ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2015, 1-10
Editors: Will be set by the publisher

ON SOME RECENT ADVANCES ON HIGH DIMENSIONAL BAYESIAN STATISTICS

Nicolas Chopin[1], Sébastien Gadat[2], Benjamin Guedj[3], Arnaud Guyader[4] and Elodie Vernet[5]

Abstract. This paper proposes to review some recent developments in Bayesian statistics for high dimensional data. After giving some brief motivations in a short introduction, we describe new advances in the understanding of Bayes posterior computation as well as theoretical contributions in non parametric and high dimensional Bayesian approaches. From an applied point of view, we describe the so-called SQMC particle method to compute posterior Bayesian law, and provide a nonparametric analysis of the widespread ABC method. On the theoretical side, we describe some recent advances in Bayesian consistency for a nonparametric hidden Markov model as well as new PAC-Bayesian results for different models of high dimensional regression.

Résumé. Nous proposons dans cet article une vue d'ensemble de récents développements en statistique bayésiennes en grande dimension. Après quelques motivations rappelées en introduction, nous présentons des avancées à la fois algorithmiques et dans la compréhension théorique de méthodes de calculs d'*a posteriori* bayésiens. En particulier, nous décrivons l'algorithme particulaire SQMC et proposons un point de vue non-paramétrique sur la méthode populaire ABC. Nous revenons ensuite également sur des contributions nouvelles en statistiques bayésiennes non paramétriques et en grandes dimensions. Dans ce contexte, nous décrivons des résultats de consistance bayésienne *a posteriori* pour des modèles non-paramétriques de Markov cachés ainsi que des résultats PAC-bayésiens pour différents modèles de régression.

1. Introduction

The analysis of Bayesian methods for high dimensional and non parametric models are at the cornerstone of some new statistical developments. Bayesian methods are tempting owing to their great generality and ability to incorporate in the statistical approach a belief of what should be the unknown quantity to be estimated (for example). It is also useful for producing efficient estimators or confidence set. It has recently attracted a lot of attention thanks to the availability of massive computational resources: in the 2000s, Bayesian works have been developed to deal with very high dimensional or even non parametric problems. Very concrete applications in biostatistics and signal processing (among others) have raised new legitimate questions that mainly concern two important points. The first one asks how should be a "good" Bayesian prior for high dimensional or

[1] ENSAE, Malakoff, France, Nicolas.Chopin@ensae.fr

[2] Toulouse School of Economics (Université Toulouse I Capitole), Toulouse, France, Sebastien.Gadat@math.univ-toulouse.fr

[3] MODAL project-team, Inria Lille - Nord Europe, Villeneuve d'Ascq, France, Benjamin.Guedj@inria.fr

[4] LSTA, Université Pierre et Marie Curie, Paris, France & Projet ASPI, Inria, Rennes, France, Arnaud.Guyader@upmc.fr

[5] Université Paris-Sud, Orsay, France, Elodie.Vernet@math.u-psud.fr

non parametric statistical model and what kind of theoretical results on the posterior distribution could we expect when the number of observations increases? The second imperative question is how to produce efficient algorithms to compute the posterior distribution and, if possible, quantify the way these numerical methods approximate this posterior distribution.

1.1. Bayes approaches

In what follows, we consider a dominated model parameterized by a set of measurable parameters $\mathbf{\Theta}$. We assume that $\mathbf{\Theta}$ is included in a metric space and each parameter $\boldsymbol{\theta} \in \mathbf{\Theta}$ defines a conditional probability distribution $\mathbb{P}(.|\theta)$. As a dominated model, all the previous laws $\mathbb{P}(.|\boldsymbol{\theta})$ are absolutely continuous with respect to a common measure denoted λ, whose density is referred to as $f(.|\boldsymbol{\theta})$.

A Bayesian prior π on $\mathbf{\Theta}$ is an initial distribution on $\mathbf{\Theta}$ that traduces a belief on the distribution of an unobserved parameter $\boldsymbol{\theta}$ living on $\mathbf{\Theta}$. We are then interested in statistical inference on $\mathbf{\Theta}$ (or in a quantity related to a distribution on $\mathbf{\Theta}$) when observing an i.i.d. sample of size n, denoted $\mathbf{y_n} := (Y_1, \dots, Y_n)$ in the sequel. A key ingredient for the analysis of the Bayesian procedures is the likelihood ratio of the sample, written as $\ell(\mathbf{y_n}|\theta)$ that satisfies $\mathbb{P}(d\mathbf{y_n}|\theta) = \ell(\mathbf{y_n}|\theta)\lambda(d\mathbf{y_n})$. This likelihood ratio is important since it permits, at least from a mathematical point of view, to compute the *posterior distribution* built using the *prior distribution* and the famous Bayes' rule:

$$\pi(\boldsymbol{\theta}|\mathbf{y_n}) = \frac{\pi(\boldsymbol{\theta})\ell(\mathbf{y_n}|\boldsymbol{\theta})}{\int_{\mathbf{\Theta}} \pi(\boldsymbol{\vartheta})\ell(\mathbf{y_n}|\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}}. \tag{1}$$

1.2. Difficulties involved in Bayesian high dimensional statistics

If the situation is clear for regular low dimensional models (see [LCY00] for a detailed introduction), Bayesian statistics raise computational and theoretical difficulties in higher ones.

1.2.1. *Posterior computation*

In order to obtain a Bayesian estimator generically given by $\mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{y_n}]$, the standard approach is to do a Monte Carlo procedure to roughly approach the former expectation: one simulates several independent values $\boldsymbol{\theta}^k \sim \pi(\boldsymbol{\theta}|\mathbf{y_n})$, making k varying between 1 and K, and compute the empirical averages, e.g.

$$\frac{1}{K}\sum_{k=1}^{K} \varphi(\boldsymbol{\theta}^k)$$

as an approximation of $\mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{y_n}] = \int_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y_n})\, d\boldsymbol{\theta}$.

A practical difficulty with this approach is that it relies on the approximation of the posterior distribution, and in most cases the denominator in (1) *is an intractable* integral. Fortunately, standard MCMC (Markov chain Monte Carlo) algorithms used to simulate from $\pi(\boldsymbol{\theta}|\mathbf{y_n})$ require evaluating the posterior density *only up to a constant,* and therefore do not require to evaluate this intractable integral. For instance, Algorithm 1 describes one step of a Gaussian Random Walk Hastings-Metropolis (RWHM) algorithm, that is, an algorithm for simulating a Markov chain that leaves invariant $\pi(\boldsymbol{\theta}|\mathbf{y_n})$, using the following proposal mechanism (assuming $\Theta = \mathbb{R}^d$): from a current point $\boldsymbol{\theta}^k$, propose new point $\boldsymbol{\theta}^\star \sim N(\boldsymbol{\theta}^k, \Sigma)$, (a random walk move), and accept/reject according to (informally) how more compatible is the proposed point to the posterior, relative to the current point. One sees that Algorithm 1 does not require evaluating the denominator of (1).

Algorithm 1 is just a simple example of possible practical approaches to Bayesian computation and various methods exist for the inference of $\boldsymbol{\theta}_0$ in this context, such as rejection algorithms [Rip06], Markov Chain Monte Carlo (MCMC) methods (e.g., the Metropolis-Hastings algorithm [MRR+53, Has70]), and Importance Sampling [Rip06]. For a comprehensive introduction to the domain, the reader is referred to the monograph [RC04].

However, in some contexts, computation of the posterior is problematic, either because the size of the data makes the calculation computationally intractable, or because calculation is impossible when using realistic

Algorithm 1 (Gaussian) Random Walk Hastings-Metropolis (RWHM) algorithm

```
Input: $\boldsymbol{\theta}^k$, $\Sigma$ (resp. a point in $\mathbb{R}^d$, and a $d \times d$ symmetric positive matrix)
Output: $\boldsymbol{\theta}^{k+1}$ (a vector in $\mathbb{R}^d$).
1: Simulate $\boldsymbol{\theta}^\star \sim N(\boldsymbol{\theta}^k, \Sigma)$.
2: With probability $1 \wedge r$ where
       $$r = \frac{\pi(\boldsymbol{\theta}^\star)\ell(\mathbf{y_n}|\boldsymbol{\theta}^\star)}{\pi(\boldsymbol{\theta}^k)\ell(\mathbf{y_n}|\boldsymbol{\theta}^k)}$$
   take $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^\star$; otherwise keep the parameter unchanged: $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k$.
```

models for how the data arises (in particular non parametric ones). Thus, despite their power and flexibility, MCMC procedures and their variants may prove irrelevant in contemporary applications involving very large dimensions or complicated models. Section 2 presents two efficient methods to handle the posterior distribution when the dimension of observation is large.

1.2.2. *Frequentist consistency of Bayesian procedures*

As already discussed above, the choice of the prior is a key issue in Bayesian statistics and is at the core of Bayesian consistency by adopting a frequentist point of view. A straightforward question is the impact of the prior π on the posterior $\pi(\cdot|\mathbf{y_n})$. That is to say, does the prior still play a role in the posterior when the number of observations increases or does it "disappear" in favor of the observations? If another prior is chosen, will the posterior be approximately the same at least when the number of observations is infinite? An answer to this question is given by the concept of posterior consistency. Studying posterior consistency implies taking a frequentist point of view, assuming that the observations come from a real parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, *i.e.*,

$$\mathbf{y_n} = (Y_1, \dots, Y_n) \text{ are distributed from } \mathbb{P}(.|\boldsymbol{\theta}_0)$$

and wondering if the posterior concentrates its mass around $\boldsymbol{\theta}_0$ when the number of observations increases.

Definition 1.1 (Consistency)**.** The posterior $\pi(\cdot|\mathbf{y_n})$ is consistent at $\boldsymbol{\theta}_0$ if for all neighborhood U of $\boldsymbol{\theta}_0$:

$$\mathbb{P}(.|\boldsymbol{\theta}_0)\text{-a.s.,} \quad \pi(U|\mathbf{y_n}) \longrightarrow 1 \quad \text{as} \quad n \longrightarrow +\infty.$$

Note that in the Bayesian literature, a neighborhood U of $\boldsymbol{\theta}$ is to be understood with respect to a topology on $(\mathbb{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$, for example the one derived from the Kullback-Leibler divergence or the weak topology associated to a countable set of bounded continuous functions. Posterior consistency may be seen as a frequentist validation of Bayesian statistics. It also ensures robustness of the posterior considering two different priors see [GR03].

The first historical answer to such a type of question is given by [Doo49]: in a very general setting, when the observations are i.i.d. and the model is identifiable, the posterior is consistent at π-almost every $\boldsymbol{\theta}_0$. The exact set of true parameters at which the posterior is consistent is not specified in this theorem and it may be topologically small. In particular, [Fre65] stated that in the nonparametric case (which is an extreme case of large dimensional setting), the couples $(\pi, \boldsymbol{\theta}_0)$ for which the posterior is consistent is very small topologically (meager). This negative result is not a reason to give up nonparametric or high dimensional Bayesian statistics: on the contrary it is a clear invitation to a careful choice of a good prior to resolve a given statistical problem.

In particular, a good choice of the prior is the price to pay to obtain some efficient behaviour of the posterior distribution in high (or even infinite) dimensional models. Finally, we can remark that Bayesian consistency is generally obtained in an asymptotic setting $n \longrightarrow +\infty$. Less is known when one is looking for a finite horizon result. One of the purposes of the PAC-Bayesian approach consists in obtaining such non asymptotic guarantees, allowing in particular to derive risk bounds for Bayesian estimators, with arbitrarily high probability (hence the acronym ***P****robably* ***A****pproximately* ***C****orrect*) in a finite horizon. Moreover, it can be shown to be efficient in large dimensional models as soon as the underlying prior is carefully chosen.

Bayesian practitioners are commonly faced with both computational and statistical issues. It is of theoretical interest to study and derive upper-bounds on the divergence between the posterior and the true distribution. Such theoretical guarantees are useful to assess the performance of Bayesian approaches. Nonetheless, from a practical perspective these bounds must be balanced with their computational counterparts, and a compromise has to be found between the statistical performance and the approximation accuracy, especially in the popular context of high dimensional statistics. Note that the study of such compromises is not explicitly addressed in this paper, and is a very active field of research (among others, see [Wai14]).

Below, we review some recent advances in Bayesian statistics in high dimensional or nonparametric situations. Section 2.3 describes a sequencial approximation algorithm of posterior distribution sampling that covers a particular case of hidden Markov models (HMM for short). In such a case, the likelihood is usually intractable and we provide an efficient way to get round of such difficulty by using a sequential quasi-Monte Carlo algorithm. Section 3 discusses on ABC algorithms and offers a nonparametric point of view for understanding the behaviour of estimators computed from ABC algorithms. In Section 4, some recent results taken from [Ver15] on posterior consistency for HMM are presented. We end the paper with Section 5, which aims at showing that the PAC-Bayesian approach adapts neatly to the high dimensional context when coupled with a suitably chosen sparsity-inducing prior.

2. Bayesian computation with SQMC

2.1. Limitations of standard RWHM

A miminal requirement to apply Algorithm 1 (and many other similar methods) is the possibility to evaluate the likelihood $\ell(\mathbf{y_n}|\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Unfortunately, there are various important cases where the likelihood itself cannot be exactly computed :

(1) Because the likelihood is an intractable integral: $\ell(\mathbf{y_n}|\boldsymbol{\theta}) = \int \ell(\mathbf{x}, \mathbf{y_n}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}$. Typically, $\mathbf{x}$ is interpreted as a latent variable in this formulation. Examples include hidden Markov models (also covered in Section 4), phylogenetic models (where $\mathbf{x}$ is a phylogeny tree, see *e.g.* [Bea10]), and more generally any model based on latent variables.

(2) Because the likelihood is *un-normalised*: $\ell(\mathbf{y_n}|\boldsymbol{\theta}) = g_{\boldsymbol{\theta}}(\mathbf{y_n})/Z(\boldsymbol{\theta})$, with $Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\theta}} g_{\boldsymbol{\theta}}(\mathbf{y_n}')\,\mathrm{d}\mathbf{y_n}'$ being intractable. Examples include Ising models, networks models [Eve12, CF13], models for point processes [GZ01], among others.

An important supplementary limitation of Algorithm 1 in a large dimensional framework is due to the very low acceptation rate produced by the MCMC dynamics. Indeed, the most of the candidates sampled at step 1 lead to a very weak acceptation ratio r and are likely to be rejected by Algorithm 1 at step 2. This problematic situation is an invitation to develop alternative methods that do not rely on acceptation/rejection scheme.

2.2. Hidden Markov models

Hidden Markov models (HMMs), also known as state-space Markov models, have been widely used in diverse fields such as speech recognition, genomics, econometrics since their introduction in [BP66]. The books [MZ97] and [CMR05] provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. HMMs are stochastic processes $(\mathbf{x}_t, \mathbf{y}_t)_{t\in\mathbb{N}}$ such that

(a) $(\mathbf{x}_t)_{t\geq 0}$ is an unobserved Markov chain,

(b) the observations $\mathbf{y}_t$'s are conditionally independent, given the $\mathbf{x}_t$'s.

The name “hidden Markov model” comes from the fact that we only observe the $\mathbf{y}_t$ component of the process and we cannot access the states $(\mathbf{x}_t)_{t\in\mathbb{N}}$ of the Markov chain. One way to fully specify such a model is as follows: $\mathbf{x}_0 \sim f^0(\mathbf{x}_0)$, and

$$\begin{aligned} \mathbf{x}_t|\mathbf{x}_{t-1} &\sim f^X(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad t\geq 1 \\ \mathbf{y}_t|\mathbf{x}_t &\sim f^Y(\mathbf{y}_t|\mathbf{x}_t), \quad t\geq 0 \end{aligned}$$

where f^0, f^X and f^Y are conditional probability densities with respect to appropriate dominating measures; in this paper, we simply assume that $\mathbf{x}_t$ take values in $\mathbb{R}^d$, and $\mathbf{y}_{\mathbf{n}_t}$ in some probability space $\mathcal{Y}$.

One may assume in addition that f^0, f^X and f^Y depend on a fixed parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $f^0_{\boldsymbol{\theta}}$, $f^X_{\boldsymbol{\theta}}$ and $f^Y_{\boldsymbol{\theta}}$, leading to the likelihood function, for data $\mathbf{y} = \mathbf{y}_{0:T}$ observed up to final time T,

$$\ell(\mathbf{y}_{0:T}|\boldsymbol{\theta}) = \int_{\mathbb{R}^{(T+1)d}} f^0_{\boldsymbol{\theta}}(\mathbf{x}_0) \prod_{t=1}^{T} f^X_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=0}^{T} f^Y_{\boldsymbol{\theta}}(\mathbf{y}_t|\mathbf{x}_t)\, \mathrm{d}\mathbf{x}_{0:T}$$

which is an integral of often very large dimension. Except in specific cases (i.e. when the state space is finite; or when the model is linear and Gaussian), this likelihood cannot be computed exactly, and some Monte Carlo integration is needed to obtain a correct approximation of this integral. For notational convenience, we omit the dependence in $\boldsymbol{\theta}$ in what follows.

2.3. Sequential Quasi-Monte Carlo and its application to hidden Markov models

2.3.1. *Particle filtering*

Particle filtering algorithms provide some very efficient methods to sample from a posterior distribution even when this distribution seems very hard to compute. A pseudo-code is given in Algorithm 2 that describes the simplest particle filtering algorithm (known as the bootstrap filter).

Note that the only requirements to implement Algorithm 2 are (i) to be able to compute $f^Y(\mathbf{y}_t|\mathbf{x}_t)$ for any $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{X} \times \mathcal{Y}$; and (ii) to be able to sample $\mathbf{x}_0 \sim f^X(dx_0)$, $\mathbf{x}_t|\mathbf{x}_{t-1} = x_{t-1} \sim f^X(dx_t|x_{t-1})$. It does not require to be able to compute the probability density $f^X(x_t|x_{t-1})$ pointwise. Thus we may apply Algorithm 2 to HMMs with a Markov process $(\mathbf{x}_t)$ described in terms of some complex algorithm for simulating from $f^X(dx_t|x_{t-1})$, and such that the density $f^X(x_t|x_{t-1})$ is intractable.

Algorithm 2 Particle filter

```
Input: $N$ (number of particles), data $\mathbf{y}_0,\dots,\mathbf{y}_T \in \mathcal{Y}$.
Operations must be performed for all $n = 1,\dots,N$.
At time 0,
    (a$_0$): Sample $\mathbf{x}_0^n \sim f^0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0$.
    (b$_0$): Compute weights $w_0^n = f^Y(\mathbf{y}_0|\mathbf{x}_0^n)$, normalised weights $W_0^n = w_0^n / \sum_{m=1}^N w_0^m$, and
        $L_0^N = \left\{N^{-1}\sum_{n=1}^N w_0^n\right\}$.
Recursively, from time $t = 1$ to time $t = T$,
    (a$_t$): Sample $a_t^1,\dots,a_t^N \in \{1,\dots,N\}$ in such a way that $\mathbb{E}\left[\sum_{m=1}^N \mathbb{I}\left(a_t^m = n\right)\right] = NW_{t-1}^n$ for all
        $n \in \{1,\dots,N\}$.
    (b$_t$): Sample $\mathbf{x}_t^n \sim f^X(\mathbf{x}_t|\mathbf{x}_{t-1}^{a_t^n})\mathrm{d}\mathbf{x}_t$.
    (c$_t$): Compute weights $w_t^n = f^Y(\mathbf{y}_t|\mathbf{x}_t^n)$, normalised weights $W_t^n = w_t^n / \sum_{m=1}^N w_t^m$, and
        $L_t^N = L_{t-1}^N\left\{N^{-1}\sum_{n=1}^N w_t^n\right\}$.
```

Let us briefly explain the construction of this algorithm. At any time t, we aim to build a discrete distribution $(x_t^n, W_t^n)_{1\le n\le N}$ that approximates the true filtering distribution. It provides some typical samples $(x_t^n)_{1\le n\le N}$ associated to a suitable sequence of weights $(W_t^n)_{1\le n\le N}$ such that the filtering distribution satisfies

$$\sum_{n=1}^{N} W_t^n \varphi(\mathbf{x}_t^n) \approx \mathbb{E}\left[\varphi(\mathbf{x}_t)|\mathbf{y}_{0:t}\right]$$

for a given function $\varphi : \mathbb{R}^d \to \mathbb{R}$. In addition, the particle filter algorithm computes a quantity L_t^N that consistently approximate the likelihood $\ell(\mathbf{y}_{0:t})$ (as $N \to +\infty$).

A simple way to motivate Algorithm 2 is through (iterated) importance sampling.

- At time 0, we generate "particles" from $f^0(\mathrm{d}\mathbf{x}_0)$, and reweight them, with weights equal to $f^Y(\mathbf{y}_0|\mathbf{x}_0^n)$, so as to target the filtering distribution

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_0) &= \frac{f^0(\mathbf{x}_0)f^Y(\mathbf{y}_0|\mathbf{x}_0)}{\ell(\mathbf{y}_0)}, \quad \ell(\mathbf{y}_0) = \int_{\mathbb{R}^d} f^0(\mathbf{x}_0)f^Y(\mathbf{y}_0|\mathbf{x}_0)\,\mathrm{d}\mathbf{x}_0, \\ &\propto f^0(\mathbf{x}_0)f^Y(\mathbf{y}_0|\mathbf{x}_0). \end{aligned}$$

Note in particular that the average of the weights is an importance sampling estimator of $\ell(\mathbf{y}_0)$:

$$L_0^N = \frac{1}{N}\sum_{n=1}^N w_0^n = \frac{1}{N}\sum_{n=1}^N f^Y(\mathbf{y}_0|\mathbf{x}_0^n) \approx \int_{\mathbb{R}^d} f^Y(\mathbf{y}_0|\mathbf{x}_0)f^0(\mathbf{x}_0)\,\mathrm{d}\mathbf{x}_0.$$

- At time $t \geq 1$, we have from the previous iteration a weighted sample $(\mathbf{x}_{t-1}^n, W_{t-1}^n)_{n=1}^N$ that targets $p(\mathbf{x}_{t-1}|\mathbf{y}_{0:t-1})$. To progress from time $t-1$ to time t, we note that

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t-1}) = p(\mathbf{x}_{t-1}|\mathbf{y}_{0:t-1})f^X(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t}) = \frac{p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t-1})f^Y(\mathbf{y}_t|\mathbf{x}_t)}{\ell(\mathbf{y}_t|\mathbf{y}_{0:t-1})} \tag{3}$$

with $\ell(\mathbf{y}_t|\mathbf{y}_{0:t-1}) = \int p(\mathbf{x}_t|\mathbf{y}_{0:t-1})f^Y(\mathbf{y}_t|\mathbf{x}_t)\,\mathrm{d}\mathbf{x}_t$. Remark that (2) uses the fact $(\mathbf{x}_t)$ is Markov, and (3) is the simple Bayes formula. We then replace in (2) the term $p(\mathbf{x}_{t-1}|\mathbf{y}_{0:t-1})$ by the random probability measure obtained at step $t-1$:

$$\sum_{n=1}^N W_{t-1}^n \delta_{\mathbf{x}_{t-1}^n}(\mathrm{d}\mathbf{x}_{t-1}).$$

This is a mixture of N Dirac masses weighted according to the random weights W_{t-1}^n. We then update our approximation of $p(\mathbf{x}_{t-1}, \mathbf{x}_t|y_{0:t-1})$ as follows:

$$\sum_{n=1}^N W_{t-1}^n \left\{\delta_{\mathbf{x}_{t-1}^n}(\mathrm{d}\mathbf{x}_{t-1}) \times f^X(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t-1})\right\}.$$

This immediately suggests Step (a_t) and (b_t) in Algorithm 2: to sample from above, first (Step (a_t)) choose ancestor $\mathbf{x}_{t-1}^m$ with probability W_{t-1}^m; call a_t^n the so chosen m; then (Step (b_t)) sample $\mathbf{x}_t^n \sim f^X(\mathbf{x}_t|\mathbf{x}_{t-1}^{a_t^n})$. Finally, in line of (3), reweight the $\mathbf{x}_t^n$ by computing $w_t^n = f^Y(\mathbf{y}_t|\mathbf{x}_t)$ (Step (c_t)). Note in particular that the average of the weights approximate the conditional likelihood $\ell(\mathbf{y}_t|\mathbf{y}_{0:t-1}) = \ell(\mathbf{y}_{0:t})/\ell(\mathbf{y}_{0:t-1})$:

$$\frac{1}{N}\sum_{n=1}^N w_t^n = \frac{1}{N}\sum_{n=1}^N f^Y(\mathbf{y}_t|\mathbf{x}_t^n) \approx \int f^Y(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{0:t-1})\,\mathrm{d}\mathbf{x}_t.$$

In practice, one way to implement Step (a_t) of Algorithm 2, known as the resampling step, is first to generate N ordered uniform variables, $u^{(1)} \leq \ldots u^{(N)}$ (see e.g. p.214 of [Dev86] for a well-known method) and next to use Algorithm 3.

The SQMC (Sequential quasi-Monte Carlo) algorithm described below will be derived from this particular interpretation of particle filtering as a sequence of $T+1$ importance sampling steps (based on random probability measures).

Algorithm 3 Resampling

```
Require: $u^{1:N}$ (such that $0 \leq u^1 \leq \ldots \leq u^N \leq 1$), $W^{1:N}$ (normalised weights)
Ensure: $a^{1:N}$ (labels in $1:N$)
  $s \leftarrow W^1$, $m \leftarrow 1$
  for $n = 1 \to N$ do
    while $s < u^n$ do
      $m \leftarrow m + 1$
      $s \leftarrow s + W^m$
    end while
    $a^n \leftarrow m$
  end for
```

2.3.2. *Quasi-Monte Carlo*

QMC (Quasi-Monte Carlo) is usually introduced as a way to approximate an integral with respect to the unit hyper-cube of dimension d:

$$\int_{[0,1]^d} \varphi(\mathbf{u})\, \mathrm{d}\mathbf{u}.$$

The standard Monte Carlo approximation of this integral is

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(\mathbf{u}^n)$$

where the $\mathbf{u}^n$ are N independent samples from the uniform distribution $\mathcal{U}\left([0,1]^d\right)$. In QMC, the same estimator is used, but the major difference relies on the fact that the points $\mathbf{u}^n$ are generated from a low discrepancy sequence. Informally, it means that for certain subsets $A \subset [0,1]^d$, the proportion of $\mathbf{u}^n$ that fall in A is close to the volume of A; in fact closer that if the $\mathbf{u}^n$ were generated randomly. For instance, for $d = 1$, one may take $\mathbf{u}^n = n/(N+1)$. Of course when $d > 1$, one needs to use more advanced strategies. An exhaustive description of these more sophisticated methods is beyond the scope of this short survey (see *e.g.* the book of [Lem09]).

We simply mention a specific convergence result: under smoothness assumption on φ, a well chosen sequence $(\mathbf{u}^n)$ exists such that

$$\left| \frac{1}{N} \sum_{n=1}^{N} \varphi(\mathbf{u}^n) - \int_{[0,1]^d} \varphi(\mathbf{u})\, \mathrm{d}\mathbf{u} \right| \leq C \frac{(\log N)^d}{N}$$

This is of course a better convergence rate than standard Monte Carlo, which is $N^{-1/2}$.

2.3.3. SQMC *(Sequential Quasi-Monte Carlo): $d = 1$*

The main difficulty when introducing QMC into particle filtering methods (and more generally in any Monte Carlo approach) relies on the necessity to rewrite the algorithm as a deterministic function of uniform variables. When this is done, one may simply replace these uniform variables by low-discrepancy sequences, as we did in the previous section.

- Let's assume that, at time 0 in Algorithm 2, the $\mathbf{x}_0^n$ are generated as $\mathbf{x}_0^n = \Gamma_0(\mathbf{u}_0^n)$, with $\mathbf{u}_0^n \sim \mathcal{U}\left([0,1]^d\right)$, and Γ_0 a certain deterministic function chosen so that $\mathbf{x}_0^n \sim f^0$; for instance, the inverse CDF. Then, one may simply replace these $\mathbf{u}_0^n$ by points generated by a low-discrepancy sequence.
- Now, consider iteration $t \geq 1$. We have seen in Section 2.3.1 that iteration t may be interpreted as an importance sampling step, where we sample the $\mathbf{x}_t^n$'s from:

$$\sum_{n=1}^{N} W_{t-1}^n \left\{ \delta_{\mathbf{x}_{t-1}^n}(\mathrm{d}\mathbf{x}_{t-1}) \times f^X(\mathrm{d}\mathbf{x}_t | \mathbf{x}_{t-1}) \right\} \tag{4}$$

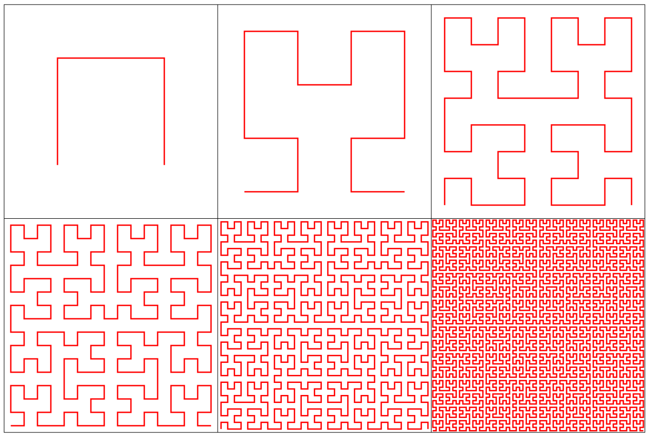

FIGURE 1. First steps of the iterative process, the limit of which is the Hilbert curve in two dimensions (Source: Wikipedia).

and reweight these new particles by $f^Y(\mathbf{y}_t|\mathbf{x}_t^n)$. Thus, we need to rewrite the simulation from (4) as a deterministic function of uniform variables. To do so, assume we have at our disposal a certain function Γ_t, such that simulating from $f^X(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t-1})$ amounts to computing $\mathbf{x}_t = \Gamma_t(\mathbf{x}_{t-1}, \mathbf{v}_t^n)$, when $\mathbf{v}_t^n \sim \mathcal{U}\left([0,1]^d\right)$.

This can be done as follows: for each $n = 1, \ldots, N$, let $\mathbf{u}_t^n \sim \mathcal{U}\left([0,1]^{d+1}\right)$, and denote $\mathbf{u}_t^n = (u_t^n, \mathbf{v}_t^n)$, with $u_t^n \in [0,1]$, $\mathbf{v}_t^n \in [0,1]^d$. Use the first component u_t^n to choose the ancestor $\mathbf{x}_{t-1}^n$, through the inverse CDF method. More precisely, (a) sort the ancestors in ascending order, i.e. find a permutation σ such that $\mathbf{x}_{t-1}^{\sigma(1)} \leq \ldots \leq \mathbf{x}_{t-1}^{\sigma(n)}$; then (b) find m such that

$$\sum_{p=1}^{m-1} W_{t-1}^{\sigma(p)} \leq u_t^n \leq \sum_{p=1}^{m} W_{t-1}^{\sigma(p)} \quad \text{(empty sum equals 0)}$$

and call a_t^n the obtained index, $a_t^n = m$. Now, to sample from $\mathbf{x}_t$ conditional on the ancestor, simply take $\mathbf{x}_t^n = \Gamma_t(\mathbf{x}_{t-1}^{a_t^n}, \mathbf{v}_t^n)$.

It is easy to see that, provided that $d = 1$ (i.e. we can indeed order the $\mathbf{x}_{t-1}^n$), the approach outlined above may be implemented in $\mathcal{O}(N \log N)$ time. If, in addition, we replace the $\mathbf{u}_t^n$'s by a low-discrepancy sequence in $[0,1]^{d+1}$, one obtains the SQMC algorithm, described in Algorithm 4.

2.3.4. SQMC *for* $d > 1$

Since the SQMC approach described in the previous section relies on the inverse CDF method, it is limited to situations where the state space is of dimension one, $d = 1$. It is nevertheless possible to extend this approach to $d > 1$, by using the Hilbert curve.

The Hilbert curve H is a continuous fractal space-filling curve, $H : [0,1] \to [0,1]^d$, with $H([0,1]) = [0,1]^d$. This curve is not a bijection, because the equation $H(x) = y$ may have more than one solution in x (for a fixed y); the set of such points y is of Lebesgue measure 0. In our framework, a crucial fact is that the function H admits however a pseudo-inverse $h : [0,1]^d \to [0,1]$, i.e. a function h such that $H(h(y)) = y$ for all $y \in [0,1]^d$. The function H is obtained as a limit of the iterative process depicted in Figure 1. We refer to the book of [Sag94] for more details on the properties of space-filling curves.

Algorithm 4 SQMC algorithm

```
At time $t = 0$,
    (a): Generate a QMC point set $\mathbf{u}_0^{1:N}$ in $[0,1]^d$, and compute $\mathbf{x}_0^n = \Gamma_0(\mathbf{u}_0^n)$ for each $n = 1, \dots, N$.
    (b): Compute $w_0^n = G_0(\mathbf{x}_0^n)$ and $W_0^n = w_0^n / \sum_{m=1}^N w_0^m$ for each $n = 1, \dots, N$.
Iteratively, from time $t = 1$ to time $t = T$,
    (a): Generate a QMC point set $\mathbf{u}_t^{1:N}$ in $[0,1]^{d+1}$; let $\mathbf{u}_t^n = (u_t^n, \mathbf{v}_t^n) \in [0,1] \times [0,1]^d$.
    (b): Hilbert sort: find permutation $\sigma_{t-1}$ such that $h \circ \psi(\mathbf{x}_{t-1}^{\sigma_{t-1}(1)}) \leq \ldots \leq h \circ \psi(\mathbf{x}_{t-1}^{\sigma_{t-1}(N)})$ if $d \geq 2$, or
        $\mathbf{x}_{t-1}^{\sigma_{t-1}(1)} \leq \ldots \leq \mathbf{x}_{t-1}^{\sigma_{t-1}(N)}$ if $d = 1$.
    (c): Find permutation $\tau$ such that $u_t^{\tau(1)} \leq ... \leq u_t^{\tau(N)}$, generate $a_{t-1}^{1:N}$ using Algorithm 3, with inputs
        $u_t^{\tau(1:N)}$ and $W_{t-1}^{\sigma_{t-1}(1:N)}$, and compute $\mathbf{x}_t^n = \Gamma_t(\mathbf{x}_{t-1}^{\sigma_{t-1}(a_{t-1}^n)}, \mathbf{v}_t^{\tau(n)})$ for each $n = 1, \dots, N$.
    (e): Compute $w_t^n = G_t(\mathbf{x}_{t-1}^{\sigma_{t-1}(a_{t-1}^n)}, \mathbf{x}_t^n)$, and $W_t^n = w_t^n / \sum_{m=1}^N w_t^m$ for each $n = 1, \dots, N$.
```

In the SQMC context, we use h to transform the N ancestors into points in $[0,1]$, before using the inverse CDF as for $d = 1$. More precisely, instead of constructing a Monte Carlo approximation of

$$\sum_{n=1}^{N} W_{t-1}^n \left\{ \delta_{\mathbf{x}_{t-1}^n}(\mathrm{d}\mathbf{x}_{t-1}) \times f^X(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t-1}) \right\}$$

we construct a low-discrepancy approximation of

$$\sum_{n=1}^{N} W_{t-1}^n \left\{ \delta_{h \circ \psi(\mathbf{x}_{t-1}^n)}(\mathrm{d}h) \times f^X(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t-1}) \right\}$$

where ψ is some user chosen transformation, from $\mathbb{R}^d$ to $[0,1]^d$, so that indeed $h \circ \psi(\mathbf{x}_{t-1}^n) \in [0,1]$. Thus, one may proceed as follows: first, find permutation σ such that $h \circ \psi(\mathbf{x}_{t-1}^{\sigma(1)}) \leq \ldots \leq h \circ \psi(\mathbf{x}_{t-1}^{\sigma(n)})$; then, exactly as before, and for each n, find m such that

$$\sum_{p=1}^{m-1} W_{t-1}^{\sigma(p)} \leq u_t^n \leq \sum_{p=1}^{m} W_{t-1}^{\sigma(p)} \quad \text{(empty sum equals 0)}$$

and set $a_t^n = n$. The rest of the Algorithm is unchanged; see Algorithm 4.

Although we have motivated the Hilbert curve in this short description as a practical way to "project" the N ancestors into $[0,1]$, there are more fundamental reasons why the Hilbert curve is a particularly convenient transformation in the context of SQMC. In a few words, the Hilbert curve (and its inverse) preserves discrepancy in some sense, that is, if the ancestors $\mathbf{x}_{t-1}^n$ have low discrepancy, then so will have the $h(\mathbf{x}_{t-1}^n)$. This point turns out to be essential when establishing the convergence properties of SQMC, (see [GC15] for a sharper description of this important point).

2.3.5. *Concluding remarks*

The main advantage of SQMC approach over standard particle filtering is the faster convergence, as $N \to \infty$. We refer to [GC15] for a formal convergence results that support this statement, and several simulation studies, where improvement factors range from 10 to 10^5 (in the sense that SMC would need 10 to 10^5 more particles to reach the same mean square error than SQMC in the considered numerical examples).

In general, the magnitude of the rate of convergence is upper-bounded by $\mathcal{O}(N^{-1})$. However, with additional assumptions on the smoothness of the function φ, [HO14] postulate that faster rates could be obtained. In particular, the authors conjecture that the order of the MSE could be $\mathcal{O}(N^{-1-2/d})$. This rate obviously deteriorates as d grows, yet it still compares favorably with the usual rate $\mathcal{O}(N^{-1})$.

More generally, QMC is now widespread in Bayesian statistics and seems to have been slightly overlooked in Bayesian computation, at least up to now. We expect that the advent of SQMC will hopefully change this state of affair.

3. A nonparametric analysis of Approximate Bayesian Computation (ABC)

3.1. Description of the method

In Bayesian statistics, other pathological situation occurs when the model is so complicated that the only task we can perform is to sample from it. This type of problem (originally arising in genetics) has motivated a drive to more approximate approaches, in particular the field of Approximate Bayesian Computation (ABC for short). In a nutshell, ABC is a family of computational techniques that offers an almost automated solution in situations where a systematic evaluation of the likelihood is computationally prohibitive, or whenever suitable likelihoods are not available. The approach was originally mentioned, but not analyzed, in [Rub84]. It was further developed in population genetics in [FL97, TBGD97, PSPLF99, BZB02], who gave the name of Approximate Bayesian Computation to a family of likelihood-free inference methods. Since its original developments, the ABC paradigm has successfully been applied to various scientific areas, ranging from archaeological science and ecology to epidemiology, stereology and protein network analysis. The recent survey [MPRR12] offers both a historical and a technical review of the domain.

Before we go into more details on ABC, further notations are required. In the sequel, we still denote $\pi(\boldsymbol{\theta})$ the density of π with respect to the Lebesgue measure on $\mathbb{R}^p$ and the (fixed) observation vector is denoted $\mathbf{y_n} = (Y_1, \ldots, Y_n)$. We assume to be given a statistic $\mathbf{S}$, taking values in $\mathbb{R}^m$. It is a function of the random variable $\boldsymbol{Y}$, with a dimension m typically much smaller than the size of $\boldsymbol{Y}$. The statistic $\mathbf{S}$ is supposed to admit a conditional density $f(\mathbf{s}|\boldsymbol{\theta})$ with respect to the Lebesgue measure on $\mathbb{R}^m$. Strictly speaking, we should write $\mathbf{S}(\boldsymbol{Y})$ instead of $\mathbf{S}$. However, since there is no ambiguity, we continue to use the latter notation. As such, the statistic $\mathbf{S}$ should be understood as a low-dimensional summary of $\boldsymbol{Y}$ and is a key ingredient to address the computational difficulties generated by the high dimensional framework generated by the number of observations in the model: $m << n$. For example, it can be a sufficient statistic for the parameter $\boldsymbol{\theta}$, but not necessarily. The conditional distribution on $\boldsymbol{\theta}$ given $\mathbf{S} = \mathbf{s}$ has a density $g(\boldsymbol{\theta}|\mathbf{s})$. According to the Bayes rule, this conditional density takes the form

$$g(\boldsymbol{\theta}|\mathbf{s}) = \frac{f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\bar{f}(\mathbf{s})}, \quad \text{where } \bar{f}(\mathbf{s}) = \int_{\mathbb{R}^p} f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

is the marginal density of $\mathbf{S}$. Finally, we denote by $\mathbf{s}_n = \mathbf{S}(\mathbf{y_n})$ the observed realization of $\mathbf{S}$ computed on the observation vector $\mathbf{y_n}$. Throughout this section, $\mathbf{s}_0$ and $\boldsymbol{y}_0$ should be considered as fixed quantities and N is the number of simulations (or particles) simulated by the ABC algorithm. In practice, the parameter N should be chosen large (typically of the order of 10^6), while k_N is commonly expressed as a percentile of N.

From a nonparametric perspective, this algorithm falls within the broad family of nearest neighbor-type procedures [FH51,LQ65,Cov68]. In order to better understand the rationale behind it, denote by $(\boldsymbol{\theta}_1, \boldsymbol{y}^1), \ldots, (\boldsymbol{\theta}_N, \boldsymbol{y}^N)$ an i.i.d. sample, with common joint distribution $\ell_n(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. This sample is clearly associated with the i.i.d. sequence $(\boldsymbol{\theta}_1, \mathbf{S}^1), \ldots, (\boldsymbol{\theta}_N, \mathbf{S}^N)$, where each pair has a density $f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Finally, let $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(k_N)}$ be the k_N-nearest neighbors of $\mathbf{s}_n$ among $\mathbf{S}^1, \ldots, \mathbf{S}^N$, and let $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k_N)}$ be the corresponding $\boldsymbol{\theta}_i$'s. With this notation, we see that the generic ABC Algorithm 5 proceeds in two steps:

1. First, simulate an N-sample $(\boldsymbol{\theta}_1, \boldsymbol{y}^1), \ldots, (\boldsymbol{\theta}_N, \boldsymbol{y}^n)$;
2. Seconds, return the variables $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k_N)}$.

Algorithm 5 Pseudo-code of a generic ABC algorithm

```
Require: A positive integer $N$, an integer $k_N$ between 1 and $N$, an observation vector $\mathbf{y_n}$ and $\mathbf{s}_n$.
Require: A sampling algorithm of $\pi$ and a sampling algorithm of observations $\boldsymbol{Y} \sim \ell_n(.|\boldsymbol{\theta})$.
  for $i = 1$ to $N$ do
    Generate $\boldsymbol{\theta}_i$ in $\boldsymbol{\Theta}$ from the prior $\pi$;
    Generate an $n$ sample $\boldsymbol{y}^i = (Y_1^i, \ldots, Y_n^i)$ from the law $\ell_n(.|\boldsymbol{\theta}_i)$.
  end for

  return The $\boldsymbol{\theta}_i$'s such that $\mathbf{S}^i = \mathbf{S}(\boldsymbol{y}^i)$ is among the $k_N$-nearest neighbors of $\mathbf{s}_n$.
```

As will become clear in Section 3.2, this simple observation opens the way to a mathematical analysis of ABC via statistical methods based on the nearest neighbors. For now, let us just specify that for a fixed $\mathbf{s}_n \in \mathbb{R}^m$, the estimate we consider to infer the posterior density $g(.|\mathbf{s}_n)$ at some point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is

$$\hat{g}_N(\boldsymbol{\theta}_0) = \frac{1}{k_N h_N^p} \sum_{j=1}^{k_N} K\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{(j)}}{h_N}\right), \tag{5}$$

where $\{h_N\}_{N\geq 0}$ is a sequence of positive real numbers (bandwidth) and K is a nonnegative Borel measurable function (kernel) on $\mathbb{R}^p$. To reduce the notational burden, we dropped the dependency of the estimate upon $\mathbf{s}_n$, keeping in mind that $\mathbf{s}_n$ is held fixed. The idea is simple: in order to estimate the posterior, just look at the k_N-nearest neighbors of $\mathbf{s}_n$ and smooth the corresponding $\boldsymbol{\theta}_j$'s around $\boldsymbol{\theta}_0$. It should be noted that (5) is a smart hybrid between a k-nearest neighbor and a kernel density estimation procedure. In particular, it is different from the Rosenblatt-type [Ros69] kernel conditional density estimates proposed in [BZB02] and analyzed in [Blu10].

To conclude this introduction, we would like to make a few comments on the topics that is **not** addressed in the following. An important part of the performance of the ABC approach, especially for high-dimensional data sets, relies upon a good choice of the summary statistic $\mathbf{S}$. In many practical applications, this statistic is picked by an expert in the field, without any particular guarantee of success. A systematic approach to choosing such a statistic, based upon a sound theoretical framework, is currently under active investigation in the Bayesian community. This important issue will not be pursued further here. As a good starting point, the interested reader is referred to [JM08], where is developed a sequential scheme for scoring statistics according to whether their inclusion in the analysis will substantially improve the quality of inference. Similarly, we not address issues regarding how to enhance efficiency of ABC and its variants, as for example with the sequential techniques of [SFT07] and [BCMR09]. Nor won't we explore the important question of ABC model choice, for which theoretical arguments are still missing [RCMP11, MPRR11]. Finally, we refer the reader to [BCG15] for details and proofs concerning the upcoming results.

3.2. Distribution of ABC outputs

We recall that $(\boldsymbol{\theta}_1, \mathbf{S}^1), \ldots, (\boldsymbol{\theta}_N, \mathbf{S}^N)$ are i.i.d. $\mathbb{R}^p \times \mathbb{R}^m$-valued random variables, with common probability density $f(\boldsymbol{\theta}, \mathbf{s}) = f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Both $\mathbb{R}^p$ and $\mathbb{R}^m$ are endowed with the Euclidean norm $\|.\|$. In this section, attention is focused on the distribution of the algorithm outputs $(\boldsymbol{\theta}_{(1)}, \mathbf{S}^{(1)}), \ldots, (\boldsymbol{\theta}_{(k_N)}, \mathbf{S}^{(k_N)})$.

In what follows, we denote by d_i the (random) distance between $\mathbf{s}_n$ and $\mathbf{S}^i$. Similarly, we let $d_{(i)}$ be the distance between $\mathbf{s}_n$ and its i-th nearest neighbor among $\mathbf{S}^1, \ldots, \mathbf{S}^N$, that is $d_{(i)} = \|\mathbf{S}^{(i)} - \mathbf{s}_n\|$. It turns out that, **conditionally on** $d_{(k_N+1)}$, one can consider the k_N-tuple $(\boldsymbol{\theta}_{(1)}, \mathbf{S}^{(1)}), \ldots, (\boldsymbol{\theta}_{(k_N)}, \mathbf{S}^{(k_N)})$ as an ordered sample drawn according to the probability density

$$\frac{\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_n\| \leq d_{(k_N+1)}]} f(\boldsymbol{\theta}, \mathbf{s})}{\displaystyle\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_n, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{s}},$$

where $\mathcal{B}_m(\mathbf{s}_n,\delta)$ stands for the closed ball in $\mathbb{R}^m$ centered at $\mathbf{s}_n$ with nonnegative radius δ. Alternatively, the (unordered) simulated values may be treated like i.i.d. realizations of variables with common density proportional to $\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_n\|\leq d_{(k_N+1)}]}f(\boldsymbol{\theta},\mathbf{s})$. Thus, given $d_{(k_N+1)}$, the accepted $\boldsymbol{\theta}_j$'s are i.i.d. realizations of the probability density

$$\frac{\displaystyle\int_{\mathcal{B}_m(\mathbf{s}_n,d_{(k_N+1)})} f(\boldsymbol{\theta},\mathbf{s})\mathrm{d}\mathbf{s}}{\displaystyle\int_{\mathbb{R}^p}\int_{\mathcal{B}_m(\mathbf{s}_n,d_{(k_N+1)})} f(\vartheta,\mathbf{s})\mathrm{d}\vartheta\mathrm{d}\mathbf{s}}.$$

Although this result is intuitively clear, its proof requires a careful mathematical analysis (see [BCG15]). Moreover, it plays a key role in the mathematical analysis of the conditional density estimate (5) associated with ABC methodology. In fact, investigating ABC in terms of nearest neighbors has other important consequences. Suppose, for example, that we are interested in estimating some finite conditional expectation $\mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{S}=\mathbf{s}_n]$, where the random variable $\varphi(\boldsymbol{\theta})$ is bounded. If $\boldsymbol{\theta}$ is itself bounded, this includes in particular the important setting where φ is polynomial and one wishes to estimate the conditional moments of $\boldsymbol{\theta}$. Then, provided $k_N/\log\log N\to\infty$ and $k_N/N\to 0$ as $N\to\infty$, *pointwise consistency* can be shown, which means that for almost all $\mathbf{s}_n$ (with respect to the distribution of $\mathbf{S}$), with probability 1,

$$\frac{1}{k_N}\sum_{j=1}^{k_N}\varphi\left(\boldsymbol{\theta}_{(j)}\right)\to\mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{S}=\mathbf{s}_n]. \tag{6}$$

The proof of this result uses a sharp statistical analysis of the nearest neighbor estimation ability. To be more precise, let us consider an i.i.d. sample $(\mathbf{X}_1,Z_1),\ldots,(\mathbf{X}_N,Z_N)$ taking values in $\mathbb{R}^m\times\mathbb{R}$, where the output variables Z_i's are bounded. Assume that the $\mathbf{X}_i$'s have a density and that our goal is to assess the regression function $r(\mathbf{x})=\mathbb{E}[Z\,|\,\mathbf{X}=\mathbf{x}]$, $\mathbf{x}\in\mathbb{R}^m$. Then the k_N-nearest neighbor regression function estimate of r takes the form

$$\hat{r}_N(\mathbf{x})=\frac{1}{k_N}\sum_{j=1}^{k_N}Z_{(j)},\quad \mathbf{x}\in\mathbb{R}^m,$$

where $Z_{(j)}$ is the Z-observation corresponding to $\mathbf{X}_{(j)}$, the j-th-closest point to $\mathbf{x}$ among $\mathbf{X}_1,\ldots,\mathbf{X}_N$. Denoting by μ the distribution of $\mathbf{X}_1$, it is stated in Theorem 3 of [Dev82] that provided $k_N/\log\log N\to\infty$ and $k_N/N\to 0$, then for μ-almost all $\mathbf{x}$, $\hat{r}_N(\mathbf{x})$ goes to $r(\mathbf{x})$ with probability 1 as N goes to ∞. This result can be transposed without further effort to our ABC setting via the correspondence $\varphi(\boldsymbol{\theta})\leftrightarrow Z$ and $\mathbf{S}\leftrightarrow\mathbf{X}$, thereby yielding (6).

3.3. Theoretical derivations

3.3.1. *Mean square error consistency*

Our next objective is to estimate the posterior density $g(\boldsymbol{\theta}_0|\mathbf{s}_n)$, $\boldsymbol{\theta}_0\in\mathbb{R}^p$. This estimation step is an important ingredient of the Bayesian analysis, whether this may be for visualization purposes or for more involved mathematical achievements. As exposed in the introduction, a natural ABC-companion estimator of $g(\boldsymbol{\theta}_0|\mathbf{s}_n)$ takes the form (5). Our goal in this section is to investigate some consistency properties of this estimator. Pointwise mean square error consistency is stated in Theorem 3.1 and mean integrated square error consistency is established in Theorem 3.2. We stress that this part of the document is concerned with minimal conditions of convergence. However, the following assumptions on the kernel is needed:

Assumption [K1] The kernel K is nonnegative and belongs to $L^1(\mathbb{R}^p)$, with $\int_{\mathbb{R}^p}K(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=1$. Moreover, the function $\boldsymbol{\theta}\in\mathbb{R}^p\longmapsto\sup_{\|\mathbf{y}\|\geq\|\boldsymbol{\theta}\|}|K(\mathbf{y})|$ is in $L^1(\mathbb{R}^p)$.

Assumption [**K1**] is in no way restrictive and is satisfied by all standard kernels such as, for example, the uniform kernel or the Gaussian kernel. In the following, we denote by λ_p (respectively, λ_m) the Lebesgue measure on $\mathbb{R}^p$ (respectively, $\mathbb{R}^m$) and set, for any positive h,

$$K_h(\boldsymbol{\theta}) = \frac{1}{h^p} K\left(\frac{\boldsymbol{\theta}}{h}\right), \quad \boldsymbol{\theta} \in \mathbb{R}^p.$$

We note once and for all that Assumption [**K1**] implies that $\int_{\mathbb{R}^p} K_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$. We are now in a position to state the two main results of this section.

Theorem 3.1 (Pointwise mean square error consistency)**.** *Assume that the kernel K is bounded and satisfies Assumption* [**K1**]*. Assume, in addition, that the joint probability density f is such that*

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \mathbf{s}) \log^+ f(\boldsymbol{\theta}, \mathbf{s})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s} < \infty. \tag{7}$$

Then, for $\lambda_p \otimes \lambda_m$-almost all $(\boldsymbol{\theta}_0, \mathbf{s}_n) \in \mathbb{R}^p \times \mathbb{R}^m$, with $\bar{f}(\mathbf{s}_n) > 0$, if $k_N \to \infty$, $k_N/N \to 0$, $h_N \to 0$ and $k_N h_N^p \to \infty$,

$$\mathbb{E}\left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_n)\right]^2 \to 0 \quad \textit{as } N \to \infty.$$

It is easy to see that assumption (7) is mild. It is for example satisfied whenever f is bounded or whenever f belongs to $L^q(\mathbb{R}^p \times \mathbb{R}^m)$ with $q > 1$. Theorem 3.2 below says that $\hat{g}_N$ is also consistent with respect to the mean integrated square error criterion. Here again, the regularity assumptions on f and π are minimal.

Theorem 3.2 (Mean integrated square error consistency)**.** *Assume that the kernel K belongs to $L^2(\mathbb{R}^p)$ and satisfies Assumption* [**K1**]*. Assume, in addition, that the joint probability density f and the prior π are in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$ and $L^2(\mathbb{R}^p)$, respectively. Then, for λ_m-almost all $\mathbf{s}_n \in \mathbb{R}^m$, with $\bar{f}(\mathbf{s}_n) > 0$, if $k_N \to \infty$, $k_N/N \to 0$, $h_N \to 0$ and $k_N h_N^p \to \infty$,*

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_n)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] \to 0 \quad \textit{as } N \to \infty.$$

3.3.2. *Rates of convergence*

In this section, we go one step further in the analysis of the ABC-companion estimate $\hat{g}_N$ by studying its mean integrated square error rates of convergence. As before, we try to keep the assumptions on unknown mathematical objects as mild as possible. We introduce the multi-index notation

$$|\beta| = \beta_1 + \ldots + \beta_n, \quad \beta! = \beta_1! \ldots \beta_n!, \quad \mathbf{x}^\beta = x_1^{\beta_1} \ldots x_n^{\beta_n}$$

for $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ and $\mathbf{x} \in \mathbb{R}^n$. If all the k-order derivatives of some function $\varphi : \mathbb{R}^n \to \mathbb{R}$ are continuous at $\mathbf{x}_0 \in \mathbb{R}^n$ then, by Schwarz's theorem, one can change the order of mixed derivatives at $\mathbf{x}_0$, and the notations

$$D^\beta \varphi(\mathbf{x}_0) = \frac{\partial^{|\beta|}\varphi(\mathbf{x}_0)}{\partial x_1^{\beta_1} \ldots \partial x_n^{\beta_n}}, \quad |\beta| \le k$$

for the higher-order partial derivatives are thus justified in this situation. Recall that the collection of all $\mathbf{s}_0 \in \mathbb{R}^m$ with $\int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} > 0$ for all $\delta > 0$ is called the support of $\bar{f}$. We shall need the following set of assumptions.

Assumption [A1] $\bar{f}$ has a compact support included in a ball of diameter $L > 0$ and is three times continuously differentiable.

Assumption [A2] The joint probability density f is in $L^2(\mathbb{R}^p \times \mathbb{R}^m)$. Moreover, for fixed $\mathbf{s}_n$, the functions

$$\boldsymbol{\theta}_0 \mapsto \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_n)}{\partial\theta_{i_1}\partial\theta_{i_2}}, \quad 1 \le i_1, i_2 \le p \qquad \text{and} \qquad \boldsymbol{\theta}_0 \mapsto \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_n)}{\partial s_j^2}, \quad 1 \le j \le m,$$

are defined and belong to $L^2(\mathbb{R}^p)$.

Assumption [A3] f is three times continuously differentiable on $\mathbb{R}^p \times \mathbb{R}^m$ and, for any β satisfying $|\beta| = 3$,

$$\sup_{\mathbf{s}\in\mathbb{R}^m} \int_{\mathbb{R}^p} \left[D^\beta f(\boldsymbol{\theta}, \mathbf{s})\right]^2 \mathrm{d}\boldsymbol{\theta} < \infty.$$

Assumption [K2] K is symmetric, is in $L^2(\mathbb{R}^p)$, and for any β such that $|\beta| \in \{1, 2, 3\}$, $\int_{\mathbb{R}^p} \left|\boldsymbol{\theta}^\beta\right| K(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < \infty$.

Recall that $\mathbf{s}_n$ is called a Lebesgue point for $\bar{f}$ if

$$\frac{1}{\lambda_m(\mathcal{B}_m(\mathbf{s}_n, \delta))} \int_{\mathcal{B}_m(\mathbf{s}_n,\delta)} \left|\bar{f}(\mathbf{s}) - \bar{f}(\mathbf{s}_n)\right| \mathrm{d}\mathbf{s} \to 0 \quad \text{as } \delta \to 0.$$

Lebesgue's differentiation theorem asserts that this is true for λ_m-almost all $\mathbf{s}_n \in \mathbb{R}^m$. If $\mathbf{s}_n$ is a Lebesgue point of $\bar{f}$ such that $\bar{f}(\mathbf{s}_n) > 0$, then it is readily seen that

$$0 < \xi_0 = \inf_{0<\delta\le L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_n,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} < \infty.$$

Let us mention that Lebesgue points are commonly encountered when dealing with nearest neighbor estimators. This was already pointed in the seminal work of [Dev82], and thereafter extended by considering "Besicovitch" conditions in [CG06]. Some recent developments in [GKM15] have even established that this kind of "minimal mass assumption" on small balls are unavoidable in general finite dimensional spaces to derive uniform consistency rates of classification (with any classifier).

Theorem 3.3 (Rates of convergence)**.** *Suppose that assumptions* [**K1**]-[**K2**] *and* [**A1**]-[**A3**] *are satisfied. Let* $\mathbf{s}_n$ *be a Lebesgue point of* $\bar{f}$ *such that* $\bar{f}(\mathbf{s}_n) > 0$*. Then, for* $m > 4$*, there exist sequences* $\{k_N\}$ *with* $k_N \propto N^{\frac{p+4}{m+p+4}}$ *and* $\{h_N\}$ *with* $h_N \propto N^{-\frac{1}{m+p+4}}$*, and a constant* C *such that*

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_n)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] = (C + \mathrm{o}(1))\, N^{-\frac{4}{m+p+4}}.$$

A few concluding remarks are in order:

(1) The constant C is explicit and depends on $\mathbf{s}_n$, ξ_0, K, $\bar{f}$ and f. We refer the reader to [BCG15] for details.

(2) Comparing this result with Theorem 4 in [BLR14] and taking $\bar{\alpha} = 2/(m+p)$, it turns out that the rate of convergence $N^{-4/(m+p+4)}$ is optimal from the minimax viewpoint.

(3) From a practical perspective, the fundamental problem is that of the joint choice of k_N and h_N in the absence of *a priori* information regarding the posterior $g(.|\mathbf{s}_n)$. Various bandwidth selection rules for conditional density estimates have been proposed in the literature [BH01, HRL04, FY04]. But most (if not all) of these procedures pertain to kernel-type estimates and are difficult to adapt to our nearest-neighbor setting. Moreover, they are tailored to global statistical performance criteria, whereas the problem here is local since $\mathbf{s}_n$ is fixed. Hence, devising a good methodology to automatically select both k_N and h_N in function of $\mathbf{s}_n$ necessitates a supplemental specific analysis.

(4) Nevertheless, Theorem 3.3 provides an insight into the proportion of simulated values which should be accepted by the algorithm. For example, a rule of thumb is obtained by taking $k_N \approx N^{(p+4)/(m+p+4)}$, so that a fraction of about $k_N/N \approx N^{-m/(m+p+4)}$ simulations should not be rejected.

(5) At last, it should be noted that the size of the statistic $\mathbf{S}$ (the integer m) can dramatically damage the convergence rate obtained in Theorem 3.3. It is thus a basic fact to choose a sufficient statistic embedded in the lowest dimensional space possible. Moreover, let us remark that the use of a low-dimensional

statistics $\mathbf{S}$ is the fundamental tool that permit to bypass the computational difficulty generated by a large number of observation n.

4. Consistency for an example of nonparametric hidden Markov model

4.1. Consistency rate of Bayesian procedures

As pointed above in the introduction, famous works (*e.g.* [Fre65] or [DF86]) have shown that in a nonparametric setting, the choice of the Bayesian prior is of primary importance to obtain Bayesian consistent procedures. A general and now usual method to prove consistency was introduced by [Sch65]. Some historical modifications can also be found in [IH81] but recent advances stand on the seminal work of [Bar88]. Roughly speaking, Bayesian consistency holds if the prior puts some mass around $\boldsymbol{\theta}_0$ (according to Kullback-Leibler neighborhoods of $\boldsymbol{\theta}_0$) and if there exist exponentially consistent tests to discriminate $\boldsymbol{\theta}_0$ against the complement of any neighborhood V of $\boldsymbol{\theta}_0$ (w.r.t. the considered topology, see Definition 1.1). intersected with the complement of a set with an exponential decreasing prior mass. This assumption acts as a penalization of complex models.

Next, a crucial step consists in choosing an adapted topology on $\boldsymbol{\Theta}$, especially when $\boldsymbol{\Theta}$ is infinite dimensional. In particular the neighborhoods mentioned above may be defined through metric on probability distributions *via* distance and weak topology on distributions, and then transferred to a topology on $\boldsymbol{\Theta}$. Indeed the property of consistency highly depends on the topology considered on $\boldsymbol{\Theta}$ and becomes more difficult to obtain with larger topologies. As an illustration, the pioneering work of [Bar88] establishes the weak consistency of Bayesian procedure for density estimation with i.i.d. observations under mild assumptions.

Theorem 4.1. *([Sch65], [GR03]) Let $\mathbf{y_n}$ be a sequence of i.i.d. observations distributed from $f_{\boldsymbol{\theta}_0}d\lambda$ and π a probability measure on the set $\mathcal{D}$ of densities with respect to λ. If for all $\epsilon > 0$,*

$$\pi\left\{f \in \mathcal{D} \ : d_{KL}(f, f_{\boldsymbol{\theta}_0}) < \epsilon\right\} > 0$$

then the posterior is consistent for the weak topology on distributions at $f_{\boldsymbol{\theta}_0}d\lambda$.

For the weak topology, the existence of the tests is a direct consequence of the Hoeffding inequality with no additional constraint. With the L_1 topology, the existence of such tests requires more effort, and relies on a link between the prior and the entropy of the model. The next result from [GR03] extends Theorem 4.1 to the case of density estimation with the L_1 topology.

Theorem 4.2. *[GR03] Let $\mathbf{y_n}$ be a sequence of i.i.d. observations distributed from $f_{\boldsymbol{\theta}_0}d\lambda$ and π a probability measure on the set $\mathcal{D}$ of densities with respect to λ. We further assume that the following conditions hold:*

i) *for all $\epsilon > 0$,*

$$\pi\left\{f \in \mathcal{D} \ : d_{KL}(f, f_{\boldsymbol{\theta}_0}) < \epsilon\right\} > 0, \tag{8}$$

ii) *for all $\delta > 0$, there exist a subset $\mathcal{F}_n$ of $\mathcal{D}$ and positive real numbers C_1 and β_1 such that*

$$\pi(\mathcal{F}_n^c) \leq C_1 \exp(-n\beta_1) \quad \text{and} \quad \sum_{n>0} N(\delta, \mathcal{F}_n, L_1)\exp(-n\delta^2/2) < \infty, \tag{9}$$

where $N(\cdot,\cdot,L_1)$ denotes the covering numbers for the L_1 topology.

Then the posterior is consistent for the L_1 topology on $\mathcal{D}$ at $f_{\boldsymbol{\theta}_0}$.

Note that different priors (based on Dirichlet or Gaussian processes) may be use in Theorem 4.2. In a sense, Bayesian consistency could be used as a guide for better prior specification. Its efficiency is controlled by the rate convergence of the associated posterior, introduced below.

Definition 4.1 (Posterior rate of convergence)**.** *The posterior $\pi(\cdot|\mathbf{y_n})$ converges with rate $\epsilon_n \to 0$ w.r.t. the distance d at $\boldsymbol{\theta}_0$ if there exists $M > 0$ such that*

$$\pi(\theta : d(\theta, \theta^*) < M\epsilon_n | Y_1, \dots, Y_n) \to 1 \quad P^{\boldsymbol{\theta}_0} a.s.$$

We refer the reader to [GGvdV00] for a detailed presentation. Note also that posterior consistency and rate of convergence, have attracted attention in various settings such as the problem of the shape invariant model (see *e.g.* [BG14]), the estimation either of a spectral density for a stationary time series or of the transition density of some ergodic Markov processes (see *e.g.* [CGR05]). More recently, [Ver15] studied the case where the observations are dependent, linked through a hidden Markov model.

4.2. The studied model: Hidden Markov Models with finite state space

We now turn back to a specific case of HMMs introduced in Section 2.3, when the hidden component $(x_t)_{t\in\mathbb{N}}$ belongs to a finite state space. We are interested in Bayesian consistency when the observations vary with the time n. We would like to stress that the context is slightly different from the one stated in Section 4.1: observations $(\mathbf{y}_t)_{1\leq t\leq n}$ are no longer independent.

Since parametric modelling of emission distributions may lead to poor results in practice, recent attention has been drawn in using nonparametric HMMs, see [YPRH11], [GCR15] and references therein. One can find some algorithms to approximate the posterior in the framework of nonparametric HMMs with finite state space (see Algorithm 6). In this section we introduce a method to study the asymptotic behaviour of the posterior in this framework.

Theoretical results for estimation procedures in nonparametric HMMs have been obtained in the recent works by [DLC12] and [GR15] in some restrictive settings. Indeed, in more general situation, even identifiability is a difficult question (addressed by [GCR15]). Frequentist asymptotic properties of estimators of HMMs parameters have been studied since the 1990s. Consistency and asymptotic normality of the maximum likelihood estimator were established in the parametric case, see [DM01], [DMR04] and [DMOvH11] for the most general results. Finally, there are only a few parametric Bayesian asymptotic results, see [dGS08] when the number of hidden states is known and [GR14a] when the number of hidden states is unknown.

We now specify the studied model (illustrated in Figure 2). We still denote the HMMs by $(x_t, \mathbf{y}_t)_{t\in\mathbb{N}}$ where $\mathbf{x}$ is a homogeneous Markov chain. In this section, the hidden states $x_t, 1 \leq t \leq n$ belong to a finite state space $\mathcal{X}$, whereas the state space is general in Section 2.2. The transition kernel is still denoted $f^X(x_t|x_{t-1})$ and it can now be simply described as a squared matrix. The conditional probability distribution of $\mathbf{y}_t$ when x_t is given, denoted by $f^Y(\mathbf{y}_t|x_t)$, is also called emission distribution.

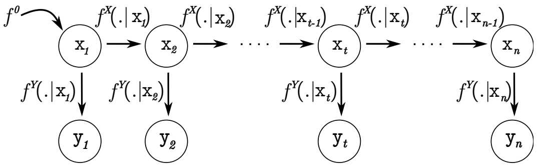

Figure 2. Evolution of the HMM with a transition kernel f^X and emission distributions $f^Y(\cdot|x)$ when $x_1 \sim f^0$.

In what follows, we will assume that f^X is strongly irreducible, meaning that there exists $\underline{q} > 0$ such that

$$\forall (i,j) \in [\![1, k]\!] \qquad f^X(j|i) \geq \underline{q}.$$

The former assumption on the transition kernel f^X implies that the Markov chain $\mathbf{x}$ possesses a unique invariant distribution f^0 with an exponential mixing rate. We also assume the chain is initialized with its invariant distribution: $x_1 \sim f^0$.

4.3. Prior structure

We assume that the number k of hidden states, as well as $\underline{q}$, are known, so that the state space of the Markov chain is set to $\{1,\dots,k\}$. In order to define the set where the prior and the posterior distributions are defined, we introduce the $(k-1)$-dimensional simplex denoted

$$\Delta_k(\underline{q}) = \{(p_1,\dots,p_k) \ : \ p_i \geq \underline{q}, \ i=1,\dots k \ ; \ \sum_{i=1}^{k} p_i = 1\}.$$

The transition matrix f^X may be identified as a k-tuple of transition distributions (the rows of the matrix), so that $f^X \in \Delta_k(\underline{q})^k$. We denote $f^0 \in \Delta_k(\underline{q})$ the invariant probability measure, that also initializes the Markov chain (*i.e.*, $x_1 \sim f^0$). Let the observation space be $\mathbb{R}^d$, and $\mathcal{F}$ be the set of probability density functions with respect to a reference measure λ on $\mathbb{R}^d$. $\mathcal{F}^k$ is the set of possible emission densities from x_t to $\mathbf{y}_t$. It means that for any $f^Y = (f^Y(\cdot|1),\dots,f^Y(\cdot|k)) \in \mathcal{F}^k$, the distribution of $\mathbf{y}_t$ conditionally on $x_t = i$ will be $f^Y(\cdot|i)d\lambda$ for each value of i between 1 and k. Note that since $\mathcal{F}$ may not be a finite dimensional set, the problem is nonparametric.

Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (f^X, f^Y) \ : \ f^X \in \Delta_k(\underline{q})^k, f^Y \in \mathcal{F}^k\}$. Remark that choosing $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ implicitly defines a transition kernel f^X and therefore a unique invariant distribution f^0. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{P}^{\boldsymbol{\theta}}$ denotes the probability distribution of $(x_t, \mathbf{y}_t)_{t\in\mathbb{N}}$ and the transitions are parametrized by $\boldsymbol{\theta}$ with an initial state x_1 sampled from the corresponding invariant distribution f^0.

We denote $P_\ell^{\boldsymbol{\theta}}$ the marginal distribution of $\mathbf{y}_1,\dots,\mathbf{y}_l$ under $\mathbb{P}^{\boldsymbol{\theta}}$ and $p_\ell^{\boldsymbol{\theta}}$ its corresponding density with respect to $\lambda^{\otimes\ell}$ under $\mathbb{P}^{\boldsymbol{\theta}}$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ associated with an initial probability μ, we have:

$$p_\ell^{\boldsymbol{\theta}}(\mathbf{y}_1,\dots,\mathbf{y}_l) = \sum_{(x_1,\dots,x_l)\in[\![1,k]\!]^\ell} f^0(x_1) f^X(x_2|x_1)\dots f^X(x_\ell, x_{\ell-1}) f^Y(\mathbf{y}_1|x_1)\dots f^Y(\mathbf{y}_\ell|x_\ell).$$

Let π denote a prior on $\boldsymbol{\Theta}$. We assume that π is a product of probability measures on $\boldsymbol{\Theta}$, $\pi = \pi_{f^X} \otimes \pi_{f^Y}$ such that π_{f^X} is a probability distribution on $\Delta_k(\underline{q})^k$ and π_{f^Y} is a probability distribution on $\mathcal{F}^k$.

4.4. Posterior consistency

4.4.1. *Topological description*

The observations are now distributed from $\mathbb{P}^{\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0 = (f_0^X, f_0^Y)$ so that $(x_t, \mathbf{y}_t)_{t\geq 1}$ is a stationary hidden Markov chain. We are interested in posterior consistency, *i.e.*, in proving that for any neighborhood U of $\boldsymbol{\theta}_0$,

$$\lim_{n\to+\infty} \pi(U|\mathbf{y_n}) = 1, \qquad \mathbb{P}^{\boldsymbol{\theta}_0} - a.s.$$

To obtain a precise definition of any neighborhood U of $\boldsymbol{\theta}_0$, we will use two different topologies as in Theorems 4.1 and 4.2. We first use the weak topology on marginal distributions $(P_\ell^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$.

Let us briefly recall the definition of a weak neighborhood of $P_\ell^{\boldsymbol{\theta}}$. For any integer N and any set of bounded continuous functions $(h_j)_{1\leq j\leq N}$ from $(\mathbb{R}^d)^\ell$ to $\mathbb{R}$, we denote

$$\mathcal{W}\left(p_\ell^{\boldsymbol{\theta}}, \epsilon, (h_j)_{1\leq j\leq N}\right) := \left\{P \ : \ \left|\int h_j dP - \int h_j p_\ell^{\boldsymbol{\theta}} d\lambda^{\otimes\ell}\right| < \epsilon, \forall j \in [\![1,N]\!]\right\}. \tag{10}$$

A weak neighborhood of $p_\ell^{\boldsymbol{\theta}}$ is a set of probability distributions O such that:

$$\exists\, N \in \mathbb{N}, \quad \exists\, \epsilon > 0, \quad \exists\, (h_j)_{1\leq j\leq N}, \quad s.t. \qquad \mathcal{W}\left(p_\ell^{\boldsymbol{\theta}}, \epsilon, (h_j)_{1\leq j\leq N}\right) \subset O.$$

Secondly, we work with the larger topology associated to the L_1-distance on the joint densities. Other topologies may be considered depending on the estimation needed, see for example [Ver15] where a product of the topologies for the transition matrix and the emission densities is also used.

4.4.2. *Main results*

In HMMs, $\mathbf{y}_t$ may not only depend on the previous observation $\mathbf{y}_{t-1}$ but also on the previous observations $\mathbf{y}_{t-2}, \dots, \mathbf{y}_1$. The generalization of the Hoeffding inequalities [Rio00] requires a level of mixing of the chain to ensure an exponential rate of concentration, hence the existence of a powerful test. Since $\boldsymbol{\theta}_0$ is such that $f_0^X \in \Delta_k(\underline{q})^k$ with a known $\underline{q}$ (non adaptive prior on $\underline{q}$), we will only consider prior π_{f^X} such that

$$\pi_{f^X} \left\{ \Delta_k(\underline{q})^k \right\} = 1. \tag{11}$$

This ensures a level of mixing of the Markov chains for the possible parameters and that the associated Markov chains are irreducible (and thus positive recurrent since $\mathcal{X}$ is finite) and admit a unique stationary probability measure.

Theorem 4.3 uses the following assumption **[N]** that lead to the posterior consistency for the weak topology, and has the same flavor as Theorem 4.1.

Assumption [N] For any $\epsilon > 0$ small enough, there exists a set $\boldsymbol{\Theta}_\epsilon \subset \boldsymbol{\Theta}$ such that $\pi(\boldsymbol{\Theta}_\epsilon) > 0$ and for any $\boldsymbol{\theta} = (f^X, f^Y) \in \boldsymbol{\Theta}_\epsilon$,

$$\begin{gathered} \|f^X - f_0^X\| < \epsilon, \\ \max_{1 \le i \le k} \int f_0^Y(y|i) \max_{1 \le j \le k} \log\left(\frac{f_0^Y(y|j)}{f^Y(y|j)}\right) \lambda(dy) < \epsilon, \\ \sum_{j=1}^k f^Y(y|j) > 0, \text{for all } y \in \mathbb{R}^d \text{ such that } \sum_{i=1}^k f_0^Y(y|i) > 0, \\ \sup_{y \,:\, \sum_{i=1}^k f_0^Y(y|i) > 0} \max_{1 \le j \le k} f^Y(y|j) < +\infty, \\ \sum_{i=1}^k \int f_0^Y(y|i) \left| \log\left(\sum_{j=1}^k f^Y(y|j)\right)\right| \lambda(dy) < +\infty. \end{gathered} \tag{12}$$

Theorem 4.3. *[Ver15] Assume that the prior π satisfies (11) and that Assumption **[N]** holds. Then for all weak neighborhood U of $P_\ell^{\boldsymbol{\theta}_0}$ as defined in (10),*

$$\lim_{n \to \infty} \pi(U|\mathbf{y_n}) = 1 \qquad \mathbb{P}^{\boldsymbol{\theta}_0} - a.s.$$

Theorem 4.3 is proved in [Ver15] using the general method introduced in [Bar88]. Assumption (11) ensures the existence of tests that discriminate the set of hypotheses $\mathbb{P}^{\boldsymbol{\theta}}$ when $\boldsymbol{\theta}$ is not in a close neighborhood of $\boldsymbol{\theta}_0$ for the weak topology. These results are derived by using a generalization of Hoeffding's inequality by [Rio00] and [GR14a]. Assumption **[N]** ensures that the prior π gives a positive weight to any Kullback-Leibler neighborhood of $\mathbb{P}^{\boldsymbol{\theta}_0}$.

It is also possible to derive a stronger result by using the L_1- norm, which defines a finer topology than the weak one. As in the case of density estimation with i.i.d. observations (Theorem 4.2), an additional assumption on the covering number implies the existence of tests that allow to discriminate $\mathbb{P}^{\boldsymbol{\theta}}$ from $\mathbb{P}^{\boldsymbol{\theta}_0}$ when dealing with the L^1 distance. For this purpose, we define the distance

$$\forall (f, g) \in \mathcal{F} \qquad d(f, g) = \max_{1 \le i \le k} \|f_i - g_i\|_1.$$

Assumption [H] For all $n > 0$, for all $\delta > 0$ there exists a set $\mathcal{F}_n \subset \mathcal{F}^k$ and positive numbers r_1, C_1 such that

$$\pi_f\big((\mathcal{F}_n)^c\big) \le C_1 e^{-nr_1} \text{ and such that } \sum_{n>0} N\left(\frac{\delta}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \exp\left(-\frac{n\delta^2 k^2 \underline{q}^2}{32l}\right) < +\infty. \tag{13}$$

Theorem 4.4. *[Ver15] Assume that the prior π satisfies (11) and that Assumption **[N]** and Assumption **[H]** hold. Then for all L_1-neighborhood U of $P_\ell^{\boldsymbol{\theta}_0}$,*

$$\lim_{n\to\infty} \pi(U|\mathbf{y_n}) = 1 \qquad \mathbb{P}^{\boldsymbol{\theta}_0} - a.s.$$

Thanks to the similarities of Assumptions (12) and (13) with Assumptions (8) and (9) respectively, it may be possible to use consistent priors in the case of density estimation with i.i.d. observations for the emission distribution in HMMs. Such examples are given in [Ver15] for instance in the case of translated emission distributions that is to say when for all $1 \le j \le k$,

$$f^Y(\cdot|j) = g(\cdot - m_j)$$

where for all $1 \le j \le k$, m_j is distributed from a probability distribution on $\mathbb{R}$ and g is a density function on $\mathbb{R}$ distributed from a mixture of Gaussians by Dirichlet process.

We complete the previous results with some insights on computational aspects. Algorithm 1 is not fitted when θ belongs to an infinite dimensional set, *i.e.*, in the nonparametric framework.

Another way of approximating the posterior distribution consists in implementing an MCMC algorithm based on Gibbs sampling. Introducing the hidden states of the HMM as latent components of the parameter allows to simplify the sampling. As a consequence, the hidden states conditionally on the observations, the transition matrix and the emission distributions may be sampled with a forward-backward algorithm, precisely described in [YPRH11]. Algorithm 6 was implemented in the nonparametric framework by [CMR05].

Algorithm 6 Gibbs sampling + forward-backward

```
Input: priors $\pi_Q$ and $\pi_f$, maximal number of iteration $I$.
Output: $(Q^0, \dots, Q^I)$ and $(f^0, \dots, f^I)$.
At time 0,
    Sample $Q^0 \sim \pi_Q$.
    Sample $f^0 \sim \pi_f$.
Recursively, from iteration $i = 1$ to iteration $i = I$,
    Sample $\mathbf{x}^i$ the hidden states with the forward-backward algorithm with transition matrix $Q^{i-1}$,
      emission distributions $f^{i-1}$ and the observations $\mathbf{y}$.
    Sample $Q^i \sim \pi(Q|\mathbf{x}^{i-1}, \mathbf{y}, f^{i-1})$.
    Sample $f^i \sim \pi(f|\mathbf{x}^{i-1}, \mathbf{y}, Q^i)$.
```

5. The PAC-Bayesian paradigm

The PAC theory consists in deriving risk bound on randomized estimators (see for example [Val84]). The PAC-Bayesian theory originates in the two seminal papers [STW97, McA99] and has been extensively formalized in the context of classification by [Cat04, Cat07] and regression by [Aud04a, Aud04b, Alq06, Alq08, AC10, AC11]. Note also the work of [See02, See03] in the framework of Gaussian processes, and the papers [SLCB+12], [AW12], [ALW13] focusing on time series and martingales. In addition, it has been worked out in the sparsity perspective more recently by [DT08, DT12, AL11, DS12, Suz12, AB13, GA13, Gue13a].

5.1. Generalities on PAC-Bayesian approaches

To illustrate the concepts behind the PAC-Bayesian approach, let us consider the standard regression model $\mathbf{y} = f_{\theta^0}(\mathbf{x}) + W$, where $\mathbf{y}$ is a real-valued response, $f_{\boldsymbol{\theta}^0} : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function depending on some parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbf{x}$ is a d-dimensional random variable and W is a real-valued noise term such that $\mathbb{E}[W|\mathbf{x}] = 0$. Let us assume that we collect an n-sized sample of i.i.d. replications of the random variable $(\mathbf{x}, \mathbf{y})$ denoted $(X_1, Y_1), \dots, (X_n, Y_n)$. For some loss function $\ell : \mathbb{R} \times \mathbb{R} \to (0, \infty)$, we define the risk (and its empirical counterpart) of some estimator $f_{\hat{\boldsymbol{\theta}}}$ of $f_{\boldsymbol{\theta}^0}$ as

$$R(f_{\hat{\boldsymbol{\theta}}}) = \mathbb{E}[\ell(\mathbf{y}, f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}))], \quad R_n(f_{\hat{\boldsymbol{\theta}}}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_{\hat{\boldsymbol{\theta}}}(X_i)).$$

Let $R^\star = R(f_{\boldsymbol{\theta}^0})$, $R^\star$ be the lowest (oracle) risk that can be reached by any predictor $f_{\boldsymbol{\theta}}$. We aim to obtain some statistical guarantees involving inequalities in deviations of the excess risk of Bayesian estimators $f_{\hat{\boldsymbol{\theta}}}$ built with a suitable choice of the prior. The nice oracle inequalities we are looking for are generally stated as follows:

$$\forall \varepsilon \in (0,1) \qquad \mathbb{P}\left[R(f_{\hat{\boldsymbol{\theta}}}) - R^\star \leq \mathrm{K} \inf_{\boldsymbol{\theta}} \Big\{ R(f_{\boldsymbol{\theta}}) - R^\star + \Delta_{n,d,\varepsilon}(\boldsymbol{\theta}) \Big\} \right] \geq 1 - \varepsilon,$$

where $K \geq 1$ is a constant and $\Delta_{n,d,\varepsilon}$ is a remainder term which decays as n grows. The message of this work is that when the ambient dimension d is large with respect to the sample size n, it is possible with a properly-chosen prior to reach convergence rates $\Delta_{n,d,\varepsilon}$ that are not too badly affected by the curse of dimensionality. Another saliant fact is that this procedure relies on very little assumption on the distribution of the variable $(\mathbf{x}, \mathbf{y})$.

We propose to investigate a semi-parametric form for the regression function, allowing for flexibility. We are interested in the situation where the unknow $f_{\boldsymbol{\theta}}$ can be sparsely decomposed in an additive model

$$f_{\boldsymbol{\theta}^0}(x_1, \dots, x_d) = \sum_{j=1}^{d} \psi_j^0(x_j),$$

and we assume that only a few of $(\psi_j^0)_{1 \leq j \leq d}$ influences the response $\mathbf{y}$.

This drives us to consider an additive model of the form

$$\left\{ f_{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_d) = \sum_{j=1}^{d} \sum_{k=1}^{m_j} \theta_{jk} \phi_k(\mathbf{x}_j) \ , \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R}^{\sum_{j=1}^{d} m_j}, \quad \|f_{\boldsymbol{\theta}}\|_\infty \leq C \right\},$$

where $\mathbf{m} = (m_1, \dots, m_d) \in \{0, \dots, M\}^d$ is a model, $\mathbb{D} = \{\phi_1, \phi_2, \dots, \phi_M\}$ is a known dictionary composed of deterministic functions (or preliminary estimators). Furthermore, C is a known constant that controls the volume of the parameters space in order to be consistent with the learning sample. This additive formulation (see for example [Sto85, HT86]) achieves a nice compromise between flexibility and interpretation.

The PAC approach produces a priori risk bounds (see [Val84]); the additional Bayesian flavor allows us to obtain a posteriori bounds. In what follows, we are especially interested in the situation of a sparse oracle $f_{\boldsymbol{\theta}^0}$ to recover. We consider $\mathcal{M}$ the set of measures on $\boldsymbol{\Theta}$ that are absolutely continuous with respect to a reference measure $d\boldsymbol{\theta}$. We wish to use a prior probability measure $\pi \in \mathcal{M}$ promoting sparsity. For this purpose, set $\lambda > 0$ and consider the following constrained optimization problem:

$$\operatorname*{arg\,min}_{\rho \in \mathcal{M}} \left\{ \int_{\boldsymbol{\Theta}} R_n(f_{\boldsymbol{\theta}}) \rho(\mathrm{d}\boldsymbol{\theta}) + \frac{\lambda}{n} d_{KL}(\rho, \pi) \right\}. \tag{14}$$

Indeed, the (frequentist) variational formulation of (14) may be interpreted as a Bayesian formulation (justifying the interpretation of π as a prior distribution). In fact, it is an exercise to check that (14) has a unique solution, which is the so-called *Gibbs posterior distribution*

$$\hat{\rho}_\lambda(\mathrm{d}\boldsymbol{\theta}) \propto \exp[-\lambda R_n(f_{\boldsymbol{\theta}})]\pi(\mathrm{d}\boldsymbol{\theta}).$$

Hence, the penalization parameter $\lambda > 0$ may be seen as an inverse temperature parameter of the Gibbs distribution. Based on the Gibbs posterior distribution $\hat{\rho}_\lambda$, two estimators are considered in this document:

$$\hat{\boldsymbol{\theta}} \sim \hat{\rho}_\lambda \quad \text{(Randomized estimator sampled with the posterior)},$$
$$\bar{\boldsymbol{\theta}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\hat{\rho}_\lambda(\mathrm{d}\boldsymbol{\theta}) = \mathbb{E}_{\hat{\rho}_\lambda}\boldsymbol{\theta} \quad \text{(Posterior mean)}.$$

As it will be shown below, PAC-Bayesian theory is a great tool to produce estimators with nearly minimax optimal properties. The first important result for PAC-Bayesian theory is the standard link between the Legendre transform of the Kullback-Leibler divergence and a Gibbs fields.

Lemma 5.1 ([Csi75]). *Let $(A, \mathcal{A})$ be a measurable space. For any probability measure μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int(\exp \circ h)\mathrm{d}\mu < \infty$,*

$$\log \int (\exp \circ h)\mathrm{d}\mu = \sup_{m \in \mathcal{M}^1_\mu(A,\mathcal{A})} \left\{ \int h\mathrm{d}m - d_{KL}(m, \mu) \right\},$$

with the convention $\infty - \infty = -\infty$. Furthermore, if h is upper-bounded on the support of μ, the supremum with respect to m in the right-hand term is reached for the Gibbs distribution g defined by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp \circ h(a)}{\int(\exp \circ h)\mathrm{d}\mu}, \quad a \in A.$$

The second important result is the Bernstein concentration inequality. PAC-Bayesian bounds depend on the Kullback-Leibler divergence and hold for any prior π. In order to obtain an optimized PAC-Bayesian estimator, we can carefully set up two parameters: the inverse temperature λ and the prior π. These two key quantities must be tailored to obtain some good oracle inequalities. In particular, we may consider a sparsity-inducing prior, such as

$$\pi_s(\boldsymbol{\theta}) \propto \sum_{\mathbf{m}} \binom{d}{|\mathbf{m}|_0}^{-1} \beta^{\sum_{j=1}^d m_j} \, \mathrm{Unif}_{\mathcal{B}_{\mathbf{m}}(C)}(\boldsymbol{\theta}),$$

where $\beta \in (0, 1)$ and $\mathcal{B}_{\mathbf{m}}(C)$ is the ℓ^1 sphere of radius C:

$$\mathcal{B}_{\mathbf{m}}(C) = \left\{ \boldsymbol{\theta}, \quad \sum_{j=1}^{d} \sum_{k=1}^{m_j} |\theta_{jk}| \leq C \right\}.$$

This prior distribution π_s defined above satisfies the desirable property to favor sparse parameters since it gives some important weight to the parameters with a low ℓ^0 norm of the coefficients $(m_j)_{1\leq j\leq d}$. Note that such priors are common in the PAC-Bayesian literature and may be traced back to [LB06].

5.2. Examples

The work [GA13] provides several practical examples detailed below.

5.2.1. *Regression models*

We consider the standard model

$$\mathbf{y} = \psi^\star(\mathbf{x}) + \mathbf{w},$$

with two mild assumptions.

Assumption [C] The noise is subexponential:

- For any integer $k \geq 2$, $\mathbb{E}[|\mathbf{w}|^k] < \infty$.
- $\mathbb{E}[\mathbf{w}|\mathbf{x}] = 0$.
- There exist two positive constants L, σ^2 such that, for any integer $k \geq 2$,

$$\mathbb{E}[|\mathbf{w}|^k|\mathbf{x}] \leq \frac{k!}{2}\sigma^2 L^{k-2}.$$

Assumption [B] $|\psi^\star|_\infty \leq C$.

In particular, Assumption **[C]** is met if $\mathbf{w}$ has a Gaussian distribution. As for Assumption **[B]**, it allows to use concentration inequalities (see for example [Mas07]). From what precedes, we obtain the following oracle inequality.

Theorem 5.2 ([GA13]). *For any $\varepsilon \in (0,1)$, set $\lambda = n\gamma/(4\sigma^2 + 4C^2)$ where $\gamma \in (0,1)$. With probability at least $1-\varepsilon$,*

$$\left.\begin{array}{l} R(f_{\hat{\boldsymbol{\theta}}}) - R(\psi^\star) \\ R(f_{\bar{\boldsymbol{\theta}}}) - R(\psi^\star) \end{array}\right\} \leq \mathrm{K}_\gamma \times \inf_{\mathbf{m}} \inf_{\boldsymbol{\theta}\in\mathcal{B}_{\mathbf{m}}(C)} \left\{ R(f_{\boldsymbol{\theta}}) - R(\psi^\star) + |\mathbf{m}|_0 \frac{\log(d/|\mathbf{m}|_0)}{n\gamma} + \frac{\log(n)}{n\gamma}\sum_{j=1}^{d} m_j + \frac{\log(2/\varepsilon)}{n\ell} \right\},$$

where $\mathrm{K}_\gamma \xrightarrow[\gamma\to 0]{} 1$ and $\mathrm{K}_\gamma \xrightarrow[\gamma\to 1]{} +\infty$.

In this result, the classical tradeoff in PAC-Bayesian-flavored oracle inequalities appear: the constant K_γ should be chosen as close as possible to 1 (in this case, the oracle inequality is said *exact* or *sharp*), but this enforces $\gamma = 0$ and the rates on the r.h.s. explode.

5.2.2. *Case of Sobolev space*

In certain function space, it is possible to derive some minimax optimality properties. Assume that ψ^* is indeed an additive form of nonparametric decomposition in a Sovolev space, for example say $\psi^\star = \sum_{j\in S^\star}\psi_j^\star$, and let $\phi_1, \phi_2, \ldots$ refer to the trigonometric basis. Assume that each of the ψ_j^*s belong to a Sobolev ellipsoid:

$$\psi_j^\star \in \mathcal{W}(r_j, \ell_j) = \left\{ f \in \mathrm{L}^2([-1,1]) : f = \sum_{k=1}^{\infty} \boldsymbol{\theta}_k\phi_k \text{ and } \sum_{i=1}^{\infty} i^{2r_j}\boldsymbol{\theta}_i^2 \leq \ell_j \right\},$$

where r_j's are unknown regularity parameters, casting our results onto the adaptive setting. We obtain the following oracle inequality.

Theorem 5.3 ([GA13]). *For any real $\varepsilon \in (0,1)$, set $\lambda = n\gamma/(4\sigma^2 + 4C^2)$ where $\gamma \in (0,1)$, with probability at least $1-\varepsilon$,*

$$\left.\begin{array}{l} R(f_{\hat{\boldsymbol{\theta}}}) - R(\psi^\star) \\ R(f_{\bar{\boldsymbol{\theta}}}) - R(\psi^\star) \end{array}\right\} \leq \mathrm{K}_\gamma \times \left\{ \sum_{j\in S^\star} \ell_j^{\frac{1}{2r_j+1}} \left(\frac{\log(n)}{2n\gamma r_j}\right)^{\frac{2r_j}{2r_j+1}} + \frac{|S^\star|\log(d/|S^\star|)}{n\gamma} + \frac{\log(2/\varepsilon)}{n\gamma} \right\}.$$

The message conveyed by these inequalities is that if there exists a sparse representation of the regression function, then the right-hand side terms of the upper bound involved in Theorem 5.3 become negligible and the excess risk of the PAC-Bayesian estimators mimics the best excess risk one could achieve in the collection. Moreover, the excess loss appears to be minimax up to a log term. Finally, note that since the ambient dimension

d only appears in logarithmic terms, the PAC-Bayesian paradigm is well-suited to the high dimensional setting, provided that a sparse representation of the regression function is available.

5.2.3. *Logistic Regression*

This PAC-Bayesian approach has been extended by [Gue13a] to the logistic regression model: $\mathbf{y} = \{\pm 1\}$, model

$$\log \frac{\mathbb{P}(\mathbf{y} = 1|\mathbf{x})}{1 - \mathbb{P}(Y = 1|\mathbf{x})} = \nu(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

The logistic loss function is thus defined as

$$\ell : (\mathbf{y}, f_{\boldsymbol{\theta}}(\mathbf{x})) \mapsto \log\left[1 + \exp(-\mathbf{y} f_{\boldsymbol{\theta}}(\mathbf{x}))\right].$$

Then the link function ν is estimated by the same collection of additive combinations of elements of the dictionary, as before. Similar oracle inequalities are provided in [Gue13a].

5.2.4. *Binary Ranking*

Note that the PAC-Bayesian theory can also be used to solve the binary ranking problem in a high-dimensional setting (see *e.g.* [GR14b]).

The bipartite ranking problem consists in learning from a sample $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ to *rank* observations $\mathbf{x}_i$, while preserving the order of their associated labels $\mathbf{y}_i \in \{\pm 1\}$. We consider this problem in the high dimensional situation, where the observations $(\mathbf{x}_i)_{1 \leq i \leq n}$ lie in a space of dimension d, possibly much larger than the sample size n. A standard approach in this context involves the introduction of a *scoring function*. We propose to estimate the optimal scoring function using the so-called Gibbs posterior distribution, which favors sparse additive estimators. This procedure appears valuable to assess the effect of each covariate on the score of an observation. Using elements from the PAC-Bayesian theory, we provide theoretical guarantees about our method, along with an implementation through MCMC.

5.3. Implementation

Note that the implementation relies on MCMC algorithms, favoring local moves of the Markov Chain. This is achieved by a so-called Subspace Carlin & Chib approach (see [CC95, PD12]), and is freely available in the R package [Gue13b], named *pacbpred* (PAC-***B****ayesian* ***Pred****iction*)[1]. A key ingredient is that we define a neighborhood relationship among the models: for any model $\mathbf{m}$, we define the possible neighborhood $\mathcal{V}_{\mathbf{m}}^+$ (resp. $\mathcal{V}_{\mathbf{m}}^-$), which includes any model sharing its covariates with $\mathbf{m}$, with an additional one (resp. minus one). The essence of the proposed algorithm is to build a symmetric random walk on the models state (as in Algorithm 1), by the means of balanced random additions and deletions of covariates. The pseudo-code of the method is detailed in Algorithm 7.

Conclusion

This paper presents some recent advances in Bayesian statistics in the context of high dimensions. HMMs are a popular modelling and within this framework, Section 4 introduces new theoretical and nonparametric results of posterior consistency. In addition, from a computational aspect, Section 2.2 provides the reader with algorithms to cope with the high dimensional settings since SQMC is shown to be at least $\mathcal{O}(N^{-1})$, whatever the dimension d is. An alternative family of algorithms consists in ABC, and the paper focuses on the need to find a low dimensional exhaustive statistics to obtain an efficient method. Lastly, the paper adapts the PAC-Bayesian tools to cope with the curse of high dimensionality: the message is that PAC-Bayesian adapts neatly to this difficult setting, provided that a sparsity-inducing prior is used. The paper presents such priors, and theoretical results along with an efficient MCMC implementation.

As such, Bayesian statistics address modern statistical problems and remain a very active field of research.

[1] http://cran.r-project.org/web/packages/pacbpred/index.html.

Algorithm 7 A subspace Carlin & Chib flavored MCMC algorithm

```
Input: finite horizon (positive integer) $T$, proposal variance $\sigma^2 > 0$, inverse temperature parameter $\lambda > 0$.
Output: a chain $(\mathbf{m}^t)_{t=1}^T$ and $(\theta^t)_{t=1}^T$.
At time $t = 2, \dots, T$,
    Pick a move: Add, delete a covariate, or stay in the current model, and form the corresponding neighborhood (in the latter case, the neighborhood is restricted to the current model itself).
    For any $\mathbf{m}$ in $\mathcal{V}^{\pm}_{\mathbf{m}_t}$, draw a candidate estimator $\tilde{\theta} \sim \mathcal{N}(\bar{\theta}, \sigma^2 \mathcal{I})$ where the Gaussian pseudoprior is centered in a candidate estimator $\bar{\theta}$[2] and where $\mathcal{I}$ stands for the identity matrix in $\mathbb{R}^{|\mathbf{m}|}$. The density of this distribution is denoted by $\varphi$.
    Pick the model $\mathbf{m}$ and candidate parameter $\tilde{\theta}$ with probability proportional to $\hat{\rho}_\lambda(\tilde{\theta})/\varphi(\tilde{\theta})$.
    Using the Metropolis-Hastings acceptance ratio
        $$\alpha = \min\left(1, \frac{\hat{\rho}_\lambda(\tilde{\theta})\varphi(\theta^{t-1})}{\hat{\rho}_\lambda(\theta^{t-1})\varphi(\tilde{\theta})}\right),$$
    set $\theta^t = \tilde{\theta}$ and $\mathbf{m}^t = \mathbf{m}$ with probability $\alpha$.
```

References

[AB13] P. Alquier and G. Biau. Sparse Single-Index Model. *Journal of Machine Learning Research*, 14:243–280, 2013.

[AC10] J.-Y. Audibert and O. Catoni. Robust linear regression through PAC-Bayesian truncation. Preprint, 2010.

[AC11] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

[AL11] P. Alquier and K. Lounici. PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

[Alq06] P. Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Pierre & Marie Curie - Paris VI, December 2006.

[Alq08] P. Alquier. PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

[ALW13] P. Alquier, X. Li, and O. Wintenberger. Prediction of Time Series by Statistical Learning: General Losses and Fast Rates. *Dependence Modeling*, 1:65–93, 2013.

[Aud04a] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré : Probabilités et Statistiques*, 40(6):685–736, November-December 2004.

[Aud04b] J.-Y. Audibert. *Théorie statistique de l' apprentissage : une approche PAC-Bayésienne*. PhD thesis, Université Pierre & Marie Curie - Paris VI, June 2004.

[AW12] P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

[Bar88] A.R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report, April 1988.

[BCG15] G. Biau, F. Cérou, and A. Guyader. New insights into approximate Bayesian computation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 51(1):376–403, 2015.

[BCMR09] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[Bea10] M. A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406, 2010.

[BG14] D. Bontemps and S. Gadat. Bayesian methods for the Shape Invariant Model. *Electronic Journal of Statistics*, 8:1522–1568, 2014.

[BH01] D.M. Bashtannyk and R.J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, 36:279–298, 2001.

[BLR14] K. Bertin, C. Lacour, and V. Rivoirard. Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincar, Probab. Statist., to appear*, 2014.

[Blu10] M. G. B. Blum. Approximate Bayesian computation: a nonparametric perspective. *J. Amer. Statist. Assoc.*, 105(491):1178–1187, 2010. With supplementary material available online.

[BP66] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[BZB02] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

[Cat04] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* école d'été de Probabilités de Saint-Flour XXXI – 2001. Springer, 2004.

[Cat07] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series.* Institute of Mathematical Statistics, 2007.

[CC95] B. P. Carlin and S. Chib. Bayesian Model choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B*, 57(3):473–484, 1995.

[CF13] A. Caimo and N. Friel. Bayesian model selection for exponential random graph models. *Social Networks*, 35(1):11–24, 2013.

[CG06] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.*, 10:340–355 (electronic), 2006.

[CGR05] N. Choudhuri, S. Ghosal, and A. Roy. Bayesian methods for function estimation. In *Bayesian thinking: modeling and computation*, volume 25 of *Handbook of Statist.*, pages 373–414. Elsevier/North-Holland, Amsterdam, 2005.

[CMR05] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models.* Springer, 2005.

[Cov68] T.M. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55, 1968.

[Csi75] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

[Dev82] L. Devroye. Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Verw. Gebiete*, 61(4):467–481, 1982.

[Dev86] L. Devroye. *Nonuniform random variate generation.* Springer-Verlag, New York, 1986.

[DF86] P Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, 14:1–26, 1986.

[dGS08] M. C. de Gunst and O. Shcherbakova. Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17(4):342–356, 2008.

[DLC12] T. Dumont and S. Le Corff. Nonparametric estimation in hidden markov models. *arxiv preprint arXiv:1209.0633*, 2012.

[DM01] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general Hidden Markov Models. *Bernoulli*, 7:381–420, 2001.

[DMOvH11] R. Douc, E. Moulines, J. Olsson, and R. van Handel. Consistency of the maximum likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, 39(1):474–513, 2011.

[DMR04] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of statistics*, 32(5):2254–2304, 2004.

[Doo49] J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[DS12] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

[DT08] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[DT12] A. S. Dalalyan and A. B. Tsybakov. Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

[Eve12] R. G Everitt. Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960, 2012.

[FH51] E. Fix and J. L. Hodges. *Discriminatory analysis—Nonparametric discrimination: Consistency properties.* Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, pages 261-279, Randolph Field, 1951.

[FL97] Y.X. Fu and W.H. Li. Estimating the age of the common ancestor of a sample of DNA sequences. *Journal of Molecular Biology and Evolution*, 14:195–199, 1997.

[Fre65] D. A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.*, 36:454–456, 1965.

[FY04] J. Fan and T.H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 94:819–834, 2004.

[GA13] B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

[GC15] M. Gerber and N. Chopin. Sequential Quasi-Monte Carlo. *Journal of the Royal Statistical Society, series B (in press)*, 77(3):509–579, 2015.

[GCR15] E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric Hidden Markov Models and applications. *Statistics and Computing, to appear*, 2015.

[GGvdV00] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[GKM15] S. Gadat, T. Klein, and C. Marteau. Classification with the nearest neighor rule in general finite dimensional spaces. *Preprint*, pages 1–42, 2015.

[GR03] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, 2003.

[GR14a] E. Gassiat and J. Rousseau. About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 2014.

[GR14b] B. Guedj and S. Robbiano. Une approche PAC-bayésienne dun problème de ranking binaire en grande dimension. In *46mes Journées de Statistique de la SFdS, Rennes*, 2014.

[GR15] E. Gassiat and J. Rousseau. Non parametric finite translation mixtures with dependent regime. *Bernoulli, to appear, arXiv preprint arXiv:1302.2345*, 2015.

[Gue13a] B. Guedj. *Agrégation d'estimateurs et de classificateurs : théorie et méthodes*. PhD thesis, Université Pierre & Marie Curie – Paris VI, 2013.

[Gue13b] B. Guedj. *pacbpred: PAC-Bayesian Estimation and Prediction in Sparse Additive Models*, February 2013. R package version 0.92.2.

[GZ01] M. G. Gu and H.-T. Zhu. Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):339–355, 2001.

[Has70] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[HO14] Zhijian He and Art B. Owen. Extensible grids: uniform sampling on a space-filling curve. Preprint, 2014.

[HRL04] P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99:1015–1026, 2004.

[HT86] T. Hastie and R. Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–318, 1986.

[IH81] I. Ibragimov and R. Has'minskii. *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics. Springer, Berlin, Heidelberg, New York, first edition, 1981.

[JM08] P. Joyce and P. Marjoran. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(26), 2008.

[LB06] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, August 2006.

[LCY00] L. Le Cam and G. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Verlag, New-York, 2000.

[Lem09] C. Lemieux. *Monte Carlo and quasi-Monte Carlo sampling*. Springer Series in Statistics. Springer, New York, 2009.

[LQ65] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36:1049–1051, 1965.

[Mas07] P. Massart. *Concentration Inequalities and Model Selection*. école d'été de Probabilités de Saint-Flour XXXIII – 2003. Springer, 2007.

[McA99] D. A. McAllester. Some PAC-Bayesian Theorems. *Machine Learning*, 37:355–363, 1999.

[MPRR11] J.-M. Marin, N. Pillai, C. P. Robert, and J. Rousseau. *Relevant statistics for Bayesian model choice*. arXiv:1110.4700, 2011.

[MPRR12] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Stat. Comput.*, 22(6):1167–1180, 2012.

[MRR$^+$53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[MZ97] I. L. MacDonald and W. Zucchini. *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall/CRC, London, UK, 1997.

[PD12] A. Petralias and P. Dellaportas. An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation*, 0(0):1–19, 2012.

[PSPLF99] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798, 1999.

[RC04] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.)*. Springer, New York, 2004.

[RCMP11] C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N.S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108:15112–15117, 2011.

[Rio00] E. Rio. Inégalités de hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(10):905–908, 2000.

[Rip06] B. D. Ripley. *Stochastic simulation*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. Reprint of the 1987 original, Wiley-Interscience Paperback Series.

[Ros69] M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York, 1969.

[Rub84] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172, 1984.

[Sag94] H. Sagan. *Space-filling curves*. Universitext. Springer-Verlag, New York, 1994.

[Sch65] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26, 1965.

[See02] M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

[See03] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

[SFT07] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765, 2007.

[SLCB$^+$12] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

[Sto85] C. J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705, 1985.

[STW97] J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997.

[Suz12] T. Suzuki. PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model. In *Proceedings of the 25th annual conference on Computational Learning Theory*, 2012.

[TBGD97] S. Tavaré, D. Balding, R. Griffith, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Ver15] E. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electron. J. Stat.*, 9:717–752, 2015.

[Wai14] M. J. Wainwright. Constrained forms of statistical minimax: Computation, communication and privacy. In *Proceedings of the International Congress of Mathematicians*, 2014.

[YPRH11] C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society*, 73:37–57, 2011.