16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ADAPTIVE METHODS FOR SEQUENTIAL IMPORTANCE SAMPLING WITH APPLICATION TO STATE SPACE MODELS

Julien Cornebise, Éric Moulines, and Jimmy Olsson[1]

Ecole Nationale Supérieure des Télécommunications,
46 Rue Barrault, 75634 Paris Cedex 13, France
email: {julien.cornebise, eric.moulines}@enst.fr

Center of Mathematical Sciences, Lund University
Box 118, SE-22100 Lund, Sweden
phone: + (46) 46 222 85 52 , email: jimmy@maths.lth.se

ABSTRACT

In this paper we discuss new adaptive proposal strategies for sequential Monte Carlo algorithms—also known as particle filters—relying on new 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 [10]) 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.

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. In this paper we develop methods for adjusting adaptively the importance sampling distribution of the particle filter.

Several authors have focused on adaptation of the size of the particle sample, e.g., by increasing the number of particles until the total weight mass reaches a positive threshold, see [12], or until the *Kullback-Leibler divergence* (KLD) between the true and estimated target distributions is below a given threshold, see [7].

Unarguably, setting an appropriate sample size is a key ingredient of any statistical estimation procedure. 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 imply the need of an unreasonably large number of samples for decreasing the variance of the estimate under a specified value.

This points to the need for adapting the importance distributions of the particle filter, e.g., by adjusting the proposal kernels. Less work has been done on this topic, with the notable exception of [14], in which the so-called *optimal kernel* is approximated, and [2], in which the expectation of a cost function, such as the mean square error or the negated *effective sample size*, is minimised over a parametric family of kernels.

Most of the algorithms described above require tools, such as the *coefficient of variation* (CV) proposed by [10], for evaluating on-line the quality of the particle swarm. In this article we justify theoretically that the CV can be used for estimating sequentially the asymptotic *chi-square distance* (CSD) between the auxiliary SMC target and importance distributions. Moreover, a new empirical estimate of the asymptotic KLD having a computational complexity which is linear in the number of particles is proposed. We also identify a type of auxiliary SMC adjustment multiplier weights which minimize these asymptotic discrepancy measures for a given proposal kernel. Finally, we use the empirical CSD and KLD estimates for designing adaptively the auxiliary particle filter importance distributions and apply the proposed algorithms to optimal filtering in state space models.

Complete proofs of the theoretical results as well as more detailed explanations and additional simulations are given in [3].

[1]This work was partly supported by the National Research Agency (ANR) under the program "ANR-05-BLAN-0299"

2. THE AUXILIARY PARTICLE FILTER

Let ν be a probability measure on some general state space $(\Xi, \mathscr{B}(\Xi))$ and let $\{(\xi_i, \omega_{N,i})\}_{i=1}^{M_N}$ be a set of *particles* on Ξ with associated weights such that $\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} f(\xi_i)$, with $\Omega_N \triangleq \sum_{i=1}^{M_N} \omega_{N,i}$, approximates expectations $\int_\Xi f(\xi)\, \nu(\mathrm{d}\xi)$ for all f in some specified class of functions. We wish to transform this sample into a new weighted particle sample approximating the probability measure

$$\mu(\cdot) \triangleq \frac{\int_\Xi L(\xi, \cdot)\, \nu(\mathrm{d}\xi)}{\int_\Xi L(\xi', \tilde{\Xi})\, \nu(\mathrm{d}\xi')} \tag{1}$$

on some other state space $(\tilde{\Xi}, \mathscr{B}(\tilde{\Xi}))$ where L is a finite transition kernel from $(\Xi, \mathscr{B}(\Xi))$ to $(\tilde{\Xi}, \mathscr{B}(\tilde{\Xi}))$. A natural strategy for achieving this is to replace ν in (1) by its particle approximation, yielding

$$\mu_N(\cdot) \triangleq \sum_{i=1}^{M_N} \frac{\omega_{N,i} L(\xi_i, \tilde{\Xi})}{\sum_{j=1}^{M_N} \omega_{N,j} L(\xi_j, \tilde{\Xi})} \left[L(\xi_i, \cdot) / L(\xi_i, \tilde{\Xi}) \right]$$

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 expensive accept-reject techniques; see [9] and [11]. This difficulty can be overcome by simulating new particles $\{\tilde{\xi}_{N,i}\}_{i=1}^{\tilde{M}_N}$ from the instrumental mixture distribution

$$\pi_N(\cdot) \triangleq \sum_{i=1}^{M_N} \frac{\omega_{N,i} \psi_{N,i}}{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}} R(\xi_i, \cdot)\,,$$

where $\{\psi_{N,i}\}_{i=1}^{M_N}$ are positive numbers referred to as *adjustment multiplier weights* and R is a markovian kernel, and associating these particles with weights $\{\mathrm{d}\mu_N / \mathrm{d}\pi_N(\tilde{\xi}_{N,i})\}_{i=1}^{\tilde{M}_N}$. In this setting, a new particle position is simulated from the stratum $R(\xi_i, \cdot)$ with probability proportional to $\omega_{N,i}\psi_{N,i}$. Unfortunately, the Radon-Nikodym derivative $\mathrm{d}\mu_N / \mathrm{d}\pi_N$ is expensive to evaluate since this involves summing over M_N terms. Thus, we introduce, as suggested by [14], an *auxiliary variable* corresponding to the selected stratum, and target instead the measure

$$\mu_N^{\mathrm{aux}}(\{i\} \times A) \triangleq \frac{\omega_{N,i} L(\xi_i, \tilde{\Xi})}{\sum_{j=1}^{M_N} \omega_{N,j} L(\xi_j, \tilde{\Xi})} \left[L(\xi_i, A) / L(\xi_i, \tilde{\Xi}) \right]$$

on the product space $\{1,\dots,M_N\}\times\Xi$. 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 $\{(I_{N,i},\tilde{\xi}_{N,i})\}_{i=1}^{\tilde{M}_N}$ of indices and particle positions from the instrumental distribution

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

and then associating each draw $(I_{N,i},\tilde{\xi}_{N,i})$ with the weight $\tilde{\omega}_{N,i}\triangleq\psi_{N,I_{N,i}}^{-1}\mathrm{d}L(\xi_{I_{N,i}},\cdot)/\mathrm{d}R(\xi_{I_{N,i}},\cdot)(\tilde{\xi}_{N,i})\propto\mathrm{d}\mu_N^{\text{aux}}/\mathrm{d}\pi_N^{\text{aux}}(I_{N,i},\tilde{\xi}_{N,i})$. Hereafter, we discard the indices and take $\{(\tilde{\xi}_{N,i},\tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ as an approximation of μ. The algorithm is summarized below.

Algorithm 1 Auxiliary particle filter (APF)

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

1: Draw $\{I_{N,i}\}_{i=1}^{\tilde{M}_N}\sim\mathscr{M}(\tilde{M}_N,\{\omega_{N,j}\psi_{N,j}/\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\}_{j=1}^{M_N})$,

2: simulate $\{\tilde{\xi}_{N,i}\}_{i=1}^{\tilde{M}_N}\sim\bigotimes_{i=1}^{\tilde{M}_N}R(\xi_{I_{N,i}},\cdot)$,

3: set $\tilde{\omega}_{N,i}\overset{\forall i}{\leftarrow}\psi_{N,I_{N,i}}^{-1}\mathrm{d}L(\xi_{I_{N,i}},\cdot)/\mathrm{d}R(\xi_{I_{N,i}},\cdot)(\tilde{\xi}_{N,i})$,

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

Note that setting, for all $1\le i\le M_N$, $\psi_{N,i}\equiv 1$ in Algorithm 1 yields the standard *bootstrap particle filter* presented in [8]. Note also that using the so-called *optimal adjustment weights* $\psi_{N,i}=\Psi^*(\xi_i)\triangleq L(\xi_i,\tilde{\Xi})$ and the *optimal kernel* $R(\xi,\cdot)=R^*(\xi,\cdot)\triangleq L(\xi,\cdot)/L(\xi,\tilde{\Xi})$ for every ξ leads to direct simulation from μ_N^{aux}. However as stated earlier these quantities are rarely available.

We may expect that the efficiency of Algorithm 1 depends highly on the choice of adjustment multiplier weights and proposal kernel. The former issue was treated by [5] who identified adjustment multiplier weights for which the increase of asymptotic variance at a single iteration of the algorithm is minimal. In this article we focus on the latter issue and discuss strategies for adaptive designing of the proposal kernel. Unlike [5], we base our methods on the results of the next section describing the asymptotic KLD and CSD between the target and importance distributions of the auxiliary SMC algorithm.

3. THEORETICAL RESULTS

From [4] we adopt the following definition.

Definition 3.1 *A weighted sample $\{(\xi_i,\omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be* consistent *for the probability measure μ and the set C if, for any $f\in\mathsf{C}$, as $N\to\infty$, $\Omega_N^{-1}\sum_{i=1}^{M_N}\omega_{N,i}f(\xi_i)\xrightarrow{\mathbb{P}}\int_\Xi f(\xi)\,\mu(\mathrm{d}\xi)$ and $\Omega_N^{-1}\max_{1\le i\le M_N}\omega_{N,i}\xrightarrow{\mathbb{P}}0$.*

We impose the following assumptions.

(A1) *The initial sample $\{(\xi_i,\omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν,C).*

(A2) *There exists a function $\Psi:\Xi\to\mathbb{R}^+$ such that $\psi_{N,i}=\Psi(\xi_i)$; moreover, $\Psi\in\mathsf{C}\cap\mathsf{L}^1(\Xi,\nu)$ and $L(\cdot,\tilde{\Xi})\in\mathsf{C}$.*

Under these assumptions we define the weight function $\Phi(\xi,\tilde{\xi})\triangleq\Psi^{-1}(\xi)\mathrm{d}L(\xi,\cdot)/\mathrm{d}R(\xi,\cdot)(\tilde{\xi})$, $(\xi,\tilde{\xi})\in\Xi\times\tilde{\Xi}$. From [5, Theorem 3.1] we obtain the following result, which describes how the consistency property is preserved through the auxiliary importance sampling operation.

Proposition 3.1 *Assume* **(A1, A2)**. *Then $\{(\tilde{\xi}_{N,i},\tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ is consistent for $(\nu,\{f\in\mathsf{L}^1(\tilde{\Xi},\mu),\int_{\tilde{\Xi}}|f(\tilde{\xi})|L(\cdot,\mathrm{d}\tilde{\xi})\in\mathsf{C}\})$.*

Let μ and ν be two probability measures on the same measurable space $(\Lambda,\mathscr{B}(\Lambda))$ such that μ is absolutely continuous with respect to ν. We then recall that the KLD and the CSD are given by $d_{\text{KL}}(\mu||\nu)\triangleq\int_\Lambda\log[\mathrm{d}\mu/\mathrm{d}\nu(\lambda)]\,\mu(\mathrm{d}\lambda)$ and $d_{\chi^2}(\mu||\nu)\triangleq\int_\Lambda[\mathrm{d}\mu/\mathrm{d}\nu(\lambda)-1]^2\,\nu(\mathrm{d}\lambda)$, respectively. We will use the following quantities to compute *empirical* estimates of the KLD and CSD between μ_N^{aux} and π_N^{aux}. Indeed, define

$$\mathscr{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N})\triangleq\tilde{\Omega}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}\log\left(\tilde{M}_N\tilde{\Omega}_N^{-1}\tilde{\omega}_{N,i}\right),$$

$$CV^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N})\triangleq\tilde{M}_N\tilde{\Omega}_N^{-2}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}^2-1,$$

where CV^2 is the square of the CV suggested by [10] as a means for detecting weight degeneracy; we then have the following result, which is the main result of this section and whose proof is available in [3].

Theorem 3.1 *Assume* **(A1, A2)**. *Then the following holds.*

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

$$\left|d_{\text{KL}}(\mu_N^{\text{aux}}||\pi_N^{\text{aux}})-\mathscr{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N})\right|\xrightarrow{\mathbb{P}}0,\quad\text{as } N\to\infty.$$

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

$$\left|d_{\chi^2}(\mu_N^{\text{aux}}||\pi_N^{\text{aux}})-CV^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N})\right|\xrightarrow{\mathbb{P}}0,\quad\text{as } N\to\infty.$$

Moreover, define the two probability measures $\mu^*(A)\triangleq\iint\nu(\mathrm{d}\xi)L(\xi,\mathrm{d}\xi')\mathbb{1}_A(\xi,\xi')/\iint\nu(\mathrm{d}\xi)L(\xi,\mathrm{d}\xi')$ and $\pi_\Psi^*(A)\triangleq\iint\nu(\mathrm{d}\xi)\Psi(\xi)R(\xi,\mathrm{d}\xi')\mathbb{1}_A(\xi,\xi')/\iint\nu(\mathrm{d}\xi)\Psi(\xi)R(\xi,\mathrm{d}\xi')$ on the product space $(\Xi\times\tilde{\Xi},\mathscr{B}(\Xi)\otimes\mathscr{B}(\tilde{\Xi}))$. As shown in the next corollary, the asymptotic KLD and CSD between the instrumental and target distributions of the particle filter can be characterised as the KLD and CSD between these distributions. In addition, it provides the adjustment multiplier weight function minimising, for a given proposal kernel R, the asymptotic KLD and CSD. Again, the proof is presented in [3].

Corollary 3.1 *Let the assumptions of Theorem 3.1 hold true. Then, as $N\to\infty$,*

$$d_{\text{KL}}(\mu_N^{\text{aux}}||\pi_N^{\text{aux}})\xrightarrow{\mathbb{P}}\eta_{\text{KL}}(\Psi)\triangleq d_{\text{KL}}(\mu^*\|\pi_\Psi^*)$$

and

$$d_{\chi^2}(\mu_N^{\text{aux}}||\pi_N^{\text{aux}})\xrightarrow{\mathbb{P}}\eta_{\chi^2}(\Psi)\triangleq d_{\chi^2}(\mu^*\|\pi_\Psi^*).$$

In addition, $\arg\min_\Psi\eta_{\text{KL}}(\Psi)=L(\xi,\tilde{\Xi})$, *and* $\arg\min_\Psi\eta_{\chi^2}(\Psi)=\Psi^*_{\chi^2,R}$, *where* $\Psi^*_{\chi^2,R}(\xi)\triangleq[\int\mathrm{d}L/\mathrm{d}R(\xi,\tilde{\xi})L(\xi,\mathrm{d}\tilde{\xi})]^{1/2}$.

Letting $R(\cdot,A)=L(\cdot,A)/L(\cdot,\tilde{\Xi})$ yields, as we may expect, a chi-square optimal adjustment multiplier weight function $\Psi^*_{\chi^2,R}(\cdot,\tilde{\Xi})=L(\cdot,\tilde{\Xi})$, which coincides with the Kullback-Leibler optimal one. In this case the importance weights are uniform, i.e. $\tilde{\omega}_{N,i}\equiv 1$.

4. ADAPTIVE IMPORTANCE SAMPLING

4.1 Adaptation by minimisation of estimated KLD and CSD

In the light of Theorem 3.1, a natural strategy for adaptive design of π_N^{aux} is to minimize the empirical estimate $\mathscr{E}$ (or CV^2) of the KLD (or CSD) under consideration over all proposal kernels belonging to some parametric family $\{R_\theta\}_{\theta\in\Theta}$. Thus, assume that there exists a random noise variable ε, having distribution λ on some measurable space $(\Lambda,\mathscr{B}(\Lambda))$, and a family $\{F_\theta\}_{\theta\in\Theta}$ of mappings from $\Xi\times\Lambda$ to $\tilde{\Xi}$ such that we are able to simulate $\tilde{\xi}\sim R_\theta(\xi,\cdot)$, for $\xi\in\Xi$, by simulating $\varepsilon\sim\lambda$ and letting $\tilde{\xi}=F_\theta(\xi,\varepsilon)$. We denote by Φ_θ the importance weight function associated with R_θ and set $\Phi_\theta\circ F_\theta(\xi,\varepsilon)\triangleq\Phi_\theta(\xi,F_\theta(\xi,\varepsilon))$.

In this setting, 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 $\{\varepsilon_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$. Now, keeping these indices and noise variables fixed, we can form an idea of how the KLD varies with θ by studying the function $\theta \mapsto \mathscr{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_{N,i}}, \varepsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$. Similarly, the CSD can be studied by using CV^2 instead of $\mathscr{E}$. This suggests an algorithm in which, as soon as the empirical KLD associated with the updated particle weights $\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}$ exceeds some given threshold κ, the particles are reproposed using R_{θ_*}, where $\theta_* = \arg\min_{\theta\in\Theta} \mathscr{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_{N,i}}, \varepsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$. The minimum θ_* exists if, e.g., the parameter space Θ is compact and the mapping $\theta \mapsto \Phi_\theta \circ F_\theta(\xi, \varepsilon)$ is continuous for all (ξ, ε), or when Θ is finite. The algorithm is summarized below, where we assume that the particles evolve according to R_{θ_0}, $\theta_0 \in \Theta$, when the adaptation operation is put on standby.

Algorithm 2 Adaptive APF

Require: (A1)

Draw $\{I_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \mathscr{M}(\tilde{M}_N, \{\omega_{N,j}\psi_{N,j}/\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\}_{j=1}^{M_N})$,
simulate $\{\varepsilon_{N,i}\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$,
if $\mathscr{E}(\{\Phi_{\theta_0} \circ F_{\theta_0}(\xi_{I_{N,i}}, \varepsilon_{N,i})\}_{i=1}^{\tilde{M}_N}) \geq \kappa$ **then**
    $\theta_* \leftarrow \arg\min_{\theta\in\Theta} \mathscr{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_{N,i}}, \varepsilon_{N,i})\}_{i=1}^{\tilde{M}_N})$,
else
    $\theta_* \leftarrow \theta_0$,
end if
set $\tilde{\xi}_{N,i} \overset{\forall i}{\leftarrow} F_{\theta_*}(\xi_{I_{N,i}}, \varepsilon_{N,i})$ and $\tilde{\omega}_{N,i} \overset{\forall i}{\leftarrow} \Phi_{\theta_*}(\xi_{I_{N,i}}, \tilde{\xi}_{N,i})$,
let $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ approximate μ.

4.2 APF adaptation by Cross-Entropy methods

Here again our aim is choose, from a parametric family, a proposal kernel which minimises the KLD between the target distribution μ_N^{aux} and the instrumental mixture distribution of the APF. Given an initial sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ approximating ν, we use importance sampling from an instrumental auxiliary distribution $\pi_{N,\theta}^{\text{aux}}$ to approximate the target auxiliary distribution μ_N^{aux}; here $\pi_{N,\theta}^{\text{aux}}$ is the straightforward modification of π_N^{aux} obtained by replacing R by R_θ, R_θ being a Markovian kernel from $(\Xi, \mathscr{B}(\Xi))$ to $(\tilde{\Xi}, \mathscr{B}(\tilde{\Xi}))$ belonging to the parametric family $\{R_\theta(\xi, \cdot) : \xi \in \Xi, \theta \in \Theta\}$.

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

$$d_{\text{KL}}(\mu_N^{\text{aux}} || \pi_{N,\theta}^{\text{aux}}) = \sum_{i=1}^{M_N} \int_{\tilde{\Xi}} \log\left(\frac{\mathrm{d}\mu_N^{\text{aux}}}{\mathrm{d}\pi_{N,\theta}^{\text{aux}}}(i, \tilde{\xi})\right) \mu_N^{\text{aux}}(i, \mathrm{d}\tilde{\xi}) \,. \quad (2)$$

In most cases, the expectation on the RHS of (2) is intractable. The main idea of the *cross-entropy* (CE) *method* (see [15]) is to approximate iteratively this expectation. Each iteration of the algorithm is split into two steps.

At iteration ℓ, denote by $\theta_N^\ell \in \Theta$ the current fit of the parameter. We then sample M_N^ℓ particles $\{(I_{N,i}^\ell, \tilde{\xi}_{N,i}^\ell)\}$ from $\pi_{N,\theta_N^\ell}^{\text{aux}}$, following Algorithm 1 with $\tilde{M}_N = \tilde{M}_N^\ell$ and $R = R_{\theta_N^\ell}$. Note that the adjustment multiplier weights are kept constant during the iterations, but this limitation may easily be removed. The second step consists in minimizing exactly the approximation

$$\theta_N^{\ell+1} \triangleq \arg\min_{\theta\in\Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_{N,i}^\ell}{\tilde{\Omega}_N^\ell} \log\left(\frac{\mathrm{d}\mu_N^{\text{aux}}}{\mathrm{d}\pi_{N,\theta}^{\text{aux}}}(I_{N,i}^\ell, \tilde{\xi}_{N,i}^\ell)\right) \quad (3)$$

of (2). In the case where the kernels L and R_θ, $\theta \in \Theta$, have densities, denoted l and r_θ, respectively, with respect to a common reference measure on $(\tilde{\Xi}, \mathscr{B}(\tilde{\Xi}))$, the minimization program (3) is equivalent to the following:

$$\theta_N^{\ell+1} \triangleq \arg\max_{\theta\in\Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_{N,i}^\ell}{\tilde{\Omega}_N^\ell} \log r_\theta(\xi_{I_{N,i}^\ell}, \tilde{\xi}_{N,i}^\ell) \,. \quad (4)$$

This algorithm is only helpful in situations where the minimization problem (3) is sufficiently simple to allow for maximization on closed form; 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 minimize numerically. This is in general the case when the function $r_\theta(\xi, \cdot)$ belongs to an exponential family for any $\xi \in \Xi$.

Since this optimization problem resembles the Monte Carlo EM algorithm, all details concerning implementation of these algorithms can be readily transposed to that context; see e.g. [13]. As seen in Section 5, convergence occurs, since we concider very simple models, within few iterations. The successive number of particles $\{\tilde{M}_N^\ell\}_{\ell=1}^L$ is also to be chosen, as a trade-off between precision of the approximation of (2) by (3) and computational cost. Numerical evidence typically shows that this number can be relatively small compared to the final size $\tilde{M}_N$, as precision is less crucial than when generating the final population from $\pi_{N,\theta_N^L}^{\text{aux}}$. Besides, it is possible (and even theoretically recommended) to increase the number of particles with the iterations, as, heuristically, high accuracy is less required in the first steps. In the current implementation in Section 5, we will show that fixing a priori the total number of iterations and using the same number of particles at each iteration $\tilde{M}_N^\ell = \tilde{M}_N/L$ can provide satisfactory results in a first run.

Algorithm 3 CE-based adaptive APF

Require: (A1)

Choose an arbitrary θ_N^0,
for $\ell = 0, \ldots, L-1$ **do**
    draw $\{I_{N,i}^\ell\}_{i=1}^{\tilde{M}_N^\ell} \sim \mathscr{M}(\tilde{M}_N^\ell, \{\omega_{N,j}\psi_{N,j}/\sum_{n=1}^{M_N}\omega_{N,n}\psi_{N,n}\}_{j=1}^{M_N})$
    simulate $\{\tilde{\xi}_{N,i}^\ell\}_{i=1}^{\tilde{M}_N^\ell} \sim \bigotimes_{i=1}^{\tilde{M}_N^\ell} R_{\theta_N^\ell}(\xi_{I_{N,i}^\ell}, \cdot)$,
    update $\tilde{\omega}_{N,i} \overset{\forall i}{\leftarrow} \Phi_{\theta_N^\ell}(\xi_{I_{N,i}^\ell}, \tilde{\xi}_{N,i}^\ell)$,
    compute, on closed form,

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

end for
run Algorithm 1 with $R = R_{\theta_N^L}$.

5. APPLICATION TO STATE SPACE MODELS

To illustrate our findings within the framework of *state space models*, we consider a first order (possibly nonlinear) autoregressive model observed in noise:

$$X_{k+1} = m(X_k) + \sigma_w(X_k)W_{k+1} \,,$$
$$Y_k = X_k + \sigma_v V_k \,,$$

where $\{W_k\}_{k=1}^\infty$ and $\{V_k\}_{k=0}^\infty$ are mutually independent sets of standard normal-distributed variables such that W_{k+1} is independent of $\{(X_i, Y_i)\}_{i=0}^k$ and V_k is independent of X_k and $\{(X_i, Y_i)\}_{i=0}^{k-1}$. In this setting, we wish to approximate the *filter distributions* $\phi_k(\cdot) \triangleq \mathbb{P}(X_k \in \cdot | Y_0, \ldots, Y_k)$, which in general lack closed form expressions, for all $k \geq 0$. By the *filtering recursion* it holds that

$$\phi_{k+1}(A) = \frac{\int_A \int_{\mathbb{R}} g(x_{k+1}, Y_{k+1})\, Q(x_k, \mathrm{d}x_{k+1})\, \phi_k(\mathrm{d}x_k)}{\int_{\mathbb{R}} \int_{\mathbb{R}} g(x_{k+1}, Y_{k+1})\, Q(x_k, \mathrm{d}x_{k+1})\, \phi_k(\mathrm{d}x_k)} \,, \quad (5)$$

where Q is the transition kernel of the *unobservable* chain $\{X_k\}_{k=0}^{\infty}$ and $g(x',\cdot)$ is the density of $\mathbb{P}(Y_k \in \cdot | X_k = x)$, that is, the distribution of the *observation* Y_k given the hidden state $X_k = x$. From (5) we conclude that this problem can, with $\nu = \phi_k$, $\mu = \phi_{k+1}$, and $L_k(x,A) = \int_A g(x',Y_{k+1})\,Q(x,\mathrm{d}x')$, be perfectly cast into the framework of Section 2, rendering sequential particle approximation of the filter measures possible. We will now consider a *fixed* timestep k, and thus drop the time index in the following.

For a model of this type, the optimal adjustment weight and the density of optimal kernel as defined in Section 2 can be expressed on closed form:

$$\Psi^*(x) = \mathcal{N}(Y_{k+1}; m(x), \sqrt{\sigma_w^2(x) + \sigma_v^2})\,, \tag{6}$$

$$r^*(x,x') = \mathcal{N}(x'; \tau(x,Y_{k+1}), \eta(x))\,, \tag{7}$$

where we have set $\mathcal{N}(x;\mu,\sigma) \triangleq \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, $\tau(x,Y_{k+1}) \triangleq [\sigma_w^2(x)Y_{k+1} + \sigma_v^2 m(x)]/[\sigma_w^2(x) + \sigma_v^2]$, and $\eta^2(x) \triangleq \sigma_w^2(x)\sigma_v^2/[\sigma_w^2(x) + \sigma_v^2]$. We may also compute the chi-square optimal adjustment multiplier weight function $\Psi^*_{\chi^2,Q}$ when the the prior kernel is used as proposal: at time k,

$$\Psi^*_{\chi^2,Q}(x) \propto \sqrt{\frac{2\sigma_v^2}{2\sigma_w^2(x) + \sigma_v^2}} \times \exp\left(-\frac{Y_{k+1}^2}{\sigma_v^2} + \frac{m(x)}{2\sigma_w^2(x) + \sigma_v^2}[2Y_{k+1} - m(x)]\right).$$

We recall from Corollary 3.1 that the optimal adjustment weight function for the KLD is given by $\Psi^*_{KL,Q}(x) = \Psi^*(x)$.

In this (deliberately chosen) simple example we will, at each timestep k, consider adaptation over the family $\{R_\theta(x,\cdot) \triangleq \mathcal{N}(\tau(x,Y_{k+1}), \theta\,\eta(x)) : x \in \mathbb{R}, \theta > 0\}$ of proposal kernels. The mode of the proposal kernel is equal to the mode of the optimal kernel, and the variance is proportional to the inverse of the Hessian of the optimal kernel at the mode. We denote by $r_\theta(x,x') \triangleq \mathcal{N}(x'; \tau(x,Y_{k+1}), \theta\,\eta(x))$ the density of $R_\theta(x,\cdot)$ with respect to the Lebesgue measure. In this setting, at every timestep k, the KLD between the target and proposal distributions is available on closed form:

$$d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_{N,\theta}^{\mathrm{aux}}) = \sum_{i=1}^{M_N} \frac{\omega_{N,i}\psi^*_{N,i}}{\sum_{j=1}^{M_N}\omega_{N,j}\psi^*_{N,j}} \times \left[\log\left(\frac{\psi^*_{N,i}\Omega_N}{\sum_{j=1}^{M_N}\omega_{N,j}\psi^*_{N,j}}\right) + \log\theta + \frac{1}{2}\left(\frac{1}{\theta^2} - 1\right)\right], \tag{8}$$

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

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

$$\theta_N^* \triangleq \arg\min_{\theta>0} d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_{N,\theta}^{\mathrm{aux}}) = 1\,, \tag{9}$$

which may also be inferred by straightforward derivation of (8) with respect to θ, and provides us with a reference to which the values found by our algorithm can be compared. Note that the proposal $\pi^{\mathrm{aux}}_{N,\theta_N^*}$ differs from the optimal instrumental distribution $\pi^{\mathrm{aux}}_{N,\theta^*_{\Psi^*}}$ by the adjustment weights used: the optimal proposal in the family considered actually uses uniform adjustment weights, $\Psi(x) = 1$, whereas as the overall optimal proposal uses optimal weights defined in (6) and is therefore equal to the target distribution μ_N^{aux}. This entails that

$$d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi^{\mathrm{aux}}_{N,\theta_N^*}) = \sum_{i=1}^{M_N} \frac{\omega_{N,i}\psi^*_{N,i}}{\sum_{j=1}^{M_N}\omega_{N,j}\psi^*_{N,j}} \log\left(\frac{\psi^*_{N,i}\Omega_N}{\sum_{j=1}^{M_N}\omega_{N,j}\psi^*_{N,j}}\right), \tag{10}$$

which is null if all the optimal weights are equal.

The implementation of Algorithm 3 is straightforward, as the optimization program (4) has the following closed form solution:

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

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

We will study the following special case of the model in question:

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

This is the classical *Gaussian autoregressive conditional heteroscedasticity* (ARCH) *model* observed in noise (see [1]). 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 [8],
(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 minimization of the empirical CSD and
(iv) the empirical KLD (Algorithm 2),
(v) an adaptive bootstrap filter using direct minimization of $d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_{N,\theta}^{\mathrm{aux}})$, see (9),
(vi) a CE-based adaptive bootstrap filter, and as a reference,
(vi) an optimal auxiliary particle filter, i.e. a filter using the optimal weights and proposal kernel defined in (6) and (7), respectively.

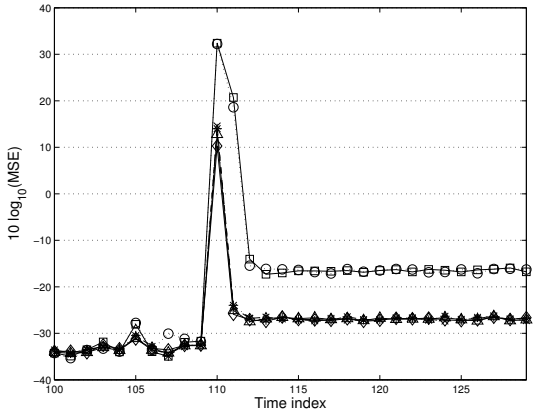
This experiment was conducted for parameters $(\beta_0, \beta_1, \sigma_v^2) = (1, 0.99, 10)$. This setting satisfies conditions upon which the ARCH(1) model is geometrically ergodic (which is $\beta_1 < 1$); the noise variance σ_v^2 is equal to $1/10$ of the stationary variance, which is equal to $\sigma_s^2 = \beta_0/(1-\beta_1)$, 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 chain, 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 of a significant likelihood, i.e., a large proportion of the total weight mass will be carried by a few particles only. We aim at illustrating that the adaptive algorithms, whose transition kernels has the same mode as the optimal transition kernel, adjust automatically their variance to the one of the latter and reach a performance comparable to that of the optimal auxiliary filter.

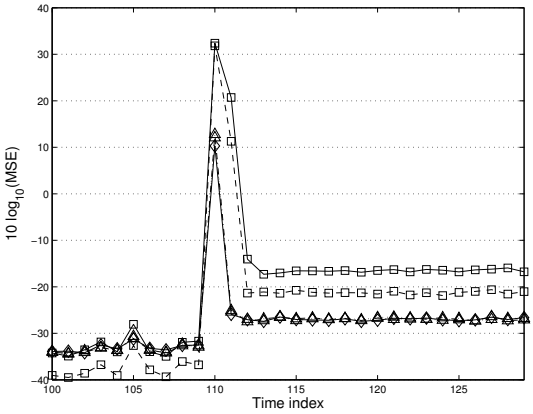
For these observation records, Figure 1 displays MSE estimates based on 500 filter means. Each filter used $5,000$ particles. The reference values used for the MSE estimates was obtained using the optimal auxiliary particle filter with as many as $500,000$ particles, which also provided a pool to initialize the filters at the stationary filtering distributions a few steps before the outlying observations.

The CE-based filter of Algorithm 3 was implemented in its simplest 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 $\{\theta_N^\ell\}_{l\geq 0}$ stabilized around the value reached in the very first step. We set $\theta_N^0 = 10$ to avoid initializing to the optimal value.

It can be seen in Figure 1(a) that using the CSD optimal weights combined the prior kernel as proposal do not improve on the plain bootstrap filter, precisely because the observations were chosen in such a way than the prior kernel was helpless. On the contrary, Figures 1(a) and 1(b) show that the adaptive schemes perform exactly



(a) Auxiliary filter based on optimal asymptotic CSD weights $\Psi^*_{\chi^2,Q}$ with prior kernel K ($\circ$), adaptive filters minimizing the empirical KLD ($*$) and CSD ($\times$), and reference filters listed below.



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

Figure 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 ($\square$), the optimal filter with weights Ψ^* and proposal kernel density r^* ($\diamond$), and a bootstrap using a proposal with parameter θ_N^* minimizing current KLD ($\triangle$ continuous line). The MSE values are computed using $N = 5,000$ particles—except for a reference bootstrap filter with $3N$ particles—and $1,000$ runs of each algorithm.

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 corresponding to the outlying constant observations, begininning at step 110. 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 stationary regime corresponding to 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 $\sum_{\ell=1}^{L} M_N^\ell / N = 1.5$—this difference is explained by the language used (Matlab), which benefits from the vectorization of the plain bootstrap filter and not from the iterative nature of the CE.

The conclusion from Figure 1(b) is that for an equal runtime, the adaptive filter outperforms, by a factor 3.5, the bootstrap filter using even 3 times more particles.

REFERENCES

[1] T. Bollerslev, R. F. Engle, and D. B. Nelson, "ARCH models", in *The Handbook of Econometrics* (R. F. Engle and D. MacFadden eds.), vol. 4, pp. 2959–3038, 1994.

[2] B. Chan, A. Doucet, A., and V. B. Tadic, "Optimization of Particle Filters Using Simultaneous Perturbation Stochastic Approximation", *Proc. IEEE ICASSP'03*, 2003.

[3] J. Cornebise, É. Moulines, and J. Olsson, "Adaptive Methods for Sequential Importance Sampling with Application to State Space Models", Technical report Lund University, 2008. URL `http://arxiv.org/abs/0803.0054`.

[4] R. Douc and É. Moulines, "Limit theorems for weighted samples with applications to sequential Monte Carlo methods", to appear in *Ann. Stat.*

[5] R. Douc, É. Moulines, and J. Olsson, "On the auxiliary particle filter", Technical report ENST, 2007. URL `http://arXiv.org/abs/0709.3448`.

[6] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte-Carlo Sampling Methods for Bayesian Filtering", S*tat. Comput*, vol. 10, 197–208, 2000.

[7] D. Fox, "Adapting the sample size in particle filters through KLD-sampling", *International Journal of Robotics Research*, vol. 22, pp. 985–1004, 2003.

[8] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to non-linear/non-Gaussian Bayesian state estimation", *IEEE Proc. Comm. Radar Signal Proc.*, vol. 140, pp. 107–113, 1993.

[9] M. Hürzeler and H. R. Künsch, "Monte Carlo approximations for general state space models", *J. Comput. Graph. Statist.*, vol. 7, pp. 175–193, 1998.

[10] A. Kong, J. S. Liu, and W. Wong, "Sequential imputation and Bayesian missing data problems", *J. Am. Statist. Assoc.*, vol. 89, pp. 278–288, 1994.

[11] H. R. Künsch, "Recursive Monte Carlo filters: algorithms and theoretical analysis", *Ann. Stat.*, vol. 33, pp. 1983–2021, 2005.

[12] F. Legland and N. Oudjane, "A sequential particle algorithm that keeps the particle system alive", Rapport de Recherche 5826, INRIA, 2006. URL `ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-5826.pdf`.

[13] R. A. Levine, and G. Casella, "Implementations of the Monte Carlo EM algorithm", *J. Comput. Graph. Statist.*, vol. 10, pp. 422–439, 2001.

[14] M. K. Pitt, and N. Shephard, "Filtering via simulation: Auxiliary particle filters", *J. Am. Statist. Assoc.*, vol. 87, pp. 493–499, 1999.

[15] R. Y. Rubinstein, and D. P. Kroese, *The Cross-Entropy Method*, Springer, 2004.