Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models

Julyan Arbel, Jean-Bernard Salomond

Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models

Julyan Arbel[*]
Inria Grenoble, Université Grenoble Alpes
julyan.arbel@inria.fr

Jean-Bernard Salomond[†]
Université Paris-Est, LAMA
jean-bernard.salomond@u-pec.fr

Abstract

In mixture models, latent variables known as allocation variables play an essential role by indicating, at each iteration, to which component of the mixture observations are linked. In sequential algorithms, these latent variables take on the interpretation of particles. We investigate the use of quasi Monte Carlo within sequential Monte Carlo methods (a technique known as sequential quasi Monte Carlo) in nonparametric mixtures for density estimation. We compare them to sequential and non sequential Monte Carlo algorithms. We highlight a critical distinction of the allocation variables exploration of the latent space under each of the three sampling approaches.

Keywords: Bayesian Nonparametrics, Density estimation, Quasi random variables, Monte Carlo methods, Sequential sampling.

1 Introduction

Sequential quasi Monte Carlo (SQMC) sampling is a novel sampling scheme proposed by Gerber and Chopin (2015) in a paper read before *The Royal Statistical Society*. It bridges ideas from quasi Monte Carlo (QMC) methods and from sequential Monte Carlo (SMC) sampling. The latter, also called particle filtering, focuses on sampling in state space models based on sequential data, while the former's initial goal is to enhance Monte Carlo efficiency by using evenly spread vectors instead of (unconstrained) random vectors. As Gerber & Chopin put it,

> *we can only think that the full potential of QMC in statistics remains underexplored.*

In this note, we explore applications of SQMC to the field of Bayesian nonparametrics. More specifically, we focus on nonparametric mixtures for density estimation. The method of nonparametric mixtures of kernels appeared in Ferguson (1973) and Lo (1984) with Dirichlet process mixtures (DPM). Posterior sampling for DPM is usually carried out by Markov chain Monte Carlo (MCMC, see Jain and Neal, 2004). See also Blei and Jordan (2006) for a variational Bayes approach. A known drawback of MCMC is its difficulty to cope with multimodality, ie the Markov chain can be trapped in local modes. Another limitation of MCMC is that it is not suited for online data acquisition, that is when data points are observed one at a time (see Caron et al., 2008). Sequential Monte Carlo methods have proved successful in addressing these issues, in general and also in the particular context of nonparametric mixture models. These models were first cast as SMC samplers by Liu (1996); MacEachern et al. (1999) as follows. Observations are spread out into unobserved clusters. Hence the cluster labels, or allocation variables, are *latent* variables which act as the *states* of the

[*]Inria Grenoble - Rhône-Alpes, 655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin, France & Laboratoire Jean Kuntzmann, Université Grenoble Alpes, http://www.julyanarbel.com/

[†]Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées (UMR 8050), UPEM, UPEC, CNRS, F-94010, Créteil, France, https://sites.google.com/site/jbsalomond/

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

observations in the context of SMC. Two features of the states in this setting are unusual in a QMC setting and will be discussed in this note: they are *discrete* and *non Markovian*. The original algorithm was later extended to particle filters by Fearnhead (2004), and further variations and applications have been proposed by Ülker et al. (2010); Carvalho et al. (2010); Griffin (2015); Tsiligkaridis and Forsythe (2015), among others.

Our goal here is to extend these existing SMC samplers for DPM to the S*Q*MC paradigm. How does SQMC fare in this nonparametric Bayes framework? The extension that we pursue is not trivial and challenging due to the two aforementioned peculiarities: (i) the state-space is discrete, unlike the examples developed by Gerber and Chopin (2015) with continuous state-spaces; (ii) the allocation variables (the states) are not Markovian. The challenges can be formulated as follows: are QMC methods still advantageous in discrete state-space settings? Is the non-Markovianity of the latent variables an impediment to practical implementation, or further to computational gains? For our investigation, we confine attention to Dirichlet process mixtures. Nevertheless, the ideas developed here reach far beyond this model.

2 Background on sampling: SMC, QMC and SQMC

Sequential inference refers to the ability to update the estimation as new observations arrive. **Sequential Monte Carlo** (SMC), or Particle filtering, is a principled technique which sequentially approximates the full posterior using particles (Doucet et al., 2001; Del Moral et al., 2006; Andrieu et al., 2010). It focuses on sequential state-space models: the density of the observations $\mathbf{y}_t$ conditionally on Markov states $\mathbf{x}_t$ in $\mathcal{X} \subseteq \mathbb{R}^d$ is given by $\mathbf{y}_t|\mathbf{x}_t \sim f^Y(\mathbf{y}_t|\mathbf{x}_t)$, with kernel

$$\mathbf{x}_0 \sim f_0^X(\mathbf{x}_0), \quad \mathbf{x}_t|\mathbf{x}_{t-1} \sim f^X(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{1}$$

The output of an SMC sampler with N particles on $(1:T)$ time period is a collection of weighted particles $\{W_T^{(n)}, \mathbf{x}_{1:T}^{(n)}\}$ for $n = 1, \ldots, N$, $W_T^{(n)} > 0$, $\sum_{n=1}^N W_T^{(n)} = 1$ which approximate the posterior distribution of the states, $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$, by a discrete distribution whose support points are the particles. Thus the posterior expectation of any function g of the states can be approximated by

$$\mathbb{E}[g(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})] \approx \sum_{n=1}^N W_T^{(n)} g(\mathbf{x}_{1:T}^{(n)})$$

with an error rate of order $\mathcal{O}_P(N^{-1/2})$.

The initial motivation of **quasi Monte Carlo** (QMC) is to use *low discrepancy* vectors instead of unconstrained random vectors in order to improve the calculation of integrals via Monte Carlo. The discrepancy $D(\mathbf{u})$ of a sample of vectors $\mathbf{u} = \mathbf{u}^{(1:N)}$ is more precisely defined as the *total variation distance* (TV)

$$D(\mathbf{u}) = d_{TV}(\mathbb{P}_{\mathbf{u}}, \lambda)$$

between the empirical measure of the sample $\mathbf{u}$, $\mathbb{P}_{\mathbf{u}}(\cdot) = \frac{1}{N}\sum_{n=1}^N \mathbb{1}(\mathbf{u}^{(n)} \in \cdot)$ and the Lebesgue measure λ. Recall that the TV distance between two probability measures P and Q is defined by

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

where $\mathcal{A}$ is a set of measurable sets, for instance Borel sets.

The reason why discrepancy is of interest to Monte Carlo sampling is perhaps best exemplified through the Koksma–Hlawka inequality which bounds the Monte Carlo error by a constant (depending on the integrand) times the discrepancy of the sample used. Hence the discrepancy alone leads the asymptotic error rate, hence the obvious aim of seeking lowest discrepancy. The best methods produce discrepancy of order $\mathcal{O}(N^{-1+\epsilon})$ for any $\epsilon > 0$, which is of course better than the Monte Carlo rate $\mathcal{O}_P(N^{-1/2})$.

In order to utilize the potential of QMC in the context of SMC, Gerber & Chopin introduce a **sequential quasi Monte Carlo** (SQMC) methodology. This assumes the existence of transforms Γ_t mapping uniform random variables to the state variables. More specifically, it is assumed that (1) can be rewritten as $\mathbf{x}_0^{(n)} = \Gamma_0(\mathbf{u}_0^{(n)})$ to generate $\mathbf{x}_0^{(n)} \sim f_0^X(\mathrm{d}\mathbf{x}_0^{(n)})$, and as $\mathbf{x}_{1:t}^{(n)} = \Gamma_t(\mathbf{x}_{1:t-1}^{(n)}, \mathbf{u}_t^{(n)})$ to generate $\mathbf{x}_{1:t}^{(n)}|\mathbf{x}_{1:t-1}^{(n)} \sim f_t^X(\mathrm{d}\mathbf{x}_{1:t}^{(n)}|\mathbf{x}_{1:t-1}^{(n)})$, where $\mathbf{u}_t^{(n)} \sim \mathcal{U}([0,1)^d)$. We show in the next section that such a deterministic function Γ_t that is easy to evaluate is available in the context of Dirichlet process mixture models, however at the cost of Markovianity.

3 Sequential quasi Monte Carlo for Infinite Mixture Models

Nonparametric mixtures are commonly used for density estimation and can be thought of as an extension of finite mixture models when the number of clusters is unknown. Observations $\mathbf{y}_{1:T}$ follow a DPM model with kernel ψ, ie $y \to \psi(y;\theta)$ are probability density functions for any θ in a parameter set Θ,

$$\mathbf{y}_t | G \overset{\text{i.i.d.}}{\sim} \int \psi(y;\theta) \mathrm{d}G(\theta), \quad t \in (1:T),$$

where G is endowed with a Dirichlet process prior $\mathrm{DP}(\alpha, G_0)$ with precision parameter α and base measure G_0. The discrete structure of the Dirichlet process induces partitions on the observations $\mathbf{y}_{1:t}$ into clusters identified by the cluster labels $\mathbf{x}_t$, or allocation variables. They link each observation $\mathbf{y}_t$ to a mixture component $\theta^*_{\mathbf{x}_t}$. For any t, $\mathbf{y}_{1:t}$ are clustered into a set of $k_t \leq t$ clusters of sizes $n_{t,j}$, for $j \in (1:k_t)$. The allocation variables can be interpreted as the *states* of $\mathbf{y}_{1:t}$ in the context of SMC. Note that they are *discrete* and elements of $\{1,\dots,k_t\}$. The *generalized Pólya urn scheme* (GPUS) is essentially a sequential formulation of this model: it describes the joint distribution of the states $\mathbf{x}_{1:t}$ by its sequence of *prior predictive distributions* $\tilde{p}_{t,j} = \mathrm{P}(\mathbf{x}_t = j | \mathbf{x}_{1:t-1})$ for $t \in (1:T)$. Interestingly, the *posterior predictive distribution* also takes the form of a GPUS $p_{t,j} = \mathrm{P}(\mathbf{x}_t = j | \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t})$. Both prior and posterior predictives, as well as details on the model, are provided in Appendix, Section A.

Our SMC sampler follows similar lines as Fearnhead (2004). The specific resampling technique, known to be unavoidable in sequential samplers, is detailed in Appendix, Section B. As stressed in Section 2, the key ingredient of SQMC is the availability of a deterministic transform Γ_t to sample $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$ with the use of an additional *quasi uniform variable*. Here, such a function is available, however at the cost of Markovianity, thus using $\mathbf{x}_{1:t-1}$ instead of $\mathbf{x}_{t-1}$. This transform utilizes the *posterior predictive distribution* of Equation 3 and takes the following form in our setting

$$\Gamma_t(\mathbf{x}^{(n)}_{1:t-1}, \mathbf{u}^{(n)}_t) = \min\left\{ j \in \{1,\dots,k^{(n)}_{t-1}+1\} : \sum_{i=1}^{j} p^{(n)}_{t,i} > \mathbf{u}^{(n)}_t \right\}$$

for any particle n, where $p^{(n)}_{t,i}$ is the associated GPUS weight (3) and with $\mathbf{u}_t \sim \mathcal{U}([0,1))$ (see Arbel and Prünster, 2015). Although the particles state-space dimension is growing with t, notice that this second argument of the Γ_t transform remains a *univariate* uniform random variable. This is of crucial importance for our methodology since SQMC samplers are known to break for large dimensional $\mathbf{u}_t$ (Gerber and Chopin, 2015). Thus, Γ_t relies on quasi uniform vectors $\mathbf{u}^{(1:N)}_t \in [0,1)^N$. Generating such a vector is straightforward, for instance by adding a random uniform variable to a regular grid of $[0,1)$ modulo 1, such as $\mathbf{u}^{(1:N)}_t = u + (0, 1/N, \dots, (N-1)/N) \bmod 1$, with $u \sim \mathcal{U}[0,1)$. This is also faster than generating N uniform random variables required in the SMC setting, since it only needs one such variable.

4 Impact of quasi Monte Carlo on diversity of the allocation variables

We compare some properties of the allocation variables trajectories $\mathbf{x}^{(n)}_{1:T}$, $n = 1,\dots,N$ of DP normal mixture models under three samplers: the non sequential Monte Carlo method (Gibbs sampler) of Neal (2000) (MC), a sequential Monte Carlo method of Griffin (2015) (SMC) and our proposed sequential quasi Monte Carlo method (SQMC)[3]. We run the three samplers on three simulated datasets of size $T = 200$, respectively with heavy-tailed, skewed, and multimodal distributions (see densities on Figure 1). The number of iterations for MC (after a burn-in period) and the number of particles in both sequential samplers is set to the same value of $N = 1000$.

For the three distributions, the posterior mean obtained for the three methods have equivalent fit (see left panel of Figure 1). However, we observe that the SMC method does not satisfactorily explore the allocation space. In particular, we observe that a single path for the allocation vector $\mathbf{x}_{1:t}$ is strongly favored by this approach up to some $t \leq T$. Using quasi random sequences is a simple way to get around this problem by insuring at each step a good spread of the random sequence. In practice, this

[3] The code is available upon request.

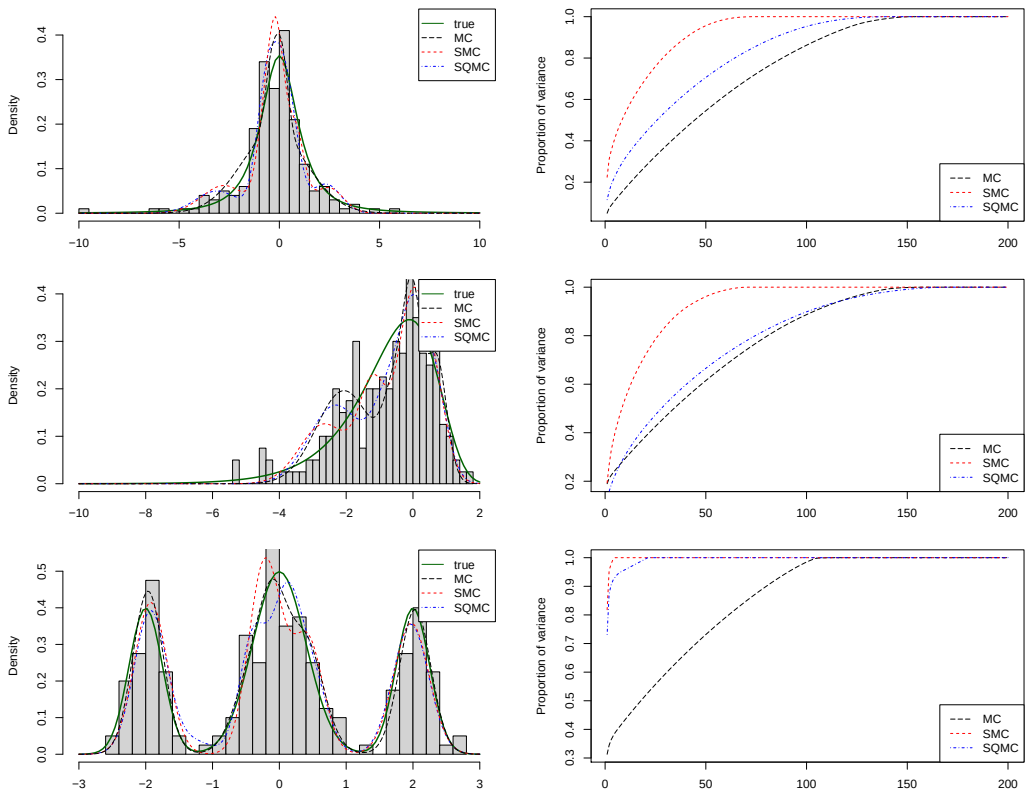

Fig. 1: Left: Density fit; Right: Diversity index (PCA) for three samplers, non sequential Monte Carlo (MC), sequential Monte Carlo (SMC) and sequential quasi Monte Carlo (SQMC). Top row: Heavy tailed distribution (student with degree of freedom two); Middle row: Skewed distribution (log-Gamma); Bottom row: Multimodal distribution with three well separated modes (mixture of normals).

makes a significant difference and we observe allocations $\mathbf{x}_{1:T}^{(n)}$ that present closer behavior in terms of dispersion to the one obtained with non sequential method. To measure that dispersion, we run a principal component analysis (PCA) on the allocation vector $\mathbf{x}_{1:T}^{(n)}$, $n = 1, \ldots, N$. The PCA gives information on how different the trajectories are from one another. We represent on the right panel of Figure 1 the proportion of variance explained by number of components in the PCA. For instance, if the samples have similar allocation vectors, then most of the variance will be on few principal components, yielding a high curve.

The first two examples, heavy-tailed and skewed distributions, cannot be written straightforwardly as a mixture of normal kernels. Thus in these cases, we do not expect a clear clustering but some heterogeneity in the sampled allocation vectors. On the contrary, the third example is exactly a mixture of normal kernels, so there should be much less variability in the allocation vector here. From the experiments, we observe that the diversity of the allocation trajectories under SQMC behaves as one would expect: it is high in the first two examples, and moderate in the third. This feature is not as clear for SMC. Additionally, having more diversity in the allocation vector is a sign of a better exploration of the latent space. This diversity for SQMC is always higher than the one for SMC, and is closer to the one obtained from a non sequential approach (MC), indicating that SQMC outperforms SMC in terms of latent space exploration.

5 Discussion

We introduce sequential quasi Monte Carlo sampling for DPM models, and discuss the applicability of the proposed methodology to a broader class of BNP mixture models. For these models, the structure of the latent space is non standard for sequential methods. We show that favoring diversity by using quasi random sequences improves the exploration of the allocation space.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Arbel, J. and Prünster, I. (2015). Discussion of "Sequential Quasi-Monte Carlo" by Gerber and Chopin. *Journal of the Royal Statistical Society. Series B*, 77:559–560.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144.

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2008). Bayesian Inference for Linear Dynamic Models With Dirichlet Process Mixtures. *IEEE Trans. Signal Process.*, 56(1):71–84.

Carvalho, C. M., Lopes, H. F., Polson, N. G., and Taddy, M. A. (2010). Particle learning for general mixtures. *Bayesian Analysis*, 5(4):709–740.

Chen, R. and Liu, J. S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

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

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Gerber, M. and Chopin, N. (2015). Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET.

Griffin, J. E. (2015). Sequential Monte Carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing, in press*.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputation. *The Annals of Statistics*, 24:910–930.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, 27(2):251–267.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585.

Tsiligkaridis, T. and Forsythe, K. (2015). Adaptive Low-Complexity Sequential Inference for Dirichlet Process Mixture Models. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 28–36. Curran Associates, Inc.

Ülker, Y., Günsel, B., and Cemgil, A. T. (2010). Sequential Monte Carlo samplers for Dirichlet process mixtures. In *International Conference on Artificial Intelligence and Statistics*, pages 876–883.

A Generalized Pólya urn schemes for infinite mixture models

The *generalized Pólya urn scheme* (GPUS) is essentially a sequential formulation of this model: it describes the joint distribution of the states $\mathbf{x}_{1:t}$ by its sequence of *prior predictive distributions* $\tilde{p}_{t,j} = \mathrm{P}(\mathbf{x}_t = j|\mathbf{x}_{1:t-1})$ for $t \in (1:T)$. For a Dirichlet process prior $DP(\alpha, G_0)$ with precision parameter α and base measure G_0

$$\tilde{p}_{t,j} \propto \begin{cases} n_{t-1,t} \text{ for } j \in (1:k_{t-1}) \\ \alpha \text{ for } j = k_{t-1}+1. \end{cases} \tag{2}$$

The *posterior predictive distribution* also takes the form of a GPUS $p_{t,j} = \mathrm{P}(\mathbf{x}_t = j|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:t})$ with

$$p_{t,j} \propto \begin{cases} n_{t-1,j}\psi_j(\mathbf{y}_t|\mathbf{x}_{1:t-1}) \text{ for } j \in (1:k_{t-1}) \\ \alpha\psi_0(\mathbf{y}_t) \text{ for } j = k_{t-1}+1 \end{cases} \tag{3}$$

where

$$\psi_j(\mathbf{y}_t|\mathbf{x}_{1:t-1}) = \frac{\int \psi(\mathbf{y}_t|\theta) \prod_{l \in \mathbf{s}_{t-1,j}} \psi(\mathbf{y}_l|\theta) G_0(\mathrm{d}\theta)}{\int \prod_{l \in \mathbf{s}_{t-1,j}} \psi(\mathbf{y}_l|\theta) G_0(\mathrm{d}\theta)}$$

and

$$\psi_0(\mathbf{y}_t) = \int \psi(\mathbf{y}_t|\theta) G_0(\mathrm{d}\theta)$$

where one denotes the allocation sets by $\mathbf{s}_{t,j} = \{l \leq t : \mathbf{x}_l = j\}$. The weights $\psi_j(\mathbf{y}_t|\mathbf{x}_{1:t-1})$ and $\psi_0(\mathbf{y}_t)$ can be calculated in conjugate cases and approximated otherwise. Note that posterior predictive distributions in the form of GPUS are also available for broader classes of random probability measures including the two-parameter Poisson–Dirichlet process (Pitman and Yor, 1997) and normalized random measures with independent increments (Regazzini et al., 2003; James et al., 2009). Algorithms are detailed by Griffin (2015).

The most widespread setting (Escobar and West, 1995; Fearnhead, 2004), that we also pursue here, considers Gaussian mixtures with a Gaussian base measure G_0. More specifically, we consider location scale mixtures where the hierarchical representation of the model assumes that observations in each cluster are draw iid from Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$. The conjugate prior for $\theta = (\mu, s)$ where $s = \sigma^{-2}$ has the conditional representation $s \sim \mathrm{Gamma}(a, b)$ and $\mu|s \sim \mathcal{N}(\eta, \tau/s)$, with fixed hyperparameters a, b, η and τ. Posterior distributions of μ_j and s_j for cluster j parameters, as well as the posterior probability of the states $\mathbf{x}_{1:t}$ are given by Fearnhead (2004).

Marginalising out the location scale parameter θ has been shown to improve substantially the efficiency of sequential samplers (Chen and Liu, 2000). Indeed, it allows us to consider only the allocations $\mathbf{x}_{1:t}$ as the states, thus removing the part on θ which would make our QMC methodology more cumbersome.

B Resampling algorithm

A typical feature of sequential methods is that the weights $\{W_t^{(n)}\}$ can become very skewed as t increases, ie most of the particles weights tend to zero while only some remain active, which

leads to an increasing variance and so-called particle degeneracy. A popular criterion for measuring degeneracy is the *effective sample size* (ESS) defined by

$$\left(\sum_{n=1}^{N} \left(W_t^{(n)}\right)^2\right)^{-1}.$$

which takes on values in $(1, N)$, and it is common practice to resample the particles when the ESS falls below a threshold. Once resampled, the particles are all given the same weight $(1/N)$. It is common to require the resampling to be *unbiased*, meaning that the expected value of any function of the particles is unchanged by the resampling. The typical resampling technique consists in sampling particles from $\mathbf{x}_{1:t}^{(1:N)}$ using the multinomial distribution $\text{Multinomial}\big(N; W_t^{(1:N)}\big)$, which leads to duplication of particle values with positive probability (see for instance Gordon et al., 1993). In the case of discrete (finite) state-space, specific resampling schemes have been tailored in order to take advantage of the finiteness of the states with the aim of reducing wasteful particle duplication. One defines as 'putative particles' the set of $M_t = \sum_{n=1}^{N} k_{t-1}^{(n)} + 1$ possible particles at step t. Computing their weights is feasible, and sampling from them instead of from the $N < M_t$ particles of step $t-1$ improves efficiency from a Rao-Blackwell point of view. However, keeping track of all possible trajectories till time t is a task beyond computational capabilities even for limited t, since the trajectories number is equal to the number of partitions of a set of size t which explodes essentially as t^t. Hence resampling appears to be unavoidable. The approach that we follow consists in the two steps

- Computing the normalized weights of the M_t putative particles, $\big(\mathbf{x}_{1:t-1}^{(n)}, j\big)$, for $n \in (1:N)$ and $j \in (1:(k_{t-1}^{(n)}+1))$, by

$$W_{t,j}^{(n)} \propto W_{t-1}^{(n)} \frac{p\big(\mathbf{x}_{1:t-1}^{(n)}, j | \mathbf{y}_{1:t}^{(n)}\big)}{p\big(\mathbf{x}_{1:t-1}^{(n)} | \mathbf{y}_{1:t-1}^{(n)}\big)}$$

  where the ratio in the right hand side, also known as the incremental weight, can be evaluated up to a normalizing constant in conjugate models.
- Choosing N among the M_t by the multinomial

$$\text{Multinomial}\big(N; \big(W_{t,j}^{(n)}, n \in (1:N),\, j \in (1:(k_{t-1}^{(n)}+1))\big)\big).$$