.23) lies only in the use of Radon-Nykodym derivatives of the two kernels rather than densities with respect to Lebesgue measure. Hereafter, we discard the indices and take $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$ as an approximation of μ . The algorithm is summarised below.

Algorithm 1 Nonadaptive APF

Require: $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ targets ν .

- 1: Draw $\{I_{N,i}\}_{i=1}^{M_N} \sim \mathcal{M}(\tilde{M}_N, \{\omega_{N,j} \psi_{N,j} / \sum_{\ell=1}^{M_N} \omega_{N,\ell} \psi_{N,\ell}\}_{j=1}^{M_N})$,
- 2: simulate $\{\tilde{\xi}_{N,i}\}_{i=1}^{M_N} \sim \otimes_{i=1}^{M_N} R(\xi_{N,I_{N,i}}, \cdot)$,
- 3: set, for all $i \in \{1, \dots, M_N\}$,

$$\tilde{\omega}_{N,i} \leftarrow \psi_{N,I_{N,i}}^{-1} dL(\xi_{N,I_{N,i}}, \cdot) / dR(\xi_{N,I_{N,i}}, \cdot)(\tilde{\xi}_{N,i}).$$

- 4: take $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$ as an approximation of μ .

4 Theoretical results

Consider the following assumptions.

- (A1) The initial sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, \mathbf{C}) .
- (A2) There exists a function $\Psi : \mathcal{E} \rightarrow \mathbb{R}^+$ such that $\psi_{N,i} = \Psi(\xi_{N,i})$; moreover, $\Psi \in \mathbf{C} \cap L^1(\mathcal{E}, \nu)$ and $L(\cdot, \tilde{\mathcal{E}}) \in \mathbf{C}$.

Under these assumptions we define for $(\xi, \tilde{\xi}) \in \mathcal{E} \times \tilde{\mathcal{E}}$ the weight function

$$\Phi(\xi, \tilde{\xi}) \triangleq \Psi^{-1}(\xi) \frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\tilde{\xi}), \tag{4.1}$$

so that for every index i , $\tilde{\omega}_{N,i} = \Phi(\xi_{N,i}, \tilde{\xi}_{N,i})$. The following result describes how the consistency property is passed through one step of the APF algorithm. A somewhat less general version of this result was also proved in Douc et al. (2008, Theorem 3.1).

Proposition 1 Assume (A1, A2). Then the weighted sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$ is consistent for $(\nu, \tilde{\mathbf{C}})$, where $\tilde{\mathbf{C}} \triangleq \{f \in L^1(\tilde{\mathcal{E}}, \mu), L(\cdot, |f|) \in \mathbf{C}\}$.

The result above is a direct consequence of Lemma 2 and the fact that the set $\tilde{\mathbf{C}}$ is proper.

Let μ and ν be two probability measures in $\mathcal{P}(\mathbf{A})$ such that μ is absolutely continuous with respect to ν . We then recall that the KLD and the CSD are, respectively, given by

$$d_{\text{KL}}(\mu \parallel \nu) \triangleq \int_{\mathbf{A}} \log[d\mu/d\nu(\lambda)] \mu(d\lambda),$$

$$d_{\chi^2}(\mu \parallel \nu) \triangleq \int_{\mathbf{A}} [d\mu/d\nu(\lambda) - 1]^2 \nu(d\lambda).$$

Define the two probability measures on the product space $(\mathcal{E} \times \tilde{\mathcal{E}}, \mathcal{B}(\mathcal{E}) \otimes \mathcal{B}(\tilde{\mathcal{E}}))$:

$$\begin{aligned} \mu^*(A) &\triangleq \frac{\nu \otimes L}{\nu L(\tilde{\mathcal{E}})}(A) \\ &= \frac{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) L(\xi, d\xi') \mathbb{1}_A(\xi, \xi')}{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) L(\xi, d\xi')}, \end{aligned} \tag{4.2}$$

$$\begin{aligned} \pi_{\Psi}^*(A) &\triangleq \frac{\nu[\Psi] \otimes R}{\nu(\Psi)}(A) \\ &= \frac{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) \Psi(\xi) R(\xi, d\xi') \mathbb{1}_A(\xi, \xi')}{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) \Psi(\xi) R(\xi, d\xi')}, \end{aligned} \tag{4.3}$$

where $A \in \mathcal{B}(\mathcal{E}) \otimes \mathcal{B}(\tilde{\mathcal{E}})$ and the outer product \otimes of a measure and a kernel is defined in (3.2).

Theorem 1 Assume (A1, A2). Then the following holds as $N \rightarrow \infty$.

(i) If $L(\cdot, |\log \Phi|) \in \mathcal{C} \cap L^1(\mathcal{E}, \nu)$, then

$$d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \xrightarrow{\mathbb{P}} d_{\text{KL}}(\mu^* \parallel \pi_\Psi^*), \tag{4.4}$$

(ii) If $L(\cdot, \Phi) \in \mathcal{C}$, then

$$d_{\chi^2}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \xrightarrow{\mathbb{P}} d_{\chi^2}(\mu^* \parallel \pi_\Psi^*), \tag{4.5}$$

Additionally, \mathcal{E} and CV^2 , defined in (2.8) and (2.4) respectively, converge to the same limits.

Theorem 2 Assume (A1, A2). Then the following holds as $N \rightarrow \infty$.

(i) If $L(\cdot, |\log \Phi|) \in \mathcal{C} \cap L^1(\mathcal{E}, \nu)$, then

$$\mathcal{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \xrightarrow{\mathbb{P}} d_{\text{KL}}(\mu^* \parallel \pi_\Psi^*). \tag{4.6}$$

(ii) If $L(\cdot, \Phi) \in \mathcal{C}$, then

$$\text{CV}^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \xrightarrow{\mathbb{P}} d_{\chi^2}(\mu^* \parallel \pi_\Psi^*). \tag{4.7}$$

Next, it is shown that the adjustment weight function can be chosen so as to minimize the RHS of (4.4) and (4.5).

Proposition 2 Assume (A1, A2). Then the following holds.

(i) If $L(\cdot, |\log \Phi|) \in \mathcal{C} \cap L^1(\mathcal{E}, \nu)$, then

$$\begin{aligned} \arg \min_{\Psi} d_{\text{KL}}(\mu^* \parallel \pi_\Psi^*) &\triangleq \Psi_{\text{KL},R}^* \\ \text{where } \Psi_{\text{KL},R}^*(\xi) &\triangleq L(\xi, \tilde{\mathcal{E}}). \end{aligned}$$

(ii) If $L(\cdot, \Phi) \in \mathcal{C}$, then

$$\begin{aligned} \arg \min_{\Psi} d_{\chi^2}(\mu^* \parallel \pi_\Psi^*) &\triangleq \Psi_{\chi^2,R}^* \\ \text{where } \Psi_{\chi^2,R}^*(\xi) &\triangleq \sqrt{\int_{\tilde{\mathcal{E}}} \frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\tilde{\xi}) L(\xi, d\tilde{\xi})}. \end{aligned}$$

It is worthwhile to notice that the optimal adjustment weights for the KLD do not depend on the proposal kernel R . The minimal value $d_{\text{KL}}(\mu^* \parallel \pi_{\Psi_{\text{KL},R}^*}^*)$ of the limiting KLD is the conditional relative entropy between μ^* and π^* .

In both cases, letting $R(\cdot, A) = L(\cdot, A)/L(\cdot, \tilde{\mathcal{E}})$ yields, as we may expect, the optimal adjustment multiplier weight function $\Psi_{\text{KL},R}^*(\cdot) = \Psi_{\chi^2,R}^*(\cdot) = L(\cdot, \tilde{\mathcal{E}})$, resulting in uniform importance weights $\tilde{\omega}_{N,i} \equiv 1$.

It is possible to relate the asymptotic CSD (4.5) between μ_N^{aux} and π_N^{aux} to the asymptotic variance of the estimator $\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i})$ of an expectation $\mu(f)$ for a given integrable target function f . More specifically, suppose that $\tilde{M}_N/M_N \rightarrow \ell \in [0, \infty]$ as $N \rightarrow \infty$ and that the initial sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ satisfies, for all f belonging to a given

class \mathcal{A} of functions, the central limit theorem

$$a_N \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} [f(\xi_{N,i}) - \mu(f)] \xrightarrow{\mathcal{D}} \mathcal{N}[0, \sigma^2(f)], \tag{4.8}$$

where the sequence $\{a_N\}_N$ is such that $a_N M_N \rightarrow \beta \in [0, \infty)$ as $N \rightarrow \infty$ and $\sigma : \mathcal{A} \rightarrow \mathbb{R}^+$ is a functional. Then the sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$ produced in Algorithm 1 is, as showed in Douc et al. (2008, Theorem 3.2), asymptotically normal for a class of functions $\tilde{\mathcal{A}}$ in the sense that, for all $f \in \tilde{\mathcal{A}}$,

$$\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} [f(\tilde{\xi}_{N,i}) - \mu(f)] \xrightarrow{\mathcal{D}} \mathcal{N}[0, \tilde{\sigma}^2[\Psi](f)],$$

where

$$\begin{aligned} \tilde{\sigma}^2[\Psi](f) &= \sigma^2\{L[\cdot, f - \mu(f)]\} / [vL(\cdot, \tilde{\mathcal{E}})]^2 \\ &\quad + \beta \ell^{-1} v(\Psi R\{\cdot, \Phi^2[f - \mu(f)]^2\}) v(\Psi) / [vL(\tilde{\mathcal{E}})]^2 \end{aligned}$$

and, recalling the definition (3.3) of a modulated measure,

$$\begin{aligned} v(\Psi R\{\cdot, \Phi^2[f - \mu(f)]^2\}) v(\Psi) / [vL(\tilde{\mathcal{E}})]^2 &= \mu^2(|f|) d_{\chi^2}\{\mu^*[|f|] / \mu^*(|f|) \parallel \pi^*\} \\ &\quad - 2\mu(f) \mu(f_+^{1/2}) d_{\chi^2}\{\mu^*[f_+^{1/2}] / \mu^*(f_+^{1/2}) \parallel \pi^*\} \\ &\quad + 2\mu(f) \mu(f_-^{1/2}) d_{\chi^2}\{\mu^*[f_-^{1/2}] / \mu^*(f_-^{1/2}) \parallel \pi^*\} \\ &\quad + \mu^2(f) d_{\chi^2}(\mu^* \parallel \pi^*) + \mu^2(|f|) - \mu^2(f). \end{aligned} \tag{4.9}$$

Here $f_+ \triangleq \max(f, 0)$ and $f_- \triangleq \max(-f, 0)$ denote the positive and negative parts of f , respectively, and $\mu^*(|f|)$ refers to the expectation of the extended function $|f| : (\xi, \tilde{\xi}) \in \mathcal{E} \times \tilde{\mathcal{E}} \mapsto |f(\tilde{\xi})| \in \mathbb{R}^+$ under μ^* (and similarly for $\mu^*(f_\pm^{1/2})$). From (4.9) we deduce that decreasing $d_{\chi^2}(\mu^* \parallel \pi^*)$ will imply a decrease of asymptotic variance if the discrepancy between μ^* and modulated measure $\mu^*[|f|] / \mu^*(|f|)$ is not too large, that is, we deal with a target function f with a regular behaviour in the support of $\mu^*(\mathcal{E} \times \cdot)$.

5 Adaptive importance sampling

5.1 APF adaptation by minimisation of estimated KLD and CSD over a parametric family

Assume that there exists a random noise variable ϵ , having distribution λ on some measurable space $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$, and a family $\{F_\theta\}_{\theta \in \Theta}$ of mappings from $\mathcal{E} \times \mathcal{A}$ to $\tilde{\mathcal{E}}$ such that we are able to simulate $\tilde{\xi} \sim R_\theta(\xi, \cdot)$, for $\xi \in \mathcal{E}$, by simulating

$\epsilon \sim \lambda$ and letting $\tilde{\xi} = F_\theta(\xi, \epsilon)$. We denote by Φ_θ the importance weight function associated with R_θ , see (4.1) and set $\Phi_\theta \circ F_\theta(\xi, \epsilon) \triangleq \Phi_\theta(\xi, F_\theta(\xi, \epsilon))$.

Assume that (A1) holds and suppose that we have simulated, as in the first step of Algorithm 1, indices $\{I_{N,i}\}_{i=1}^{\tilde{M}_N}$ and noise variables $\{\epsilon_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$. Now, keeping these indices and noise variables fixed, we can form an idea of how the KLD varies with θ via the mapping $\theta \mapsto \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{N,I_{N,i}}, \epsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$. Similarly, the CSD can be studied by using CV^2 instead of \mathcal{E} . This suggests an algorithm in which the particles are reproposed using $R_{\theta_N^*}$, with $\theta_N^* = \arg \min_{\theta \in \Theta} \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{N,I_{N,i}}, \epsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$.

This procedure is summarised in Algorithm 2, and its modification for minimisation of the empirical CSD is straightforward.

Algorithm 2 Adaptive APF

Require: (A1)

- 1: Draw $\{I_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \mathcal{M}(\tilde{M}_N, \{\omega_{N,j} \psi_{N,j} / \sum_{\ell=1}^{\tilde{M}_N} \omega_{N,\ell} \psi_{N,\ell}\}_{j=1}^{\tilde{M}_N})$,
- 2: simulate $\{\epsilon_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$,
- 3: $\theta_N^* \leftarrow \arg \min_{\theta \in \Theta} \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{N,I_{N,i}}, \epsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$,
- 4: set

$$\tilde{\xi}_{N,i} \stackrel{\forall i}{\leftarrow} F_{\theta_N^*}(\xi_{N,I_{N,i}}, \epsilon_{N,i})$$

- 5: update

$$\tilde{\omega}_{N,i} \stackrel{\forall i}{\leftarrow} \Phi_{\theta_N^*}(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}),$$

- 6: let $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ approximate μ .

Remark 2 A slight modification of Algorithm 2, lowering the added computational burden, is to apply the adaptation mechanism only when the estimated KLD (or CSD) is above a chosen threshold.

Remark 3 It is possible to establish a law of large numbers as well as a central limit theorem for the algorithm above, similarly to what has been done for the nonadaptive auxiliary particle filter in Douc et al. (2008) and Olsson et al. (2007).

More specifically, suppose again that (4.8) holds for similar $(A, \beta, \sigma(\cdot))$ and that $\tilde{M}_N / M_N \rightarrow \ell \in [0, \infty]$ as $N \rightarrow \infty$. Then the sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ produced in Algorithm 2 is asymptotically normal for a class of functions \tilde{A} in the sense that, for all $f \in \tilde{A}$,

$$\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} [f(\tilde{\xi}_{N,i}) - \mu(f)] \xrightarrow{\mathcal{D}} \mathcal{N}[0, \tilde{\sigma}_{\theta_N^*}^2(f)],$$

where

$$\begin{aligned} \tilde{\sigma}_{\theta_N^*}^2(f) &\triangleq \beta \ell^{-1} \nu(\Psi R_{\theta_N^*} \{ \cdot, \Phi_{\theta_N^*}^2 [f - \mu(f)]^2 \}) \nu(\Psi) / [\nu L(\tilde{\Xi})]^2 \\ &\quad + \sigma^2(L\{ \cdot, [f - \mu(f)] \}) / [\nu L(\tilde{\Xi})]^2 \end{aligned}$$

and θ_N^* minimises the asymptotic KLD. The complete proof of this result is however omitted for brevity.

5.2 APF adaptation by cross-entropy (CE) methods

Here we construct an algorithm which selects a proposal kernel from a parametric family in a way that minimises the KLD between the instrumental mixture distribution and the target mixture μ_N^{aux} (defined in (3.5)). Thus, recall that we are given an initial sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{\tilde{M}_N}$; we then use IS to approximate the target auxiliary distribution μ_N^{aux} by sampling from the instrumental auxiliary distribution

$$\pi_{N,\theta}^{\text{aux}}(\{i\} \times A) \triangleq \frac{\omega_{N,i} \psi_{N,i}}{\sum_{j=1}^{\tilde{M}_N} \omega_{N,j} \psi_{N,j}} R_\theta(\xi_{N,i}, A), \tag{5.1}$$

which is a straightforward modification of (3.6) where R is replaced by R_θ , that is, a Markovian kernel from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\tilde{\mathcal{X}}, \mathcal{B}(\tilde{\mathcal{X}}))$ belonging to the parametric family $\{R_\theta(\xi, \cdot) : \xi \in \mathcal{X}, \theta \in \Theta\}$.

We aim at finding the parameter θ^* which realises the minimum of $\theta \mapsto d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta}^{\text{aux}})$ over the parameter space Θ , where

$$\begin{aligned} d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta}^{\text{aux}}) &= \sum_{i=1}^{\tilde{M}_N} \int_{\tilde{\mathcal{X}}} \log \left(\frac{d\mu_N^{\text{aux}}}{d\pi_{N,\theta}^{\text{aux}}} (i, \tilde{\xi}) \right) \mu_N^{\text{aux}}(i, d\tilde{\xi}). \end{aligned} \tag{5.2}$$

Since the expectation in (5.2) is intractable in most cases, the key idea is to approximate it iteratively using IS from more and more accurate approximations—this idea has been successfully used in CE methods; see e.g. Rubinstein and Kroese (2004). At iteration ℓ , denote by $\theta_N^\ell \in \Theta$ the current fit of the parameter. Each iteration of the algorithm is split into two steps: In the first step we sample, following Algorithm 1 with $\tilde{M}_N = \tilde{M}_N^\ell$ and $R = R_{\theta_N^\ell}$, M_N^ℓ particles $\{(I_{N,i}^\ell, \tilde{\xi}_{N,i}^\ell)\}_{i=1}^{M_N^\ell}$ from $\pi_{N,\theta_N^\ell}^{\text{aux}}$. Note that the adjustment multiplier weights are kept constant during the iterations, a limitation which is however not necessary. The second step consists of computing the exact solution

$$\theta_N^{\ell+1} \triangleq \arg \min_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_{N,i}^\ell}{\tilde{\Omega}_N^\ell} \log \left(\frac{d\mu_N^{\text{aux}}}{d\pi_{N,\theta}^{\text{aux}}} (I_{N,i}^\ell, \tilde{\xi}_{N,i}^\ell) \right) \tag{5.3}$$

to the problem of minimising the Monte Carlo approximation of (5.2). In the case where the kernels L and R_θ have densities, denoted by l and r_θ , respectively, with respect to a common reference measure on $(\tilde{\mathcal{X}}, \mathcal{B}(\tilde{\mathcal{X}}))$, the minimisation program (5.3) is equivalent to

$$\theta_N^{\ell+1} \triangleq \arg \max_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_{N,i}^\ell}{\tilde{\Omega}_N^\ell} \log r_\theta(\xi_{I_{N,i}^\ell}, \tilde{\xi}_{N,i}^\ell). \tag{5.4}$$

This algorithm is helpful only in situations where the minimisation problem (5.3) is sufficiently simple for allowing for closed-form minimisation; this happens, for example, if the objective function is a convex combination of concave functions, whose minimum either admits a (simple) closed-form expression or is straightforward to minimise numerically. As mentioned in Sect. 2.1, this is generally the case when the function $r_\theta(\xi, \cdot)$ belongs to an exponential family for any $\xi \in \tilde{\mathcal{X}}$.

Since this optimisation problem closely resembles the Monte Carlo EM algorithm, all the implementation details of these algorithms can be readily transposed to that context; see for example Levine and Casella (2001), Eickhoff et al. (2004), and Levine and Fan (2004). Because we use very simple models, convergence occurs, as seen in Sect. 6, within only few iterations. When choosing the successive particle sample sizes $\{\tilde{M}_N^\ell\}_{\ell=1}^L$, we are facing a trade-off between precision of the approximation (5.3) of (5.2) and computational cost. Numerical evidence typically shows that these sizes may, as high precision is less crucial here than when generating the final population from $\pi_{N,\theta_N^L}^{\text{aux}}$, be relatively small compared to the final size \tilde{M}_N . Besides, it is possible (and even theoretically recommended) to increase the number of particles with the iteration index, since, heuristically, high accuracy is less required at the first steps. In the current implementation in Sect. 6, we will show that fixing a priori the total number of iterations and using the same number $\tilde{M}_N^\ell = \tilde{M}_N/L$ of particles at each iteration may provide satisfactory results in a first run.

6 Application to state space models

For an illustration of our findings we return to the framework of state space models in Sect. 2.2 and apply the CE-adaptation-based particle method to *filtering* in nonlinear state space models of type

$$\begin{aligned} X_{k+1} &= m(X_k) + \sigma_w(X_k)W_{k+1}, \quad k \geq 0, \\ Y_k &= X_k + \sigma_v V_k, \quad k \geq 0, \end{aligned} \tag{6.1}$$

where the parameter σ_v and the \mathbb{R} -valued functions (m, σ_w) are known, and $\{W_k\}_{k=1}^\infty$ and $\{V_k\}_{k=0}^\infty$ are mutually independent sequences of independent standard normal-distributed

Algorithm 3 CE-based adaptive APF

Require: $\{(\xi_i, \omega_i)\}_{i=1}^{M_N}$ targets ν .

- 1: Choose an arbitrary θ_N^0 ,
- 2: **for** $\ell = 0, \dots, L - 1$ **do** \triangleright More intricate criteria are sensible
- 3: draw

$$\{I_{N,i}^\ell\}_{i=1}^{\tilde{M}_N^\ell} \sim \mathcal{M}(\tilde{M}_N^\ell, \{\omega_j \psi_{N,j} / \sum_{n=1}^{M_N} \omega_n \psi_{N,n}\}_{j=1}^{M_N}),$$

- 4: simulate $\{\tilde{\xi}_{N,i}^\ell\}_{i=1}^{\tilde{M}_N^\ell} \sim \otimes_{i=1}^{\tilde{M}_N^\ell} R_{\theta_N^\ell}(\xi_{I_{N,i}^\ell}, \cdot)$,
- 5: update

$$\tilde{\omega}_{N,i} \stackrel{\forall i}{\leftarrow} \Phi_{\theta_N^\ell}(\xi_{I_{N,i}^\ell}, \tilde{\xi}_{N,i}^\ell),$$

- 6: compute, with available closed-form,

$$\theta_N^{\ell+1} \triangleq \arg \min_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_{N,i}^\ell}{\tilde{\Omega}_N^\ell} \log \left(\frac{d\mu_N^{\text{aux}}}{d\tau_{N,\theta}^{\text{aux}}} (I_{N,i}^\ell, \tilde{\xi}_{N,i}^\ell) \right),$$

- 7: **end for**
- 8: run Algorithm 1 with $R = R_{\theta_N^L}$.

variables. In this setting, we wish to approximate the filter distributions $\{\phi_{k|k}\}_{k \geq 0}$, defined in Sect. 2.2 as the posterior distributions of X_k given $Y_{0:k}$ (recall that $Y_{0:k} \triangleq (Y_0, \dots, Y_k)$), which in general lack closed-form expressions. For models of this type, the optimal weight and density defined in (2.26) and (2.27), respectively, can be expressed in closed-form:

$$\Psi_k^*(x) = \mathcal{N}(Y_{k+1}; m(x), \sqrt{\sigma_w^2(x) + \sigma_v^2}), \tag{6.2}$$

where $\mathcal{N}(x; \mu, \sigma) \triangleq \exp(-(x - \mu)^2 / (2\sigma^2)) / \sqrt{2\pi\sigma^2}$ and

$$r_k^*(x, x') = \mathcal{N}(x'; \tau(x, Y_{k+1}), \eta(x)), \tag{6.3}$$

with

$$\tau(x, Y_{k+1}) \triangleq \frac{\sigma_w^2(x)Y_{k+1} + \sigma_v^2 m(x)}{\sigma_w^2(x) + \sigma_v^2},$$

$$\eta^2(x) \triangleq \frac{\sigma_w^2(x)\sigma_v^2}{\sigma_w^2(x) + \sigma_v^2}.$$

We may also compute the chi-square optimal adjustment multiplier weight function $\Psi_{\chi^2, Q}^*$ when the prior kernel is used as proposal: at time k ,

$$\begin{aligned} \Psi_{\chi^2, Q}^*(x) &\propto \sqrt{\frac{2\sigma_v^2}{2\sigma_w^2(x) + \sigma_v^2}} \\ &\times \exp \left(-\frac{Y_{k+1}^2}{\sigma_v^2} + \frac{m(x)}{2\sigma_w^2(x) + \sigma_v^2} [2Y_{k+1} - m(x)] \right). \end{aligned} \tag{6.4}$$

We recall from Proposition 2 that the optimal adjustment weight function for the KLD is given by $\Psi_{KL,Q}^*(x) = \Psi_k^*(x)$.

In these intentionally chosen simple example we will consider, at each timestep k , adaption over the family

$$\{R_\theta(x, \cdot) \triangleq \mathcal{N}(\tau(x, Y_{k+1}), \theta\eta(x)) : x \in \mathbb{R}, \theta > 0\} \quad (6.5)$$

of proposal kernels. In addition, we keep the adjustment weights constant, that is $\Psi(x) = 1$.

The mode of each proposal kernel is centered at the mode of the optimal kernel, and the variance is proportional to the inverse of the Hessian of the optimal kernel at the mode. Let $r_\theta(x, x') \triangleq \mathcal{N}(x'; \tau(x, Y_{k+1}), \theta\eta(x))$ denote the density of $R_\theta(x, \cdot)$ with respect to the Lebesgue measure. In this setting, at every timestep k , a closed-form expression of the KLD between the target and proposal distributions is available:

$$\begin{aligned} d_{KL}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta}^{\text{aux}}) &= \sum_{i=1}^{M_N} \frac{\omega_{N,i} \psi_{N,i}^*}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}^*} \\ &\times \left[\log \left(\frac{\psi_{N,i}^* \Omega_N}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}^*} \right) + \log \theta \right. \\ &\left. + \frac{1}{2} \left(\frac{1}{\theta^2} - 1 \right) \right], \end{aligned} \quad (6.6)$$

where we set $\psi_{N,i}^* \triangleq \Psi^*(\xi_{N,i})$ and $\Omega_N = \sum_{i=1}^{M_N} \omega_{N,i}$.

As we are scaling the optimal standard deviation, it is obvious that

$$\theta_N^* \triangleq \arg \min_{\theta > 0} d_{KL}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta}^{\text{aux}}) = 1, \quad (6.7)$$

which may also be inferred by straightforward derivation of (6.6) with respect to θ . This provides us with a reference to which the parameter values found by our algorithm can be compared. Note that the instrumental distribution $\pi_{N,\theta_N^*}^{\text{aux}}$ differs from the target distribution μ_N^{aux} by the adjustment weights used: recall that every instrumental distribution in the family considered has uniform adjustment weights, $\Psi(x) = 1$, whereas the overall optimal proposal has, since it is equal to the target distribution μ_N^{aux} , the optimal weights defined in (6.2). This entails that

$$\begin{aligned} d_{KL}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta_N^*}^{\text{aux}}) &= \sum_{i=1}^{M_N} \omega_{N,i} \frac{\psi_{N,i}^*}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}^*} \log \left(\frac{\psi_{N,i}^* \Omega_N}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}^*} \right), \end{aligned} \quad (6.8)$$

which is zero if all the optimal weights are equal.

The implementation of Algorithm 3 is straightforward as the optimisation program (5.4) has the following closed-form solution:

$$\theta_N^{\ell+1} = \left\{ \sum_{i=1}^{M_N} \frac{\tilde{\omega}_{N,i}^{\theta_N^\ell}}{\tilde{\Omega}_N^{\theta_N^\ell} \eta^2_{N,I_{N,i}^{\theta_N^\ell}}} \left(\tilde{\xi}_{N,i}^{\theta_N^\ell} - \tau_{N,I_{N,i}^{\theta_N^\ell}} \right)^2 \right\}^{1/2}, \quad (6.9)$$

where $\tau_{N,i} \triangleq \tau(\xi_{N,i}, Y_{k+1})$ and $\eta_{N,i}^2 \triangleq \eta^2(\xi_{N,i})$. This is a typical case where the family of proposal kernels allows for efficient minimisation. Richer families sharing this property may also be used, but here we are voluntarily willing to keep this toy example as simple as possible.

We will study the following special case of the model (6.1):

$$m(x) \equiv 0, \quad \sigma_w(x) = \sqrt{\beta_0 + \beta_1 x^2}.$$

This is the classical Gaussian *autoregressive conditional heteroscedasticity* (ARCH) model observed in noise (see Bollerslev et al. 1994). In this case an experiment was conducted where we compared:

- (i) A plain nonadaptive particle filter for which $\Psi \equiv 1$, that is, the bootstrap particle filter of Gordon et al. (1993),
- (ii) An auxiliary filter based on the prior kernel and chi-square optimal weights $\Psi_{\chi^2,Q}^*$,
- (iii) Adaptive bootstrap filters with uniform adjustment multiplier weights using numerical minimisation of the empirical CSD and
- (iv) The empirical KLD (Algorithm 2),
- (v) An adaptive bootstrap filter using direct minimisation of $d_{KL}(\mu_N^{\text{aux}} \parallel \pi_{N,\theta}^{\text{aux}})$, see (6.7),
- (vi) A CE-based adaptive bootstrap filter, and as a reference,
- (vi) An optimal auxiliary particle filter, i.e. a filter using the optimal weight and proposal kernel defined in (6.2) and (6.3), respectively.

This experiment was conducted for the parameter set $(\beta_0, \beta_1, \sigma_v^2) = (1, 0.99, 10)$, yielding (since $\beta_1 < 1$) a geometrically ergodic ARCH(1) model (see Chen and Chen 2000, Theorem 1); the noise variance σ_v^2 is equal to 1/10 of the stationary variance, which here is equal to $\sigma_s^2 = \beta_0/(1 - \beta_1)$, of the state process.

In order to design a challenging test of the adaptation procedures we set, after having run a hundred burn-in iterations to reach stationarity of the hidden states, the observations to be constantly equal to $Y_k = 6\sigma_s$ for every $k \geq 110$. We expect that the bootstrap filter, having a proposal transition kernel with constant mean $m(x) = 0$, will have a large mean square error (MSE) due a poor number of particles in regions where the likelihood is significant. We aim at illustrating that the adaptive algorithms, whose transition kernels

have the same mode as the optimal transition kernel, adjust automatically the variance of the proposals to that of the optimal kernel and reach performances comparable to that of the optimal auxiliary filter.

For these observation records, Fig. 1 displays MSEs estimates based on 500 filter means. Each filter used 5,000 particles. The reference values used for the MSE estimates were obtained using the optimal auxiliary particle filter with as many as 500,000 particles. This also provided a set from which the initial particles of every filter were drawn, hence allowing for initialisation at the filter distribution a few steps before the outlying observations.

The CE-based filter of Algorithm 3 was implemented in its most simple form, with the inside loop using a constant number of $M_N^\ell = N/10 = 500$ particles and only $L = 5$ iterations: a simple prefatory study of the model indicated that the Markov chain $\{\theta_N^\ell\}_{l \geq 0}$ stabilised around the value reached in the very first step. We set $\theta_N^0 = 10$ to avoid initialising at the optimal value.

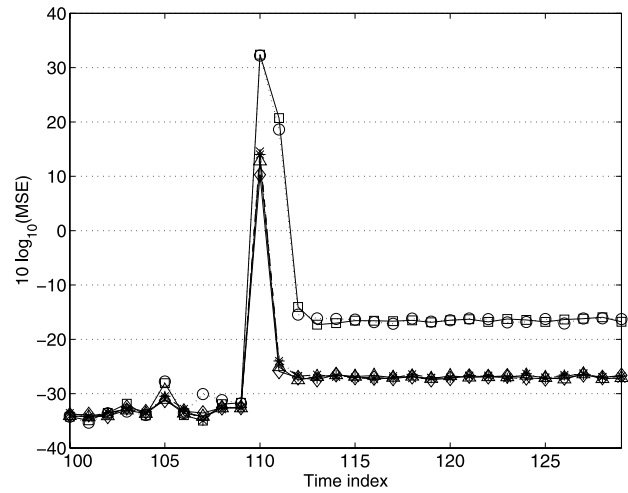
It can be seen in Fig. 1a that using the CSD optimal weights combined with the prior kernel as proposal does not improve on the plain bootstrap filter, precisely because the observations were chosen in such a way that the prior kernel was helpless. On the contrary, Figs. 1a and 1b show that the adaptive schemes perform exactly similarly to the optimal filter: they all success in finding the optimal scale of the standard deviation, and using uniform adjustment weights instead of optimal ones does not impact much.

We observe clearly a change of regime, beginning at step 110, corresponding to the outlying constant observations. The adaptive filters recover from the changepoint in one timestep, whereas the bootstrap filter needs several. More important is that the adaptive filters (as well as the optimal one) reduce, in the regime of the outlying observations, the MSE of the bootstrap filter by a factor 10.

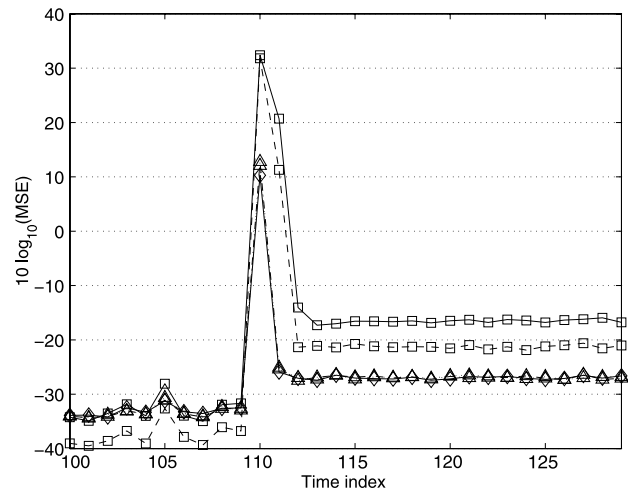
Moreover, for a comparison with fixed simulation budget, we ran a bootstrap filter with $3N = 15,000$ particles. This corresponds to the same simulation budget as the CE-based adaptive scheme with N particles, which is, in this setting, the fastest of our adaptive algorithms. In our setting, the CE-based filter is measured to expand the plain bootstrap runtime by a factor 3, although a basic study of algorithmic complexity shows that this factor should be closer to $\sum_{\ell=1}^L M_N^\ell / N = 1.5$ —the difference rises from Matlab benefitting from the vectorisation of the plain bootstrap filter, not from the iterative nature of the CE.

The conclusion drawn from Fig. 1b is that for an equal runtime, the adaptive filter outperforms, by a factor 3.5, the bootstrap filter using even three times more particles.

Acknowledgements The authors are grateful to Prof. Paul Fearnhead for encouragements and useful recommendations, and to the anonymous reviewers for insightful comments and suggestions that improved the presentation of the paper.



(a) Auxiliary filter based on chi-square optimal weights $\Psi_{\chi^2, Q}^*$ and prior kernel K (o), adaptive filters minimising the empirical KLD (*) and CSD (x), and reference filters listed below.



(b) CE-based adaption (Δ , dash-dotted line), bootstrap filter with $3N$ particles (\square , dashed line), and reference filters listed below.

Fig. 1 Plot of MSE performances (on log-scale) on the ARCH model with $(\beta_0, \beta_1, \sigma_v^2) = (1, 0.99, 10)$. Reference filters common to both plots are: the bootstrap filter (\square , continuous line), the optimal filter with weights Ψ^* and proposal kernel density r^* (\diamond), and a bootstrap filter using a proposal with parameter θ_N^* minimising the current KLD (Δ , continuous line). The MSE values are computed using $N = 5,000$ particles—except for the reference bootstrap using $3N$ particles (\square , dashed line)—and 1,000 runs of each algorithm

Appendix A: Proofs

A.1 Proof of Theorems 1 and 2

We preface the proofs of Theorems 1 and 2 with the following two lemmas.

Lemma 1 Assume (A2). Then the following identities hold.

- (i) $d_{\text{KL}}(\mu^* \parallel \pi_{\Psi}^*) = \nu \otimes L\{\log[\Phi \nu(\Psi)/\nu L(\tilde{\mathcal{E}})]\}/\nu L(\tilde{\mathcal{E}})$,
- (ii) $d_{\chi^2}(\mu^* \parallel \pi_{\Psi}^*) = \nu(\Psi) \nu \otimes L(\Phi)/[\nu L(\tilde{\mathcal{E}})]^2 - 1$.

Proof We denote by $q(\xi, \xi')$ the Radon-Nikodym derivative of the probability measure μ^* with respect to $\nu \otimes R$ (where the outer product \otimes of a measure and a kernel is defined in (3.2)), that is,

$$q(\xi, \xi') \triangleq \frac{\frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\xi')}{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) L(\xi, d\xi')}, \tag{A.1}$$

and by $p(\xi)$ the Radon-Nikodym derivative of the probability measure π^* with respect to $\nu \otimes R$:

$$p(\xi) = \frac{\Psi(\xi)}{\nu(\Psi)}. \tag{A.2}$$

Using the notation above and definition (4.1) of the weight function Φ , we have

$$\frac{\Phi(\xi, \xi') \nu(\Psi)}{\nu L(\tilde{\mathcal{E}})} = \frac{\nu(\Psi) \frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\xi')}{\Psi(\xi) \nu L(\tilde{\mathcal{E}})} = p^{-1}(\xi) q(\xi, \xi').$$

This implies that

$$\begin{aligned} d_{\text{KL}}(\mu^* \parallel \pi_{\Psi}^*) &= \iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) R(\xi, d\xi') q(\xi, \xi') \\ &\quad \times \log(p^{-1}(\xi) q(\xi, \xi')) \\ &= \nu \otimes L\{\log[\Phi \nu(\Psi)/\nu L(\tilde{\mathcal{E}})]\}/\nu L(\tilde{\mathcal{E}}), \end{aligned}$$

which establishes assertion (i). Similarly, we may write

$$\begin{aligned} d_{\chi^2}(\mu^* \parallel \pi_{\Psi}^*) &= \iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(d\xi) R(\xi, d\xi') p^{-1}(\xi) q^2(\xi, \xi') - 1 \\ &= \frac{\iint_{\mathcal{E} \times \tilde{\mathcal{E}}} \nu(\Psi) \nu(d\xi) R(\xi, d\xi') [\frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\xi')]^2 \Psi^{-1}(\xi)}{[\nu L(\tilde{\mathcal{E}})]^2} - 1 \\ &= \nu(\Psi) \nu \otimes L(\Phi)/[\nu L(\tilde{\mathcal{E}})]^2 - 1, \end{aligned}$$

showing assertion (ii). □

Lemma 2 Assume (A1, A2) and let $\mathbf{C}^* \triangleq \{f \in \mathbb{B}(\mathcal{E} \times \tilde{\mathcal{E}}) : L(\cdot, |f|) \in \mathbf{C} \cap L^1(\mathcal{E}, \nu)\}$. Then, for all $f \in \mathbf{C}^*$, as $N \rightarrow \infty$,

$$\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \nu \otimes L(f)/\nu L(\tilde{\mathcal{E}})$$

Proof It is enough to prove that

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \nu \otimes L(f)/\nu(\Psi), \tag{A.3}$$

for all $f \in \mathbf{C}^*$; indeed, since the function $f \equiv 1$ belongs to \mathbf{C}^* under (A2), the result of the lemma will follow from (A.3) by Slutsky’s theorem. Define the measure $\varphi(A) \triangleq \nu(\Psi \mathbb{1}_A)/\nu(\Psi)$, with $A \in \mathcal{B}(\mathcal{E})$. By applying Theorem 1 in Douc and Moulines (2008) we conclude that the weighted sample $\{(\xi_{N,i}, \psi_{N,i})\}_{i=1}^{\tilde{M}_N}$ is consistent for $(\varphi, \{f \in L^1(\mathcal{E}, \varphi) : \Psi|f| \in \mathbf{C}\})$. Moreover, by Theorem 2 in the same paper this is also true for the uniformly weighted sample $\{(\xi_{N,I_{N,i}}, 1)\}_{i=1}^{\tilde{M}_N}$ (see the proof of Theorem 3.1 in Douc et al. 2008 for details). By definition, for $f \in \mathbf{C}^*$, $\varphi \otimes R(\Phi|f|)\nu(\Psi) = \nu \otimes L(|f|) < \infty$ and $\Psi R(\cdot, \Phi|f|) = L(\cdot, |f|) \in \mathbf{C}$. Hence, we conclude that $R(\cdot, \Phi|f|)$ and thus $R(\cdot, \Phi f)$ belong to the proper set $\{f \in L^1(\mathcal{E}, \varphi) : \Psi|f| \in \mathbf{C}\}$. This implies the convergence

$$\begin{aligned} &\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E} \left[\tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \middle| \mathcal{F}_N \right] \\ &= \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} R(\xi_{N,I_{N,i}}, \Phi f) \xrightarrow{\mathbb{P}} \varphi \otimes R(\Phi f) \\ &= \nu \otimes L(f)/\nu(\Psi), \end{aligned} \tag{A.4}$$

where $\mathcal{F}_N \triangleq \sigma(\{\xi_{N,I_{N,i}}\}_{i=1}^{\tilde{M}_N})$ denotes the σ -algebra generated by the selected particles. It thus suffices to establish that

$$\begin{aligned} &\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \left\{ \mathbb{E} \left[\tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \middle| \mathcal{F}_N \right] \right. \\ &\quad \left. - \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \right\} \xrightarrow{\mathbb{P}} 0, \end{aligned} \tag{A.5}$$

and we do this, following the lines of the proof of Theorem 1 in Douc and Moulines (2008), by verifying the two conditions of Theorem 11 in the same work. The sequence

$$\left\{ \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E} \left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \middle| \mathcal{F}_N \right] \right\}_N$$

is tight since it tends to $\nu \otimes L(|f|)/\nu(\Psi)$ in probability (cf. (A.4)). Thus, the first condition is satisfied. To verify the second condition, take $\epsilon > 0$ and consider, for any $C > 0$, the decomposition

$$\begin{aligned} &\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E} \left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \right. \\ &\quad \left. \times \mathbb{1}_{\{\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \geq \epsilon\}} \middle| \mathcal{F}_N \right] \end{aligned}$$

$$\begin{aligned} &\leq \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} R(\xi_{N,IN,i}, \Phi|f|\mathbb{1}_{\{\Phi|f|\geq C\}}) \\ &\quad + \mathbb{1}_{\{\epsilon\tilde{M}_N < C\}} \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E} \left[\tilde{\omega}_{N,i} |f(\xi_{N,IN,i}, \tilde{\xi}_{N,i})| \middle| \mathcal{F}_N \right]. \end{aligned}$$

Since $R(\cdot, \Phi f)$ belongs to the proper set $\{f \in L^1(\tilde{\mathcal{E}}, \varphi) : \Psi|f| \in \mathbb{C}\}$, so does the function $R(\cdot, \Phi|f|\mathbb{1}_{\{\Phi|f|\geq C\}})$. Thus, since the indicator $\mathbb{1}_{\{\epsilon\tilde{M}_N < C\}}$ tends to zero, we conclude that the upper bound above has the limit $\varphi \otimes R(\Phi|f|\mathbb{1}_{\{\Phi|f|\geq C\}})$; however, by dominated convergence this limit can be made arbitrarily small by increasing C . Hence

$$\begin{aligned} &\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E} \left[\tilde{\omega}_{N,i} |f(\xi_{N,IN,i}, \tilde{\xi}_{N,i})| \right. \\ &\quad \left. \times \mathbb{1}_{\{|\tilde{\omega}_{N,i} |f(\xi_{N,IN,i}, \tilde{\xi}_{N,i})| \geq \epsilon\}} \middle| \mathcal{F}_N \right] \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

which verifies the second condition of Theorem 11 in Douc and Moulines (2008). Thus, (A.5) follows. \square

Proof of Theorem 1 We start with (i). In the light of Lemma 1 we establish the limit

$$d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \xrightarrow{\mathbb{P}} \nu \otimes L\{\log[\Phi \nu(\Psi)/\nu L(\tilde{\mathcal{E}})]\}/\nu L(\tilde{\mathcal{E}}), \tag{A.6}$$

as $N \rightarrow \infty$. Hence, recall the definition (given in Sect. 4) of the KLD and write, for any index $m \in \{1, \dots, \tilde{M}_N\}$,

$$\begin{aligned} &d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \\ &= \sum_{i=1}^{M_N} \mathbb{E}_{\mu_N^{\text{aux}}} \left[\log \Phi(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i \right] \\ &\quad \times \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) + \log \left[\frac{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}}{\sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, \tilde{\mathcal{E}})} \right], \end{aligned} \tag{A.7}$$

where $\mathbb{E}_{\mu_N^{\text{aux}}}$ denotes the expectation associated with the random measure μ_N^{aux} . For each term of the sum in (A.7) we have

$$\begin{aligned} &\mathbb{E}_{\mu_N^{\text{aux}}} \left[\log \Phi(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i \right] \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) \\ &= \frac{\omega_{N,i} L(\xi_{N,i}, \log \Phi)}{\sum_{j=1}^{M_N} \omega_{N,j} L(\xi_{N,j}, \tilde{\mathcal{E}})}, \end{aligned}$$

and by using the consistency of $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ (under (A1)) we obtain the limit

$$\begin{aligned} &\sum_{i=1}^{M_N} \mathbb{E}_{\mu_N^{\text{aux}}} \left[\log \Phi(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i \right] \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) \\ &\quad \xrightarrow{\mathbb{P}} \nu \otimes L(\log \Phi)/\nu L(\tilde{\mathcal{E}}), \end{aligned}$$

where we used that $L(\cdot, |\log \Phi|) \in \mathbb{C}$ by assumption, implying, since \mathbb{C} is proper, $L(\cdot, \log \Phi) \in \mathbb{C}$. Moreover, under (A2), by the continuous mapping theorem,

$$\log \left[\frac{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}}{\sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, \tilde{\mathcal{E}})} \right] \xrightarrow{\mathbb{P}} \log[\nu(\Psi)/\nu L(\tilde{\mathcal{E}})],$$

yielding

$$\begin{aligned} &d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \\ &\quad \xrightarrow{\mathbb{P}} \nu \otimes L(\log \Phi)/\nu L(\tilde{\mathcal{E}}) + \log[\nu(\Psi)/\nu L(\tilde{\mathcal{E}})] \\ &= \nu \otimes L\{\log[\Phi \nu(\Psi)/\nu L(\tilde{\mathcal{E}})]\}/\nu L(\tilde{\mathcal{E}}), \end{aligned}$$

which establishes (A.6) and, consequently, (i).

To prove (ii) we show that

$$d_{\chi^2}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \xrightarrow{\mathbb{P}} \nu(\Psi) \nu \otimes L(\Phi)/[\nu L(\tilde{\mathcal{E}})]^2 - 1 \tag{A.8}$$

and apply Lemma 1. Thus, recall the definition of the CSD and write, for any index $m \in \{1, \dots, \tilde{M}_N\}$,

$$\begin{aligned} &d_{\chi^2}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) = \mathbb{E}_{\mu_N^{\text{aux}}} \left[\frac{d\mu_N^{\text{aux}}}{d\pi_N^{\text{aux}}}(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \right] - 1 \\ &= \sum_{i=1}^{M_N} \mathbb{E}_{\mu_N^{\text{aux}}} \left[\frac{d\mu_N^{\text{aux}}}{d\pi_N^{\text{aux}}}(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i \right] \\ &\quad \times \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) - 1. \end{aligned}$$

Here

$$\begin{aligned} &\mathbb{E}_{\mu_N^{\text{aux}}} \left[\frac{d\mu_N^{\text{aux}}}{d\pi_N^{\text{aux}}}(\xi_{N,IN,m}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i \right] \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) \\ &= \omega_{N,i} L(\xi_{N,i}, \Phi) \left[\sum_{j=1}^{M_N} \omega_{N,j} L(\xi_{N,j}, \tilde{\mathcal{E}}) \right]^{-2} \\ &\quad \times \sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}, \end{aligned}$$

and using the consistency of $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ yields the limit

$$\sum_{i=1}^{M_N} \mathbb{E}_{\mu_N^{\text{aux}}} \left[\frac{d\mu_N^{\text{aux}}}{d\pi_N^{\text{aux}}}(\xi_{N, I_{N,m}}, \tilde{\xi}_{N,m}) \Big| I_{N,m} = i \right] \mu_N^{\text{aux}}(\{i\} \times \tilde{\mathcal{E}}) \xrightarrow{\mathbb{P}} \nu(\Psi) \nu \otimes L(\Phi) / [\nu L(\tilde{\mathcal{E}})]^2.$$

which proves (A.8). This completes the proof of (ii). \square

Proof of Theorem 2 Applying directly Lemma 2 for $f = \log \Phi$ (which belongs to \mathbf{C}^* by assumption) and the limit (A.3) for $f \equiv 1$ yields, by the continuous mapping theorem,

$$\begin{aligned} & \mathcal{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \\ &= \tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} \log \tilde{\omega}_{N,i} + \log(\tilde{M}_N \tilde{\Omega}_N^{-1}) \\ &\xrightarrow{\mathbb{P}} \nu \otimes L(\log \Phi) / \nu L(\tilde{\mathcal{E}}) + \log[\nu(\Psi) / \nu L(\tilde{\mathcal{E}})] \\ &= \nu \otimes L\{\log[\Phi \nu(\Psi) / \nu L(\tilde{\mathcal{E}})]\} / \nu L(\tilde{\mathcal{E}}). \end{aligned}$$

Now, we complete the proof of assertion (i) by applying Lemma 1.

We turn to (ii). Since Φ belongs to \mathbf{C}^* by assumption, we obtain, by applying Lemma 2 together with (A.3),

$$\begin{aligned} & \text{CV}^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \\ &= (\tilde{M}_N \tilde{\Omega}_N^{-1}) \tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i}^2 - 1 \\ &\xrightarrow{\mathbb{P}} \nu_{\text{KL}}(\Psi) \triangleq \nu(\Psi) \nu \otimes L(\Phi) / [\nu L(\tilde{\mathcal{E}})]^2 - 1. \end{aligned} \quad (\text{A.9})$$

From this (ii) follows via Lemma 1. \square

A.2 Proof of Proposition 2

Define by $q(\xi) \triangleq \int_{\tilde{\mathcal{E}}} R(\xi, d\xi') q(\xi, \xi')$ the marginal density of the measure on $(\tilde{\mathcal{E}}, \mathcal{B}(\tilde{\mathcal{E}}))$, $A \in \mathcal{B}(\tilde{\mathcal{E}}) \mapsto \mu^*(A \times \tilde{\mathcal{E}})$. We denote by $q(\xi'|\xi) = q(\xi, \xi')/q(\xi)$ the conditional distribution. By the chain rule of the entropy, (the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other), we may split the KLD between μ^* and π^* as follows,

$$\begin{aligned} & d_{\text{KL}}(\mu_N^{\text{aux}} \parallel \pi_N^{\text{aux}}) \\ &= \int_{\tilde{\mathcal{E}}} \nu(d\xi) q(\xi) \log(p^{-1}(\xi) q(\xi)) \\ &\quad + \iint_{\tilde{\mathcal{E}} \times \tilde{\mathcal{E}}} \nu(d\xi) R(\xi, d\xi') q(\xi, \xi') \log q(\xi|\xi'). \end{aligned}$$

The second term in the RHS of the previous equation does not depend on the adjustment multiplier weight Ψ . The first

term is canceled if we set $p = q$, i.e. if

$$\frac{\Psi(\xi)}{\nu(\Psi)} = \int_{\tilde{\mathcal{E}}} R(\xi, d\xi') q(\xi, \xi') = \frac{L(\xi, \tilde{\mathcal{E}})}{\int_{\tilde{\mathcal{E}}} \nu(d\xi) L(\xi, \tilde{\mathcal{E}})},$$

which establishes assertion (i).

Consider now assertion (ii). Note first that

$$\begin{aligned} & \iint_{\tilde{\mathcal{E}} \times \tilde{\mathcal{E}}} \nu(d\xi) R(\xi, d\xi') p^{-1}(\xi) q^2(\xi, \xi') - 1 \\ &= \int_{\tilde{\mathcal{E}}} \nu(d\xi) p^{-1}(\xi) g^2(\xi) - 1 \\ &= \nu^2(g) \left\{ \int_{\tilde{\mathcal{E}}} \nu(d\xi) \frac{g^2(\xi)}{p(\xi) \nu^2(g)} - 1 \right\} + \nu^2(g) - 1, \end{aligned} \quad (\text{A.10})$$

where

$$g^2(\xi) = \int_{\tilde{\mathcal{E}}} R(\xi, d\xi') q^2(\xi, \xi').$$

The first term on the RHS of (A.10) is the CSD between the probability distributions associated with the densities $g/\nu(g)$ and $\Psi/\nu(\Psi)$ with respect to ν . The second term does not depend on Ψ and the optimal value of the adjustment multiplier weight is obtained by canceling the first term. This establishes assertion (ii).

References

Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice Hall, New York (1979)

Andrieu, C., Davy, M., Doucet, A.: Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions. *IEEE Trans. Signal Process.* **51**(7), 1762–1770 (2003)

Arouna, B.: Robbins-monro algorithms and variance reduction in finance. *J. Comput. Finance* **7**(2) (2004)

Bollerslev, T., Engle, R.F., Nelson, D.: ARCH models. In: Engle, R.F., McFadden, D. (eds.) *Handbook of Econometrics*, pp. 2959–3038. North-Holland, Amsterdam (1994)

Cappé, O., Moulines, E., Rydén, T.: *Inference in Hidden Markov Models*. Springer, Berlin (2005), <http://www.tsi.enst.fr/~cappel/ihmm/>

Cappé, O., Douc, R., Guillin, A., Marin, J.M., Robert, C.P.: Adaptive importance sampling in general mixture classes. *Stat. Comput.* (2008, this issue)

Carpenter, J., Clifford, P., Fearnhead, P.: An improved particle filter for non-linear problems. *IEE Proc. Radar Sonar Navig.* **146**, 2–7 (1999)

Chen, M., Chen, G.: Geometric ergodicity of nonlinear autoregressive models with changing conditional variances. *Can. J. Stat.* **28**(3), 605–613 (2000)

Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. Wiley, New York (1991)

de Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**, 19–67 (2005)

Douc, R., Moulines, E.: Limit theorems for weighted samples with applications to sequential Monte Carlo. *Ann. Stat.* **36** (2008, to appear), [arXiv:math.ST/0507042](https://arxiv.org/abs/math/0507042)

Douc, R., Moulines, E., Olsson, J.: On the auxiliary particle filter. *Probab. Math. Stat.* **28**(2) (2008)

- Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10**, 197–208 (2000)
- Doucet, A., De Freitas, N., Gordon, N. (eds.): *Sequential Monte Carlo Methods in Practice*. Springer, New York (2001)
- Eickhoff, J.C., Zhu, J., Amemiya, Y.: On the simulation size and the convergence of the Monte Carlo EM algorithm via likelihood-based distances. *Stat. Probab. Lett.* **67**(2), 161–171 (2004)
- Evans, M., Swartz, T.: Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems. *Stat. Sci.* **10**, 254–272 (1995)
- Fearnhead, P.: Computational methods for complex stochastic systems: a review of some alternatives to mcmc. *Stat. Comput.* **18**, 151–171 (2008)
- Fearnhead, P., Liu, Z.: On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B* **69**(4), 590–605 (2007)
- Fort, G., Moulines, E.: Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Stat.* **31**(4), 1220–1259 (2003)
- Fox, D.: Adapting the sample size in particle filters through KLD-sampling. *Int. J. Rob. Res.* **22**(11), 985–1004 (2003)
- Geweke, J.: Bayesian inference in econometric models using Monte-Carlo integration. *Econometrica* **57**(6), 1317–1339 (1989)
- Givens, G., Raftery, A.: Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Am. Stat. Assoc.* **91**(433), 132–141 (1996)
- Gordon, N., Salmond, D., Smith, A.F.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F Radar Signal Process.* **140**, 107–113 (1993)
- Ho, Y.C., Lee, R.C.K.: A Bayesian approach to problems in stochastic estimation and control. *IEEE Trans. Autom. Control* **9**(4), 333–339 (1964)
- Hu, X.L., Schon, T.B., Ljung, L.: A basic convergence result for particle filtering. *IEEE Trans. Signal Process.* **56**(4), 1337–1348 (2008)
- Hürzeler, M., Künsch, H.R.: Monte Carlo approximations for general state-space models. *J. Comput. Graph Stat.* **7**, 175–193 (1998)
- Kailath, T., Sayed, A., Hassibi, B.: *Linear Estimation*. Prentice Hall, New York (2000)
- Kong, A., Liu, J.S., Wong, W.: Sequential imputation and Bayesian missing data problems. *J. Am. Stat. Assoc.* **89**(278–288), 590–599 (1994)
- Künsch, H.R.: Recursive Monte-Carlo filters: algorithms and theoretical analysis. *Ann. Stat.* **33**(5), 1983–2021 (2005), [arXiv:math.ST/0602211](https://arxiv.org/abs/math/0602211)
- Legland, F., Oudjane, N.: A sequential algorithm that keeps the particle system alive. *Tech. rep., Rapport de recherche 5826*, INRIA (2006), [ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-5826.pdf](http://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-5826.pdf)
- Levine, R.A., Casella, G.: Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Stat.* **10**(3), 422–439 (2001)
- Levine, R.A., Fan, J.: An automated (Markov chain) Monte Carlo EM algorithm. *J. Stat. Comput. Simul.* **74**(5), 349–359 (2004)
- Liu, J.: *Monte Carlo Strategies in Scientific Computing*. Springer, Berlin (2004)
- Oh, M.S., Berger, J.O.: Adaptive importance sampling in Monte Carlo integration. *J. Stat. Comput. Simul.* **41**(3–4), 143–168 (1992)
- Oh, M.S., Berger, J.O.: Integration of multimodal functions by Monte Carlo importance sampling. *J. Am. Stat. Assoc.* **88**(422), 450–456 (1993)
- Olsson, J., Moulines, E., Douc, R.: Improving the performance of the two-stage sampling particle filter: a statistical perspective. In: *Proceedings of the IEEE/SP 14th Workshop on Statistical Signal Processing*, Madison, USA, pp. 284–288 (2007)
- Pitt, M.K., Shephard, N.: Filtering via simulation: Auxiliary particle filters. *J. Am. Stat. Assoc.* **94**(446), 590–599 (1999)
- Ristic, B., Arulampalam, M., Gordon, A.: *Beyond Kalman Filters: Particle Filters for Target Tracking*. Artech House, Norwood (2004)
- Rubinstein, R.Y., Kroese, D.P.: *The Cross-Entropy Method*. Springer, Berlin (2004)
- Shen, C., van den Hengel, A., Dick, A., Brooks, M.J.: Enhanced importance sampling: unscented auxiliary particle filtering for visual tracking. In: *AI 2004: Advances in Artificial Intelligence. Lecture Notes in Comput. Sci.*, vol. 3339, pp. 180–191. Springer, Berlin (2004)
- Shephard, N., Pitt, M.: Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**(3), 653–667 (1997), erratum in **91**, 249–250 (2004)
- Soto, A.: Self adaptive particle filter. In: Kaelbling, L.P., Saffiotti, A. (eds.) *Proceedings of the 19th International Joint Conferences on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, pp. 1398–1406 (2005), <http://www.ijcai.org/proceedings05.php>
- Stephens, M., Donnelly, P.: Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**(4), 605–655 (2000), with discussion and a reply by the authors
- Straka, O., Simandl, M.: Particle filter adaptation based on efficient sample size. In: *Proceedings of the 14th IFAC Symposium on System Identification*, Newcastle, Australia, pp. 991–996 (2006)
- Van der Vaart, A.W.: *Asymptotic Statistics*. Cambridge University Press, Cambridge (1998)
- Wei, G.C.G., Tanner, M.A.: A Monte-Carlo implementation of the EM algorithm and the poor man's Data Augmentation algorithms. *J. Am. Stat. Assoc.* **85**, 699–704 (1991)