In this paper we introduce perturbed Bayesian inference, a new Bayesian based approach for online parameter inference. Given a sequence of stationary observations \((Y_t)_{t \geq 1}\), a parametric model \(\{f_\theta, \theta \in \mathbb{R}^d\}\) and \(\theta^* := \arg\max_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\log f_\theta(Y_1)\right]\), the sequence \((\tilde{\pi}^N_t)_{t \geq 1}\) of perturbed posterior distributions has the following properties: (i) \(\tilde{\pi}^N_t\) does not depend on \((Y_s)_{s > t}\); (ii) the time and space complexity of computing \(\tilde{\pi}^N_t\) from \(\tilde{\pi}^N_{t-1}\) and \(Y_t\) is at most \(cN\), where \(c < +\infty\) is independent of \(t\); and (iii) for \(N\) large enough \(\tilde{\pi}^N_t\) converges almost surely as \(t \to +\infty\) to \(\theta^*\) at rate \(\log(t)^{(1+\varepsilon)/2}t^{-1/2}\), with \(\varepsilon > 0\) arbitrary and under classical conditions that can be found in the literature on maximum likelihood estimation and on Bayesian asymptotics.

**Keywords:** Bayesian inference, online inference, streaming data

1 Introduction

In many modern applications a large number of observations arrive continuously and need to be processed in real time, either because it is impracticable to store the data or because a decision should be made and/or revised as soon as possible as you data arrives. This is for instance the case with digital financial transactions data, where the number of observations per day frequently exceeds the million and where online fraud detection is of obvious importance (Zhang et al., 2018). In this context, the notion of data stream is more appropriate than that of a dataset, which supposes infrequent updates (O’callaghan et al., 2002). Following Henzinger et al. (1998), we informally refer to a data stream as a sequence of observations that can be read only once and in the order in which they arrive. The data stream model is also relevant for large datasets, where the number of observations is such that each of them can only be read a small number of times for practical considerations (O’callaghan et al., 2002).

Beyond computations of simple descriptive statistics, statistical inference from data streams is a challenging task. This is particularly true for parameter estimation in parametric models, the focus of this paper. Indeed, current approaches to online parameter...
estimation either lack theoretical guarantee (such as streaming variational Bayes methods, [Broderick et al., 2013]), or require a computational effort that grows with the sample size (such as the IBIS algorithm of [Chopin, 2002] or are valid only for a very small class of models (such as methods based on stochastic approximations, see e.g. [Toulis et al., 2017; Zhou and Said 2018], or the conditional density filtering algorithm of [Guhaniyogi et al. 2018]). We refer the reader to the latter reference for a more detailed literature review on that topic.

Let \((Y_t)_{t \geq 1}\) be a stationary sequence and let \(\mathcal{M} = \{f_\theta, \theta \in \mathbb{R}^d\}\) be a parametric model for \(Y_1\). We let \(\theta^\star \in \mathbb{R}^d\) denote the parameter value that minimizes the Kullback-Leibler (KL) divergence between the true distribution of \(Y_1\), denoted by \(f^\star\), and \(\mathcal{M}\). We do not assume the model to be well specified, i.e. that \(f_{\theta^\star} = f^\star\). If we let \(\pi_0\) be a prior distribution for \(\theta\), then the posterior distribution associated with the observations \((Y_s)_{t \geq 1}\), for any \(t > 0\), can be written recursively as

\[
\pi_t := \Psi_t(\pi_{t-1})
\]

for some mapping \(\Psi_t\) depending only on \(Y_t\) (see Definition 1 below). Hence, Bayesian inference is well suited to streaming data. Moreover, \(\pi_t\) has good convergence properties, and notably, it converges to \(\theta^\star\) (in the sense of Definition 2 below) at rate \(t^{-\alpha}\) almost surely, for all \(\alpha \in (0, 1/2)\), under regularity conditions (e.g. [Sriram et al., 2013]). Unfortunately, these desirable properties of Bayesian inference can rarely be exploited in practice. Indeed, \(\Psi_t\) is in general intractable and thus computing \(\pi_t\) usually requires an infinite computational budget.

In this paper we introduce perturbed Bayesian inference (PBI) as a practical alternative to Bayesian inference for online parameter estimation. The proposed method provides an algorithmic definition of sequences of perturbed posterior distributions (PPDs) in which each iteration incorporates information coming from a new observation. Perturbed Bayesian inference has strong theoretical guarantees for a large class of statistical models, with our main result establishing that a finite computational budget is enough to enable the almost sure convergence of the sequence of PPDs to \(\theta^\star\) at rate \(t^{-\alpha}\), for every \(\alpha \in (0, 1/2)\). Moreover, with a communication requirement of \(O(\log(t))\) to process \(t\) observations, the algorithm underpinning the definition of the PPDs can be efficiently parallelized. This is important given the ever-increasing emphasis on parallel architectures in modern computing systems.

Perturbed Bayesian inference is based on two simple observations that hold if \(\pi_0\) has a finite support. In such a case \(\pi_t\) is (i) fully computable but (ii) will usually not concentrate on \(\theta^\star\) as \(t\) increases. The idea of PBI is thus to approximate \(\pi_0\) by \(\tilde{\pi}_0^N\) having a finite support of size \(N \geq 1\) to take advantage of (i) and, in order to correct for (ii), to propagate \(\tilde{\pi}_0^N\) using a perturbed version \((\tilde{\Psi}_t)_{t \geq 1}\) of the sequence \((\Psi_t)_{t \geq 1}\), where the perturbations enable the support of \(\tilde{\pi}_t^N := \tilde{\Psi}_t \circ \cdots \circ \tilde{\Psi}_1(\tilde{\pi}_0^N)\) to evolve over time.

The perturbed Bayes updates \((\tilde{\Psi}_t)_{t \geq 1}\) explicitly depend on the properties of the Bayes updates and, in particular, assume that \((\Psi_t)_{t \geq 1}\) is such that \((\pi_t)_{t \geq 1}\) converges to \(\theta^\star\) at the usual \(t^{-1/2}\) rate. Notice that this property of Bayes updates is not only an assumption for our convergence results to hold but is assumed in the algorithmic definition of the
PPDs. Conditions under which the posterior distribution converges at that rate can be found e.g. in Kleijn and van der Vaart (2012) for i.i.d. observations and in Ghosal and van der Vaart (2007) for non i.i.d. data.

In the following subsections we summarize the main properties of the proposed PPDs, discuss some related approaches, and finally conclude this introduction with the general set-up that we will consider throughout the paper.

1.1 Summary of the main results

We define the perturbed Bayes updates \( (\tilde{\Psi}_t)_{t \geq 1} \) in a manner which ensures that, for a given \( \tilde{\pi}_N^0 \), the sequence \( (\tilde{\pi}_N^t)_{t \geq 1} \) of PPDs, where \( \tilde{\pi}_N^t = \tilde{\Psi}_t(\tilde{\pi}_N^{t-1}) \) for all \( t > 0 \), has the following properties:

Property 1. For any \( t \geq 1 \), \( \tilde{\pi}_N^t \) is independent of \( (Y_s)_{s > t} \).

Property 2. For any \( t \geq 1 \), provided that \( f_{\theta}(y) \) can be computed at a finite cost, the time and space complexity of computing \( \tilde{\pi}_N^t \) from \( (\tilde{\pi}_N^{t-1}, Y_t) \) is at most \( cN \), where \( c < +\infty \) is independent of \( t \).

Property 3. Under classical conditions in maximum likelihood estimation and Bayesian asymptotics, for any \( N \) large enough and i.i.d. observations \( (Y_t)_{t \geq 1} \), \( \tilde{\pi}_N^t \) converges almost surely as \( t \to +\infty \) to \( \theta_* \) at rate \( \log(t)^{(1+\varepsilon)/2}t^{-1/2} \), with \( \varepsilon > 0 \) arbitrary.

Property 1 ensures the applicability of the method to streaming data while Property 2 ensures that the computation cost and memory requirements do not grow with the number of observations. Property 3 is the main convergence property of PBI and has important practical corollaries, such as the almost sure convergence of the point estimate \( \hat{\theta}_N^t := \int \theta \tilde{\pi}_N^t(d\theta) \) to \( \theta_* \) at rate \( \log(t)^{(1+\varepsilon)/2}t^{-1/2} \). Under the classical conditions mentioned in Property 3, \( \pi_t \) converges to \( \theta_* \) at rate \( t^{-1/2} \) in probability while both \( \pi_t \) (Proposition 4 below) and \( \tilde{\pi}_N^t \) converge to \( \theta_* \) at rate \( t^{-\alpha} \) almost surely, for every \( \alpha \in (0, 1/2) \).

Property 2 implies that, in the definition of the perturbed Bayes update \( \tilde{\Psi}_t \), the likelihood function \( \theta \mapsto f_{\theta}(Y_t) \) is the only information about \( \theta_* \) that \( Y_t \) is assumed to contain. In particular, the differentiability of \( f_{\theta} \) is not required, making PBI applicable on a wide class of models. This weak requirement has however a cost in terms of computational resources needed for Property 3 to hold. As explained in Section 2.2 knowing just \( \theta \mapsto f_{\theta}(Y_t) \) enables us to establish Property 3 only for \( N \geq K_*^d \), where \( K_* \geq 2 \) is some problem specific integer.

Although we refer to \( (\tilde{\Psi}_t)_{t \geq 1} \) as the sequence of perturbed Bayes updates and to \( \tilde{\pi}_N^t \) as the PPD to simplify the presentation, we define in fact a whole class of perturbed Bayes updates and corresponding PPDs in this paper. A particular instance of the perturbed Bayesian approach, that borrows ideas from mean-field variational inference (Blei et al., 2017) and that we term mean-field perturbed Bayesian inference by analogy, is considered for the numerical study of Section 4. Mean-field PBI can be implemented for any \( N \geq 2d \) and numerical results show that the corresponding PPDs may have good convergence behaviour for values of \( N \) much smaller than \( 2^d \).
1.2 Related approaches

The approach introduced in this work falls in the category of particle based methods, where a set of $N$ random variables $\{\theta_t^n\}_{n=1}^N$, or particles, are propagated as new observations arrive. In current particle based methods for online inference (e.g. Chopin (2002)) the particles $\{\theta_t^n\}_{n=1}^N$ are used to define an approximation $\hat{\pi}_t^N$ of the posterior distribution $\pi_t$, in the sense that $\hat{\pi}_t^N \to \pi_t$ as $N \to +\infty$. This convergence generally comes at the price of increasing time and space complexity as $t$ increases since it requires to revisit all the past observations each time a new particle system is generated (see e.g. Crisan et al. (2018) for attempts to solve this issue). Requiring that $\hat{\pi}_t^N$ approximates $\pi_t$ for all $t \geq 1$ also comes at the price of high communication cost. For instance, in Chopin (2002), Crisan et al. (2018) a certain function of the likelihood evaluations $\{f_{\theta_t}(Y_t)\}_{n=1}^N$, called the effective sample size, is computed at each time step to ensure that the approximation error does not explode as $t$ increases. This operation requires interaction between particles and the communication cost thus induced may reduce considerably the gain brought by parallel computations. On the contrary, PBI does not aim at approximating the posterior distribution at a finite time; i.e. in general we do not have $\hat{\pi}_t^N \to \pi_t$ as $N \to +\infty$. This is one of the reasons why Property 2 holds and why the algorithm defining $\hat{\pi}_t^N$ can be efficiently parallelized.

The last important remark to be made is that the concentration of $\hat{\pi}_t^N$ on $\theta_*$ is guaranteed only in the limiting regime $N = +\infty$ while in PBI we only need $N$ to be sufficiently large.

1.3 Set-up and preliminaries

All random variables are defined on a probability space $((\Omega, \mathcal{F}, \mathbb{P})$ with associated expectation operator $\mathbb{E}$. We denote by $\mathcal{B}(A)$ the Borel $\sigma$-algebra on the set $A$. The observations form a sequence $\{Y_t\}_{t \geq 1}$ of random variables where each $Y_t$ takes values in a measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. Hereafter the index $t$ of the observations is referred to as the time. The sequence $\{Y_t\}_{t \geq 1}$ is assumed to be a stationary process under $\mathbb{P}$ with a common marginal distribution $f_\theta(y)dy$, where $dy$ is some $\sigma$-finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. Let $\{f_\theta, \theta \in \Theta\}$ be the statistical model of interest, such that for any $\theta \in \Theta \subset \mathbb{R}^d$, $f_\theta$ is a probability density function on $\mathcal{Y}$ (w.r.t. $dy$). Below we denote by $d\theta$ the Lebesgue measure on $(\Theta, \mathcal{B}(\Theta))$ and by $\mathcal{P}(\Theta)$ the set of probability measures on $(\Theta, \mathcal{B}(\Theta))$.

The following assumption is considered to hold throughout the paper.

Assumption A1. The Kullback-Leibler divergence $KL : \Theta \to [0, +\infty)$, defined by $KL(\theta) := \mathbb{E}[\log(f_\theta(Y_1)/f_{\theta_0}(Y_1))]$, $\theta \in \Theta$, has a unique minimum at $\theta_* \in \Theta$. NB: We do not require $f_{\theta_*} = f_\theta$.

To simplify the presentation we assume that $\Theta = \mathbb{R}^d$ and that $f_\theta(y) > 0$ for all $(\theta, y) \in \Theta \times \mathcal{Y}$. The latter condition ensures that the recursion (1) is well-defined for any prior distribution $\pi_0 \in \mathcal{P}(\Theta)$ and any realization $(y_t)_{t \geq 1}$ of $(Y_t)_{t \geq 1}$, while the assumption $\Theta = \mathbb{R}^d$ will greatly simplify the algorithmic definition of the PPDs introduced in the next section. However, neither of these conditions is needed for the method to be applicable nor is required for our theoretical results to hold.
The mappings \((\Psi_t)_{t \geq 1}\) introduced in \([1]\) are formally defined as follows.

**Definition 1.** For integer \(t \geq 1\), the mapping \(\Psi_t : \Omega \times \mathcal{P}(\Theta) \to \mathcal{P}(\Theta)\) defined by

\[
\Psi_t^\omega(\eta)(d\theta) := \frac{f_\theta(Y_t(\omega))}{\int_{\Theta} f_{\theta'}(Y_t(\omega)) \eta(d\theta')} \eta(d\theta), \quad (\omega, \eta) \in \Omega \times \mathcal{P}(\Theta)
\]

is called the Bayes update associated with observation \(Y_t\). For any integers \(0 \leq s < t\) we write \(\Psi_{s:t} := \Psi_t \circ \cdots \circ \Psi_{s+1}\).

Let \(\pi_0 \in \mathcal{P}(\Theta)\) be a prior distribution. Then, by \([1]\) and Definition \([1]\) we have \(\pi_t = \Psi_{0:t}(\pi_0)\) and \(\pi_t = \Psi_{s:t}(\pi_s)\) for all \(s \in \{0, \ldots, t-1\}\) and \(t \geq 1\).

The formal definition of the convergence rate of a probability measure to a singular point, mentioned above, is as follows.

**Definition 2.** Let \((\eta_t)_{t \geq 1}\) be a random sequence of probability measures on \(\Theta\). We say that \((\eta_t)_{t \geq 1}\) converges to \(\theta_*\) are rate \(r_t \to 0\) in probability (resp. almost surely) if for any sequence \((M_t)_{t \geq 1}\) in \(\mathbb{R}_{\geq 0}\) such that \(\lim_{t \to +\infty} M_t = +\infty\), we have, \(\eta_t(\|\theta - \theta_*\| \geq M_t r_t) \to 0\) as \(t \to +\infty\), where the convergence holds in probability (resp. almost surely).

Under suitable assumptions on \((Y_t)_{t \geq 1}\) and the Bayesian model at hand it has been proved that \(\Psi_{0:t}(\pi_0)\) converges to \(\theta_*\) at rate \(t^{-1/2}\) in probability while a strictly slower rate is needed to ensure the almost sure convergence of \(\pi_t\) to \(\theta_*\) (see e.g. Ghosal and van der Vaart \([2007]\)). In particular, under Assumptions \([A1-A5]\) used to establish Property \([3]\) with \([A2-A5]\) stated in Section \([3]\), for every \(\alpha \in (0, 1/2)\) the posterior distribution \(\pi_t\) converges to \(\theta_*\) at rate \(t^{-\alpha}\) almost surely, as stated in the next result.

**Proposition 1.** Assume \([A1-A2]\) and let \(\pi_0 \in \mathcal{P}(\Theta)\) be such that \(\pi_0(U) > 0\) for some neighbourhood \(U\) of \(\theta_*\). Then, \(\pi_t\) converges to \(\theta_*\) at rate \(t^{-1/2}\) in probability and, for every \(\alpha \in (0, 1/2)\), \(\pi_t\) converges to \(\theta_*\) at rate \(t^{-\alpha}\) almost surely.

Proposition \([1]\) is a consequence of Kleijn and van der Vaart \([2012]\) Theorem 3.1 and Theorem 3.3) and hence its proof is omitted.

The condition in Proposition \([1]\) that \(\pi_0\) puts positive mass on a neighbourhood \(U\) of \(\theta_*\) is necessary for the convergence of \(\Psi_{0:t}(\pi_0)\) towards \(\theta_*\). In particular, for a prior distribution \(\mu = N^{-1} \sum_{n=1}^{N} \delta_{\theta_n}\) with a finite support, where \(\delta_{\theta}\) denotes a point mass located at \(\theta\), it is not difficult to check that as \(t\) increases \(\Psi_{0:t}(\mu)\) concentrates on the element of \(\{\theta_0, \ldots, \theta_N\}\) minimising the KL divergence between \(f_\theta\) and \(f_\theta\), that is,

\[
\Psi_{s:(s+t)}(\mu) = \delta_{\tilde{\theta}_0}, \quad \text{where } \tilde{\theta}_0 \in \arg\min_{\theta \in \{\theta_0, \ldots, \theta_N\}} KL(\theta) \quad \mathbb{P} - \text{a.s.}, \quad \forall s \geq 0. \quad (2)
\]

### 1.4 Additional notation and outline of the paper

Let \(|x|\) be the maximum norm of the vector \(x \in \mathbb{R}^d\), \(|A|\) be the matrix norm of the \(d \times d\) matrix \(A\) induced by the Euclidean norm and \(|C|\) be the cardinality of the set \(C \subset \Theta\). For \(\epsilon > 0\) and \(\theta \in \Theta\) let \(B_\epsilon(\theta)\) be the ball of size \(\epsilon > 0\) for the maximum norm and, for
For a given initial distribution \( \pi_0^N = \frac{1}{N} \sum_{n=1}^{N} \delta_{\theta_0^n} \) the algorithmic definition of the corresponding sequence \((\tilde{\pi}_t^N)_{t \geq 1}\) of PPDs is given in Algorithm 1. Algorithm 1 depends on random mappings \(F, \tilde{F}, G, \tilde{G}\) whose roles will be first informally explained in Section 2.1 after which rigorous formal specifications are given in the form of Conditions C1-C4 in Section 2.2.

2 Perturbed Bayesian inference

As mentioned in the introductory section, the main idea of PBI is to propagate \(\tilde{\pi}_0^N\) using a perturbed version \((\Psi_t)_{t \geq 1}\) of the sequence \((\tilde{\pi}_t^N)_{t \geq 1}\), where the perturbations enable the support of \(\Psi_{0:t}(\tilde{\pi}_0^N)\) to be updated over time. To guarantee that the resulting sequence of PPDs, as defined in Algorithm 1, concentrates on \(\theta_*\), we choose these perturbations to be sufficiently significant to ensure a proper exploration of the parameter space and yet sufficiently insignificant to preserve the concentration property (2).

Our strategy for controlling the magnitude of the perturbations in order to find the right balance is to define a perturbation schedule \((t_p+1)_{p \geq 1}\) which determines the times when the actual perturbations occur. More formally, we define \(\Psi_t = \Psi_t^*\) for all \(t \geq 1\) such that \(t \notin (t_p+1)_{p \geq 1}\) and we let \(\Psi_t\) be updated after \((t_p+1)_{p \geq 1}\). By doing this, \((\Psi_t)_{t \geq 1}\) will retain the concentration property (2) of \((\tilde{\pi}_t^N)_{t \geq 1}\) between consecutive perturbation times while the occasional perturbations provide opportunities to correct the support of the PPDs towards \(\theta_*\). In Algorithm 1 an explicit definition of \((t_p)_{p \geq 1}\) is given on Line 2 and is justified in the next subsection. Notice that \((t_p)_{p \geq 1}\) is such that \(\log(t_p) = \mathcal{O}(p)\), i.e. the time between consecutive perturbations increases exponentially fast in \(p\). Since, in Algorithm 1 the operations performed at every time \(t \in (t_p+2):t_{p+1}\) are embarrassingly parallel, the definition of \((t_p)_{p \geq 1}\) is therefore such that the communication requirement of PBI to process \(t\) observations is \(\mathcal{O}(\log(t))\).

At every perturbation time \(t_p + 1\) the support \(\theta_t^{1-N}\) of \(\tilde{\pi}_{t_p+1}^N\) is computed before the information of the latest observation \(Y_{t_p+1}\) is incorporated (hence we write \(\theta_t^{1-N}\) instead of \(\theta_t^{1-N}\)). The size of the support does not vary over time, and thus the support of \(\tilde{\pi}_t^N\) contains \(N\) points for all \(t \geq 0\).
Algorithm 1 Perturbed Bayes posterior distributions

Input: Integers \((N, M, t_1) \in \mathbb{N}^3\), real numbers \((\varepsilon, \epsilon_0, \beta) \in \mathbb{R}_{>0}^3\), \(\kappa \in (0, 1)\) and \(\rho > 2\), starting values \(\theta_0^{1:N} \in \Theta^N\) and \(\vartheta_0^{1:(N+M)} \in \Theta^{N+M}\)

// Define
1: Let \(\tilde{N} = N + M\), \(\tilde{\chi} = (N, M, \kappa, \beta)\)
2: Let \(t_0 = 0\) and \(t_p = t_{p-1} + \lfloor (\kappa^{-2} - 1)t_{p-1} \rfloor \lor t_1\) for \(p \in \mathbb{N}\)
3: Let \(\epsilon_p = \epsilon_0 (1 \land (p^{-1} \rho \log(p+1))^{\frac{1}{\rho+3}})\) for \(p \in \mathbb{N}\)
4: Let \(c_0 = 1\) and \(c_p = \left(\frac{1 + \kappa}{2\kappa}\right)^p \lor p^{\frac{1}{\rho+2}}\) for \(p \in \mathbb{N}\)

// Initialization
5: Let \(\bar{t}_p = \tilde{N}^{-1} \sum_{n=1}^{\tilde{N}} \vartheta_0^n, w_0^{1:N} = (1, \ldots, 1), \tilde{w}_0^{1:\tilde{N}} = (1, \ldots, 1), q_0 = 0\) and \(\xi_p = p = 1\)
6: for \(t \geq 1\) do
7: if \(t = t_p + 1\) then
8: // Perturbation
9: Let \(\bar{t}_p = G(t_p, \bar{w}_p^{1:N}, \tilde{\theta}_p^{1:N}, \tilde{\chi})\) and \(\bar{t}_p = \tilde{G}(t_p, \epsilon_{p-1}, \bar{t}_p^{1:1}, \tilde{w}_p^{1:\tilde{N}}, \tilde{\theta}_p^{1:\tilde{N}}, \tilde{\chi})\)
10: if \(\|\bar{t}_p - \bar{t}_p\| \leq 2\epsilon_p\) then
11: Let \(q_p = q_{p-1} + 1\), \(\xi_p = \kappa(c_{p-1}/c_p)\) and \(\bar{t}_p = \bar{t}_p\)
12: else
13: Let \(q_p = 1\), \(\xi_p = \epsilon_p\) and \(\bar{t}_p = \bar{t}_p\)
14: end if
15: Let \(\theta_1^{1:N} = F(t_p, \bar{t}_p, \xi_p, N)\) and \(\theta^{1:N} = \tilde{F}(t_p, \bar{t}_p, \epsilon_p, \tilde{\chi})\)
16: Let \(w_p^{1:N} = (1, \ldots, 1)\) and \(\tilde{w}_p^{1:\tilde{N}} = (1, \ldots, 1)\)
17: \(p \leftarrow p + 1\)
18: end if
19: // Bayes updates
20: Let \(\tilde{\pi}_p^n = \tilde{w}_p^{n-1} f_{\theta_p}^n(Y_t)\) for \(n \in 1:\tilde{N}\)
21: Let \(w_p^n = w_p^{n-1} f_{\theta_p}^{n-1}(Y_t)\) for \(n \in 1:N\)
22: Let \(\tilde{\pi}_p^n = \sum_{n=1}^{\tilde{N}} \frac{w_p^n}{\sum_{m=1}^{\tilde{N}} w_p^m} \delta_{\theta_p}^n\)
An important feature of Algorithm 1 is that the PPD support is updated by means of auxiliary PPDs. To understand why this mechanism is introduced, assume that we want to compute \( \theta_{t_p}^{1:N} \) from \( \tilde{\pi}_{t_p}^N \) only. For now, assume \( \Theta \) compact, that \( \bar{\theta}_{t_p} \) is the expectation of \( \theta \) under \( \tilde{\pi}_{t_p}^N \) and that \( \theta_{t_p}^{1:N} \overset{\text{ld}}{\rightarrow} N_d(\bar{\theta}_{t_p}, \sigma_p^2 I_d) \) for some \( \sigma_p > 0 \). Then, to guarantee that \( \tilde{\pi}_{t_p}^N \) becomes increasingly concentrated as \( t \to +\infty \), one must clearly have \( \sigma_p \to 0 \) as \( p \to +\infty \).

On the other hand, a classical argument in the literature on global optimization literature (see e.g. Ingber and Rosen, 1992, Section 3.) imposes to have \( \sigma_p \geq f_N(t_p)^{-1/d} \), for some strictly increasing function \( f_N : \mathbb{R}_+ \to \mathbb{R}_+ \), in order to guarantee that the concentration of \( \tilde{\pi}_{t_p}^N \) is towards the right target \( \theta_* \). Since the subsequence \( (\tilde{\pi}_{t_p+1}^N)_{p \geq 1} \) cannot converge to \( \theta_* \) at a rate faster than \( \sigma_p \), which depends on the dimension, we end up with a sequence of PPDs with poor statistical properties.

To avoid this limitation, in Algorithm 1 the exploration of the parameter space is performed by a sequence of auxiliary PPDs. Its interactions with \( \tilde{\pi}_{t_p}^N \) guides the sequence of PPDs towards \( \theta_* \) and, this way, the rate at which the support of \( \tilde{\pi}_{t_p}^N \) can concentrate over time (the scaling parameter \( \sigma_p \) in the above example) can be made independent of the dimension of the parameter space, as in Property 3.

This mechanism is explained with more details in Section 2.1.2.

### 2.1.1 Perturbation scheduling depending on the radius of the support

The definition of the perturbation schedule \((t_p + 1)_{p \geq 1}\) is closely connected with the definition of the support of the PPD in Algorithm 1 or more specifically with the radius of the support, as we will now explain.

On Line 14 of Algorithm 1 the support \( \theta_{t_p}^{1:N} \) of \( \tilde{\pi}_{t_{p+1}}^N \) is defined through a random mapping \( E \) in a manner (see Condition C1 in Section 2.2.1) which guarantees \( \theta_{t_p}^{1:N} \) being almost surely contained in a ball of radius \( \xi_p \) such that

\[
\max_{n \in 1:N} \min_{n \neq m} \| \theta_{t_p}^n - \theta_{t_p}^m \| \leq r_p := \frac{\xi_p}{K},
\]

where \( K \geq 2 \) is an integer depending on \( N \). Therefore, the elements of \( \theta_{t_p}^{1:N} \) can be interpreted to represent a discretization of a ball of radius \( \xi_p \) with a resolution defined by \( K \) (large \( K \) corresponds to high resolution).

Let \( n_* = \arg\min_{n \in 1:N} KL(\theta_{t_p}^n) \) and assume that \( \| \theta_{t_p}^{n_*} - \theta_* \| \leq r_p \). Then, by (2), the probability measure \( \Psi_{(t_p + 1);(t_p + s)}(\tilde{\pi}_{t_p+1}^N) \) concentrates on \( \theta_{t_p}^{n_*} \) as \( s \) increases, implying that, for an arbitrary \( \delta \in (0, 1) \), the probability \( \tau_{t_p+s}^{\delta} \) that the mass of \( \Psi_{(t_p + 1);(t_p + s)}(\tilde{\pi}_{t_p+1}^N) \) on the set \( B_{\xi_p/K}(\theta_*) \) is at least \( \delta \) converges to one. For fixed \( s \), \( \tau_{t_p+s}^{\delta} \) is an increasing function of \( r_p \). Therefore, for a given \( \tau \in (0, 1) \), the number of observations \( s_{r_p} \) needed to ensure that \( \tau_{t_p+s_{r_p}}^{\delta} \geq \tau \) decreases with \( r_p \). In other words, fewer observations are required with low resolution. The precise value of \( s_{r_p} \) depends on the convergence properties of the Bayes updates \( \Psi_{t_{p+1}} \) and, in PBI, it is assumed that \( \Psi_{t_{p+1}} \) enables the posterior distribution to converge to \( \theta_* \) at rate \( t^{-1/2} \).

With this in mind, the perturbation schedule \((t_p)_{p \geq 1}\) and the radii \((\xi_p)_{p \geq 1}\) are defined on Lines 2, 10 and 12 so that, for every \( \delta > 0 \) and \( p \geq 1 \), the difference \( t_{p+1} - t_p \) is such...
that $\tau_{t_p+(t_{p+1}-t_p)}^\delta \geq \tau_p^\delta$ for a suitable choice of sequence $(\tau_p^\delta)_{p \geq 1}$ that converges to one as $p \to +\infty$. For reasons explained in Section 2.1.2, $(\tau_p^\delta)_{p \geq 1}$ must converge sufficiently fast to ensure that the convergence rate of the PPD induced by $(t_p)_{p \geq 1}$ and $(\xi_p)_{p \geq 1}$ holds almost surely. Algorithmically, the dependency between $(t_p)_{p \geq 1}$ and $(\xi_p)_{p \geq 1}$ is manifested by the parameter $\kappa$ appearing in the definitions of both $(t_p)_{p \geq 1}$ and $(\xi_p)_{p \geq 1}$, and we see that $\xi_p$ is increasing and $t_{p+1} - t_p$ is decreasing in $\kappa$.

2.1.2 Updating the support of the perturbed posterior distribution

The support of the PPD is updated by adjusting its radius and location. In Algorithm 1, this is done on Lines 9–13 which also define how the information from the auxiliary PPDs is incorporated.

The sequence $(\tilde{\eta}_p)_{p \geq 1}$ of auxiliary PPDs, with support of size $\tilde{N} = N + M$ (with $M \geq 1$ being parameter of Algorithm 1), is defined by

$$\tilde{\eta}_t = \sum_{n=1}^{\tilde{N}} \frac{\tilde{w}_t^n}{\sum_{m=1}^{\tilde{N}} \tilde{w}_t^m} \delta_{\tilde{\theta}_t^{\eta}}(\tilde{\theta}_t), \quad p \in \mathbb{N}, \ t \in \{t_{p-1} + 1, \ldots, t_p\}$$

where $\tilde{w}_t^n$ is for all $n \in 1: \tilde{N}$ and $t \geq 1$ as defined on Line 18 of Algorithm 1 and $\tilde{\theta}_t^{\eta}$ is computed according to Line 14 through a random mapping $\tilde{F}$. This mapping is such that $\tilde{\eta}_t^{\tilde{N}}$ depends on $\tilde{\eta}_p$ only and hence, as explained before, the subsequence $(\tilde{\eta}_p)_{p \geq 1}$ can only concentrate on $\theta_*$ at a poor, dimension dependent, rate. However, this is not a cause for concern as the role of these auxiliary PPDs is only to find the neighbourhood of $\theta_*$ and to guide the sequence $(\tilde{\pi}_t^N)_{t \geq 1}$ towards it, as we will now explain.

Under Condition C1 below, the support of $\tilde{\pi}_{t_p+1}^N$ is almost surely included in the ball $B_{\tilde{\xi}_p}(\tilde{\theta}_p)$, whose radius $\tilde{\xi}_p$ and centre $\tilde{\theta}_p$ are determined by using information from $\tilde{\eta}_p$ (Lines 9–13). The centre $\tilde{\theta}_p$ is set equal to one of two estimates of $\theta_*$, namely $\hat{\theta}_p$ and $\bar{\theta}_p$, that are computed on Line 8 by random functions $G$ and $\tilde{G}$ and are derived from $\tilde{\pi}_p$ and $\tilde{\eta}_p$, respectively. The decision on which estimate to choose as the centre of $\tilde{\theta}_p$ is based on the distance $|\tilde{\eta}_p - \tilde{\theta}_p|$ according to the following rationale.

To fix the ideas, let us assume that $P(\tilde{\eta}_p \in B_{\tilde{\xi}_p}(\theta_*)) = 1$. While this assumption may not always hold, it should be noted that the sequence $(\epsilon_p)_{p \geq 1}$, on Line 3, is defined to converge to zero at a very slow rate to ensure the fast convergence of $P(\tilde{\eta}_p \in B_{\tilde{\xi}_p}(\theta_*))$ to one as $p \to +\infty$. In this case, $|\tilde{\eta}_p - \tilde{\theta}_p| > 2\epsilon_p$ implies that $\tilde{\theta}_p \notin B_{\tilde{\xi}_p}(\theta_*)$, and hence $\tilde{\theta}_p$ is closer to $\theta_*$ than $\tilde{\theta}_p$. Therefore, the support of $\tilde{\pi}_{t_p+1}^N$ is centred at $\tilde{\theta}_p = \bar{\theta}_p$, and the radius $\tilde{\xi}_p$ of the ball in which it is contained is set to $\epsilon_p$, the level of precision of the estimate $\tilde{\theta}_p$. If, on the other hand, $|\tilde{\theta}_p - \tilde{\eta}_p| \leq 2\epsilon_p$ we cannot conclude from the available information whether $\tilde{\theta}_p$ is a better estimate than $\hat{\theta}_p$, i.e., whether $|\tilde{\theta}_p - \theta_*|$ is larger than $|\tilde{\theta}_p - \theta_*|$ or not. Then, on Line 14 the information coming from $\tilde{\eta}_p$ is not used to compute the support and $\tilde{\pi}_{t_p+1}^N$ will be centred at $\tilde{\theta}_p = \hat{\theta}_p$. In this case, $\xi_p$ will be set to $\kappa(\epsilon_p/\epsilon_{q_p-1})\xi_{p-1}$, which requires some further explanation.

The variable $q_p$, defined on Lines 10 and 12, counts the number of consecutive perturbations whereby the information from the auxiliary PPD has not been used, i.e. $p - q_p$
is the largest integer $s \in 1:(p - 1)$ for which $\xi_s = \epsilon_s$. Then, by using the definition of $(c_p)_{p \geq 1}$ on Line 4, which depends on the parameter $\varepsilon > 0$, we have
\[
\xi_p = \mathcal{O}\left( \log(t_p)^{\frac{1+s}{2}} t_p^{-1/2} \right)
\] (3)
as $q_p \to +\infty$ (i.e. if $p = q_p + s$ for some $s \geq 1$), which explains the logarithmic factor appearing in Property 3.

To understand the rate in (3), and hence the definition of $(c_p)_{p \geq 0}$, remark that $q_p \to +\infty$ if an only if $\xi_p = \epsilon_p$, or equivalently, $\|\hat{\theta}_p - \bar{\theta}_p\| > 2\epsilon_p$, for only finitely many $p \geq 1$. This condition intuitively requires that the two estimators $\hat{\theta}_p$ and $\bar{\theta}_p$ converge to $\theta_*$ almost surely, and more importantly, as $\hat{\theta}_p$ is derived from $\hat{\pi}_t^N$ only, that $(\hat{\pi}_t^N)_{t \geq 1}$ converges to $\theta_*$ almost surely. As shown in Proposition 4 under the assumptions used to establish Property 3, the posterior distribution converges almost surely to $\theta_*$ at rate $t^{-\alpha}$, where $\alpha \in (0,1/2)$ is arbitrary. The support of $\hat{\pi}_t^N$ being included in a ball of size $\xi_p$, we see that the rate in (3) is optimal in the sense that, for arbitrary $\alpha \in (0,1/2)$, the PPD also converges to $\theta_*$ at rate $t^{-\alpha}$, almost surely.

Notice that by the definition on Line 4 $(c_p)_{p \geq 0}$ is strictly increasing and $c_p K^p < 1$ for all $p \geq 1$. These two conditions are not necessary but facilitate the theoretical analysis.

2.2 Random mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$

So far we have informally introduced the roles played by these random functions used in Algorithm 4. Functions $G$ and $\tilde{G}$ generate estimates of $\theta_*$ that are used for parametrising the generation of the supports for both PPD processes, $(\hat{\pi}_t^N)_{t \geq 1}$ and $(\tilde{\eta}_t)_{t \geq 1}$, which in turn is implemented by the random mappings $F$ and $\tilde{F}$, respectively. We will now make the definition of $\hat{\pi}_t^N$ more precise by imposing formal conditions on $F$, $\tilde{F}$ $G$ and $\tilde{G}$ that are notably designed to enable Properties 3 to hold.

As mentioned in the introductory section, by Property 2, PBI only requires the ability to evaluate $f_\theta$ point-wise to be used in practice. This imposes the condition $N \geq 2^d$ to guarantee the concentration of $\hat{\pi}_t^N$ on $\theta_*$. To understand this constraint, assume that $\hat{\theta}_p \in B_{\xi_p}(\theta_*)$ in Algorithm 4. Then, a necessary condition to ensure that, for some $\delta \in (0,1)$, the perturbed posterior distribution $\hat{\pi}_t^N$ can have a nearly unit mass on $B_{\xi_p}(\theta_*)$ is that at least one point in its support belongs to $B_{\xi_p}(\theta_*)$. Since no information about how to move towards $\theta_*$ from $\hat{\theta}_p$ is assumed, the only option is to require that, for some integer $K \geq \delta^{-1}$, each of the $K^d$ hypercubes of equal volume that partition $B_{\xi_p}(\theta_*)$ is assigned a positive mass by $\hat{\pi}_t^N$. This requires to take $N \geq K^d \geq 2^d$.

The conditions introduced in the following are assumed to hold for arbitrary
\[
\mu \in \Theta, \ (t, N, M) \in \mathbb{N}^3, \ (w^{1:N}, w^{1:\tilde{N}}, \kappa, \beta, \epsilon) \in \mathbb{R}^{3+N+\tilde{N}}, \ (\theta^{1:N}, \theta^{1:\tilde{N}}) \in \Theta^{N+\tilde{N}}
\]
where we recall that $\tilde{N} = N + M$. We also define $\tilde{\chi} = (N, M, \kappa, \beta)$ and
\[
K_N := \max \{k \in \mathbb{N} : N \geq k^d\}.
\] (4)

10
2.2.1 Conditions on the support generation mappings $F$ and $\tilde{F}$

**Condition C1.** The random variable $\theta_1^{1:N} := F(t, \mu, \epsilon, N)$ takes values in $\Theta^N$ and satisfies, for some $C \in \mathbb{R}_{>0}$,

1. $\theta_1^{1:N}$ is independent of $(Y_s)_{s>t}$.

2. $\mathbb{P}(1 \leq \sum_{n=1}^{N} 1(\theta_1^n \in B_{j\epsilon/K_N(\mu)}) \leq C) = 1$ for all $j \in 1:K^d$ and for $N \geq 2^d$, with $K_N$ defined in (4).

3. $\mathbb{P}(\forall n \in 1:N, \theta_1^n \in B_{\epsilon}(\mu)) = 1$ for $N \geq 2^d$.

Condition C1 is required to ensure Property 1. Under C1.2, each of $K_N^d$ hypercubes of equal size partitioning $B_\epsilon(\mu)$ contains at least one (see the discussion above) and at most $C$ elements of $\theta_1^{1:N}$. The upper bound $C$ is used to guarantee that increasing $M$ and $N$ does not deteriorate the statistical properties of the PPD (Theorem 3 below). Finally, C1.3 implies that, for every $N \geq 2$ and with probability one, the support of $\hat{\pi}_{t_0+1}$ is included in the set $B_{\epsilon}(\hat{\theta}_{t_0})$. This condition stabilizes the behaviour of the PPDs between perturbation times $t_p + 1$ and $t_{p+1} + 1$ and thus enables convergence results for the whole sequence $(\hat{\pi}_n^{N})_{t \geq 1}$ instead of only for the subsequence $(\hat{\pi}_{t_p}^{N})_{p \geq 1}$.

We also introduce the following additional condition which is of particular interest.

**Condition C1*.** The $\Theta^N$ valued random variable $\theta_1^{1:N} := F(t, \mu, \epsilon, N)$ is such that, for $N \geq 2^d$ and with $K_N$ as in (4), $\mathbb{P}(\exists n \in 1:N, \theta_1^n = z_j) = 1$ for every $j \in 1:K^d$, where $z_j$ denotes the centroid of $B_{j\epsilon/K_N(\mu)}$.

Under C1*, with probability one, $\hat{\pi}_{t_p+1}(B_{\epsilon p/K}(\theta_*)) > 0$ whenever $\hat{\theta}_{t_p} \in B_{\epsilon p}(\theta_*)$ while, in that case, C1 only guarantees that with probability one, $\hat{\pi}_{t_p+1}(B_{2\epsilon p/K}(\theta_*)) > 0$. A consequence of this simple observation is that the extra condition C1* enables us to establish Property 3 for $N \geq K^d$ while, assuming C1 only and for the same constant $K_*$, we need $N \geq (2K_*)^d$ (Theorem 2 below).

**Condition C2.** The random variable $\tilde{\theta}_1^{1:}\tilde{N} := \tilde{F}(t, \mu, \epsilon, \tilde{\chi})$ takes values in $\Theta^{N}$ and satisfies, for some $C \in \mathbb{R}_{>0}$,

1. $\tilde{\theta}_1^{1:}\tilde{N}$ is independent of $(Y_s)_{s>t}$.

2. $\mathbb{P}(1 \leq \sum_{n=1}^{N} 1(\theta_1^n \in B_{j\epsilon/K_N(\mu)}) \leq C) = 1$ for all $j \in 1:K^d$ and for $N \geq 2^d$, with $K_N$ defined in (4).

3. $\tilde{\theta}_1^{N+1} \sim t_\nu(g(\mu), \Sigma_i)$ is generated independently with $\nu > 0$, $g : \Theta \rightarrow \Theta$ a measurable function such that $g(\Theta)$ is a bounded set and $\Sigma_i$ is a random variable, taking values in the space of symmetric positive definite matrices, such that $\mathbb{P}(\|\Sigma_i\|^1/2 \geq \gamma_i, \|\Sigma_i\| < C, \|\Sigma_i\| \|\Sigma_i^{-1}\| < C) = 1$ with $(\gamma_i)_{t \geq 1} \in \mathbb{R}_{>0}$ such that $\lim_{p \rightarrow +\infty} \gamma_i^{-\nu}(p^{-1} \log(p+1)) \tilde{\pi}_t = 0$.
Condition \(C_2.1\) and \(C_2.2\) are similar to \(C_1.1\) and \(C_1.2\). Condition \(C_2.3\) is imposed to ensure that, for every perturbation time \(t_p + 1\), we have \(\tilde{\eta}_{t_p+1}(B_{\epsilon_p}(\theta_*)) > 0\) with a sufficiently high probability to guide the support of the PPD towards \(\theta_*\). The rationale behind the \(g\) function is to guarantee that this probability is lower bounded uniformly in \(\bar{t}_p\). For instance, \(g\) can be defined as

\[
g(\theta) = \arg\min_{\theta' \in [-L,L]^d} \|\theta' - \theta\|, \quad \theta \in \Theta
\]  

(5)

where \(L \in \mathbb{R}_{>0}\). Conditions on the sequence \((\gamma_t)_{t \geq 1}\) are trivially met when \(\|\Sigma_t\|^{1/2} = \gamma\) for all \(t\) and some \(\gamma > 0\). Taking a sequence \((\Sigma_t)_{t \geq 1}\) satisfying \(\underline{C}_2\) for some \((\gamma_t)_{t \geq 1}\) that converges to zero as \(t \to +\infty\) is necessary to enable the whole sequence \((\tilde{\eta}_t)_{t \geq 1}\) of auxiliary PPDs to concentrate on \(\theta_*\), and not only the subsequence \((\tilde{\eta}_{t_p})_{p \geq 1}\). However, such a requirement is unnecessary, since only the subsequence interacts with \((\tilde{\pi}_t^N)_{t \geq 1}\).

Notice that under \(C_2\) the last \(M - 1\) elements of the support \(\bar{v}_{t_p}^{N+M}\) of \(\tilde{\eta}_{t_p+1}\) only need to be independent of the observations \((Y_t)_{t > t_p}\). Thus, any estimator of \(\theta_*\) based on observations \((Y_t)_{t=1}^{t_p}\) can be incorporated in the support of the auxiliary PPD, with the goal of finding quickly a small neighbourhood of \(\theta_*\). For instance, one can choose \(M = 2\) and, for some \(N_{\exp} \gg 1\) and exploratory random variables \(\bar{v}_{t_p}^{1:N_{\exp}}\) independent of \((Y_t)_{t > t_p-1}\), set

\[
\bar{v}_{t_p}^{N} = \bar{v}_{t_p}^{n_p}, \quad n_p \in \arg\max_{n \in 1:N_{\exp}} \prod_{s=t_p+1}^{t_p} f_{\tilde{\eta}_s}(Y_s).
\]  

(6)

Remark 1. Under \(C_2.1\) and \(C_2.2\) the sequence \((\bar{v}_{t_p}^{1:N})_{p \geq 1}\) can explore the parameter space since \(\sum_{p=1}^{\infty} \epsilon_p = +\infty\). However, these two conditions are not sufficient to ensure the convergence of \((\tilde{\eta}_p)_{p \geq 1}\) to \(\theta_*\), notably when the KL divergence is multi-modal.

The following extra condition is introduced for the same reason we considered the extra condition \(C_1.2\) for the mapping \(F\).

Condition \(C_2.3\). The \(\Theta^N\) valued random variable \(\bar{v}_t^{1:N} := \tilde{F}(t, \mu, \epsilon, \bar{\chi})\) is such that, for \(N \geq 2^d\) and with \(K_N\) as in \(\{\}\), \(\mathbb{P}(\exists n \in 1:N, \bar{v}_t^n = z_j) = 1\) for every \(j \in 1:K_N^d\), where \(z_j\) denotes the centroid of \(B_{\epsilon}\).

2.2.2 Conditions on the point-estimate mappings \(G\) and \(\tilde{G}\)

Condition \(C_3\). The random variable \(\bar{\theta}_t := G(t,\bar{w}^{1:N},\bar{\theta}^{1:N},N)\) takes values in \(\Theta\) and satisfies

1. \(\bar{\theta}_t\) is independent of \((Y_s)_{s > t}\).
2. \(\mathbb{P}(\bar{\theta}_t = \sum_{n=1}^{N} \frac{w^n}{\sum_{m=1}^{N} w^m} \bar{\theta}^n) = 1\) for \(N \geq 2^d\).
Condition C3.1 is similar to C1.1 and C2.1 while C3.2 implies that $\bar{\theta}_p$ is the expectation of $\theta$ under the perturbed posterior distribution $\pi^{N}_{t_p}$ when $N \geq 2^d$. While Theorem 1 which establishes Property 3 may still hold for other definitions of $\bar{\theta}_p$, Condition C3.2 has the advantage of enabling the derivation of a convergence result for the PPD that holds uniformly for all $M \geq 1$ and $N$ sufficient large (Theorem 3 below).

We first introduce the following simple condition on $\tilde{G}$.

**Condition C4*.** The random variable $\tilde{\vartheta}_t := \tilde{G}(t, \epsilon, \mu, \tilde{w}^{1: N}, \vartheta^{1: N}, \tilde{x})$ takes its values in $\Theta$ and satisfies

1. $\tilde{\vartheta}_t$ is independent of $(Y_s)_{s>t}$.
2. $\mathbb{P}(\tilde{\vartheta}_t = \vartheta^{n'}, n' \in \text{argmax}_{n \in 1: N} \tilde{w}^n) = 1$ for $N \geq 2^d$.

Under $\mathcal{C}4^*$, the estimate $\tilde{\vartheta}_p$ is equal to the mode of the auxiliary perturbed posterior distribution $\tilde{\eta}_p$. As we shall see below, our main convergence result, Theorem 1 holds under $\mathcal{C}4^*$. However, $\mathcal{C}4^*$ is not sufficient to study theoretically whether increasing $M$ or $N$ can have a negative impact on the statistical properties of $\tilde{\eta}_p$. Indeed, whenever $\tilde{\eta}_{p-1}(B_{\epsilon_p}(\vartheta_*)) > 0$ there is in general a positive probability that the point in the support of $\tilde{\eta}_p$ having the largest weight does not belong to $B_{\epsilon_p}(\vartheta_*)$ and thus, under $\mathcal{C}2$ and $\mathcal{C}4^*$, that $\tilde{\eta}_{p-1}(B_{\epsilon_p}(\vartheta_*)) = 0$. Given the random nature of the likelihood function $\theta \mapsto \prod_{s=t_{p-1}+1}^{t_p} f_\theta(Y_s)$, and the weak assumptions imposed on $\hat{\epsilon}$, it is difficult to rule out the possibility of the probability of this undesirable event being an increasing function in $M$ or $N$. More precisely, the difficulty under $\mathcal{C}4^*$ is to find a sequence $(\delta_p)_{p \geq 1}$ such that $\delta_p \to 1$ and, for all $M \geq 1$ and $N$ sufficiently large,

$$\mathbb{P}(\tilde{\vartheta}_p \in B_{\epsilon_p}(\vartheta_*) \mid \tilde{\vartheta}_{p-1} \in B_{\epsilon_p}(\vartheta_*)) \geq \delta_p, \quad \forall p \geq 1. \quad (7)$$

The existence of such a sequence $(\delta_p)_{p \geq 1}$ is needed in our analysis to ensure that the convergence behaviour of the PPD does not deteriorate as $M$ or $N$ increases.

The following Condition C4 on $\tilde{G}$ enables us to derive a convergence result for the PPD that holds uniformly for all $M \geq 1$ and $N$ large enough. Moreover, this condition motivates the introduction of the specific mapping $\tilde{G}$ considered in the numerical study of Section 4.

**Condition C4.** The random variable $\tilde{\vartheta}_t := \tilde{G}(t, \epsilon, \mu, \tilde{w}^{1: N}, \vartheta^{1: N}, \tilde{x})$ takes its values in $\Theta$ and satisfies, for some $\Delta \in (0, 1)$, $(\zeta_1, \zeta_3) \in \mathbb{R}^2_{> 0}$ and $(\zeta_2, \zeta_4) \in (0, 1)^2$,

1. $\tilde{\vartheta}_t$ is independent of $(Y_s)_{s>t}$.
2. $\mathbb{P}(\tilde{\vartheta}_t = \tilde{\vartheta}_t^{(1)} 1(Z_t > \Delta) + \tilde{\vartheta}_t^{(2)} 1(Z_t \leq \Delta)) = 1$ for $N \geq 2^d$, where $\tilde{\vartheta}_t^{(2)} = \vartheta^{n'}$, with $n' \in \text{argmax}_{n \in 1: N} \tilde{w}^n$,

$$Z_t = \frac{\sum_{n=1}^{N} \tilde{w}^n \mathbb{1}(\vartheta^{n} \in B_{1+\kappa}^{(\Delta)}(\mu))}{\sum_{m=1}^{N} \tilde{w}^m}, \quad \tilde{\vartheta}_t^{(1)} = \frac{\sum_{n \in 1: N \cup J} \hat{a}_n(J) \tilde{w}^n \vartheta^{n}}{\sum_{m \in 1: N \cup J} \hat{a}_m(J) \tilde{w}^m}$$
and, with $J = \{ n \in (N + 1): \tilde{N} s. t. \vartheta^n \in B_{(1+2\epsilon)}(\mu) \}$,

$$a_n = \begin{cases} \frac{\zeta_1 M}{N}, & n \in 1:N \\ \zeta_2 M, & n = N + 1, \tilde{a}_n(J) = \frac{\zeta_4 (1/|J|)}{N}, & n \in J \cap \{N + 1\} \\ (1 - \zeta_2), & n > N + 1 \end{cases}$$

The stabilizing effect of $C_4$ is somewhat technical. Further details are given in Section S2 of the Supplementary Material A to explain why, under $C_3$, the inequality (7) holds for some $\delta_p \to 1$ and uniformly in $M \geq 1$ and $N$ sufficiently large.

### 3 Convergence results

#### 3.1 Assumptions

We shall consider the following assumptions on the model $\{f_\theta, \theta \in \Theta\}$ and on the true distribution $f^\star$ of the observations.

**Assumption A2.** The observations $(Y_t)_{t \geq 1}$ are i.i.d. under $P$.

**Assumption A3.** There exists an open neighbourhood $U$ of $\theta^\star$ and a square-integrable function $m_{\theta^\star}: \mathcal{Y} \to \mathbb{R}$ such that, for all $\theta_1, \theta_2 \in U$,

$$\left| \log \frac{f_{\theta_1}(Y_1)}{f_{\theta_2}(Y_1)} \right| \leq m_{\theta^\star}(Y_1) \|\theta_1 - \theta_2\|, \quad \mathbb{P} - a.s.$$  

We assume that $\mathbb{E}[f_\theta(Y_1)/f_{\theta^\star}(Y_1)] < +\infty$ for all $\theta \in U$ and that $\mathbb{E}[e^{s m_{\theta^\star}(Y_1)}] < +\infty$ for some $s > 0$.

**Assumption A4.** For every $\epsilon > 0$ there exists a sequence of measurable functions $(\phi_t)_{t \geq 1}$, with $\phi_t: \mathcal{Y} \to \{0, 1\}$, such that

$$\lim_{t \to +\infty} \mathbb{E}[\phi_t(Y_1)] = 0, \quad \lim_{t \to +\infty} \sup_{t: \|\theta - \theta^\star\| \geq \epsilon} \mu^t_\theta(1 - \phi_t) = 0,$$

where, for $\theta \in \Theta$ and $t \geq 1$, we denote by $\mu^t_\theta$ the measure on $\mathcal{Y}$ defined by $\mu^t_\theta(A) = \mathbb{E}[1_A(Y_1:t) \prod_{s=1}^t \frac{f_\theta(Y_s)}{f_{\theta^\star}(Y_s)}]$ for all $A \in \mathcal{B}(\mathcal{Y})$.

Assumptions A3 and A4 are borrowed from Kleijn and van der Vaart (2012), where a Bernstein-von-Mises theorem for misspecified models is established. Kleijn and van der Vaart (2012) use these conditions to show the existence for misspecified models of tests for complements of shrinking balls around $\theta^\star$ versus $f^\star$ that have a uniform exponential power. Such tests are classical tools to establish a concentration rate for the posterior distributions. Following closely Kleijn and van der Vaart (2012, Theorem 3.1), these tests are used in a similar a way to study precisely the convergence property (2) of Bayes updates (Lemma S2 in the Supplementary Material A) on which the concentration of the PPD between two successive perturbation times relies, as explained in Section 2.1.
Assumption A5. For every $y \in \mathcal{Y}$ the function $\theta \mapsto \log f_{\theta}(y)$ is three times continuously differentiable on some neighbourhood $U$ of $\theta_*$, with first derivative $\dot{\theta}_{\theta}(y) \in \mathbb{R}^d$ and second derivative $\ddot{\theta}_{\theta}(y) \in \mathbb{R}^{d \times d}$, and there exists a measurable function $\dddot{m}_{\theta} : \mathcal{Y} \rightarrow \mathbb{R}^+$ such that, for all $\theta \in U$,

$$\left| \frac{\partial^3 \log f_{\theta}(Y_1)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| \leq \dddot{m}_{\theta_0}(Y_1), \quad \forall (i, j, k) \in \{1, \ldots, d\}^3, \quad \mathbb{P} - a.s. \quad (8)$$

We assume that $\mathbb{E}[\dddot{m}_{\theta}(Y_1)] < +\infty$, the matrix $\mathbb{E}[\dot{\theta}_{\theta}(Y_1)\dot{\theta}_{\theta}(Y_1)^T]$ is invertible and the matrix $\mathbb{E}[\ddot{\theta}_{\theta}(Y_1)]$ is negative definite.

Assumption A6. $\lim_{t \to +\infty} \mathbb{P}(\sup_{\theta \in \Theta} \left| \frac{1}{t} \sum_{s=1}^{t} \log f_{\theta}(Y_s) - \mathbb{E}[\log f_{\theta}(Y_1)] \right| \geq \epsilon) = 0$ for all $\epsilon > 0$.

Assumption A7. $\|\hat{\theta}_{t, \text{mle}} - \theta_*\| = O_p(t^{-\frac{1}{2}})$ where, for every $t \geq 1$, $\hat{\theta}_{t, \text{mle}} : \mathcal{Y}^t \rightarrow \Theta$ is defined by

$$\hat{\theta}_{t, \text{mle}}(y) = \arg\max_{\theta \in \tilde{U}} \sum_{s=1}^{t} \log f_{\theta}(y_s), \quad y \in \mathcal{Y}^t.$$ with $\tilde{U}$ a closed ball containing a neighbourhood of $\theta_*$.

Assumptions A5-A7 are used to ensure that, for every $p \geq 1$, the mode of the auxiliary perturbed posterior distribution $\tilde{m}_{\theta_p}$ is in the ball $B_{\epsilon_p}(\theta_*)$ with sufficiently large probability. This property is needed to show that $\mathbb{P}(\tilde{m}_{\theta_p} \in B_{\epsilon_p}(\theta_*))$ converges quickly to one (Theorem S1 in the Supplementary Material A) and thus that the auxiliary PPDs efficiently guide the sequence $(\tilde{\pi}_N^T)_{t \geq 1}$ towards $\theta_*$, as explained in Section 2.1.2.

Assumption A5 contains some of the classical conditions to prove the asymptotic normality of the MLE (van der Vaart, 1998, Section 5.6, p.67) while Assumption A6 is a standard requirement to establish the consistency of this estimator (van der Vaart, 1998, Theorem 5.7, page 45). It would be appealing to replace assumptions A5 and A6 by modern, weaker, assumptions such as those presented in van der Vaart (1998, Chapter 5) for establishing the consistency and asymptotic normality of the MLE. One of the main merits of these assumptions is to not require the differentiability of $f_{\theta}$, a property that is not used in our definition of PPDs. However, in the context of this paper it is not clear how they can be used instead of A5 and A6 and we leave this question for future research. Lastly, Kleijn and van der Vaart (2012, Lemma 2.2) provide conditions under which A7 holds.

### 3.2 Convergence rate of perturbed posterior distributions

Our main result is the following theorem which establishes that Property [3] holds under the assumptions listed in Section 3.1.

**Theorem 1.** Assume A1-A7. Then, there exists a finite constant $L_* \geq 1$, that depends only on the model $\{f_{\theta}, \theta \in \Theta\}$ and on $f_\star$, such that under $C_1$, $C_2$ and for all $N \geq \left(2\inf\{k \in \mathbb{N} : k > \kappa^{-1}L_*\}\right)^d$ the sequence of perturbed posterior distributions
(\hat{\pi}_t^N)_{t \geq 1} defined in Algorithm 1 is such that, for every sequence \((M_t)_{t \geq 1}\) in \(\mathbb{R}_{>0}\) with \(\lim_{t \to +\infty} M_t = +\infty\), we have

\[
\lim_{t \to +\infty} \hat{\pi}_t^N \left( \|\theta - \theta_*\| \geq M_t \log(t)^{\frac{1+\varepsilon}{2}} \right) = 0, \quad \mathbb{P} - a.s.
\]

In addition, the result still holds if \(C4\) is replaced by \(C4'\).

Remark that this result imposes no conditions on the parameter \(\varepsilon > 0\) and hence it can be taken arbitrary small. We also note that the convergence rate given in Theorem 1 is such that, for arbitrary \(\alpha \in (0, 1/2)\), \(\hat{\pi}_t^N\) converges to \(\theta_*\) at rate \(t^{-\alpha}\) almost surely. Thus, Theorem 1 is consistent with the almost sure convergence result of Proposition 1.

As mentioned in Section 2.2, the requirement on \(N\) to ensure Property 3 can be significantly weakened by requiring that the random mappings \(F\) and \(\bar{F}\) satisfy the additional conditions \(C1'\) and \(C2'\) respectively. Henceforth, we write \(x \neq y\) if and only if \(x, y \in \mathbb{R}^d\) is such that \(x_t \neq y_t\) for all \(t \in 1:d\).

**Theorem 2.** Assume \(A2, A7\) and let \(L_* > 1\) be as in Theorem 1. Then, under \(C1, C4, C1', C2'\) and if \(N^{-1} \sum_{n=1}^{N} \theta_0^n + \theta_*\) and \(N^{-1} \sum_{n=1}^{N} \theta_0^n = \theta_*\), the conclusions of Theorem 1 hold for all \(N \geq (\inf\{k \in \mathbb{N} : k > \kappa^{-1} L_*\})^d\).

Notice that if \(\theta_0^{1:N}\) (resp. \(\theta_0^{1:N}\)) is a realization of a continuous random variable on \(\Theta^N\) (resp. \(\Theta^N\)) then the condition \(N^{-1} \sum_{n=1}^{N} \theta_0^n = \theta_*\) (resp. \(N^{-1} \sum_{n=1}^{N} \theta_0^n = \theta_*\)) of Theorem 2 holds with probability one.

The constant \(L_* \geq 1\) introduced in Theorem 1 is related to the curvature of the KL divergence around \(\theta_*\). Informally speaking, this constant guarantees that \(KL(\theta) < KL(\theta')\) for all \(\theta \in B_{\kappa \xi_{p-1}/L_*}(\theta_*)\) and all \(\theta' \notin B_{\kappa \xi_{p-1}/L_*}(\theta_*)\), implying that, with high probability, \(\hat{\pi}_t^N(B_{\kappa \xi_{p-1}/L_*}(\theta_*))\) is close to one whenever we have \(\hat{\pi}_t^N(B_{\kappa \xi_{p-1}/L_*}(\theta_*)) > 0\) (see Section 2.1.1). Under \(C1, C2\) the condition on \(N\) appearing in Theorem 1 is such that

\[
\hat{\pi}_t^{N}, (B_{\xi_{p-1}/L_*}(\theta_*)) > 0 \implies \hat{\pi}_{t_1}^{L_1+1}(B_{\kappa \xi_{p-1}/L_*}(\theta_*)) > 0
\]

and therefore, as \(\xi_p \geq \kappa \xi_{p-1}\) for every \(p \geq 1\), it follows that, with high probability, \(\hat{\pi}_t^N(B_{\xi_p}(\theta_*)) \approx 1\) whenever \(\hat{\pi}_{t_1}^{L_1+1}(B_{\kappa \xi_{p-1}/L_*}(\theta_*)) > 0\), which is necessary for Property 3 to hold.

The lower bounds for \(N\) given in Theorems 1 and 2 suggest that \(\kappa\) controls the trade-off between statistical and computational efficiency; small \(\kappa\) means fewer perturbations but large computational cost, i.e. large \(N\), while large \(\kappa\) means frequent perturbations with low computational cost. However, this intuitive connection between the frequency of perturbations and the statistical properties of the PPD remains to be theoretically established. Lastly, it should be clear that increasing \(N\) above the thresholds given in Theorems 1 and 2 will not improve the statistical properties of the PPD which are constrained by number of steps \(t_p - t_{p-1}\) between two successive perturbations of the Bayes updates (see the discussion in Section 2.1.1).
3.3 Point-estimators

Let us now analyse the asymptotic behaviour of point-estimators derived from the sequence of PPDs defined in Algorithm 1. In particular, we focus on the asymptotic behaviour of the sequence \( (\hat{\theta}_t^N)_{t \geq 1} \) of perturbed posterior means, defined by

\[
\hat{\theta}_t^N = \int_\Theta \pi_t^N(d\theta), \quad t \in \mathbb{N}.
\]

However, since for all \( p \geq 1 \) and \( t \in (t_p - 1):t_p \) the support of \( \pi_t^N \) is almost surely included in a ball of size \( \xi_p \), the result presented below also holds for other classical Bayesian estimators such as the mode or the median of \( \pi_t^N \).

The following result is a direct consequence of Theorems 1 and 2.

Corollary 1. Assume A1-A7, let \( L^\star > 1 \) be as in Theorem 1 and assume that one of the following two conditions holds:

- \( N \geq (2 \inf\{k \in \mathbb{N} : k > \kappa^{-1} L^\star\})^d \) and C1-C4 hold.
- \( N \geq (\inf\{k \in \mathbb{N} : k > \kappa^{-1} L^\star\})^d \), C1-C4 hold, and the additional conditions of Theorem 2 hold.

Then,

\[
\lim_{t \to +\infty} \log(t)^{1+\varepsilon} t^{1/2} \|\theta^\star - \hat{\theta}_t^N\| < +\infty, \quad \mathbb{P} - a.s.
\]

In addition the result still holds if C4 is replaced by C4*.

3.4 Stability with respect to \( M \) and \( N \)

In Sections 3.2 and 3.3 we provided results on the behaviour of \( \pi_t^N \) as \( t \to +\infty \). However, \( \pi_t^N \) is not only a function of the observations \( (Y_s)_{s=1}^t \) but also of \( M \) and \( N \), and it is therefore of interest to study its behaviour also when \( M \) or \( N \) increases. Indeed, as the constant \( L^\star \) appearing in the statement of Theorem 1 is unknown, it is important to ensure that the PPD has some good statistical properties that hold uniformly for all \( N \) large enough. Following the discussion of Section 2.2.1, increasing \( M \) should allow the sequence of PPDs to find a small neighbourhood of \( \theta^\star \) earlier and hence to accelerate the convergence. Consequently, it is useful to make sure that this intuitive strategy for improving the inference cannot actually worsen the convergence behaviour of the PPD.

The assertion that the convergence behaviour of the PPD does not worsen with \( M \) and \( N \) is made rigorous by the following theorem. Recall from Section 2.2 that for this reason we introduced the more complicated Condition C4 to replace C4*.

Theorem 3. Assume A1-A7, let \( L^\star > 1 \) be as in Theorem 1 and assume that C1-C4 hold. Then, there exists a constant \( C \in \mathbb{R} \) such that

\[
\lim_{p \to +\infty} \inf_{(M,N) \in Q^N} \mathbb{P} \left( \sup_{t \geq t_p} \pi_t^N \left( \{ \|\theta - \theta^\star\| \geq C \log(t)^{1+\varepsilon} t^{-1/2} \} \right) = 0 \right) = 1
\]

17
\[
\lim_{\mu \to +\infty} \inf_{(M,N) \in Q_\kappa} \mathbb{P}\left( \sup_{t \geq t_p} \log(t) - \frac{\mu}{2} t^{1/2} \| \hat{\theta}_t^N - \theta^* \| \leq C \right) = 1 \tag{11}
\]

with \( Q_\kappa = \{(M,N) \in \mathbb{N}^2 : N \geq \left( 2 \inf\{ k \in \mathbb{N} : k > \kappa^{-1}L_* \} \right)^d \} \). If, in addition, the extra conditions of Theorem 2 hold then (10)-(11) hold for
\[
Q_\kappa = \{(M,N) \in \mathbb{N}^2 : N \geq \left( \inf\{ k \in \mathbb{N} : k > \kappa^{-1}L_* \} \right)^d \}.
\]

### 4 Numerical study

In this section we illustrate the performance of PBI for online parameter estimation in Bayesian quantile regression [Yu and Moyeed (2001)] and mixture of logistic regression models. Quantile regression models are challenging due to non-differentiable and potentially multi-modal likelihood functions. With mixture models, parameter inference is a difficult task because the corresponding likelihood function is usually multi-modal (Melnykov et al., 2010). An algorithm was proposed in Kaptein and Ketelaar (2018) for online parameter estimation in mixture of logistic regression models, but the algorithm is not guaranteed to convergence to \( \theta^* \).

All the results presented in this section are based on a single run of Algorithm 1 using a fixed seed of the underlying random number generator. The tests were performed on Intel Core i7-6700 CPU at 3.40GHz \( \times 8 \) and the running time per observation is less than 0.008 second in all the experiments.

#### 4.1 Implementation of Algorithm 1

We consider in this section a generic implementation of Algorithm 1 that can be readily used for any model and is by no means optimised, in any sense, for the specific problem at hand. However, in practice it may be useful to tune its input parameters as well as the support generation and point-estimate mappings for the problem at hand.

When \( N < 2^d \) the computational budget is not sufficient to explore all the directions of the parameter space. By this exploration we mean the partitioning of \( B_{\mu}^p(\hat{\theta}_t) \) (resp. \( B_{\mu}^p(\tilde{\theta}_t) \)) into \( K^d > 1 \) hypercubes of equal volume so that each hypercube contains at least one point of the support of \( \tilde{\pi}_N^{t_p+1} \) (resp. \( \tilde{\eta}_{t_p+1} \)). In this case, the directions of the parameter space to explore must be specified by the support generation mappings \( F \) and \( \tilde{F} \).

Inspired by the mean-field variational approach [Blei et al., 2017] we use in this section support generation mappings \( F \) and \( \tilde{F} \), and associated point-estimate mappings \( G \) and \( \tilde{G} \), that rely on the existence of a good mean-field approximation of \( f_\theta \) to decide which directions of the parameter space to explore. This particular instance of the PBI framework, that we term mean-field perturbed Bayesian inference, is presented in detail in the Supplementary Material B and below we only outline its main features.
4.1.1 Mean-field perturbed Bayesian inference

For a non-empty subset \( u \subset 1:d \) let \( f'_{u, \theta} : \mathcal{Y} \to \mathbb{R} \) be an arbitrary measurable function that depends only on \( \theta_u := (\theta_i, i \in u) \).

The basic idea underlying mean-field PBI is to specify the mappings \( F, \tilde{F}, G \) and \( \tilde{G} \) in such a way that, given a partition \( \{S_r\}_{r=1}^R \) of \( 1:d \) and a model \( \{f'_r, \theta \in \Theta\} \) such that \( f'_\theta = \prod_{r=1}^R f'_{S_r, \theta} \), the inference about \( \theta \), obtained by running Algorithm 1 with \( f_\theta = f'_\theta \) is equivalent (in term of asymptotic properties) to the one obtained by running, for all \( r \in 1:R \), Algorithm 1 with \( f_\theta = f'_r \). The number \( R \) of elements in the partition \( \{S_r\}_{r=1}^R \) is the smallest integer in \( 1:d \) such that, for all \( r \in 1:R \), the exploration of all the directions of the \( |S_r| \)-dimensional space in which \( \theta_{S_r} \) takes values can be explored (in the above sense). Consequently, \( R \) depends on the available computational budget and is such that \( N \geq \sum_{r=1}^R 2^{|S_r|} \). The mappings \( F, \tilde{F}, G \) and \( \tilde{G} \) underpinning mean-field PBI are therefore well-defined for any \( N \geq 2d \). Moreover, they satisfy Conditions C1-C4 of Section 2.2.

An important feature of mean-field PBI is that it only requires to specify the partition \( \{S_r\}_{r=1}^R \) of \( 1:d \) and not to actually approximate \( f_\theta \) in any sense, which enables Property 3 to hold when the mean-field assumption \( \{S_r\}_{r=1}^R \) is correct (i.e. when \( f_\theta \) can be written as \( f_\theta = \prod_{r=1}^R f'_{S_r, \theta} \)) and \( N \geq Kd \) for some \( K \geq 2 \). Moreover, since no approximation of \( f_\theta \) is required, the PPD may still have good convergence properties when there exists a sufficiently good (but not exact) mean-field approximation \( \prod_{r=1}^R f'_{S_r, \theta} \) of \( f_\theta \) (see below for numerical results supporting this point). Lastly, in mean-field PBI we can easily bypass the difficult choice of the partition \( \{S_r\}_{r=1}^R \) by making its construction fully automatic and data driven (in which case the partition of \( 1:d \) evolves over time). This is the approach we follow in this section.

4.1.2 Mappings specification and starting values

All the numerical results are obtained with a version of Algorithm 1 deploying the mean-field approach described above, with \( M = 2 \) (when \( N \geq 2^d \)). The mappings \( F, \tilde{F}, G \) and \( \tilde{G} \) are such that the extra conditions C1* and C2* hold, and the function \( g \) in C2 is as in [5] with \( L = 500 \). For every \( p \geq 1 \) an auxiliary point set \( \tilde{\eta}_{t-1}^{1:N_{aux}} \) is sampled from \( \mu_{N_{aux},t-1} \), a probability measure on \( \Theta_{t-1}^{N_{aux}} \) that is independent of \( (Y_t)_{t>t_{p-1}} \). This point set is used to define the partition of \( 1:d \) required by the mean-field perturbed Bayesian approach, the element \( \tilde{\eta}_{t}^N \) of \( \tilde{\eta}_{t,p+1} \) (as suggested in Section 2.2.1) and the matrix \( \Sigma_{t_p} \) that appears in C2B. More details about the implementation of Algorithm 1 considered in this section are provided in the Supplementary Material B.

The initial values \( \theta_0^N \) and \( \tilde{\theta}_0^N \) in Algorithm 2 are independent such that

\[
\theta_0^N \sim \mathcal{N}_d(\mu_0, I_d), \quad \tilde{\theta}_0^N \sim \mathcal{N}_d(\mu_0, I_d)
\]  

where the parameter \( \mu_0 \) will be specified later depending on the model.
Table 1: Input parameters of Algorithm 1

| parameter | $\epsilon_0$ | $\kappa$ | $\Delta$ | $\varrho$ | $\beta$ | $\varepsilon$ | $\zeta_1$ | $\zeta_2$ | $\zeta_3$ | $\zeta_4$ |
|-----------|---------------|----------|----------|----------|--------|-------------|---------|---------|---------|---------|
| value     | 1             | 0.9      | 0.95     | 2.1      | 0.01   | 0.1         | 1       | 0.5     | 1       | 0.5     |

4.1.3 Parameter values

The results presented in this section are obtained when the input parameters of Algorithm 1 and those appearing in [3] are set to the values listed in Table 1. The choice $\Delta = 0.95$ is made to facilitate the exploration of the parameter space while choosing $\varepsilon = 0.1$ makes the logarithmic term in Property 3 small. We set $\kappa = 0.9$ as large $\kappa$ allows to reduce the minimum value of $N$ for which the convergence results of Section 3 hold. The parameters $\varrho$, $\beta$ and $\epsilon_0$ enter only in the definition of $(\epsilon_p)_{p \geq 1}$ and we propose to choose their values by setting first $\varrho$ and $\beta$ close to their admissible lower bounds and then tuning the sequence $(\epsilon_p)_{p \geq 1}$ through the parameter $\epsilon_0$, whose impact on that sequence is easier to understand. Following this approach we have $(\varrho, \beta) = (2.1, 0.01)$. The values of $(\zeta_1, \ldots, \zeta_4)$ are somewhat arbitrarily set to $(1, 0.5, 1, 0.5)$.

The parameter $\epsilon_0$ plays a crucial role since, for every $p \geq 1$, both $\epsilon_p$ and $\xi_p$ are proportional to this parameter. Large $\epsilon_0$ may improve the exploration of the parameter space and help guiding the PPD quickly towards $\theta_*$ (see Remark 1). This in turn may decrease the number of observations $t^*$ needed to have $\xi_p \neq \epsilon_p$ for all $t_p \geq t^*$, and thus improve the concentration of the PPD around $\theta_*$ for all $t$ large enough. On the other hand, when the initial values in Algorithm 1 are close to $\theta_*$, increasing $\epsilon_0$ will have at best a small impact on $t^*$ and may increase $\xi_p$ for all $t_p$ large enough, reducing the concentration of the PPD around $\theta_*$. We have set a default value $\epsilon_0 = 1$ in Table 1, but the impact of $\epsilon_0$ on the behaviour of the PPD is studied in the first example presented below.

The difference $t_{p+1} - t_p$ between two successive perturbations is approximatively proportional to $t_1$. Consequently, $t_1$ controls the forgetting of the initial state of Algorithm 1 if $t_1$ is small the support of the PPD is updated frequently while the updates are infrequent for large $t_1$ is large. This last input parameter of Algorithm 1 will be set as the same time as the prior parameter $\mu_0$.

4.2 Bayesian quantile regression models

Let $(Z_t)_{t \geq 1}$ and $(X_t)_{t \geq 1}$ be sequences of i.i.d. random variables taking values in $\mathbb{R}$ and $\mathbb{R}^{d_x}$, respectively, and define $Y_t = (Z_t, X_t)$ for all $t \geq 1$. We assume that, for every $q \in (0, 1)$, the $q$-th conditional quantile function of $Z_t$ given $X_t$ belongs to $\{\mu(\theta, \cdot), \theta \in \Theta\}$, with $\Theta = \mathbb{R}^d$. For $q \in (0, 1)$ let $\rho_q(u) = (|u| + (2q - 1)u)/2$, with $u \in \mathbb{R}$,

$$g_{q, \theta}(z|x) = q(q - 1) \exp \left\{ - \rho_q(z - \mu(\theta, x)) \right\}$$

and $\theta_*^{(q)} \in \arg\min_{\theta \in \Theta} \mathbb{E}[\rho_q(Z_1 - \mu(\theta, X_1))]$. Notice that $g_{q, \theta}(|x|$ is the density of an asymmetric Laplace distribution and that, although $g_{q, \theta}(\cdot)$ is not a density function on $\mathcal{Y} := \mathbb{R}^{d_x+1}$, we can still perform (perturbed) Bayesian inference with $f_\theta(y) = g_{q, \theta}(z|x)$.
As assumed by PBI, for every $q \in (0, 1)$ and under appropriate conditions, the posterior distribution $\pi_{q,t}(d\theta) \propto \prod_{s=1}^{t} g_q(\theta_s | X_s) \pi_0(d\theta)$ converges to $\theta^{(q)}_t$ at rate $t^{-1/2}$ (Sriram et al., 2013).

We generate $T$ observations as follows

$$Z_t = \mu(\theta^{(1/2)}_t, X_t) + v_t, \quad v_t \sim N_1(0, 1), \quad t \geq 1$$

where different distributions of $X_t$ will be considered. Below $\mu_0 = \theta^{(1/2)}_t - 10$ and, to minimize further the impact of the initial values we let $t_1 = 5$, so that after only 5 observations the support of the PPD is updated a first time. The results for this test are given in Figures 1c-1d, where the estimation error $\|\hat{\theta}_t - \theta^{(q)}_t\|$ is reported for $t \in (t_p)p \geq 1\cap T$ and various simulation settings.

### 4.2.1 Non-linear model 1

Inspired by Geraci (2017) we let

$$\mu(\theta, x) = \theta_1 + \frac{\theta_1 - \theta_4 + x_1}{1 + e(\theta_2 + x_2 - x_3)/\theta_5}, \quad X_1 \sim N_2\left(0, \left(\begin{array}{c} 4 - 2 \\ -4 \end{array}\right)\right) \otimes U(0, 20)$$

and $\theta^{(1/2)}_t = (70, 10, 3, 10)$. Simulations are done for $T = 10^7$ and $q = 0.5$. The results for this model are given in Figures 1a-1c where $N_{aux} = 1.000$.

In Figure 1c the behaviour of the estimation error is as predicted by Corollary 1 for $N = 2^d$ and $N = 8^d$, although for $N = 2^d$ a much larger sample size is needed to enter in the asymptotic regime. Figure 1b compares the estimation errors obtained for $\epsilon_0 \in \{0.2, 1\}$ when $N = 8^d$. As explained in Section 4.1.3 the time needed by the PPD to be in a small neighbourhood of $\theta^{(1/2)}_t$ is smaller for $\epsilon_0 = 1$ than for $\epsilon_0 = 0.2$. However, in this example, the initial values used in Algorithm 1 are sufficiently close to $\theta^{(1/2)}_t$ so that, for $t$ large enough, the estimation error is the smallest for $\epsilon_0 = 0.2$.

### 4.2.2 Non-linear model 2

Inspired by Hunter and Lange (2000), we let

$$\mu(\theta, x) = \sum_{i=1}^{d} \left( e^{-x;\theta^2} + x_i \theta_{d-i+1} \right), \quad \theta^{(1/2)}_t = 1_{d \times 1}, \quad X_1 \sim \delta_1 \otimes U(0, 1)^{d-1}$$

where $1_{d \times d_2}$ denotes the $d \times d_2$ matrix whose entries are all equal to one. Simulations are done with $T = 3 \times 10^6$, $d \in \{7, 25\}$ and $q = 0.5$. The results for this model are given in Figures 2a-2b where $N_{aux} = 20.000$.

For $d = 7$, the convergence rate of $\hat{\theta}_t^N$ observed in Figure 2b is as predicted by Corollary 1 when $N = 4^d$ and slightly slower than the expected $\log(t)^{(1+\epsilon)/2}t^{-1/2}$ rate when $N = 2^d$. The results for $d = 25$ are given in Figure 2b where $N = 30.000$. Notice that at every perturbation time $t_p + 1$, the mean-field PBI requires the set $1:d$ to be partitioned into two subsets. In Figure 2b we observe that the estimation error decreases roughly as $t^{-1/4}$.  

21
Figure 1: Estimation errors for the quantile regression models and the mixture of logisict regression models. The simulation setup is as described in the text. Plots (a)-(b) are for the non-linear model 1. In (a) the results are for $N = 8^4$ (solid line) and $N = 2^4$. In (b) the results are for $\epsilon_0 = 1$ (solid line) and $\epsilon_0 = 0.2$. Plots (c)-(d) are for the non-linear model 2. In (c) the results are for $N = 2^d$ (solid line) and $N = 4^d$. Plots (e)-(f) are for the linear model. Results are for $q = 0.5$ (solid lines) and $q = 0.05$. Plots (g)-(i) are for the mixture model. In (g) the results are for $N = 4^d$ (solid line) and $N = 2^d$. 
This rate is slower than the one predicted in Corollary \[\text{I}\] but, being much smaller than \(2^{25}\), \(N\) does not satisfy the assumption of Corollary \[\text{I}\]. Lastly, Figure \[\text{III}\] illustrates a potential issue that one may encounter when taking a small value for the parameter \(t_1\), namely that the estimation error \(\|\hat{\theta}_t^N - \theta_*(1/2)\|\) first increases significantly before converging to 0. This initial increase of the estimation error arises because the \(t_1 = 5\) observations used to update the support of the PPDs during the first few iterations of Algorithm \[\text{I}\] do not bring enough information about \(\theta_*\) to guide the PPDs towards this correct value.

4.2.3 Linear model

We let \(d = 20\) and

\[
\mu(\theta, x) = \sum_{i=1}^{d} \theta_i x_i, \quad \theta_*^{(1/2)} \sim \mathcal{U}(1, 5)^d, \quad X_1 \sim \delta_1 \otimes N_{d-1}(0, \Sigma_x).
\]

For this model we take \(N = 35000\) so that, at every perturbation time \(t_p + 1\), the mean-field PBI requires to partition the set \(1:d\) into two subsets.

To illustrate how the performance of mean-field PBI depends on the strength of the mean-field assumption \(\{S_{t_p+1, \tau}\}^2_{\tau=1}\) made at every perturbation time \(t_p + 1\) we perform simulations for \(\Sigma_x \in \{\Sigma_{x,0}, \Sigma_{x,1}\}\) where, for \(\rho \geq 0\), the matrix \(\Sigma_{x,\rho}\) is defined as follows

\[
\tilde{\Sigma}_{x,\rho} = \begin{pmatrix} A^T A & \rho 1_{10 \times 9} \\ \rho 1_{9 \times 10} & \tilde{A}^T \tilde{A} \end{pmatrix}, \quad \Sigma_{x,\rho} = \tilde{\Sigma}_{x,\rho} / \|\tilde{\Sigma}_{x,\rho}\|_{\max}
\]

where the elements of the matrices \(A\) and \(\tilde{A}\) are i.i.d. random draws from the \(\mathcal{U}(0, 1)\) distribution. Under the covariance matrix \(\Sigma_{x,0}\) we see that \(X_{t,2:10}\) and \(X_{t,11:20}\) are independent while under \(\Sigma_{x,1}\) we have

\[
\min_{i \in \{2:10\}, j \in \{11:20\}} |\text{cor}(X_{t,i}, X_{t,j})| \approx 0.11, \quad \max_{i \in \{2:10\}, j \in \{11:20\}} |\text{cor}(X_{t,i}, X_{t,j})| \approx 0.3.
\]

The results for this model and for the quantiles \(q \in \{0.05, 0.5\}\) are given in Figures \[\text{III-III}\] where \(T = 3 \times 10^6\) and \(N_{\text{aux}} = 40000\).

When \(\Sigma_x = \Sigma_{x,0}\) and \(q = 0.5\), the estimation error behaves in Figure \[\text{III}\] as predicted by Corollary \[\text{I}\] despite of the fact that \(N < 2^d\). When \(q = 0.05\) it is not clear whether the convergence rate is slower than for \(q = 0.5\) (e.g. because the asymmetry in the density function \(g_q(\cdot|x)\) brought by taking \(q \neq 0.5\) makes the mean-field assumption stronger) or if more observations are needed to enter in the asymptotic regime (e.g. because the observations are less informative to estimate the 5\% quantile than the median). For \(q \in \{0.05, 0.5\}\) the final partition of \(1:d\) built by the algorithm is \(\{2:10, \{1, 11:20\}\}\), which is consistent with the fact that \(X_{t,2:10}\) and \(X_{t,11:20}\) are independent when \(\Sigma_x = \Sigma_{x,0}\). For \(\Sigma_x = \Sigma_{x,1}\) the behaviour of the PPD is worse than for \(\Sigma_x = \Sigma_{x,0}\), since in this case the estimation error decreases as \(t^{-1/4}\) in Figure \[\text{III}\]. These numerical results suggest that the strength of the mean-field assumption is influenced by the correlation between \(X_{t,2:10}\) and \(X_{t,11:20}\), which is quite intuitive.
4.3 Mixture of logistic regression models

Let \((Z_t)_{t \geq 1}\) and \((X_t)_{t \geq 1}\) be i.i.d sequences of random variables taking values in \(\{0, 1\}\) and \(\mathbb{R}^d_e\), respectively, and define \(Y_t = (Z_t, X_t)\) for all \(t \geq 1\). We assume that, for some \(J \in \mathbb{N}\), the probability density function of \(Z_t\) given \(X_t\) belongs to \(\{g_{J, \theta}(\cdot), \theta \in \Theta\}\), where \(\Theta = \mathbb{R}^{(d_x + 1)J - 1}\) and, for \(\theta \in \Theta\) and with \(\theta_0 = 0\),

\[
g_{J, \theta}(z|x) = \sum_{j=1}^{J} \frac{\exp(\theta_j - 1)}{\sum_{j'=1}^{J} \exp(\theta_{j'} - 1)} \left( z + (1 - z) \exp\left( - \sum_{i=1}^{d_x} x_i \theta_{j - 1 + (j - 1)d_x + i} \right) \right).
\]

Contrary to what is assumed in \cite{A1}, the minimizer of the KL divergence is not unique because the likelihood function is invariant to any relabelling of the mixture components. Therefore, a direct application of Algorithm 1 would result in a sequence of PPDs whose support would jump infinitely many times from one global minimum of the KL divergence to another. In this case, \(\xi_p = \epsilon_p\) infinitely often and thus Property 3 cannot hold. To avoid this issue, Line 9 of Algorithm 1 is replaced by the following Line 9’

9’: if \(\exists j \in 1:J!\) such that \(\|\hat{\theta}_p - \bar{\theta}_p(j)\| \leq 2\epsilon_p\) then

where \(\{\bar{\theta}_p(j)\}_{j=1}^J\) is a set of \(J\) distinct elements such that, for all \((z, x) \in \mathcal{Y}\) and all \(j, j' \in 1:J\), we have \(g_{J, \bar{\theta}_p(j)}(z|x) = g_{J, \bar{\theta}_p(j')}(z|x)\).

For this model we take \(t_1 = 100\) and \(\mu_0 = (0, \ldots, 0)\).

4.3.1 Simulated data

We take \(J = 2\) and we generate the observations as follows

\[
Z_t|X_t \sim g_{J, \theta_s}(z|X_t), \quad X_t \sim \delta_1 \otimes \mathcal{N}(0, I_{d_x - 1})
\]

where different values of \(d_x\) and \(\theta_s\) such that \(\alpha_{*2} := \exp(\theta_{s1})/(1 + \exp(\theta_{s1})) = 0.7\) will be considered. The number of mixture components \(J = 2\) is assumed to be known.

Figure 12 shows the estimation error \(\|\hat{\theta}_t - \theta_s\|\) for \(t \in (t_p)_{p \geq 1} \cap 1:10^6\) when \(d_x = 3\), \(\theta_s\) is such that \(\mathbb{P}(Z_1 = 1) \approx 0.06\) and \(N_{aux} = 100000\). The results are as predicted by Corollary 1 for \(N = 4^d\) but not for \(N = 2^d\). In Figure 13, where \(d_x = 6\), \(\theta_s\) is such that \(\mathbb{P}(Z_1 = 1) \approx 0.22\), \(N = 2^d\) and \(N_{aux} = 20000\), the estimation error decreases as \(t^{-2/7}\) and thus, as per in Figure 12, the computational effort is not enough for the estimation error to decrease as \(\log(t)(1 + t)^2t^{-1/2}\). Lastly, in Figure 14 we consider the case \(d_x = 10\), \(\theta_s\) such that \(\mathbb{P}(Z_1 = 1) \approx 0.22\), \(N = 40000\) and \(N_{aux} = 20000\). In this scenario, where \(N \ll 2^d = 2^{21}\) so that the lower bound on \(N\) assumed in Corollary 1 cannot be satisfied, the estimation error decreases as \(t^{-1/3}\).

4.3.2 Application to E-commerce data

We end this section by applying PBI to analyse the behaviour of e-commerce customers from click streams data. The use of mixture of logistic regression models in this context
Table 2: Estimated parameter values for the real data example

| Mixture weight | Intercept | Number of clicks | % of clicks in category S | % clicks in category B | Max. time spent on an item (in sec.) |
|----------------|-----------|------------------|--------------------------|-----------------------|--------------------------------------|
| 0.76           | -3.57     | -2.5             | 0.31                     | -0.08                 | -0.24                                |
| 0.24           | -2.83     | 0.38             | -0.31                    | 0.04                  | 0                                    |

has been considered e.g. in Kaptein and Ketelaar (2018). The data used in this experiment are those of the RecSys Challenge 2015 (available at http://recsys.yoochoose.net/challenge.html) where the clicked items are divided into three categories: special offers (S), brand (B) and regular items. After having removed missing observations the dataset contains records for 8,918,543 customers collected during six consecutive months and where about 5% of them bought an item.

To keep the problem simple we assume the existence of $J = 2$ groups and consider four features to explain the probability that an e-shopper buys an item (see Table 2). About 75% of the observations are randomly selected to estimate the model parameters while the remaining observations are used as a test set to assess the fitted model. Before applying Algorithm 1 we randomly permute the observations to make them ‘more similar’ to a sample from a stationary process.

The estimated parameter values, obtained for $N = N_{aux} = 40,000$, are given in Table 2. Based on the Nielsen Norman Group classification of e-shoppers (www.nngroup.com/articles/ecommerce-shoppers), the results can be interpreted as showing that about 76% of the customers are browsers or bargain hunters while about 24% are product focused or researchers. Indeed, for the former group the probability of buying an item increases with the percentage of clicks in the category special offers and decreases with the total number of clicks (a large number of clicks can be possibly interpreted as a sign that a leisurely shopper do not find anything interesting to buy). Moreover, it seems reasonable to assume that the purchasing decision of browsers/bargain hunters is mostly impulsive, which is supported by the negative relationship between the duration of the longest stay on a given page and the probability of purchasing an item. For the second group of customers we observe a larger estimated value for the intercept and lower estimated values for the other parameters. These estimates are coherent with the expected behaviour of product focused/researchers since, to a large extent, their purchasing decisions is made before visiting the site.

Figure 2a aims at assessing the convergence of $\hat{\theta}_t^N$ by plotting $\|\hat{\theta}_{t+1}^N - \hat{\theta}_t^N\|$ as a function of $p$. The results in this figure show that this quantity decreases as $\log(t_p)^{(1+\epsilon)/2}t_p^{-1/2}$ for $p$ large enough which, given the definition of the estimator $\hat{\theta}_t^N$, shows that $\hat{\theta}_t^N$ concentrates on some parameter value as $t$ increases. Note that this convergence rate for $\|\hat{\theta}_{tp+1}^N - \hat{\theta}_tp^N\|$ is consistent with Property 3. In Figure 2b we use the mixture model and the parameter values of Table 2 to estimate, for every e-shoppers belonging to the test set, the probability of buying an item. The results in this plot show that the estimated probabilities tend to be larger for customers that did buy an item than for those who did not, and thus that the fitted model provides meaningful predictions (at least at a
5 Conclusion

In this paper we introduced perturbed Bayesian inference as a practical alternative to Bayesian inference for online parameter estimation. Our main result establishes that a finite computational budget is enough to enable the almost sure convergence of $\tilde{\pi}_t^{N}$ to $\theta_*$ at rate $t^{-\alpha}$, for every $\alpha \in (0, 1/2)$. More precisely, we show that there exists a model dependent constant $N_* \in \mathbb{N}$ such that this latter convergence property holds for all $N \geq N_*$. When $N < N_*$, numerical results suggest that the PPD may still converge to $\theta_*$ but typically this will happen at a slower rate. The link between the computational budget, the complexity of the model (e.g. as measured by $N_*$) and the convergence rate of the PPD is an interesting open problem.

The lower bound $N_*$ on $N$ grows exponentially with the dimension $d$ of the parameter space, the reason being that PBI only requires the ability to evaluate the likelihood function point-wise to be applicable. On the other hand, estimators based on stochastic approximations have proved that gradient information allows to build computationally cheap and efficient methods for online parameter estimation for models with convex KL divergence (Toulis and Airoldi, 2015). Based on these promising results, future research should investigate the possibility of using this information within the perturbed Bayesian framework to reduce the computational cost needed for the PPDs to enjoy the aforementioned convergence behaviour.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013). Stream-
ing variational bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–551.

Crisan, D., Miguez, J., et al. (2018). Nested particle filters for online parameter estimation in discrete-time state-space markov models. *Bernoulli*, 24(4A):3039–3086.

Geraci, M. (2017). Nonlinear quantile mixed models. *arXiv preprint arXiv:1712.09981*.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *Ann. Stat.*, 35(1):192–225.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2018). Bayesian conditional density filtering. *Journal of Computational and Graphical Statistics*, (just-accepted).

Henzinger, M. R., Raghavan, P., and Rajagopalan, S. (1998). Computing on data streams. *External memory algorithms*, 50:107–118.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77.

Ingber, L. and Rosen, B. (1992). Genetic algorithms and very fast simulated reannealing: A comparison. *Mathematical and computer modelling*, 16:87–87.

Kaptein, M. and Ketelaar, P. (2018). Maximum likelihood estimation of a finite mixture of logistic regression models in a continuous data stream. *arXiv preprint arXiv:1802.10529*.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Melnykov, V., Maitra, R., et al. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116.

O’callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2002). Streaming-data algorithms for high-quality clustering. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 685–694. IEEE.

Sriram, K., Ramamoorthi, R., Ghosh, P., et al. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, 8(2):479–504.

Toulis, P. and Airoldi, E. M. (2015). Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Statistics and computing*, 25(4):781–795.

Toulis, P., Airoldi, E. M., et al. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727.
van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zhang, R., Zheng, F., and Min, W. (2018). Sequential behavioral data processing using deep learning and the markov transition field in online fraud detection. *arXiv preprint arXiv:1808.05329*.

Zhou, J. and Said, S. (2018). Recursive parameter estimation in a riemannian manifold. *arXiv preprint arXiv:1805.06811*. 
Supplementary Material A: Proofs

Mathieu Gerber* Kari Heine†

Contents

S1 Additional notation S1
S2 Stabilizing effect of Condition C4: Informal explanation S2
S3 Preliminary results S3
S3.1 A technical lemma .......................................................... S3
S3.2 Convergence rate of Bayes updates ..................................... S4
S3.3 An MLE type result ............................................................ S5
S3.4 Preliminary results for the auxiliary sequence of PPDs .......... S5
S3.5 A preliminary result for the sequence of PPDs ...................... S14
S3.6 Interactions between the PPDs and the auxiliary PPDs .......... S15
S4 Proof of Theorem 1 S18
S5 Proof of Theorem 2 S20
S6 Proof of Corollary 1 S21
S7 Proof of Theorem 3 S21
S8 Proof of Sections S3.2 and S3.3 S22
S8.1 Two direct implications of Assumptions A1, A2 and A5 .......... S22
S8.2 Proof of the result of Section S3.2 ...................................... S24
S8.3 Proof of the result of Section S3.3 ...................................... S29

S1 Additional notation

Let $\sigma(X)$ be $\sigma$-algebra generated by the random variable $X$ and $N_0 = \{0\} \cup \mathbb{N}$.

*School of Mathematics, University of Bristol, UK.
†Department of Mathematical Sciences, University of Bath, UK.
S2 Stabilizing effect of Condition C4: Informal explanation

To understand the stabilizing effect of C4 we explain in the following why, under that condition, the inequality (7) holds for some \( \delta_p \to 1 \) and uniformly in \( M \geq 1 \) and \( N \) sufficiently large.

Let \( \bar{\eta}_t^{(2)} \) be the mode of \( \tilde{\eta}_t \) and, assuming \( M > 1 \), define a random probability measure

\[
\mu_{p-1} = \sum_{n=1}^{N} \frac{a_n \delta_p^{\eta_{p-1}}}{\sum_{n'=1}^{N} a_{n'}} \tag{S.1}
\]
on \( \Theta \) so that, under C4,

\[
\bar{\eta}_t = \bar{\eta}_t^{(2)} \iff \Psi_{t_{p-1}:t_p}(\mu_{p-1})(B_{(1+\kappa)\epsilon_{p-1}}(\bar{\eta}_{t_{p-1}})) \leq \Delta. \tag{S.2}
\]

Remark now that under C2.2, and for \( N \) large enough, \( \mu_{p-1}(B_{\kappa\epsilon_{p-1}}(\theta_*)) > 0 \) whenever \( \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*) \) and therefore, using the concentration property (2) of the Bayes updates, we have, as \( p \to +\infty \),

\[
\tau_{M,N}^{(p)} : = \mathbb{P}(\Psi_{t_{p-1}:t_p}(\mu_{p-1})(B_{(1+\kappa)\epsilon_{p-1}}(\bar{\eta}_{t_{p-1}})) > \Delta \mid \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)) \to 1.
\]

Moreover, whenever \( \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*) \) we also have

\[
\Psi_{t_{p-1}:t_p}(\mu_{p-1})(B_{(1+\kappa)\epsilon_{p-1}}(\bar{\eta}_{t_{p-1}})) \geq \Psi_{t_{p-1}:t_p}(\mu_{p-1})(B_{\kappa\epsilon_{p-1}}(\theta_*))
\]

and consequently it follows that the conditional probability that (S.2) holds is upper bounded by

\[
\mathbb{P}(\bar{\eta}_t = \bar{\eta}_t^{(2)} \mid \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)) \leq 1 - \tau_{M,N}^{(p)}.
\]

To control the probability \( \tau_{M,N}^{(p)} \) we can leverage the rich literature on the concentration of posterior distributions (see the introductory section for references). In particular, under suitable conditions, it can be shown (Lemma S2 below) that if there exists a constant \( c > 0 \) such that, for all \( M \geq 1 \) and \( N \geq 1 \),

\[
\mathbb{P}(\mu_{p-1}(B_{\kappa\epsilon_{p-1}}(\theta_*)) \geq c \mid \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)) = 1, \quad \forall p \geq 1 \tag{S.3}
\]

then there exists a sequence \( (\delta_p^{(2)})_{p \geq 1} \) such that \( \delta_p^{(2)} \to 1 \) and such that, for all \( M \geq 1 \) and \( N \) sufficiently large, \( \tau_{M,N}^{(p)} \geq \delta_p^{(2)} \) for all \( p \geq 1 \). Note that under (S.3), whenever \( \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*) \), the prior probability \( \mu_{p-1}(B_{\kappa\epsilon_{p-1}}(\theta_*)) \) does not decrease to zero as \( M \) or \( N \) increases. The definition of \( \mu_{p-1} \) in (S.1) is such that, under C2.1-C2.2 and for \( N \) large enough, (S.3) holds and thus, for all \( M \geq 1 \) and \( N \) sufficiently large,

\[
\mathbb{P}(\bar{\eta}_t = \bar{\eta}_t^{(2)} \mid \bar{\eta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)) \leq 1 - \delta_p^{(2)}, \quad \forall p \geq 1. \tag{S.4}
\]
Moreover, under C4, whenever (S.2) does not hold, \( \hat{\theta}_{tp} \) is equal to \( \bar{\theta}_{tp}^{(1)} \), the expectation of \( \theta \) under \( \Psi_{tp-1;tp}(\hat{\mu}_{p-1}) \), where

\[
\hat{\mu}_{p-1} = \sum_{n \in 1:N \cup J_{p-1}} \frac{\hat{a}_n(J_{p-1})\delta_{\bar{\theta}_{tp-1}^n}}{\sum_{n' \in 1:N \cup J_{p-1}} \hat{a}_{n'}(J_{p-1})}
\]

with \( J_{p-1} = \{ n \in (N+1):(N+M), \bar{\theta}_{tp-1}^n \in B_{(1+2\kappa)}\epsilon_{p-1}(\bar{\theta}_{tp-1}) \} \). Remark that because \( \hat{\mu}_{p-1} \) is a probability measure on \( B_{(1+2\kappa)}\epsilon_{p-1}(\bar{\theta}_{tp-1}) \) we have

\[
\|\bar{\theta}_{tp}^{(1)} - \theta_*\| \leq \kappa\epsilon_{p-1} + 2(1 + \kappa)\epsilon_{p-1}\Psi_{tp-1;tp}(\hat{\mu}_{p-1})(B_{c\kappa\epsilon_{p-1}}(\bar{\theta}_{tp-1}))
\]

whenever \( \bar{\theta}_{tp-1} \in B_{\epsilon_p-1}(\theta_*) \). Then, on the one hand,

\[
\mathbb{P}(\bar{\theta}_{tp}^{(1)} \notin B_{\epsilon_p}(\theta_*)) \mathbb{P}(\bar{\theta}_{tp-1} \in B_{\epsilon_p-1}(\theta_*)) \leq \mathbb{P}(\Psi_{tp-1;tp}(\bar{\theta}_{tp}^{(1)})(B_{c\kappa\epsilon_{p-1}}(\bar{\theta}_{tp-1})) = \mathbb{P}(\Psi_{tp-1;tp}(\hat{\mu}_{p-1})(B_{c\kappa\epsilon_{p-1}}(\bar{\theta}_{tp-1})) \geq \frac{c}{2(1+\kappa)} \mathbb{P}(\bar{\theta}_{tp-1} \in B_{\epsilon_p-1}(\theta_*))
\]

where, using the definition of \( (\epsilon_p)_{p \geq 1} \), the second inequality holds for \( p \) large enough and for some \( c \in (0,1) \). On the other hand, the coefficients \( \hat{a}_n(J_{p-1}) \) appearing in the definition of \( \hat{\mu}_{p-1} \) play the same role as the coefficients \( a_n \) used to define \( \mu_{p-1} \) and thus, as explained above and under C2.1-C2.2, enable us to show (see the proof of Lemma S4 below) that for some sequence \( (\delta_{p}^{(1)})_{p \geq 1} \) such that \( \delta_{p}^{(1)} \to 1 \) we have, for all \( M \geq 1 \) and \( N \) large enough,

\[
\mathbb{P}(\bar{\theta}_{tp}^{(1)} \in B_{\epsilon_p}(\theta_*)) \geq \delta_{p}^{(1)}, \quad \forall p \geq 1.
\]

Together with (S.4), this shows that (7) holds with \( \delta_p = \delta_p^{(1)} + \delta_p^{(2)} - 1 \) (see Lemma S4 below).

**Remark S1.** In the definition of \( J \) in C4, the radius of the ball around \( \mu \) comes from the analysis of \( \mathbb{P}(\bar{\theta}_{tp} \in B_{\epsilon_p}(\theta_*)) \mathbb{P}(\bar{\theta}_{tp-1} \notin B_{\epsilon_p-1}(\theta_*)) \), given in Lemma S5 below, which in addition to the probability in (7) is the second key quantity that we need to control to establish the convergence results of Section 3.

### S3 Preliminary results

#### S3.1 A technical lemma

The following result seems to be common knowledge but its original proof is difficult track back.

**Lemma S1.** Let \( A \in \mathcal{F} \) be such that \( \mathbb{P}(A) > 0 \). Then,

\[
\sup_{B \in \mathcal{F}} |\mathbb{P}(B|A) - \mathbb{P}(B)| = 1 - \mathbb{P}(A).
\]

S3
Proof. Let $B \in \mathcal{F}$ and note that $P(B \mid A) - P(B) = b(1 - a)/a - c$ where

$$a = P(A), \quad b = P(B \cap A), \quad c = P(B \setminus A).$$

Then, because $0 \leq b \leq a$ and $0 \leq c \leq 1 - a$, we have

$$-(1 - a) \leq -c \leq P(B \mid A) - P(B) \leq b(1 - a)/a \leq 1 - a$$

and thus

$$\sup_{B \in \mathcal{F}} |P(B \mid A) - P(B)| = 1 - a = 1 - P(A).$$

\[\square\]

S3.2 Convergence rate of Bayes updates

**Lemma S2.** Assume A1-A5, let $(\tilde{c}_j)_{j \geq 0}$ be a sequence in $(0,1]$ and, for every $(p,j) \in \mathbb{N}^2$, let $\tilde{c}_{p,j} = \tilde{c}_j \epsilon_p$ and $p_{p,j+1} = t_{p,j+1} - t_{p,j}$. Let

$$K_\kappa = \inf \{ k \in \mathbb{N} : k > \kappa^{-1} L_* \}, \quad L_* = D^{-1/2}(E[m_{\theta_*}^2] \vee d\|V_{\theta_*}\|)^{1/2} \quad (S.6)$$

with $m_{\theta}$ as in A3, $V_{\theta_*}$ as in Lemma S11 and $D \in \mathbb{R}_{>0}$ as in Lemma S13. Let $\lambda > 0$ and $\bar{\alpha} > 0$ be such that

$$K_\kappa \geq \kappa^{-1} L_*(1 + \bar{\alpha})^{1/2}(1 + \lambda)^{1/2}. \quad (S.7)$$

Then, for every $\alpha \in (0, \bar{\alpha})$ and $\zeta \in (0,1)$, there exists a constant $\bar{c} \in \mathbb{R}_{>0}$ such that, for every $(p,j) \in \mathbb{N}^2$ and $\eta \in \mathcal{P}(\Theta)$ with $\eta(B_{\epsilon_{p,j-1}/K_*}(\theta_*)) \geq \zeta$, we have, $\mathbb{P}$-a.s.,

$$E[\Psi_{t_{p,j+1}:t_{p,j}}(\eta)(B_{\kappa \epsilon_{p,j-1}/K_*}(\theta_*)) \mid \sigma(Y_{1:t_{p,j-1}})] \leq \bar{c}(\tau_{p,j}^{-1/2} + \tau_{p,j}^{-1} \tilde{c}_{p,j-1}^{-2})$$

where $\kappa_\alpha = \kappa \sqrt{(1 + \alpha)/(1 + \bar{\alpha})}$. See Section S8.2 for a proof.

**Corollary S1.** Let $L_*$ be as defined in (S.6). Then, $L_* \geq 1$.

**Proof.** Let $K_\kappa$ be as defined in (S.6) and assume that $L_* \in (0,1)$. Then, there exist $\lambda > 0$ and $\bar{\alpha} > 0$ such that (S.7) holds and such that, for some $\alpha \in (0, \bar{\alpha})$, we have

$$K_\kappa^{-1} > \kappa_\alpha = \kappa \sqrt{\frac{1 + \alpha}{1 + \bar{\alpha}}} \geq \frac{\kappa}{\sqrt{(1 + \alpha)/(1 + \lambda)}}.$$ 

In this case it is easily checked that there exists a probability measure $\eta' \in \mathcal{P}(\Theta)$ such that we have both $\eta(B_{\epsilon_{p,j-1}/K_*}(\theta_*)) \geq \zeta$ and $\eta'(B_{\kappa \epsilon_{p,j-1}/K_*}(\theta_*)) = 1$, implying that $\Psi_{t_{p,j+1}:t_{p,j}}(\eta')(B_{\kappa \epsilon_{p,j-1}/K_*}(\theta_*)) = 1$, $\mathbb{P}$-almost surely. This contradicts the conclusion of Lemma S2 and the result follows. \[\square\]
S3.3 An MLE type result

Lemma S3. Assume A1-A2, A5-A7 and C2, and let \((\bar{\epsilon}_p)_{p \geq 0}\) be a sequence in \(\mathbb{R}_{>0}\) verifying \(\lim_{p \to +\infty} (t_p - t_{p-1})^{1/2} \bar{\epsilon}_p = +\infty\) and \(\lim_{p \to +\infty} \bar{\epsilon}_p = 0\). Then, there exist

- constants \(p \in \mathbb{N}, \zeta \in \mathbb{R}_{>0}\) and \(c_* \in (0,1)\)
- \(a\) sequence \((\Omega_{p,\bar{\epsilon}_p})_{p \geq 1}\) in \(\mathcal{F}\) such that \(\lim_{p \to +\infty} \mathbb{P}(\Omega_{p,\bar{\epsilon}_p}) = 1\) and such that for all \(p \geq 1\) the random set \(\Omega_{p,\bar{\epsilon}_p}\) is \(\sigma(Y_{(t_p-1):t_p})\) measurable

such that we have for all \(M \geq 1, N \geq 1, p \geq p\) and \(\mathbb{P}\)-a.s.,

\[
\mathbb{P}(\delta^{(2)}_p \in B_{t_p}(\theta_*) | \Omega_{p,\bar{\epsilon}_p}, \sigma(Y_{1:t_p}, \bar{\theta}_{t_0:t_{p-2}}, \theta_0^{1:N})) \\
\geq \mathbb{P}(\exists \kappa > 1: M \text{ s.t. } \delta_{t_{p-1}}^{N+1} \in B_{c_* \bar{\epsilon}_p}(\theta_*) | \Omega_{p,\bar{\epsilon}_p}, \sigma(Y_{1:t_p}, \bar{\theta}_{t_0:t_{p-2}}, \theta_0^{1:N})) \\
\geq \mathbb{P}(\delta_{t_{p-1}}^{N+1} \in B_{c_* \bar{\epsilon}_p}(\theta_*) | \Omega_{p,\bar{\epsilon}_p}, \sigma(Y_{1:t_p}, \bar{\theta}_{t_0:t_{p-2}}, \theta_0^{1:N})) \\
\geq c_* \kappa \gamma \tau_p \dagger_{p-1}^{d_p}
\]

where \(\delta_{t_p}^{(2)} = \delta_{t_{p-1}}^{(2)}\) with \(n_p \in \arg\max_{1:N} \tilde{w}^n_{t_p}\).

See Section S8.3 for a proof.

S3.4 Preliminary results for the auxiliary sequence of PPDs

S3.4.1 Under Conditions C2 and C4

Lemma S4. Assume A1-A5, C2 and C4. Then, there exists a constant \(c \in \mathbb{R}_{>0}\) such that

\[
\sup_{(M,N) \in \mathcal{Q}_n} \mathbb{P}(\delta_{t_p} \notin B_{t_p}(\theta_*) | \delta_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)) \leq c (t_p - t_{p-1})^{-1/2}, \quad \forall p \in \mathbb{N}
\]

where \(\mathcal{Q}_n := \{ (M, N) \in \mathbb{N}^2 : N \geq (2K_n)^d \} \) with \(K_n\) as defined in (S.6).

Proof. Let \((M, N) \in \mathcal{Q}_n\). Below \(p \in \mathbb{N}\) is a constant (independent of \(M\) and \(N\)) whose value can change from one expression to another.

For every \(p \in \mathbb{N}\) let \(\tau_p = t_p - t_{p-1}\). Let \(\lambda > 0\) and \(\bar{\alpha} > 0\) be such that (S.7) holds, \(\kappa = \kappa \sqrt{1 + \bar{\alpha}/2}/(1 + \bar{\alpha})\) and \(\nu : [0,1] \to \mathbb{R}_{>0}\) be such that

\[
\mathbb{E}(|\Psi_{t_{p-1}:t_p}(\eta)| B_{c \tau_p^{1/2}}(\theta_*)) \leq \nu(c) \tau_p^{-1/2}
\]

for every \(p \geq 1\) and \(\eta \in \mathcal{P}(\theta)\) verifying \(\eta(B_{t_p^{-1/2}K_n}(\theta_*)) \geq c\). Note that under A1-A5 such a mapping \(\nu : [0,1] \to \mathbb{R}_{>0}\) exists by Lemma S2 (with \(\bar{c}_1 = 1\) and \(j = 1\)) and because \(\tau_p^{-1/2} \leq \bar{c} \tau_{p-1}^{-1}\) for all \(p \geq 1\) and some constant \(\bar{c} < +\infty\).

For every \(p \in \mathbb{N}\) let \(\bar{\delta}_{t_p}^{(2)}\) be as in Lemma S3 and

\[
J_{p-1} = \{ n \in (N+1):\tilde{N} \text{ s.t. } \bar{\theta}_{t_{p-1}}^{n} \in B_{(1+2\bar{\alpha})\tau_{p-1}^{-1}}(\delta_{t_{p-1}}) \}, \quad \tilde{J}_{p-1} = 1:N \cup J_{p-1}
\]

S5.
\[ \bar{v}_{tp}^{(1)} = \sum_{n \in J_{p-1}} \tilde{W}_{tp}^n v_{tp-1}^n, \quad Z_p = \sum_{n=1}^{\tilde{N}} a_n \hat{W}_{tp}^n \mathbb{1}(\bar{v}_{tp-1}^n \in B_{(1+\epsilon)p-1}(\bar{v}_{tp-1})) \] (S.10)

and

\[ \tilde{W}_{tp}^n = \sum_{m \in J_{p-1}} \hat{w}_{tp}^m \hat{a}_{n}(J_{p-1}) \], \quad n \in \tilde{J}_{p-1} \]

where \( a_1: \tilde{N} \) and \( \hat{a}_1: \tilde{N}(J) \) are as in C4, for \( J \subset (N + 1): \tilde{N} \).

Then, under C4 and for every \( p \geq 1 \),

\[ \mathbb{P}(\bar{v}_{tp} \not\in B_{tp}(\theta_*)) | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*) \]

\[ = \mathbb{P}(\hat{v}_{tp}^{(1)} \not\in B_{tp}(\theta_*), Z_p > \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \]

\[ + \mathbb{P}(\hat{v}_{tp}^{(2)} \not\in B_{tp}(\theta_*), Z_p \leq \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \]

\[ \leq \mathbb{P}(\bar{v}_{tp} \not\in B_{\epsilon p}(\theta_*)) | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \]

\[ + \mathbb{P}(Z_p \leq \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \].

In the remainder of the proof we find upper bounds for the two terms appearing on the r.h.s. of (S.11).

For every \( p \geq 1 \) let

\[ \mu_{p-1} = \sum_{n=1}^{\tilde{N}} a_n \frac{\alpha_{n}}{(1 + \xi_1)M - (1 - \xi_2)\delta_{\epsilon p-1}.} \] (S.12)

Notice that \( \mu_{p-1} \) is a random probability measure on \( \Theta \).

Then, for the second term on the r.h.s. of (S.11) we have, for all \( p \geq 1 \),

\[ \mathbb{P}(Z_p \leq \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \]

\[ \leq \mathbb{P}(\bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*), \mu_{p-1}(B_{\epsilon p-1/\kappa}(\theta_*)) \geq 1 - \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)). \]

Since \((M, N) \in \mathcal{Q}_n\) we have \( N \geq (2K_n)^d \) and thus, under C2.2 and for some constant \( \tilde{\epsilon} > 0 \),

\[ \mathbb{P}(\bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) = \mathbb{P}(\bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*), \mu_{p-1}(B_{\epsilon p-1/\kappa}(\theta_*)) \geq \frac{\tilde{\epsilon} \xi_1}{1 + \xi_1} K_n^{-d}). \]

Consequently, under A2, C2.1, C2.2 and C4, and using (S.8) and Markov’s inequality, we have for all \( p \geq 1 \),

\[ \mathbb{P}(Z_p \leq \Delta | \bar{v}_{tp-1} \in B_{\epsilon p-1}(\theta_*)) \leq \frac{\tilde{\epsilon} \xi_1 K_n^{-d}/(1 + \xi_1)}{1 - \Delta} \tau_{p-1/2}. \] (S.13)

To find an upper bound the second term in (S.11) we define for every \( p \geq 1 \),

\[ \hat{\mu}_{p-1} = \sum_{n \in J_{p-1}} \hat{W}_{tp}^n \delta_{\epsilon p-1}. \] (S.14)
Note that $\hat{\mu}_{p-1}$ is a random probability measure on $\Theta$.
Then, for all $p \geq 1$ we have
\[
\mathbb{P}\left(\bar{\vartheta}_{1}^{(1)} \notin B_{p}(\theta_{*}) \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right)
\]
\[
= \mathbb{P}\left(\left\| \sum_{n \in J_{p}} W_{n}^{t}(\vartheta_{p-1}^{n} - \theta_{*}) \mathbb{I}(\vartheta_{p-1}^{n} \in B_{p}(\theta_{*})) + \sum_{n \in J_{p}} W_{n}^{t}(\vartheta_{p-1}^{n} - \theta_{*}) \mathbb{I}(\vartheta_{p-1}^{n} \notin B_{p}(\theta_{*})) \right\| \geq \epsilon_{p} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right) \quad (S.15)
\]
\[
\leq \mathbb{P}\left(\kappa\epsilon_{p-1} + 2(1 + \kappa)\epsilon_{p-1} \mathbb{P}_{t}^{(1)}(B_{p}^{c}(\theta_{*})) \geq \epsilon_{p} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right)
\]
\[
= \mathbb{P}\left(\Psi_{t}^{(1)}(\hat{\mu}_{p-1})(B_{p}^{c}(\theta_{*})) \geq \frac{\epsilon_{p} - \kappa\epsilon_{p-1}}{2(1 + \kappa)\epsilon_{p-1}} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right).
\]
To proceed further remark that $(\epsilon_{p})_{p \geq 1}$ is such that $\epsilon_{p} \geq \kappa\epsilon_{p-1}$ for all $p \geq p_{0}$ and therefore, for very $p \geq p_{0}$ we have
\[
\mathbb{P}\left(\bar{\vartheta}_{t}^{(1)} \notin B_{p}(\theta_{*}) \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right)
\]
\[
\leq \mathbb{P}\left(\Psi_{t}^{(1)}(\hat{\mu}_{p-1})(B_{p}^{c}(\theta_{*})) \geq \frac{\kappa\epsilon_{p-1} - \kappa\epsilon_{p-1}}{2(1 + \kappa)\epsilon_{p-1}} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right) \quad (S.16)
\]
\[
= \mathbb{P}\left(\Psi_{t}^{(1)}(\hat{\mu}_{p-1})(B_{p}^{c}(\theta_{*})) \geq c_{\kappa} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right)
\]
with $c_{\kappa} = \frac{\kappa - \kappa}{2(1 + \kappa)} > 0$.
Since $(M, N) \in \mathcal{Q}_{K}$ we have $N \geq (2K_{\kappa})^{d}$ and thus, under C2.2,
\[
\mathbb{P}(\bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})) = \mathbb{P}(\bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*}), \hat{\mu}_{p-1}(B_{p-1}/K_{\kappa}(\theta_{*})) \geq \frac{\zeta_{3}}{1 + \zeta_{3}K_{\kappa}^{d}})
\]
with $\zeta_{3} > 0$ as per above. Therefore, under A2, C2.1, C2.2 and C4, and using (S.8) and Markov's inequality, this shows that, for any $p \geq p_{0}$,
\[
\mathbb{P}(\bar{\vartheta}_{t}^{(1)} \notin B_{p}(\theta_{*}) \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*}))
\]
\[
\leq \mathbb{P}\left(\Psi_{t}^{(1)}(\hat{\mu}_{p-1})(B_{p}^{c}(\theta_{*})) \geq c_{\kappa} \mid \bar{\vartheta}_{p-1} \in B_{p-1}(\theta_{*})\right)
\]
\[
\leq \frac{\epsilon_{p} + \zeta_{3}K_{\kappa}^{d} / (1 + \zeta_{3})}{c_{\kappa}} t_{p}^{-1/2}.
\]
Together with (S.11) and (S.13), this completes the proof. \qed

**Lemma S5.** Assume A1-A7, C2 and C4. Then, there exist constants $c \in \mathbb{R}_{>0}$ and $p \in \mathbb{N}$ such that
\[
\sup_{(N, M) \in \mathcal{Q}_{\kappa}} \mathbb{P}(\bar{\vartheta}_{t}^{(1)} \notin B_{p}(\theta_{*}) \mid \bar{\vartheta}_{p-1} \notin B_{p-1}(\theta_{*})) \leq 1 - c\gamma_{p-1}^{p} \epsilon_{p}^{d}, \quad \forall p \geq p_{0}
\]
with $\mathcal{Q}_{\kappa}$ as defined in Lemma S4.
Proof. Let \((M, N) \in Q_κ\). Below, \(c \in \mathbb{R}_{>0}\) and \(p \in \mathbb{N}\) are two constants (independent of \(M\) and \(N\)) whose values can change from one expression to another.

Let \(c_\ast > 0\) be as in the statement of Lemma S3 and, for \(p \geq 1\), let \(\varepsilon_p = c_\ast \varepsilon_{p-1}\) and let \(Ω_{p, \varepsilon_p}\) be as in the statement of Lemma S3. Without loss of generality, we assume below that \(c_\ast \leq \min(\kappa/2, K_{\kappa}^{-1})\) (with \(K_{\kappa}/2\) as defined in (S.6)).

Then, under A2 and C2.1, we have for every \(p \geq 1\)

\[
\mathbb{P}(\tilde{d}_p \notin B_{\varepsilon_p}(\theta_\ast)| \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast))
= 1 - \mathbb{P}(\tilde{d}_p \in B_{\varepsilon_p}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast))
\leq 1 - \mathbb{P}(\tilde{d}_p \in B_{\varepsilon_p}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p}) \mathbb{P}(\Omega_{p, \varepsilon_p})
\]

(S.17)

where, under C4,

\[
\mathbb{P}(\tilde{d}_p \in B_{\varepsilon_p}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
= \mathbb{P}(\tilde{d}_{p}^{(1)} \in B_{\varepsilon_p}(\theta_\ast), Z_p > \Delta | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
+ \mathbb{P}(\tilde{d}_{p}^{(2)} \in B_{\varepsilon_p}(\theta_\ast), Z_p \leq \Delta | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
\]

(S.18)

with \(Z_p\) and \(\tilde{d}_{p}^{(1)}\) defined in (S.10) and \(\tilde{d}_{p}^{(2)}\) defined in the statement of Lemma S3. We now find lower bounds for the two terms on the r.h.s. of (S.18).

For every \(p \geq 1\) let \(Ω′_{p-1} = \{\tilde{d}_{p-1}^{N+1} \in B_{c_\ast \varepsilon_{p-1}}(\theta_\ast), \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast)\}\) and

\[
b_p = \mathbb{P}(Z_p \leq \Delta | Ω′_{p-1}, \Omega_{p, \varepsilon_p}).
\]

Then, for the second probability appearing on r.h.s. of (S.18) we have, for all \(p \geq 1\),

\[
\mathbb{P}(\tilde{d}_{p}^{(2)} \in B_{\varepsilon_p}(\theta_\ast), Z_p \leq \Delta | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
\geq b_p \mathbb{P}(\tilde{d}_{p}^{(2)} \in B_{\varepsilon_p}(\theta_\ast), Z_p \leq \Delta, Ω′_{p-1}, \Omega_{p, \varepsilon_p})
\times \mathbb{P}(\tilde{d}_{p-1}^{N+1} \in B_{c_\ast \varepsilon_{p-1}}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
\]

(S.19)

where, by Lemma S3, for all \(p \geq p\) we have both

\[
\mathbb{P}(\tilde{d}_{p-1}^{N+1} \in B_{c_\ast \varepsilon_{p-1}}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p}) \geq c_\ast \varepsilon_{p-1} \gamma_{p-1}^d
\]

(S.20)

and (with \(p\) is such that \(\varepsilon_p \geq \kappa \varepsilon_{p-1} \geq \varepsilon_p\) for all \(p \geq p\))

\[
\mathbb{P}(\tilde{d}_{p}^{(2)} \in B_{\varepsilon_p}(\theta_\ast), Z_p \leq \Delta, Ω′_{p-1}, \Omega_{p, \varepsilon_p}) = 1.
\]

(S.21)

Then, combining (S.19)-(S.21) yields

\[
\mathbb{P}(\tilde{d}_{p}^{(2)} \in B_{\varepsilon_p}(\theta_\ast), Z_p \leq \Delta | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p}) \geq b_p \varepsilon_{p-1} \gamma_{p-1}^d, \quad \forall p \geq p.
\]

(S.22)

Next, for the first probability appearing on the r.h.s. of (S.18), we have

\[
\mathbb{P}(\tilde{d}_{p}^{(1)} \in B_{\varepsilon_p}(\theta_\ast), Z_p > \Delta | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
\geq (1 - b_p) \mathbb{P}(\tilde{d}_{p}^{(1)} \in B_{\varepsilon_p}(\theta_\ast), Z_p > \Delta, Ω′_{p-1}, \Omega_{p, \varepsilon_p})
\times \mathbb{P}(\tilde{d}_{p-1}^{N+1} \in B_{c_\ast \varepsilon_{p-1}}(\theta_\ast) | \tilde{d}_{p-1} \notin B_{\varepsilon_{p-1}}(\theta_\ast), \Omega_{p, \varepsilon_p})
\geq (1 - b_p) \varepsilon_{p-1} \gamma_{p-1}^d
\]

(S.23)
where the last inequality uses (S.20) and holds for $p \geq p$. We now find a lower bound for the second term on the r.h.s. of (S.23).

Let $\delta \in (0,1)$ be such that $\delta + \Delta > 1$ (note that such a $\delta$ exists because $\Delta > 0$ under C4) and, for every $p \geq 1$, let $\Psi_p = \Psi_{t_p-1:t_p}(\mu_{p-1})(B_{\kappa p_{-1}/2}(\theta_*))$ with $\mu_{p-1}$ defined in (S.12). Then,

$$P(\tilde{j}^{(1)}_{t_p} \in B_{\kappa p-1}(\theta_*) | Z_p > \Delta, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}) \geq b_{p,1} b_{p,2}$$

(S.24)

where

$$b_{p,1} := P(\tilde{j}^{(1)}_{t_p} \in B_{\kappa p-1}(\theta_*) | \Psi_p > \delta, Z_p > \Delta, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p})$$

$$b_{p,2} = P(\tilde{\Psi}_p > \delta | Z_p > \Delta, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p})$$

and we now show that for every $(c_1, c_2) \in (0,1)^2$ there exists a $p_{c_1,c_2} \in N$ (independent of $M$ and $N$) such that

$$b_{p,1} \geq (1 - b_{p}) - c_1, \quad b_{p,2} \geq c_2 - b_{p}, \quad \forall p \geq p_{c_1,c_2}.$$  

(S.25)

We start by studying $b_{p,1}$. Because $\Delta + \delta > 1$, for all $p \geq 1$ we have

$$\{\tilde{\Psi}_p > \delta, Z_p > \Delta, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}\} = \{\tilde{\Psi}_p > \delta, Z_p > \Delta, B_{(1+\kappa)p_{-1}}(\theta_*) \cap B_{(1+\kappa)p_{-1}}(\tilde{j}_{t_p-1}) \neq \emptyset, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}\}$$

(S.26)

$$= \{\tilde{\Psi}_p > \delta, Z_p > \Delta, \tilde{j}_{t_p-1} \in B_{(2+3\kappa)p_{-1}/2}(\theta_*), \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}\}$$

where the last equality uses (S.26) and the inequality follows from similar computations as in (S.15) and (S.16).

Therefore, using (S.27), for all $p \geq p$

$$b_{p,1} \geq P(\Psi_p > 1 - c'_\kappa | \tilde{\Psi}_p > \delta, Z_p > \Delta, \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p})$$

$$\geq P(\tilde{\Psi}_p > 1 - c'_\kappa | \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p})$$

$$+ P(\tilde{\Psi}_p > \delta, Z_p > \Delta | \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}) - 1$$

$$\geq P(\tilde{\Psi}_p > 1 - c'_\kappa | \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}) + P(\tilde{\Psi}_p > \delta | \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p})$$

(S.28)

$$+ P(Z_p > \Delta | \tilde{\Omega}_{p-1}, \Omega'_{p-1}, \Omega_{p,\tilde{\epsilon}_p}) - 2$$

S9
where the penultimate inequality uses Lemma S1 while the last inequality uses the fact that \( P(A \cap B) \geq P(A) + P(B) - 1 \) for all \( A, B \in \mathcal{F} \) (Fréchet’s inequality). We now find a lower bound for the each term appearing on the r.h.s. of (S.28).

To this end remark that, under C4,

\[
\Omega_{p-1} \cap \Omega_{p-1}' = \big\{ \hat{\Omega}_{p-1}, \Omega_{p-1}', \mu_{p-1}(B_{c_{\ast}p^{-1}}(\theta_{\ast})) \geq \frac{\zeta_4}{1 + \zeta_3} \big\} \quad \text{(S.29)}
\]

\[
\Omega_{p-1}' = \big\{ \Omega_{p-1}', \mu_{p-1}(B_{c_{\ast}p^{-1}}(\theta_{\ast})) \geq \frac{\zeta_2}{1 + \zeta_1} \big\}.
\]

Recall that \( c_{\ast} \leq K_{\kappa/2} \) and let \( \tau_p = t_p - t_{p-1} \).

Then, for the first term on the r.h.s. of (S.28), for all \( p \geq 1 \),

\[
\mathbb{P}\left( \tilde{\psi}_p > 1 - c'_{\kappa} | \tilde{\Omega}_{p-1}, \Omega_{p-1}', \Omega_{p_{\tilde{c}_p}} \right)
\geq \mathbb{P}\left( \tilde{\psi}_p > 1 - c'_{\kappa} | \tilde{\Omega}_{p-1}, \Omega_{p-1}' \right) - \mathbb{P}(\Omega_{p_{\tilde{c}_p}})
\geq 1 - \frac{v(\zeta_4/(1 + \zeta_3))}{\kappa - \delta} \tau_p^{-1/2} - \mathbb{P}(\Omega_{p_{\tilde{c}_p}}), \quad \text{(S.30)}
\]

with \( v() \) as in the proof of Lemma S4. The first inequality uses Lemma S1, A2, C2 and C4 while the second inequality uses in addition (S.29), Lemma S2 and Markov’s inequality.

Similarly, for the second probability on the r.h.s. of (S.28) we have for all \( p \geq 1 \)

\[
\mathbb{P}\left( \tilde{\psi}_p > \delta | \tilde{\Omega}_{p-1}, \Omega_{p-1}', \Omega_{p_{\tilde{c}_p}} \right)
\geq \mathbb{P}\left( \tilde{\psi}_p > \delta | \tilde{\Omega}_{p-1}, \Omega_{p-1}' \right) - \mathbb{P}(\Omega_{p_{\tilde{c}_p}})
\geq 1 - \frac{v(\zeta_2/(1 + \zeta_1))}{1 - \delta} \tau_p^{-1/2} - \mathbb{P}(\Omega_{p_{\tilde{c}_p}}), \quad \text{(S.31)}
\]

To find a lower bound for the third term on the r.h.s. of (S.28) remark first that, for all \( p \geq 1 \)

\[
\mathbb{P}\left( Z_p > \Delta, \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \cap \Omega_{p_{\tilde{c}_p}} \right) \leq \mathbb{P}\left( Z_p > \Delta | \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \cap \Omega_{p_{\tilde{c}_p}} \right)
\leq \mathbb{P}\left( Z_p > \Delta | \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \right) + \mathbb{P}(\Omega_{p_{\tilde{c}_p}})
\leq \mathbb{P}\left( 1 - \tilde{\psi}_p > \Delta | \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \right) + \mathbb{P}(\Omega_{p_{\tilde{c}_p}})
\leq \frac{v(\zeta_2/(1 + \zeta_1))}{\Delta} \tau_p^{-1/2} + \mathbb{P}(\Omega_{p_{\tilde{c}_p}}),
\]

where the second inequality uses Lemma S1, the third inequality uses the fact, since \( c_{\ast} \leq \kappa/2 \), we have \( \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \cap \Omega_{p_{\tilde{c}_p}} \) and the last inequality uses similar computations as in (S.31).

Therefore, for all \( p \geq 1 \),

\[
\mathbb{P}\left( Z_p > \Delta | \tilde{\Omega}_{p-1}, \Omega_{p-1}', \Omega_{p_{\tilde{c}_p}} \right)
\geq \mathbb{P}\left( Z_p > \Delta, \tilde{\Omega}_{p-1}^c \cap \Omega_{p-1}' \cap \Omega_{p_{\tilde{c}_p}} \right) \quad \text{(S.32)}
\]

\[
\geq \mathbb{P}\left( Z_p > \Delta | \Omega_{p-1}', \Omega_{p_{\tilde{c}_p}} \right) - \frac{v(\zeta_2/(1 + \zeta_1))}{\Delta} \tau_p^{-1/2} - \mathbb{P}(\Omega_{p_{\tilde{c}_p}}).
\]
Using (S.28), (S.30), (S.31), (S.32) and the fact that \( \lim_{p \to +\infty} P(\Omega_{p,c}^c) = 0 \) we conclude that for all \( c_1 \in (0,1) \) there exists a \( p_{c_1} \in \mathbb{N} \) (independent of \( M \) and \( N \)) such that \( b_{p,1} \geq (1 - b_p) - c_1 \) for all \( p \geq p_{c_1} \).

To find a lower bound for \( b_{p,2} \) note that

\[
\begin{align*}
b_{p,2} &= \mathbb{P}(\tilde{\psi}_p \in \Delta | Z_p \geq \Gamma, \Omega_{p-1}^c, \Omega_{p,\tilde{c}_p}) \\
&\geq \mathbb{P}(\tilde{\psi}_p \in \Delta | \Omega_{p-1}^c, \Omega_{p,\tilde{c}_p}) - \mathbb{P}(Z_p \leq \Gamma | \Omega_{p-1}^c, \Omega_{p,\tilde{c}_p}) \\
&\geq 1 - \frac{v(\xi_1/(1 + \xi_1))}{2} \frac{\tau_{p-1/2}}{2} - \mathbb{P}(\Omega_{p,\tilde{c}_p}^c) - b_p
\end{align*}
\]

where the first inequality uses Lemma S1 while the second inequality uses similar computations as in (S.31). Since \( \lim_{p \to +\infty} \mathbb{P}(\Omega_{p,\tilde{c}_p}^c) = 0 \), this shows that for all \( c_2 \in (0,1) \) there exists a \( p_{c_2} \geq 1 \) (independent of \( M \) and \( N \)) such that \( b_{p,2} \geq c_2 - b_p \) for all \( p \geq p_{c_2} \).

This concludes to show (S.25).

To conclude the proof let \( c \in (0,1) \) and \( f : [0,1] \to (0,+\infty) \) be the mapping defined by

\[
f(b) = b + (1 - b)(c - b)^2, \quad b \in [0,1].
\]

Notice that \( f \) is continuous and strictly positive on \( [0,1] \) so that \( \min_{b \in [0,1]} f(b) > 0 \). Therefore, using (S.18) and (S.22)-(S.25) (with \( c_1 = 1 - c \) and \( c_2 = c \)), we have for all \( p \geq p_2 \)

\[
\mathbb{P}(\tilde{\theta}_{t_{p-1}} \notin B_{t_{p-1}}(\theta_*), \Omega_{p,\tilde{c}_p}) \geq \xi \gamma_{t_{p-1}}^{\nu_{p-1}} \frac{d}{1 - \xi} (b_{p} + (1 - b_{p})(c - b_{p})^2) \\
\geq \xi \gamma_{t_{p-1}}^{\nu_{p-1}} \frac{d}{1 - \xi}.
\]

Consequently, using (S.17), for all \( p \geq p_2 \)

\[
\mathbb{P}(\tilde{\theta}_{t_{p}} \notin B_{t_{p}}(\theta_*)) \mid \tilde{\theta}_{t_{p-1}} \notin B_{t_{p-1}}(\theta_*) \leq 1 - \xi \gamma_{t_{p-1}}^{\nu_{p-1}} \frac{d}{1 - \xi} P(\Omega_{p,\tilde{c}_p}) \\
\leq 1 - \xi \gamma_{t_{p-1}}^{\nu_{p-1}} \frac{d}{1 - \xi}
\]

where, since \( \lim_{p \to +\infty} \mathbb{P}(\Omega_{p-1,r}^c) = 1 \), the last inequality holds for \( p \) sufficiently large.

The proof is complete. \( \square \)

**Theorem S1.** Assume A1-A7, C2 and C4. Then, there exists a constant \( \bar{c} \in \mathbb{R}_{>0} \) such that

\[
\sup_{p \geq 1} \rho^{p} \sup_{(M,N) \in Q_\kappa} \mathbb{P}(\tilde{\theta}_{t_{p}} \notin B_{t_{p}}(\theta_*)) \leq \bar{c}
\]

with \( Q_\kappa \) as define in Lemma S4.

**Proof.** Let \( (M,N) \in Q_\kappa \). Below \( \bar{c} > 0 \) and \( p \geq 1 \) are two finite constants (independent of \( M \) and \( N \)) whose values can change from one expression to another.

For every \( p \geq 1 \) we define

\[
b_p = \mathbb{P}(\tilde{\theta}_{t_{p}} \notin B_{t_{p}}(\theta_*)), \quad x_p = \mathbb{P}(\tilde{\theta}_{t_{p}} \notin B_{t_{p}}(\theta_*)) \mid \tilde{\theta}_{t_{p-1}} \notin B_{t_{p-1}}(\theta_*)
\]

and

\[
y_p = \mathbb{P}(\tilde{\theta}_{t_{p}} \notin B_{t_{p}}(\theta_*)) \mid \tilde{\theta}_{t_{p-1}} \in B_{t_{p-1}}(\theta_*)).
\]

S11
Then, with the convention that empty products equal one, we have

\[
b_p = b_0 \prod_{s=1}^{p} (x_s - y_s) + \sum_{s=1}^{p} y_s \prod_{j=s+1}^{p} (x_j - y_j), \quad \forall p \geq 1
\]  

(S.33)

and in the remainder of the proof we find an upper bound for each of the two terms on
the r.h.s. of (S.33).

Remark first that (with the convention that empty sums are null)

\[
t_1 \kappa^{-2(p-1)} \leq t_p \leq t_1 \kappa^{-2(p-1)} + \sum_{i=0}^{p-2} \kappa^{-2i} \leq \kappa^{-2(p-1)} (t_1 + (\kappa^{-2} - 1)^{-1}), \quad \forall p \geq 1
\]  

(S.34)

so that, by Lemma S4 and using the shorthand \( \tau_p = t_p - t_{p-1} \),

\[
y_p \leq \tilde{c} \tau_p^{-1/2} \leq \tilde{c} \kappa^p \quad \forall p \geq 1.
\]  

(S.35)

By Lemma S5, \( x_p \leq 1 - \tilde{c}^{-1} \gamma^\nu_{t_{p-1}} \epsilon_{p-1}^d \) for all \( p \geq p \). Therefore, as \( \lim_{p \to +\infty} y_p = 0 \) by (S.35), while \( \lim_{p \to +\infty} \gamma^\nu_{t_{p-1}} \epsilon_{p-1}^d = 0 \) under C2, it follows that

\[
|x_p - y_p| \leq |1 - \tilde{c}^{-1} \gamma^\nu_{t_{p-1}} \epsilon_{p-1}^d|, \quad \forall p \geq 1.
\]

Under C2, and using (S.34), \( \beta > 0 \) and \( (\gamma_t)_{t \geq 1} \) are such that \( \lim_{p \to +\infty} \epsilon_p / \gamma^\nu_{t_{p-1}} = 0 \), and thus

\[
|x_p - y_p| \leq 1 - \epsilon_{p-1}^{d + \beta}, \quad \forall p \geq p
\]  

(S.36)

where \( \epsilon_p = \left( p^{-\theta} \log(p + 1) \right)^{\frac{1}{p+1}} \) for every \( p \geq 1 \). Henceforth \( p \) is taken sufficiently large so that

\[
\epsilon_p < \min_{1 \leq s < p} \epsilon_s, \quad \forall p \geq p
\]  

(S.37)

Note that such a \( p \) exists since the sequence \( (\epsilon_p)_{p \geq 1} \) is non-increasing for \( p \) large enough.

Then, for all \( p > p \) and \( s \in \{p-1, \ldots, p-1\} \), we have

\[
\prod_{j=s+1}^{p} |x_j - y_j| \leq \exp \left( - \sum_{j=s+1}^{p} \epsilon_{j-1}^{d + \beta} \right) \leq \exp \left( - (p-s) \epsilon_{p-1}^{d + \beta} \right)
\]

\[
\leq \exp \left( - \left( 1 - \epsilon_p^\theta \right) \log(p^\theta) \right)
\]

\[= p^{-\theta} \exp \left( \frac{s}{p} \log(p^\theta) \right)
\]

where the first inequality uses (S.36) and the second inequality uses (S.37). This shows that

\[
\prod_{j=s+1}^{p} |x_j - y_j| \leq \tilde{c} p^{-\theta} \exp \left( \frac{s}{p} \log(p^\theta) \right), \quad \forall p \geq 1, \quad \forall s \in \{0, \ldots, p-1\}.
\]  

(S.38)

S12
Applying (S.38) with $s = 0$ yields, for the first term appearing on the r.h.s. of (S.33),

$$\prod_{s=1}^{p} |x_s - y_s| \leq \bar{c} p^{-e}, \quad \forall p \geq 2. \tag{S.39}$$

For the second term on the r.h.s. of (S.33) we have for all $p \geq 1$ and $s \in \{0, \ldots, p-1\}$, using (S.35) and (S.38),

$$y_s \prod_{j=s+1}^{p} |x_j - y_j| \leq \bar{c} \kappa^s p^{-e} \exp \left( \frac{s}{p} \log(p^e) \right) = \bar{c} p^{-e} \exp \left( -s \left( \log(\kappa^{-1}) - \frac{\log(p^e)}{p} \right) \right).$$

Let $p$ be large enough so that $\log(p^e)/p \leq \log(\kappa^{-1})/2$ for all $p \geq p_0$. Then, for $p \geq p_0$, we have

$$\sum_{s=1}^{p} y_s \prod_{j=s+1}^{p} |x_j - y_j| \leq \bar{c} p^{-e} \sum_{s=1}^{\infty} \exp \left( -\frac{s}{2} \log(\kappa^{-2}) \right) \leq \bar{c} p^{-e}. \tag{S.40}$$

Combining (S.33), (S.39) and (S.40) yields the result.

\[\square\]

**S3.4.2 Under Conditions C2 and C4**

**Lemma S6.** Assume A1-A5, C2, C4* and let $Q_\kappa$ be as in Lemma S4. Then, for every $(M,N) \in Q_\kappa$ there exists a constant $\bar{c} \in \mathbb{R}_{>0}$ such that

$$P(\tilde{d}_p \not\in B_{\epsilon_p}(\theta_*) \mid \tilde{d}_{p-1} \in B_{\epsilon_{p-1}}(\theta_*)) \leq \bar{c} (t_p - t_{p-1})^{-1/2}, \quad \forall p \in \mathbb{N}.$$

**Proof.** For every $p \geq 1$ let $\mu'_p = \tilde{d}_p = \tilde{d}_p^{(2)} = 1$ for all $p \geq 1$, with $\tilde{d}_p^{(2)}$ as in Lemma S3. For every $p \geq 1$ let $n_p \in 1: \tilde{N}$ be such that $\tilde{d}_p^{(2)} = \tilde{d}_p^{n_p}$ and note that $\tilde{w}_p^{n_p} \geq \tilde{N}^{-1}$.

Then, for every $p \geq 1$ we have

$$P(\tilde{d}_p \not\in B_{\epsilon_p}(\theta_*) \mid \tilde{d}_{p-1} \in B_{\epsilon_{p-1}}(\theta_*))$$

$$= P(\tilde{d}_p^{(2)} \not\in B_{\epsilon_p}(\theta_*) \mid \tilde{d}_{p-1} \in B_{\epsilon_{p-1}}(\theta_*))$$

$$\leq P(\Psi_{t_{p-1}:t_p}(\mu'_p)(B_{\epsilon_p}^{c}(\theta_*)) \geq \tilde{N}^{-1} \mid \tilde{d}_{t_{p-1}} \in B_{\epsilon_{t_{p-1}}}(\theta_*))$$

$$\leq \tilde{N} E \left[ \Psi_{t_{p-1}:t_p}(\mu'_p)(B_{\epsilon_p}^{c}(\theta_*)) \mid \tilde{d}_{t_{p-1}} \in B_{\epsilon_{t_{p-1}}}(\theta_*) \right]$$

$$\leq \tilde{N} v(c)(t_p - t_{p-1})^{-1/2}$$

with $v(\cdot)$ as in the proof of Lemma S4 and where the penultimate inequality uses Markov’s inequality while the last inequality holds under A2, C2 and C4* and for some constant $c \in \mathbb{R}_{>0}$. The proof is complete. \[\square\]
Lemma S7. Assume A1-A7, C2 and C4∗. Then, there exist constants \( c \in \mathbb{R}_{>0} \) and \( \bar{p} \in \mathbb{N} \) such that
\[
\sup_{(N,M) \in Q_\kappa} \mathbb{P}(\hat{\theta}_{t_p} \notin B_{\epsilon_p}(\theta_\ast)) \leq 1 - \frac{c^p \epsilon_p^d}{\kappa^p c_{j-1} \epsilon_p}, \quad \forall p \geq \bar{p}
\]
with \( Q_\kappa \) as defined in Lemma S4.

Proof. Under C4∗, \( \mathbb{P}(\hat{\theta}_{t_p} = \bar{\theta}^{(2)}_{t_p}) = 1 \) for all \( p \geq 1 \), with \( \bar{\theta}^{(2)}_{t_p} \) as in Lemma S3. Therefore, under A1-A7, C2 and C4∗, the results is a direct consequence of Lemma S3.

Theorem S2. Assume A1-A7, C2, C4∗ and let \( Q_\kappa \) be as in Lemma S4. Then, for every \((M, N) \in Q_\kappa\) there exists a constant \( \bar{c} \in \mathbb{R}_{>0} \) such that
\[
\sup_{p \geq 1} p^\alpha \mathbb{P}(\hat{\theta}_{t_p} \notin B_{\epsilon_p}(\theta_\ast)) \leq \bar{c}.
\]

Proof. The proof is identical to that of Theorem S1, where Lemme S6 and S7 are used in place of Lemme S4 and S5.

S3.5 A preliminary result for the sequence of PPDs

Lemma S8. Assume A1-A5, C1 and C3. Then, there exists a constant \( \bar{c} \in \mathbb{R}_{>0} \) such that, for every \((j, p) \in \mathbb{N}^2\),
\[
\sup_{(M,N) \in Q_\kappa} \mathbb{P}(\hat{\theta}_{t_{p+j}} \notin B_{\epsilon_p}(\theta_\ast)) \leq \bar{c}((\epsilon_p^2 t_p)^{-1} + \kappa^j t_p^{-1/2})
\]
with \( Q_\kappa \) as defined in Lemma S4.

Proof. Let \((M, N) \in Q_\kappa\) and, for every \( p \geq 1 \), let \( \eta_{p-1} = \frac{1}{N} \sum_{n=1}^{N} \delta_{\hat{\theta}_{t_{p-1}}^n} \). Note that, since \((M, N) \in Q_\kappa\) we have \( N \geq (2K_\kappa)^d \), with \( K_\kappa \) defined in (S.6), and thus we have, under C1.2,
\[
\mathbb{P}(\hat{\theta}_{t_{p+j}} \in B_{\epsilon_p}(\theta_\ast), \xi_{p+j-1} = \kappa^j \epsilon_p) = \mathbb{P}(\hat{\theta}_{t_{p+j}} \in B_{\epsilon_p}(\theta_\ast), \xi_{p+j-1} = \kappa^j \epsilon_p, \eta_{p-1} \geq \bar{c}K_\kappa^{-d})
\]
with \( \bar{c} > 0 \) as in the proof of Lemma S4.

Let \( \bar{\alpha} > 0 \) be as in Lemma S2, \( \kappa = \kappa \sqrt{(1 + \bar{\alpha}/2)/(1 + \bar{\alpha})} \) and note that, since \( c_j \geq c_{j-1} \) for all \( j \geq 1 \), we have
\[
\frac{(\kappa \epsilon_p c_j - \kappa^j \epsilon_p)}{\kappa^j \epsilon_p} = \frac{c_j}{c_{j-1}} - \kappa \geq \bar{\alpha}, \quad \forall (p, j) \in \mathbb{N}^2.
\]
Therefore, under the assumptions of the lemma, and using similar computations as in (S.15) and (S.16), to show the result it is enough to show that there exists a constant \( \tilde{c}_2 \in \mathbb{R}_{>0} \) (independent of \( M \) and \( N \)) such that, for every \((j,p)\) ∈ \( \mathbb{N}^2 \),

\[
\mathbb{E}\left[ \Psi_{t_{p+j-1} t_{p+j}}(\eta_{p-1}) \left( B_{(c_j^{p-1})^{-1} \epsilon_{p-1}}(\sigma_j) \right) \right] \geq \tilde{c}_2 \left( (c_j^{p-1} \epsilon_j^2 t_p)^{-1} + \kappa_j t_p^{-1/2} \right)^{-1}.
\] (S.41)

Under the assumptions of the lemma, (S.41) follows from Lemma S2 (with \( \tilde{c}_j = c_j \kappa_j^j \in (0,1] \) for all \( j \geq 0 \)) and the proof is complete.

\[ \square \]

### S3.6 Interactions between the PPDs and the auxiliary PPDs

#### S3.6.1 Under Conditions C1-C4

**Lemma S9.** Assume A1-A7 and C1-C4. Then,

\[
\sum_{p=1}^{\infty} \sup_{(M,N) \in \mathcal{Q}_N} \mathbb{P}(S_p = 1) < +\infty
\]

with \( \mathcal{Q}_N \) as defined in Lemma S4 and \( S_p = 1_{(2\epsilon_p, +\infty)}(\|\tilde{v}_p - \tilde{v}_p\|) \) for every \( p \in \mathbb{N}_0 \).

**Proof.** Let \((M,N) \in \mathcal{Q}_N\) and notice that, under the assumptions of the theorem, all the conditions of Theorem S1 and of Lemma S8 are verified. Below \( \tilde{c} \in \mathbb{R}_{>0} \) and \( p \in \mathbb{N} \) are constants (independent of \( M \) and \( N \)) whose values can change from one expression to another.

For every \((p,j) \in \mathbb{N}_0^2\) let \( \tilde{c}_{p,j} = c_j \kappa_j^j \epsilon_p \) and \( \tilde{S}_p = \inf \left\{ s \geq 1 \text{ such that } S_{p+s} = 1 \right\} \).

Then,

\[
\sum_{p=1}^{\infty} \sup_{(M,N) \in \mathcal{Q}_N} \mathbb{P}(S_p = 1) = \sum_{p=1}^{\infty} \sup_{(M,N) \in \mathcal{Q}_N} \sum_{k=1}^{\infty} \mathbb{P}(S_p = 1, \tilde{S}_p = k) \\
= \sum_{p=1}^{\infty} \sup_{(M,N) \in \mathcal{Q}_N} \sum_{k=1}^{\infty} \mathbb{P}(S_p = 1, \tilde{v}_{p+k} \in B_{(\epsilon_p)}(\theta_j), \tilde{S}_p = k) \\
+ \sum_{p=1}^{\infty} \sup_{(M,N) \in \mathcal{Q}_N} \sum_{k=1}^{\infty} \mathbb{P}(S_p = 1, \tilde{v}_{p+k} \notin B_{(\epsilon_p)}(\theta_j), \tilde{S}_p = k) \quad \text{(S.42)}
\]

and in the remainder of the proof we show that the two double series appearing on the r.h.s. of (S.42) are finite.
By Theorem S1, \( p^\mu \sup_{(M,N)\in Q_\kappa} \mathbb{P}(\bar{\vartheta}_{t\rho} \notin B_{\varrho}(\theta_\star)) \leq \bar{c} \) so that, for the first double series in (S.42), we have

\[
\sup_{j=1}^{\infty} \sum_{p=1}^{\infty} \sum_{k=1}^{\infty} \mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star)) \leq \bar{c} \sum_{p=1}^{\infty} \sum_{k=1}^{\infty} \frac{1}{(p+k)^\rho} \leq \bar{c} \sum_{p=1}^{\infty} \sum_{k=1}^{\infty} \frac{1}{(p^2+k^2)^{\rho/2}} \leq \bar{c}
\]

where the last inequality holds because \( \rho \geq 2 \) (see e.g. Borwein and Borwein, 1987, p.305).

We now study the second double series in (S.42). To this end let \( p \geq p_1 \) be large enough so that, for all \( p \geq p_1 \), we have \( \bar{\epsilon}_{p,j} \leq \epsilon_{p+j} \) for all \( j \geq 1 \). Remark that such a \( p \) exists because \( \sup_{j \geq 1} c_j \kappa^j \leq 1 \) while \( \lim_{p \to +\infty} \epsilon_p/\epsilon_{p-1} = 1 \). Then, for all \( p \geq p_1 \),

\[
\sup_{j=1}^{\infty} \sum_{k=1}^{\infty} \mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), S_p = 1, \bar{S}_p = k) \leq \mathbb{P}(\bar{\vartheta}_{t\rho+1} \notin B_{\varrho+1}(\theta_\star), S_p = 1, \bar{S}_p = 1) + \sum_{k=2}^{\infty} \mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), S_p = 1, \bar{S}_p = k) \tag{S.44}
\]

Below we find an upper bound for the two terms on the r.h.s. of (S.44).

For the first term we have, for all \( p \geq p_1 \),

\[
\mathbb{P}(\bar{\vartheta}_{t\rho+1} \notin B_{\varrho+1}(\theta_\star), S_p = 1, \bar{S}_p = 1) = \mathbb{P}(\bar{\vartheta}_{t\rho+1} \notin B_{\varrho+1}(\theta_\star), \bar{\vartheta}_{t\rho} \in B_{\varrho}(\theta_\star), \bar{\vartheta}_{t\rho} = \bar{\vartheta}_{t\rho}, \xi_p = \epsilon_p, S_p = 1, \bar{S}_p = 1) + \mathbb{P}(\bar{\vartheta}_{t\rho+1} \notin B_{\varrho+1}(\theta_\star), \bar{\vartheta}_{t\rho} \notin B_{\varrho}(\theta_\star), S_p = 1, \bar{S}_p = 1) \leq \mathbb{P}(\bar{\vartheta}_{t\rho+1} \notin B_{\varrho+1}(\theta_\star), \bar{\vartheta}_{t\rho} \in B_{\varrho}(\theta_\star), \xi_p = \epsilon_p) + \mathbb{P}(\bar{\vartheta}_{t\rho} \notin B_{\varrho}(\theta_\star)) \leq \bar{c} \left( \epsilon_p 2^{-1} + t_p^{-1/2} + p^{-\rho} \right)
\]

where the last inequality uses Theorem S1 and Lemma S8, and the fact that \( \epsilon_p = \bar{\epsilon}_{p,0} \).

We now find an upper bound for the second term on the r.h.s. of (S.44). For every \( k \geq 2 \) we have

\[
\{ S_p = 1, \bar{S}_p = k \} \subset \{ \xi_{p+j} = \bar{\epsilon}_{p,j}, \bar{\vartheta}_{p+j} = \bar{\vartheta}_{p+j} \} \quad \forall j = 1, \ldots, (k-1)
\]

and therefore, for all \( p \geq 1 \) and \( k \geq 2 \),

\[
\begin{align*}
\mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), S_p = 1, \bar{S}_p = k) &= \mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), S_p = 1, \bar{S}_p = k, \bar{\vartheta}_{t\rho+k-1} \in B_{\varrho+k-1}(\theta_\star), \xi_{p+k-1} = \bar{\epsilon}_{p,k-1}) + \mathbb{P}(\bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), S_p = 1, \bar{S}_p = k, \bar{\vartheta}_{t\rho+k-1} \notin B_{\varrho+k-1}(\theta_\star), \bar{\vartheta}_{t\rho+k-1} = \bar{\vartheta}_{t\rho+k-1}) \\
&\leq \mathbb{P}(S_p = k, \bar{\vartheta}_{t\rho+k} \notin B_{\varrho+k}(\theta_\star), \bar{\vartheta}_{t\rho+k-1} \in B_{\varrho+k-1}(\theta_\star), \xi_{p+k-1} = \bar{\epsilon}_{p,k-1}) + \mathbb{P}(\bar{\vartheta}_{t\rho+k-1} \notin B_{\varrho+k-1}(\theta_\star), S_p = 1, \bar{S}_p = k).
\end{align*}
\]
To proceed further we define, for every \((p, j, k) \in \mathbb{N}^3\),
\[
b_{j,k,p} = \mathbb{P}(\bar{S}_p = k, \hat{\theta}_{t_{p+j}} \not\in B_{\bar{\epsilon}_{p,j}}(\theta^*) | \hat{\theta}_{t_{p+j-1}} \in B_{\bar{\epsilon}_{p,j-1}}(\theta^*), \xi_{p+j-1} = \bar{\epsilon}_{p,j-1})
\]
so that, for all \(p \geq 1\),
\[
\sum_{k=2}^{\infty} \mathbb{P}(\hat{\theta}_{t_{p+k}} \not\in B_{\bar{\epsilon}_{p,k}}(\theta^*), S_p = 1, \bar{S}_p = k) \\
\leq \sum_{k=2}^{\infty} \sum_{j=2}^{k} b_{j,k,p} + \sum_{k=2}^{\infty} \mathbb{P}(\hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*), S_p = 1, \bar{S}_p = k).
\]
(S.46)

For the first double series on the r.h.s. of (S.46) we have, for every \(p \geq 1\) and \(\bar{k} \in \mathbb{N}\),
\[
\sum_{k=2}^{\bar{k}} \sum_{j=2}^{k} b_{j,k,p} = \sum_{j=2}^{\bar{k}} \sum_{k=2}^{\infty} b_{j,k,p} \\
\leq \sum_{j=2}^{\bar{k}} \sum_{k=1}^{\infty} b_{j,k,p} \\
= \sum_{j=2}^{\bar{k}} \mathbb{P}(\hat{\theta}_{t_{p+j}} \not\in B_{\bar{\epsilon}_{p,j}}(\theta^*) | \hat{\theta}_{t_{p+j-1}} \in B_{\bar{\epsilon}_{p,j-1}}(\theta^*), \xi_{p+j-1} = \bar{\epsilon}_{p,j-1}) \\
\leq \bar{c} \left( \frac{1}{\epsilon_p^{4t_{p}}} \sum_{j=1}^{\infty} c_{j-1}^2 + t_{p}^{-\frac{1}{2}} \sum_{j=1}^{\infty} \kappa_j^j \right)
\]
where the last inequality uses Lemma S8. Therefore, because \(\sum_{j=1}^{\infty} c_{j-1}^2 < +\infty\) and \(\sum_{j=1}^{\infty} \kappa_j^j < +\infty\), it follows that
\[
\sum_{k=2}^{\infty} \sum_{j=2}^{k} b_{j,k,p} \leq \bar{c} \left( \epsilon_p^{-2t_{p}^{-1}} + t_{p}^{-\frac{1}{2}} \right).
\]
(S.47)

To prepare bounding the second double series on the r.h.s. of (S.46) note that, for all \(p \geq 1\) and \(k \geq 2\), we have
\[
\mathbb{P}(\hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*), S_p = 1, \bar{S}_p = k) \\
= \mathbb{P}(\hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*), S_p = 1, \bar{S}_p = k, \hat{\theta}_{t_{p}} = \bar{\theta}_{t_{p}}, \xi_p = \epsilon_p) \\
\leq \mathbb{P}(\bar{S}_p = k, \hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*) | \hat{\theta}_{t_{p}} \in B_{\epsilon_p}(\theta^*), \xi_p = \epsilon_p) + \mathbb{P}(\hat{\theta}_{t_{p}} \not\in B_{\epsilon_p}(\theta^*), \bar{S}_p = k)
\]
so that
\[
\sum_{k=2}^{\infty} \mathbb{P}(\hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*), S_p = 1, \bar{S}_p = k) \\
\leq \mathbb{P}(\hat{\theta}_{t_{p}} \not\in B_{\epsilon_p}(\theta^*)) + \mathbb{P}(\hat{\theta}_{t_{p+1}} \not\in B_{\bar{\epsilon}_{p,1}}(\theta^*) | \hat{\theta}_{t_{p}} \in B_{\epsilon_p}(\theta^*), \xi_p = \epsilon_p) \\
\leq \bar{c} \left( \epsilon_p^{-e} + \left( \epsilon_p^{-2t_{p}^{-1}} + t_{p}^{-1/2} \right) \right)
\]
(S.48)
where the last inequality uses Theorem S1 and Lemma S8.

Therefore, by (S.44)-(S.48),
\[
\sum_{k=1}^{\infty} P(\tilde{\theta}_{tp+k} \not\in B_{tp+k}(\theta_\star), S_p = 1, \bar{S}_p = k) \leq \bar{c}(\epsilon^{-2} t_p^{-1} + t_p^{-1/2} + p^{-\theta}), \quad \forall p \geq p
\]
and thus, since \(\sum_{p=1}^{\infty} \epsilon_p^{-2} t_p^{-1} < +\infty\) and \(\sum_{p=1}^{\infty} t_p^{-1/2} < +\infty\), it follows that
\[
\sum_{p=1}^{\infty} \sup_{(M,N) \in Q_\kappa} \sum_{k=1}^{\infty} P(\tilde{\theta}_{tp+k} \not\in B_{tp+k}(\theta_\star), S_p = 1, \bar{S}_p = k) \leq \bar{c}.
\]

Together with (S.42) and (S.43), this shows the result. \(\Box\)

**S3.6.2 Under Conditions C1-C3 and C4**

**Lemma S10.** Assume A1-A7, C1-C3, C4 and let \(Q_\kappa\) be as in Lemma S4. Then, for every \((M,N) \in Q_\kappa\), we have \(\sum_{p=1}^{\infty} P(S_p = 1) < +\infty\) with \(S_p\) as defined in Lemma S9.

**Proof.** The proof of this result is similar to that of Lemma S9, where Theorem S2 is used in placed of Theorem S1. \(\Box\)

**S4 Proof of Theorem 1**

**Proof.** Below we only prove the first part of the theorem since, to prove the second part, it suffices to replace Theorem S1 by Theorem S2 and Lemma S9 by Lemma S10 in what follows.

Let \(M\) and \(N\) be as in the statement of the theorem, where \(L_\star\) is as defined in the statement of Lemma S2. Notice that \(L_\star \geq 1\) by Corollary S1. Since \((M,N) \in Q_\kappa\), with \(Q_\kappa\) as in Lemma S4, under the assumptions made in the first part of the theorem all the conditions of Theorem S1 and of Lemma S9 are fulfilled.

For every \(p \geq 1\) let \(S_p : \Omega \to \{0,1\}\) be as in the statement of Lemma S9 and \(\Omega_1 \in \mathcal{F}\) be a set of \(\mathbb{P}\)-probability one such that, for all \(\omega \in \Omega_1\),
\[
p_\omega := \inf \{p \geq 1 : S_{p+k}(\omega) = 0, \forall k \in \mathbb{N}_0\} < +\infty, \quad \lim_{p \to +\infty} \hat{\theta}_p = \theta_\star.
\]
(S.49)

Note that such a set exists by Theorem S1 and Lemma S9.

Let \(\Omega_2 \in \mathcal{F}\) be a set of \(\mathbb{P}\)-probability one such that, for every \(\omega \in \Omega_2\), we have
\[
\theta_{tp}^0(\omega) \in B_{tp}(\hat{\theta}_p), \quad \forall n \in 1 : N, \quad \forall p \in \mathbb{N}.
\]
(S.50)

Note that, because \(N \geq 2^d\), such a set exists under C1.

As a first step we show that there exist constants \(C \in \mathbb{R}\) and \(\tilde{p} \in \mathbb{N}\) such that, for every \(\omega \in \tilde{\Omega} := \Omega_1 \cap \Omega_2\), we have
\[
\|\hat{\theta}_p^\omega - \theta_\star\| \leq C \tilde{p}^{1+d} t_p^{-1/2}, \quad \forall p \geq p_\omega + \tilde{p}.
\]
(S.51)
To show (S.51) let \( \omega \in \tilde{\Omega} \). Then,
\[
\| \hat{\theta}_{tp}^\omega - \theta^* \| \leq 2 \sum_{k=p}^{\infty} \xi_k(\omega), \quad \forall p \geq p_\omega. \tag{S.52}
\]

Indeed, \( \omega \) is such that, for all \( p \geq p_\omega \), (S.50) holds while \( \hat{\theta}_{tp}^\omega = \bar{\theta}_{tp}^\omega \), and thus (S.52) is a necessary condition to have \( \lim_{p \to +\infty} \hat{\theta}_{tp}^\omega = \theta^* \).

To proceed further let \( k \geq 1 \) be such that \( c_k = k^{(1+\varepsilon)/2} \) for all \( k \geq k \). Then, for all \( p \geq p_\omega + k + 1 \), we have
\[
\sum_{k=p}^{\infty} \xi_k(\omega) = \epsilon_{p_\omega} \sum_{k=p}^{\infty} c_{k-p_\omega} \kappa^{k-p_\omega}
= \epsilon_{p_\omega} \sum_{k=p}^{\infty} (k - p_\omega)^{1+\varepsilon} \kappa^{k-p_\omega}
= \epsilon_{p_\omega} \kappa^{p-p_\omega} \sum_{k=0}^{\infty} (k + p - p_\omega)^{1+\varepsilon} \kappa^k \tag{S.53}
\]
\[
= \epsilon_{p_\omega} \kappa^{p-p_\omega} \sum_{k=0}^{\infty} \exp \left( \frac{1 + \varepsilon}{2} \log(k + p - p_\omega) - k \log(\kappa^{-1}) \right)
\leq \epsilon_{p_\omega} (p - p_\omega)^{1+\varepsilon} \kappa^{p-p_\omega} \sum_{k=0}^{\infty} \exp \left( k \left( \frac{1 + \varepsilon}{2(p - p_\omega)} - \log(\kappa^{-1}) \right) \right)
\]

where the last inequality uses the fact that \( \log(x + y) \leq \log(y) + \frac{x}{y} \) for all \( x \geq 0 \) and \( y > 0 \).

Let \( \tilde{p} = (1 + \varepsilon) \log(\kappa^{-1})^{-1} + k + 2 \) so that
\[
\frac{1 + \varepsilon}{2(p - p_\omega)} \leq \frac{1}{2} \log(\kappa^{-1}), \quad \forall p \geq \tilde{p}_\omega + \tilde{p}.
\]

Note also that, using (S.34),
\[
\kappa^{p-p_\omega} \leq t_p^{-1/2}((\kappa^{-2} - 1)^{-1} + t_p)^{1/2}, \quad \forall p \geq \tilde{p}_\omega + \tilde{p}.
\]

Therefore, using (S.53), we have
\[
2 \sum_{k=p}^{\infty} \xi_k(\omega) \leq p^{1+\varepsilon} t_p^{-1/2} C_1, \quad \forall p \geq \tilde{p}_\omega + \tilde{p} \tag{S.54}
\]

with
\[
C_1 = \sup_{\rho \geq 1} 2((\kappa^{-2} - 1)^{-1} + t_p)^{1/2} \epsilon_{\rho} \kappa^{-p} \sum_{k=0}^{\infty} \exp \left( - \frac{k}{2} \log(\kappa^{-1}) \right) < +\infty. \tag{S.55}
\]

S19
Note that \( C_1 \) is indeed finite since, by (S.53), \( t_p = O(\kappa^{-2p}) \). Together with (S.52), (S.54) shows (S.51).

To proceed further remark that, by (S.50) and (S.51), for every \( \omega \in \tilde{\Omega} \) we have

\[
\max_{n \in 1:N} \| \theta_{t_p}^{p,\omega} - \theta_n \| \leq 3C_1 p^{1+2} t_p^{-1/2}, \quad \forall p \geq p_\omega + \bar{p}
\]  

(S.56)

implying that, for every \( \omega \in \tilde{\Omega} \),

\[
\bar{\pi}_t^{N,\omega}\left( \{ \| \theta - \theta_n \| \geq 4C_1 p^{1+2} t_p^{-1/2} \} \right) = 0, \quad \forall t \in (t_p + 1), \quad \forall p \geq p_\omega + \bar{p}.
\]  

(S.57)

To conclude the proof note that, for every \( p \geq 2 \),

\[
\frac{p^{1+2}}{t_p^{1/2}} t_p^{-1/2} = \log(t) \frac{p^{1+2}}{t_p^{1/2}} t_p^{-1/2} \leq \log(t) \frac{p^{1+2}}{t_p^{1/2}} t_p^{-1/2} \leq \frac{p}{2 \log(t_1) + (p-2) \log(\kappa^{-1})} (t^{-1} t_p^{-1})^{1/2}
\]  

(S.58)

where the penultimate inequality uses (S.34) and where

\[
C_2 = \sup_{p>2} \left( \frac{p}{2 \log(t_1) + (p-2) \log(\kappa^{-1})} (t^{-1} t_p^{-1})^{1/2} < +\infty.
\]  

(S.59)

Together with (S.57), (S.58) shows that for every \( \omega \in \tilde{\Omega} \)

\[
\bar{\pi}_t^{N,\omega}\left( \{ \| \theta - \theta_n \| \geq 4C_1 C_2 \log(t) \frac{1+2}{1} t^{-1/2} \} \right) = 0, \quad \forall t \geq t_{p_\omega + \bar{p}}.
\]  

(S.60)

The proof is complete upon noting that \( P(\tilde{\Omega}) = P(\Omega_1 \cap \Omega_2) = 1 \) because \( P(\Omega_1) = P(\Omega_2) = 1 \).

\[ \square \]

**S5 Proof of Theorem 2**

**Proof.** By assumption, we have \( N^{-1} \sum_{n=1}^N \theta_{t_p} = \theta_* \) and \( \hat{N}^{-1} \sum_{n=1}^{\hat{N}} \tilde{\theta}_{t_p} = \theta_* \) and thus

\[
P(\forall p \geq 0, \hat{\theta}_{t_p} = \theta_*, \tilde{\theta}_{t_p} = \theta_*) = 1.
\]

Consequently, under C1* (resp. C2*), for any \( p \geq 1, \epsilon > 0 \) and \( N \geq 2^d \), with \( P \) probability one the set \( \hat{\theta}_{t_p} \) (resp. \( \tilde{\theta}_{t_p} \)) contains at least one point in \( B_{\epsilon/\kappa}(\theta_*) \) whenever \( \hat{\theta}_{t_p} \in B_{\epsilon}(\theta_*) \) (resp. \( \tilde{\theta}_{t_p} \in B_{\epsilon}(\theta_*) \)), with \( \kappa \) as in Lemma S2. This simple observation readily shows that Theorems S1-S2 and Lemma S9-S10 hold when, instead of being as defined in Lemma S4, the set \( Q_\kappa \) is defined by \( Q_\kappa = \{(M,N) \in \mathbb{N}^2 : N \geq K^d_\kappa \} \). The result then follows from the computations made in the proof of Theorem 1. \[ \square \]
S6 Proof of Corollary 1

Proof. Below we only prove the result under C1-C4. To prove the result under C1-C3 and C4* it suffices to replace Theorem S1 by Theorem S2 and Lemma S9 by Lemma S10 in what follows. The result under the additional conditions C1* and C2* follows from a similar argument as the one used in the proof of Theorem 2.

Let \((M, N) \in Q_\kappa\), with \(Q_\kappa\) as in Lemma S4, and \(\tilde{\Omega} \in F\) with \(\mathbb{P}(\tilde{\Omega}) = 1\) as in the proof of Theorem 1.

Then, by (S.56), for all \(\omega \in \tilde{\Omega}\) we have

\[
\|\hat{\theta}^N,\omega_{t_p} - \theta_*\| \leq 3C_1 p^{1+\varepsilon} t_p^{-1/2}, \quad \forall p \geq p_\omega + \tilde{p}
\]

with \((p_\omega, \tilde{p}) \in \mathbb{N}\) and \(C_1 \in \mathbb{R}\) as in the proof of Theorem 1.

Consequently, using the fact that (S.50) and (S.54) hold for every \(\omega \in \tilde{\Omega}\), it follows that for every \(\omega \in \tilde{\Omega}\), \(p \geq p_\omega + \tilde{p}\) and \(t \in \{t_p + 1, \ldots, t_{p+1} - 1\}\) we have

\[
\|\hat{\theta}^N,\omega_{t} - \theta_*\| \leq \|\hat{\theta}^N,\omega_{t_p} - \theta_*\| + \xi_p(\omega) \leq 3C_1 p^{1+\varepsilon} t_p^{-1/2} + \xi_p(\omega) \leq 4C_1 p^{1+\varepsilon} t_p^{-1/2}.
\]

Together with (S.58), this shows that for every \(\omega \in \tilde{\Omega}\)

\[
\|\hat{\theta}^N,\omega_{t} - \theta_*\| \leq 4C_1 C_2 \log(t)^{1+\varepsilon} t^{-1/2}, \quad \forall t \geq t_{p\omega+\tilde{p}} \quad (S.61)
\]

and the proof is complete.

S7 Proof of Theorem 3

Proof. Below we only prove the result under C1-C4, the result under the additional conditions C1* and C2* then follows from a similar argument as the one used in the proof of Theorem 2.

Let \(\Omega_1 \in \mathcal{F}\) with \(\mathbb{P}(\Omega_1) = 1\) be such that, for all \(\omega \in \Omega_1\),

\[
\lim_{p \to +\infty} \hat{\theta}_{t_p}^\omega = \theta_*, \quad \forall (M, N) \in Q_\kappa.
\]

Note that such a set exists by Theorem S1 and Lemma S9. Let \(\Omega_2 \in \mathcal{F}\) be such that \(\mathbb{P}(\Omega_2) = 1\) and such that (S.50) holds for all \(\omega \in \Omega_2\) and all \((M, N) \in Q_\kappa\).

Let \((M, N) \in Q_\kappa\) and, for every \(p \geq 1\) let \(S_p\) be as in the statement of Lemma S9 and let \(\Omega_p^{(N, M)} = \{S_{p'} = 0, \forall p' \geq p\}\). For \(\omega \in \Omega\) let

\[
p_\omega^{(N, M)} := \inf \{p \geq 1 : S_{p+k}(\omega) = 0, \forall k \in \mathbb{N}_0\}
\]

and remark that \(\Omega_p^{(N, M)} = \{p_\omega^{(N, M)} \leq p\}\).

Let \(\tilde{\Omega} \in \mathcal{F}\) be such that \(\mathbb{P}(\tilde{\Omega}) = 1\) and such that (S.60) holds for every \((M, N) \in Q_\kappa\).

Then, the computations in the proof of Theorem 1 shows that

\[
\tilde{\Omega}_p^{(N, M)} := \Omega_1 \cap \Omega_2 \cap \Omega_p^{(N, M)} \subset \tilde{\Omega}'.
\]
and thus, by (S.60) and (S.61), for every $\omega \in \tilde{\Omega}^{(N,M)}$ we have

$$\tilde{\pi}_N^\omega \left( \{ \theta \in \Theta : \|\theta - \theta_*\| \geq 4C_1 C_2 \log(t) \frac{1 + \epsilon}{2} t^{-1/2} \} \right) = 0, \ \forall t \geq t_p + \tilde{p} \geq t_p^{(N,M)} + \tilde{p}$$

and

$$\|\hat{\theta}_t^N - \theta_*\| \leq 4C_1 C_2 \log(t) \frac{1 + \epsilon}{2} t^{-1/2} \ \forall t \geq t_p + \tilde{p} \geq t_p^{(N,M)} + \tilde{p}$$

where $\tilde{p} \in \mathbb{N}$ is as in (S.60) while $C_1$ and $C_2$ are defined in (S.55) and (S.59), respectively. Since $P(\tilde{\Omega}^{(N,M)}_p) = P(\tilde{\Omega}^{(N,M)}_p)$, this shows that

$$P\left( \sup_{t \geq t_p} \tilde{\pi}_t^N \left( \{ \theta \in \Theta : \|\theta - \theta_*\| \geq 4C_1 C_2 \log(t) \frac{1 + \epsilon}{2} t^{-1/2} \} \right) = 0 \right) = P(\Omega^{(N,M)}_{p - \tilde{p}}), \ \forall p > \tilde{p}$$

and

$$P\left( \sup_{t \geq t_p} \log(t) \frac{1 + \epsilon}{2} t^{1/2} \|\hat{\theta}_t^N - \theta_*\| \leq 4C_1 C_2 \right) = P(\Omega^{(N,M)}_{p - \tilde{p}}), \ \forall p > \tilde{p}$$

where, for every $(M, N) \in \mathbb{Q}_n$ and $p \geq 1,$

$$P(\Omega^{(N,M)}_p) = 1 - P(\exists p' \geq p, S_{p'} = 1) \geq 1 - \sum_{p' = p}^{\infty} P(S_{p'} = 1) \geq 1 - \sum_{p' = p}^{\infty} \sup_{(M, N) \in \mathbb{Q}_n} P(S_{p'} = 1).$$

By Lemma S9, $\lim_{p \to +\infty} \sum_{p' = p}^{\infty} \sup_{(M, N) \in \mathbb{Q}_n} P(S_{p'} = 1) = 0$ and thus

$$\lim_{p \to +\infty} \inf_{(M, N) \in \mathbb{Q}_n} P\left( \sup_{t \geq t_p} \tilde{\pi}_t^N \left( \{ \theta \in \Theta : \|\theta - \theta_*\| \geq 4C_1 C_2 \log(t) \frac{1 + \epsilon}{2} t^{-1/2} \} \right) = 0 \right) = 1$$

$$\lim_{p \to +\infty} \inf_{(M, N) \in \mathbb{Q}_n} P\left( \sup_{t \geq t_p} \log(t) \frac{1 + \epsilon}{2} t^{1/2} \|\hat{\theta}_t^N - \theta_*\| \leq 4C_1 C_2 \right) = 1.$$
Proof. The result is a direct consequence of A1 and A5, and the proof is omitted to save space. \(\square\)

For a \(d \times d\) matrix \(A\) let \(A_{i,j}\) be its entry \((i,j)\). We then have the following result.

**Lemma S12.** Assume A1, A2 and A5 and let \(V_0\) be as in Lemma S11. Then, for every \(\epsilon > 0\) there exists a constant \(v_\epsilon \in \mathbb{R}_{>0}\) such that

\[
\lim_{t \to +\infty} \mathbb{P}\left( \max_{(i,j) \in \{1,\ldots,d\}^2} \sup_{\theta \in B_{v_\epsilon}(\theta_*)} \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_\theta(Y_s)_{ij} - (-V_0)_{ij} \right| \geq \epsilon \right) = 0.
\]

**Proof.** Let \(\tilde{m}_{\theta_*}\) be as in A5 and \(v_\epsilon \in \mathbb{R}_{>0}\) be such that we have both \(B_{v_\epsilon}(\theta_*) \subset U\) and \(v_\epsilon \leq \epsilon(4\mathbb{E}[\tilde{m}_{\theta_*}(Y_1)])^{-1}\), with \(U\) as in A5. Notice that such a \(v_\epsilon\) exists since, under A5, we have \(\mathbb{E}[\tilde{m}_{\theta_*}(Y_1)] < +\infty\). Note also that

\[
\mathbb{P}\left( \max_{(i,j) \in \{1,\ldots,d\}^2} \sup_{\theta \in B_{v_\epsilon}(\theta_*)} \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_\theta(Y_s)_{ij} - (-V_0)_{ij} \right| \geq \epsilon \right)
\leq \mathbb{P}\left( \max_{(i,j) \in \{1,\ldots,d\}^2} \sup_{\theta \in B_{v_\epsilon}(\theta_*)} \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_\theta(Y_s)_{ij} - \frac{1}{t} \sum_{s=1}^t \tilde{b}_{\theta_*}(Y_s)_{ij} \right| \geq \epsilon/2 \right)
+ \sum_{(i,j) \in \{1,\ldots,d\}^2} \mathbb{P}\left( \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_{\theta_*}(Y_s)_{ij} - \mathbb{E}[\tilde{b}_{\theta_*}(Y_1)]_{ij} \right| \geq \epsilon/2 \right)
\]

where, under A2 and A5, and by the law of large numbers, the last term converges to zero as \(t \to +\infty\). To show that the first term also converges to 0 as \(t \to +\infty\) note that, under A5 and using the mean value theorem,

\[
\max_{(i,j) \in \{1,\ldots,d\}^2} \sup_{\theta \in B_{v_\epsilon}(\theta_*)} \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_\theta(Y_s)_{ij} - \frac{1}{t} \sum_{s=1}^t \tilde{b}_{\theta_*}(Y_s)_{ij} \right| \leq v_\epsilon \frac{1}{t} \sum_{s=1}^t \tilde{m}_{\theta_*}(Y_s), \quad \mathbb{P} \text{ a.s.}
\]

Thus,

\[
\mathbb{P}\left( \max_{(i,j) \in \{1,\ldots,d\}^2} \sup_{\theta \in B_{v_\epsilon}(\theta_*)} \left| \frac{1}{t} \sum_{s=1}^t \tilde{b}_\theta(Y_s)_{ij} - \frac{1}{t} \sum_{s=1}^t \tilde{b}_{\theta_*}(Y_s)_{ij} \right| \geq \frac{\epsilon}{2} \right)
\leq \mathbb{P}\left( v_\epsilon \frac{1}{t} \sum_{s=1}^t \tilde{m}_{\theta_*}(Y_s) \geq \frac{\epsilon}{2} \right)
\leq \mathbb{P}\left( v_\epsilon \frac{1}{t} \sum_{s=1}^t \tilde{m}_{\theta_*}(Y_s) - \mathbb{E}[\tilde{m}_{\theta_*}(Y_1)] \geq \frac{\epsilon}{2} - v_\epsilon \mathbb{E}[\tilde{m}_{\theta_*}(Y_1)] \right)
\leq \mathbb{P}\left( \frac{1}{t} \sum_{s=1}^t \tilde{m}_{\theta_*}(Y_s) - \mathbb{E}[\tilde{m}_{\theta_*}(Y_1)] \geq \frac{\epsilon}{4} \right)
\]

where the penultimate inequality uses the fact that \(v_\epsilon \leq \epsilon(4\mathbb{E}[\tilde{m}_{\theta_*}(Y_1)])^{-1}\). Under A2 and A5, and by the law of large numbers, the last probability converges to zero as \(t \to +\infty\). The proof is complete. \(\square\)
S8.2 Proof of the result of Section S3.2

We first prove the following lemma which is a consequence of Kleijn and van der Vaart (2012, Theorem 3.3).

**Lemma S13.** Assume A1-A5. Then, there exist constants \((c_*, \delta_*, D, \tilde{D}) \in \mathbb{R}^4_{>0}\) such that, for all sequence \((M_t)_{t \geq 1}\) in \(\mathbb{R}_{>0}\) verifying \(\lim_{t \to +\infty} M_t = +\infty\) and \(\lim_{t \to +\infty} M_t t^{-1/2} = 0\), there exists a sequence of measurable functions \((\psi_t)_{t \geq 1}\), with \(\psi_t : \mathcal{Y}^t \to \{0, 1\}\) for all \(t \geq 1\), such that

\[
\mathbb{E}[\psi_t(Y_{1:t})] \leq \tilde{D}^{-1} \left( t^{-1/2} + \frac{e^{-\tilde{D} M_t}}{\sqrt{M_t}} \right), \quad \forall t \geq 1
\]

while, for every \(t \geq 1\) such that \(M_t^{-1} \leq \delta_*\) and \(\theta \in \Theta\) such that \(\|\theta - \theta_*\| \geq M_t t^{-1/2}\), we have

\[
\mu^L_{\theta}(1 - \psi_t(Y_{1:t})) \leq e^{-t D(\|\theta - \theta_*\|^2/\epsilon_*^2)}
\]

with the measure \(\mu^L_{\theta}\) defined in A4.

**Proof.** To show the result we need to explicit the sequence \((\psi_t)_{t \geq 1}\) used in the proof of Kleijn and van der Vaart (2012, Theorem 3.3), noticing that under A1-A5 and by Lemma S11, all the assumptions of this latter are verified.

To this end, for \(L > 0\) we let \(\tilde{l}_t^L : \mathcal{Y} \to \mathbb{R}^d\) be such that, for all \(y \in \mathcal{Y}\) and \(i \in 1:d, (\tilde{l}_t^L(y))_i = L\) if \(|(\tilde{l}_t^L(y))_i| \geq L\) and \((\tilde{l}_t^L(y))_i = (\hat{l}_t^L(y))_i\) otherwise.

For every \(t \geq 1\) let \(\psi_t^L : \mathcal{Y}^t \to \{0, 1\}\) be defined by

\[
\psi_t^L(y) = \mathbb{I} \left( \left\| \frac{1}{t} \sum_{s=1}^{t} \tilde{l}_s^L(y_s) - \mathbb{E}[\tilde{l}_s^L(Y_1)] \right\| > \sqrt{M_t/t} \right), \quad y \in \mathcal{Y}^t.
\]

Then, from the computations in the proof of Kleijn and van der Vaart (2012, Theorem 3.3), for small enough finite constants \(c_1 > 0, \delta_*> 0\) and \(\epsilon_* > 0\) and large enough finite constant \(L_* > 0\), we have

\[
\mu^L_{\theta}(1 - \psi_t^L(Y_{1:t})) \leq e^{-c_1 t \|\theta - \theta_*\|}, \quad \forall \theta \in \{\theta' \in \Theta : M_t/\sqrt{t} \leq \|\theta' - \theta_*\| \leq \epsilon_*\}
\]

for all \(t\) such that \(M_t^{-1} \leq \delta_*\). We note from the computations in the proof of Kleijn and van der Vaart (2012, Theorem 3.3) that the constants \(c_1, \delta_*, \epsilon_*\) and \(L_*\) can be made independent of the sequence \((M_t)_{t \geq 1}\).

Next, by Kleijn and van der Vaart (2012, Lemma 3.3), there exists a sequence of tests \((\psi_{2,t})_{t \geq 1}\) and constants \((c_2, c_3) \in \mathbb{R}^2_{>0}\) such that

\[
\mathbb{E}[\psi_{2,t}(Y_{1:t})] \leq e^{-tc_2}, \quad \sup_{\{\theta : \|\theta - \theta_*\| > \epsilon_*\}} \mu^L_{\theta}(1 - \psi_{2,t}(Y_{1:t})) \leq e^{-tc_3}, \quad \forall t \geq 1.
\]

Notice that \(c_2\) and \(c_3\) do not depend on \((M_t)_{t \geq 1}\).

Let \(t D = c_1 \wedge c_3\) and, for every \(t \geq 1\), let \(\psi_t = \psi_t^L \vee \psi_{2,t}\), so that the sequence \((\psi_t)_{t \geq 1}\) verifies the second part of the lemma.
To show that this sequence also verifies the first part of the lemma we define, for every $i \in 1:d$,

$$X_{t,i} = \sqrt{\frac{1}{t}} \sum_{s=1}^{t} \frac{(i_{\theta_{s}^{*}}^{L}(Y_s))_{i} - \mathbb{E}((i_{\theta_{s}^{*}}^{L}(Y_1))_{i})}{\sigma_i}, \quad \sigma_i = \sqrt{\mathbb{E}((i_{\theta_{s}^{*}}^{L}(Y_1))_{i}^2) - \mathbb{E}((i_{\theta_{s}^{*}}^{L}(Y_1))_{i})^2}.$$ 

Then, as $\sup_{y \in Y} \|i_{\theta_{s}^{*}}^{L}(y)\| \leq L < +\infty$, we have by the Berry-Esseen theorem (Shiryaev, 1996, p. 374)

$$\max_{i \in \{1,\ldots,d\}} \sup_{x \in \mathbb{R}} |\mathbb{P}(X_{t,i} \leq x) - \Phi(x)| \leq \frac{c_48L^3}{\sigma^3t^{1/2}}, \quad \tilde{\sigma} := \min_{i \in 1:d} \sigma_i, \quad \forall t \geq 1$$

for some constant $c_4 \in \mathbb{R}$ and where $\Phi : \mathbb{R} \to (0,1)$ denotes the c.d.f. of the $\mathcal{N}(0,1)$ distribution. We implicitly assume here that $\tilde{\sigma} > 0$ since other the result of the lemma trivially hold.

Then,

$$\mathbb{E}[\psi_{1,i}^{L}(Y_{1:t})] = \mathbb{P}\left(\sqrt{\frac{1}{t}} \left| \sum_{s=1}^{t} i_{\theta_{s}^{*}}^{L}(Y_s) - \mathbb{E}[i_{\theta_{s}^{*}}^{L}(Y_1)] \right| > \sqrt{M_t}\right)$$

$$\leq \sum_{i=1}^{d} \mathbb{P}(X_{t,i} > \sqrt{M_t} \sigma_i^{-1})$$

$$= \sum_{i=1}^{d} \mathbb{P}(X_{t,i} > \sqrt{M_t} \sigma_i^{-1}) + \sum_{i=1}^{d} \mathbb{P}(X_{t,i} < -\sqrt{M_t} \sigma_i^{-1})$$

$$\leq 2d \frac{c_48L^3}{\tilde{\sigma}^3t^{1/2}} + 2d\Phi\left(-\sqrt{M_t}/\tilde{\sigma}\right)$$

$$\leq 2d \frac{c_48L^3}{\tilde{\sigma}^3t^{1/2}} + 2d\tilde{\sigma} e^{\frac{-M_t}{2\tilde{\sigma}^2}}/\sqrt{M_t2\pi}$$

where the last inequality uses the fact that

$$\Phi(-x) = 1 - \Phi(x) \leq \frac{e^{-x^2/2}}{x\sqrt{2\pi}}, \quad \forall x \in \mathbb{R}_{>0}.$$ 

Therefore,

$$\mathbb{E}[\psi_{t}(Y_{1:t})] \leq \mathbb{E}[\psi_{1,i}^{L}(Y_{1:t})] + \mathbb{E}[\psi_{2,i}(Y_{1:t})] \leq 2d \frac{c_48L^3}{\tilde{\sigma}^3t^{1/2}} + 2d\tilde{\sigma} e^{\frac{-M_t}{2\tilde{\sigma}^2}}/\sqrt{M_t2\pi} + e^{-tc_2}, \quad \forall t \geq 1$$

and the result follows by taking $\tilde{D} \in (0,(2\tilde{\sigma}^2)^{-1})$ sufficiently small. \hfill \Box

**Proof of Lemma S2.** The proof is based on the proof of Kleijn and van der Vaart (2012, Theorem 3.1).
Under A2 the observations \((Y_t)_{t \geq 1}\) are i.i.d. and thus, \(\mathbb{P}\)-almost surely,
\[
E\left[\Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)}) \mid \sigma(Y_{1:(t_{p+j-1})})\right] = E\left[\Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)})\right].
\]
(S.62)

Below we find an upper bound for the second expectation.

First, notice that under A1-A5 the assumptions of Kleijn and van der Vaart (2012, Lemma 3.1 and Lemma 3.2) are verified. These two results play a key role in what follows.

For every \((p, j) \in \mathbb{N}^2\) let \(M_{p,j} = \kappa \alpha_{t_{p,j}}^{-1/2} \epsilon_{t_{p,j}}\) and \(\epsilon_{t_{p,j}} = \kappa \alpha_1^{1/2} \epsilon_{t_{p,1}}\). Then, using the fact that \(\hat{c}_j \in (0, 1]\) for all \(j \geq 1\) and the definition of \((t_p, \epsilon_p)_{p \geq 1}\) given in Algorithm 1, it is easily checked that
\[
\lim_{p \to +\infty} M_{p,j} = +\infty, \quad \lim_{p \to +\infty} M_{p,j}^{-1/2} = 0, \quad \forall j \in \mathbb{N}.
\]
(S.63)

Thus, for every \(j \in \mathbb{N}\), there exists a sequence \((M_{t_{p,j}}')_{t \geq 1}\) in \(\mathbb{R}_{\geq 0}\) such that \(M_{t_{p,j}}' = M_{p,j}\) for all \(p \in \mathbb{N}\) and such that \(\lim_{t \to +\infty} M_{t_{p,j}}' = +\infty\) and \(\lim_{t \to +\infty} M_{t_{p,j}}'^{-1/2} = 0\).

For every \(j \in \mathbb{N}\) let \((\psi_{t_{p,j}}^{(j)})_{t \geq 1}\) be a sequence of tests such that the result of Lemma S13 holds for the sequence \((M_{t_{p,j}}')_{t \geq 1}\) and let \(W_{p,j} = \psi_{t_{p,j}}^{(j)}(Y_{(t_{p,j}+1):t_{p,j+1}})\), for every \(p \in \mathbb{N}\).

Let \((\delta_*, \epsilon_*) \in \mathbb{R}^2_*\) be as in Lemma S13 and \(p^{(1)} \in \mathbb{N}\) be such that
\[
\epsilon_{t_{p,j}}' < \epsilon_*, \quad M_{p,j}^{-1} \leq \delta_*, \quad \forall p \geq p^{(1)}, \quad \forall j \in \mathbb{N}.
\]
Note that such a \(p^{(1)}\) exists since \((\hat{c}_j)_{j \geq 1}\) and \((t_p, \epsilon_p)_{p \geq 1}\) are such that
\[
\lim_{p \to +\infty} \sup_{j \geq 1} M_{p,j}^{-1} = 0, \quad \lim_{p \to +\infty} \sup_{j \geq 1} \hat{c}_{p,j} = 0.
\]

We also define for every \(\delta > 0\)
\[
B(\delta) = \left\{ \theta \in \Theta : -E\left[\log \frac{f_\theta}{f_{\theta_*}}(Y_1)\right] \leq \delta^2, \quad E\left[\left(\log \frac{f_\theta}{f_{\theta_*}}(Y_1)\right)^2\right] \leq \delta^2 \right\}
\]
and, for every \((p, j) \in \mathbb{N}^2\),
\[
\Xi_{p,j} = \left\{ \int_{\theta = \hat{c}_{p,j}}^{t_{p,j}} \prod_{s = t_{p,j-1}+1}^{t_{p,j}} \frac{f_\theta}{f_{\theta_*}}(Y_s) \eta(d\theta) \leq \eta(B(a_{p,j})) e^{-\tau_{p,j} a_{p,j}^2(1+\lambda)} \right\}
\]
(S.64)

where \(a_{p,j} = M_{p,j} \epsilon_{p,j}\) for all \((p, j) \in \mathbb{N}^2\), \(M = \sqrt{D/(1+\alpha)(1+\lambda)}\).

Then, for all \(p \geq p^{(1)}\) and \(j \geq 1\),
\[
E[\Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)})] = E[W_{p,j} \Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)})] + E[(1 - W_{p,j}) \Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)})] \leq E[W_{p,j}] + E[(1 - W_{p,j}) \Xi_{p,j}] \Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)})]
\]
(S.65)
\[
+ E[(1 - W_{p,j}) \Xi_{p,j}^\complement] \Psi_{t_{p+j-1}:t_{p+j}}(\eta)(B_{\epsilon_{t_{p+j-1}}^c(\theta_*)} \backslash B_{\epsilon_{t_{p,j}}^c(\theta_*)})] + 2P(\Xi_{p,j})
\]

S26
and in the remainder of this proof we find upper bounds for each term on the r.h.s. of (S.65).

Let \( p^{(2)} \in \mathbb{N} \) be such
\[
a_{p,j}^2 (1 + \lambda) - D \epsilon^2_k \leq -a_{p,j}^2 (1 + \lambda), \quad \forall p \geq p^{(2)}, \quad \forall j \in \mathbb{N}.
\]

Note that such an \( p^{(2)} \) exists since, for every \((p, j) \in \mathbb{N}^2\), \( \tau_{p,j}^{-1/2} M_{p,j} \leq \kappa \epsilon_{p,j-1} \leq \kappa \epsilon_p \)
where \( \lim_{p \to \infty} \epsilon_p = 0 \).

Then, by (S.64) and Lemma S13, for every \( p \geq p^{(1)} \lor p^{(2)} \) and \( j \geq 1 \), we have, under A2,
\[
E[(1 - W_{p,j}) \mathbb{I}_{\bar{p}}(p_{j+p+1}) \eta(B^c_{\alpha}(\theta_*))] \\
\leq e^{\tau_{p,j} a_{p,j}^2 (1 + \lambda)} \int_{B^c_{\alpha}(\theta_*)} \mu_{\theta}^{(1)}(1 - \psi_{\tau_{p,j}}(Y_{(t_{p,j+1} + 1)}) \eta(d\theta)) \\
\leq e^{\tau_{p,j} a_{p,j}^2 (1 + \lambda) - D \epsilon^2_k} \eta(B(a_{p,j})) \\
\leq e^{-\tau_{p,j} a_{p,j}^2 (1 + \lambda)} \frac{\eta(B(a_{p,j}))}{\eta(B(a_{p,j}))}
\]
where, for any \( t \geq 1 \), \( \mu_{\theta}^{(1)} \) is as in A4.

Next, we find a lower bound for \( \eta(B(a_{p,j})) \) by following Kleijn and van der Vaart (2012, Lemma 3.2). Under A3 and A5, and for \( \|\theta - \theta_*\| \) small enough, we have
\[
- E \left[ \log \frac{f_\theta(Y_1)}{f_{\theta_*}(Y_1)} \right] \leq d \|V_{\theta_*}\| \|\theta - \theta_*\|^2
\]
and
\[
E \left[ \left( \log \frac{f_\theta(Y_1)}{f_{\theta_*}(Y_1)} \right)^2 \right] \leq E[m_{\theta_*}^2(Y_1)] \|\theta - \theta_*\|^2.
\]
Then, for \( \delta > 0 \) small enough, \( \{\theta : \|\theta - \theta_*\| < C_* \delta \} \subset B(\delta) \) for all \( \delta \in (0, \delta) \) and with \( C_* = \left( E[m_{\theta_*}^2] \lor d \|V_{\theta_*}\| \right)^{-1/2} \).

Let \( p^{(3)} \in \mathbb{N} \) be such that \( M \kappa \epsilon_{p-1} < \delta \) for all \( p \geq p^{(3)} \) and note that, as \( \alpha > 0 \) and \( \lambda > 0 \) are such that
\[
K_c \geq \kappa^{-1} L_*(1 + \alpha)(1 + \lambda) = \frac{(1 + \alpha)(1 + \lambda)}{\kappa C_* D^{1/2}}
\]
we have \( M^{-1} C_*^{-1} K_c^{-1} \leq (1 + \alpha)/(1 + \alpha) \kappa \). Therefore, for all \( p \geq p^{(3)} \),
\[
M^{-1} \delta > \kappa \epsilon_{p,j-1} > \kappa \alpha \epsilon_{p,j-1} = \sqrt{\frac{1 + \alpha}{1 + \alpha} \kappa \epsilon_{p,j-1}} \geq M^{-1} C_*^{-1} K_c^{-1} \epsilon_{p,j-1}, \quad \forall j \in \mathbb{N}
\]
and thus \( a_{p,j} = M \kappa a \epsilon_{p,j-1} \in [C_*^{-1} K_c^{-1} \epsilon_{p,j-1}, \delta) \) for all \( p \geq p^{(3)} \) and \( j \geq 1 \). Consequently, for every \( p \geq p^{(3)} \) and \( j \geq 1 \),
\[
\eta(B(a_{p,j})) \geq \eta(B_{C_* a_{p,j}(\theta_*)}) \geq \eta(B_{\epsilon_{p,j-1}/K_c(\theta_*)})
\]
so that, for all \( p \geq p^{(1)} \lor p^{(2)} \lor p^{(3)} \),

\[
\mathbb{E} \left[ (1 - W_{p,j}) I_{\Xi_{p,j}} \Psi_{t_{p+1:d_{p+1}}}(\eta) \left( B^*_p(\theta^*_p) \right) \right] \leq \frac{e^{-M^2M_{p,j}^2(1+\lambda)}}{\eta(B_{t_{p,j-1}/K_s}(\theta^*_p))}, \quad \forall j \in \mathbb{N}. \quad \text{(S.66)}
\]

Next we consider the third term on the r.h.s. of (S.65). Let \( j \geq 1 \) and \( I_{p,j} \) be the smallest integer such that \((I_{p,j} + 1)e^\epsilon_{p,j}^\prime > \epsilon^*_p\). Then, following the computations in Kleijn and van der Vaart (2012, Theorem 3.1) and using similar computations as per above we have, for every \( p \geq p^{(1)} \lor p^{(2)} \),

\[
\mathbb{E} \left[ (1 - W_{p,j}) I_{\Xi_{p,j}} \Psi_{t_{p+1:d_{p+1}}}(\eta) \left( B^*_p(\theta^*_p) \right) \right] \leq \frac{\sum_{i=1}^{I_{p,j}} e^{\tau_{p,j}^2(1+\lambda) - \tau_{p,j} Di^2(\epsilon_{p,j}^\prime)^2} \eta\left( \{ \theta : i \epsilon_{p,j}^\prime \leq \| \theta - \theta^*_p \| \leq (i + 1) \epsilon_{p,j}^\prime \} \right)}{\eta(B(a_{p,j}))}
\]

where \( \tilde{c}_1 = \sum_{i=1}^{\infty} e^{-D(i^2 - 1)} < +\infty \). Then, using the definition of \( M \), this shows that for every \( p \geq p^{(1)} \lor p^{(2)} \),

\[
\mathbb{E} \left[ (1 - W_{p,j}) I_{\Xi_{p,j}} \Psi_{t_{p+1:d_{p+1}}}(\eta) \left( B^*_p(\theta^*_p) \right) \right] \leq \frac{\tilde{c}_1 e^{-\frac{D^2}{1+\alpha}M_{p,j}^2}}{\eta(B_{t_{p,j-1}/K_s}(\theta^*_p))}, \quad \forall j \in \mathbb{N}. \quad \text{(S.67)}
\]

For the last term on the r.h.s. of (S.65) we have by using Kleijn and van der Vaart (2012, Lemma 3.1),

\[
P(\Xi_{p,j}) \leq \frac{(1 + \alpha)(1 + \lambda)}{D \lambda^2 M_{p,j}^2}, \quad \forall (p, j) \in \mathbb{N}^2 \quad \text{(S.68)}
\]

and finally, by Lemma S13, for the first term on the r.h.s. of (S.65)

\[
\mathbb{E}[W_{p,j}] \leq \tilde{D}^{-1} \left( \tau_{p,j}^{-1/2} + e^{-\tilde{D} M_{p,j}} \sqrt{M_{p,j}} \right), \quad \forall (p, j) \in \mathbb{N}^2. \quad \text{(S.69)}
\]

By combining (S.65)-(S.69), we have for every \( p \geq p^{(1)} \lor p^{(2)} \lor p^{(3)} \) and \( j \in \mathbb{N} \),

\[
\mathbb{E} \left[ \Psi_{t_{p+1:d_{p+1}}}(\eta) \left( B^*_p(\theta^*_p) \right) \right] \leq \tilde{D}^{-1} \tau_{p,j}^{-1/2} + \tilde{D}^{-1} e^{-\tilde{D} M_{p,j}} \sqrt{M_{p,j}} + \frac{2(1 + \alpha)(1 + \lambda)}{D \lambda^2 M_{p,j}^2} + \frac{e^{-\frac{D}{1+\alpha}M_{p,j}^2} + \tilde{c}_1 e^{-\frac{D^2}{1+\alpha}M_{p,j}^2}}{\eta(B_{t_{p,j-1}/K_s}(\theta^*_p))}. \quad \text{(S.70)}
\]
To conclude the proof, remark that \( \lim_{p \to +\infty} \inf_{j \geq 1} M_{p,j}^{-1} = +\infty \) so that there exists a \( p^{(4)} \in \mathbb{N} \) such that, for all \( p \geq p^{(4)} \),

\[
\frac{e^{-DM_{p,j}}}{\sqrt{M_{p,j}}} \leq M_{p,j}^{-2}, \quad \forall j \in \mathbb{N}.
\]

Together with (S.70), and using the fact that \( e^{-x} \leq x^{-1} \) for all \( x > 0 \), this shows that, for every \( p \geq p^{(1)} \lor p^{(2)} \lor p^{(3)} \lor p^{(4)} \),

\[
\mathbb{E}\left[ \Psi_{t^{p,j}+1: t^{p,j}}(\eta) \left( B^c_{p,j}(\theta_*) \right) \right] 
\leq \tilde{D}^{-1} \tau_{p,j}^{1/2} + M_{p,j}^{-2} \left( \tilde{D}^{-1} + \frac{2(1 + \alpha)(1 + \lambda)}{D \lambda^2} + \frac{\frac{1 + \alpha}{D} + \bar{c}_1 \frac{1 + \alpha}{\lambda}}{\eta(B_{p,j-1} / K_\epsilon(\theta_*))} \right), \quad \forall j \geq 1
\]

and the result follows. \( \square \)

**S8.3 Proof of the result of Section S3.3**

We start with some additional notation that will be used in the proof of Lemma S3.

**S8.3.1 Additional notation**

Let \( U \subset \Theta \) be as in A5 and, for every \( p \in \mathbb{N} \), let \( \tau_p = t_p - t_{p-1}, L_p : \Theta \times \mathcal{Y}_p \to \mathbb{R} \), and \( H_p : \Theta \times \mathcal{Y}_p \to \mathbb{R}^{d \times d} \) be respectively defined by

\[
L_p(\theta, y) = \frac{1}{\tau_p} \sum_{s=1}^{\tau_p} \log f_{\theta}(y_s), \quad (\theta, y) \in \Theta \times \mathcal{Y}_p,
\]

and

\[
H_p(\theta, y) = -\frac{1}{\tau_p} \sum_{s=1}^{\tau_p} \tilde{L}_{\theta}(y_s), \quad (\theta, y) \in U \times \mathcal{Y}_p.
\]

For every \( (p, \omega) \in \mathbb{N} \times \Omega \), let

\[
Y^{\omega,p} = \left( Y_{s}(\omega), s \in \{t_{p-1} + 1, \ldots, t_p\} \right), \quad Y^p = \left( Y_{s}, s \in \{t_{p-1} + 1, \ldots, t_p\} \right),
\]

and, for every \( \epsilon > 0 \) and \( \delta > 0 \), and with \( \hat{U} \) as in A7, let

\[
U_{p,\epsilon} = \left\{ \theta \in B_{\epsilon}(\theta_*): L_p(\theta, y) \geq K_{p,\epsilon}(y) \right\}, \quad K_{p,\epsilon}(y) = \sup_{\theta \in \hat{U} \setminus B_{\epsilon}(\theta_*)} L_p(\theta, y),
\]

and

\[
\Omega_{p,\epsilon,\delta} = \Omega' \cap \Omega_{p,\epsilon}^{(1)} \cap \Omega_{p,\delta}^{(2)} \cap \Omega_{p,\delta}^{(3)} \cap \Omega_{p,\delta}^{(4)} \cap \Omega_{p,\delta}^{(5)}
\]
where $\Omega' \in \mathcal{F}$ is such that $\mathbb{P}(\Omega') = 1$ and such that (8) in A5 holds for all $\omega \in \Omega'$ and where

$$
\Omega^{(1)}_{p,\epsilon} = \{ \omega \in \Omega : \hat{\theta}_{\text{mle}}(Y_{\omega}^{\epsilon}) \in B_{\epsilon}(\theta_*) \},
$$

$$
\Omega^{(2)}_{p} = \{ \omega \in \Omega : \max_{(i,j) \in \{1, \ldots, d\}^2} \frac{1}{\tau_p} \sum_{s=t_p-1}^{t_p} \hat{\theta}_s(Y_{\omega}^{\epsilon})_{ij} - E[\hat{\theta}_s(Y_1)]_{ij} \leq \delta_p \},
$$

$$
\Omega^{(3)}_{p,\delta} = \{ \omega \in \Omega : \sup_{\theta \in B_{\epsilon}(\theta_*)} \max_{(i,j) \in \{1, \ldots, d\}^2} \frac{1}{\tau_p} \sum_{s=t_p-1}^{t_p} \hat{\theta}_s(Y_{\omega}^{\epsilon})_{ij} - (V_0)_{ij} \leq \delta \}, \quad \text{(S.75)}
$$

$$
\Omega^{(4)}_{p,\delta} = \{ \omega \in \Omega : \sup_{\theta \in \Theta} |L_p(\theta, Y_{\omega}^{\epsilon}) - E[\log f_0(Y_1)]| \leq \delta \},
$$

$$
\Omega^{(5)}_{p,\delta} = \{ \omega \in \Omega : \frac{1}{\tau_p} \sum_{s=t_p-1}^{t_p} \tilde{m}_{\theta_0}(Y_{s}^{\epsilon}) \leq E[\tilde{m}_{\theta_0}(Y_1)] + \delta \}.
$$

In (S.75), $\hat{\theta}_s$ and $\tilde{m}_{\theta_0}$ are as in A5, $V_0$ is as in Lemma S11 and, for every $t \in \mathbb{N}$ and $y \in \mathcal{Y}^t$, $\hat{\theta}_{\text{mle}}(y)$, is as in A7. In the definition of $\Omega^{(3)}_{p,\delta}$, $v_\delta$ is as in the statement Lemma S12 (for $\epsilon = \delta$) while, in the definition of $\Omega^{(2)}_{p}$, $(\delta_p)_{p \geq 1}$ is a sequence in $\mathbb{R}_{>0}$ such that $\lim_{p \to +\infty} \delta_p = 0$ and

$$
\lim_{p \to +\infty} \mathbb{P}\left( \{ \omega \in \Omega : \max_{(i,j) \in \{1, \ldots, d\}^2} \frac{1}{\tau_p} \sum_{s=t_p-1}^{t_p} \hat{\theta}_s(Y_{\omega}^{\epsilon})_{ij} - E[\hat{\theta}_s(Y_1)]_{ij} \leq \delta_p \} \right) = 1.
$$

Note that such a sequence $(\delta_p)_{p \geq 1}$ exists under A2 and A6 since $(t_p)_{p \geq 0}$ is such that $\lim_{p \to +\infty} (t_p - t_p - 1) = +\infty$.

The following lemma is a direct consequence of Assumptions A1-A7 and of Lemma S12.

**Lemma S14.** Assume A2-A7, let $\delta > 0$ and $(\epsilon_p)_{p \geq 0}$ be a sequence in $\mathbb{R}_{>0}$ such that $\lim_{p \to +\infty} (t_p - t_p - 1)^{1/2} \epsilon_p = +\infty$. Then, $\lim_{p \to +\infty} \mathbb{P}(\Omega_{p,\epsilon_p}) = 1$.

**S8.3.2 Proof of the lemma**

**Proof of the Lemma S3.** Let $\delta > 0$ be sufficiently small so that, for every $p \geq 1$ and $\omega \in \Omega^{(3)}_{p,\delta}$, the mapping $\theta \mapsto L_p(\theta, Y_{\omega}^{\epsilon})$ is strictly concave on $B_{v_\delta}(\theta_*)$. Note that such a $\delta > 0$ exists because $-V_\theta$ is negative definite by Lemma S11. Without loss of generality we assume below that $\hat{U} \subset B_{v_\delta}(\theta_*)$.

Let $(\epsilon'_p)_{p \geq 1}$ be a sequence in $\mathbb{R}_{>0}$ such that $\lim_{p \to +\infty} (t_p - t_p - 1)^{1/2} \epsilon'_p = +\infty$ and $\lim_{p \to +\infty} \epsilon'_p / \epsilon_p = 0$, and let $\Omega_p = \Omega_{p,\epsilon'_p}$. We assume henceforth that $p$ is such that $\epsilon'_p < \epsilon_p / 2$. 

S30
We start by first showing that
\[ \bigcup_{n=1}^{M} \{ \omega \in \Omega_p : \bar{\vartheta}_{t_p}^{N+n}((\omega)) \in U_{p,\bar{t}_p}(Y_{\omega,p}) \} \]
\subset \{ \omega \in \Omega_p : \bar{\vartheta}_{t_p}^{(2)}((\omega)) \in B_{\bar{t}_p}(\theta_*) \} \tag{S.76}

so that, \( \mathbb{P} \)-almost surely,
\[ \mathbb{P}(\bar{\vartheta}_{t_p}^{(2)} \in B_{\bar{t}_p}(\theta_*) | \Omega_p, \sigma(Y_{1:t_p}, \bar{\vartheta}_{t_0:t_{p-2}}, \theta_{1:N}^{t_0:t_{p-1}})) \]
\[ \geq \mathbb{P}(\exists n \in 1 : M \text{ s.t. } \vartheta_{t_p-1}^{N+n} \in U_{p,\bar{t}_p}(Y)^p | \Omega_p, \sigma(Y_{1:t_p}, \bar{\vartheta}_{t_0:t_{p-2}}, \theta_{1:N}^{t_0:t_{p-1}})) \tag{S.77} \]

In the remainder of the proof we show (S.76) and study probability appearing in the r.h.s. of (S.77).

For (S.76) to hold, it suffices by the definition of \( \bar{\vartheta}_{t_p}^{(2)} \) to show that
\[ L_p(\theta, Y_{\omega,p}) \geq L_p(\theta', Y_{\omega,p}), \tag{S.78} \]
for all \((\theta, \theta') \in U_{p,\bar{t}_p}(Y_{\omega,p}) \times \Theta \setminus B_{\bar{t}_p}(\theta_*)\), but by the definitions of \( K_{p,\bar{t}_p} \) and \( U_{p,\bar{t}_p} \), we only need show (S.78) for \((\theta, \theta') \in U_{p,\bar{t}_p}(Y_{\omega,p}) \times \Theta \setminus \bar{U}\).

For a sufficiently small \( \delta > 0 \) there exists an open neighbourhood \( U_\delta \subset \bar{U} \) of \( \theta_* \) such that
\[ \inf_{\theta \in U_\delta} \mathbb{E}[\log f_\theta(Y_1)] \geq 3\delta + \sup_{\theta \in \Theta \setminus \bar{U}} \mathbb{E}[\log f_\theta(Y_1)], \]
because, under A5, the mapping \( \theta \mapsto \mathbb{E}[\log f_\theta(Y_1)] \) is continuous on \( \bar{U} \) and, under A1, \( \mathbb{E}[\log f_\theta(Y_1)] > \mathbb{E}[\log f_\theta(Y_1)] \) for all \( \theta \in \Theta \setminus \{\theta_*\} \). Consequently, for every \( \omega \in \Omega_p \subset \Omega_p^{(4)} \),
\[ \inf_{\theta \in U_\delta} L_p(\theta, Y_{\omega,p}) \geq \inf_{\theta \in U_\delta} \mathbb{E}[\log f_\theta(Y_1)] - \delta \geq 2\delta + \sup_{\theta \in \Theta \setminus \bar{U}} \mathbb{E}[\log f_\theta(Y_1)] \]
\[ \geq \delta + \sup_{\theta \in \Theta \setminus \bar{U}} L_r(\theta, Y_{\omega,p}), \]
and because \( U_{p,\bar{t}_p}(Y_{\omega,p}) \subset U_\delta \), we conclude that (S.78) holds for \((\theta, \theta') \in U_{p,\bar{t}_p}(Y_{\omega,p}) \times \Theta \setminus \bar{U}\), whenever \( p \) is sufficiently large so that \( B_{\bar{t}_p}(\theta_*) \subset U_\delta \). Hence we also have (S.76) and (S.77).

Next we show there exists \( c_* \in \mathbb{R}_{>0} \) such that, that for sufficiently large \( p \),
\[ \bigg\{ \omega \in \Omega : B_{c_* \bar{t}_p}(\theta_*) \subset U_{p,\bar{t}_p}(Y_{\omega,p}) \bigg\} \cap \Omega_p = \Omega_p. \tag{S.79} \]

To simplify the notation in what follows let \( \hat{\theta}_{p,mle} = \hat{\theta}_{t_p,mle}(Y_{\omega,p}) \) and \( H_p = H_p(\hat{\theta}_{p,mle}, Y_{\omega,p}) \).
As preliminary computations to establish (S.79) we first show that there exists a sequence \((v_p)_{p \geq 1} \in \mathbb{R}_{>0} \) such that \( \lim_{p \to +\infty} v_p/\bar{t}_p^2 = 0 \) and such that, for \( p \) large enough,
\[ \left| L_p(\theta, Y_{\omega,p}) - L_p(\hat{\theta}_{p,mle}, Y_{\omega,p}) + \frac{(\theta - \hat{\theta}_{p,mle})^T V_{\theta_*}(\theta - \hat{\theta}_{p,mle})}{2} \right| \leq v_p \tag{S.80} \]
for all \((\omega, \theta) \in \Omega_p \times \overline{B_{p}(\theta_*)}\), with \(\overline{B_{p}(\theta_*)}\) the closure of \(B_{p}(\theta_*)\).

Let \(\theta \in \overline{B_{p}(\theta_*)}\) and let us consider, for \(\omega \in \Omega_p\), a second order Taylor expansion of \(L_p(\theta, Y^{\omega,p})\) at \(\hat{\theta}_{p,mle}^\omega\). We assume henceforth that \(p\) is such that \(\overline{B_{p}(\theta_*)} \cap \hat{U} = \hat{B}_{p}(\theta_*)\).

Then, for \(\omega \in \Omega_p \cap \Omega_p^{(1)} \cap \Omega_p^{(3)}\), \(\hat{\theta}_{p,mle}^\omega\) is an interior extremal point of a concave function so that the first order term must be zero. Therefore, for every \(\omega \in \Omega_p\),

\[
L_p(\theta, Y^{\omega,p}) = L_p(\hat{\theta}_{p,mle}^\omega, Y^{\omega,p}) - \frac{1}{2}(\theta - \hat{\theta}_{p,mle}^\omega)^T H_p^\omega(\theta - \hat{\theta}_{p,mle}^\omega) + R_1^{\omega,p}(\theta) \tag{S.81}
\]

where, for some constant \(\bar{c} < +\infty\) and because \(\Omega_p \subset \Omega_p^{(1)} \cap \Omega_p^{(5)}\) (recall that \(p\) is such that \(\epsilon'_p < \epsilon_p/2\)) the remainder \(R_1^{\omega,p}(\theta)\) is such that

\[
|R_1^{\omega,p}(\theta)| \leq \bar{c} \|\theta - \hat{\theta}_{p,mle}^\omega\|^3 \leq \bar{c} (3\epsilon_p/2)^3. \tag{S.82}
\]

Moreover, for every \((i, j) \in \{1, \ldots, d\}^2\) and \(\omega \in \Omega_p\), the mean value theorem yields

\[
\frac{1}{t_p - t_{p-1}} \sum_{s=t_{p-1}+1}^{t_p} \tilde{\eta}_{\omega,p,mle}^{ij}(Y_{s}^{\omega,p}) = \frac{1}{t_p - t_{p-1}} \sum_{s=t_{p-1}+1}^{t_p} \tilde{\eta}_{\omega,p}^{ij}(Y_{s}^{\omega,p}) + \tilde{R}^{\omega,p}_{ij} \tag{S.83}
\]

where, as \(\Omega_p \subset \Omega_p^{(1)} \cap \Omega_p^{(5)}\) and by assuming \(\bar{c}\) sufficiently large,

\[
|\tilde{R}^{\omega,p}_{ij}| \leq \bar{c} \|\hat{\theta}_{p,mle}^\omega - \theta_*\| \leq \bar{c} \bar{\epsilon}_p. \tag{S.84}
\]

Let \(\tilde{R}^{\omega,p} = (\tilde{R}^{\omega,p}_{ij})_{i,j=1}^d\). Then, by (S.72) and (S.83)

\[
H_p^\omega = -E[\tilde{\eta}_\omega(Y_1)] - \tilde{R}^{\omega,p} - \frac{1}{t_p - t_{p-1}} \sum_{s=t_{p-1}+1}^{t_p} \tilde{\eta}_{\omega}^{ij}(Y_{s}^{\omega,p}) + E[\tilde{\eta}_{\omega}(Y_1)].
\]

Hence by (S.84) and the fact that \(\Omega_p \subset \Omega_p^{(2)}\), we have for every \(\omega \in \Omega_p\),

\[
(\theta - \hat{\theta}_{p,mle}^\omega)^T (-E[\tilde{\eta}_\omega(Y_1)])(\theta - \hat{\theta}_{p,mle}^\omega) - d^2 \bar{c} \epsilon_p^3 - d^2 \epsilon_p^3 \delta_p
\]

\[
\leq (\theta - \hat{\theta}_{p,mle}^\omega)^T H_p^\omega(\theta - \hat{\theta}_{p,mle}^\omega)
\]

\[
\leq (\theta - \hat{\theta}_{p,mle}^\omega)^T (-E[\tilde{\eta}_\omega(Y_1)])(\theta - \hat{\theta}_{p,mle}^\omega) + d^2 \bar{c} \epsilon_p^3 + d^2 \epsilon_p^3 \delta_p. \tag{S.85}
\]

By Lemma S11, \(E[\tilde{\eta}_\omega(Y_1)] = -V_\theta\) and therefore, letting \(v_p = \bar{c} (3\epsilon_p/2)^3 + \frac{1}{2}(d^2 \bar{c} \epsilon_p^3 + d^2 \epsilon_p^3 \delta_p)\) and noting that \(\lim_{p \to +\infty} v_p/\epsilon_p^2 = 0\), (S.81), (S.82) and (S.85) show (S.80).

Next, for every \(\omega \in \Omega_p\) let \(\theta^{p,\omega} \in \{\theta \in \Theta : \|\theta - \theta_*\| = \epsilon_p\}\) be such that \(L_p(\theta^{p,\omega}, Y^{\omega,p}) = K_{p,\epsilon_p}(Y^{\omega,p})\). Note that such a \(\theta^{p,\omega}\) exists since the set \(\hat{U} \setminus \overline{B_{p}(\theta_*)}\) is compact and, under A5 and for every \(\omega \in \Omega_p \cap \Omega_p^{(1)} \cap \Omega_p^{(3)}\), the mapping \(\theta \mapsto L_p(\theta, Y^{\omega,p})\) is strictly concave on \(\hat{U}\) and attains its maximum on \(\overline{B_{p}(\theta_*)} \subset \hat{B}_{p,2}(\theta_*)\).
Then, by (S.80) and using the fact that, by Lemma S11, $V_{\theta^*}$ is positive definite, for $p$ large enough and every $\omega \in \Omega_p$, we have

$$K_{p,\hat{\theta}}(Y^{\omega,p}) \leq L_p(\hat{\theta}_{p,mle}^{\omega}, Y^{\omega,p}) - \frac{(\theta^p - \hat{\theta}_{p,mle}^{\omega})^T V_{\theta^*} (\theta^p - \hat{\theta}_{p,mle}^{\omega})}{2} + v_p \quad (S.86)$$

where $\sigma_{\min}^2 (V_{\theta^*}) > 0$ is the smallest eigenvalue of $V_{\theta^*}$.

To proceed further let $c \in (0, 1)$ and $\hat{\theta} \in B_{\epsilon_p}(\theta^*)$. Then, by (S.80), for $p$ large enough and every $\omega \in \Omega_p$, we have

$$L_p(\hat{\theta}, Y^{\omega,p}) \geq L_p(\hat{\theta}_{p,mle}^{\omega}, Y^{\omega,p}) - \frac{1}{2} \| \hat{\theta} - \hat{\theta}_{p,mle}^{\omega} \|^2 \| V_{\theta^*} \| - v_p \quad (S.87)$$

with $v_p' = \frac{1}{2} \| V_{\theta^*} \| (c^2_p)^2 + v_p$. Note that $\lim_{p \to +\infty} v_p'/c_p^2 = 0$.

Therefore, by (S.86) and (S.87), for $p$ large enough and all $\omega \in \Omega_p$, a sufficient condition to have $L_p(\hat{\theta}, Y^{\omega,p}) > K_{p,\hat{\theta}}(Y^{\omega,p})$ for every $\hat{\theta} \in B_{\epsilon_p}(\theta^*)$ is that

$$c^2 < \frac{\sigma_{\min}(V_{\theta^*})}{4 \| V_{\theta^*} \|} \frac{2 \| V_{\theta^*} \| v_p + v_p'}{c_p^2}.$$ 

Therefore, since $\lim_{p \to +\infty} (v_p + v_p')/c_p^2 = 0$, this shows (S.79) for $c_p = \sqrt{\sigma_{\min}(V_{\theta^*})/(8 \| V_{\theta^*} \|)}$.

To complete the proof let $\Omega_p' = \{ \omega \in \Omega : B_{c_\epsilon_p}(\theta^*) \subset U_{p,\hat{\theta}}(Y^{\omega,p}) \}$ so that, by (S.77) and (S.79), for $p$ large enough and $P$-almost surely,

$$P(\bar{\omega}^{(2)} \in B_{\epsilon_p}(\theta^*) | \Omega_p, \sigma(Y_{1:p}, \hat{\theta}_{1:p-2}^{1:N}, \hat{\theta}_{1:p-1}^{1:N})) \geq P(\exists n \in 1 : M \text{ s.t. } \hat{\theta}_{p-1}^{N+n} \in B_{\epsilon_p}(\theta^*) | \Omega_p', \Omega_p, \sigma(Y_{1:p}, \hat{\theta}_{1:p-2}^{1:N}, \hat{\theta}_{1:p-1}^{1:N})) \quad (S.88)$$

showing the first part of the lemma with $\Omega_p, \hat{\theta} = \Omega_p$.

To complete the proof we find a lower bound for the probability appearing on the r.h.s. of the equality sign.

Let $c_p = \Gamma((\nu + d)/2)/\Gamma(\nu/2)(\pi

\nu)^d/2$ and $c' = \| \theta^* \| \vee \sup_{\mu \in \Theta} \| g(\mu) \|$, with $g$ as in C2. Since, $P$-almost surely,

$$x^T \Sigma_{t-1}^{-1} x \leq d \| \Sigma_{t-1}^{-1} \| \| x \|^2, \quad \forall x \in \mathbb{R}^d,$$
\[ |\Sigma_{t_{p-1}}| \leq \|\Sigma_{t_{p-1}}\|^d \] and supp\(\geq 0\) \(\|\Sigma_{t_{p}}\| < +\infty\), we have, \(P\)-almost surely,

\[
\inf_{\{\|\theta\| \leq c', \|\theta'\| \leq c'\}} t_\nu(\theta; \theta', \Sigma_{t_{p-1}}) \geq c_\nu |\Sigma_{t_{p-1}}|^{-\frac{1}{2}} \left(1 + \nu^{-1}(2c')^2d\|\Sigma_{t_{p-1}}^{-1}\|\right)^{-\frac{c+d}{2}} \\
\geq \xi \|\Sigma_{t_{p-1}}\|^{-\frac{d}{2}}\|\Sigma_{t_{p-1}}^{-1}\|^{-\frac{c+d}{2}} \\
\geq \xi \gamma_{t_{p-1}}^\nu
\]

for a constant \(\xi > 0\) and where the last inequality holds under C2.

For \(p\) large enough, \(\{\vartheta_{t_{p-1}}^{N+1} \in B_{c_\epsilon \tilde{\epsilon}_p}(\theta_*)\} \subseteq \{\|\vartheta_{t_{p-1}}^{N+1}\| \leq c'\}\) so that, using (S.88) and for \(p\) sufficiently large, we have \(P\)-almost surely

\[
P(\vartheta_{t_{p}}^{(2)} \in B_{\tilde{\epsilon}_p}(\theta_*)) \mid \Omega_p, \sigma(Y_{1:t_{p}}, \theta_{1:N}^{1:t_{p}}, \theta_{0:t_{p}}^{1:N}) \\
\geq P(\exists n \in 1 : M \text{ s.t. } \vartheta_{t_{p-1}}^{N+n} \in B_{c_\epsilon \tilde{\epsilon}_p}(\theta_*) \mid \Omega_p, \sigma(Y_{1:t_{p}}, \theta_{1:N}^{1:t_{p}}, \theta_{0:t_{p}}^{1:N})) \\
\geq P(\vartheta_{t_{p-1}}^{N+1} \in B_{c_\epsilon \tilde{\epsilon}_p}(\theta_*)) \mid \Omega_p, \sigma(Y_{1:t_{p}}, \theta_{1:N}^{1:t_{p}}, \theta_{0:t_{p}}^{1:N})) \\
\geq \xi \gamma_{t_{p-1}}^\nu (c_\epsilon \tilde{\epsilon}_p)^d.
\]

The proof is complete upon noting that the above computations do not depend on \(N\) and \(M\). \(\square\)

References

Borwein, J. M. and Borwein, P. B. (1987). *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity.* Wiley, New York.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Shiryaev, A. N. (1996). *Probability.* Springer.
Supplementary Material B: Complement of Section 4

Mathieu Gerber* Kari Heine†

Contents

S1. Mean-field perturbed Bayesian inference
S1.1. Mean-field mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$: Basic idea ................. S1
S1.2. Data driven mean-field perturbed Bayesian inference ...................... S3

S2. Details about the implementation of Algorithm 1 used in Section 4 ......... S5
S2.1. Building the partitions ................................................. S5
S2.2. Support generating mappings ......................................... S5

A. Mean-field Mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$: Formal definition S6
A.1. Mean-field support generation mappings $F$ and $\tilde{F}$ .................... S6
A.2. Mean-field point estimate mappings $G$ and $\tilde{G}$ ........................ S7

S1. Mean-field perturbed Bayesian inference
S1.1. Mean-field mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$: Basic idea

Conditions MF1-MF4 on the random mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$ that underpin mean-field perturbed Bayesian inference have strong similarities with Conditions C1-C4 listed in Section 2.2 and, for that reason, their precise statements is postponed to Appendix A of this supplementary material. In this subsection we only explain the rational behind these conditions.

Let $N \geq 2d$ and

$$R_N := \min \left\{ R \in \mathbb{N} : \exists s \in \mathbb{N}^R, \sum_{r=1}^{R} s_r = d, N \geq \sum_{r=1}^{R} 2^{s_r} \right\} \quad (S.1)$$

*School of Mathematics, University of Bristol, UK.
†Department of Mathematical Sciences, University of Bath, UK.
Under MF1, for given $t \geq 1$, $\mu \in \Theta$ and $\epsilon > 0$, the mean-field support generation mapping $F$ first defines a partition $\{S_{t,r}\}_{r=1}^{R_N}$ of $1:d$ and then performs a $K_N$-exhaustive exploration of the $R_N$ subsets

$$B^{S_{t,r}}_e(\mu) := \{ \theta \in B_e(\mu) : \theta_i = \mu_i, \forall i \notin S_{t,r} \}, \quad r = 1, \ldots, R_N.$$  \hfill (S.3)

By a $K$-exhaustive exploration of the subsets defined in (S.3) we mean that the random variables $\hat{\theta}_{1:|S_{t,r}|}^{t:N} := F(t, \mu, \epsilon, N)$ is such that, for all $r \in 1:R_N$, each of the $K^{|S_{t,r}|} > 1$ hypercubes partitioning the $|S_{t,r}|$-dimensional hypercube $B^{S_{t,r}}_e(\mu)$ into hypercubes of equal volume contains at least one element of $\hat{\theta}_{1:|S_{t,r}|}^{t:N}$ with $\mathbb{P}$-probability one (see Figure S1 for an illustration). The integer $R_N$ is therefore the size of the smallest partition of $1:d$ for which a $K$-exhaustive exploration of the corresponding $R_N$ subspaces defined in (S.3) is possible for the given value of $N$. Notice that the definition of $R_N$ in (1) ensures that $R_N = 1$ when $N \geq 2^d$ while the definition of $K_N$ in (S.2) agrees with the definition of $K_N$ in (4) for all $N \geq 2^d$.

Conditions MF2.1 and MF2.2 on $\tilde{F}$ are similar to Condition MF1 on $F$ and are such that $F$ and $\tilde{F}$ define the same partition $\{S_{t,r}\}_{r=1}^{R_N}$ of $1:d$. Condition MF2.3, which assumes $M \geq (R_N + 1)M'$ for some $M' \in \mathbb{N}$, and Conditions MF3 and MF4 on the point-estimate mappings $G$ and $\tilde{G}$, are such that Property 3 holds for some models $\{f_\theta, \theta \in \Theta\}$ and for some $N < 2^d$, as we will now explain.

To this end assume that $F$ and $\tilde{F}$ are such that $\{S_{t,r}\}_{r=1}^{R_N} = \{S_r\}_{r=1}^{R_N}$ for all $t \geq 0$ and that the mean-field assumption is correct, i.e.

$$f_\theta(y) = \prod_{r=1}^{R_N} f'_{S_r, \theta}(y), \quad \forall (y, \theta) \in \mathcal{Y} \times \Theta.$$  \hfill (S.4)
Next, for every $r \in 1:R_N$, let $(\tilde{\pi}_N^{r,t})_{t \geq 1}$ be the sequence of PPDs defined in Algorithm 1, with $f_\theta = f_{S_r,\theta}^{\prime}$, $M$ replaced by $M'$, and with the support generation and point-estimate mappings such that C1-C4 hold. Let $(\pi_t^N)_{t \geq 1}$ be the sequence of distributions on $\Theta$ with support $\{(\theta_{1,t}^N, \ldots, \theta_{R_N,t}^N)\}_{n=1}^N$ and (unnormalized) weights $\{(\prod_{r=1}^{R_N} w_{r,t}^n)\}_{n=1}^N$ at time $t \in (t_p + 1):t_{p+1}$, where $\theta_{r,t}^N$ is the support of $\pi_{r,t}^N$ and $w_{r,t}^N$ the corresponding (unnormalized) weights. Notice that, under appropriate conditions on $(f_{S_r,\theta})_{r=1}^{R_N}$ and on $N$, Property 3 holds for the sequence $(\tilde{\pi}_t^N)_{t \geq 1}$.

Then, Conditions MF1-MF3 enable Algorithm 1 to define, under (S.4), a sequence of PPDs $(\tilde{\pi}_t^N)_{t \geq 1}$ that is identical to $(\pi_t^N)_{t \geq 1}$ (in the sense that the random sequences $(\tilde{\pi}_t^N)_{t \geq 1}$ and $(\pi_t^N)_{t \geq 1}$ have the same distribution) and hence that may satisfy Property 3 for some $N < 2^d$. If it would have been possible to define MF4 such that, under MF1-MF4, the sequence of PPDs defined in Algorithm 1 is identical (in the above sense) to $(\pi_t^N)_{t \geq 1}$, we follow a slightly different approach in this section for the reason that we now explain. If such a condition MF4 on $\tilde{G}$ is somewhat optimal when (S.4) holds, the constraints on the definition of the point-estimate $\hat{\theta}_{t_p}$ it imposes lead to a poor exploration of the parameter space since, under MF4 and conditionally to all the random variables involved in Algorithm 1 up to time $t_{p-1}$, with probability one $\hat{\theta}_{t_p}$ is in a subset of $\Theta$ having Lebesgue measure zero. Consequently, under only small deviations from the mean-field assumption, the property that, for every $\epsilon > 0$ the set $B_\epsilon(\theta_*)$ is visited infinitely often with probability one, may fail under MF4 (see Section 2.1 for reasons why this property is needed).

Informally speaking, Condition MF4 is designed to maximize the ability of the PPDs to explore the parameter space while preserving the convergence properties of $(\pi_t^N)_{t \geq 1}$ when (S.4) holds. Indeed, under MF1-MF4, when (S.4) is satisfied and under appropriate assumptions on $(f_{S_r,\theta})_{r=1}^{R_N}$, it can be shown that there exists an integer $\tilde{K}_* \geq 2$ such that the results of Section 3 hold when $N \geq R_N \tilde{K}_*^\text{max}_{r \in 1:R_N} |S_r|$.

At this stage it is worth mentioning that the definition of $R_N$ in (S.1) may not be satisfactory for some models. This is for instance the case when there exists a partition $(S_t^r)_{r=1}^{R_N}$ of $1:d$ such that $f_\theta = \prod_{r=1}^{R_N} f_{S_r,\theta}^{\prime}$ since, in this scenario, it is sensible to avoid to have an $R_N < R$ that imposes to take a partition $(S_{t,r})_{r=1}^{R_N}$ for which there exist $r \in 1:R$ and distinct $r_1, r_2 \in 1:R_N$ such that we have both $S_{r_1}^r \cap S_{r_1} \neq \emptyset$ and $S_{r_2}^r \cap S_{r_2} \neq \emptyset$. The implicit assumption underlying the definition (S.1) of $R_N$ is that no particular structure of the model is known, and that reducing $R_N$ as $N$ increases reduces the strength of the mean-field assumption and thus improves the statistical properties of the PPDs.

**S1.2. Data driven mean-field perturbed Bayesian inference**

In the previous subsection the partition $(S_{t,r})_{r=1}^{R_N}$ defined by the mappings $F$ and $\tilde{F}$ depends on $t$ and is thus allowed to evolve over time, offering the possibility to learn in the course of Algorithm 1 a partition $(\tilde{S}_r)_{r=1}^{R_N}$ that ‘minimizes’ the impact of the mean-field assumption on the inference. When no particular structure of the model at hand is known, such an automatic and data driven partitioning of the parameter $\theta$ is a sensible strategy.
Let \( \mu_{t_{p-1}}(d\theta) = \otimes_{i=1}^{d} \mu_{i,t_{p-1}}(d\theta_i) \) be a probability measure on \( \Theta \), independent of \( (Y_t)_{t \geq t_{p-1}} \), and let \( \pi_{\tau|t_{p-1}} = \Psi_{t_{p-1}:t_{p-1}+\tau}(\mu_{t_{p-1}}) \) for every \( \tau \in \mathbb{N} \). Then, at perturbation time \( t_p + 1 \) we propose to define \( \{S_{t_{p+1},r}\}_{r=1}^{R_N} \) as

\[
\{S_{t_{p+1},r}\}_{r=1}^{R_N} \in \arg\min_{\{S_r\}_{r=1}^{R_N}} \sum_{r \neq r'} \sum_{i \in S_r} \sum_{i' \in S_{r'}} |\rho_{p,ii'}|
\]

(S.5)

where \( (\rho_{p,ij})_{i,j=1}^{d} \) is the correlation matrix of \( \theta \) under \( \pi_{\tau|t_{p-1}} \), with \( \tau_p \in (1:(t_p - t_{p-1})) \) such that \( \tau_p \to +\infty \) as \( p \to +\infty \), \( \mathbb{P} \)-almost surely. The rational behind this approach is simple. On the one hand, when (S.4) holds then the random variables \( \{\theta_{t_{r}}\}_{r=1}^{R_N} \) are independent under \( \pi_{\tau|t_{p-1}} \) and thus \( \{S_{t_{p+1},r}\}_{r=1}^{R_N} = \{S_r\}_{r=1}^{R_N} \) in (S.5). On the other hand, this latter observation suggests that, when defined as in (S.5), the partition \( \{S_{t_{p+1},r}\}_{r=1}^{R_N} \) should be close to \( \{S_r\}_{r=1}^{R_N} \) when \( f_{\theta} \approx \prod_{r=1}^{R_N} f_{\theta_{t_{r}}} \). The condition \( \tau_p \to +\infty \) is made to allow the resulting sequence \( \{(S_{t_{p+r}})_{r=1}^{R_N}\}_{p \geq 1} \) to converge to some fixed partition \( \{S_r\}_{r=1}^{R_N} \) of \( d \).

To make this approach useful in practice, we however need, at every perturbation time \( t_p + 1 \), to have an approximation \( \tilde{\rho}_p \) of the unknown matrix \( \rho_p := (\rho_{p,ij})_{i,j=1}^{d} \). We propose to take for \( \tilde{\rho}_p \) a Monte Carlo estimate of \( \rho_p \). To this end, for some \( N_{mf} \in \mathbb{N} \) we sample \( \tilde{\rho}_{t_{p-1}}^{1:N_{mf}} \sim \mu_{t_{p-1}} \), compute the weights

\[
W_p^n = \frac{\prod_{s=t_{p-1}+1}^{t_{p-1}+\tau_p} f_{\tilde{\rho}_{t_{p-1}}}(Y_s)}{\sum_{n'=1}^{N_{mf}} \prod_{s=t_{p-1}+1}^{t_{p-1}+\tau_p} f_{\tilde{\rho}_{t_{p-1}}'}(Y_s)} \quad n = 1, \ldots, N_{mf}
\]

and define \( \hat{\rho}_p = (\hat{\rho}_{p,ij})_{i,j=1}^{d} \) as the empirical correlation matrix of the weighted point set \( \{(W_p^n, \tilde{\rho}_{t_{p-1}}^{n})\}_{n=1}^{N_{mf}} \). The quantity \( \hat{\rho}_p \) is an importance sampling estimate (Robert and Casella, 2004, Section 3.3) of \( \rho_p \) and, following the standard approach in the Monte Carlo literature, \( \tau_p \) is taken so that, with high probability, the effective sample size \( \text{ESS}_p = (\sum_{n=1}^{N_{mf}} (W_p^n)^2)^{-1} \) is larger than \( cN_{mf} \), with \( c \in (0,1) \) a tuning parameter. A precise definition of \( \mu_{t_{p-1}} \) and \( \tau_p \) is given in Section S2 below.

Given the Monte Carlo estimate \( \hat{\rho}_p \) of the matrix \( \rho_p \), we can in practice define the partition \( \{S_{t_{p+1},r}\}_{r=1}^{R_N} \) as the solution of the minimum \( R_N \)-cut problem

\[
\arg\min_{\{S_r\}_{r=1}^{R_N}} \sum_{r \neq r'} \sum_{i \in S_r} \sum_{i' \in S_{r'}} |\hat{\rho}_{p,ii'}|
\]

(S.6)

at the cost of \( O(d^{R_N}) \) operations (Goldschmidt and Hochbaum, 1988). Alternatively, \( \{S_{t_{p+1},r}\}_{r=1}^{R_N} \) can be defined as the solution of a fast approximate algorithm to solve (S.6), such as the one proposed in Saran and Vazirani (1995) that requires only \( O(d) \) operations.
S2. Details about the implementation of Algorithm 1 used in Section 4

For every \( p \geq 1 \) the probability distribution \( \mu_{\text{aux},dp} \) used to sample \( \tilde{\vartheta}_{tp}^{1:Naux} \) is such that (with \( \Sigma_{tp} \), the matrix in C2.3 when \( t = tp \), see below)

\[
\tilde{\vartheta}_{tp}^{n} \sim \mathcal{U}(\tilde{\vartheta}_{tp}^{(2)} - \xi_{p}, \tilde{\vartheta}_{tp}^{(2)} + \xi_{p}), \quad n \in 1:Nm_{mf}, \quad \tilde{\vartheta}_{tp}^{n} \sim \mathcal{N}(\tilde{\vartheta}_{tp}^{(2)}, \Sigma_{tp}), \quad n \in (Nm_{mf} + 1):Naux
\]

where \( N_{mf} = \lfloor Naux/2 \rfloor \).

For \( (p, \tau) \in \mathbb{N}^{2} \) let

\[
\tilde{w}_{p,\tau}^{n} = \prod_{s=tp-1+\tau}^{tp} f_{\vartheta_{p-1}}(Y_{s}), \quad n \in 1:Naux, \quad W_{p,\tau}^{n} = \frac{\tilde{w}_{p,\tau}^{n}}{\sum_{n'=1}^{Nm_{mf}} \tilde{w}_{p,\tau}^{n'}}, \quad n \in 1:Nm_{mf}.
\]

S2.1. Building the partitions

At time \( tp + 1 \) the partition \( \{S_{tp+1,r}\}_{r=1}^{R_{\text{aux}}} \) is built using the approach proposed in Section S1.2, with \( \mu_{p-1} \) being the \( \mathcal{U}(\tilde{\vartheta}_{tp}^{(2)} - \xi_{p}, \tilde{\vartheta}_{tp}^{(2)} + \xi_{p}) \) distribution, \( W_{p}^{1:Naux} = W_{p,\tau}^{1:Nm_{mf}} \) and where

\[
\tau_{p} = \lfloor (tp - tp-1)^{1}/T_{p} \rfloor
\]

with \( T_{0} = 3 \) \( \mathbb{P} \)-almost surely and, for every \( p \geq 1 \),

\[
T_{p} = \begin{cases} 
T_{p-1} + 0.1, \quad \text{ESS}_{p-1,\tau_{p-1}} < [Nm_{mf}/2]/4 \\
\max(1, T_{p-1} - 0.1), \quad \text{ESS}_{p-1,\tau_{p-1}} > [Nm_{mf}/2]/4 \quad (S.7) \\
T_{p-1}, \quad \text{otherwise.}
\end{cases}
\]

The definition of \( T_{p} \) in (S.7) is such that the effective sample size \( \text{ESS}_{p,\tau_{p}} \), is ‘most of the time’ smaller than \( 3[Nm_{mf}/2]/4 \) and larger than \( [Nm_{mf}/2]/4 \). Notice that typically \( T_{p} > 2 \) since it is expected that \( \xi_{p} \) is of order \( (tp - tp-1)^{-1/2} \) (discarding the logarithmic term) while the Bayes updates are such that the posterior distribution concentrates on \( \theta_{*} \) at rate \( t^{-1/2} \).

S2.2. Support generating mappings

For \( t \in (tp)_{p \geq 1} \) the matrix \( \Sigma_{t} \) in C2 is defined by \( \Sigma_{tp} = 10 \hat{\rho}_{p} \) (with \( \hat{\rho}_{p} \) as defined in Section S1.2) and the element \( \vartheta_{tp}^{N} \) of the support of \( \tilde{\vartheta}_{tp+1} \) is such that

\[
\prod_{s=tp-1+1}^{tp} f_{\vartheta_{p}}(Y_{s}) \geq \max \left( \max_{1:Naux} \tilde{w}_{tp}^{n}, \max_{n \in 1:N} w_{tp}^{n} \right).
\]

Notice that, in (6), we therefore have \( N_{\text{exp}} = Naux + N + \tilde{N} \).

The support generating mapping \( F \) is such that, in MF1,

\[
\mathbb{P}(\exists n \in 1:N, \theta_{t}^{n} = z_{j,r}) = 1, \quad \forall j \in 1:K_{r,N}, \quad \forall r \in 1:R_{N},
\]

\[
S5
\]
where $z_{j,r}$ denotes the centroid of the $|S_{t,r}|$-dimensional hypercube $B_{j,r/K_{r,N}}^{S_{t,r}}(\mu)$ and the integers $\{K_{r,N}\}_{r=1}^{R_N}$ are such that

$$\min_{r \in 1:R_N} K_{r,N} \geq K_N, \quad N \geq \sum_{r=1}^{R_N} \nu_{|S_{t,r}|}$$

and $N < \sum_{r=1}^{R_N} (K_{r,N} + i_r)^{|S_{t,r}|}$ for all $(i_1, \ldots, i_{R_N}) \in \{0,1\}^{R_N}$ such that $\sum_{r=1}^{R_N} i_r = 1$. If $N > \sum_{r=1}^{R_N} \nu_{|S_{t,r}|}$, the remaining points are randomly sampled so that MF1 holds.

The support generating mapping $\tilde{F}$ is such that, in MF2, the first $N$ points of $\vartheta_1^1: \tilde{N}$ are generated in a similar way.

### A. Mean-field Mappings $F$, $\tilde{F}$, $G$ and $\tilde{G}$: Formal definition

For $\mu \in \Theta$, $u \subset 1:d$ a non-empty set, $k \in \mathbb{N}$ and $\epsilon > 0$, let $B_u^k(\mu)$ be as in (S.3) and $\{B_{j/k}^u(k)\}_{j=1}^{k|u|}$ be the partition $B_u^k(\mu)$ into $k|u|$ hyper-rectangles such that, for all $j \in 1:k|u|$, the hypercube $\{x_u \in \mathbb{R}^{|u|} : (x_u, \mu) \in B_{j/k}^u(k)\}$ has volume $(2\epsilon/k)^{|u|}$.

Here and below, for $x_u \in \mathbb{R}^{|u|}$ and $x' \in \mathbb{R}^{d-|u|}$, when $y \in \mathbb{R}^d$ is written as $y = (x_u, x'_{d\setminus u})$ it should be understood that $y_i = x_i$ if $i \in u$ and $y_i = x'_i$ if $i \notin u$.

The conditions introduced in the following are assumed to hold for arbitrary $N \geq 2d$, partition $\{S_{t,r}\}_{r=1}^{R_N}$ of $1:d$ such that $\sum_{r=1}^{R_N} \nu_{|S_{t,r}|} \leq N$ and

$$\mu \in \Theta, \quad (t, M') \in \mathbb{N}^2, \quad (w^1:N, \tilde{w}^1:N, \kappa, \beta, \epsilon) \in \mathbb{R}^{3+N}, \quad (\vartheta^1:N, \vartheta^1:\tilde{N}) \in \Theta^{N+\tilde{N}}$$

with $\tilde{N} = N + (R_N + 1)M'$. We also define $M = (R_N + 1)M'$ and $\bar{\chi} = (N, M, \kappa, \beta)$.

#### A.1. Mean-field support generation mappings $F$ and $\tilde{F}$

**Condition MF1.** Let $R_N$ and $K_N$ be as in (S.1) and (S.2). The $\Theta^N$ valued random variable $\vartheta^1:N := F(t, \mu, \epsilon, N)$ satisfies, for some $C \in \mathbb{R}_{>0}$,

1. $\vartheta^1:N$ is independent of $(Y_s)_{s>1}$.
2. $\mathbb{P} \left( 1 \leq \sum_{n=1}^{N} 1(\vartheta^1_n \in B_{j,K_{r,N}}^{S_{t,r}}(\mu)) \leq C \right) = 1$ for all $r \in 1:R_N$ and $j \in 1:K_{N}^{S_{t,r}}$.
3. $\mathbb{P} \left( \forall n \in 1:N, \exists r \in 1:R_N \ s.t. \ \vartheta^1_n \in B_{r/K_{r,N}}^{S_{t,r}}(\mu) \right) = 1$.

**Condition MF2.** Let $R_N$ and $K_N$ be as in (S.1) and (S.2). The $\Theta^{\tilde{N}}$ valued random variable $\vartheta^1:\tilde{N} := \tilde{F}(t, \mu, \epsilon, \bar{\chi})$ satisfies, for some $C \in \mathbb{R}_{>0}$,

1. $\vartheta^1:\tilde{N}$ is independent of $(Y_s)_{s>1}$.
2. $\mathbb{P} \left( 1 \leq \sum_{n=1}^{N} 1(\vartheta^1_n \in B_{j,K_{r,N}}^{S_{t,r}}(\mu)) \leq C \right) = 1$ for all $r \in 1:R_N$ and $j \in 1:K_{N}^{S_{t,r}}$.
3. $\vartheta^1(\tilde{N}+rn) = (\mu_{S_{t,r}}, \vartheta^1_{S_{t,r}}(n+RN'M'+n))$ for all $r \in 1:R_N$ and $n \in 1:M'$, where $\vartheta^1_n \sim t_{\nu}(g(\mu), \Sigma_t)$, with $g(\cdot)$ and $\Sigma_t$ as in C2.3.
A.2. Mean-field point estimate mappings $G$ and $\tilde{G}$

For $r \in 1:RN$ let

$$I_r = \{ n \in (1:N) : \theta^n \in B_{t,w}^{S_{r,r}}(\mu) \}, \quad \tilde{I}_r = \{ n \in (1:(N + R_NM')) : \theta^n \in B_{t,w}^{S_{r,r}}(\mu) \}.$$ 

Condition MF3. Let $R_N$ be as in (S.1). The $\Theta$ valued random variable $\tilde{\theta}_t := G(t,w^{1:N},\theta^{1:N},N)$ satisfies

1. $\tilde{\theta}_t$ is independent of $(Y_s)_{s>t}$.
2. $\mathbb{P}(\forall r \in 1:R_N, \tilde{\theta}_{t,S_{r,r}} = \sum_{n \in 1:I_r} \mathbb{E}[\theta^n_{S_{r,r}}]) = 1.$

Condition MF4. Let $R_N$ be as in (S.1). The $\Theta$ valued random variable $\tilde{\theta}_t := \tilde{G}(t,\epsilon,\mu,\bar{w}^{1:N},\bar{N},\bar{\lambda})$ satisfies, for some $\Delta \in (0,1)$, $(\zeta_1, \zeta_3) \in \mathbb{R}^2_{>0}$ and $(\zeta_2, \zeta_4) \in (0,1)^2$,

1. $\tilde{\theta}_t$ is independent of $(Y_s)_{s>t}$.
2. $\mathbb{P}(\tilde{\theta}_t = \tilde{\theta}_t^{(1)}(Z_t > \Delta) + \tilde{\theta}_t^{(2)}(Z_t \leq \Delta)) = 1$ with $\tilde{\theta}_t^{(2)} = \theta_{n'}$, where $n' \in \arg\max_{n \in 1:N} \bar{w}^n$, $\tilde{\theta}_t^{(1)} = (\tilde{\theta}_t^{(1)}, r \in 1:R_N)$, $Z_t = \min_{r \in 1:R_N} Z_r$, where, for $r \in 1:R_N$,

$$Z_r = \sum_{n \in 1:I_r} \frac{\tilde{a}_n^{(1)} \bar{w}^n}{\sum_{n' \in 1:I_r} \tilde{a}_n^{(1)}} \mathbb{I}(\theta^n \in B_{(1+n)r}^{S_{r,r}}(\mu)),$$

$$\tilde{a}_n^{(1)} = \begin{cases} \frac{\zeta_1 M'}{|I_r|}, & n \in I_r \\ \zeta_2 M', & n = N + r \\ (1 - \zeta_2), & n > N + r \end{cases}$$

and

$$\tilde{\theta}_t^{(1)} = \sum_{n \in 1:I_r \cup J_r} \frac{\tilde{a}_n^{(2)}(J_r) \bar{w}^n}{\sum_{n' \in 1:I_r \cup J_r} \tilde{a}_n^{(2)}(J_r) \bar{w}^n} \mathbb{I}(\theta^n \in B_{(1+n)r}^{S_{r,r}}(\mu)),$$

$$\tilde{a}_n^{(2)}(J_r) = \begin{cases} \frac{\zeta_3 |J_r|}{|I_r|}, & n \in I_r \\ \zeta_4 |J_r|, & n \in J_r \cap \{ N + r \} \\ (1 - \zeta_4), & n \in J_r \setminus \{ N + r \} \end{cases}$$

with $J_r = \{ n \in (N + r):(N + rM') \text{s.t. } \theta^n \in B_{(1+2n)r}^{S_{r,r}}(\mu) \}.$

References

Goldschmidt, O. and Hochbaum, D. S. (1988). Polynomial algorithm for the $k$-cut problem. In *Foundations of Computer Science, 1988., 29th Annual Symposium on*, pages 444–451. IEEE.

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

Saran, H. and Vazirani, V. V. (1995). Finding $k$ cuts within twice the optimal. *SIAM Journal on Computing, 24*(1):101–108.