SMC WITH ADAPTIVE RESAMPLING: LARGE SAMPLE ASYMPTOTICS

Élise Arnaud and François Le Gland

Université Joseph Fourier, Laboratoire Jean Kuntzman and INRIA Rhône Alpes, France
INRIA Rennes and IRISA, France

ABSTRACT

A longstanding problem in sequential Monte Carlo (SMC) is to mathematically prove the popular belief that resampling does improve the performance of the estimation (this of course is not always true, and the real question is to clarify classes of problems where resampling helps). A more pragmatic answer to the problem is to use adaptive procedures that have been proposed on the basis of heuristic considerations, where resampling is performed only when it is felt necessary, i.e. when some criterion (effective number of particles, entropy of the sample, etc.) reaches some prescribed threshold. It still remains to mathematically prove the efficiency of such adaptive procedures. The contribution of this paper is to propose an approach, based on a representation in terms of multiplicative functionals (in which importance weights are treated as particles, roughly speaking) to obtain the asymptotic variance of adaptive resampling procedures, when the sample size goes to infinity. It is then possible to see the impact of the threshold on the asymptotic variance, at least in the Gaussian case, where the resampling criterion has an explicit expressions in the large sample asymptotics.

1. INTRODUCTION

Consider the unnormalized and normalized Feynman–Kac distributions defined on the set E by

$$\langle \gamma_n, \phi \rangle = \int_E \cdots \int_E \phi(x_n)\, \gamma_0(dx_0) \prod_{k=1}^{n} R_k(x_{k-1}, dx_k) \,,$$

and

$$\langle \mu_n, \phi \rangle = \frac{\langle \gamma_n, \phi \rangle}{\langle \gamma_n, 1 \rangle} \,,$$

respectively, characterized by

- the nonnegative measure $\gamma_0(dx)$,
- and the nonnegative kernels $R_k(x, dx')$,

for any $k = 1, \cdots, n$. Associated with this integral representation is the recurrence relation $\gamma_k = \gamma_{k-1}\, R_k$, for any $k = 1, \cdots, n$. In full generality, it is always possible to decompose the nonnegative measure

$$\gamma_0(dx) = W_0(x)\, p_0(dx) \,, \tag{1}$$

in terms of a nonnegative function and a normalized probability distribution, and to decompose the nonnegative kernel

$$R_k(x, dx') = W_k(x, x')\, P_k(x, dx') \,, \tag{2}$$

in terms of a nonnegative function and a normalized Markov kernel, for any $k = 1, \cdots, n$. Given the decompositions (1) and (2) introduce the nonnegative measure and the nonnegative kernels

$$\gamma_0^{\square}(dx) = |W_0(x)|^2\, p_0(dx) \,,$$

and

$$R_k^{\square}(x, dx') = |W_k(x, x')|^2\, P_k(x, dx') \,,$$

respectively, for any $k = 1, \cdots, n$. With the decompositions (1) and (2) is associated the equivalent probabilistic representation

$$\langle \gamma_n, \phi \rangle = \mathbb{E}[\phi(X_n) \prod_{k=0}^{n} W_k(X_{k-1}, X_k)\,] \,, \tag{3}$$

where $\{X_k \,,\ k = 0, 1, \cdots, n\}$ is a Markov chain taking values in E and characterized by

- its initial probability distribution $p_0(dx)$,
- and its transition probabilities $P_k(x, dx')$,

for any $k = 1, \cdots, n$, and where $W_k(x, x')$ is a bounded nonnegative function for any $k = 0, 1, \cdots, n$, with the abuse of notation $W_0(x, x') = W_0(x')$ to be used throughout the paper for $k = 0$.

Starting from the probabilistic representation (3) for the unnormalized distribution, a first type of Monte Carlo approximation can be designed as

$$\langle \gamma_n, \phi \rangle \approx \frac{1}{N} \sum_{i=1}^{N} \phi(\xi_n^i) \prod_{k=0}^{n} W_k(\xi_{k-1}^i, \xi_k^i) \,,$$

978-1-4244-2710-9/09/$25.00

for any bounded measurable function ϕ, where independently for any $i = 1, \cdots, N$

$$(\xi_0^i, \cdots, \xi_n^i) \sim p_0(dx_0)\ P_1(x_0, dx_1) \cdots P_n(x_{n-1}, dx_n)\ ,$$

i.e. $\xi_{0:n}^i = (\xi_0^i, \cdots, \xi_n^i)$ is distributed as a sample path of the Markov chain. This immediately results in a self–normalized approximation

$$\mu_n \approx \mu_n^N = \frac{\gamma_n^N}{\langle \gamma_n^N, 1 \rangle} = \sum_{i=1}^N w_n^i\ \delta_{\xi_n^i}\ ,$$

of the normalized distribution, where

$$w_n^i \propto \prod_{k=0}^n W_k(\xi_{k-1}^i, \xi_k^i)\ ,$$

in terms of a weighted empirical probability distribution characterized by sample positions $(\xi_n^1, \cdots, \xi_n^N)$ and sample weights $(w_n^1, \cdots, w_n^N)$, which can be computed recursively as follows

$$\xi_k^i \sim P_k(\xi_{k-1}^i, dx') \quad \text{and} \quad w_k^i \propto w_{k-1}^i\ W_k(\xi_{k-1}^i, \xi_k^i)\ ,$$

for any $i = 1, \cdots, N$. There is a well–known limitation with this first type of Monte Carlo approximation: indeed, it appears after some iterations that a few sample paths have a significant weight, while the remaining sample paths have a negligible weight, which means that only a few sample paths are really contributing to the approximation. A possible remedy to this degeneracy problem is to use the weights to resample from the current weighted empirical probability distributions, so that samples with a high weight are replicated while samples with a low weight are discarded. Several different resampling strategies are available. Resampling introduces extra additional randomness in the sample. A more efficient implementation is to resample not at each iteration of the algorithm, but only at some time instants, and the question that naturally arises is when to resample, i.e. how to take the decision to resample. Two different approaches can be used to address this issue:

- evaluate the asymptotic variance, as the sample size N goes to infinity, of the algorithm that resamples at some prescribed times $0 \leq t_1 < \cdots < t_p < n$, and optimize the expression of the asymptotic variance with respect to the number and the location of the resampling times,
- evaluate the asymptotic variance, as the sample size N goes to infinity, of the adaptive algorithm [1] that resamples at times where some heuristic rule is met, e.g. if the effective sample size (or alternatively the entropy of the weighted sample) drops below a prescribed level, and study the impact of the prescribed level, or threshold.

The first approach relies on considering the Markov chain with values in path space, and has been presented at the workshop on adaptive Monte Carlo methods held in Fleurance in July 2007. The second approach relies on considering another Markov chain, with values in the (space, weight) product space, along the lines of [2].

2. NONLINEAR ADAPTIVE MARKOV MODEL

On the product set $E \times [0, \infty)$, define the probability distribution

$$\mu_0^{\mathrm{e}}(dx_0, dv_0) = p_0(dx_0)\ \delta_{W_0(x_0)}(dv_0)\ ,$$

and the nonnegative kernels

$$R_k^{\mathrm{e}}(\mu, x, v, dx', dv') =$$

$$= g_{k-1}^{\mathrm{e}}(\mu, v)\ \underbrace{P_k(x, dx')\ \delta_{f_k^{\mathrm{e}}(\mu, x, v, x')}(dv')}_{P_k^{\mathrm{e}}(\mu, x, v, dx', dv')}\ ,$$

where

$$f_k^{\mathrm{e}}(\mu, x, v, x') = \begin{cases} v\ W_k(x, x')\ , & \text{if } \mu \notin D, \\ W_k(x, x')\ , & \text{if } \mu \in D, \end{cases}$$

and

$$g_{k-1}^{\mathrm{e}}(\mu, v) = \begin{cases} 1\ , & \text{if } \mu \notin D, \\ v\ , & \text{if } \mu \in D, \end{cases}$$

by definition, so that the identity

$$g_{k-1}^{\mathrm{e}}(\mu, v)\ f_k^{\mathrm{e}}(\mu, x, v, x') = v\ W_k(x, x')\ ,$$

holds for any $k = 1, \cdots, n$. Clearly

$$R_k^{\mathrm{e}}(\mu, x, v, dx', dv') =$$

$$= 1_{(\mu \notin D)}\ \underbrace{P_k(x, dx')\ \delta_{v\, W_k(x, x')}(dv')}_{P_k^{\mathrm{imp}}(x, v, dx', dv')}$$

$$+ 1_{(\mu \in D)}\ \underbrace{v\ P_k(x, dx')\ \delta_{W_k(x, x')}(dv')}_{R_k^{\mathrm{red}}(x, v, dx', dv')}\ ,$$

and define

$$g_{k-1}^{\mathrm{red}}(x, v) = v\ ,$$

and

$$P_k^{\mathrm{red}}(x, v, dx', dv') = P_k(x, dx')\ \delta_{W_k(x, x')}(dv')\ ,$$

for any $k = 1, \cdots, n$. Let $e(v) = v$ for any $v \geq 0$, by definition.

Example 2.1. For instance

$$D = \{\mu \in \mathcal{P} \,:\, \frac{\langle \mu, 1 \otimes e^2 \rangle}{\langle \mu, 1 \otimes e \rangle^2} \geq c\} \,,$$

where $c \geq 1$ is a threshold to be fixed.

Consider next the unnormalized and normalized Feynman–Kac distributions defined on the product set $E \times [0, \infty)$ by

$$\langle \gamma_n^{\mathrm{e}}, F \rangle \;=\; \int_E \int_0^\infty \cdots \int_E \int_0^\infty F(x_n, v_n)\; \mu_0^{\mathrm{e}}(dx_0, dv_0)$$

$$\prod_{k=1}^{n} R_k^{\mathrm{e}}(\mu_{k-1}^{\mathrm{e}}, x_{k-1}, v_{k-1}, dx_k, dv_k)$$

and

$$\langle \mu_n^{\mathrm{e}}, F \rangle = \frac{\langle \gamma_n^{\mathrm{e}}, F \rangle}{\langle \gamma_n^{\mathrm{e}}, 1 \rangle} \,,$$

respectively. Associated with this integral representation is the recurrence relation

$$\gamma_k^{\mathrm{e}} \;=\; \gamma_{k-1}^{\mathrm{e}}\, R_k^{\mathrm{e}}(\mu_{k-1}^{\mathrm{e}})$$

$$\;=\; \mu_{k-1}^{\mathrm{e}}\, R_k^{\mathrm{e}}(\mu_{k-1}^{\mathrm{e}})\; \langle \gamma_{k-1}^{\mathrm{e}}, 1 \rangle$$

$$\;=\; [\, 1_{(\mu_{k-1}^{\mathrm{e}} \notin D)}\; \mu_{k-1}^{\mathrm{e}}\, P_k^{\mathrm{imp}}$$

$$+\, 1_{(\mu_{k-1}^{\mathrm{e}} \in D)}\; (g_{k-1}^{\mathrm{red}}\; \mu_{k-1}^{\mathrm{e}})\; P_k^{\mathrm{red}}\,]\; \langle \gamma_{k-1}^{\mathrm{e}}, 1 \rangle \,,$$

for the unnormalized distribution, hence

$$\langle \gamma_k^{\mathrm{e}}, 1 \rangle \;=\; [\, 1_{(\mu_{k-1}^{\mathrm{e}} \notin D)}$$

$$+\, 1_{(\mu_{k-1}^{\mathrm{e}} \in D)}\; \langle \mu_{k-1}^{\mathrm{e}}, g_{k-1}^{\mathrm{red}} \rangle\,]\; \langle \gamma_{k-1}^{\mathrm{e}}, 1 \rangle \,,$$

for the normalizing constant, and

$$\mu_k^{\mathrm{e}} \;=\; 1_{(\mu_{k-1}^{\mathrm{e}} \notin D)}\; \mu_{k-1}^{\mathrm{e}}\, P_k^{\mathrm{imp}}$$

$$+\, 1_{(\mu_{k-1}^{\mathrm{e}} \in D)}\; (g_{k-1}^{\mathrm{red}} \cdot \mu_{k-1}^{\mathrm{e}})\; P_k^{\mathrm{red}} \,,$$

for the normalized distribution, for any $k = 1, \cdots, n$.

The Feynman–Kac distributions for the more general nonlinear Markov model defined on the product set $E \times [0, \infty)$ includes as a special case the Feynman–Kac distributions for the original model and other unnormalized distributions defined on the set E, for appropriate choices of test functions. Indeed, introduce

$$\Gamma_k = \langle \gamma_k^{\mathrm{e}}, 1 \rangle \,, \qquad \langle \gamma_k^{(1)}, \phi \rangle = \langle \gamma_k^{\mathrm{e}}, \phi \otimes e \rangle$$

and

$$\langle \gamma_k^{(2)}, \phi \rangle = \langle \gamma_k^{\mathrm{e}}, \phi \otimes e^2 \rangle \,,$$

and notice that the statistics that governs redistribution can also be expressed in terms of normalizing constants as follows

$$c_k = \frac{\langle \mu_k^{\mathrm{e}}, 1 \otimes e^2 \rangle}{\langle \mu_k^{\mathrm{e}}, 1 \otimes e \rangle^2} = \frac{\langle \gamma_k^{(2)}, 1 \rangle\; \Gamma_k}{\langle \gamma_k, 1 \rangle^2} \,, \tag{4}$$

for any $k = 0, 1, \cdots, n$.

Proposition 2.2. *The two unnormalized distributions defined above satisfy the following recurrence relations*

$$\gamma_k^{(1)} = \gamma_{k-1}^{(1)}\, R_k \,,$$

with initial condition $\gamma_0^{(1)} = \gamma_0$, which implies that $\gamma_k^{(1)} = \gamma_k$, and

$$\gamma_k^{(2)} = 1_{(\mu_{k-1}^{\mathrm{e}} \notin D)}\; \gamma_{k-1}^{(2)}\, R_k^{\square} + 1_{(\mu_{k-1}^{\mathrm{e}} \in D)}\; \gamma_{k-1}\, R_k^{\square} \,,$$

with initial condition $\gamma_0^{(2)} = \gamma_0^{\square}$, and the normalizing constant satisfies the following recurrence relation

$$\Gamma_k = 1_{(\mu_{k-1}^{\mathrm{e}} \notin D)}\; \Gamma_{k-1} + 1_{(\mu_{k-1}^{\mathrm{e}} \in D)}\; \langle \gamma_{k-1}, 1 \rangle \,,$$

with initial condition $\Gamma_0 = 1$.

3. PARTICLE APPROXIMATION

Introducing a particle approximation of the form

$$\mu_{k-1}^{\mathrm{e}} \approx \mu_{k-1}^{\mathrm{e},N} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\xi_{k-1}^i,\, v_{k-1}^i)} \,,$$

yields : if $\mu_{k-1}^{\mathrm{e},N} \notin D$, then

$$\mu_{k-1}^{\mathrm{e},N}\, P_k^{\mathrm{imp}}(dx', dv') =$$

$$= \frac{1}{N} \sum_{i=1}^{N} \underbrace{P_k(\xi_{k-1}^i, dx')\; \delta_{v_{k-1}^i}\, W_k(\xi_{k-1}^i, x')(dv')}_{m_k^i(dx', dv')} \,,$$

hence the approximation

$$\mu_k^{\mathrm{e},N} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\xi_k^i,\, v_k^i)} \,,$$

where independently for any $i = 1, \cdots, N$ the random variable (ξ_k^i, v_k^i) is distributed according to $m_k^i(dx', dv')$, or equivalently

$$\xi_k^i \sim P_k(\xi_{k-1}^i, dx') \quad \text{and} \quad v_k^i = v_{k-1}^i\, W_k(\xi_{k-1}^i, \xi_k^i) \,,$$

and if $\mu_{k-1}^{\mathrm{e},N} \in D$, then

$$g_{k-1}^{\mathrm{red}} \cdot \mu_{k-1}^{\mathrm{e},N} = \sum_{i=1}^{N} w_{k-1}^i\; \delta_{(\xi_{k-1}^i,\, v_{k-1}^i)} \,,$$

with $w_{k-1}^i \propto v_{k-1}^i$, and

$$(g_{k-1}^{\mathrm{red}} \cdot \mu_{k-1}^{\mathrm{e},N})\, P_k^{\mathrm{red}}(dx', dv') =$$

$$= \underbrace{\sum_{i=1}^N w_{k-1}^i \; P_k(\xi_{k-1}^i, dx') \; \delta_{W_k(\xi_{k-1}^i, x')}(dv')}_{\bar{m}_k(dx', dv')} ,$$

hence the approximation

$$\mu_k^{\mathrm{e},N} = \frac{1}{N} \sum_{i=1}^N \delta_{(\xi_k^i, v_k^i)} ,$$

where independently for any $i = 1, \cdots, N$ the random variable (ξ_k^i, v_k^i) is distributed according to $\bar{m}_k(dx', dv')$, or equivalently

$$\xi_k^i \sim P_k(\xi_{k-1}^{\tau_k^i}, dx') \quad \text{and} \quad v_k^i = W_k(\xi_{k-1}^{\tau_k^i}, \xi_k^i) ,$$

where the index $\tau_k^i \in \{1, \cdots, N\}$ is selected according to the weights $(w_{k-1}^1, \cdots, w_{k-1}^N)$, e.g. using multinomial sampling or some other sampling strategy from a discrete probability distribution.

Example 3.1. Notice that the *effective sample size* can be expressed in terms of the proposed particle approximation. Indeed

$$\frac{\langle \mu_{k-1}^{\mathrm{e},N}, 1 \otimes e^2 \rangle}{\langle \mu_{k-1}^{\mathrm{e},N}, 1 \otimes e \rangle^2} = \frac{\frac{1}{N} \sum_{i=1}^N |v_{k-1}^i|^2}{|\frac{1}{N} \sum_{i=1}^N v_{k-1}^i|^2} = N \sum_{i=1}^N |w_{k-1}^i|^2 ,$$

i.e.

$$\frac{\langle \mu_{k-1}^{\mathrm{e},N}, 1 \otimes e^2 \rangle}{\langle \mu_{k-1}^{\mathrm{e},N}, 1 \otimes e \rangle^2} = \frac{N}{N_{\mathrm{eff}}} \geq 1 ,$$

where N_{eff} is the *effective sample size*.

This particle approximation for the nonlinear adaptive Markov model coincides with the classical particle approximation [1] with adaptive resampling, and it satisfies a CLT that can be proved using standard techniques.

4. CENTRAL LIMIT THEOREM

Let $\mathbb{T} = \{t_1, \cdots, t_p\} \subset \{0, 1, \cdots, n-1\}$ denotes the set of redistribution times, i.e. $t \in \mathbb{T}$ iff $\mu_t^{\mathrm{e}} \in D$, and let $\mathbb{T}^+ = \mathbb{T} \cup \{n\} = \{t_1, \cdots, t_p, t_{p+1}\}$ with $t_{p+1} = n$ by convention.

Theorem 4.1. *For the normalization constant and for the normalized distribution, it holds*

$$\sqrt{N} \, [\, \frac{\langle \gamma_n^N, 1 \rangle}{\langle \gamma_n, 1 \rangle} - 1 \,] \Longrightarrow \mathcal{N}(0, V_n) ,$$

and

$$\sqrt{N} \, \langle \mu_n^N - \mu_n, \phi \rangle \Longrightarrow \mathcal{N}(0, v_n(\phi)) ,$$

in distribution as $N \uparrow \infty$, for any bounded measurable function ϕ, with the following expression for the asymptotic variances

$$V_n = \sum_{t \in \mathbb{T}^+} [\, \frac{\langle \mu_t^{(2)}, |R_{t+1:n} \, 1|^2 \rangle}{\langle \mu_t, R_{t+1:n} \, 1 \rangle^2} \, c_t - 1 \,] ,$$

and

$$v_n(\phi) = \sum_{t \in \mathbb{T}^+} \frac{\langle \mu_t^{(2)}, |R_{t+1:n} \, (\phi - \langle \mu_n, \phi \rangle)|^2 \rangle}{\langle \mu_t, R_{t+1:n} \, 1 \rangle^2} \, c_t ,$$

respectively, where

$$R_{k:n} \, \phi(x) \; = \; \mathbb{E}[\phi(X_n) \prod_{p=k}^n W_p(X_{p-1}, X_p) \mid X_{k-1} = x] ,$$

for any $k = 1, \cdots, n+1$, with the convention $R_{n+1:n} \, \phi(x) = \phi(x)$ for any $x \in E$.

This result is of theoretical more than practical interest, since computing the asymptotic variances V_n or $v_n(\phi)$ for the estimation of $\langle \gamma_n, 1 \rangle$ or $\langle \mu_n, \phi \rangle$, respectively, is as complicated as computing the desired object itself. Still, it is possible in some simple cases to compute the exact expression of the asymptotic variance and to compute the exact expression of the limiting number of redistributions, as a function of the threshold c. Such curves can be obtained for instance for a one dimensional linear filtering problem, where the normalizing constant $\langle \gamma_n, 1 \rangle$, to be interpreted as the likelihood of the model, can be computed exactly in terms of a Kalman filter, but where the asymptotic variance V_n and the limiting number of redistributions can also be computed exactly, and it is also possible to visualize the impact of the ratio standard deviation of state noise / standard deviation of observation noise.

5. REFERENCES

[1] Arnaud Doucet, Simon J. Godsill, and Christophe Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.

[2] François Le Gland, "Combined use of importance weights and resampling weights in sequential Monte Carlo methods," *ESAIM : Proceedings*, vol. 19, pp. 85–100 (electronic), 2007.