Approximation of hidden Markov models by mixtures of experts with application to particle filtering

Jimmy Olsson **Jonas Ströjby**
Division of Mathematical Statistics, Centre for Mathematical Sciences, Lund University,
Box 118, SE-22100 Lund, Sweden, e-mail: {jimmy, strojby}@maths.lth.se.

Abstract

Selecting conveniently the proposal kernel and the adjustment multiplier weights of the auxiliary particle filter may increase significantly the accuracy and computational efficiency of the method. However, in practice the optimal proposal kernel and multiplier weights are seldom known. In this paper we present a simulation-based method for constructing offline an approximation of these quantities that makes the filter close to fully adapted at a reasonable computational cost. The approximation is constructed as a mixture of experts optimised through an efficient stochastic approximation algorithm. The method is illustrated on two simulated examples.

1 Introduction

Sequential Monte Carlo (SMC) *methods* (see e.g. Doucet et al. (2001)) has emerged as a powerful tool for handling nonlinear filtering problems. The interest in these techniques has increased dramatically over recent years and several significant improvements of the plain *bootstrap particle filter* have been proposed. The perhaps most versatile of the SMC algorithms is the *auxiliary particle filter* (APF) introduced in Pitt and Shephard (1999) and analysed theoretically in Douc et al. (2008) and Johansen and Doucet (2008). The APF allows for more flexibility in the way the particles are evolved by introducing a set of *adjustment multiplier weights*. These weights are used at the resampling operation as a tool for eliminating/duplicating particles having *presumably* small/large importance weights at the subsequent mutation operation. In this way computational efficiency is gained. When the particles are mutated according to the so-called *optimal proposal kernel*, being the distribution of a next state conditional on the current state *as well as* the next observation, and the adjustment multipliers are proportional to the density of the next observation given the current state, the inherent instrumental and target distributions of the particle filter coincide and the filter is referred to as *fully adapted*.

Unfortunately, the optimal proposal kernel and adjustment multipliers are available on closed-form only for simple models, such as Gaussian models with linear measurement equation. Various approximations of the optimal proposal kernel has thus been suggested in the literature; see e.g. Doucet et al. (2000) a methods based on extended Kalman filters and Chan et al. (2003) for an approach similar in spirit to our approach. Cornebise et al. (2009) suggest to approximate the optimal proposal kernel at each step by a *mixture of experts*, giving a very well adapted proposal kernel at a reasonable computational cost. However, none of the mentioned works addresses the optimal adjustment multiplier weights which have seen limited interest in the literature.

Thus we focus on the joint transition density of the measurements and the hidden states, an idea that is also suggested in Johansen and Doucet (2008). The proposed method, which adopts some of the techniques proposed in Cornebise et al. (2009), produces a global parametric approximation of the joint transition density under the assumption that the hidden chain has a stationary distribution. An interesting feature of the algorithm is that it does not require any expression of the transition kernel of the hidden chain as long as its transitions can be simulated. Through the approximation of the joint transition density of the bivariate process provided by the algorithm, *both* the optimal adjustment multiplier weights and the optimal proposal kernel are readily available. The approximation is calculated offline and expressed as a state depen-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

dent mixture model, where the weights and the means depend on the previous state, similar to so-called *hierarchical mixture of experts*, see Jordan and Jacobs (1994). The proposal kernel and adjustment multiplier function obtained through marginalisation are therefore global functions, available for any value of the previous state and the measurement. Since the estimation is performed offline, the resulting APF incurs almost no additional computational cost. As a byproduct an approximation of the transition density of the latent process is also provided. This is relevant for models for which simulation from the transition density is possible but where the transition densities lack closed-form expressions. For many models used in practice, for instance most partially observed diffusion and Lévy processes, this is in fact the case. Since only simulation is required, the linearised dynamics simulated on a dense grid (such as the Euler scheme) suffices to be able to produce good density approximation.

The article is organized as follows. In Section 2 we introduce some notation and basic concepts. In Section 3 we discuss optimal filtering in general and describe briefly the APF. The optimal proposal kernel and the optimal adjustment multiplier weights are introduced. Our proposed algorithm is introduced in Section 4 and in the implementation part, Section 5, the algorithm is demonstrated on two examples.

2 Preliminaries

A *hidden Markov model* (HMM) is a stochastic model on two levels, where a non-observable Markov chain of the bottom level is partially observed through an observation sequence of the top level. More specifically, let Q be Markov kernel on some space $\mathsf{X} \subseteq \mathbb{R}^{d_X}$ equipped with the associated Borel σ-field $\mathcal{B}(\mathsf{X})$ and let G be a Markov kernel from X to some other state space $\mathsf{Y} \subseteq \mathbb{R}^{d_Y}$ equipped with the associated Borel σ-field $\mathcal{B}(\mathsf{Y})$. Now, define the Markov kernel

$$\mathbb{H}(x, y, A) \stackrel{\text{def}}{=} \iint_A Q(x, \mathrm{d}x')\, G(x', \mathrm{d}y')$$

on the product space $(\mathsf{X}\times\mathsf{Y}, \mathcal{B}(\mathsf{X})\otimes\mathcal{B}(\mathsf{Y}))$. For any initial distribution χ on $\mathcal{B}(\mathsf{X})$ we denote by $\mathbb{P}_\chi$ and $\mathbb{E}_\chi$ the probability distribution and associated expectation of the time homogenous Markov chain with initial distribution $\iint_A G(x, \mathrm{d}y)\, \chi(\mathrm{d}x)$ and transition kernel $\mathbb{H}$ on the canonical space $(\mathsf{X}\times\mathsf{Y})^{\mathbb{N}}$ equipped with the σ-field $(\mathcal{B}(\mathsf{X}) \otimes \mathcal{B}(\mathsf{Y}))^{\otimes\mathbb{N}}$. We denote by $Z \stackrel{\text{def}}{=} \{(X_k, Y_k)\}_{k\geq 0}$ the associated process, where $X \stackrel{\text{def}}{=} \{X_k\}_{k\geq 0}$ are the hidden states and $Y \stackrel{\text{def}}{=} \{Y_k\}_{k\geq 0}$ are the observations. As a consequence of our definition, the observed values of Y are, conditionally on the latent states X, independent with conditional distribution $Y_k|X \sim G(X_k, \cdot)$. We will throughout this paper assume that the Markov kernel Q is ϕ-irreducible, positive recurrent, and strongly aperiodic, and we denote by π its stationary distribution. Set $\bar{\mathbb{P}} \stackrel{\text{def}}{=} \mathbb{P}_\pi$ and $\bar{\mathbb{E}} \stackrel{\text{def}}{=} \mathbb{E}_\pi$. It is easily seen that under $\bar{\mathbb{P}}$, also the bivariate process Z is stationary with stationary distribution $\bar{\pi}(A) \stackrel{\text{def}}{=} \iint_A G(x, \mathrm{d}y)\, \pi(\mathrm{d}x)$.

Throughout this paper we will assume that the measures $Q(x, \cdot)$ and $G(x, \cdot)$ are, for any $x \in \mathsf{X}$, absolutely continuous with respect to the Lebesgue measure λ, and we denote, respectively, by q and g the corresponding densities. Under this assumption, each measure $\mathbb{H}(x, y, \cdot)$ has, for any $(x, y) \in \mathsf{X} \times \mathsf{Y}$, a density function with respect to the Lebesgue measure as well, and we denote this density by $p(\cdot|x, y)$. For any set $A \in \mathcal{B}(\mathsf{X}) \otimes \mathcal{B}(\mathsf{Y})$, the function $(x, y) \mapsto p(A|x, y)$ does not depend on y, and the restriction, which we denote by the same symbol, of this mapping to X is thus well defined. As described in the introduction, the aim of the present paper is to approximate the transition density p by a mixture of experts under the assumption that the bivariate process Z can be simulated. Denote by

$$\begin{aligned} p(y_{k+1}|x_k) &\stackrel{\text{def}}{=} \int p(x_{k+1}, y_{k+1}|x_k)\, \lambda(\mathrm{d}x_{k+1}) \,, \\ p(x_{k+1}|x_k, y_{k+1}) &\stackrel{\text{def}}{=} p(x_{k+1}, y_{k+1}|x_k)/p(y_{k+1}|x_k) \end{aligned} \tag{2.1}$$

the densities of the conditional distribution of Y_{k+1} given X_k and the conditional distribution of X_{k+1} given X_k as well as Y_{k+1}, respectively. The latter distribution is usually referred to as the *optimal kernel*.

Finally, as a measure of closeness of two distributions we use the *Kullback-Leibler divergence* (KLD): Let $(\mathsf{Z}, \mathcal{B}(\mathsf{Z}))$ be some state space and let μ_1 and μ_2 be two probability measures on $\mathcal{B}(\mathsf{Z})$ such that $\mu_1 \ll \mu_2$; then the KLD $d_{\mathrm{KL}}(\mu_1\|\mu_2)$ between μ_1 and μ_2 is defined by

$$d_{\mathrm{KL}}(\mu_1\|\mu_2) \stackrel{\text{def}}{=} \int \log \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(z)\, \mu_1(\mathrm{d}z) \,.$$

Other measures, such as the χ^2-*distance*, of closeness are of course possible, but the KLD turns out to be very convenient in conjunction with the exponential families used in this paper, since this makes it possible to optimise most parameters on closed-form; see Section 4.

3 Particle filters

3.1 Optimal filtering

To motivate why approximation of the density p is important we discuss the use of particle filters for filter-

ing in HMMs. Let $Y_{0:n} = (Y_0, Y_1, \ldots, Y_n)$ be a given record (similar vector notation will be used also for other quantites) of observations. Then the *filtering distribution* at time n is defined by the conditional probability

$$\phi_n(A) \stackrel{\text{def}}{=} \mathbb{P}_\chi(X_n \in A | Y_{0:n}) = \frac{\int \mathbb{1}_A(x_n) \prod_{k=0}^{n-1} Q(x_k, \mathrm{d}x_{k+1}) g(x_{k+1}, Y_{k+1})\, g(x_0, Y_0)\, \chi(\mathrm{d}x_0)}{\int \prod_{k=0}^{n-1} Q(x'_k, \mathrm{d}x'_{k+1}) g(x'_{k+1}, Y_{k+1})\, g(x'_0, Y_0)\, \chi(\mathrm{d}x'_0)}$$

for A belonging to $\mathcal{B}(\mathsf{X})$. Computing the filtering distribution is essential when estimating the hidden states or performing any inference on unknown model parameters. Under the assumptions above, ϕ_n has a well defined density with respect to the Lebesgue measure. By inspecting the definition above, we conclude that the flow $\{\phi_n\}_{n \geq 0}$ of filter distributions can be expressed recursively according to

$$\phi_{n+1}(A) = \frac{\iint \mathbb{1}_A(x_{n+1}) Q(x_n, \mathrm{d}x_{n+1}) g(x_{n+1}, Y_{n+1})\, \phi_n(\mathrm{d}x_n)}{\iint Q(x'_n, \mathrm{d}x'_{n+1}) g(x'_{n+1}, Y_{n+1})\, \phi_n(\mathrm{d}x'_n)} = \frac{\int p(Y_{n+1}|x_n) \int_A p(x_{n+1}|x_n, Y_{n+1})\, \lambda(\mathrm{d}x_{n+1})\, \phi_n(\mathrm{d}x_n)}{\int p(Y_{n+1}|x'_n)\, \phi_n(\mathrm{d}x'_n)}, \tag{3.1}$$

where the densities $p(y_{n+1}|x_n)$ and $p(x_{n+1}|x_n, y_{n+1})$ are defined in (2.1). Equation (3.1), usually referred to as the *filtering recursion*, is however only deceptively simple since closed-form solutions to this recursion can be obtain in only in two cases, i.e., when the HMM is linear/Gaussian or when X is a finite set. In the general case we are thus referred to simulation-based techniques producing Monte Carlo approximations of these posterior distributions.

3.2 Particle filters

Assume that we have at hand a sample $\{(\xi_n^i, \omega_n^i)\}_{i=1}^N$ of *particles* $\{\xi_n^i\}_{i=1}^N$ and associated importance weights $\{\omega_n^i\}_{i=1}^N$ targeting the measure ϕ_n in the sense that

$$\sum_{i=1}^N \frac{\omega_n^i}{\sum_{\ell=1}^N \omega_n^\ell} f(\xi_n^i) \approx \int_{\mathsf{X}} f(x)\, \phi_n(\mathrm{d}x)\ . \tag{3.2}$$

for a large class of target functions f on X. We wish to transform $\{(\xi_n^i, \omega_n^i)\}_{i=1}^N$ into a new weighted particle sample $\{(\xi_{n+1}^i, \omega_{n+1}^i)\}_{i=1}^N$ approximating the filter ϕ_{n+1} at the next time step. We hence plug the particle approximation (3.2) into the filtering recursion (3.1), yielding the approximation

$$\phi_{n+1}^N(A) \stackrel{\text{def}}{=} \sum_{i=1}^N \frac{\omega_n^i p(Y_{n+1}|\xi_n^i)}{\sum_{\ell=1}^N \omega_n^\ell p(Y_{n+1}|\xi_n^\ell)} \int_A p(x_{n+1}|Y_{n+1}, \xi_n^i)$$

of $\phi_{n+1}(A)$. Here $p(Y_{n+1}|\xi_n^i) \stackrel{\text{def}}{=} p(Y_{n+1}|x_n)|_{x_n = \xi_n^i}$ and similarly for $p(x_{n+1}|Y_{n+1}, \xi_n^i)$. Note that ϕ_{n+1}^N has a well defined density, which we denote by the same symbol, with respect to the Lebesgue measure; this density is proportional to the function

$$x_{n+1} \mapsto \sum_{i=1}^N \omega_n^i q(\xi_n^i, x_{n+1}) g(x_{n+1}, Y_{n+1})\ .$$

Simulating N draws from the mixture ϕ_{n+1}^N would yield the desired set of particles. However, simulating from ϕ_{n+1}^N is in general not easily performed and requires expensive accept-reject techniques. Thus we apply instead importance sampling using the instrumental distribution

$$\pi_{n+1}^N(A) \stackrel{\text{def}}{=} \sum_{i=1}^N \frac{\omega_n^i \psi_n^i}{\sum_{\ell=1}^N \omega_n^\ell \psi_n^\ell} R_n(\xi_n^i, A)\ , \tag{3.3}$$

where A is in $\mathcal{B}(\mathsf{X})$, and associate each drawn particle ξ_{n+1}^i with an importance weight ω_{n+1}^i proportional to $\phi_{n+1}^N(x_{n+1}) / \pi_{n+1}^N(x_{n+1})|_{x_{n+1} = \xi_{n+1}^i}$. Here $\{\psi_n^i\}_{i=1}^N$ are nonnegative numbers referred to as *adjustment multiplier weights* and R_n is a Markovian kernel having transition density r_n with respect to the Lebesgue measure. Hence the density

$$\pi_{n+1}^N(x_{n+1}) \propto \sum_{i=1}^N \omega_n^i \psi_n^i r_n(\xi_n^i, x_{n+1})$$

is well defined. It is assumed that R_n dominates the optimal kernel. The instrumental kernel R_n as well as the adjustment multipliers may depend on the new observation Y_{k+1}, and ideally one would take $r_n(x_n, x_{n+1}) \equiv p(x_{n+1}|Y_{n+1}, x_n)$ and $\psi_n^i = p(Y_{n+1}|\xi_n^i)$ for all i, in which case the target distribution ϕ_{n+1}^N and the instrumental distribution π_{n+1}^N of the particle filter *coincide*. In this case the particle filter is usually referred to as *fully adapted*. This shows clearly the importance of being able to approximate p (and thus the marginals (2.1)) with good precision. A problem with the described approach is that computing any weight ω_{n+1}^i requires the evaluation of a sum of N terms, yielding an algorithm of $\mathcal{O}(N^2)$ computational complexity. To cope with this problem, we follow Pitt and Shephard (1999) and introduce an auxiliary variable I corresponding to the selected mixture component and target instead the auxiliary distribution

$$\bar{\phi}_{n+1}^N(\{i\} \times A) \stackrel{\text{def}}{=} \frac{\omega_n^i p(Y_{n+1}|\xi_n^i)}{\sum_{\ell=1}^N \omega_n^\ell p(Y_{n+1}|\xi_n^\ell)} \int_A p(x_{n+1}|Y_{n+1}, \xi_n^i)$$

on the product space $\{1, \ldots N\} \times \mathsf{X}$. The auxiliary distribution $\bar{\phi}_{n+1}^N$ has a density function proportional to $\omega_n^i q(\xi_n^i, x_{n+1}) g(x_{n+1}, Y_{n+1})$. Note that $\bar{\phi}_{n+1}^N$

has ϕ_{n+1}^N as marginal distribution, i.e., $\phi_{n+1}^N(A) = \sum_{i=1}^N \bar{\phi}_{n+1}^N(\{i\} \times A)$. As auxiliary instrumental distribution we take

$$\bar{\pi}_{n+1}^N(\{i\} \times A) \stackrel{\text{def}}{=} \frac{\omega^i \psi_k^i}{\sum_{\ell=1}^N \omega_n^\ell \psi_n^\ell} R_n(\xi_n^i, A)$$

and sample pairs $\{(I_{n+1}^i, \xi_{n+1}^i)\}_{i=1}^N$ of indices and particle positions from $\bar{\pi}_{n+1}^N$ and assign each draw the importance weight

$$\omega_{n+1}^i \stackrel{\text{def}}{=} \frac{q(\xi_n^{I_{n+1}^i}, \xi_{n+1}^i) g(\xi_{n+1}^i, Y_{n+1})}{\psi_n^{I_{n+1}^i} r_n(\xi_n^{I_{n+1}^i}, \xi_{n+1}^i)} . \tag{3.4}$$

Finally we take $\{(\xi_{n+1}^i, \omega_{n+1}^i)\}_{i=1}^N$ as an approximation of ϕ_{n+1} and discard the indices. Note that no sum appears in the expression (3.4) of the importance weights; by introducing the auxiliary variable we have hence, satisfactorily, obtained an algorithm of linear (in the number of particles) computational cost. Obviously, the efficiency of the algorithm depends heavily on how we choose the adjustment multiplier weights and the proposal kernel, and in the next section we discuss how to obtain a close to fully adapted particle filter by means of approximating offline the transition density p (and thus all its marginals) by a mixture of experts.

4 Approximation of $\mathbb{H}$ by a mixture of experts

We turn to the problem of approximating the optimal proposal kernel and the optimal adjustment multiplier weights given by (2.1). Naturally, we propose to construct a parametric approximation of these via the joint distribution p of the state and measurements. Inspired by Cornebise et al. (2009) we approximate this joint density by a mixture of experts, which is fitted to simulated data by means of an online EM algorithm. More specifically, we represent the density $p(z|x)$ by a mixture of form

$$p_\theta(z|x) = \sum_{j=1}^M \alpha_j(x;\gamma) \rho(z; x, \lambda_j) , \tag{4.1}$$

where $\gamma = (\gamma_1^T, \ldots, \gamma_M^T)^T$ and $\theta \stackrel{\text{def}}{=} (\gamma^T, \lambda_1^T, \ldots, \lambda_M^T)^T$ is a vector of parameters, $\{\alpha_j\}_{j=1}^M$ are weighting functions, and each $\rho(\cdot, \lambda_j)$ is Markovian transition kernel from X to $\mathsf{X} \times \mathsf{Y}$. We denote by $\boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$ the set of possible parameters. The weighting functions $\{\alpha_j\}_{j=1}^M$ are required to sum to one to ensure that (4.1) is a density. For the theoretical exposition, assume that the weighting functions are constant, i.e., $\alpha_j(x;\gamma) \equiv \alpha_j$ with $\sum_{j=1}^M \alpha_j = 1$. In this case the model is often referred to as a *mixture of regressions*. More general weighting functions will be considered later on.

Assumption 4.1. *The mixture kernels $\rho(\cdot, \lambda_j)$ are of form*

$$\rho(z; x, \lambda_j) = h(x, z) \exp\left(-\psi(\lambda_j) + \langle U(\bar{x}, z), \phi(\lambda_j)\rangle\right) , \tag{4.2}$$

where U is a vector of sufficient statistics that do not depend on any parameters, $\langle \cdot, \cdot \rangle$ denotes the scalar product, $\bar{x}$ the extended vector $(1, x)^T$, ψ and ϕ are functions of the parameters only, and h is a function independent of the parameters.

In the implementation part (Section 5) we will make use of the Gaussian densities with mean $\beta_j \bar{x}$ and covariance matrix $\Sigma_j = (\Sigma_j^1, \ldots, \Sigma_j^{d_Z})$ and denote jointly $\lambda_j = (\beta_j^T, (\Sigma_j^1)^T, \ldots, (\Sigma_j^{d_Z})^T)^T$. In addition, $U(\bar{x}, z) = (1, zz^T, \bar{x}\bar{x}^T, z\bar{x}^T)$ in the Gaussian case. It is however worth to notice that all results obtained in this paper hold for the more general family of *integrated curved exponentials* such as student's t-distribution.

In order to be able to perform quick optimisation we augment the state space with the index J of the mixture component, resulting in the auxiliary density

$$\bar{p}_\theta(z, j|x) = \alpha_j(x;\gamma) \rho(z; x, \lambda_j) . \tag{4.3}$$

Also we define the conditional mixture weights, or *responsibilities*, as

$$\bar{p}_\theta(j|x, z) = \frac{\alpha_j(x;\gamma) \rho(z; x, \lambda_j)}{\sum_{i=1}^M \alpha_i(x;\gamma) \rho(z; x, \lambda_i)} . \tag{4.4}$$

Now, let family $\mathcal{H}$ be a family of Markovian kernels from X to $\mathsf{X} \times \mathsf{Y}$ where each $\tilde{\mathbb{H}} \in \mathcal{H}$ is such that $\tilde{\mathbb{H}}(x, \cdot)$ dominates $\mathbb{H}(x, \cdot)$ for all $x \in \mathsf{X}$. We say that a kernel $\tilde{\mathbb{H}}^*$ belonging to $\mathcal{H}$ is $\mathcal{H}$-*optimal* if it holds that

$$\begin{aligned} \tilde{\mathbb{H}}^* &= \arg\min_{\tilde{\mathbb{H}} \in \mathcal{H}} \bar{\mathbb{E}} \left[d_{\mathrm{KL}}(\mathbb{H}(X_0, \cdot) \| \tilde{\mathbb{H}}(X_0, \cdot)) \right] \\ &= \arg\min_{\tilde{\mathbb{H}} \in \mathcal{H}} \iint \log \left(\frac{\mathrm{d}\mathbb{H}(x, \cdot)}{\mathrm{d}\tilde{\mathbb{H}}(x, \cdot)}(z) \right) \mathbb{H}(x, \mathrm{d}z) \, \pi(\mathrm{d}x) , \end{aligned} \tag{4.5}$$

i.e. the expected value of the KLD under the stationary distribution of the hidden Markov chain. Now assume that each kernel $\tilde{\mathbb{H}}$ in $\mathcal{H}$ has a transition density h with respect to the Lebesgue measure. Then

$$\begin{aligned} &\arg\min_{\tilde{\mathbb{H}} \in \mathcal{H}} \bar{\mathbb{E}} \left[d_{\mathrm{KL}}(\mathbb{H}(X_0, \cdot) \| \tilde{\mathbb{H}}(X_0, \cdot)) \right] \\ &\quad = \arg\min_{\tilde{\mathbb{H}} \in \mathcal{H}} \iint \log \frac{p(z|x)}{h(z|x)} \, \mathbb{H}(x, \mathrm{d}z) \, \pi(\mathrm{d}x) \\ &\quad = \arg\min_{\tilde{\mathbb{H}} \in \mathcal{H}} \left\{ \iint \log p(z|x) \, \mathbb{H}(x, \mathrm{d}z) \, \pi(\mathrm{d}x) \right. \\ &\qquad \left. - \iint \log h(z|x) \, \mathbb{H}(x, \mathrm{d}z) \, \pi(\mathrm{d}x) \right\} , \end{aligned}$$

where the first term on the right hand side does not depend on h. In the following we let $\mathcal{H}$ be the family of mixtures p_θ of form (4.1), and thus the optimisation problem (4.5) can be alternatively expressed as

$$\arg\max_{\theta\in\boldsymbol{\Theta}} \iint \log p_\theta(z|x)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\;. \tag{4.6}$$

Calculating exactly expectations under the measures π and $\mathbb{H}$ is in general not possible, since the stationary distribution is not known on closed-form. In the following we discuss how the intricate maximisation problem (4.6) can be handled within the framework of *missing data problems* by means of stochastic approximation methods. Thus, in the following we assume that the function

$$\begin{aligned}\mathcal{Q}(\theta;\theta^\ell) &\stackrel{\mathrm{def}}{=} \int\sum_{j=1}^M \log\bar{p}_\theta(z,j|x)\bar{p}_{\theta^\ell}(j|x,z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\\ &\cong \sum_{j=1}^M \log\alpha_j \int \bar{p}_{\theta^\ell}(j|x,z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\\ &\quad -\sum_{j=1}^M \psi(\lambda_j)\int \bar{p}_{\theta^\ell}(j|x,z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\\ &\quad +\sum_{j=1}^M \left\langle \int \bar{p}_{\theta^\ell}(j|x,z)U(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x), \phi(\lambda_j)\right\rangle\;,\end{aligned}$$

where $\cong$ means equality up constant that is independent of the parameter θ, has a unique global maximum over $\boldsymbol{\Theta}$ for any value of the sufficient statistics

$$s_{i,j}(\theta^\ell) \stackrel{\mathrm{def}}{=} \int \bar{p}_{\theta^\ell}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\;,$$

where u_i denotes the ith submatrix of U. In addition, we set $s_{0,j}(\theta^\ell) \stackrel{\mathrm{def}}{=} \int \bar{p}_{\theta^\ell}(j|x,z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)$. We collect all these statistics in a structure which we denote by $s(\theta^\ell)$ and denote this maximum by $\bar{\theta}(s)$. In the Gaussian case the maxima are given by $\alpha_j = s_{1,j}(\theta^\ell)$, $\beta_j = s_{4,j}(\theta^\ell)s_{3,j}^{-1}(\theta^\ell)$ and

$$\Sigma_j = \frac{s_{2,j}(\theta^\ell) - s_{4,j}(\theta^\ell)s_{3,j}^{-1}(\theta^\ell)s_{4,j}^T(\theta^\ell)}{s_{1,j}(\theta^\ell)}\;.$$

The following result is instrumental for the method we use for solving (4.6). In order to keep the arguments lucid we skip some of the technical details, and refer the interested reader to a companion paper. Define the so-called *mean field* $s\mapsto H(s)$ as a structure containing all the mappings

$$H_{i,j}(s) \stackrel{\mathrm{def}}{=} \int \bar{p}_{\bar{\theta}(s)}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x) - s \tag{4.7}$$

on the space of all possible values of the sufficient statistics. We then have the next result.

Proposition 4.1. *Under weak assumptions the follwong holds. If s^* is a root of the mean field H in the sense that $H(s^*)=0$, then $\theta^* = \bar{\theta}(s^*)$ satisfies $\nabla_\theta\bar{\mathbb{E}}[d_{\mathrm{KL}}(\mathbb{H}(X_0,\cdot)\|\mathbb{H}_\theta(X_0,\cdot))]|_{\theta=\theta^*}=0$. Conversely, if θ^* is a stationary point in the same sense, then the structure s^* contaning all*

$$\begin{aligned}s^*_{0,j} &\stackrel{\mathrm{def}}{=} \int \bar{p}_{\theta^*}(j|x,z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\;,\\ s^*_{i,j} &\stackrel{\mathrm{def}}{=} \int \bar{p}_{\theta^*}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\end{aligned}$$

is a root of H.

The proof follows the lines of the proof of Proposition 1 in Cappe and Moulines (2009). The dual problem of (4.6) is thus to find a root of the mean field H, a task that is well suited for the classical *Robbins-Monroe* stochastic approximation procedure

$$\hat{s}_{\ell+1} = \hat{s}_\ell + \gamma_{\ell+1}(H(\hat{s}_\ell) + \xi_{\ell+1})\;,$$

where $\{\gamma_\ell\}_{\ell\geq 1}$ is a decreasing sequence such that

$$\lim_{\ell\to\infty}\gamma_\ell = 0\;,\qquad \sum_{\ell=1}^\infty \gamma_\ell = \infty\;, \tag{4.8}$$

and $\{\xi_\ell\}_{\ell\geq 0}$ is a sequence of Markovian stochastic perturbations; the sum $H(\hat{s}_\ell)+\xi_{\ell+1}$ can thus be viewed as a noisy observation of $H(\hat{s}_\ell)$. In order to obtain such noisy observations, we use that $Q^\ell(x_0,\cdot)$ approaches π as ℓ increases under rather weak conditions, e.g. that the latent chain X is *Harris recurrent*. The convergence holds in general for any initial distribution χ. This yields the approximation

$$\begin{aligned}&\int \bar{p}_{\bar{\theta}(s)}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\pi(\mathrm{d}x)\\ &\quad\approx \int \bar{p}_{\bar{\theta}(s)}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\chi Q^\ell(\mathrm{d}x)\;.\end{aligned} \tag{4.9}$$

In order to approximate the right integral, assume that we are given sets of independent draws $\{X_\ell^k\}_{k=1}^K$ and $\{Z_{\ell+1}^k\}_{k=1}^K$ where $X_\ell^k\sim\chi Q^\ell$ and $Z_{\ell+1}^k\sim\mathbb{H}(X_\ell^k,\cdot)$. We then form the Monte Carlo estimate

$$\begin{aligned}&\int \bar{p}_{\bar{\theta}(s)}(j|x,z)u_i(\bar{x},z)\,\mathbb{H}(x,\mathrm{d}z)\,\chi Q^\ell(\mathrm{d}x)\\ &\qquad\approx \frac{1}{K}\sum_{k=1}^K \bar{p}_{\bar{\theta}(s)}(j|X_\ell^k,Z_{\ell+1}^k)u_i(\bar{X}_\ell^k,Z_{\ell+1}^k)\;.\end{aligned}$$

The proposed method thus involves the simulation of K latent chains X^k evolving independently. In addition, the $Z_{\ell+1}^k$'s are simulated independently on these chains. In this setting, each member of noise sequence

$\{\xi_\ell\}_{\ell\geq 1}$ contains elements of form

$$\xi_\ell^{i,j} \stackrel{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K} \bar{p}_{\bar{\theta}(\hat{s}_{\ell-1})}(j|X_\ell^k, Z_{\ell+1}^k) u_i(\bar{X}_\ell^k, Z_{\ell+1}^k) - \int \bar{p}_{\bar{\theta}(\hat{s}_{\ell-1})}(j|x,z) u_i(\bar{x},z)\, \mathbb{H}(x,\mathrm{d}z)\, \pi(\mathrm{d}x) \ .$$

Finally, this gives us the algorithm

$$\begin{aligned} &\hat{s}_{\ell+1}^{i,j} = \hat{s}_\ell^{i,j} + \\ &\gamma_{\ell+1}\left(\frac{1}{K}\sum_{k=1}^{K} \bar{p}_{\hat{\theta}_\ell}(j|X_\ell^k, Z_{\ell+1}^k) u_i(\bar{X}_\ell, Z_{\ell+1}^k) - \hat{s}_\ell^{i,j}\right) \ , \\ &\hat{\theta}_{\ell+1} = \bar{\theta}(\hat{s}_{\ell+1}) \ , \end{aligned} \tag{4.10}$$

where $\hat{s}_\ell$ contains all the $\hat{s}_\ell^{i,j}$'s.

In order to discuss the convergence issues of the scheme we assume that the chain X mixes geometrically fast, i.e.$\|\chi Q^n - \pi\|_{\mathrm{TV}} \leq C\rho^n$ for any initial distribution χ and constants $C < \infty$ and $0 < \rho < 1$. Such a geometrical forgetting property is satisfied by a large class of models and can be verified by checking the so-called *Foster-Lyapunov drift condition*, see i.e. Cappé et al. (2005). A detailed study of the asymptotic properties of the algorithm (4.10) is beyond the scope of this note and can be found in a forthcoming paper; however, the main steps consist of (1) showing that $s \mapsto w(s) \stackrel{\text{def}}{=} \bar{\mathbb{E}}[d_{\mathrm{KL}}(\mathbb{H}(X_0,\cdot)\|\mathbb{H}_{\bar{\theta}(s)}(X_0,\cdot))]$ is a *Lyapunov* function for the mean field H, i.e.$\langle \nabla_s w(s), H(s)\rangle \leq 0$ with equality if and only if $H(s) = 0$, and (2) establishing that $\limsup_n \sup_{k\geq n} |\sum_{\ell=n}^k \gamma_\ell \xi_\ell|$ vanishes almost surely. Since the perturbations are Markovian, theory presented e.g. Duflo (1997), chapter 9.2.3 can be employed in order to show this. In practice logistic weights $\alpha(x;\gamma)$ are used. In this case the normalized weights are convex in the parameters and thus easily optimized. The proof will be analogous but more involved.

5 Simulation study

We illustrate our method on two simulated examples. In this part we consider the more general framwork of *logistic* weighting functions $\alpha_j(x;\gamma) = \exp(\gamma_j^T \bar{x})/\sum_{i=1}^M \exp(\gamma_i^T \bar{x})$. In this case the optimum $\bar{\theta}(\hat{s}_\ell)$ cannot be found analytically and it is thus necessary to apply some convenient optimisation procedure. We omit the details for brevity.

Example 1. For a first order (possibly nonlinear) autoregressive model

$$\begin{aligned} X_{k+1} &= m(X_k) + \sigma_w(X_k) W_{k+1} \ , \\ Y_k &= X_k + \sigma_v V_k \ , \end{aligned} \tag{5.1}$$

the optimal proposal kernel and the multiplier adjustment weighting function are obtainable on closed-form, which makes the model well suited for an initial assessment of our algorithm. As a special case of (5.1), we consider here the well known *ARCH model* observed in noise:

$$\begin{aligned} X_{k+1} &= \sqrt{\beta_0 + \beta_1 X_k} W_{k+1} \ , \\ Y_k &= X_k + \sigma_v V_k \ , \end{aligned}$$

where $\beta_0 = 1$, $\beta_1 = 0.5$, and $\sigma_v = 0.25$. In this setting, we estimate $\mathbb{H}$ using $M = 9$ mixture components. We let the learning loop (4.10) run for $n = 20,000$ iterations and use $K = 10$ realizations of the latent chain. The Robbins-Monroe sequence $\gamma_\ell = 1/\ell^{0.6}$ is used, and the gradient descent stepsize is $\delta = 0.01$. Both the parameters and the chain are started at random values and we use a 20 step burn-in phase before starting to update the parameters. In Figure 1, the approximated weighting function is shown and compared to the exact one for fixed observations $(-2.9653, -0.3891, 0.3703, 3.2077)$ obtained by selecting the center points of each quartile from a simulated sample comprising $1,000$ values. In these graphs it can be seen that the approximated weight functions follow rather closely the exact ones, especially in the support of the stationary distribution. As a second experiment, filtering of the ARCH process is performed using the APF based on approximated as well as exact optimal proposal kernels and importance weight functions. The outcome is compared to that of the vanilla bootstrap filter. The study is performed for 200 observations using 500 particles. In Figure 2 the cumulative sums of the sorted normalised weights are displayed, each line representing one of the 200 time-steps; (a) displays the weight distribution of the bootstrap filter while (b) is the distribution of the APF based on the mixture approximation. From this figure it is evident that the algorithm provides a close to fully adapted filter that drastically outperforms the bootstrap filter. For the fully adapted optimal filter, the distribution is of course always a straight line, indicating uniform weights at all time-steps. Finally, note that despite the fixed variances in the mixture components, a very efficient approximation of a stochastic volatility models such as ARCH may be constructed.

Example 2. In this example we consider a vector valued autoregressive model with nonlinear measurement equation:

$$\begin{aligned} X_{k+1} &= A_0 + A_1 X_k + \Sigma_w W_{k+1} \ , \\ Y_k &= \begin{pmatrix} |X_k^{(1)}| \\ |X_k^{(2)}| \end{pmatrix} + \Sigma_v V_k \ , \end{aligned}$$

where $A_0 = (0,0)^T$, $A_1 = ((0.5,0)^T, (0,-0.5)^T)$, and $(\Sigma_w, \Sigma_v) = (1, 0.25)$. In this case we estimate the den-

sity of $\mathbb{H}$ using $M = 12$ mixture components. As in the previous example we let the online-EM loop run for $n = 20,000$ iterations and use $K = 10$ latent chain trajectories. Alse here the Robbins-Monroe step size is set to $\gamma_\ell = 1/\ell^{0.6}$ and the gradient descent step size to $\delta = 0.01$. Both the parameters and the chain are started at random values and we use a 20 step burn-in phase before the estimation algorithm is triggered. In this case, no closed-form expressions of the optimal kernel and importance weight function are available. Thus, in Figures 3–6 the mixture-approximated proposal kernel and weighting function are shown together with *estimates*, obtained by means of truncated Gaussian kernel density estimation using $10,000$ simulations and bandwidth 0.05, of the optimal ones. In each of these pictures, X_0 is set to each of the column vectors of

$$\begin{pmatrix} -0.7923 & -1.0077 & 1.9120 & 0.6653 \\ 1.8676 & -1.2014 & -0.9948 & 1.9506 \end{pmatrix} \quad (5.2)$$

and Y_1 to each of the column vectors of

$$\begin{pmatrix} 0.3717 & 0.9695 & 1.5260 & 2.3871 \\ 2.9020 & 0.9471 & 0.1831 & 2.7523 \end{pmatrix}, \quad (5.3)$$

where the latter corresponds to the center points of each quartile (with respect to the first component of Y) in a simulated sample of 1000 points, and the former are X-values of the same trajectory (and located at the preceding time-step) as each of these Y's. For each of the fixed (X_0, Y_1)-pairs, the kernel estimation-based as well as the mixture-based approximations of the optimal kernel density $x_1 \mapsto p(x_1|X_0, Y_1)$ and the optimal importance weight function $x_0 \mapsto p(Y_1|x_0)$ are plotted in 2D. As in the previous example, the figures display a nice agreement between the two different approximations, indicating that the mixture-parameters are learned well also in this bivariate case.

Finally, filtering of the hidden process is performed using the APF based on the mixture-optimised proposal kernels and weighting functions. The performence is again compared to that of the vanilla bootstrap filter. In this case a data record comprising 100 observations was swept using 500 particles. In Figure 2 (c-d) the cumulative sums of the resulting sorted normalised particle weights are displayed, each line representing one of the 200 time-steps. The outcome shows again that adjusting a mixture of experts leads to a significantly improved particle filter with a clear advantage over the plain bootstrap filter.

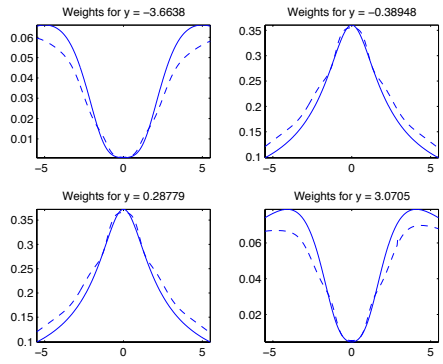

Figure 1: Estimated and true weighting functions at four different time-steps for the ARCH model. The unbroken/dashed lines represent the exact/approximated importance weight functions, respectively. The simulation is based on 500 particles, 9 mixture components, and $20,000$ iterations of the online-EM loop (4.10).

References

Cappe, O. and Moulines, E. (2009). Online EM algorithm for latent data models, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **3**(71), 593–613.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

Chan, B., Doucet, A., and Tadic, V. (2003). Optimisation of particle filters using simultaneous perturbation stochastic approximation, *In proceedings of IEEE ICASSP*.

Cornebise, J., Moulines, E., and Olsson, J. (2009). Approximating the optimal kernel in sequential monte-carlo methods by means of mixture of experts. To be submitted.

Douc, R., Moulines, E., and Olsson, J. (2008). Optimality of the auxiliary particle filter, *Probabability and Mathematical Statistics*, **29**(1), 1–29.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On Sequential Monte Carlo Sampling Methods for Bayesian Filtering, *Statistics and Computing*, **10**, 197–208.

Duflo, M. (1997). *Random Iterative Models*. Springer Verlag, Berlin.

Johansen, A. M. and Doucet, A. (2008). A note on auxiliary particle filters, *Statistics and Probability Letters*.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the em algorithm, *Neural computation*, **6**, 181–214.

Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters, *J. Am. Statist. Assoc.*, **87**, 493–499.

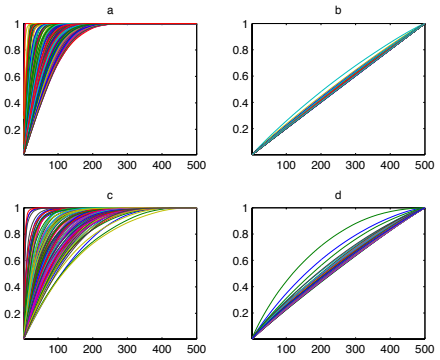

Figure 2: Cumulative sums of sorted normalised particle weights for APFs using mixture-based approximations of the optimal adjustment multipliers and proposal kernel ((b) and (d), corresponding to Example 1 and 2, respectively) and plain bootstrap filters ((a) and (c), corresponding to Example 1 and 2, respectively) for the two models. Each line corresponds to each of the 200 time steps and the particle population size was set to 500 for both models.

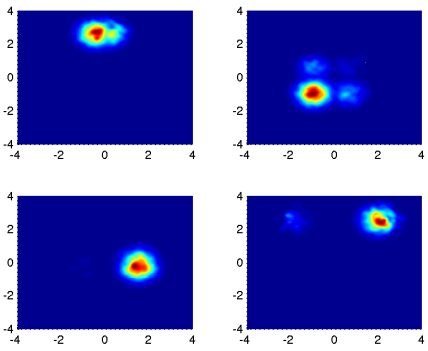

Figure 3: Estimation of the optimal proposal density function $x_1 \mapsto p(x_1|X_0, Y_1)$ obtained by means of truncated Gaussian kernel density estimation using $10,000$ simulations and bandwidth 0.05 for each of the pairs (X_0, Y_1) in (5.2) and (5.3).

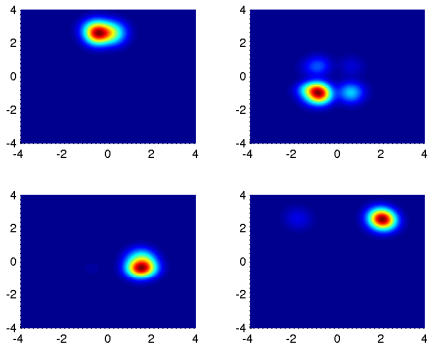

Figure 4: Estimation of the optimal proposal density function $x_1 \mapsto p(x_1|X_0, Y_1)$ obtained by adaptation of a mixture of experts with 12 components using a training sequence of length $20,000$. The approximation is plotted for each of the pairs (X_0, Y_1) in (5.2) and (5.3).

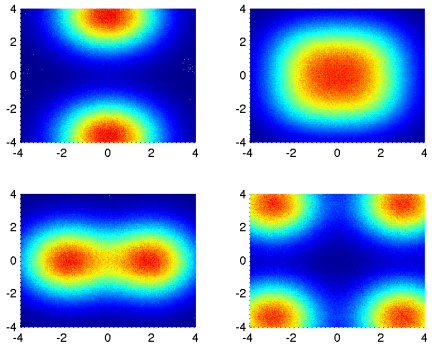

Figure 5: As in Figure 3, but for the approximated optimal adjustment weight function $x_0 \mapsto p(Y_1|x_0)$ instead.

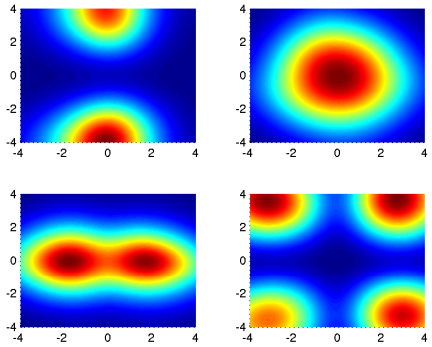

Figure 6: As in Figure 4, but for the approximated optimal adjustment weight function $x_0 \mapsto p(Y_1|x_0)$ instead.