Asymptotic Behaviour of the Posterior Distribution in Approximate Bayesian Computation

BY THOMAS A. DEAN[1], SUMEETPAL S. SINGH[2] & AJAY JASRA[3]

[1]Dogtooth, Cambridge, UK. E-Mail: *thomas.dean@cantab.net*

[2]Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK. E-Mail: *sss40@cam.ac.uk*

[3]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955, KSA. E-Mail: *ajay.jasra@kaust.edu.sa*

Abstract

Approximate Bayesian computation (ABC) is a popular technique for approximating likelihoods and is often used in parameter estimation when the likelihood functions are analytically intractable. In the context of Hidden Markov Models (HMMs), we analyse the asymptotic behaviour of the posterior distribution in ABC based Bayesian parameter estimation. In particular we show that Bernstein-von Mises type results still hold but that the resulting posterior is biased in the sense that it concentrates around a point in parameter space that differs from the true parameter value. Furthermore we obtain precise rates for the size of this bias with respect to a natural accuracy parameter of the ABC method. Finally we discuss, via a numerical example, the implications of our results for the practical implementation of ABC.

Key words: Parameter Estimation, Hidden Markov model, Approximate Bayesian Computation

1 Introduction

One of the most fundamental problems in statistics is that of parameter estimation. Suppose that one has a collection of probability laws $\mathbb{P}_\theta$ parametrised by a collection of parameter vectors $\theta \in \Theta$. Suppose further that one has data $\hat{Z}$ generated by a process distributed according to some law $\mathbb{P}_{\theta^*}$ where the exact value of $\theta^* \in \Theta$ is unknown. The problem of parameter estimation is to infer the value of the unknown parameter vector θ^* from the data $\hat{Z}$. Many standard methods for estimating the value of θ^* are based upon using the likelihood function $p_\theta(\hat{Z})$. For example Bayesian approaches use the likelihood to reweight some prior distribution to obtain a posterior distribution on the space of parameter vectors that represents ones sense of certainty of any given parameter vector being equal to θ^*. Alternatively one may take a frequentist approach and estimate θ^* with the parameter vector which maximises the value of the corresponding likelihood (ie. maximum likelihood estimation (MLE)).

Of course these approaches all rely on one being able to compute the likelihood functions $p_\theta(\hat{Z})$, either exactly or numerically. However, in a wide range of applications this is not possible, either because no analytic expression for the likelihoods exists or else because computing them is computationally intractable. Despite this one is often still able, in such cases, to generate random variables distributed according to the corresponding laws $\mathbb{P}_\theta$. This has led to the development of methods in which θ^* is estimated by implementing a standard likelihood based parameter estimator using some principled approximation to the likelihood instead of the true likelihood function itself. In general these approximations are estimated using Monte Carlo simulation based on generating samples from the relevant probability distributions.

A method which has become very popular in practice and on which we shall focus our attention for the rest of this paper is approximate Bayesian computation (ABC). A non-exhaustive list of references for applications of the method includes: McKinley et al. (2009); Peters et al. (2010); Pritchard et al. (1999); Ratmann et al. (2009); Tavre et al. (1997). The standard ABC approach to approximating the likelihood is as follows. Suppose that the distributions $\mathbb{P}_\theta$ all have a density $p_\theta(\cdot)$ on some space $\mathbb{R}^m$ w.r.t. some dominating measure μ. Furthermore suppose that the functions $p_\theta(\cdot)$ cannot be evaluated directly but that one can generate random variables distributed according to the laws $\mathbb{P}_\theta$. Given some data $\hat{Z}$ the general ABC approach to approximating the values of the likelihood functions $p_\theta(\hat{Z})$ is to choose a metric $d(\cdot,\cdot)$ on $\mathbb{R}^m$ and a tolerance parameter $\epsilon > 0$ and for all $\theta \in \Theta$ approximate the likelihood $p_\theta(\hat{Z})$ with

$$p_\theta^\epsilon(\hat{Z}) \triangleq \frac{1}{\mu\left(B_{\hat{Z}}^\epsilon\right)} \mathbb{P}_\theta\left(d(\hat{Z}, Z) \leq \epsilon\right). \tag{1}$$

where $B_{\hat{Z}}^\epsilon$ denotes the set of all points at a distance less than or equal to ϵ from $\hat{Z}$. Typically the probabilities (1) are then estimated via naïve Monte Carlo simulation.

Intuitively,

$$\frac{1}{\mu\left(B_{\hat{Z}}^\epsilon\right)} \mathbb{P}_\theta\left(d(\hat{Z}, Z) \leq \epsilon\right) \approx p_\theta(\hat{Z})$$

and thus for sufficiently small ϵ the quantity (1) provides a good approximation to the likelihood. In general there is a trade off in making ϵ small to obtain a 'good' approximation versus keeping ϵ large to ensure that one can obtain a reasonable estimate of (1) via Monte Carlo.

Although the ABC approximation to the likelihood can be used to replace the likelihood function in any likelihood based parameter estimation procedure it is most commonly used in the context of Bayesian parameter inference and so we shall focus on this one particular application of ABC for the rest of this paper. In Bayesian parameter inference one expresses the information about the parameter vector θ^* contained in the data $\hat{Z}$ in terms of a posterior distribution on the space Θ defined by

$$\pi(\theta) \propto \pi^0(\theta)\, p_\theta(\hat{Z}) \tag{2}$$

where π^0 is some prior distribution representing ones initial knowledge about the parameter vector θ^*. When using the ABC approximation to the likelihood function in the context of Bayesian parameter inference the information about the parameter vector θ^* contained in the data $\hat{Z}$ is expressed by the approximate Bayesian posterior distribution

$$\pi^\epsilon(\theta) \propto \pi^0(\theta)\, p^\epsilon_\theta(\hat{Z}). \tag{3}$$

Hence forth we shall refer to the estimator in (3) as the ABC Bayesian parameter estimator.

Clearly in general the posterior distributions (2) and (3) will differ. There are numerous works on the theoretical behaviour of ABC such as in Biau et al. (2015); Dean et al. (2014); Fearnhead and Prangle (2012). We stress, however, that a lot of work has been done in the context where the observations have an independence structure and the focus of this work is in the intrinsically more challenging hidden Markov model scenario. We follow the approach taken in Dean et al. (2014) in which the asymptotic behaviour of the MLE implemented with the ABC approximation to the likelihood (henceforth ABC MLE) was studied. The analysis in this paper is based on the observation that the ABC approximation to the likelihood can be considered as being equal to the likelihood function of a perturbed probability distribution. Using this observation it was shown that ABC MLE in some sense inherits its behaviour from the standard MLE but that the resulting estimator has an innate asymptotic bias. Furthermore, it is shown that this bias can be made arbitrarily small by choosing a sufficiently small values of the ABC parameter ϵ.

The results in Dean et al. (2014) concerning the asymptotic behaviour of ABC MLE provide a mathematical justification of this method analgous to that provided for the standard MLE by the results concerning asymptotic consistency and normality. The aim of this paper is to develop an equivalent and equally rigorous mathematical justification of the use of ABC in the context of Bayesian parameter estimation. In particular we shall develop an understanding the resulting estimators large data set asymptotic properties. This will then allow us to provide a mathematical justification for the ABC Bayesian parameter estimator analogous to those provided for standard likelihood based estimators by the usual results concerning their asymptotic properties.

We shall do this by establishing Berstein-von Mises type results for Bayesian parameter estimation implemented with ABC approximations to the likelihood. Moreover we shall show that the resulting posterior distributions are asymptotically biased in the sense that they concentrate around a point in parameter space that differs from the true parameter value θ^*. Further we shall derive rates for the size of this asymptotic bias as a function of the ABC parameter ϵ. In the next section we provide an outline of the approach that we shall take to this problem. We remark that some existing work for ABC asymptotics includes: Frazier et al. (2018); Li & Fearnhead (2018).

1.1 Contributions and Structure

In this paper we shall study the asymptotic behaviour of ABC Bayesian parameter estimation when used in parameter estimation for hidden Markov models. This will be convenient as (as we will show) the Markovian context imbues the ABC MLE with a particularly nice mathematical structure. Furthermore, as HMMs are used as statistical models in a wide range of applications including Bioinformatics (e.g. Durbin et al. (1998)), Econometrics (e.g. Kim et al. (1998)) and Population genetics (e.g. Felsenstein and Churchill (1996)) (see also Cappé et al. (2005) for a recent overview), the class of models thus considered is sufficently general to be of genuine practical interest.

For the purpose of this paper a HMM will be considered to be a pair of discrete-time stochastic processes, $\{X_k\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$. The hidden process, $\{X_k\}_{k\geq 0}$, is a homogenous Markov chain taking values in some Polish space $\mathcal{X}$ and the observed process $\{Y_k\}_{k\geq 0}$ takes values in $\mathbb{R}^m$ for some $m \geq 1$. Conditional on X_k the observations Y_k are statistically independent of the random variables $Y_0, \ldots, Y_{k-1}; X_0, \ldots, X_{k-1}$.

In many models the densities of the conditional laws of the observed process w.r.t. the hidden state either have no known analytic expression or else are computationally intractable. In this case it follows that standard methods

to estimating the likelihoods of the observed process, eg. SMC, can no longer be used and that an alternative approach like ABC must be used. For a more detailed discussion of this point see Dean et al. (2014).

For the rest of this paper we shall consider performing ABC Bayesian parameter estimation for HMMs using the following specialization of the standard ABC likelihood approximation (1), proposed in Jasra et al. (2012), for when the observations are generated by a HMM. Specifically, given a sequence of observations $\hat{Y}_1, \ldots, \hat{Y}_n$ from a HMM, we shall approximate the corresponding likelihood functions with the probabilities

$$\mathbb{P}_\theta \left(Y_1 \in B^\epsilon_{\hat{Y}_1}, \ldots, Y_n \in B^\epsilon_{\hat{Y}_n} \right) \tag{4}$$

where for all $y \in \mathbb{R}^m$, B^ϵ_y denotes the ball of radius ϵ centered around the point y. The benefit of this approach is that it retains the Markovian structure of the model. This facilitates both simpler Markov chain Monte Carlo (MCMC) (e.g. McKinley et al. (2009)) and sequential Monte Carlo (SMC) (e.g. Jasra et al. (2012)) implementation of the ABC approximation. Furthermore the resulting approximation has a structure which is particularly tractable to mathematical analysis.

We shall begin by analysing the behaviour of the ABC Bayesian estimator in the case that the conditional laws of the observed state $\{Y_k\}_{k\geq 0}$ are absolutely continuous w.r.t. Lebesgue measure. The analysis will be based upon the observation, made in Dean et al. (2014), that the ABC approximation to the likelihood is equal to the likelihood of a perturbed HMM. Using this observation we shall establish the following results.

Firstly we show that the resulting ABC Bayesian posterior distributions obey a Bernstein-von Mises type theorem. Furthermore we shall show that these posteriors are asymptotically biased in the sense that as the number of data points goes to infinity the resulting posterior distributions concentrate about a point in parameter space that differs from the true parameter value. Finally we show that the size of the asymptotic bias goes to zero as ϵ tends to zero and further that under mild differentiability conditions on the conditional laws of the observations w.r.t. the observation state parameter one can obtain precise rates for the size of the asymptotic bias w.r.t. ϵ.

We then drop the assumption that the conditional laws of the observed state are absolutely continuous w.r.t. Lebesgue measure and show that the posterior distributions again obey a Bernstein-von Mises type theorem and are again asymptotically biased with a bias whose size goes to zero as ϵ tends to zero. Moreover we show that in the general case one can again derive a precise expression for the order of the size of the asymptotic bias w.r.t. ϵ. We also demonstrate, via examples, that in both cases the rates we derive for the size of the asymptotic bias are tight in the sense that they are the best possible rates that may be obtained under such general conditions.

We note that in practice one typically works with a summary statistic of the data set rather than the entire data set, especially when the observations $\{Y_k\}_{k\geq 0}$ take values in some high dimensional space. So far we have implicitly assumed that one is working with the complete data. For ease of exposition we shall persist with this assumption throughout the main part of the paper, leaving discussion of the conditions under which the results we derive will continue to hold when one uses summary statistics to a dedicated section near the end.

Finally we note that the results in this paper can be used to significantly extend those in Dean et al. (2014). Firstly we note that the results in this paper can be used to show that those derived in Dean et al. (2014) hold under far weaker conditions than are assumed in that paper. Secondly the techniques used in this paper can also be used to derive rates for the size of the asymptotic bias of the ABC MLE. Lastly one can use the results established in this paper to show that for sufficiently small ϵ the ABC MLE has an asymptotic normality type property. These points are all discussed in more detail at appropriate points in the text, see in particular Remarks 3.4 and 3.4. We also note that there are practical ways to implement ABC in the context of interest; see Yildirim et al. (2015).

This paper is structured as follows. In Section 2 the notation and assumptions are given. In Section 3 we establish the main results of the paper concerning the asymptotic behaviour of the ABC Bayesian parameter estimator as well as their extension to the case when one works with a summary statistic of the data set. Supporting technical lemmas and proofs of the theoretical results are housed in the appendices.

2 Notation and Assumptions

2.1 Notation and Main Assumptions

Throughout this paper we shall use lower case letters x, y, z to denote dummy variables and upper case letters X, Y, Z to denote random variables. Observations of a random variable, i.e. data, will be denoted by $\hat{Y}$. Given any $\epsilon > 0$ and $y \in \mathbb{R}^m$ we shall let B^ϵ_y denote the closed ball of radius ϵ centered on the point y and let $\mathcal{U}_{B^\epsilon_y}$ denote the uniform distribution on B^ϵ_y. For any $A \subset \mathbb{R}^m$ the indicator function of A will be denoted by $\mathbb{I}_A$.

In what follows we need to refer to various different scalar, vector and matrix norms. Given a scalar z and a vector a we shall let $|z|$ and $|a|$ denote the standard Euclidean scalar and vector norms respectively. For any

$d_1 \times d_2$ two dimensional matrix M we shall let $\|M\| = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |m_{i,j}|$ where for each pair i, j, $m_{i,j}$ denotes the i,j^{th} entry of the matrix M. Similarly, for any $d_1 \times d_2 \times d_3$ three dimensional matrix M we shall let $\|M\| = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} |m_{i,j,k}|$. We note that although using $|\cdot|$ (likewise $\|\cdot\|$) to denote multiple norms is an abuse of notation there is in practice no loss of clarity as the precise meaning of these terms will always be made clear by the context in which they are used.

For any vector of variables a we shall let ∇_a denote the gradient operator with respect to a. Moreover given vectors of variables a, b, c of dimensions d_1, d_2 and d_3 we shall let $\nabla_a \nabla_b$ and $\nabla_a \nabla_b \nabla_c$ denote the $d_1 \times d_2$ and $d_1 \times d_2 \times d_3$ matrices of partial derivatives with entries given by $\frac{\partial^2}{\partial a_i b_j}$ and $\frac{\partial^3}{\partial a_i b_j c_k}$ respectively. Further, for any vector of variables a we shall let ∇_a^2 and ∇_a^3 denote $\nabla_a \nabla_a$ and $\nabla_a \nabla_a \nabla_a$ respectively.

It is assumed that for any HMM the hidden state $\{X_k\}_{k \geq 0}$ is time-homogenous and takes values in a compact Polish space $\mathcal{X}$ with associated Borel σ-field $\mathcal{B}(\mathcal{X})$. Throughout this paper it will be assumed that we have a collection of HMMs all defined on the same state space and parametrised by some parameter vector θ taking values in a *connected* compact set $\Theta \subseteq \mathbb{R}^d$. Furthermore we shall reserve θ^* to denote the 'true' value of the parameter vector θ. For each $\theta \in \Theta$ we shall let $Q_\theta(x, \cdot)$ denote the transition kernel of the corresponding Markov chain and for each $x \in \mathcal{X}$ and $\theta \in \Theta$ we assume that $Q_\theta(x, \cdot)$ has a density $q_\theta(x, \cdot)$ w.r.t. some common finite dominating measure μ on $\mathcal{X}$. The initial distribution of the hidden state will be denoted by π_0, i.e. X_0 has distribution π_0.

We also assume that the observations $\{Y_k\}_{k \geq 0}$ take values in a state space $\mathcal{Y} \subset \mathbb{R}^m$ for some $m \geq 1$. Furthermore, for each k we assume that the random variable Y_k is conditionally independent of $(\ldots, X_{k-1}; X_{k+1}, \ldots)$ and $(\ldots, Y_{k-1}; Y_{k+1}, \ldots)$ given X_k and that the conditional laws have densities $g_\theta(y|x)$ w.r.t. some common σ-finite dominating measure ν. We further assume that for every θ the joint chain $\{X_k, Y_k\}_{k \geq 0}$ is positive Harris recurrent and has a unique invariant distribution π_θ. For each $\theta \in \Theta$ we shall let $\overline{\mathbb{P}}_\theta$ denote the law of stationary distribution of the corresponding HMM and $\overline{E}_\theta$ denote expectations with respect to the stationary distribution $\overline{\mathbb{P}}_\theta$.

We shall frequently have to refer to various kinds of both finite, infinite and doubly infinite sequences. For brevity the following shorthand notations are used. For any pair of integers $k \leq n$, $Y_{k:n}$ denotes the sequence of random variables $Y_k, \ldots, Y_n$; $Y_{-\infty:k}$ denotes the sequence $\ldots, Y_k$; $Y_{n:\infty}$ denotes the sequence $Y_n, \ldots$ and $Y_{-\infty:k;n:\infty}$ denotes the sequence $\ldots, Y_k; Y_n, \ldots$. Further given a measure μ on a Polish space $\mathcal{X}$ we let $\int \cdot \, \mu(dx_{1:n})$ denote integration w.r.t. the n-fold product measure $\mu^{\otimes n}$ on the n-fold product space $\mathcal{X}^n$. Moreover, given a function $f(x_1, \ldots, x_n) : \mathcal{X}^n \to \mathbb{R}$ and integers $1 \leq k \leq l \leq n$, we shall let $\int f(\cdot) \mu(dx_{1:k;l:n})$ denote the partial integrals $\int_{\mathcal{X}^{n-l+k}} f(\cdot) \mu(dx_1) \cdots \mu(dx_k) \mu(dx_l) \cdots \mu(dx_n)$.

Finally, we note that the asymptotic results that we prove for the ABC posterior distribution hold independently of the initial condition of the hidden state process $\{X_k\}_{k \geq 0}$. Thus, in order to keep the presentation as concise as possible we shall suppress the presence of the initial condition of the hidden state except in those instances where it needs to be referred to explicitly. In particular we shall always suppress in our notation the dependence of the likelihood $p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n)$ and ABC approximate likelihood $p_\theta^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n)$ on the initial condition of the hidden state X_0.

2.2 Particular Assumptions

In addition to the assumptions above, the following particular assumptions are made at various points in the article.

(**A1**) For all $y \in \mathcal{Y}$, $x, x' \in \mathcal{X}$, the mappings $\theta \to q_\theta(x, x')$ and $\theta \to g_\theta(y|x)$ are three times continuously differentiable w.r.t. θ.

(**A2**) There exist constants $\underline{c}_1, \overline{c}_1 \in (0, \infty)$ such that for every $y \in \mathcal{Y}$, $x, x' \in \mathcal{X}$, $\theta \in \Theta$

$$\begin{aligned} \underline{c}_1 \leq q_\theta(x, x') \leq \overline{c}_1, \\ g_\theta(y|x) \leq \overline{c}_1. \end{aligned} \tag{5}$$

(**A3**) There exists a constant $\overline{c}_2 \in (0, \infty)$ such that for every $x, x' \in \mathcal{X}$, $\theta \in \Theta$

$$|\nabla_\theta \log q_\theta(x, x')|, |\nabla_\theta^2 \log q_\theta(x, x')| \leq \overline{c}_2.$$

(**A4**) For all $\theta \in \Theta$ and $y \in \mathcal{Y}$

$$0 < \int_{\mathcal{X}} g_\theta(y|x) \, \mu(dx) < \infty. \tag{6}$$

(**A5**) For any $K > 0$

$$\overline{E}_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla_\theta \log g_\theta\left(Y+z|x\right)\right\|^3\right],$$
$$\overline{E}_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla^2_\theta \log g_\theta\left(Y+z|x\right)\right\|^2\right], \tag{7}$$
$$\overline{E}_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla^3_\theta \log g_\theta\left(Y+z|x\right)\right\|\right] \leq \infty.$$

Remark 2.1. *Assumptions (A1)-(A6) are similar to those used in Douc et al. (2004) to prove consistency of the MLE for HMMs. We use similar assumptions in this paper as, broadly speaking, our approach will be to show that the ABC Bayesian paramter estimator inherits its properties from the standard Bayesian parameter estimator* (2)*. However the methods and emphasis of this paper differ from those in Douc et al. (2004) and as a result the assumptions we require have a slightly different flavour. In particular we shall require slightly stronger conditions on the differentiability of the conditional densities $g_\theta(y|x)$ but slightly weaker conditions on their integrability. For more details see Remarks 3.2 and 3.3.*

3 Approximate Bayesian Computation

In this section we present our results on the asymptotic behaviour of the ABC Bayesian parameter estimator when used to perform parameter estimation for HMMs. The key component of the analysis is the observation that the ABC approximate likelihood $p^\epsilon_\theta(\cdots)$ defined in (4) is (up to some suitable rescaling) equal to the likelihood of the data under the perturbed HMM

$$\{X^\epsilon_i, Y^\epsilon_i\}_{i\geq 1} \triangleq \{X_i, Y_i + \epsilon Z_i\}_{i\geq 1} \tag{8}$$

where $\{X_i, Y_i\}$ is equal to the original HMM corresponding to the law $\mathbb{P}_\theta$ and $\{Z_i\}_{i\geq 1}$ is a sequence of i.i.d. random variables uniformly distributed on the unit ball in $\mathbb{R}^m$. Moreover one can show that under the assumptions of Section 2.1 that for any $\epsilon > 0$ the transition kernels and conditional laws of the corresponding perturbed HMM have densities

$$q^\epsilon\left(x, x'\right) \triangleq q\left(x, x'\right) \tag{9}$$

and

$$g^\epsilon\left(y|x\right) \triangleq \frac{\int_{B^\epsilon_y} g\left(z|x\right)\nu(dz)}{\int_{B^\epsilon_y}\nu\left(dz\right)} \tag{10}$$

respectively w.r.t. the dominating measures μ and $\nu * \mathcal{U}_{B^\epsilon_y}$ where $*$ denotes convolution. For more details see Dean et al. (2014).

In the next section we shall study the implications of this probabilistic structure when the ABC Bayesian parameter estimator is used to perform parametric inference for HMMs whose conditional laws are absolutely continuous w.r.t. Lebesgue measure.

3.1 ABC for Bayesian Parameter Inference

Throughout this section we shall assume that the conditional laws of the observed state $\{Y_k\}_{k\geq 0}$ are absolutely continuous w.r.t. Lebesgue measure. Recall that in standard Bayesian parameter inference one expresses the information about the parameter vector θ^* contained in the data $\hat{Y}_1, \ldots, \hat{Y}_n$ in terms of the posterior distribution

$$\pi^n\left(\theta\right) \propto \pi^0\left(\theta\right) p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n) \tag{11}$$

where π^0 is some suitable prior. (Not to be confused with the initial distribution π_0 of X_0.) When performing ABC Bayesian parameter estimation using the ABC approximation defined in (4) the information about the parameter vector θ^* will be expressed by the approximate Bayesian posterior distribution

$$\pi^{\epsilon,n}\left(\theta\right) \propto \pi^0\left(\theta\right) p^\epsilon_\theta(\hat{Y}_1, \ldots, \hat{Y}_n). \tag{12}$$

Clearly in general the posterior distributions (11) and (12) will differ. In order to understand the qualitative differences between the two it is instructive to study the following simple example.

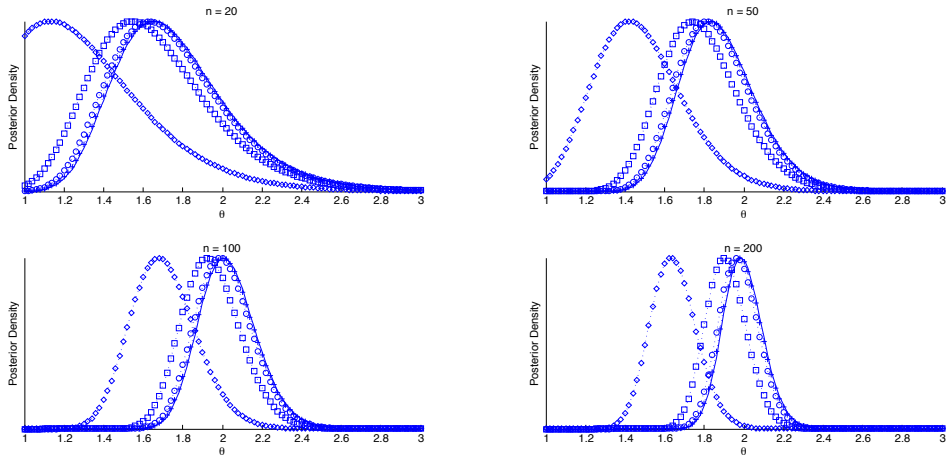

Figure 1: Unnormalized Posterior Densities of Variance Parameter for IID Gaussian Random Variables

Example 3.1. *Consider a sequence of i.i.d. random variables* $\{X_i\}_{i\geq 1}$ *such that for each* i, $X_i \sim N\left(0, \theta^2\right)$. *Suppose that we wish to use the ABC Bayesian posterior* (12) *to infer the value of the variance parameter of the random variables* $\{X_i\}_{i\geq 1}$ *given data* $\hat{X_1}, \ldots, \hat{X_n}$ *and a prior* $\pi_0(\theta) \sim Unif\,(1,3)$ *when the true value of the variance parameter* $\theta^* = 2$. *In this case the ABC approximation to the posterior becomes*

$$\pi^{\epsilon,n}\left(\theta\right) \propto \prod_{i=1}^{n} P_\theta(\hat{X_i} - \epsilon \leq X_i \leq \hat{X_i} + \epsilon).$$

Plots showing unnormalised versions of the resulting posterior distributions for this estimator and for the standard Bayesian parameter estimator given in (11) *for various values of* n *are shown in Figure 1. Each graph has a plot of the standard Bayesian posterior* (11) *(solid line) and of the ABC Bayesian posteriors* (12) *for values of* ϵ *equal to* $0.25, 0.5, 1$ *and* 2 *(dashed lines with crosses, circles squares and diamonds respectively). Since the posterior densities are unnormalised they have all w.l.o.g. been plotted as having equal heights.*

One can see that for both the standard Bayesian and ABC Bayesian parameter estimators the posterior distributions concentrate as the amount of data increases. However as the value of n increases, while the posterior (11) should concentrate around the true value of the variance parameter, in contrast the posteriors (12) concentrate around a quantity which underestimates the true value of the variance by an amount which increases with ϵ.

Intuitively, this can be understood in the following manner. It follows from (8) that for all θ the ABC approximation (4) approximates the likelihood of the data with the likelihood function of a perturbed HMM for which the value of the variance of the observed state is greater than for the corresponding unperturbed HMM. Thus it follows that if the parameter estimator (12) tries to 'match' variances, in the sense that it tends to favour those values of θ for which the variance of the observed state of the corresponding perturbed HMMs (8) matches that of the observed data then it will be systematically biased towards parameter values for which the corresponding unperturbed models underestimate the true value of the variance of the observed state of the HMM which generated the data.

The above example illustrates how performing Bayesian inference using the ABC approximation to the likelihood will lead to a biased estimate of the model parameters. We will now formulate this notion in a more mathematically rigorous manner by comparing the asymptotic behaviour of both the standard Bayesian parameter estimate and is ABC counterpart. We shall start by studying the asymptotic behaviour of the former. Using standard arguments one can show that the standard Bayesian parameter estimator obeys the following Bernstein-Von Mises type theorem (see Borwanker et al. (1971) for more details). The proof is deferred to Appendix B.

Theorem 3.1. *Suppose that one has a collection of HMMs parameterized by some parameter vector* $\theta \in \Theta$ *and that one is given data* $\hat{Y_1}, \ldots, \hat{Y_n}$ *generated by the HMM corresponding to the unknown parameter vector* θ^* *and that one tries to infer the true value of* θ^* *using the exact Bayesian posterior* (11). *Suppose further that the following conditions hold:*

1. *There exists a twice continuously differentiable function* $l\left(\theta\right) : \Theta \to \mathbb{R}$ *such that* $\overline{\mathbb{P}}_{\theta^*}$ *a.s.*

$$\frac{1}{n}\left(\log p_\theta(\hat{Y_1}, \ldots, \hat{Y_n}) - \log p_{\theta^*}(\hat{Y_1}, \ldots, \hat{Y_n})\right) \to l\left(\theta\right) \tag{13}$$

uniformly in θ.

2. *The function* $l(\theta)$ *has a unique maximum at* $\theta = \theta^*$.

Then the standard MLE, henceforth denoted $\hat{\theta}_n$, *is asymptotically consistent. Moreover suppose that the following extra conditions are satisfied:*

3. $\overline{\mathbb{P}}_{\theta^*}$ *a.s.*

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{|\theta - \theta^*| \leq \delta} \left| \nabla_\theta^2 \frac{1}{n} \log p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n) - \nabla_\theta^2 l(\theta^*) \right| = 0. \tag{14}$$

4.

$$\nabla_\theta^2 l(\theta^*) < 0. \tag{15}$$

Then if the parameter vector θ^* *belongs to the interior of* Θ *and the prior distribution* π^0 *is absolutely continuous w.r.t. Lebesgue measure and has a continuous density on* Θ *which is non zero in a neighbourhood of the true parameter value* θ^* *one has for all initial conditions* π_0 *that*

$$\lim \left\| \pi^n - N\left(\hat{\theta}_n, -\frac{\nabla_\theta^2 l(\theta^*)^{-1}}{n} \right) \right\|_{TV} = 0$$

$\overline{\mathbb{P}}$ *a.s.*.

Theorem 3.1 immediately implies the following more standard Bernstein-von Mises result.

Corollary 3.1. *Suppose that the conditions of Theorem 3.1 hold. For any n let* $\tilde{\pi}^n$ *be the (random) probability law on the space* $\mathbb{R}^m$ *with density proportional to* $\pi^n(\frac{\theta}{\sqrt{n}} + \hat{\theta}_n)$. *Then one has for all initial conditions* π_0 *that* $\overline{\mathbb{P}}$ *a.s. the sequence of random laws* $\tilde{\pi}^n$ *converge in total variation to* $N(0, -\nabla_\theta^2 l(\theta^*)^{-1})$.

One can make several observations about the assumptions of Theorem 3.1 as well as its consquences.

Remark 3.1. *The quantity* $\nabla_\theta^2 l(\theta^*)$ *is equal to the asymptotic Fisher information matrix* $I(\theta^*)$ *of the corresponding collection of HMMs.*

Remark 3.2. *We now give sufficient conditions for 1-4 of Theorem 3.1. Lemma B.9 in Appendix B proves conditions 1 and 3 hold if the underlying collection of HMMs satisfy assumptions (A1)-(A5). Furthermore it follows from Douc et al. (2004) that given assumptions (A1)-(A2) we have that condition 2 in Theorem 3.1 holds if the collection of HMMs obey the following additional assumption:*

(**A6**) $\theta = \theta^*$ *if and only if* $\overline{\mathbb{P}}_\theta^Y = \overline{\mathbb{P}}_{\theta^*}^Y$.

For a discussion of when condition 4 of Theorem 3.1 holds see **?**.

Remark 3.3. *Conditions 1-3 of Theorem 3.1 are sufficient to guarantee that the corresponding maximum likelihood estimator (MLE) is asymptotically consistent. As Lemma B.9 establishes these conditions, it thus provides a different approach to studying the asymptotic behaviour of the MLE based on analysing the asymptotic behaviour of the corresponding likelihood surface. This contrasts with the standard approach to studying the asymptotic behaviour of the MLE which involves showing that the mean log likelihood functions converge to the relative entropies of the collection of HMMs w.r.t. the true HMM in a sufficiently uniform manner, see Douc et al. (2004) for more details. However, in the context of ABC, the limits of the ABC approximations to the likelihood functions can no longer be interpreted as relative entropies but can still be understood as defining a suitable limiting approximate likelihood surface, see Remark 3.4 for more details. We note that compared to the approach taken in Douc et al. (2004) the likelihood surface approach of Lemma B.9 requires slightly more stringent conditions on the differentiability of the conditional densities of the HMM but slightly less stringent conditions on their integrability.*

In order to compare the performances of the two estimators (11) and (12) we next derive a Bernstein-von Mises type result for the ABC Bayesian posterior. The key step is to show that under suitable assumptions the ABC approximate likelihood surface will satisfy conditions analogous to those for the exact likelihood surface given by 1-4 in Theorem 3.1. This is the content of the following theorem whose proof is again deferred to Appendix B.

Theorem 3.2. *Suppose that one has a collection of HMMs parameterized by some parameter vector $\theta \in \Theta$ and that one is given data $\hat{Y}_1, \ldots, \hat{Y}_n$ generated by the HMM corresponding to an unknown parameter vector θ^* and that one tries to infer the true value of θ^* using the ABC approximate Bayesian posterior* (12)*. Suppose further that the collection of HMMs satisfies assumptions (A1)-(A5) and that the mean relative log likelihood functions* (13) *satisfy conditions 1-4 of Theorem 3.1. Then for every $\epsilon > 0$ there exists a twice continuously differentiable function $l^\epsilon(\theta) : \Theta \to \mathbb{R}$ such that for all initial conditions π_0 one has that $\bar{\mathbb{P}}_{\theta^*}$ a.s.*

$$\frac{1}{n}\left(\log p_\theta^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n) - \log p_{\theta^*}^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n)\right) \to l^\epsilon(\theta) \tag{16}$$

uniformly in θ. For sufficiently small values of ϵ the function $l^\epsilon(\theta)$ will have a unique maximum at some point $\theta^{*,\epsilon}$ such that $\nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon}) < 0$ and

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{|\theta - \theta^{*,\epsilon}| \leq \delta} \left| \nabla_\theta^2 \frac{1}{n} \log p_\theta^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n) - \nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon}) \right| = 0. \tag{17}$$

Furthermore $\theta^{*,\epsilon} \to \theta^*$ and $\nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon}) \to \nabla_\theta^2 l(\theta^*)$ as $\epsilon \to 0$.

Remark 3.4. *In this paper we have chosen to restrict our attention to the situation where the ABC approximation to the likelihood is used in the context of Bayesian parameter estimation. However as already noted above the ABC approximation can be used within various different approaches to parameter estimation. In particular it can be used in the context of maximum likelihood estimation. Doing so results in the following ABC maximum likelihood estimator (ABC MLE)*

$$\hat{\theta}_{n,\epsilon} \triangleq \arg\max_{\theta \in \Theta} \left\{ p_\theta^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n) \right\}. \tag{18}$$

The properties of the resulting estimator were extensively analysed in Dean et al. (2014) where it was shown that

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \left| \hat{\theta}_{n,\epsilon} - \theta^* \right| = 0. \tag{19}$$

In Dean et al. (2014) the asymptotic result (19) *was proved by studying the limits of the ABC approximate log likelihoods $p_\theta^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n)$. In standard MLE the limits of the mean log likelihood functions are analysed by considering the relative entropies of various members of the parameteric family of distributions under consideration. Since the limits of the ABC approximate mean log likelihood functions have no obvious interpretations in terms of relative entropies very strong assumptions concerning the behaviour of the underlying HMMs had to be made to ensure that these limits existed and were suitably well behaved. In contrast since it is clear that Theorem 3.2 immediately implies* (19) *it follows that one can obtain this result under much weaker conditions by instead considering the limiting behaviour of the ABC approximate likelihood surface. Thus Theorem 3.2 provides a significant relaxation of the results in Dean et al. (2014).*

Furthermore Theorem 3.2 can be used to extend the results in that paper in two significant ways. Firstly it follows from Theorem 3.2 that for small enough values of ϵ the ABC MLE $\hat{\theta}_{n,\epsilon}$ will $\bar{\mathbb{P}}_{\theta^*}$ a.s. converge to the unique (deterministic) limit point $\theta^{*,\epsilon}$ as $n \to \infty$. This is a much tighter result than (19) *which, for any value of $\epsilon > 0$, allows for the ABC MLE to have more than one (and possibly infinitely many) accumulation points as $n \to \infty$. Secondly the results in Dean et al. (2014) provide no insight into the asymptotic distribution of the ABC MLE $\hat{\theta}_{n,\epsilon}$ about its limit point $\theta^{*,\epsilon}$. It contrast, given Theorem 3.2, one can use standard arguments to show that under assumptions (A1)-(A5) one has that for sufficiently small ϵ that*

$$\sqrt{n}\nabla_\theta \left(\log p_{\theta^{*,\epsilon}}^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n) - \log p_{\theta^*}^\epsilon(\hat{Y}_1, \ldots, \hat{Y}_n)\right) \to N(0, \nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon}))$$

and hence, using equation (17) *in Theorem 3.2, that*

$$\sqrt{n}\left(\hat{\theta}_{n,\epsilon} - \theta^{*,\epsilon}\right) \to N(0, \nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon})^{-1}). \tag{20}$$

For more details see Douc et al. (2004).

Given Theorem 3.2 we can easily derive the following theorem which provides an analogous result to that of Theorem 3.1 for the ABC Bayesian estimator. We again defer the proof until Appendix B.

Theorem 3.3. *Suppose that one has a collection of HMMs parameterized by some parameter vector $\theta \in \Theta$ and that one is given data $\hat{Y}_1, \ldots, \hat{Y}_n$ generated by the HMM corresponding to an unknown parameter vector θ^* and that one tries to infer the true value of θ^* using the ABC approximate Bayesian posterior* (12)*. Suppose further that the collection of HMMs satisfy assumptions (A1)-(A5) and that the mean relative log likelihood functions* (13) *satisfy conditions 1-4 of Theorem 3.1. Then if the parameter vector θ^* belongs to the interior of Θ one has that for sufficiently small values of ϵ that for all initial conditions π_0*

$$\lim \left\| \pi^{\epsilon,n} - N\left(\hat{\theta}_{n,\epsilon}, \frac{\nabla_\theta^2 l^\epsilon \left(\theta^{*,\epsilon}\right)^{-1}}{\sqrt{n}} \right) \right\|_{TV} = 0$$

and

$$\lim_{n\to\infty} \hat{\theta}_{n,\epsilon} = \theta^{*,\epsilon}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. where $l^\epsilon(\theta)$ and $\theta^{*,\epsilon}$ are as in Theorem 3.2 and where the for each n the random variable $\hat{\theta}_{n,\epsilon}$ is equal to the ABC MLE defined in (18)*. The convergence of the renormalised posterior is again with respect to the total variation norm.*

Theorem 3.3 immediately implies the following Bernstein-von Mises result for the ABC Bayesian parameter estimator.

Corollary 3.2. *Suppose that the conditions of Theorem 3.3 hold. For any n let $\tilde{\pi}_{\epsilon,n}$ be the (random) probability law on the space $\mathbb{R}^m$ with density proportional to $\pi_{\epsilon,n}(\frac{\theta-\hat{\theta}_{n,\epsilon}}{\sqrt{n}})$ if $\frac{1}{\sqrt{n}}(\theta-\hat{\theta}_{n,\epsilon}) \in \Theta$ and equal to zero otherwise. Then one has for all initial conditions π_0 that $\overline{\mathbb{P}}$ a.s. the sequence of random laws $\tilde{\pi}_{\epsilon,n}$ converge weakly to a $N(0, \nabla_\theta^2 l^\epsilon \left(\theta^{*,\epsilon}\right)^{-1})$ random variable.*

Remark 3.5. *Given the interpretation in* (8) *of the ABC approximation to the likelihood as being the likelihood of the data under a perturbed HMM it follows from that Theorem 3.3 that a Bernstein-Von Mises type theorem still holds when one performs parameter estimation using misspecified models in the sense of White (1982).*

Theorems 3.2 and 3.3 show that asymptotically the true Bayesian and ABC Bayesian posteriors are exponentially concentrated around the points θ^* and $\theta^{*,\epsilon}$ respectively and thus that the difference between them will asymptotically be of the same order as $|\theta^{*,\epsilon} - \theta^*|$ (with respect to some suitable metric that respects the topology of weak convergence - for example the Prokhorov or Lipschitz-dual norms) as $n \to \infty$. Furthermore these results show that $|\theta^{*,\epsilon} - \theta^*| \to 0$ as $\epsilon \to 0$.

It is natural to ask at what rate does $\theta^{*,\epsilon} \to \theta^*$ as $\epsilon \to 0$. We begin our answer to this question by revisiting Example 3.1.

Example 3.2. *We return again to the model in Example 3.1 and consider for each ϵ the quantity $\theta^{*,\epsilon}$ as defined in Theorem 3.2. Recall that is both the point around which the ABC Bayesian posterior concentrates and the limit point for the corresponding ABC MLE. In figure 2 we give plots of both $\theta^{*,\epsilon}$ as a function of ϵ (crosses) and of the corresponding best fit quadratic curve (solid line).*

Figure 2 suggests that for ϵ sufficiently small the size of the asymptotic bias of the ABC Bayesian (and ABC MLE) estimator should be of order ϵ^2. The next theorem shows that under the following extra mild assumptions on the differentiability of the conditional likelihood functions $g_\theta(y|x)$ w.r.t. the observed state variable y this is indeed the case.

(**A7**) The functions $g_\theta(y|x)$ and $\nabla_\theta g_\theta(y|x)$ are twice continuously differentible w.r.t. the variable y and for all $K > 0$

$$\begin{aligned} &E_{\theta^*}\left[\sup_{x\in\mathcal{X}} \sup_{\theta\in\Theta} \sup_{z\in B_K^0} \left| \frac{\nabla_y^2 g_\theta(Y+z|x)}{g_\theta(Y|x)} \right|^2 \right], \\ &E_{\theta^*}\left[\sup_{x\in\mathcal{X}} \sup_{\theta\in\Theta} \sup_{z\in B_K^0} \left| \frac{\nabla_y^2 \left(\nabla_\theta g_\theta(Y+z|x)\right)}{g_\theta(Y|x)} \right|^2 \right] < \infty. \end{aligned} \tag{21}$$

Theorem 3.4. *Suppose that in addition to all of the assumptions of Theorem 3.3 one has that assumption (A7) above also holds. Then there exists a vector $\Delta\theta^*$ such that for ϵ sufficiently small*

$$\theta^{\epsilon,*} - \theta^* = \epsilon^2 \Delta\theta^* + o(\epsilon^2). \tag{22}$$

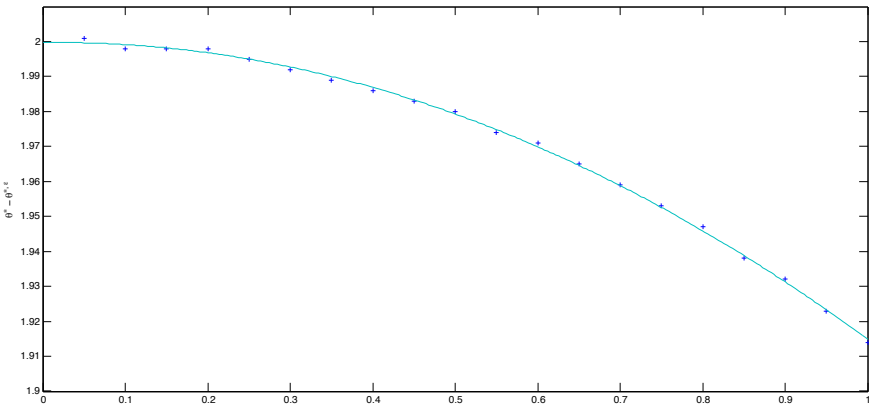

Figure 2: Plot of $\theta^{*,\epsilon}$ as a function of ϵ for IID Gaussian Random Variables

Remark 3.6. *It follows from the proof of Theorem 3.4 that one has the following exact expression for the asymptotic bias term $\Delta\theta^*$:*

$$\Delta\theta^* = -\nabla_\theta^2 l(\theta^*)^{-1}\Bigg(\sum_{i=-\infty}^{\infty} E_{\theta^*}\left[E_{\pi_{\theta,i}}\left[\nabla_\theta(\log g_\theta(Y_i|X_i)q_\theta(X_{i-1},X_i))\right]\right]$$
$$+E_{\theta^*}\left[E_\theta\left[\frac{\nabla_\theta g_\theta(Y_0|X_0)}{g_\theta(Y_0|X_0)}\frac{\nabla_y^2 g_\theta(Y_0|X_0)}{g_\theta(Y_0|X_0)} - \frac{\nabla_y^2\nabla_\theta g_\theta(Y_0|X_0)}{g_\theta(Y_0|X_0)}|Y_{-\infty:\infty}\right]\right]\Bigg)$$

where the random signed measures $\pi_{\theta,i}$ are as in Lemma C.15. See Appendix C for more details.

Remark 3.7. *Theorem 3.4 provides a considerable improvement on the rates obtained in Dean et al. (2014) for the size of the asymptotic bias of the ABC MLE w.r.t. ϵ. The rates obtained in that paper are much less tight than the ones obtained above and were derived under much more restrictive conditions.*

3.2 ABC for Bayesian Parameter Inference - the General Case

In the previous section we studied the behaviour of the ABC Bayesian parameter estimator for HMMs in the special case that the conditional laws of the HMMs are absolutely continuous w.r.t. Lebesgue measure. In particular we showed that the resulting estimators obey a Bernstein-von Mises type result with an asymptotic bias whose size goes to zero as ϵ goes to zero. Furthermore we showed that under mild differentiability conditions the size of the asymptotic bias is of order ϵ^2 for sufficiently small ϵ.

In this section we shall show that analogous results still hold for the more general case where one drops the assumption that the conditional laws of the HMMs are absolutely continuous w.r.t. Lebesgue measure. We start by considering Theorems 3.2 and 3.3. A careful reading of the proofs of these theorems shows that they are independent of the assumption that the conditional laws of the HMMs are absolutely continuous w.r.t. Lebesgue measure and thus that they still hold in the general case. Hence the only thing that remains is to understand the rate at which the asymptotic bias of the ABC Bayesian (and ABC MLE) estimator goes to zero as ϵ goes to zero. In order to do this it is instructive to consider the following example.

Example 3.3. *Let π^1 be the distribution on the set of diadic numbers of the form $\frac{1}{2^{2k}}; k = 0, 1, \ldots$ given by $\pi^1(\frac{1}{2^{2k}}) = \frac{1}{2^{k+1}}$ for all k and let π^2 be the distribution on the set of diadic numbers of the form $\frac{1}{2^{2k+1}}$ given by $\pi^2(\frac{1}{2^{2k+1}}) = \frac{1}{2^{k+1}}$ for all $k = 0, 1, \ldots$. Furthermore let $\{\pi_\alpha\}_{\alpha\in[0,1]}$ be the set of distributions defined such that for all α, $\pi_\alpha = \alpha\pi^1 + (1-\alpha)\pi^2$.*

It is clear that the distributions π_α satisfy the conditions of Theorem 3.2 and hence that for any ϵ the limiting approximate mean log likelihood surface $l^\epsilon(\alpha)$ exist and is well defined. Further if we assume that the true value of the parameter is equal to $\alpha^* = \frac{1}{2}$ then it is easy to show that $\nabla_\alpha^2 l(\alpha^*) \neq 0$ and that for every $k \geq 2$

$$\nabla_\alpha l^{\frac{1}{2^{2k}}}(\alpha^*) = \frac{1}{8}\frac{1}{2^{2k}},\ \nabla_\alpha l^{\frac{1}{2^{2k+1}}}(\alpha^*) = -\frac{1}{4}\frac{1}{2^{2k+1}}$$

from which it follows that for all k

$$\alpha^{*,\frac{1}{2^{2k}}} - \alpha^* = -\frac{1}{\nabla_\alpha^2 l(\alpha^*)} \frac{1}{8} \frac{1}{2^{2k}} + o(\frac{1}{2^{2k}}),$$
$$\alpha^{*,\frac{1}{2^{2k+1}}} - \alpha^* = \frac{1}{\nabla_\alpha^2 l(\alpha^*)} \frac{1}{4} \frac{1}{2^{2k+1}} + o(\frac{1}{2^{2k+1}}).$$

The above example shows that in the general case one should expect that the size of the asymptotic bias will be at least $O(\epsilon)$ and that the limit $\frac{\theta^{*,\epsilon}-\theta^*}{\epsilon}$ will not be well defined. The next theorem shows that in general the behaviour of the asymptotic bias will be no worse than this. In order for it to hold we need to make the following differentiability assumptions.

(**A8**) The functions $g_\theta(y|x)$ and $\nabla_\theta g_\theta(y|x)$ are twice continuously differential w.r.t. the variable y and for all $K > 0$

$$E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B_K^0}\left|\frac{\nabla_y g_\theta(Y+z|x)}{g_\theta(Y|x)}\right|^2\right],$$
$$E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B_K^0}\left|\frac{\nabla_y\left(\nabla_\theta g_\theta(Y+z|x)\right)}{g_\theta(Y|x)}\right|^2\right] < \infty. \tag{23}$$

Theorem 3.5. *Suppose that in addition to all of the assumptions of Theorem 3.3 one has that assumption (A8) above also holds. Then*

$$\theta^{\epsilon,*} - \theta^* = O(\epsilon). \tag{24}$$

The proof of Theorem 3.5 is deferred to Appendix D.

3.3 Sufficient Statistics

So far we have assumed that one is working with the complete data sequence $\hat{Y}_1, \ldots, \hat{Y}_n$. In practice the full data set is often too high-dimensional and instead one performs inference using a summary statistic $\mathcal{S}(\hat{Y}_1, \ldots, \hat{Y}_n)$ where $\mathcal{S}(\cdots)$ is some mapping from $\mathbb{R}^m \times \cdots \mathbb{R}^m$ to a lower dimensional Euclidean space, see for example Tavre et al. (1997).

In general this mapping will destroy the Markovian structure of the data and the results so far derived will not be applicable to ABC based parameter inference conducted using the corresponding summary statistic. However in practice it is often the case that the mapping $\mathcal{S}(\cdots)$ is of the form $\mathcal{S}(\hat{Y}_1, \ldots, \hat{Y}_n) = S(\hat{Y}_1), \ldots, S(\hat{Y}_n)$ for some function $S(\cdot)$ that maps from $\mathbb{R}^m$ to a space $\mathbb{R}^{m'}$ of lower dimension. When this is true it is easy to see that the Markovian structure of the data is preserved. Moreover suppose that assumptions (A1)-(A5) as well as any of (A7) and (A8) as appropriate hold for the underlying HMM. It is easy to see that these assumptions will be preserved by any 'reasonably smooth' mapping $S(\cdot)$. (For the sake of brevity we shall not provide a rigorous formulation of this statement.) Hence it follows that if the mapping $S(\cdot)$ preserves the identifiability of the system, that is to say if assumption (A6) also holds for the HMMs with observations $S(Y_1), S(Y_2), \ldots$, then all the relevant results of this section will continue to hold for the ABC Bayesian parameter estimator implemented with the sufficient statistic $S(\cdot)$.

Appendix A: Auxillary Results

In this section we present some results that will be needed in the proofs of Theorems 3.1 , 3.2 , 3.3 and 3.4. The first two lemmas are standard result from real analysis which we state without proof.

Lemma A.1. *Let a compact set $G \subset \mathbb{R}^u$ be given and a sequence of continuously differentiable functions $f_n : G \to \mathbb{R}^v$, $n \geq 1$, such that the sequence $\nabla f_n(z)$ is Cauchy uniformly in z. Let the function $g(z)$ be the limit of $\nabla f_n(z)$.*

Assume also that $f_n(z^*)$ is Cauchy for some $z^* \in G$. Then there exists a continuously differentiable function f such that $f_n(z) \to f(z)$ uniformly in z and $\nabla f(z) = g(z)$.

Lemma A.2. *Let a compact set $G \subset \mathbb{R}^u$ and some constant $L > 0$ be given. Suppose that there exists a continuous function $f : G \to \mathbb{R}^v$ and sequence of continuous functions $f_n : G \to \mathbb{R}^v$, $n \geq 1$, such that for all n the function f_n is L Lipschitz continuous. Then $f_n \to f$ uniformly in G if and only if $f_n \to f$ pointwise on a countable dense subset of G.*

The next two Lemmas are standard results from the theory of uniformly ergodic Markov chains, see for example Del Moral (2004); Dean et al. (2014).

Lemma A.3. *Suppose that the transition kernel and conditional likelihoods of some HMM $\{X_i, Y_i\}_{i\geq 1}$ satisfy assumption (A2). Then there exists some $0 < \rho < 1$ such that for all $f \in L_\infty$, for all constants $a \vee b < r$ and $s < l \wedge m$ and for all $x, x' \in \mathcal{X}$ one has that*

$$\begin{aligned}\Big|E\left[f\left(X_r,\ldots,X_s\right)|X_a = x; Y_k,\ldots,Y_l\right] - E\left[f\left(X_r,\ldots,X_s\right)|X_b = x'; Y_k,\ldots,Y_m\right]\Big| \\ \leq 4\rho^{(l-s)\wedge(m-s)\wedge(r-a)\wedge(r-b))}\left\|f\right\|_\infty. \end{aligned} \tag{A-25}$$

Moreover

$$E\left[f\left(X_0,\ldots,X_r\right)|\ldots,Y_{-1},Y_0,Y_1,\ldots\right]$$

$(r \geq 0)$ is well defined for any doubly infinite sequences $\ldots,Y_{-1},Y_0,Y_1,\ldots$ and $f \in L_\infty$. For $l,k \geq 0$,

$$\begin{aligned}\sup_x \Big|E\left[f\left(X_0,\ldots,X_r\right)|X_{-l} = x; Y_{-l},\ldots,Y_{r+k}\right] \\ - E\left[f\left(X_0,\ldots,X_r\right)|\ldots,Y_{-1},Y_0,Y_1,\ldots\right]\Big| \leq 4\rho^{l\wedge k}\left\|f\right\|_\infty \end{aligned} \tag{A-26}$$

and constant ρ depends only on the quantities $\underline{c}$ and $\overline{c}$ appearing in (5).

Lemma A.4. *Suppose that the transition kernel and conditional likelihoods of some HMM $\{X_i, Y_i\}_{i\geq 1}$ satisfy assumption (A2). Then for any $f, g \in L_\infty$, all $l,k \geq 0$, all sequences $Y_{-l},\ldots,Y_k$ and $-l \leq r \leq s \leq k$ one has that*

$$\begin{aligned}\Big|E\left[f\left(X_r\right)g\left(X_s\right)|Y_{-l},\ldots,Y_k\right] - E\left[f\left(X_r\right)|Y_{-l},\ldots,Y_k\right]E\left[g\left(X_s\right)|Y_{-l},\ldots,Y_k\right]\Big| \\ \leq 4\rho^{s-r}\left\|f\right\|_\infty\left\|g\right\|_\infty \end{aligned} \tag{A-27}$$

where ρ is as in Lemma A.3. Note that the value of ρ again depends only on the quantities $\underline{c}$ and $\overline{c}$ appearing in (5).

The fifth lemma is essentially a corollary and extension of Propositions 4 and 5 in Douc et al. (2004).

Lemma A.5. *Suppose that one has two collections of HMMs both defined on the same state spaces and parameterised by vectors $\theta \in \Theta$ and $\hat{\theta} \in \hat{\Theta}$ respectively. Suppose further that both collections of HMMs satisfy assumption (A2) with the same values of $\underline{c}$ and $\overline{c}$. Finally suppose that some parameter vector $\theta^* \in \Theta$ is given and let $\{X_i, Y_i\}_{i\geq 1}$ denote the corresponding stochastic process.*

Given measurable functions $\phi_1, \phi_2, \phi_3 : \hat{\Theta} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $y \in \mathcal{Y}$, $k < l$ and $s \in \{1,2,3\}$ let

$$\left\|\phi_s\right\|_\infty(y) \triangleq \sup_{\hat{\theta}\in\hat{\Theta}}\sup_x\left|\phi_s\left(\theta,x,y\right)\right|, \; \phi_{s;k:l}(\theta) \triangleq \sum_{i=k+1}^{l}\phi_s\left(\theta,X_i,Y_i\right).$$

Further for any $n > 0$ define the random variables $\Delta_{0,n}$, $\Gamma_{0,n}$, $\Psi_{0,n}$ and $\Omega_{0,n}$ by

$$\Delta_{0,n}(\hat{\theta}) \triangleq E_{\hat{\theta}}\Big[\phi_{1;-n:0}(\theta)\Big|Y_{-n:0}\Big] - E_{\hat{\theta}}\Big[\phi_{1;-n:-1}(\theta)\Big|Y_{-n:-1}\Big],$$

$$\begin{aligned}\Gamma_{0,n}(\hat{\theta}) \triangleq{}& E_{\hat{\theta}}\Big[\phi_{1;-n:0}(\theta)\phi_{2;-n:0}(\theta)\Big|Y_{0:-n}\Big] - E_{\hat{\theta}}\Big[\phi_{1;-n:-1}(\theta)\phi_{2;-n:-1}(\theta)\Big|Y_{-n:-1}\Big] \\ &+ E_{\hat{\theta}}\Big[\phi_{1;-n:-1}(\theta)\Big|Y_{-n:-1}\Big]E_{\hat{\theta}}\Big[\phi_{2;-n:-1}(\theta)\Big|Y_{-n:-1}\Big] \\ &- E_{\hat{\theta}}\Big[\phi_{1;-n:0}(\theta)\Big|Y_{-n:-0}\Big]E_{\hat{\theta}}\Big[\phi_{2;-n:0}(\theta)\Big|Y_{-n:-0}\Big],\end{aligned}$$

$$\begin{aligned}\Psi_{0,n}(\hat{\theta}) \triangleq{}& E_{\hat{\theta}}\Big[\phi_{1;-n:0}(\theta)\phi_{2;-n:0}(\theta)\Big|Y_{0:-n}\Big]E_{\hat{\theta}}\Big[\phi_{3;-n:0}(\theta)\Big|Y_{-n:-0}\Big] \\ &- E_{\hat{\theta}}\Big[\phi_{1;-n:0}(\theta)\Big|Y_{-n:-0}\Big]E_{\hat{\theta}}\Big[\phi_{2;-n:0}(\theta)\Big|Y_{-n:-0}\Big]E_{\hat{\theta}}\Big[\phi_{3;-n:0}(\theta)\Big|Y_{-n:-0}\Big] \\ &+ E_{\hat{\theta}}\Big[\phi_{1;-n:-1}(\theta)\Big|Y_{-n:-1}\Big]E_{\hat{\theta}}\Big[\phi_{2;-n:-1}(\theta)\Big|Y_{-n:-1}\Big]E_{\hat{\theta}}\Big[\phi_{3;-n:-1}(\theta)\Big|Y_{-n:-1}\Big] \\ &- E_{\hat{\theta}}\Big[\phi_{1;-n:-1}(\theta)\phi_{2;-n:-1}(\theta)\Big|Y_{-n:-1}\Big]E_{\hat{\theta}}\Big[\phi_{3;-n:-1}(\theta)\Big|Y_{-n:-1}\Big],\end{aligned}$$

and

$$\begin{aligned}\Omega_{0,n}(\hat{\theta}) \triangleq\ & E_{\hat{\theta}}\left[\phi_{1;-n:0}(\theta)\phi_{2;-n:0}(\theta)\phi_{3;-n:0}(\theta)\middle|Y_{-n:-0}\right] \\ & - E_{\hat{\theta}}\left[\phi_{1;-n:0}(\theta)\middle|Y_{-n:-0}\right]E_{\hat{\theta}}\left[\phi_{2;-n:0}(\theta)\middle|Y_{-n:-0}\right]E_{\hat{\theta}}\left[\phi_{3;-n:0}(\theta)\middle|Y_{-n:-0}\right] \\ & + E_{\hat{\theta}}\left[\phi_{1;-n:-1}(\theta)\middle|Y_{-n:-1}\right]E_{\hat{\theta}}\left[\phi_{2;-n:-1}(\theta)\middle|Y_{-n:-1}\right]E_{\hat{\theta}}\left[\phi_{3;-n:-1}(\theta)\middle|Y_{-n:-1}\right] \\ & - E_{\hat{\theta}}\left[\phi_{1;-n:-1}(\theta)\phi_{2;-n:-1}(\theta)\phi_{3;-n:-1}(\theta)\middle|Y_{-n:-1}\right].\end{aligned}$$

Then

(i) if $\|\phi_1\|_\infty \in L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$ *there exists an integrable random variable* $\Delta_{0,\infty}$ *such that*

$$\Delta_{0,n}(\theta) \to \Delta_{0,\infty}(\theta) \tag{A-28}$$

in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$.

(ii) if $\|\phi_1\|_\infty, \|\phi_2\|_\infty \in L^2\left(\overline{\mathbb{P}}_{\theta^*}\right)$ *then there exists an integrable random variable* $\Gamma_{0,\infty}$ *such that*

$$\Gamma_{0,n}(\theta) \to \Gamma_{0,\infty}(\theta) \tag{A-29}$$

in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$.

(iii) if $\|\phi_1\|_\infty, \|\phi_2\|_\infty, \|\phi_3\|_\infty \in L^3\left(\overline{\mathbb{P}}_{\theta^*}\right)$ *then there exist integrable random variables* $\Psi_{0,\infty}$ *and* $\Omega_{0,\infty}$ *such that*

$$\Psi_{0,n}(\theta) \to \Psi_{0,\infty}(\theta),\ \Omega_{0,n}(\theta) \to \Omega_{0,\infty}(\theta) \tag{A-30}$$

in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$.

Moreover there exist constants $C < \infty$ *and* $0 < \rho < 1$ *which depend only on* $\underline{c}$ *and* $\overline{c}$ *such that for all* $n \geq m > 0$

$$\left.\begin{array}{l} E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}}\left|\Delta_{0,n}(\hat{\theta}) - \Delta_{0,m}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}}\left|\Gamma_{0,n}(\hat{\theta}) - \Gamma_{0,m}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}}\left|\Psi_{0,n}(\hat{\theta}) - \Psi_{0,m}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}}\left|\Omega_{0,n}(\hat{\theta}) - \Omega_{0,m}(\hat{\theta})\right|\right] \end{array}\right\} \leq C\rho^m \left\{\begin{array}{c} E_{\theta^*}\left[\|\phi_1\|_\infty\right] \\ \sup_{s\in\{1,2\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \\ \sup_{s\in\{1,2,3\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \\ \sup_{s\in\{1,2,3\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \end{array}\right. \tag{A-31}$$

and

$$\left.\begin{array}{l} E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta},n>0}\left|\Delta_{0,n}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta},n>0}\left|\Gamma_{0,n}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta},n>0}\left|\Psi_{0,n}(\hat{\theta})\right|\right] \\ E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta},n>0}\left|\Omega_{0,n}(\hat{\theta})\right|\right] \end{array}\right\} \leq C \left\{\begin{array}{c} E_{\theta^*}\left[\|\phi_1\|_\infty\right] \\ \sup_{s\in\{1,2\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \\ \sup_{s\in\{1,2,3\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \\ \sup_{s\in\{1,2,3\}} E_{\theta^*}\left[\|\phi_s\|_\infty\right] \end{array}\right. . \tag{A-32}$$

The proof of this lemma follows very closely the proofs of Propositions 4 and 5 in Douc et al. (2004) hence we restrict ourselves to giving only the essential details.

Proof. We shall just provide proofs for the results concerning the quantities $\Omega_{0,n}(\hat{\theta})$. The proofs of the other results follow in an identical fashion.

We begin by proving (A-31). For any $n \geq m > 0$ we have that

$$\begin{aligned}\sup_{\hat{\theta}\in\hat{\Theta}}\left|\Omega_{0,n}(\hat{\theta}) - \Omega_{0,m}(\hat{\theta})\right| \leq\ & \alpha + \beta_{1,2,3} + \beta_{2,3,1} + \beta_{3,1,2} + \gamma_{1,2,3} + \gamma_{2,3,1} + \gamma_{3,1,2} + \delta \\ & + \psi_{1,2,3} + \psi_{2,3,1} + \psi_{3,1,2} + \kappa_{1,2,3} + \kappa_{2,3,1} + \kappa_{3,1,2} + \omega_{1,2,3} + \omega_{2,3,1} + \omega_{3,1,2}\end{aligned}$$

where

$$\begin{aligned}
\alpha = \sum_{i,j,k=-m+1}^{-1} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| & E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\phi_2(\theta, X_j, Y_j)\phi_3(\theta, X_k, Y_k)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_j, Y_j)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_k, Y_k)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\phi_2(\theta, X_j, Y_j)\phi_3(\theta, X_k, Y_k)\big|Y_{-n:-1}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\big|Y_{-n:-1}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_j, Y_j)\big|Y_{-n:-1}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_k, Y_k)\big|Y_{-n:-1}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\phi_2(\theta, X_j, Y_j)\phi_3(\theta, X_k, Y_k)\big|Y_{-m:0}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_j, Y_j)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_k, Y_k)\big|Y_{-m:0}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\phi_2(\theta, X_j, Y_j)\phi_3(\theta, X_k, Y_k)\big|Y_{-m:-1}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_i, Y_i)\big|Y_{-m:-1}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_j, Y_j)\big|Y_{-m:-1}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_k, Y_k)\big|Y_{-m:-1}\Big]\Big|,
\end{aligned} \tag{A-33}$$

$$\begin{aligned}
\beta_{a,b,c} = \sum_{i,j=-m+1}^{-1} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| & E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_i, Y_i)\phi_c(\theta, X_j, Y_j)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_j, Y_j)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_i, Y_i)\phi_c(\theta, X_j, Y_j)\big|Y_{-m:0}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_i, Y_i)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_j, Y_j)\big|Y_{-m:0}\Big]\Big|,
\end{aligned}$$

$$\begin{aligned}
\gamma_{a,b,c} = \sum_{i=-m+1}^{-1} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| & E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_0, Y_0)\phi_c(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_0, Y_0)\phi_c(\theta, X_i, Y_i)\big|Y_{-m:0}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_i, Y_i)\big|Y_{-m:0}\Big]\Big|,
\end{aligned}$$

$$\begin{aligned}
\delta = \sup_{\hat{\theta}\in\hat{\Theta}} \Big| & E_{\hat{\theta}}\Big[\phi_1(\theta, X_0, Y_0)\phi_2(\theta, X_0, Y_0)\phi_3(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_0, Y_0)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_1(\theta, X_0, Y_0)\phi_2(\theta, X_0, Y_0)\phi_3(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_1(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_2(\theta, X_0, Y_0)\big|Y_{-m:0}\Big] E_{\hat{\theta}}\Big[\phi_3(\theta, X_0, Y_0)\big|Y_{-m:0}\Big]\Big|,
\end{aligned}$$

$$\begin{aligned}
\psi_{a,b,c} = \sum_{i=-n+1}^{-m} \sum_{j,k=-n+1}^{-1} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| & E_{\hat{\theta}}\Big[\phi_a(\theta, X_i, Y_i)\phi_b(\theta, X_j, Y_j)\phi_c(\theta, X_k, Y_k)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_j, Y_j)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_k, Y_k)\big|Y_{-n:0}\Big] \\
& - E_{\hat{\theta}}\Big[\phi_a(\theta, X_i, Y_i)\phi_b(\theta, X_j, Y_j)\phi_c(\theta, X_k, Y_k)\big|Y_{-n:-1}\Big] \\
& + E_{\hat{\theta}}\Big[\phi_a(\theta, X_i, Y_i)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_j, Y_j)\big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_k, Y_k)\big|Y_{-n:-1}\Big]\Big|,
\end{aligned}$$

$$\kappa_{a,b,c} = \sum_{i=-n+1}^{-m} \sum_{j=-n+1}^{-1} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_i, Y_i)\phi_c(\theta, X_j, Y_j)\Big|Y_{-n:0}\Big]$$
$$- E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_i, Y_i)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_j, Y_j)\Big|Y_{-n:0}\Big]$$
$$- E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_i, Y_i)\phi_c(\theta, X_j, Y_j)\Big|Y_{-n:-1}\Big]$$
$$+ E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_i, Y_i)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_j, Y_j)\Big|Y_{-n:-1}\Big]\Big|,$$

and

$$\omega_{a,b,c} = \sum_{i=-n+1}^{-m} \sup_{\hat{\theta}\in\hat{\Theta}} \Big| E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_0, Y_0)\phi_c(\theta, X_i, Y_i)\Big|Y_{-n:0}\Big]$$
$$- E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_i, Y_i)\Big|Y_{-n:0}\Big]$$
$$- E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\phi_b(\theta, X_0, Y_0)\phi_c(\theta, X_i, Y_i)\Big|Y_{-n:-1}\Big]$$
$$+ E_{\hat{\theta}}\Big[\phi_a(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_b(\theta, X_0, Y_0)\Big|Y_{-n:0}\Big] E_{\hat{\theta}}\Big[\phi_c(\theta, X_i, Y_i)\Big|Y_{-n:-1}\Big]\Big|.$$

If $\|\phi_1\|_\infty, \|\phi_2\|_\infty, \|\phi_3\|_\infty \in L^3\left(\bar{\mathbb{P}}_{\theta^*}\right)$ then it follows that $\sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_1(\theta, x, Y)|$, $\sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_2(\theta, x, Y)|$ and $\sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_3(\theta, x, Y)|$ are all finite $\bar{\mathbb{P}}_{\theta^*}$ a.s.. Thus that we can apply Lemmas A.3 and A.4 to each individual term in the sum on the right hand side of (A-33) to get that for any $-m+1 \leq i, j \leq -1$ the corresponding term is bounded by

$$\sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_1(\theta, x, Y_i)| \sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_2(\theta, x, Y_j)| \sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} |\phi_3(\theta, x, Y_k)|$$
$$\times 16\left(\rho^{|i-j|\wedge|j-k|\wedge|k-i|} \wedge \rho^{i\wedge j\wedge k-m} \wedge \rho^{1-i\vee j\vee k}\right). \tag{A-34}$$

where the quantity ρ is as in Lemma A.3 and hence can be determined purely as a function of the quantities $\underline{c}$ and $\bar{c}$. It then immediately follows from the expression for α and (A-34) that there exist C_α and $0 < \rho_\alpha < 1$ which are functions purely of $\underline{c}$ and $\bar{c}$ such that for all $n \geq m > 0$

$$E_{\theta^*}[\alpha] \leq C_\alpha \rho_\alpha^m \sup_{s\in\{1,2,3,\}} E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}} \sup_{x\in\mathcal{X}} \left|\phi_s(\theta, x, Y)^2\right|\right]. \tag{A-35}$$

Clearly one can prove analogous results for the expected values $E_{\theta^*}[\beta_{a,b,c}]$, $E_{\theta^*}[\gamma_{a,b,c}]$, $E_{\theta^*}[\delta]$, $E_{\theta^*}[\psi_{a,b,c}]$, $E_{\theta^*}[\kappa_{a,b,c}]$ and $E_{\theta^*}[\omega_{a,b,c}]$ from which (A-31) immediately follows. Moreover it is easy to see that (A-32) follows from (A-31) by using that

$$E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Gamma_{0,1}\left(\hat{\theta}\right)\right|\right] < \infty$$

and

$$E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Omega_{0,n}\left(\hat{\theta}\right)\right|\right] = E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Omega_{0,1}\left(\hat{\theta}\right)\right|\right] + \sum_{r=2}^{n} E_{\theta^*}\left[\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Omega_{0,r}\left(\hat{\theta}\right) - \Omega_{0,r-1}\left(\hat{\theta}\right)\right|\right]$$

for all $n > 1$.

Finally we note that the existence of a limit in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ of the sequence of random variables $\Omega_{0,n}(\hat{\theta})$ is again an immediate consequence of (A-31). □

The next lemma is a well known result concerned with the question of when the differentiation and expectation operators can be interchanged. However since we shall need to refer to it we state it here explicitly for clarity of exposition. Its proof is a simple application of the dominated convergence theorem.

Lemma A.6. *Let a Polish space $\mathcal{X}$, a positive σ-finite measure μ on $\mathcal{X}$, a compact set Γ and a function $f(\gamma, x) : \Gamma \times \mathcal{X} \to \mathbb{R}$ be given. Suppose that f is everywhere differentiable w.r.t. to γ and*

$$\int_{\mathcal{X}} \sup_{\gamma \in \Gamma} \left| \frac{\partial}{\partial \gamma} f(\gamma, x) \right| \mu(dx) < \infty.$$

Then $\int_{\mathcal{X}} f(\gamma, x)\mu(dx)$ is everywhere differentiable w.r.t. to γ and

$$\frac{\partial}{\partial \gamma} \int_{\mathcal{X}} f(\gamma, x)\mu(dx) = \int_{\mathcal{X}} \frac{\partial}{\partial \gamma} f(\gamma, x)\mu(dx).$$

The next Lemma is a statement of the Fisher identity and the Louis missing information principle (see for example Douc et al. (2004)) plus an extension of these to third order derivatives of the log likelihood function. Under assumptions (A1)-(A5) its proof is a standard application of Lemma A.6 which we leave to the reader.

Lemma A.7. *Suppose that assumptions (A1)-(A5) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$ where for each $\theta \in \Theta$ we let $g_\theta(y|x)$ and $q_\theta(x', x)$ denote the densities of the conditional law and transition kernel of the corresponding HMM. For any $\epsilon \geq 0$ let $g^\epsilon_\theta(y|x)$ denote the density of the conditional law of the corresponding perturbed HMM defined in* (10)*. By convention we let $g^0_\theta(y|x) = g_\theta(y|x)$.*

For any $\theta \in \Theta$, $\epsilon \geq 0$ and $n > 0$ let $\psi(\theta, x, x', y) = \log g^\epsilon_\theta(y|x')\, q_\theta(x, x')$ and following the notation of Lemma A.5 let $\psi_n(\theta) \triangleq \sum_{i=1}^n \psi(\theta, X_{i-1}, X_i, Y_i)$. Then one has that for any $\theta \in \Theta$ and $\epsilon \geq 0$ the log ABC approximate likelihood function $\log p^\epsilon_\theta(\cdots)$ is three times differentiable and

$$\nabla_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n) = E_{\theta^\epsilon}\left[\nabla_\theta \psi_n(\theta) \middle| Y_{1:n}\right], \tag{A-36}$$

$$\nabla^2_\theta \frac{1}{n} \log p^\epsilon_\theta(Y_1, \ldots, Y_n) = E_{\theta^\epsilon}\left[\nabla^2_\theta \psi_n \middle| Y_{1:n}\right] + E_{\theta^\epsilon}\left[(\nabla_\theta \psi_n)^2 \middle| Y_{1:n}\right] - E_{\theta^\epsilon}\left[\nabla_\theta \psi_n \middle| Y_{1:n}\right]^2, \tag{A-37}$$

and

$$\begin{aligned}\nabla^3_\theta \frac{1}{n} \log p^\epsilon_\theta(Y_1, \ldots, Y_n) &= E_{\theta^\epsilon}\left[\nabla^3_\theta \psi_n \middle| Y_{1:n}\right] + 3E_{\theta^\epsilon}\left[\nabla^2_\theta \psi_n \nabla_\theta \psi_n \middle| Y_{1:n}\right] \\ &- 3E_{\theta^\epsilon}\left[\nabla^2_\theta \psi_n \middle| Y_{1:n}\right] E_{\theta^\epsilon}\left[\nabla_\theta \psi_n \middle| Y_{1:n}\right] - 3E_{\theta^\epsilon}\left[(\nabla_\theta \psi_n)^2 \middle| Y_{1:n}\right] E_{\theta^\epsilon}\left[\nabla_\theta \psi_n \middle| Y_{1:n}\right] \\ &+ E_{\theta^\epsilon}\left[(\nabla_\theta \psi_n)^3 \middle| Y_{1:n}\right] + 2E_{\theta^\epsilon}\left[\nabla_\theta \psi_n \middle| Y_{1:n}\right]^3 \end{aligned} \tag{A-38}$$

where $E_{\theta^\epsilon}[\cdot|\cdot]$ denotes conditional expectation w.r.t. the law of the perturbed HMM defined by (8)*.*

The final Lemma shows that the Fisher identity, Louis missing information principle etc. also hold in a mean sense.

Lemma A.8. *Suppose that assumptions (A1)-(A5) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Then one has that for any $\theta \in \Theta$ and $\epsilon \geq 0$ the log ABC approximate likelihood function $\log p^\epsilon_\theta(\cdots)$ is three times differentiable and*

$$\nabla_\theta E_{\theta^*}\left[\log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right] = E_{\theta^*}\left[\nabla_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right], \tag{A-39}$$

$$\nabla^2_\theta E_{\theta^*}\left[\log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right] = E_{\theta^*}\left[\nabla^2_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right], \tag{A-40}$$

and

$$\nabla^3_\theta E_{\theta^*}\left[\log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right] = E_{\theta^*}\left[\nabla^3_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right]. \tag{A-41}$$

Proof. Equations (A-39), (A-40) and (A-41) will follow immediately from Lemmas A.6 and A.8 if one can establish that

$$E_{\theta^*}\left[\sup_{\theta \in \Theta} |\nabla_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)|\right], E_{\theta^*}\left[\sup_{\theta \in \Theta} \left|\nabla^2_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right|\right],$$
$$E_{\theta^*}\left[\sup_{\theta \in \Theta} \left|\nabla^3_\theta \log p^\epsilon_\theta(\, Y_1, \ldots, Y_n)\right|\right] < \infty.$$

However this follows immediately from assumptions (A1)-(A5) and expressions (A-36)-(A-38). □

Appendix B: Proofs of Theorems 3.1 , 3.2 and 3.3

We start with the the proof of Theorem 3.1.

Proof of Theorem 3.1 . Consistency of the MLE can be deduced from conditions 1-2 by standard arguments.

Let p_0 denote the density of the prior π_0 w.r.t. Lebesgue measure. It follows from conditions 1-4 of the theorem that for any $\delta > 0$ there exist some $\eta > 0$ such that

$$\sup_{\theta,\theta' \in B^{\eta}_{\theta^*}} \frac{p_0(\theta)}{p_0(\theta')} \leq (1+\delta), \tag{B-42}$$

and some and $\gamma > 0$ such that $\overline{\mathbb{P}}_{\theta^*}$ a.s. there exists some n' such that for all $n \geq n'$

$$\pi^n(\Theta / B^{\eta}_{\theta^*}) \leq e^{-n\gamma}, \tag{B-43}$$

and

$$\sup_{\theta \in B^{\eta}_{\theta^*}} \sup_{v \in \mathbb{R}^d} \frac{\left| v^T \nabla^2_{\theta} \frac{1}{n} \left(\log p_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n) - \log p_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n) \right) v - v^T \nabla^2_{\theta} l\left(\theta^*\right) v \right|}{\left| v^T \nabla^2_{\theta} l\left(\theta^*\right) v \right|} \leq \delta. \tag{B-44}$$

Using the fact that $\nabla_{\theta} \log p_{\hat{\theta}_n}(\hat{Y}_1, \ldots, \hat{Y}_n) = 0$ it follows from the consistency of the MLE and Taylor's theorem that $\overline{\mathbb{P}}_{\theta^*}$ a.s. one has that for all n sufficiently large that for all $\theta \in B^{\eta}_{\theta^*}$

$$\log p_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n) - \log p_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n) = \frac{1}{2} n (\theta - \hat{\theta}_n)^T \nabla^2_{\theta} l\left(\theta^*\right) (\theta - \hat{\theta}_n) + R_n(\theta) \tag{B-45}$$

where for all $\theta \in B^{\eta}_{\theta^*}$ the remainder term $R_n(\theta)$ is bounded by

$$\sup_{\theta' \in B^{\eta}_{\theta^*}} \frac{1}{2} \left| (\theta - \hat{\theta}_n)^T \nabla^2_{\theta} \log p_{\theta'}(\hat{Y}_1, \ldots, \hat{Y}_n)(\theta - \hat{\theta}_n) - n(\theta - \hat{\theta}_n)^T \nabla^2_{\theta} l\left(\theta^*\right) (\theta - \hat{\theta}_n) \right|. \tag{B-46}$$

It then follows from (B-44), (B-45) and (B-46) that $\overline{\mathbb{P}}_{\theta^*}$ a.s. one has that for all n sufficiently large that

$$e^{\frac{1}{2} n (\theta - \hat{\theta}_n)^T \nabla^2_{\theta} l(\theta^*)(\theta - \hat{\theta}_n)(1+\delta)} \leq \frac{p_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n)}{p_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n)} \leq e^{\frac{1}{2} n (\theta - \hat{\theta}_n)^T \nabla^2_{\theta} l(\theta^*)(\theta - \hat{\theta}_n)(1-\delta)} \tag{B-47}$$

for all $\theta \in B^{\eta}_{\theta^*}$. The result now follows from (B-42), (B-43) and (B-47). □

The proofs of Theorems 3.2 and 3.3 rely on the following three lemmas which show firstly that the relative mean log ABC likelihood surface

$$\theta \to \frac{1}{n} \left(\log p^{\epsilon}_{\theta}(Y_1, \ldots, Y_n) - \log p^{\epsilon}_{\theta^*}(Y_1, \ldots, Y_n) \right)$$

converges uniformly $\bar{\mathbb{P}}_{\theta^*}$ a.s. to the surface defined by some function $l^{\epsilon}(\theta)$ and secondly that the curvature of the limiting function $l^{\epsilon}(\theta)$ converges to that of the function $l(\theta)$ defined in (13) as $\epsilon \to 0$. The proofs of these lemmas are deferred to Section B.4.

Lemma B.9. *Suppose that assumptions (A1)-(A5) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Then for any $\epsilon \geq 0$ there exists a three times continuously differentiable function $l^{\epsilon}(\theta)$ such that*

$$\lim_{n \to \infty} \sup_{\theta \in \Theta} \left| \frac{1}{n} \left(\log p^{\epsilon}_{\theta}(Y_1, \ldots, Y_n) - \log p^{\epsilon}_{\theta^*}(Y_1, \ldots, Y_n) \right) - l^{\epsilon}(\theta) \right| = 0 \tag{B-48}$$

*$\overline{\mathbb{P}}_{\theta^*}$ a.s. where for all θ and ϵ, $p^{\epsilon}_{\theta}(\cdots)$ denotes the ABC approximate likelihood function defined in (**??**). By convention we define $p^{0}_{\theta}(\cdots)$ to be equal to the true likelihood function $p_{\theta}(\cdots)$. Moreover there exists some constant $0 < K < \infty$ such that for all $\theta \in \Theta$ and $\epsilon \geq 0$*

$$\nabla_{\theta} l^{\epsilon}(\theta), \nabla^2_{\theta} l^{\epsilon}(\theta), \nabla^3_{\theta} l^{\epsilon}(\theta) \leq K \tag{B-49}$$

Further one has that $\overline{\mathbb{P}}_{\theta^*}$ *a.s. there exists some* n' *such that for all* $n \geq n'$

$$\left.\begin{array}{l}\sup_{\theta\in\Theta}\left\|\frac{1}{n}\nabla_\theta\big(\log p_\theta^\epsilon(Y_1,\ldots,Y_n)-\log p_{\theta^*}^\epsilon(Y_1,\ldots,Y_n)\big)\right\|\\ \sup_{\theta\in\Theta}\left\|\frac{1}{n}\nabla_\theta^2\big(\log p_\theta^\epsilon(Y_1,\ldots,Y_n)-\log p_{\theta^*}^\epsilon(Y_1,\ldots,Y_n)\big)\right\|\\ \sup_{\theta\in\Theta}\left\|\frac{1}{n}\nabla_\theta^3\big(\log p_\theta^\epsilon(Y_1,\ldots,Y_n)-\log p_{\theta^*}^\epsilon(Y_1,\ldots,Y_n)\big)\right\|\end{array}\right\}\leq K \tag{B-50}$$

for the same K *as in* (B-49).

Lemma B.10. *Suppose that assumptions (A1)-(A5) hold for a collection of HMMs parametrised by some vector* $\theta\in\Theta$ *and for any* $\epsilon>0$ *let* $l^\epsilon(\theta)$ *be equal to the corresponding limit function defined in Lemma B.9. Then for all* $\theta\in\Theta$ *one has that*

$$\lim_{\epsilon\to 0}\nabla_\theta l^\epsilon(\theta)=\nabla_\theta l(\theta).$$

Lemma B.11. *Suppose that assumptions (A1)-(A5) hold for a collection of HMMs parametrised by some vector* $\theta\in\Theta$ *and for any* $\epsilon>0$ *let* $l^\epsilon(\theta)$ *be equal to the corresponding limit function defined in Lemma B.9. Then for all* $\theta\in\Theta$ *one has that*

$$\lim_{\epsilon\to 0}\nabla_\theta^2 l^\epsilon(\theta)=\nabla_\theta^2 l(\theta).$$

Proof of Theorem 3.2. It follows immediately from (B-48), (B-50) and Lemma A.2 that for all ϵ

$$\lim_{n\to\infty}\frac{1}{n}\left(\log p_\theta^\epsilon(Y_1,\ldots,Y_n)-\log p_{\theta^*}^\epsilon(Y_1,\ldots,Y_n)\right)=l^\epsilon(\theta) \tag{B-51}$$

uniformly in θ $\overline{\mathbb{P}}$ a.s. where l^ϵ is the limiting function defined in Lemma B-48. Furthermore we have from (B-50) and Lemma B.10 that

$$\lim_{\epsilon\to 0}\sup_{\theta\in\theta}|l^\epsilon(\theta)-l(\theta)|=0. \tag{B-52}$$

For any $\epsilon>0$ let $\mathcal{M}_\epsilon$ denote the set of maximizers of l^ϵ. We begin by noting that since l^ϵ is continuous for all ϵ the sets $\mathcal{M}_\epsilon$ are always non-empty. Since by assumption $l(\theta)$ has a unique maximum at θ^* and is continuous in Θ it follows from (B-52) that

$$\lim_{\epsilon\to 0}\sup_{\theta\in\mathcal{M}_\epsilon}|\theta-\theta^*|=0. \tag{B-53}$$

Suppose that for all ϵ sufficiently small the set $\mathcal{M}_\epsilon$ consists of a single element

$$\mathcal{M}_\epsilon=\{\theta^{*,\epsilon}\} \tag{B-54}$$

for some $\theta^{*,\epsilon}\in\Theta$. It would then follow from (B-53) that $\theta^{*,\epsilon}\to\theta^*$ and hence from (B-50) and Lemma B.11 that $\nabla_\theta^2 l^\epsilon(\theta^{*,\epsilon})\to\nabla_\theta^2 l(\theta^*)$ as $\epsilon\to 0$. Thus in order to complete the proof of the theorem it is sufficient to prove that (B-54) holds for sufficiently small ϵ.

We start by noting that from Lemma B.11 we have that

$$\lim_{\epsilon\to 0}\left\|\nabla_\theta^2 l^\epsilon(\theta^*)-\nabla_\theta^2 l(\theta^*)\right\|=0. \tag{B-55}$$

Since we have by Lemma B.9 that there exists some finite constant K such that

$$\left\|\nabla_\theta^2 l^\epsilon(\theta')-\nabla_\theta^2 l^\epsilon(\theta)\right\|\leq K|\theta'-\theta| \tag{B-56}$$

for all ϵ and $\theta,\theta'\in\Theta$ it follows from (B-55) and the assumption that $\nabla_\theta^2 l(\theta^*)<0$ that there exists some $\eta>0$ such that

$$\lim_{\epsilon\to 0}\sup_{\theta\in B_{\theta^*}^\eta}\sup_{v\in\mathbb{R}^d}\frac{\left|\nabla_\theta^2 l^\epsilon(\theta)v-\nabla_\theta^2 l(\theta^*)v\right|}{|\nabla_\theta^2 l(\theta^*)v|}\leq\frac{1}{2}. \tag{B-57}$$

It then follows from (B-57) and a simple application of Taylor's theorem that

$$\lim_{\epsilon\to 0}\inf_{\theta,\theta'\in B_{\theta^*}^\eta:\theta\neq\theta'}\frac{|\nabla_\theta l^\epsilon(\theta)-\nabla_\theta l^\epsilon(\theta')|}{|\nabla_\theta^2 l(\theta^*)(\theta-\theta')|}>0 \tag{B-58}$$

and hence that for sufficiently small ϵ that there is at most one $\theta\in B_{\theta^*}^\eta$ such that $\nabla_\theta l^\epsilon(\theta)=0$. Equation (B-54) and hence the proof of the theorem now follows from the preceding observation and (B-53). □

Proof of Theorem 3.3. Given the results in Theorem 3.2 the proof of Theorem 3.3 follows in exactly the same way as the proof of Theorem 3.1. We leave the details to the reader. □

B.4 Proofs of Lemmas B.9, B.10 and B.11

In order to complete this section we need to provide the proofs of Lemmas B.9, B.10 and B.11. We start with the proof of Lemma B.9.

Proof of Lemma B.9. For any n the gradient of the log ABC likelihood may be decomposed into the following telescoping sum

$$\nabla_\theta \frac{1}{n} \log p_\theta^\epsilon(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} \big(\nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big). \tag{B-59}$$

It follows from (A-36) and (A-28) that for all $\theta \in \Theta$ and $\epsilon \geq 0$ there exists some $\sigma(Y_{-\infty:0})$ measurable random variable $R_\theta^\epsilon(Y_{-\infty:0})$ such that

$$\nabla_\theta \log p_\theta^\epsilon(Y_{-n}, \ldots, Y_0) - \nabla_\theta \log p_\theta^\epsilon Y_{-n}, \ldots, Y_{-1}) \to R_\theta^\epsilon(Y_{-\infty:0}) \tag{B-60}$$

in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$ as $n \to \infty$. Furthermore it follows from (A-36) and (A-31) that there exist constants $C < \infty$ and $0 < \rho < 1$ such that for all $\theta \in \Theta$, $\epsilon \geq 0$ and $n > 0$

$$E_{\theta^*}\left[\sup_{k \geq n} \left| \nabla_\theta \log p_\theta^\epsilon(Y_{-k}, \ldots, Y_0) - \nabla_\theta \log p_\theta^\epsilon(Y_{-k}, \ldots, Y_{-1}) - R_\theta^\epsilon(Y_{-\infty:0}) \right| \right] \leq C\rho^n.$$

Thus we have from the ergodic theorem that for any $m > n > 0$

$$\begin{aligned}
&\limsup_{m\to\infty} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0}) \right] \right| \\
&\leq \limsup_{m\to\infty} \left| \frac{1}{m} \sum_{i=1}^{n} \big(\nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0}) \right] \right| \\
&+ \limsup_{m\to\infty} \left| \frac{1}{m} \sum_{i=n+1}^{m} R_\theta^\epsilon(Y_{-\infty:i}) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0}) \right] \right| \\
&+ \limsup_{m\to\infty} \frac{1}{m} \sum_{i=n+1}^{m} \sup_{k \geq -1} \left| \nabla_\theta \log p_\theta^\epsilon(Y_{-k}, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_{-k}, \ldots, Y_{i-1}) - R_\theta^\epsilon(Y_{-\infty:i}) \right| \\
&\leq C\rho^n.
\end{aligned} \tag{B-61}$$

It now follows from (B-59) and (B-61) that

$$\nabla_\theta \frac{1}{n} \log p_\theta^\epsilon(Y_1, \ldots, Y_n) \to E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0}) \right] \tag{B-62}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. Moreover it follows from (A-32) and the ergodic theorem that

$$\begin{aligned}
&\limsup_{m\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big) \right| \\
&\leq \limsup_{m\to\infty} \frac{1}{m} \sum_{i=1}^{n} \sup_{\theta\in\Theta} \sup_{k \leq i-1} \left| \big(\nabla_\theta \log p_\theta^\epsilon(Y_k, \ldots, Y_i) - \nabla_\theta \log p_\theta^\epsilon(Y_k, \ldots, Y_{i-1}) \big) \right| \\
&= C
\end{aligned} \tag{B-63}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. Similarly one can show that

$$\begin{aligned}
&\limsup_{m\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\nabla_\theta^2 \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta^2 \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big) \right|, \\
&\limsup_{m\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\nabla_\theta^3 \log p_\theta^\epsilon(Y_1, \ldots, Y_i) - \nabla_\theta^3 \log p_\theta^\epsilon(Y_1, \ldots, Y_{i-1}) \big) \right| \leq C
\end{aligned} \tag{B-64}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. and that for any $\theta \in \Theta$ and $\epsilon \geq 0$ there exist $\sigma(Y_{-\infty:0})$ measurable random variable $S^\epsilon_\theta(Y_{-\infty:0})$ and $T^\epsilon_\theta(Y_{-\infty:0})$ such that

$$\begin{aligned}\nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_{-1}) \to S^\epsilon_\theta(Y_{-\infty:0}),\\ \nabla^3_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla^3_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_{-1}) \to T^\epsilon_\theta(Y_{-\infty:0})\end{aligned} \tag{B-65}$$

in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$ and

$$\nabla^2_\theta \frac{1}{n} \log p^\epsilon_\theta(Y_1,\ldots,Y_n) \to E_{\theta^*}\left[S^\epsilon_\theta(Y_{-\infty:0})\right], \nabla^3_\theta \frac{1}{n} \log p^\epsilon_\theta(Y_1,\ldots,Y_n) \to E_{\theta^*}\left[T^\epsilon_\theta(Y_{-\infty:0})\right] \tag{B-66}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s..

Clearly (B-50) follows from (B-63) and (B-64). Furthermore, under the assumption that (B-48) holds (for a three times differentiable l^ϵ) it is easy to see that (B-49) is then an immediate consequence of (B-50), (B-62) and (B-66). Thus in order to complete the proof of the lemma it remains to show (B-48).

We start by noting that it follows from Lemma A.8 that for any $\theta \in \Theta$, $\epsilon \geq 0$ and $n > 0$

$$\begin{aligned}&\nabla_\theta E_{\theta^*}\left[\left(\nabla_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0)\right)\right]\\ &\quad = E_{\theta^*}\left[\left(\nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0)\right)\right]\end{aligned} \tag{B-67}$$

and

$$\begin{aligned}&\nabla_\theta E_{\theta^*}\left[\left(\nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla^2_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0)\right)\right]\\ &\quad = E_{\theta^*}\left[\left(\nabla^3_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0) - \nabla^3_\theta \log p^\epsilon_\theta(Y_{-n},\ldots,Y_0)\right)\right].\end{aligned} \tag{B-68}$$

Since (A-31) implies that the convergence in (B-60) and (B-65) is uniform w.r.t. θ it follows from Lemma A.1 that $E_{\theta^*}\left[R^\epsilon_\theta(Y_{-\infty:0})\right]$ and $E_{\theta^*}\left[S^\epsilon_\theta(Y_{-\infty:0})\right]$ are differentiable and that $\nabla_\theta E_{\theta^*}\left[R^\epsilon_\theta(Y_{-\infty:0})\right] = E_{\theta^*}\left[S^\epsilon_\theta(Y_{-\infty:0})\right]$ and $\nabla_\theta E_{\theta^*}\left[S^\epsilon_\theta(Y_{-\infty:0})\right] = E_{\theta^*}\left[T^\epsilon_\theta(Y_{-\infty:0})\right]$.

One more application of Lemma A.1 will complete the proof of (B-48) if we can show that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{n} \nabla_\theta \left(\log p^\epsilon_\theta(Y_1,\ldots,Y_n) - \log p^\epsilon_{\theta^*}(Y_1,\ldots,Y_n)\right) - E_{\theta^*}\left[R^\epsilon_\theta(Y_{-\infty:0})\right] \right| = 0 \tag{B-69}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. (Note that the gradient of the empirical functions are converging uniformly while the functions themselves are Cauchy at θ^*.) (B-50) implies that $\overline{\mathbb{P}}_{\theta^*}$ a.s. the sequence of random variables $\frac{1}{n}\nabla_\theta \left(\log p^\epsilon_\theta(Y_1,\ldots,Y_n) - \log p^\epsilon_{\theta^*}(Y_1,\ldots,Y_n)\right)$ are eventually uniformly Lipschitz continuous w.r.t. θ. Furthermore equation (B-62) also implies that there exists a countable dense set of Θ, say D, such that $\overline{\mathbb{P}}_{\theta^*}$ a.s. the sequence of random variables $\frac{1}{n}\nabla_\theta \left(\log p^\epsilon_\theta(Y_1,\ldots,Y_n) - \log p^\epsilon_{\theta^*}(Y_1,\ldots,Y_n)\right)$ converge to $E_{\theta^*}\left[R^\epsilon_\theta(Y_{-\infty:0})\right]$ for all $\theta \in D$. We can now deduce (B-69) from the two preceding observations and a direct application of Lemma A.2.

□

In remains to provide the proofs of Lemmas B.10 and B.11. Since the proofs of these two lemmas are almost identical we prove only Lemma B.10.

Proof of Lemma B.10. We start by stating some properties of the perturbed conditional likelihood (10) that will needed in the sequel. First note that it follows from assumptions (A2) and (A5) and Lemma A.6 that

$$\nabla_\theta g^\epsilon_\theta(y|x) \triangleq \frac{\int_{B^\epsilon_y} \nabla_\theta g_\theta(z|x)\,\nu(dz)}{\int_{B^\epsilon_y} \nu(dz)}. \tag{B-70}$$

Furthermore since

$$\int_{B^\epsilon_y} \nabla_\theta g_\theta(z|x)\,\nu(dz) \leq \sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B^\epsilon_y} \left(\frac{\nabla_\theta g_\theta(z|x)}{g_\theta(z|x)}\right) \times \int_{B^\epsilon_y} g_\theta(z|x)\,\nu(dz)$$

it follows from (B-70) and assumption (A5) that for any $\epsilon > 0$

$$\bar{E}_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left\|\nabla_\theta \log g_\theta^\epsilon\left(Y|x\right)\right\|\right] < \infty. \tag{B-71}$$

We are now ready to begin the proof of Lemma B.10 proper. It follows from (A-31), (A-36), (B-59), (B-60), (B-69) and assumptions (A1)-(A5) that for any $\delta > 0$ there exists n' such that

$$\left|\nabla_\theta l^\epsilon\left(\theta\right) - \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_\theta^\epsilon(Y_1,\ldots,Y_n)\right]\right| \leq \delta$$

for all $n \geq n'$, θ and $\epsilon \geq 0$. Thus in order to prove the result it is sufficient to show that

$$\lim_{\epsilon\to 0}\bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_\theta^\epsilon(Y_1,\ldots,Y_n)\right] = \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_\theta(Y_1,\ldots,Y_n)\right]$$

for all θ and hence by (A-36) that

$$\begin{aligned}&\lim_{\epsilon\to 0}\bar{E}_{\theta^*}\left[E_{\theta^\epsilon}\left[\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\right]\\ &\quad = \bar{E}_{\theta^*}\left[E_{\theta}\left[\nabla_\theta\left(\log g_\theta\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\right]\end{aligned} \tag{B-72}$$

for all θ and $1 \leq k \leq n$. Assumption (A3) implies that

$$\left|\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)\right| \leq \sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\nabla_\theta \log g_\theta^\epsilon\left(Y_k|x\right)\right| + K$$

for some finite positive constant that is independent of θ and ϵ and hence it follows from (B-71) and the dominated convergence theorem that in order to prove (B-72) it is sufficient to show that

$$\begin{aligned}&\lim_{\epsilon\to 0}E_{\theta^\epsilon}\left[\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\\ &\quad = E_{\theta}\left[\nabla_\theta\left(\log g_\theta\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\end{aligned} \tag{B-73}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. Recall that

$$\begin{aligned}&E_{\theta^\epsilon}\left[\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\\ &= \frac{\int_{\mathcal{X}^n}\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|x_k\right)q_\theta\left(x_{k-1},x_k\right)\right)\prod_{i=1}^n\left(g_\theta^\epsilon\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_1)\cdots\mu(dx_n)}{\int_{\mathcal{X}^n}\prod_{i=1}^n\left(g_\theta^\epsilon\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_1)\cdots\mu(dx_n)}\end{aligned} \tag{B-74}$$

and

$$\begin{aligned}&E_{\theta}\left[\nabla_\theta\left(\log g_\theta\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\\ &= \frac{\int_{\mathcal{X}^n}\nabla_\theta\left(\log g_\theta\left(Y_k|x_k\right)q_\theta\left(x_{k-1},x_k\right)\right)\prod_{i=1}^n\left(g_\theta\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_1)\cdots\mu(dx_n)}{\int_{\mathcal{X}^n}\prod_{i=1}^n\left(g_\theta\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_1)\cdots\mu(dx_n)}.\end{aligned} \tag{B-75}$$

Since by assumptions (A2) and (A5) we have for any $K > 0$ that

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla_\theta g_\theta^\epsilon\left(Y+z|x\right)\right\| < \infty$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. it follows that we can use the Lebesgue differentiation theorem (see for example Wheeden and Zygmund (1977)) to deduce that for μ a.s. all $x \in \mathcal{X}$ that

$$\nabla_\theta g_\theta^\epsilon\left(Y_k|x\right) \to \nabla_\theta g_\theta\left(Y_k|x\right),\ g_\theta^\epsilon\left(Y_k|x\right) \to g_\theta\left(Y_k|x\right) \tag{B-76}$$

ν a.s.. Standard arguments show that for all $\theta \in \Theta$ the set $\mathcal{N}(\theta) \in \mathcal{X}\times\mathcal{Y}$ defined by

$$\mathcal{N}(\theta) = \left\{x, y : \lim_{\epsilon\to 0} g_\theta^\epsilon\left(y|x\right) = g_\theta\left(y|x\right)\right\}$$

is $\mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y})$ measurable. Furthermore it follows from (B-76) that the set $\mathcal{N}(\theta)$ has $\mu \times \nu$ full measure and hence that $\overline{\mathbb{P}}_{\theta^*}$ a.s.

$$\lim_{\epsilon \to 0} g_\theta^\epsilon (Y|x) = g_\theta (Y|x) \quad \nu \text{ a.s..} \tag{B-77}$$

Similarly one can show that $\overline{\mathbb{P}}_{\theta^*}$ a.s.

$$\lim_{\epsilon \to 0} \nabla_\theta g_\theta^\epsilon (Y|x) = \nabla_\theta g_\theta (Y|x) \quad \nu \text{ a.s..} \tag{B-78}$$

It follows from assumptions (A2) and (A5) that

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \nabla_\theta \left(\log g_\theta^\epsilon (Y_k|x_k) \, q_\theta (x_{k-1}, x_k) \right) \prod_{i=1}^{n} \left(g_\theta^\epsilon (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) < \infty \tag{B-79}$$

and

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \prod_{i=1}^{n} \left(g_\theta^\epsilon (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) < \infty \tag{B-80}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. and hence from (B-77) and (B-78) and from the dominated convergence theorem that

$$\begin{aligned} &\int_{\mathcal{X}^n} \nabla_\theta \left(\log g_\theta^\epsilon (Y_k|x_k) \, q_\theta (x_{k-1}, x_k) \right) \prod_{i=1}^{n} \left(g_\theta^\epsilon (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) \mu(dx_1) \cdots \mu(dx_n) \to \\ &\int_{\mathcal{X}^n} \nabla_\theta \left(\log g_\theta (Y_k|x_k) \, q_\theta (x_{k-1}, x_k) \right) \prod_{i=1}^{n} \left(g_\theta (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) \mu(dx_1) \cdots \mu(dx_n) \end{aligned} \tag{B-81}$$

and

$$\begin{aligned} &\int_{\mathcal{X}^n} \prod_{i=1}^{n} \left(g_\theta^\epsilon (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) \mu(dx_1) \cdots \mu(dx_n) \to \\ &\int_{\mathcal{X}^n} \prod_{i=1}^{n} \left(g_\theta (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) \mu(dx_1) \cdots \mu(dx_n) \end{aligned} \tag{B-82}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. Since by assumption (A4) we have that

$$\int_{\mathcal{X}^n} \prod_{i=1}^{n} \left(g_\theta (Y_i|x_i) \, q_\theta (x_{i-1}, x_i) \right) \mu(dx_1) \cdots \mu(dx_n) > 0$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. it now follows that (B-73) is implied by (B-81) and (B-82). □

Appendix C: Proof of Theorem 3.4

The proof of Theorem 3.4 is based on the following lemma whose proof is deferred to Section C.5.

Lemma C.12. *Suppose that assumptions (A1)-(A5) and (A7) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Furthermore for any $\theta \in \Theta$ and $\epsilon > 0$ let $l^\epsilon(\theta)$ be equal to the corresponding limit function defined in Lemma B.9. Then for all θ there exists a vector V_θ such that for all $\epsilon > 0$*

$$\nabla_\theta l^\epsilon (\theta^*) - \nabla_\theta l (\theta^*) = \epsilon^2 V_\theta + o(\epsilon^2). \tag{C-83}$$

Proof of Theorem 3.4. We have by Theorem 3.2 and Lemma B.9 that for any $\eta > 0$ there exists some ϵ_η such that for all $\epsilon \leq \epsilon_\eta$ the surface $l^\epsilon(\theta)$ has a unique maximum at some point $\theta^{*,\epsilon} \in B_{\theta^*}^\eta$, is differentiable in $B_{\theta^*}^\eta$ and

$$\nabla_\theta l^\epsilon (\theta^{*,\epsilon}) = 0. \tag{C-84}$$

For all ϵ let $\tilde{\theta}^\epsilon = \theta^* - \epsilon^2 \nabla_\theta^2 l(\theta^*)^{-1} V_{\theta^*}$ where V_{θ^*} is as in Lemma C.12. It follows from (B-50), Lemma B.11 and (C-83) that

$$\nabla_\theta l^\epsilon(\tilde{\theta}^\epsilon) = o(\epsilon^2). \tag{C-85}$$

We then have from (B-58), (C-84) and (C-85) that

$$\left|\tilde{\theta}^\epsilon - \theta^{*,\epsilon}\right| = o(\epsilon^2)$$

which concludes the proof of the theorem. □

C.5 Proof of Lemma C.12

A central role in the proof of Lemma C.12 will be played by the following time inhomogeneous versions of the perturbed HMM (8).

Suppose that one has a collection of HMMs parametrised by some parameter vector $\theta \in \Theta$ and that for each value of θ the conditional laws and transition kernels of the corresponding HMM have densities $g_\theta(y|x)$ and $q_\theta(x, x')$ respectively. Given some $\theta \in \Theta$ and $\epsilon > 0$ we define the HMM $\left\{X_i^+, Y_i^+\right\}_{i\in\{\ldots,-1,0,1,\ldots\}}$ to be the time inhomogeneous HMM such that at each time i the process has transition kernel $q_\theta\left(i-1, x_{i-1}, x_i\right)$ and conditional law $g_{\theta,i}^{\epsilon,+}\left(y|x\right)$ where

$$q_\theta\left(i-1, x_{i-1}, x_i\right) = q_\theta\left(x_{i-1}, x_i\right), \quad g_{\theta,i}^{\epsilon,+}\left(y|x\right) = \begin{cases} g_\theta^\epsilon\left(y|x\right) & \text{if } i > 0 \\ g_\theta\left(y|x\right) & \text{otherwise} \end{cases} \tag{C-86}$$

and where the quantity $g_\theta^\epsilon\left(y|x\right)$ in (C-86) is equal to the ABC perturbed conditional likelihood defined in (10). Similarly we define the HMM $\left\{X_i^-, Y_i^-\right\}_{i\in\{\ldots,-1,0,1,\ldots\}}$ to be the time inhomogeneous HMM such that at each time i the process has transition kernel $q_\theta\left(i-1, x_{i-1}, x_i\right)$ and conditional law $g_{\theta,i}^{\epsilon,-}\left(y|x\right)$ where

$$q_\theta\left(i-1, x_{i-1}, x_i\right) = q_\theta\left(x_{i-1}, x_i\right), \quad g_{\theta,i}^{\epsilon,-}\left(y|x\right) = \begin{cases} g_\theta^\epsilon\left(y|x\right) & \text{if } i \geq 0 \\ g_\theta\left(y|x\right) & \text{otherwise} \end{cases}. \tag{C-87}$$

In what follows we shall be interested in the law of the HMM $\left\{X_i^+, Y_i^+\right\}$ as the time of the initial distribution tends to $-\infty$. Clearly the restriction of the resulting law to the set of all times less than or zero should be equal to that of the stationary distribution $\overline{\mathbb{P}}_\theta$ of the corresponding unperturbed HMM while in general one would expect the two laws to diverge for later times. This leads us to define, for all $\theta \in \Theta$ and $\epsilon > 0$, the distribution $P_{\theta^{\epsilon,+}}$ on the space $(\mathcal{X} \times \mathcal{Y})^\infty$ by

$$P_{\theta^{\epsilon,+}}\left(A\right) = \overline{\mathbb{P}}_\theta\left(A\right) \tag{C-88}$$

for all $A \in \sigma(X_{-\infty:0} \times Y_{-\infty:0})$ and

$$P_{\theta^{\epsilon,+}}\left(A|X_{-\infty:0}, Y_{-\infty:0}\right) = P_{\theta^{\epsilon,+}}\left(A|X_0\right) = \overline{\mathbb{P}}_{\theta^\epsilon}\left(A|X_0\right) \tag{C-89}$$

for all $A \in \sigma(X_{1:\infty} \times Y_{1:\infty})$ where $\overline{\mathbb{P}}_{\theta^\epsilon}$ denotes the stationary distribution of the perturbed HMM defined in (8). Similarly we define the distribution $P_{\theta^{\epsilon,-}}$ on the space $(\mathcal{X} \times \mathcal{Y})^\infty$ by

$$P_{\theta^{\epsilon,-}}\left(A\right) = \overline{\mathbb{P}}_\theta\left(A\right) \tag{C-90}$$

for all $A \in \sigma(X_{-\infty:-1} \times Y_{-\infty:-1})$ and

$$P_{\theta^{\epsilon,-}}\left(A|X_{-\infty:-1}, Y_{-\infty:-1}\right) = P_{\theta^{\epsilon,-}}\left(A|X_{-1}\right) = \overline{\mathbb{P}}_{\theta^\epsilon}\left(A|X_{-1}\right) \tag{C-91}$$

for all $A \in \sigma(X_{0:\infty} \times Y_{0:\infty})$. Finally for all θ and $\epsilon > 0$ we shall let $p_{\theta^{\epsilon,+}}(\cdots)$, $E_{\theta^{\epsilon,+}}[\cdot]$ and $E_{\theta^{\epsilon,+}}[\cdot|\cdot]$ and $p_{\theta^{\epsilon,-}}(\cdots)$, $E_{\theta^{\epsilon,-}}[\cdot]$ and $E_{\theta^{\epsilon,-}}[\cdot|\cdot]$ denote the likelihood functions and expectation and conditional expectation operators w.r.t. to the laws $P_{\theta^{\epsilon,+}}$ and $P_{\theta^{\epsilon,-}}$ respectively.

Remark C.8. *Using the same techniques as were used to prove Lemma A.3 one can show that analogous results hold for the inhomogeneous HMMs defined in* (C-86) *and* (C-87)*. See for example Cappé et al. (2005) for more details.*

Remark C.9. *It follows from* (C-86)-(C-91) *and Remark C.8 that for all* $r \neq 0$

$$\|P_{\theta^{\epsilon,+}}(X_r|Y_{-\infty:\infty}) = P_{\theta^{\epsilon,-}}(X_r|Y_{-\infty:\infty})\|_{TV} \leq 4\rho^{|r|}.$$

The following relation, which is an immediate consequence of the definitions of $P_{\theta^\epsilon,+}$ and $P_{\theta^\epsilon,-}$, will prove very useful:

$$P_{\theta^\epsilon,+}(X_{-\infty:\infty}|Y_{-\infty:-1;1:\infty}) = P_{\theta^\epsilon,-}(X_{-\infty:\infty}|Y_{-\infty:-1;1:\infty}). \tag{C-92}$$

The next three results show how the inhomogeneous perturbed HMMs above relate to the limiting behaviour of the gradients of the log likelihood surfaces $l^\epsilon(\theta)$.

Lemma C.13. *Suppose that assumptions (A1)-(A5) and (A7) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Furthermore for any $\theta \in \Theta$ and $\epsilon > 0$ let $l^\epsilon(\theta)$ be equal to the corresponding limit function defined in Lemma B.9. Then for all $\epsilon > 0$ and $\theta \in \Theta$*

$$\begin{aligned}
\nabla_\theta l^\epsilon(\theta) - \nabla_\theta l(\theta) = \sum_{i=-\infty}^{-1} \Big(& E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta(Y_i|X_i)\, q_\theta(X_{i-1}, X_i)\right)|Y_{\infty:\infty}\right]\right] \\
& - E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g_\theta(Y_i|X_i)\, q_\theta(X_{i-1}, X_i)\right)|Y_{\infty:\infty}\right]\right]\Big) \\
+ \sum_{i=1}^{\infty} \Big(& E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g^\epsilon_\theta(Y_i|X_i)\, q_\theta(X_{i-1}, X_i)\right)|Y_{\infty:\infty}\right]\right] \\
& - E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g^\epsilon_\theta(Y_i|X_i)\, q_\theta(X_{i-1}, X_i)\right)|Y_{\infty:\infty}\right]\right]\Big) \\
& + E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta(Y_0|X_0)\, q_\theta(X_{-1}, X_0)\right)|Y_{\infty:\infty}\right]\right] \\
& - E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g^\epsilon_\theta(Y_0|X_0)\, q_\theta(X_{-1}, X_0)\right)|Y_{\infty:\infty}\right]\right].
\end{aligned} \tag{C-93}$$

Proof of Lemma C.13. First we recall that by (A-31), (A-36), (B-59), (B-60), (B-69) and assumptions (A1)-(A5) we have that

$$\nabla_\theta l^\epsilon(\theta) - \nabla_\theta l(\theta) = \lim_{n\to\infty} \frac{1}{n} \left(E_{\theta^*}\left[\nabla_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right] - E_{\theta^*}\left[\nabla_\theta \log p_\theta(Y_1, \ldots, Y_n)\right]\right). \tag{C-94}$$

Next note that by definition it follows that

$$p_\theta(y_1, \ldots, y_n) = p_{\theta^{\epsilon,+}}(Y_{-n+1} = y_1, \ldots, Y_0 = y_n)$$

and thus

$$E_{\theta^*}\left[\nabla_\theta \log p_\theta(Y_1, \ldots, Y_n)\right] = E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_{-n+1}, \ldots, Y_0)\right]. \tag{C-95}$$

Similarly one can show that

$$E_{\theta^*}\left[\nabla_\theta \log p^\epsilon_\theta(Y_1, \ldots, Y_n)\right] = E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{n-1}, \ldots, Y_0)\right]. \tag{C-96}$$

Moreover it follows from (C-86) and (C-87) that for any $l < 0 < k$

$$E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_l, \ldots, Y_{k+1})\right] = E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{l-1}, \ldots, Y_k)\right]. \tag{C-97}$$

It now follows from (C-94), (C-95), (C-96) and (C-97) that

$$\begin{aligned}
\nabla_\theta l^\epsilon(\theta) - \nabla_\theta l(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big(& E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_{-n+i}, \ldots, Y_{i-1})\right] \\
& - E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{-n+i}, \ldots, Y_{i-1})\right]\Big).
\end{aligned} \tag{C-98}$$

Finally we note that in the light of Remark C.8 it follows that results analogous to Lemmas A.3, A.4 and A.7 hold for the conditional laws and expecations of the HMMs defined in (C-86) and (C-87) and for the corresponding likelihood functions $p_{\theta^{\epsilon,+}}(\cdots)$ and $p_{\theta^{\epsilon,-}}(\cdots)$ and thus we can use exactly the same kind of reasoning as was used to prove Lemma A.5 to show that the limit on the right hand side of (C-98) is equal to the right hand side of (C-93). □

Lemma C.14. *Suppose that assumptions (A1)-(A5) and (A7) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Then for any $\theta \in \Theta$ and integers $l \leq k$*

$$\begin{aligned}
\lim_{\epsilon \to 0} \|P_{\theta^{\epsilon,+}}(X_l, \ldots, X_k | Y_{\infty:\infty}) - P_\theta(X_l, \ldots, X_k | Y_{\infty:\infty})\|_{TV} = 0, \\
\lim_{\epsilon \to 0} \|P_{\theta^{\epsilon,-}}(X_l, \ldots, X_k | Y_{\infty:\infty}) - P_\theta(X_l, \ldots, X_k | Y_{\infty:\infty})\|_{TV} = 0, \\
\lim_{\epsilon \to 0} \|P_{\theta^{\epsilon,+}}(X_l, \ldots, X_k | Y_{0:\infty}) - P_\theta(X_l, \ldots, X_k | Y_{0:\infty})\|_{TV} = 0, \\
\lim_{\epsilon \to 0} \|P_{\theta^{\epsilon,-}}(X_l, \ldots, X_k | Y_{\infty:0}) - P_\theta(X_l, \ldots, X_k | Y_{\infty:0})\|_{TV} = 0
\end{aligned} \tag{C-99}$$

for $\overline{\mathbb{P}}_{\theta^*}$ a.s. all doubly infinite sequences $\ldots, Y_{-1}, Y_0, Y_1, \ldots$.

Proof. We will just prove the result for the conditional probabilities $P_{\theta^{\epsilon,+}}(X_l, X_k | Y_{0:\infty})$. The proofs of the other results are identical and we leave it to the reader to fill in the details. First note that by the definition of the total variation norm and by Lemma A.3 we have that for all $r > k$

$$\begin{aligned}
\|P_{\theta^{\epsilon,+}}(X_l, \ldots, X_k | Y_{0:\infty}) - P_{\theta^{\epsilon,+}}(X_l, \ldots, X_k | Y_{0:r})\|_{TV} \leq 4\rho^{r-k} \\
\|P_\theta(X_l, \ldots, X_k | Y_{0:\infty}) - P_\theta(X_l, \ldots, X_k | Y_{0:r})\|_{TV} \leq 4\rho^{r-k}.
\end{aligned} \tag{C-100}$$

Thus in order to prove the result it suffices to prove that

$$\lim_{\epsilon \to 0} \|P_{\theta^{\epsilon,+}}(X_l, \ldots, X_k | Y_{0:r}) - P_\theta(X_l, \ldots, X_k | Y_{0:r})\|_{TV} = 0$$

for every $r > k$.

Thus it is sufficient to show that

$$\lim_{\epsilon \to 0} \sup_{f \in l_\infty} |E_{\theta^{\epsilon,+}}[f(X_l, \ldots, X_k) | Y_{0:r}] - E_\theta[f(X_l, \ldots, X_k) | Y_{0:r}]| = 0 \tag{C-101}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. As in the proof of Lemma B.10 we can express the conditional expectations of interest as

$$\begin{aligned}
&E_{\theta^{\epsilon,+}}[f(X_l, \ldots, X_k) | Y_{0:r}] \\
&= \frac{\int_{\mathcal{X}^{r-l+1}} f(X_{l:k}) g_\theta(Y_0|x_0) \prod_{i=1}^r g_\theta^\epsilon(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r)}{\int_{\mathcal{X}^{r-l+1}} g_\theta(Y_0|x_0) \prod_{i=1}^r g_\theta^\epsilon(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r)}
\end{aligned} \tag{C-102}$$

and

$$\begin{aligned}
&E_\theta[f(X_l, \ldots, X_k) | Y_{0:r}] \\
&= \frac{\int_{\mathcal{X}^{r-l+1}} f(X_{l:k}) \prod_{i=0}^r g_\theta(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r)}{\int_{\mathcal{X}^{r-l+1}} \prod_{i=0}^r g_\theta(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r)}
\end{aligned} \tag{C-103}$$

Furthermore by arguing in the same way as in the proof of (B-82) we have that

$$\begin{aligned}
\int_{\mathcal{X}^{r-l+1}} g_\theta(Y_0|x_0) \prod_{i=1}^r g_\theta^\epsilon(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r) \to \\
\int_{\mathcal{X}^{r-l+1}} \prod_{i=0}^r g_\theta(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r).
\end{aligned} \tag{C-104}$$

Next we note that

$$\begin{aligned}
&\sup_{f \in L_\infty} \left| \int_{\mathcal{X}^{r-l+1}} f(X_{l:k}) g_\theta(Y_0|x_0) \prod_{i=1}^r g_\theta^\epsilon(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r) \right. \\
&\quad \left. - \int_{\mathcal{X}^{r-l+1}} f(X_{l:k}) \prod_{i=0}^r g_\theta(Y_i|x_i) \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r) \right| \\
&\leq \int_{\mathcal{X}^{r-l+1}} g_\theta(Y_0|x_0) \prod_{i=1}^r \left| g_\theta(Y_i|x_i) - g_\theta^\epsilon(Y_i|x_i) \right| \prod_{j=l}^r q_\theta(x_{j-1}, x_j) \overline{\mathbb{P}}_\theta(dx_{j-1}) \mu(dx_j) \cdots \mu(dx_r).
\end{aligned} \tag{C-105}$$

However it follows from (B-77) and dominated convergence theorem that $\overline{\mathbb{P}}_{\theta^*}$ a.s. the right hand side of (C-105) goes to 0 as $\epsilon \to 0$. Equation (C-101) is then a consequence of (C-102)-(C-105). □

Lemma C.15. *Suppose that assumptions (A1)-(A5) and (A7) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Furthermore for any $h \in L^2(\overline{\mathbb{P}}_{\theta^*})$ let*

$$L_\infty^h = \left\{ f : \mathcal{Y} \times \mathcal{X}^2 \to \mathbb{R} : \sup_{x,x' \in \mathcal{X}^2} |f(Y,x,x')| \leq h(Y) \quad \overline{\mathbb{P}}_{\theta^*} a.s. \right\}.$$

Then there exists a finite constant C such that for all $h \in L^2(\overline{\mathbb{P}}_{\theta^*})$, $\theta \in \Theta$, $\epsilon > 0$ and integers i

$$\sup_{f \in L_\infty^h} E_{\theta^*} \left[\left| E_{\theta^{\epsilon,+}} \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] - E_{\theta^{\epsilon,-}} \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] \right| \right] \leq C\epsilon^2 \rho^{|i|} E_{\theta^*} \left[h^2 \right]^{\frac{1}{2}}. \tag{C-106}$$

Moreover, for all $\theta \in \Theta$ there exists a sequence of $\sigma(Y_{-\infty:\infty})$ measurable random finite signed measures $\pi_{\theta,i}$ such that for all $h \in L^2(\overline{\mathbb{P}}_{\theta^*})$, i and $\epsilon > 0$

$$\sup_{f \in L_\infty^h} E_{\theta^*} \left[\left| E_{\pi_{\theta,i}} \left[f(Y_i, x, x') \right] \right| \right] \leq C\rho^{|i|} E_{\theta^*} \left[h^2 \right]^{\frac{1}{2}}. \tag{C-107}$$

and

$$\frac{1}{\epsilon^2} \sup_{f \in L_\infty^h} E_{\theta^*} \Big[\Big| E_{\theta^{\epsilon,+}} \left[f(Y_i, X_{i-1}, X_i | Y_{\infty:\infty} \right] - E_{\theta^{\epsilon,-}} \left[f(Y_i, X_{i-1}, X_i | Y_{\infty:\infty} \right] - \epsilon^2 E_{\pi_{\theta,i}} \left[f(Y_i, x, x') \right] \Big| \Big] \to 0 \tag{C-108}$$

as $\epsilon \to 0$.

Proof. Throughout this proof we shall make extensive use of the following simple fact: for all i and $f \in L_\infty^h$

$$E_{\theta^{\epsilon,+}} \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] = \frac{E_{\theta^{\epsilon,+}} \left[f(Y_i, X_{i-1}, X_i) g_\theta \left(Y_0 | X_0 \right) | Y_{\infty:-1;1:\infty} \right]}{E_{\theta^{\epsilon,+}} \left[g_\theta \left(Y_0 | X_0 \right) | Y_{\infty:-1;1:\infty} \right]}. \tag{C-109}$$

We note that an analogous result holds for the conditional expectations $E_{\theta^{\epsilon,-}} \left[\cdot | Y_{\infty:\infty} \right]$. We also note that by a simple application of Taylor's theorem we have that there exist constants K_1 and K_2 such that

$$g_\theta^\epsilon \left(y | x \right) = g_\theta \left(y | x \right) + K_1 \epsilon^2 \nabla_y^2 g_\theta \left(y | x \right) + r(\theta, \epsilon, x, y) \tag{C-110}$$

where the remainder term $r(\theta, \epsilon, x, y)$ is bounded by

$$K_2 \epsilon^2 \sup_{z \in B_0^\epsilon} \left| \nabla_y^2 g_\theta \left(y + z | x \right) - \nabla_y^2 g_\theta \left(y | x \right) \right| \tag{C-111}$$

for all x, y and ϵ.

We shall show that the lemma holds true with the signed measures $\pi_{\theta,i}$ defined such that for all θ, i and $f \in L_\infty$

$$E_{\pi_{\theta,i}} \left[f(Y_i, X_{i-1}, X_i) \right] = E_\theta \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] E_\theta \left[\frac{K_1 \nabla_y^2 g_\theta \left(Y_0 | X_0 \right)}{g_\theta \left(Y_0 | X_0 \right)} | Y_{\infty:\infty} \right] - E_\theta \left[f(Y_i, X_{i-1}, X_i) \frac{K_1 \nabla_y^2 g_\theta \left(Y_0 | X_0 \right)}{g_\theta \left(Y_0 | X_0 \right)} | Y_{\infty:\infty} \right]. \tag{C-112}$$

We have by (C-109) that for any $h \in L^2(\overline{\mathbb{P}}_{\theta^*})$ and all i

$$\begin{aligned} &\sup_{f \in L_\infty^h} E_{\theta^*} \left[\left| E_{\theta^{\epsilon,+}} \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] - E_{\theta^{\epsilon,-}} \left[f(Y_i, X_{i-1}, X_i) | Y_{\infty:\infty} \right] \right| \right] \\ &= \sup_{f \in L_\infty^h} E_{\theta^*} \left[\left| \frac{E_{\theta^{\epsilon,+}} \left[f_i g_\theta | Y_{\infty:-1;1:\infty} \right]}{E_{\theta^{\epsilon,+}} \left[g_\theta | Y_{\infty:-1;1:\infty} \right]} - \frac{E_{\theta^{\epsilon,-}} \left[f_i g_\theta^\epsilon | Y_{\infty:-1;1:\infty} \right]}{E_{\theta^{\epsilon,-}} \left[g_\theta^\epsilon | Y_{\infty:-1;1:\infty} \right]} \right| \right] \end{aligned} \tag{C-113}$$

where we have used the shorthand $g_\theta = g_\theta(Y_0|X_0)$, $f_i = f(Y_i, X_{i-1}, X_i)$ etc.. Simple algebra and (C-92) show that the right hand side of (C-113) is bounded by

$$\begin{aligned}
&\sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|\frac{E_{\theta^{\epsilon,+}}[f_i g_\theta|Y_{\infty:-1;1:\infty}]}{E_{\theta^{\epsilon,+}}[g_\theta|Y_{\infty:-1;1:\infty}]} - \frac{E_{\theta^{\epsilon,-}}[f_i g_\theta^\epsilon|Y_{\infty:-1;1:\infty}]}{E_{\theta^{\epsilon,-}}[g_\theta|Y_{\infty:-1;1:\infty}]}\right.\right.\\
&\left.\left.+\frac{E_{\theta^{\epsilon,-}}[f_i g_\theta^\epsilon|Y_{\infty:-1;1:\infty}]}{E_{\theta^{\epsilon,-}}[g_\theta^\epsilon|Y_{\infty:-1;1:\infty}]}\left(\frac{E_{\theta^{\epsilon,-}}[g_\theta^\epsilon|Y_{\infty:-1;1:\infty}] - E_{\theta^{\epsilon,-}}[g_\theta|Y_{\infty:-1;1:\infty}]}{E_{\theta^{\epsilon,-}}[g_\theta|Y_{\infty:-1;1:\infty}]}\right)\right|\right]\\
&= \sup_{f\in L_\infty^h} E_{\theta^*}\left[\left| -\epsilon^2 E_{\theta^{\epsilon,+}}\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}f_i|Y_{\infty:\infty}\right]\right.\right.\\
&\left.\left.+E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}]\left(\epsilon^2 E_{\theta^{\epsilon,+}}\left[K_1\frac{1}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] + E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right]\right)\right|\right] \qquad \text{(C-114)}
\end{aligned}$$

where K_1 is as in (C-110), $\nabla_y^2 g_\theta = \nabla_y^2 g_\theta(Y_0|X_0)$ and $r(\theta,\epsilon) = r(\theta,\epsilon,X_0,Y_0)$ and $r(\theta,\epsilon,x,y)$ is as in (C-110).

The next step is to bound the terms on the right hand side of (C-114). First note that by Hölder's inequality, Lemma A.4 and Remarks C.8 and C.9 we have that

$$\begin{aligned}
&\sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}]E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[f_i\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right]\right|\right]\\
&= \sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|\left(E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}] - E_{\theta^{\epsilon,+}}[f_i|Y_{\infty:\infty}]\right)E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right]\right.\right.\\
&\left.\left.+E_{\theta^{\epsilon,+}}[f_i|Y_{\infty:\infty}]E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[f_i\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right]\right|\right]\\
&\le 8\rho^{|i|-1}E_{\theta^*}\left[\sup_{x\in\mathcal{X}}|r(\theta,\epsilon,x,Y_0)|^2\right]^{\frac{1}{2}}E_{\theta^*}\left[h^2\right]^{\frac{1}{2}}. \qquad \text{(C-115)}
\end{aligned}$$

Similarly one may show that

$$\begin{aligned}
&\sup_{f\in L_\infty^h}\epsilon^2 E_{\theta^*}\left[\left|E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}]E_{\theta^{\epsilon,+}}\left[K_1\frac{1}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right]\right|\right],\\
&\sup_{f\in L_\infty^h}\epsilon^2 E_{\theta^*}\left[\left|E_\theta[f_i|Y_{\infty:\infty}]E_\theta\left[K_1\frac{1}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_\theta\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right]\right|\right]\\
&\le 8\epsilon^2\rho^{|i|-1}E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\left|K_1\frac{\nabla_y^2 g_\theta(Y_0|x)}{g_\theta(Y_0|x)}\right|\right]E_{\theta^*}\left[h^2\right]^{\frac{1}{2}}. \qquad \text{(C-116)}
\end{aligned}$$

Equations (C-106) and (C-107) now follow from (C-112), (C-114), (C-115) and (C-116). We have by (C-111) and (C-115) and assumption (A7) that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}\sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}]E_{\theta^{\epsilon,+}}\left[\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[f_i\frac{r(\theta,\epsilon)}{g_\theta}|Y_{\infty:\infty}\right]\right|\right] = 0$$

and hence it follows from (C-112) and (C-114)that in order to prove (C-108) it is sufficient to show that

$$\begin{aligned}
&\lim_{\epsilon\to 0}\sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|E_\theta\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_\theta[f_i|Y_{\infty:\infty}]E_\theta\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right]\right.\right.\\
&\left.\left.+E_{\theta^{\epsilon,-}}[f_i|Y_{\infty:\infty}]E_{\theta^{\epsilon,+}}\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right]\right|\right] = 0. \qquad \text{(C-117)}
\end{aligned}$$

Finally we note that

$$\begin{aligned}
&\lim_{\epsilon\to 0}\sup_{f\in L_\infty^h} E_{\theta^*}\left[\left|E_\theta\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,+}}\left[K_1\frac{f_i}{g_\theta}\nabla_y^2 g_\theta|Y_{\infty:\infty}\right]\right|\right]\\
&\le \lim_{\epsilon\to 0}K_1 E_{\theta^*}\left[|h|\sup_{x\in\mathcal{X}}\left|\frac{1}{g_\theta}\nabla_y^2 g_\theta\right|\left\|P_\theta(X_{i-1},X_1|Y_{\infty:\infty}) - P_{\theta^{\epsilon,+}}(X_{i-1},X_1|Y_{\infty:\infty})\right\|_{TV}\right]
\end{aligned}$$

which is equal to zero by Lemma C.14. A similar result holds for the remaining terms in (C-117) and thus the proof is complete. □

Proof of Lemma C.12. It follows from (C-93) that in order to prove the lemma it is sufficient to show that there exists a constant $C < \infty$ and a constant $0 < \rho < 1$ such that

$$\begin{aligned}&\Big| E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g_\theta\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\right)|Y_{\infty:\infty}\right]\right]\Big| \leq C\epsilon^2\rho^{|i|}\end{aligned} \tag{C-118}$$

for all $i \neq 0$ and that there exists a sequence of vectors $\ldots, V_{-1}, V_0, V_1, \ldots$ such that

$$|V_i| \leq C\rho^{|i|} \tag{C-119}$$

for all i,

$$\begin{aligned}&E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] = \epsilon^2 V_1 + o(\epsilon^2)\end{aligned} \tag{C-120}$$

and

$$\begin{aligned}&E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g_\theta\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\right)|Y_{\infty:\infty}\right]\right] = \epsilon^2 V_i + o(\epsilon^2)\end{aligned} \tag{C-121}$$

for all $i \neq 0$.

We shall show that (C-118)-(C-121) hold with the sequence of vectors

$$V_i \triangleq E_{\pi_{\theta,i}}\left[\nabla_\theta \left(\log g_\theta\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\right)\right]$$

where for all i the signed measure $\pi_{\theta,i}$ is as in Lemma C.15. We start by noting that, with the sequence of vectors V_i defined above, equations (C-118), (C-119) and (C-121) are immediate consequences of Lemma C.15. Furthermore since

$$\begin{aligned}&E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &= E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &+ E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,-}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right]\end{aligned} \tag{C-122}$$

it follows from Lemma C.15 that in order to prove (C-120) it is sufficient to show that

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}\Big(E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right] \\ &- E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\nabla_\theta \left(\log g^\epsilon_\theta\left(Y_0|X_0\right) q_\theta\left(X_{-1},X_0\right)\right)|Y_{\infty:\infty}\right]\right]\Big)\end{aligned} \tag{C-123}$$

exists and is finite. We first observe that by (C-109) we have that (C-123) is equal to

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}E_{\theta^*}\left[E_{\theta^{\epsilon,+}}\left[\frac{\nabla_\theta g_\theta\left(Y_0|X_0\right)}{g_\theta\left(Y_0|X_0\right)} - \frac{\nabla_\theta g^\epsilon_\theta\left(Y_0|X_0\right)}{g^\epsilon_\theta\left(Y_0|X_0\right)}|Y_{\infty:\infty}\right]\right].$$

The result now follows from (C-110), assumptions (A2), (A3), (A5) and (A7), the dominated convergence theorem and the fact that

$$\begin{aligned}&\frac{\nabla_\theta g_\theta\left(Y_0|X_0\right)}{g_\theta\left(Y_0|X_0\right)} - \frac{\nabla_\theta g^\epsilon_\theta\left(Y_0|X_0\right)}{g^\epsilon_\theta\left(Y_0|X_0\right)} = \left(\frac{\nabla_\theta g_\theta\left(Y_0|X_0\right)}{g_\theta\left(Y_0|X_0\right)} - \frac{\nabla_\theta g^\epsilon_\theta\left(Y_0|X_0\right)}{g_\theta\left(Y_0|X_0\right)}\right) \\ &+ \frac{\nabla_\theta g^\epsilon_\theta\left(Y_0|X_0\right)}{g^\epsilon_\theta\left(Y_0|X_0\right)}\left(\frac{g^\epsilon_\theta\left(Y_0|X_0\right) - g_\theta\left(Y_0|X_0\right)}{g_\theta\left(Y_0|X_0\right)}\right).\end{aligned}$$

□

Appendix D: Proof of Theorem 3.5

The proof of Theorem 3.5 is very similar to that of Theorem 3.4 and so we shall just provide a sketch of the most important points. Firstly, given any $h \in L^2(\overline{\mathbb{P}}_{\theta^*})$, recall the definition of L^h_∞ in Lemma C.15. Using exactly the same methods as were used to prove that lemma one can show that for all θ and ϵ there exist a finite constant C and a sequence of signed measures $\pi_{\theta,i,\epsilon}$ such that

$$E_{\theta^{\epsilon,+}}\left[f(Y_i, X_{i-1}, X_i)|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon,-}}\left[f(Y_i, X_{i-1}, X_i)|Y_{\infty:\infty}\right] = \epsilon E_{\pi_{\theta,i,\epsilon}}\left[f(Y_i, x, x')\right]$$

and

$$\sup_{f \in L^h_\infty} E_{\theta^*}\left[\left|E_{\pi_{\theta,i}}\left[f(Y_i, x, x')\right]\right|\right] \leq C\rho^{|i|} E_{\theta^*}\left[h^2\right]^{\frac{1}{2}}.$$

One can then use the above result along with Lemmas C.13 and C.14 to show that

$$\nabla_\theta l^\epsilon(\theta^*) - \nabla_\theta l(\theta^*) = O(\epsilon). \tag{D-124}$$

Theorem 3.5 can then be proved from (D-124) in exactly the same way that Theorem 3.4 is proved using (C-83). We leave the details to the reader.

References

G. Biau, F. Cerou, and A. Guyader. New insights into approximate Bayesian computation. *Ann. Inst. Henri Poincare*, 51:376–403, 2015.

J. Borwanker, G. Kallianpur, and B.L.S. Prakasa Rao. The bernstein-von mises theorem for markov processes. *Ann. Math. Stat.*, 42:1241–1253, 1971.

O. Cappé, T. Rydén, and E. Moulines. *Inference in Hidden Markov Models*. Springer-Verlag: New York, 2005.

T.A. Dean, S.S. Singh, A. Jasra, and G.W. Peters. Parameter estimation for hidden markov models with intractable likelihoods. *Scand. J. Statist.*, 41:970–987, 2014.

P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag: New York., 2004.

R. Douc, E. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *Ann. Statist.*, 32:2254–2304, 2004.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. CUP: Cambridge., 1998.

P. Fearnhead and D. Prangle. Semi-automatic approximate bayesian computation. *J. R. Statist. Soc. B*, 74: 419-474–2304, 2012.

J. Felsenstein and G.A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.

D. T. Frazier, G. Martin, C. Robert, and J. Rousseau. Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105:593–607., 2018.

A. Jasra, S. S. Singh, J. S. Martin, and E. McCoy. Filtering via approximate bayesian computation. *Stat. Comput,*, 22:1223–1237, 2012..

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with arch models. *Rev. Econom. Stud.*, 65:361–393, 1998.

W. Li, and P. Fearnhead. On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, 105:285–299., 2018.

J. McKinley, C. Cook, and R. Deardon. Inference for epidemic models without likelihooods. *Intl. J. Biostat.*, 5, 2009.

G. Peters, M. W. Wüthrich, and P. Shevchenko. Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance Math. Econom.*, 47:36–51., 2010.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and P. Feldman. Population growth of human y chromosome microsatellites. *Mol. Biol. Evol.*, 16:1791–1798., 1999.

O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. USA*, 106:10576–10581., 2009.

S. Tavre, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145:505–518., 1997.

D. Volný. Approximating martingales and the central limit theorem for strictly stationary processes. *Stoch. Proc. Appl.*, 44:41–74, 1993.

M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC, 1995.

R. L. Wheeden and A. Zygmund. *Measure and Integral; An Introduction to Real Analysis*. Marcel Dekker, New York., 1977.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25., 1982.

S. Yildirim, S. S. Singh, T. A. Dean and A. Jasra. Parameter estimation in hidden Markov models with intractable likelihoods via sequential Monte Carlo. *J. Comp. Graph. Stat.*, 24:846–865, 2015.