ESAIM: PROCEEDINGS AND SURVEYS, **Vol. ?, 2014, 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 statistiques 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* bayésien. 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. This evolution also guided by very concrete applications in biostatistics and signal processing (among others) has raised new natural questions that mainly concern two important points. The first one asks how should be a "good" Bayesian prior

[1] ENSAE, 3 Avenue Pierre Larousse, 92245 Malakoff, France

[2] Toulouse School of Economics (Université Toulouse I Capitole), 21 allées de Brienne, 31000 Toulouse, France

[3] Modal project-team, Inria Lille - Nord Europe. 40 avenue du Halley, 59650 Villeneuve dAscq

[4] LSTA, Université Pierre et Marie Curie & Projet ASPI, INRIA Rennes, France

[5] Université Paris-Sud, 91405 Orsay Cedex France

for high dimensional or 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 make it possible the computation of the posterior distribution and, if possible, quantify the way these numerical methods approximate this posterior distribution.

1.1. **Bayes approaches**

In what follows, we will consider a dominated model parameterized by a set of measurable parameters $\boldsymbol{\Theta}$. We will assume that $\boldsymbol{\Theta}$ is included in a metric space and each parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 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 will be referred to as $f(.|\boldsymbol{\theta})$.

A Bayesian prior π on $\boldsymbol{\Theta}$ is an initial distribution on $\boldsymbol{\Theta}$ that traduces a belief on the distribution of an unobserved parameter $\boldsymbol{\theta}$ living on $\boldsymbol{\Theta}$. We are then interested in statistical inference on $\boldsymbol{\Theta}$ (or in a quantity related to a distribution on $\boldsymbol{\Theta}$) when observing an i.i.d. sample of size n, denoted $\mathbf{y_n} := (Y_1, \ldots, Y_n)$ 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:

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

We will see in the sequel some very nice results about the behaviour of the posterior distribution, which thus permit to compute certain quantities (e.g. mean or moments) of the posterior distribution and therefore to perform Bayesian inference.

1.2. **Posterior computation**

1.2.1. *Bayesian inference*

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 monographs [RC04] and [MR07]. However, in some contexts, computation of the posterior is problematic, either because the size

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: $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k$.
```

of the data makes the calculation computationally intractable, or because calculation is impossible when using realistic models for how the data arises. Thus, despite their power and flexibility, MCMC procedures and their variants may prove irrelevant in a growing number of contemporary applications involving very large dimensions or complicated models. This computational burden typically arises in fields such as ecology, population genetics and image analysis, just to name a few.

1.2.2. *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 3), 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.

1.2.3. *Approximate Bayesian Computation methods*

Another 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. There are too many references to be included here, but the recent survey [MPRR12] offers both a historical and a technical review of the domain.

1.3. Consistency of Bayesian procedures

1.3.1. *Frequentist point of view*

As already discussed above, the choice of the prior is a key issue in Bayesian statistics. It can be important for computational reasons since it may help a lot to use some particular conjugate prior/posterior to accelerate the evaluation of Bayes estimators (see *e.g.* [GCSR04]). It is also at the core of Bayesian consistency by adopting a frequentist point of view. A natural 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 (meaning that $n \longrightarrow +\infty$).

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

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

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] proved that in the nonparametric case, 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.

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 on any closed neighborhood of $\boldsymbol{\theta}_0$ and if there exist exponentially consistent tests to discriminate $\boldsymbol{\theta}_0$ against the complementary of all neighborhood of $\boldsymbol{\theta}_0$ (for the considered topology) intersected with a set with an exponential decreasing prior mass. An important underlying concept resides on the topology considered on $\boldsymbol{\Theta}$. In particular the neighborhoods mentioned above are generally 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 $\boldsymbol{\Theta}$ (through the neighborhoods U considered). The finer the topology is, the more difficult it is to prove the existence of the tests and posterior consistency As an example, famous applications of the results stated in [Bar88], in the case of density estimation with i.i.d. observations lead to Theorems 1.2 and 1.3. Here, neighborhoods of $\boldsymbol{\theta}_0$ are defined through the Kullback-Leibler divergence between $\mathbb{P}(.|\boldsymbol{\theta})$ and $\mathbb{P}(.|\boldsymbol{\theta}_0)$ and the prior should put a positive weight a on such neighborhood.

Theorem 1.2. *([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 $\mathcal{D}$ at $f_{\boldsymbol{\theta}_0} d\lambda$.

For the weak topology, the existence of the tests is a direct consequence of the Hoeffding inequality without any additional constraint. Considering now a finer topology, namely the L_1 one, it is more difficult to prove the existence of such statistical tests. Particularly, the tests exist if the prior puts mainly its mass on not "too big" set (in the sense of covering numbers $N(., \cdot, \cdot)$). In particular, it is still possible to deal with the L_1 topology in the framework of density estimation with i.i.d. observations. It can be shown that if the prior puts mainly its mass on not a "too big" set (in the sense of covering numbers), then an exponentially consistent test exists. Such consequence is stated in the next result in the framework of density estimation.

Theorem 1.3. *[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 $\epsilon > 0$,*

$$\pi \left\{ f \in \mathcal{D} \ : d_{KL}(f, f_{\boldsymbol{\theta}_0}) < \epsilon \right\} > 0. \tag{2}$$

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

$$\pi(\mathcal{F}_n^c) \leq C_1 \exp(-n\beta_1) \quad \text{and} \quad \sum_{n>0} N(\delta, \mathcal{F}_n, L_1) \exp(-n\delta^2/2) < \infty \tag{3}$$

.

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

These last theorems can be applied to different priors based on Dirichlet or Gaussian processes (see *e.g.* [GR03]).

1.3.2. *Consistency rate of Bayesian procedures*

Finally, consistency is a tool for choosing the prior for a given estimation. To ensure more precisely the behavior of the posterior distribution, the rate of convergence of the posterior can be studied. The usual method to study the rates of convergence is given in [GGvdV00], it mainly relies on the method used to prove consistency and requires, as exhibited in the last Theorem, a fine upper bound on the complexity of the set where the estimation problem is located. For example, it is commonly necessary to upper bound the covering numbers (in the Hellinger or Kullback sense) of a particular set $\{\mathbb{P}(.|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and similarly control the weight of closed neiborhoods of $\mathbb{P}(.|\boldsymbol{\theta}_0)$. These methods have been applied in many various situations 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]). Recently, [Ver13] has studied the case where the observations are dependent, namely linked through a hidden Markov model.

1.3.3. *PAC-Bayesian approaches*

We can remark in the two previous paragraphs that both consistency and consistency rates are generally obtained in an asymptotic setting $n \longrightarrow +\infty$ although less is known when one is looking for a finite horizon result. In a nutshell, the PAC-Bayesian approach consists in a technical toolbox, 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. The core of the PAC-Bayesian scheme is the concentration of the empirical excess risk of a Bayesian estimator towards its risk. This is obtained by the means of Bernstein-like concentration inequalities in the following.

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 [ALW12, AW12, SLCB$^+$12] 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].

Below, we will review some recent advances in Bayesian statistics in high dimensional or nonparametric situations. Section 2.1 will describe 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 will provide an efficient way to get round of such difficulty by using a sequential quasi-Monte Carlo algorithm. Section 2.2 will discuss on ABC algorithms and will offer a nonparametric point of view for understanding the behaviour of estimators computed from ABC algorithms. In Section 3, some recent results taken from [Ver13] on posterior consistency for HMM are presented. We end the paper with Section 4, which aims at showing that the PAC-Bayesian approach adapts neatly to the high dimensional context when coupled with a suitable chosen sparsity-inducing prior.

2. Bayesian computation

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

2.1.1. *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 to 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 will simply assume that $\mathbf{x}_t$, resp. $\mathbf{y}_t$, take values in $\mathbb{R}^{d_x}$, resp. $\mathbb{R}^{d_y}$.

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,

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

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 require some form of Monte Carlo integration. For notational convenience, we will omit the dependence in $\boldsymbol{\theta}$ in what follows.

2.1.2. *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 $\mathbf{x}_0 \sim f^X(dx_0)$, $\mathbf{x}_t|\mathbf{x}_{t-1} = x_{t-1} \sim f^X(dx_t|x_{t-1})$. In particular, some complicate models are such that the density $f^X(x_t|x_{t-1})$ of the Markov transition is intractable, but Algorithm 2 can still be implemented in this case (provided we can at least sample from f^X).

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

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

for a given function $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$. In addition, the particle filter algorithm computes a quantity L^N_t that mimics an approximation of the likelihood $\ell(\mathbf{y}_{0:t})$. By approximation, we mean consistent estimation, as $N \to +\infty$ (under appropriate conditions).

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

Algorithm 2 Particle filter

```
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 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
        $L_t^N = L_{t-1}^N \left\{N^{-1} \sum_{n=1}^N w_t^n\right\}$.
```

- 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

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

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

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

- 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

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t-1}) = p(\mathbf{x}_{t-1}|\mathbf{y}_{0:t-1})f^X(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{4}$$

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t}) = \frac{p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t-1})f^Y(\mathbf{y}_t|\mathbf{x}_t)}{\ell(\mathbf{y}_t|\mathbf{y}_{0:t-1})} \tag{5}$$

  with $\ell(\mathbf{y}_t|\mathbf{y}_{0:t-1}) = \int p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{0:t-1})f^Y(\mathbf{y}_t|\mathbf{x}_t)\,\mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{x}_t$. Remark that (4) uses the fact $(\mathbf{x}_t)$ is Markov, and (5) is the simple Bayes formula. We then replace in (4) the term $p(\mathbf{x}_{t-1}|\mathbf{y}_{0:t-1})$ by the random probability measure obtained at step $t-1$:

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

  It is a mixture of N Dirac masses weighted according to the random weights W_{t-1}^n that traduce the likelihood of observations $\mathbf{y}_t$ given $\mathbf{x}_t^n$ (the weights increase with the conditional likelihood of $\mathbf{y}_t$ given $\mathbf{x}_t^n$). It is thus natural to 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 (5), 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, also 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.

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 $n = 1 \rightarrow N$ do
    while $s < u^n$ do
      $m \leftarrow m + 1$
      $s \leftarrow s + W^m$
    end while
    $a^n \leftarrow m$
  end for
```

The SQMC 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).

2.1.3. *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 will 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.

2.1.4. *SQMC (Sequential Quasi-Monte Carlo): $d_x = 1$*

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 $\mathbf{u}_0^n \sim \mathcal{U}\left([0,1]^{d_x}\right)$, 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 2.1.2 that iteration t may be interpreted as an importance sampling step, where we sample the $\mathbf{x}_t^n$'s from:

$$\sum_{n=1}^{N} W_{t-1}^n \left\{ \delta_{\mathbf{x}_{t-1}^n}(\mathrm{d}\mathbf{x}_{t-1}) \times f^X(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t-1}) \right\} \tag{6}$$

  and reweight these new particles by $f^Y(\mathbf{y}_t|\mathbf{x}_t^n)$. Thus, we need to rewrite the simulation from (6) as a deterministic function of uniforms. 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 compute $\mathbf{x}_t = \Gamma_t(\mathbf{x}_{t-1}, \mathbf{v}_t^n)$, when $\mathbf{v}_t^n \sim \mathcal{U}\left([0,1]^{d_x}\right)$.

  This can be done as follows: for each $n = 1, \ldots, N$, let $\mathbf{u}_t^n \sim \mathcal{U}\left([0,1]^{d_x+1}\right)$, and denote $\mathbf{u}_t^n = (u_t^n, \mathbf{v}_t^n)$, with $u_t^n \in [0,1]$, $\mathbf{v}_t^n \in [0,1]^{d_x}$. 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 so 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 $d_x = 1$ (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 $[0,1]^{d_x+1}$, one obtains the SQMC algorithm, described in Algorithm 4. (SQMC stands for Sequential Quasi Monte Carlo.)

2.1.5. *SQMC for $d_x > 1$*

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, $d_x = 1$. It is nevertheless possible to extend this approach to $d_x > 1$, 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, the interesting point 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 $x \in [0,1]^d$. The function H is obtained as a limit of the iterative process depicted by Figure 1. We refer to the book of [Sag94] for more details on the properties of space-filling curves.

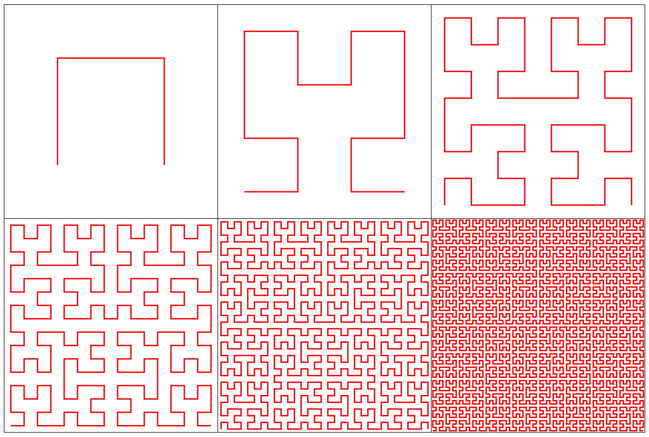

FIGURE 1. First steps of the iterative process, the limit of which is the Hilbert curve in two dimensions (Source: Wikipedia).

In the SQMC context, we will use h to transform the N ancestors into points in $[0,1]$, before using the inverse CDF as for $d_x = 1$. 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 $\mathbb{R}^{d_x}$ 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 [GC14] for a sharper description of this important point).

2.1.6. *Concluding remarks*

The main advantage of SQMC approach over standard particle filtering is the faster convergence, as $N \to \infty$. We refer to [GC14] 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).

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 $h \circ \psi(\mathbf{x}_{t-1}^{\sigma_{t-1}(1)}) \leq \dots \leq h \circ \psi(\mathbf{x}_{t-1}^{\sigma_{t-1}(N)})$ if $d \geq 2$, or
        $\mathbf{x}_{t-1}^{\sigma_{t-1}(1)} \leq \dots \leq \mathbf{x}_{t-1}^{\sigma_{t-1}(N)}$ 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 $n = 1, \dots, N$.
```

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.

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

Let us recall that $\ell(\boldsymbol{Y}|\boldsymbol{\theta})$ refers to the distribution (likelihood) of the random variable $\boldsymbol{Y}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is an unknown parameter that we wish to estimate, with a prior distribution π. 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 $\boldsymbol{y}_0 = (Y_1^0, \dots, Y_n^0)$.

Before we go into more details on ABC, some more notations are required. 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 dimension 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 $\boldsymbol{Y}$. For example, it can be a sufficient statistic for the parameter $\boldsymbol{\theta}$, but not necessarily. Assuming that the prior distribution on $\boldsymbol{\theta}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, 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 $\mathbf{s}_0 = \mathbf{S}(\boldsymbol{y}_0)$ the observed realization of $\mathbf{S}$ computed on the data set $\boldsymbol{y}_0$. Throughout the document, $\mathbf{s}_0$ and $\boldsymbol{y}_0$ should be considered as fixed quantities and N will be the number of simulations (or particles) simulated by the ABC algorithm. As stressed in [MPRR12], a classical formulation of ABC is the following one:

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 $\boldsymbol{y}_0$ and $\mathbf{s}_0$.
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 $\boldsymbol{y}^i = (Y_1^i, \dots, Y_n^i)$ 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 $\mathbf{s}_0$.
```

In practice, the parameter N should be chosen very large (typically of the order of 10^6), while k_N is commonly expressed as a percentile of N. Thus, for example, the choice $N = 10^6$ and a percentile $k_N/N = 0.1\%$ allow to retain 1000 simulated $\boldsymbol{\theta}_i$'s.

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 $(\boldsymbol{\theta}_1, \boldsymbol{y}^1), \dots, (\boldsymbol{\theta}_N, \boldsymbol{y}^N)$ an i.i.d. sample, with common joint distribution $\ell_n(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. This sample is naturally associated with the i.i.d. sequence $(\boldsymbol{\theta}_1, \mathbf{S}^1), \dots, (\boldsymbol{\theta}_N, \mathbf{S}^N)$, where each pair has a density $f(\mathbf{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Finally, let $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(k_N)}$ be the k_N-nearest neighbors of $\mathbf{s}_0$ among $\mathbf{S}^1, \dots, \mathbf{S}^N$, and let $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k_N)}$ 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 $(\boldsymbol{\theta}_1, \boldsymbol{y}^1), \dots, (\boldsymbol{\theta}_N, \boldsymbol{y}^n)$;
(2) Seconds, return the variables $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k_N)}$.

As will become clear in Section 2.2.1, 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 $\mathbf{s}_0 \in \mathbb{R}^m$, the estimate we will consider to infer the posterior density $g(.|\mathbf{s}_0)$ at some point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is

$$\hat{g}_N(\boldsymbol{\theta}_0) = \frac{1}{k_N h_N^p} \sum_{j=1}^{k_N} K\left(\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{(j)}}{h_N}\right), \tag{7}$$

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 $\mathbf{s}_0$, keeping in mind that $\mathbf{s}_0$ is held fixed. The idea is simple: in order to estimate the posterior, just look at the k_N-nearest neighbors of $\mathbf{s}_0$ and smooth the corresponding $\boldsymbol{\theta}_j$'s around $\boldsymbol{\theta}_0$. It should be noted that (7) 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 will **not** be 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], who develop a sequential scheme for scoring statistics according to whether their inclusion in the analysis will substantially improve the quality of inference. Similarly, we will 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 [BCG12] for details and proofs concerning the upcoming results.

2.2.1. *Distribution of* ABC *outputs*

We recall that $(\boldsymbol{\theta}_1, \mathbf{S}^1), \dots, (\boldsymbol{\theta}_N, \mathbf{S}^N)$ 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 $(\boldsymbol{\theta}_{(1)}, \mathbf{S}^{(1)}), \dots, (\boldsymbol{\theta}_{(k_N)}, \mathbf{S}^{(k_N)})$.

In what follows, we denote by d_i the (random) distance between $\mathbf{s}_0$ and $\mathbf{S}^i$. Similarly, we let $d_{(i)}$ be the distance between $\mathbf{s}_0$ and its i-th nearest neighbor among $\mathbf{S}^1, \dots, \mathbf{S}^N$, that is $d_{(i)} = \|\mathbf{S}^{(i)} - \mathbf{s}_0\|$. It turns out that, **conditionally on** $d_{(k_N+1)}$, one can consider the k_N-tuple $(\boldsymbol{\theta}_{(1)}, \mathbf{S}^{(1)}), \dots, (\boldsymbol{\theta}_{(k_N)}, \mathbf{S}^{(k_N)})$ as an ordered sample drawn according to the probability density

$$\frac{\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_0\| \leq d_{(k_N+1)}]} f(\boldsymbol{\theta}, \mathbf{s})}{\displaystyle\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{s}},$$

where $\mathcal{B}_m(\mathbf{s}_0, \delta)$ stands for the closed ball in $\mathbb{R}^m$ centered at $\mathbf{s}_0$ with nonnegative radius δ. Alternatively, the (unordered) simulated values may be treated like i.i.d. realizations of variables with common density proportional to $\mathbf{1}_{[\|\mathbf{s}-\mathbf{s}_0\| \leq d_{(k_N+1)}]} f(\boldsymbol{\theta}, \mathbf{s})$. Thus, given $d_{(k_N+1)}$, the accepted $\boldsymbol{\theta}_j$'s are i.i.d. realizations of the probability density

$$\frac{\displaystyle\int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\mathbf{s}}{\displaystyle\int_{\mathbb{R}^p} \int_{\mathcal{B}_m(\mathbf{s}_0, d_{(k_N+1)})} f(\vartheta, \mathbf{s}) \mathrm{d}\vartheta \mathrm{d}\mathbf{s}}.$$

Although this conclusion is intuitively clear, its proof requires a careful mathematical analysis (see [BCG12]). Moreover, it plays a key role in the mathematical analysis of the conditional density estimate (7) 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 $\mathbb{E}[\varphi(\boldsymbol{\theta}) | \mathbf{S} = \mathbf{s}_0]$, where the random variable $\varphi(\boldsymbol{\theta})$ is bounded. If $\boldsymbol{\theta}$ is itself bounded, it 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 $N \to \infty$, it can be shown the *pointwise consistency*, which means that for almost all $\mathbf{s}_0$ (with respect to the distribution of $\mathbf{S}$), with probability 1,

$$\frac{1}{k_N} \sum_{j=1}^{k_N} \varphi\left(\boldsymbol{\theta}_{(j)}\right) \to \mathbb{E}[\varphi(\boldsymbol{\theta}) | \mathbf{S} = \mathbf{s}_0]. \tag{8}$$

The proof of such a 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-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 proved 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 stating (8).

2.2.2. *Mean square error consistency*

Our next objective is to estimate the posterior density $g(\boldsymbol{\theta}_0 | \mathbf{s}_0)$, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. This estimation step is an important ingredient of the Bayesian analysis, whether this may be for visualization purposes or more involved mathematical achievements. As exposed in the introduction, a natural ABC-companion estimate of $g(\boldsymbol{\theta}_0 | \mathbf{s}_0)$ takes the form (7). Our goal in this section is to investigate some consistency properties of this estimate. Pointwise mean square error consistency is proved in Theorem 2.1 and mean integrated square error consistency is established in Theorem 2.2. We stress that this part of the document is concerned with minimal conditions of convergence. However, the following assumptions on the kernel will be 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 set [**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 2.1 (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*

$$\int_{\mathbb{R}^p} \int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \mathbf{s}) \log^+ f(\boldsymbol{\theta}, \mathbf{s}) \mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{s} < \infty. \tag{9}$$

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

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

It is easy to see that assumption (9) 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 2.2 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 2.2 (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 $\mathbf{s}_0 \in \mathbb{R}^m$, with $\bar{f}(\mathbf{s}_0) > 0$, if $k_N \to \infty$, $k_N/N \to 0$, $h_N \to 0$ and $k_N h_N^p \to \infty$,*

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

2.2.3. *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 keep trying to alleviate the assumptions on the 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 $\mathbf{s}_0$, the functions

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

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 $\mathbf{s}_0$ is called a Lebesgue point if

$$\frac{1}{\lambda_m(\mathcal{B}_m(\mathbf{s}_0, \delta))} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \left|\bar{f}(\mathbf{s}) - \bar{f}(\mathbf{s}_0)\right| \mathrm{d}\mathbf{s} \to 0 \quad \text{as } \delta \to 0.$$

Lebesgue's differentiation theorem asserts that this is true for λ_m-almost all $\mathbf{s}_0 \in \mathbb{R}^m$. If $\mathbf{s}_0$ is a Lebesgue point of $\bar{f}$ such that $\bar{f}(\mathbf{s}_0) > 0$, then it is readily seen that

$$0 < \xi_0 = \inf_{0<\delta\leq L} \frac{1}{\delta^m} \int_{\mathcal{B}_m(\mathbf{s}_0,\delta)} \bar{f}(\mathbf{s})\mathrm{d}\mathbf{s} < \infty.$$

Let us mention that Lebesgue points are commonly encountered when dealing with Nearest Neighbor rule. This was already pointed in the seminal work of [Dev82], and thereafter extended by considering "Besicovitch" conditions in [CG06]. Some recent developments in [GKM14] 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 2.3 (Rates of convergence)**.** *Suppose that assumptions* **[K1]**-**[K2]** *and* **[A1]**-**[A3]** *are satisfied. Let* $\mathbf{s}_0$ *be a Lebesgue point of* $\bar{f}$ *such that* $\bar{f}(\mathbf{s}_0) > 0$*. Denote*

$$\phi_1(\boldsymbol{\theta}_0, \mathbf{s}_0) = \frac{1}{2} \sum_{i_1,i_2=1}^{p} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\partial\theta_{i_1}\partial\theta_{i_2}} \int_{\mathbb{R}^p} \theta_{i_1}\theta_{i_2} K(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \quad \textit{and} \quad \Phi_1(\mathbf{s}_0) = \frac{1}{\bar{f}^2(\mathbf{s}_0)} \int_{\mathbb{R}^p} \phi_1^2(\boldsymbol{\theta}_0, \mathbf{s}_0)\mathrm{d}\boldsymbol{\theta}_0,$$

$$\phi_2(\boldsymbol{\theta}_0, \mathbf{s}_0) = \frac{1}{2m+4} \sum_{j=1}^{m} \frac{\partial^2 f(\boldsymbol{\theta}_0, \mathbf{s}_0)}{\partial s_j^2} \quad \textit{and} \quad \Phi_2(\mathbf{s}_0) = \frac{1}{\bar{f}^4(\mathbf{s}_0)} \int_{\mathbb{R}^p} \left[\phi_2(\boldsymbol{\theta}_0, \mathbf{s}_0)\bar{f}(\mathbf{s}_0) - \phi_3(\mathbf{s}_0) f(\boldsymbol{\theta}_0, \mathbf{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0$$

$$\phi_3(\mathbf{s}_0) = \frac{1}{2m+4} \sum_{j=1}^{m} \frac{\partial^2 \bar{f}(\mathbf{s}_0)}{\partial s_j^2} \quad \textit{and} \quad \Phi_3(\mathbf{s}_0) = \frac{2}{\bar{f}^3(\mathbf{s}_0)} \int_{\mathbb{R}^p} \phi_1(\boldsymbol{\theta}_0, \mathbf{s}_0) \left[\phi_2(\boldsymbol{\theta}_0, \mathbf{s}_0)\bar{f}(\mathbf{s}_0) - \phi_3(\mathbf{s}_0) f(\boldsymbol{\theta}_0, \mathbf{s}_0)\right] \mathrm{d}\boldsymbol{\theta}_0.$$

Then, for $m > 4$*, there exist sequences* $\{k_N\}$ *with* $k_N \propto N^{\frac{p+4}{m+p+4}}$ *and* $\{h_N\}$ *with* $h_N \propto N^{-\frac{1}{m+p+4}}$ *such that*

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \left[\hat{g}_N(\boldsymbol{\theta}_0) - g(\boldsymbol{\theta}_0|\mathbf{s}_0)\right]^2 \mathrm{d}\boldsymbol{\theta}_0\right] = \left(\frac{m\Phi_1(\mathbf{s}_0)}{\xi_0^{4/m}(m-4)} + \Phi_2(\mathbf{s}_0) + \frac{m\Phi_3(\mathbf{s}_0)}{\xi_0^{2/m}(m-2)} + \int_{\mathbb{R}^p} K^2(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \mathrm{o}(1)\right) N^{-\frac{4}{m+p+4}}.$$

Three concluding remarks are in order:

(1) 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 $g(.|\mathbf{s}_0)$. 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 $\mathbf{s}_0$ is fixed. Hence, devising a good methodology to automatically select both k_N and h_N in function of $\mathbf{s}_0$ necessitates a supplemental specific analysis.

(2) Nevertheless, Theorem 2.3 provides an insight into the proportion of simulated values which should be accepted by the algorithm. For example, a rough 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.

(3) 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 2.3. It is thus a basic fact to choose a sufficient statistic embedded in the lowest dimensional space possible.

3. Consistency for an example of nonparametric hidden Markov model

3.1. The studied model: hidden Markov models with finite state space

We now turn back to a specific case of HMMs introduced in Section 2.1, when then hidden component $(\mathbf{x}_t)_{t\in\mathbb{N}}$ lives in a finite state space. We are interested in Bayesian consistency results when the observation time (denoted n in what follows) is increasing. We would like to emphasize that such a result should be obtained in a different context as those stated in Section 1.3.1: observations $(\mathbf{y}_t)_{1\leq t\leq n}$ are no longer independent here and a significant amount of work is needed to reach Bayesian consistency.

Frequentist asymptotic properties of estimators of HMMs parameters have been studied since the 1990s. Consistency and asymptotic normality of the maximum likelihood estimator have been established in the parametric case, see [DM01], [DMR04] and [DMOvH11] for the most general consistency results up to now. As to parametric Bayesian asymptotic results, there are only a few recent results, see [dGS08] when the number of hidden states is known and [GR14a] when the number of hidden states is unknown. Because parametric modeling of emission distributions may lead to poor results in practice, in particular for clustering purposes, recent interest in using non parametric HMMs appeared in applications, see [YPRH11], [GCR14] and references therein. Theoretical results for estimation procedures in non parametric HMMs have been obtained only very recently such as in [DLC12] and in [GR13] since even identifiability remained an open problem (see [GCR14]).

The studied model is specified here and can be visualized in Figure 2. We still denote the HMMs $(\mathbf{x}_t, \mathbf{y}_t)_{t\in\mathbb{N}}$ where $\mathbf{x}$ is a homogeneous Markov chain whose transition kernel was previously denoted $f^X(\mathbf{x}_t|\mathbf{x}_{t-1})$. This kernel is now simply described as a squared matrix Q since we assume in this paragraph the finiteness of the state space $\mathcal{X}$ where $\mathbf{x}$ is living. In the meantime, the conditional probability distribution of $\mathbf{y}_t$ when $\mathbf{x}_t$ is given was previously denoted $f^Y(\mathbf{y}_t|\mathbf{x}_t)$ and is now shortened as $f_{\mathbf{x}_t}(\mathbf{y}_t)$.

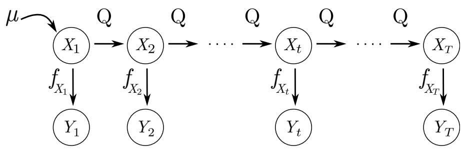

Figure 2. Schematic evolution of the HMM with a transition kernel Q and a conditional distribution $f_{\mathbf{x}}$ when $\mathbf{x}_1 \sim \mu$.

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

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

The former assumption on the transition kernel Q implies that the Markov chain $\mathbf{x}$ possesses a unique invariant distribution μ with an exponential mixing rate. In the meantime, we also assume the chain is initialized with its invariant distribution: $\mathbf{x}_1 \sim \mu$.

3.2. Prior structure

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

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

The transition matrix Q may be identified as a k-uple of transition distributions (the lines of the matrix), so that $Q \in \Delta_k(\underline{q})^k$. We denote $\mu \in \Delta_k(\underline{q})$ the invariant probability measure, that also initializes the Markov chain at time 1: $\mathbf{x}_1 \sim \mu$. We assume that the observation space is $\mathbb{R}^d$ endowed with its Borel sigma field. Let $\mathcal{F}$ 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 $\mathbf{x}_t$ to $\mathbf{y}_t$. It means that for any $f=(f_1,\ldots,f_k)\in\mathcal{F}^k$, the distribution of $\mathbf{y}_t$ conditionally to $\mathbf{x}_t = i$ will be $f_i d\lambda$ for each value of i between 1 and k.

Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (Q,f)\ :\ Q\in\Delta_k(\underline{q})^k, f\in\mathcal{F}^k\}$. Remark that a particular $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ implicitely defines a transition kernel Q and therefore a unique invariant distribution μ. For any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\mathbb{P}^{\boldsymbol{\theta}}$ denotes the probability distribution of $(\mathbf{x}_t,\mathbf{y}_t)_{t\in\mathbb{N}}$ when the transitions are parametrized by $\boldsymbol{\theta}$ and when the initial state $\mathbf{x}_1$ is distributed according to the invariant distribution μ.

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

$$p_l^{\boldsymbol{\theta}}(y_1,\ldots,y_l) = \sum_{(x_1,\ldots,x_l)\in[\![1,k]\!]^l} \mu_{x_1} Q_{x_1,x_2}\ldots Q_{x_{l-1},x_l} f_{x_1}(y_1)\ldots f_{x_l}(y_l).$$

Let π denotes a prior on $\boldsymbol{\Theta}$, we assume that π is a product of probability measures on $\boldsymbol{\Theta}$, $\pi = \pi_Q\otimes\pi_f$ such that π_Q is a probability distribution on $\Delta_k(\underline{q})^k$ and π_f is a probability distribution on $\mathcal{F}^k$.

3.3. Posterior consistency

3.3.1. *Topological description*

The observations are now distributed from $\mathbb{P}^{\boldsymbol{\theta}_0}$ where $\boldsymbol{\theta}_0 = (Q^0, f^0)$ so that the distribution of $(\mathbf{x}_t,\mathbf{y}_t)_{t\geq 1}$ follows a stationary HMM. We are interested in posterior consistency, that is to prove that for all neighborhood U of $\boldsymbol{\theta}_0$, with $\mathbb{P}^{\boldsymbol{\theta}_0}$ almost surely:

$$\lim_{n\to+\infty}\pi(U|\mathbf{y_n}) = 1.$$

To make the former equality meaningful, it is necessary to define a neighborhood concept of $\boldsymbol{\theta}_0$ and a topology has to be chosen for a precise definition of U. We choose to study posterior consistency for the problem of density estimation, *i.e.* we want to know if the posterior concentrates its mass around the parameters such that the associated distribution P_l^{θ} of l consecutive observations is closed to the one associated to the true parameter. We will use two different topologies as in Theorems 1.2 and 1.3. We first use the weak topology on marginal distributions $(P_l^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$.

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

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

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

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

We will also work with the finer topology associated to the L_1-distance on the joint densities. Other topologies may be considered depending on the estimation needed, see for example [Ver13] where a product of the topologies for the transition matrix and the emission densities is also used.

3.3.2. *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 $\mathbf{y}_{t-2}, \ldots, \mathbf{y}_1$. The generalization of the Hoeffding inequalities [Rio00] requires a level of mixing of the chain to ensure an exponential rate of concentration (and then the existence of a powerful test between two hypotheses). Since $\boldsymbol{\theta}_0$ is such that $Q^0 \in \Delta_k(\underline{q})^k$ with a known $\underline{q}$ (non adaptive prior on $\underline{q}$), we will only consider prior π_Q such that

$$\pi_Q\left\{\Delta_k(\underline{q})^k\right\} = 1 \text{ and } \min_{1\le i\le k} \mu_i \ge \underline{q}. \tag{11}$$

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 here) and admit a unique stationary probability measure.

Theorem 3.1 describes a set of assumptions that lead to the posterior consistency for the weak topology, and may be compared to Theorem 1.2. The following assumption on the neighborhood of $\boldsymbol{\theta}^0$ is used:

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

$$\|Q - Q^0\| < \epsilon, \tag{12a}$$

$$\max_{1\le i\le k} \int f_i^0(y) \max_{1\le j\le k} \log\left(\frac{f_j^0(y)}{f_j(y)}\right) \lambda(dy) < \epsilon, \tag{12b}$$

$$\text{For all } y \in \mathbb{R}^d \text{ such that } \sum_{i=1}^k f_i^0(y) > 0 \Longrightarrow \sum_{j=1}^k f_j(y) > 0, \tag{12c}$$

$$\sup_{y\,:\,\sum_{i=1}^k f_i^0(y)>0} \max_{1\le j\le k} f_j(y) < +\infty, \tag{12d}$$

$$\sum_{i=1}^k \int f_i^0(y) \left|\log\left(\sum_{j=1}^k f_j(y)\right)\right| \lambda(dy) < +\infty \tag{12e}$$

Theorem 3.1. *[Ver13] Assume that the prior π satisfies* (11) *and that* ***Assumption [N]*** *holds, then for all weak neighborhood U of $P_l^{\boldsymbol{\theta}_0}$ (see* (10)*),*

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

Theorem 3.1 is proved in [Ver13] 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 closed 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 1.3), an additional assumption on the covering number implies the existence of tests that permit to discriminate $\mathbb{P}^{\boldsymbol{\theta}}$ to $\mathbb{P}^{\boldsymbol{\theta}_0}$ 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

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

Theorem 3.2. *[Ver13] Assume that the prior π satisfies* (11) *and that* ***Assumption [N]*** *and* ***Assumption [H]*** *hold, then for all L_1-neighborhood U of $P_l^{\boldsymbol{\theta}_0}$, there exists $r > 0$ such that*

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

Thanks to the similarities of Assumptions (12b) and (13) with Assumptions (2) and (3) 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 [Ver13] for instance in the case of translated emission distributions that is to say when for all $1 \leq j \leq k$,

$$f_j(\cdot) = g(\cdot - m_j)$$

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

4. The PAC-Bayesian paradigm

4.1. Generality 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. 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

$$R(f_{\hat{\boldsymbol{\theta}}}) = \mathbb{E}[\ell(\mathbf{y}, f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}))], \quad R_n(f_{\hat{\boldsymbol{\theta}}}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_{\hat{\boldsymbol{\theta}}}(X_i)). \tag{14}$$

Let $R^\star = R(f_{\boldsymbol{\theta}^0})$, $R^\star$ is of course 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:

$$\forall \varepsilon \in (0,1) \qquad \mathbb{P}\left[R(f_{\hat{\boldsymbol{\theta}}}) - R^\star \leq \mathrm{K} \inf_{\boldsymbol{\theta}}\Big\{R(f_{\boldsymbol{\theta}}) - R^\star + \Delta_{n,d,M,\varepsilon}(\boldsymbol{\theta})\Big\}\right] \geq 1 - \varepsilon, \tag{15}$$

where $K \geq 1$ is a constant and $\Delta_{n,d,M,\varepsilon}$ 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 $\Delta_{n,d,M,\varepsilon}$ 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 naturally drives us to consider an additive model of the form

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

where $\mathbf{m} = (m_1,\ldots,m_d) \in \{0,\ldots,M\}^d$ is a model, $\mathbb{D} = \{\phi_1,\phi_2,\ldots,\phi_M\}$ 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. As usual, 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 naturally wish to use a prior probability measure $\pi\in\mathcal{M}$ promoting sparsity. For this purpose, we consider the following constrained optimization problem:

$$\operatorname*{arg\,min}_{\rho\in\mathcal{M}} \left\{ \int_{\boldsymbol{\Theta}} R_n(f_{\boldsymbol{\theta}})\rho(\mathrm{d}\boldsymbol{\theta}) + \frac{\lambda}{n}\mathcal{KL}(\rho,\pi) \right\}, \tag{16}$$

where the Kullback-Leibler divergence is defined as

$$\forall\rho\in\mathcal{M} \qquad \mathcal{KL}(\rho,\pi) = \int_{\boldsymbol{\Theta}} \log\left[\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\boldsymbol{\theta})\right]\rho(\mathrm{d}\boldsymbol{\theta}).$$

Indeed, the (frequentist) variational formulation of (16) 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 (16) 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. From 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 shown 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 4.1 ([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 $\int(\exp\circ h)\mathrm{d}\mu < \infty$,*

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

with the convention $\infty-\infty=-\infty$. Further, 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 following concentration inequality (see *e.g.* [Mas07]).

Lemma 4.2 (Bernstein's inequality)**.** *Let $(T_i)_{i=1}^n$ be a collection of real independent random variables. Assume there exist two positive constants v and w such that*

$$\sum_{i=1}^n \mathbb{E}T_i^2 \leq v,$$

and for any integer $k \geq 3$,

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \leq \frac{k!}{2} v w^{k-2}.$$

Then, for any $\gamma \in \left(0, \frac{1}{w}\right)$,

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^n (T_i - \mathbb{E}T_i)\right)\right] \leq \exp\left(\frac{v\gamma^2}{2(1-w\gamma)}\right).$$

PAC-Bayesian bounds depend on the Kullback-Leibler divergence and hold for any prior π. In order to obtain an optimized PAC-Bayesian estimator, two levers are at our disposal: the inverse temperature parameter λ and the prior π. These two key quantities must be well-tailored to obtain some good oracle inequalities. In particular, we may consider a sparsity-inducing prior, such as

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

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 nice property to favor sparse parameters (it gives a highest mass to the parameters with a low ℓ^0 norm of the coefficients $(m_j)_{1\leq j\leq d}$).

4.2. Examples

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

4.2.1. *Regression models*

We consider the standard model

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

with two mild assumptions.

Assumption 1: 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 2: $|\psi^\star|_\infty \leq C$.

In particular, Assumption 1 is met if $\mathbf{w}$ has a Gaussian distribution. As for Assumption 2, it allows to use concentration inequalities such as Lemma 4.2. From what precedes, we obtain the following oracle inequality.

Theorem 4.3 ([GA13]). *For any $\varepsilon \in (0,1)$, any $0 < \lambda < n/(4\sigma^2 + 4C^2)$, with probability at least $1-\varepsilon$,*

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

where $\mathrm{K}_\lambda \xrightarrow[\lambda \to 0]{} 1$ and $\mathrm{K}_\lambda \xrightarrow[\lambda \to n/(4\sigma^2 + C^2)]{} +\infty$.

4.2.2. *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 $\phi_1, \phi_2, \dots$ 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 4.4 ([GA13]). *For any real $\varepsilon \in (0,1)$, any $0 < \lambda < n/(4\sigma^2 + 4C^2)$, with probability at least $1-\varepsilon$,*

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

The message carried by these inequalities is that if there exists a sparse representation of the regression function, then the right-hand side terms 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.

4.2.3. *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].

4.2.4. *Binary Ranking*

Note that the PAC-Bayesian tools can also be usedto 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.

4.3. 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-Bayesian Pred**iction*)[1].

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.

[ALW12] P. Alquier, X. Li, and O. Wintenberger. Prediction of Time Series by Statistical Learning: General Losses and Fast Rates. Preprint, 2012.

[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 lapprentissage : 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.

[BCG12] G. Biau, F. Cérou, and A. Guyader. New Insights into Approximate Bayesian Computation. *arXiv:1207.6461*, 2012.

[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.

[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.

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

[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.

[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.

[GC14] M. Gerber and N. Chopin. Sequential Quasi-Monte Carlo. *Journal of the Royal Statistical Society, series B (in press)*, 2014.

[GCR14] E. Gassiat, A. Cleynen, and S. Robin. Finite state space non parametric Hidden Markov Models are in general identifiable. *Statistics and Computing, to appear*, 2014.

[GCSR04] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

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

[GKM14] S. Gadat, T. Klein, and C. Marteau. Classification with the nearest neighor rule in general finite dimensional spaces. *Preprint*, pages 1–53, 2014.

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

[GR13] E. Gassiat and J. Rousseau. Non parametric finite translation mixtures with dependent regime. *arXiv preprint arXiv:1302.2345*, 2013.

[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 problme de ranking binaire en grande dimension. In *46mes Journées de Statistique de la SFdS, Rennes*, 2014.

[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.

[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.

[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.

[MR07] J.-M. Marin and C. P. Robert. *Bayesian core: a practical approach to computational Bayesian statistics.* Springer Texts in Statistics. Springer, New York, 2007.

[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.

[Ver13] E. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *ArXiv e-prints*, November 2013.

[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.