Adaptive Methods for Sequential Importance Sampling

J. Cornebise[a], **É. Moulines**[a], **J. Olsson**[b]
[a] TELECOM ParisTech, France
[b] Lund University, Sweden

MAS 2008
August 29th, 2008
Rennes, France

Outline

We are here $\longrightarrow$ •

Sequential Monte Carlo methods: a short overview

Sequential Monte Carlo (SMC) methods (or *particle filters*) . . .

- . . . approximate a sequence of probability distributions by weighted empirical measures associated with random *particles*. The weighted particle sample is updated recursively in time.
- . . . have developed rapidly since the seminal paper by Gordon *et al.* (1993). See Doucet *et al.* (2001) for a survey and Del Moral (2004) for a comprehensive treatment of theoretical aspects.
- . . . consist in two main operations: *mutation* (importance sampling) and *selection* (resampling), and loads of papers on how to refine these operations have been written during the last decade.

SMC in a nutshell

Assume that the sample $\{(\xi_i, \omega_i)\}_{i=1}^N$, the ξ_i's being referred to as particles, approximates the probability measure ν on Ξ such that

$$\sum_{i=1}^{N} \frac{\omega_i}{\Omega_N} f(\xi_i) \approx \int f(\xi)\, \nu(\mathrm{d}\xi) \;,$$

with $\Omega_N = \sum_{j=1}^N \omega_j$, for a class of target functions f.

Main task: update the sample in order to approximate instead the probability measure

$$\mu(A) = \frac{\int L(\xi, A)\, \nu(\mathrm{d}\xi)}{\int L(\xi, \tilde{\Xi})\, \nu(\mathrm{d}\xi)}$$

on $\tilde{\Xi}$, where L is a finite transition kernel from Ξ to $\tilde{\Xi}$.

SMC in a nutshell

Natural strategy: plug the particle approximation into the transition formula:

$$\mu(A) \approx \mu_N(A) = \frac{\sum_{i=1}^{N} \frac{\omega_i}{\Omega_N} L(\xi_i, A)}{\sum_{i=1}^{N} \frac{\omega_i}{\Omega_N} L(\xi_i, \tilde{\Xi})} = \sum_{i=1}^{N} \frac{\omega_i L(\xi_i, \tilde{\Xi})}{\sum_{j=1}^{N} \omega_j L(\xi_j, \tilde{\Xi})} \left[L(\xi_i, \cdot)/L(\xi_i, \tilde{\Xi}) \right]$$

and try to sample new particles from this *mixture* in order to approximate μ. Repeat recursively.

Problem: sampling from μ_N is in general infeasible without using accept-reject methods—might be expensive.

The auxiliary particle filter

Auxiliary particle filter (Pitt & Shephard, 1999): make instead draws $\{\tilde{\xi}_i\}_{i=1}^N$ from the instrumental mixture distribution

$$\pi_N(A) = \sum_{i=1}^N \frac{\omega_i \psi_i}{\sum_{j=1}^N \omega_j \psi_j} R(\xi_i, A)$$

by

1. drawing indices I_i multinomially w.r.t. the *adjusted* weights $\{\omega_i \psi_i\}_{i=1}^N$, with $\psi_i = \Psi(\xi_i)$, and
2. simulating $\tilde{\xi}_i \sim R(\xi_{I_i}, \cdot)$.

Finally, associate the draws with weights

$$\tilde{\omega}_i = \frac{\mathrm{d}\mu_N}{\mathrm{d}\pi_N}(\tilde{\xi}_i) \ .$$

The auxiliary particle filter

Problem: evaluating $\mathrm{d}\mu_N/\mathrm{d}\pi_N(\tilde{\xi}_i)$ is expensive.

Solution: introduce I_i as *auxiliary variable* and target instead

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

on $\{1, \ldots, N\} \times \tilde{\boldsymbol{\Xi}}$ by sampling, as above, $\{(\tilde{\xi}_i, I_i)\}_{i=1}^N$ from

$$\pi_N^{\mathrm{aux}}(\{i\} \times A) = \frac{\omega_i \psi_i}{\sum_{j=1}^{M_N} \omega_j \psi_j} R(\xi_i, A) \; .$$

and assigning each pair $(\tilde{\xi}_i, I_i)$ the weight

$$\tilde{\omega}_i = \Psi^{-1}(\xi_{I_i}) \frac{\mathrm{d}L(\xi_{I_i}, \cdot)}{\mathrm{d}R(\xi_{I_i}, \cdot)}(\tilde{\xi}_i) \propto \frac{\mathrm{d}\mu_N^{\mathrm{aux}}}{\mathrm{d}\pi_N^{\mathrm{aux}}}(\tilde{\xi}_i, I_i) \; .$$

The auxiliary particle filter

Marginalisation: now, since

$$\mu_N(A) = \sum_{i=1}^{N} \mu_N^{\text{aux}}(\{i\} \times A) \ ,$$

discard the I_i's and let the sample $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^N$ approximate μ_N.

Important issue: how should R be chosen, given adjustment weights $\{\psi_i\}_{i=1}^N$?

The **aim of this work** is to design R adaptively.

We are here $\longrightarrow$ •

SMC adaptation: a short overview

Adaptation: tuning on-line and automatically the key parameters of SMC algorithms.

Examples of work in the past:

- Adapting the **particles sample size**: Legland & Oudjane (2006) avoid degeneracy of the weights; KLD-Sampling by Fox (2003) (refined in Soto, 2005, and Straka & Simandl, 2006).
- Adapting the **proposal distribution**: Pitt & Shephard (1999) and Doucet, Godsill & Andrieu (2000) approximate the optimal kernel $L(\xi, \cdot)/L(\xi, \tilde{\Xi})$.

Kullback-Leibler divergence, chi-square distance

Here we focus on **minimising the following distances** between μ_N^{aux} and π_N^{aux}. Let $\eta \ll \lambda$ be probability measures on $\mathcal{S}$ and recall the Kullback-Leibler divergence (KLD) and the chi-square distance (CSD) between η and λ:

Definition: KLD and CSD

$$d_{\mathrm{KL}}(\eta||\lambda) = \int_{\mathcal{S}} \log \frac{\mathrm{d}\eta}{\mathrm{d}\lambda}(s)\, \eta(\mathrm{d}s) \;,$$

$$d_{\chi^2}(\eta||\lambda) = \int_{\mathcal{S}} \left[\frac{\mathrm{d}\eta}{\mathrm{d}\lambda}(s) - 1\right]^2 \lambda(\mathrm{d}s) \;.$$

Shannon entropy, coefficient of variation

The following quantities will be used for estimating the KLD and CSD between μ_N^{aux} and π_N^{aux}.

Quality criterions

$$\mathcal{E}_N = \sum_{i=1}^{N} \frac{\tilde{\omega}_i}{\tilde{\Omega}_N} \log\left(\frac{N\tilde{\omega}_i}{\tilde{\Omega}_N}\right) ,$$

$$\mathrm{CV}_N^2 = N \sum_{i=1}^{N} \left(\frac{\tilde{\omega}_i}{\tilde{\Omega}_N}\right)^2 - 1 ,$$

where $\tilde{\Omega}_N = \sum_{i=1}^{N} \tilde{\omega}_i$. Here

- $\mathcal{E}_N$ is (almost) the negated *Shannon entropy* of the importance weights.
- CV_N^2 is the squared *coefficient of variation* (Kong, Liu & Wong, 1994) of the importance weights.

Used to measure weight degeneracy: are minimal when the $\tilde{\omega}_i$'s are equal and maximal when only one particle carries all weight.

Limiting measures

Define, on the product space $\Xi \times \tilde{\Xi}$:

Two fundamental measures

$$\mu^*(A) = \frac{\iint_A L(\xi, \mathrm{d}\tilde{\xi})\, \nu(\mathrm{d}\xi)}{\iint L(\xi, \mathrm{d}\tilde{\xi})\, \nu(\mathrm{d}\xi)} ,$$

$$\pi^*[\Psi](A) = \frac{\iint_A \Psi(\xi)\, R(\xi, \mathrm{d}\tilde{\xi})\, \nu(\mathrm{d}\xi)}{\iint \Psi(\xi)\, \nu(\mathrm{d}\xi)} .$$

μ^* and $\pi^*[\Psi]$ can be interpreted as the *asymptotic* target and proposal distributions associated with the auxiliary particle model.

Convergence results

Now the following can be established using convergence results obtained by Douc & Moulines (2007):

Theorem

Under weak technical assumptions, as $N \to \infty$,

$$d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) \xrightarrow{\mathbb{P}} d_{\mathrm{KL}}\left(\mu^* \| \pi^*[\Psi]\right) ,$$
$$d_{\chi^2}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) \xrightarrow{\mathbb{P}} d_{\chi^2}\left(\mu^* \| \pi^*[\Psi]\right) .$$

Theorem

Under similar assumptions, as $N \to \infty$,

$$\mathcal{E}_N \xrightarrow{\mathbb{P}} d_{\mathrm{KL}}\left(\mu^* \| \pi^*[\Psi]\right) ,$$
$$\mathrm{CV}_N^2 \xrightarrow{\mathbb{P}} d_{\chi^2}\left(\mu^* \| \pi^*[\Psi]\right) .$$

Asymptotically optimal weights

By examining carefully $d_{\mathrm{KL}}\left(\mu^* \| \pi^*[\Psi]\right)$ and $d_{\chi^2}\left(\mu^* \| \pi^*[\Psi]\right)$ we also derive the *optimal* adjustment weights Ψ minimising the asymptotic KLD and CSD for a given proposal kernel R:

Corollary

Under weak technical conditions, the following holds:

i) The Kullback-Leibler optimal weight function is given by

$$\arg\min_{\Psi} d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) = L(\xi, \tilde{\Xi}) \; ;$$

i) the chi-square optimal weight function is given by

$$\arg\min_{\Psi} d_{\chi^2}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) = \sqrt{\int \frac{\mathrm{d}L(\xi, \cdot)}{\mathrm{d}R(\xi, \cdot)}(\tilde{\xi})\, L(\xi, \mathrm{d}\tilde{\xi})} \; .$$

Interestingly, the Kullback-Leibler optimal weight does not depend on R!

KLD/CSD-based adaptation: key idea

These insights can be used for efficient adaptive design of π_N^{aux}:

Key idea

Minimise $d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}})$ or $d_{\chi^2}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}})$ over a family $\{R_\theta\}_{\theta \in \Theta}$ of proposal kernels.

This leads to the instrumental distribution

$$\pi_{N,\theta}^{\mathrm{aux}}(\{i\} \times A) = \frac{\omega_i \psi_i}{\sum_{j=1}^N \omega_j \psi_j} R_\theta(\xi_i, A) \ .$$

KLD/CSD-based adaptation: key idea

These insights can be used for efficient adaptive design of π_N^{aux}:

Key idea

Minimise $d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}}||\pi_N^{\mathrm{aux}})$ or $d_{\chi^2}(\mu_N^{\mathrm{aux}}||\pi_N^{\mathrm{aux}})$ over a family $\{R_\theta\}_{\theta\in\Theta}$ of proposal kernels.

This leads to the instrumental distribution

$$\pi_{N,\theta}^{\mathrm{aux}}(\{i\}\times A) = \frac{\omega_i\psi_i}{\sum_{j=1}^N \omega_j\psi_j} R_\theta(\xi_i, A) \; .$$

Two possible approaches:

- Minimise $\mathcal{E}_N$ or CV_N^2 by fixing random seed (not detailed here).
- Minimise

$$d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}}||\pi_{N,\theta}^{\mathrm{aux}}) = \sum_{i=1}^N \int \log\left(\frac{\mathrm{d}\mu_N^{\mathrm{aux}}}{\mathrm{d}\pi_{N,\theta}^{\mathrm{aux}}}(i,\tilde{\xi})\right) \mu_N^{\mathrm{aux}}(\{i\}\times \mathrm{d}\tilde{\xi}) \; .$$

  via a *cross-entropy* approach (Rubinstein & Kroese, 2004).

Cross-entropy-based SMC adaptation

Iteration ℓ, current parameter fit $\theta_\ell \in \Theta$.

- **IS approximation (E-step)**: Sample M^ℓ draws $\{(I_i^\ell, \tilde{\xi}_i^\ell)\}_{i=1}^{M^\ell}$ from $\pi_{N,\theta_\ell}^{\mathrm{aux}}$.
- **Optimisation (M-step)**: Solve

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

  or equivalently, assuming that $R_\theta(\xi, \cdot)$ has a density $r_\theta(\xi, \cdot)$,

$$\theta_{\ell+1} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{M^\ell} \frac{\tilde{\omega}_i^\ell}{\tilde{\Omega}_N^\ell} \log r_\theta(\xi_{I_i^\ell}, \tilde{\xi}_i^\ell) .$$

Choosing $\{r_\theta\}_{\theta\in\Theta}$ intelligently, *e.g.*, as a *normal exponential family*, render closed-form (and thus quick) minimisation possible.

Theoretical justication in progress. . .

We are here ⟶ •

State space models

Consider a *state space model* (or *hidden Markov model*):

$$X_{n+1}|X_n \sim Q_n(X_n, \cdot) \ ,$$
$$Y_n|X_n \sim g_n(X_n, \cdot) \ ,$$

for $n \geq 0$, and the problem of computing the *filter distributions*

$$\phi_n(A) = \mathbb{P}(X_n \in A|Y_{0:n}) \ .$$

This problem can, via the *filtering recursion*, be perfectly cast into our framework since

$$\phi_{n+1}(A) = \frac{\int L_n(\xi, A)\, \phi_n(\mathrm{d}\xi)}{\int L_n(\xi, \tilde{\Xi})\, \phi_n(\mathrm{d}\xi)} \ ,$$

with

$$L_n(\xi, A) = \int_A g_n(\tilde{\xi}, Y_{n+1})\, Q_n(\xi, \mathrm{d}\tilde{\xi}) \ .$$

ARCH model observed in noise

As a special case, consider, for $n \geq 0$, the following *ARCH model* observed in noise:

$$X_{n+1} = W_{n+1}\sqrt{\beta_0 + \beta_1 X_n^2}\ ,$$
$$Y_n = X_n + \sigma_v V_n\ ,$$

with $\{W_n\}_{n\geq 1}$ and $\{V_n\}_{n\geq 0}$ being white noise sequences, and $(\beta_0, \beta_1, \sigma_V^2) = (1, 0.99, 10)$.

- $\{X_n\}_{n\geq 0}$ is stationary with stationary variance σ_{s}^2.
- This model belongs to a family of models for which direct simulation from the optimal kernels $L_n(\xi, \cdot)/L_n(\xi, \tilde{\Xi})$ is feasible and closed-form expressions of the optimal weights $\Psi_n(\xi) = L_n(\xi, \tilde{\Xi})$ are tractable (*full adaptation*); this is interesting from a **comparative** point of view.

CE-adaptation within the ARCH framework

Let $m(\xi, Y_{n+1})$ and $\eta(\xi)$ denote mean and standard deviation of the—in this case Gaussian—optimal kernel $L_n(\xi, \cdot)/L_n(\xi, \tilde{\Xi})$. We consider adaptation of a proposal $\pi^{\text{aux}}_{N,\theta}$, with $\psi_i \equiv 1$, over the family $\{R_\theta\}_{\theta \in (0,c]}$, where

$$R_\theta(\xi, \cdot) = \mathcal{N}(m(\xi, Y_{k+1}), \theta\eta(\xi)) \ ,$$

and study how well the cross-entropy updates

$$\theta_{\ell+1} = \sqrt{\sum_{i=1}^{M^\ell} \frac{\tilde{\omega}_i^\ell}{\tilde{\Omega}_N^\ell \eta^2(\xi_{I_i^\ell})} \left[\tilde{\xi}_i^\ell - m(\xi_{I_i^\ell}, Y_{n+1})\right]^2}$$

minimise the—for this toy example explicitely known—Kullback-Leibler distance

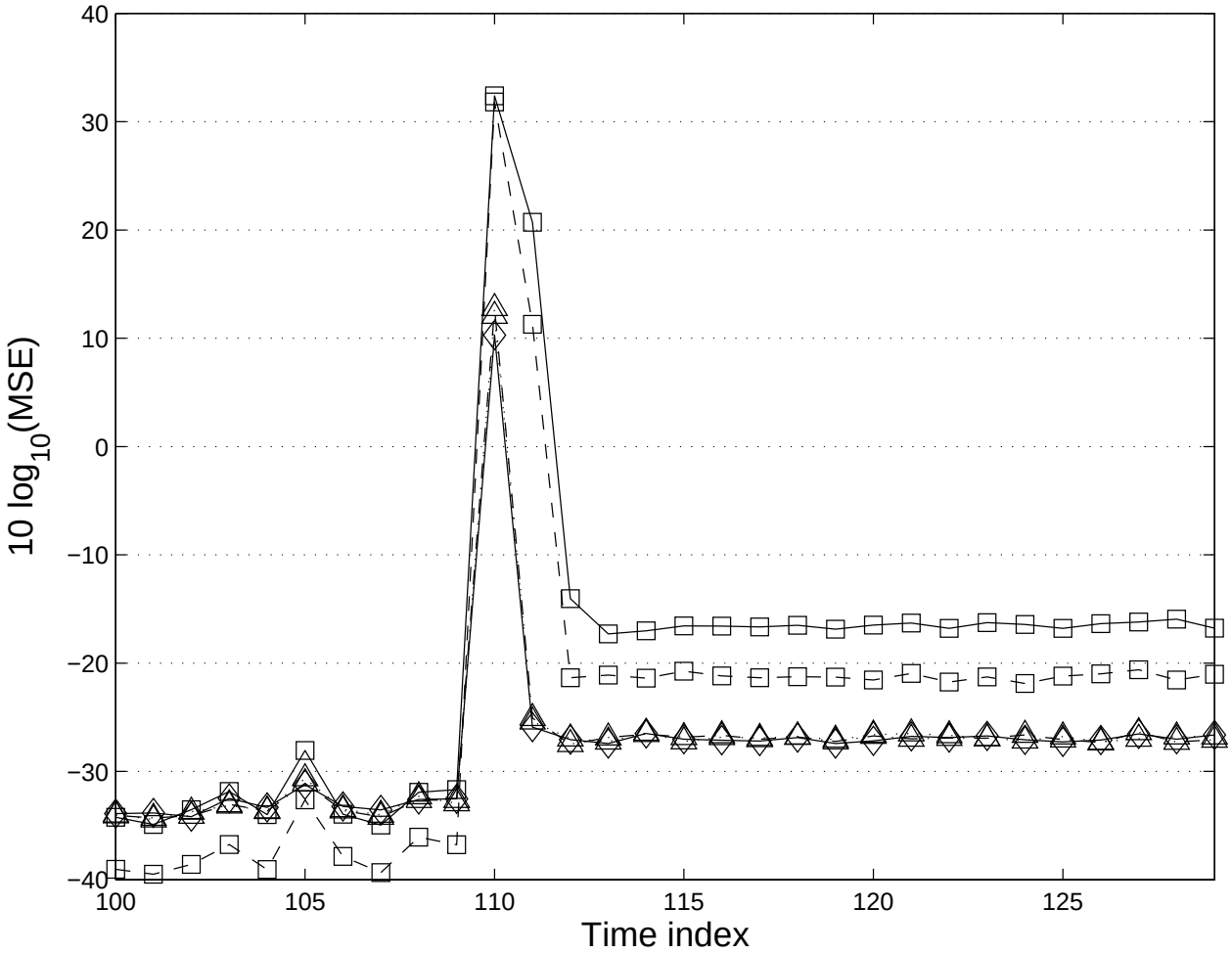
$$\begin{aligned} &d_{\mathrm{KL}}(\mu^{\text{aux}}_N || \pi^{\text{aux}}_{N,\theta}) \\ &= \sum_{i=1}^N \frac{\omega_i L_n(\xi_i, \tilde{\Xi})}{\sum_{j=1}^N \omega_j L_n(\xi_j, \tilde{\Xi})} \left[\log\left(\frac{L_n(\xi_i, \tilde{\Xi})\Omega_N}{\sum_{j=1}^N \omega_j L_n(\xi_j, \tilde{\Xi})}\right) + \log\theta + \frac{1}{2}\left(\frac{1}{\theta^2} - 1\right)\right] . \end{aligned}$$

CE-adaptation within the ARCH framework

In order to make the problem challenging, we filtered an observation sequence with constant outliers $Y_n \equiv 6\sigma_s$ for $n \geq 110$.

In this setting we compared the following particle filters:

1. plain nonadaptive bootstrap particle filter (using $\Psi \equiv 1$ and $R = Q$),
2. plain nonadaptive bootstrap particle filter with $3N$ particles,
3. fully adapted filter with optimal weights and kernel,
4. adaptive bootstrap filter minimising directly $d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_{N,\theta}^{\mathrm{aux}})$, yielding $\theta = 1$ for all time steps,
5. CE-based adaptive bootstrap filter with $\theta^0 = 10$.



MSE values (in Decibel) for cross-entropy-based adaptative filter ($\triangle$, $-\cdot$) and bootstrap filter with $3N$ particles ($\square$, $--$). **Reference filters:** bootstrap filter ($\square$, continuous line), fully adapted filter ($\diamond$), and bootstrap filter with proposal parameter minimising the current KLD ($\triangle$, continuous line).

Thanks for the attention

Any remark, question, or suggestion is warmly welcomed!

jimmy@maths.lth.se
http://www.maths.lth.se/matstat/staff/jimmy/