ASYMPTOTIC BEHAVIOUR OF APPROXIMATE BAYESIAN ESTIMATORS

By Thomas. A. Dean* and Sumeetpal S. Singh*

University of Cambridge

Although approximate Bayesian computation (ABC) has become a popular technique for performing parameter estimation when the likelihood functions are analytically intractable there has not as yet been a complete investigation of the theoretical properties of the resulting estimators. In this paper we give a theoretical analysis of the asymptotic properties of ABC based parameter estimators for hidden Markov models and show that ABC based estimators satisfy asymptotically biased versions of the standard results in the statistical literature.

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 likeilhood 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 expres-

*T.A. Dean and S.S. Singh's research is funded by the Engineering and Physical Sciences Research Council (EP/G037590/1) whose support is gratefully acknowledged. This

AMS 2000 subject classifications: Primary 62M09; secondary 62B99, 62F12, 65C05

Keywords and phrases: Parameter Estimation, Hidden Markov Model, Maximum Likelihood, Approximate Bayesian Computation, Sequential Monte Carlo

sion 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 recently 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]. See also [Sisson and Fan, to be published] for a review on computational methodology. 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

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

Typically the probabilities (1) are themselves estimated using Monte Carlo techniques. A particularly appealing feature of the ABC methodology is that, despite the methods name, the resulting approximations to the likelihoods may then be used in any likelihood based parameter inference methodology the user desires.

Intuitively, the justification for the ABC approximation is that for sufficiently small ϵ

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

where $B_{\hat{Z}}^\epsilon$ denotes the d-ball of radius ϵ around the point $\hat{Z}$ and thus the probabilities (1) will provide a good approximation to the likelihood, up to the value of some renormalising factor which is independent of θ and hence can be ignored.

Clearly in general the estimators based on ABC approximations to the likelihood will differ from those based on the exact value of the likelihood function, however although the use of ABC has become commonplace there has to date been little investigation of the precise nature of the theoretical properties of ABC based estimators. One notable exception is [Fearnhead and Prangle, 2010]. In this paper the authors consider the problem of finding the optimal choice, for a given data set, of summary statistic and ϵ in order to minimise the mean square error of the resulting ABC posterior distribution on parameter space. Unfortunately the resulting optimal choice of summary statistic involves computing a conditional expectation w.r.t. the unknown posterior distribution and hence it can only be computed approximately and not exactly. Further the analysis is done only for fixed size data sets and the asymptotic properties of the ABC estimator are left unexplored.

An alternative approach is taken in [Dean et al., 2010] 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., 2010] 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. However they do not establish any asymptotic normality type properties of this estimator and there are as yet no analogous results for the ABC Bayesian parameter estimator. The aim of this paper is to bridge these theoretical gaps by showing that the standard results in likelihood based parameter estimation, that is to say asymptotic consistency, asymptotic normality and Bernstein-von Mises type theorems, also hold in a suitably modified version for parameter estimators based on ABC approximations to the likelihood. In the next section we provide an outline of the approach that we shall take to proving these results.

1.1. *Contributions and Structure.* In this paper we shall study the asymptotic behaviour of ABC parameter estimators when used to perform inference for hidden Markov models. This will be convenient as (as we will show) the Markovian context imbues the ABC approximations 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 the rest of this paper we shall consider performing ABC based parameter estimation for HMMs using the following specialization of the standard ABC likelihood approximation (1), proposed in [Jasra et al., 2010], 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

$$\mathbb{P}_\theta\left(Y_1 \in B^\epsilon_{\hat{Y}_1}, \ldots, Y_n \in B^\epsilon_{\hat{Y}_n}\right) \tag{2}$$

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., 2010]) implementation of the ABC approximation. Furthermore the resulting approximation has a structure which is particularly tractable to mathematical analysis.

The purpose of this paper is to show that one can prove results about the asymptotic behaviour of ABC based parameter estimators analogous to the standard results in the literature concerning the asymptotic behaviour of estimators based on the exact value of the likelihood. In particular we show that one can develop a theoretical justification of ABC parameter estimation procedures based on their large sample properties analogous to those provided for Bayesian and maximum likelihood based procedures by the standard Bernstein-von Mises and asymptotic consistency and normality results respectively. Our approach is based on the observation in [Dean et al., 2010] that ABC can be considered as performing parameter estimation using the likelihoods of a collection of perturbed HMMs which suggests that in some sense ABC based parameter estimators should inherit their behaviour from the standard statistical estimators. We first show that unlike the MLE, which is asymptotically consistent, the ABC MLE estimator has an innate asymptotic bias in the sense that the value of the estimator converges to the wrong point in parameter space as the number of observations tends to infinity. Moreover we show that asymptotically the ABC MLE is normally distributed around this biased estimate. Secondly we show that the resulting ABC Bayesian posterior distributions obey a Bernstein-von Mises type theorem but that the posteriors are again asymptotically biased in the sense that as the number of data points goes to infinity the resulting posterior distributions concentrate about the limit of the ABC MLE rather than the true parameter value. Finally we show that the size of the asymptotic bias of both the ABC Bayesian and ABC MLE estimators goes to zero as ϵ tends to zero and under mild regularity conditions we obtain sharp rates for this convergence. Together these results show that ABC based parameter estimates are asymptotically biased with a bias which can be made arbitrarily small by taking a suitable choice of ϵ and thus provide a rigorous justification for performing statistical inference based on ABC approximations to the likelihood.

We note that the results in this paper extend those in [Dean et al., 2010] in several ways. In particular we provide a much sharper analysis of the ABC MLE than that contained in [Dean et al., 2010]. The crucial difference between the current paper and [Dean et al., 2010] is that it is not possible using the techniques of [Dean et al., 2010] to show that the ABC MLE has a unique limit point. In contrast, in this paper we show that for sufficiently small values of ϵ the ABC MLE has one and only one limit point. This then enables us to extend the scope of the analysis in [Dean et al., 2010] to include asymptotic normality results for the ABC MLE and Bernstein-von Mises type results for ABC based Bayesian estimators.

This paper is structured as follows. In Section 2 the notation and assumptions are given and in Section 3 we present our main results concerning the asymptotic behaviour of ABC. The article is summarized in Section 4 and supporting technical lemmas and proofs of some of the theoretical results are housed in the four 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_y^ϵ denote the closed ball of radius ϵ centered on the point y and let $\mathcal{U}_{B_y^\epsilon}$ denote the uniform distribution on B_y^ϵ. 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 and for any matrix M we shall let $\|M\|$ denote the Frobenius norm. We note that although using $|\cdot|$ 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 gradiant 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 $d_1 \times d_2 \times d_3$ matricies 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. Further given vectors u, v, w we shall let $u * v$ and $u * v * w$ denote the outer products of u, v and u, v, w and u^{*2} and u^{*3} denote the outer products $u * u$ and $u * u * u$ 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 $\Theta \in \mathbb{R}^d$. 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.

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 $\overline{\mathbb{E}}_\theta$ 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 $\mathcal{X}^n$.

For any two probability measures μ_1, μ_2 on a measurable space $(E, \mathscr{E})$ we let $\|\mu_1 - \mu_2\|_{TV}$ denote the total variation distance between them. For all $p \in [1, \infty)$ we let $L_p(\mu)$ denote the set of real valued measurable functions satisfying $\int |f(x)|^p \, \mu(dx) < \infty$.

Finally we note that when writing the likelihood $p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n)$ of a sequence of observations $\hat{Y}_1, \ldots, \hat{Y}_n$ we shall typically suppress the dependence of the likelihood function on the the initial condition of the hidden state of the process unless we specifically need to refer to it in which case we shall write the likelihood as $p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n | X_0 = x)$.

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

(**A1**) The parameter vector θ^* belongs to the interior of Θ and $\theta = \theta^*$ if and only if $\overline{\mathbb{P}}_\theta(\ldots, Y_{-1}, Y_0, Y_1, \ldots) = \overline{\mathbb{P}}_{\theta^*}(\ldots, Y_{-1}, Y_0, Y_1, \ldots)$.

(**A2**) 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.

(**A3**) There exist constants $\underline{c}_1, \overline{c}_1 \in (0, \infty)$ such that for every $y \in \mathcal{Y}$, $x, x' \in \mathcal{X}$, $\theta \in \Theta$

$$\begin{aligned} \underline{c}_1 \leq q_\theta(x, x') &\leq \overline{c}_1, \\ g_\theta(y|\, x) &\leq \overline{c}_1. \end{aligned} \tag{3}$$

(**A4**) There exists a constant $\overline{c}_2 \in (0, \infty)$ such that for every $y \in \mathcal{Y}$, $x, x' \in \mathcal{X}$, $\theta \in \Theta$

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

(**A5**) For all $\theta \in \Theta$

$$0 < \int_{\mathcal{X}} g_\theta\,(y|x)\, \mu(dx) < \infty \tag{4}$$

for all $y \in \mathcal{Y}$.

(**A6**) For any $K > 0$

$$E_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla_\theta \log g_\theta\left(Y+z|x\right)\right\|^3\right],$$

$$E_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla_\theta^2 \log g_\theta\left(Y+z|x\right)\right\|^2\right], \tag{5}$$

$$E_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\sup_{z\in B_0^K}\left\|\nabla_\theta^3 \log g_\theta\left(Y+z|x\right)\right\|\right] \leq \infty.$$

REMARK 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 parameter estimators inherit their properties from standard statistical estimators. 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 particluar we shall require slightly stronger conditions on the differentiability of the conditonal densities $g_\theta(y|x)$ but slightly weaker conditions on their integrability.*

REMARK 2. *In general assumptions (A3)-(A6) will hold when the state space $\mathcal{X}$ is compact. However we expect that the behaviours predicted by Theorems 2, 3, 4 and 5 will provide a good qualitative guide to the behaviour of ABC MLE in practice even in cases where the underlying HMMs do not satisfy these assumptions.*

3. Approximate Bayesian Computation.

3.1. *Structure of ABC Estimators.* Suppose that a collection of HMMs

$$\{X_k, Y_k\}_{k\geq 0} \tag{6}$$

parameterised by some $\theta \in \Theta$ are given. For any sequence of observations $\hat{Y}_1, \ldots, \hat{Y}_n$ for $\theta \in \Theta$ let $p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n)$ denote the likelihood of the observations under the corresponding HMM (6). Following [Jasra et al., 2010] we consider approximating $p_\theta(\hat{Y}_1, \ldots, \hat{Y}_n)$ by the ABC approximation,

$$\mathbb{P}_\theta\left(Y_1 \in B^\epsilon_{\hat{Y}_1}, \ldots, Y_n \in B^\epsilon_{\hat{Y}_n}\right)$$
$$= \int_{\mathcal{X}^{n+1}\times\mathcal{Y}^n}\left[\prod_{k=1}^n q_\theta(x_{k-1}, x_k)\mathbb{I}_{B^\epsilon_{\hat{Y}_k}}(y_k) g_\theta(y_k|x_k)\right]\pi_0(dx_0)\,\mu(dx_{1:n})\nu(dy_{1:n}). \tag{7}$$

The purpose of this paper is to analyse the asymptotic properties of likelihood based parameter estimators implemented using the ABC approximate likelihoods (7). The key to our analysis is the following observation, see [Dean et al., 2010] for more details;

$$\int_{\mathcal{X}^{n+1}\times\mathcal{Y}^n}\left[\prod_{k=1}^{n} q_\theta(x_{k-1},x_k)\mathbb{I}_{B^\epsilon_{\hat{Y}_k}}(y_k)g_\theta(y_k|x_k)\right]\pi_0(dx_0)\,\mu(dx_{1:n})\nu(dy_{1:n})$$

$$\propto\int_{\mathcal{X}^{n+1}}\left[\prod_{k=1}^{n} q_\theta(x_{k-1},x_k)g^\epsilon_\theta(\hat{Y}_k|x_k)\right]\pi_0(dx_0)\mu(dx_{1:n}) \tag{8}$$

where

$$g^\epsilon_\theta(y|x)=\frac{1}{\nu\left(B^\epsilon_y\right)}\int_{B^\epsilon_y} g_\theta(y'|x)\,\nu(dy'). \tag{9}$$

The crucial point is that the quantity $g^\epsilon_\theta(y|x)$ defined in (9) is the density of the measure obtained by convolving the measure corresponding to $g_\theta(y|x)$ with $\mathcal{U}_{B^\epsilon_0}$ where the density is taken w.r.t. the new dominating measure obtained by convolving ν with $\mathcal{U}_{B^\epsilon_0}$. One can then immediately see that the quantities $q_\theta(x,x')$ and $g^\epsilon_\theta(y|x)$ appearing in (8) are the transition kernels and conditional laws respectively for a perturbed HMM $\{X_k,Y^\epsilon_k\}_{k\geq 0}$ defined such that it is equal in law to the process

$$\{X_k,Y_k+\epsilon Z_k\}_{k\geq 0} \tag{10}$$

where $\{X_k,Y_k\}_{k\geq 0}$ is the original HMM and the $\{Z_k\}_{k\geq 0}$ are an i.i.d. sequence of $\mathcal{U}_{B^1_0}$ distributed random variables.

3.2. *Theoretical Results.* It follows that performing statistical inference using the ABC approximations to the likelihood is equivalent to performing inference using a misspecified collection of models. It is well known (see for example [White, 1982]) that this will in general lead to biased estimates of the true parameter value. In the rest of this paper we shall investigate the theoretical consequences of this for ABC based parameter estimators.

We start by showing that almost surely the ABC MLE will converge, with increasing sample size, to a given point in parameter space that is not equal to the true parameter value (more generally the set of accumulation points will belong to a given subset of parameter space) and hence that the ABC MLE is asymptotically biased (Theorem 2). Further, we show that these accumulation points must lie in some neighbourhood of the true parameter value and that the size of this neighbourhood shrinks to zero as ϵ goes to zero. Next we show that for sufficiently small values of ϵ the ABC MLE has a unique limit point and that asymptotically the ABC MLE is normally distributed about this point with a variance that is proportional to $\frac{1}{n}$ (Theorem 3). Third we show that aymptotically the ABC Bayesian posterior converges to that of a Normal random variable, centered on the location of the ABC MLE and with variance again proportional to $\frac{1}{n}$ (Theorem 4). Finally we show that under certain Lipschitz conditions one can obtain a rate for the decrease in the size of the asymptotic bias of the ABC parameter estimators (Theorem 5).

These results show that the error of ABC based parameter estimators may be decomposed into two parts. A bias component whose size depends on ϵ and a variance component whose size is proportional to $\frac{1}{\sqrt{n}}$. Furthermore they show that the size of the bias can be made arbitrarily small by a suitable choice of ϵ. Thus taken together the results show that the accuracy of estimators based on ABC approximations to the likelihood can be made to be arbitrarily close to that of estimators based on the exact value of the likelihood, providing a rigourous mathematical justification for the ABC methodology.

We note that there are two important technical issues that arise in the proofs of these results. Firstly, as noted in [Dean et al., 2010], one cannot simply analyse the behaviour of the ABC MLE by extending the parameter space Θ to include ϵ and then applying standard results from the theory of MLE because the perturbed likelihoods $g_\theta^\epsilon(y|x)$ are in some sense insufficiently continuous. Instead one has to establish that in some sense the Lebesgue differentiation theorem still holds upon taking asymptotic limits.

Secondly we note that because the dominating measures of the original and perturbed HMMs are no longer necessarily mutually absolutely continuous with respect to each other we can no longer take the standard approach to analysing likelihood based estimators by studying the limits of

$$\lim_{n\to\infty}\frac{1}{n}\log p_\theta^\epsilon(\hat{Y}_1,\ldots,\hat{Y}_n)$$

and interpreting them in terms of Kullback-Leibler distances. To avoid this problem we instead show that for any ϵ the relative mean log likelihood surfaces (considered as functions of θ)

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

almost surely converge to some limiting surface $l^\epsilon(\theta)$. The behaviour of ABC based parameter estimators can then be understood by examining the behaviour of the corresponding limiting log likelihood surfaces. The key result in doing so is the following whose proof is deferred until Appendix B.

THEOREM 1. *Suppose that one has a collection of HMMs parameterized by some parameter vector $\theta \in \Theta$ that satisfy assumptions (A1)-(A6). For any $\epsilon \geq 0$ let $p^{\epsilon}_{\theta}(\cdots)$ denote the likelihood function w.r.t. the perturbed HMMs (10) (and where by definition we let $p^{0}_{\theta}(\cdots)$ denote the likelihood function of the original HMM (6)). Let data $\hat{Y}_1, \ldots, \hat{Y}_n$ generated by the HMM corresponding to an unknown parameter vector θ^* be given. Then for every $\epsilon \geq 0$ there exists a twice continuously differentiable function $l^{\epsilon}(\theta) : \Theta \to \mathbb{R}$ such that for all $x \in \mathcal{X}$ one has that $\bar{\mathbb{P}}_{\theta^*}$ a.s.*

(11)

$$\begin{aligned}
\frac{1}{n}\Big(\log p^{\epsilon}_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n|X_0 = x) - \log p^{\epsilon}_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n)|X_0 = x\Big) &\to l^{\epsilon}(\theta) \\
\frac{1}{n}\nabla_{\theta}\Big(\log p^{\epsilon}_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n|X_0 = x) - \log p^{\epsilon}_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n|X_0 = x)\Big) &\to \nabla_{\theta} l^{\epsilon}(\theta) \\
\frac{1}{n}\nabla^2_{\theta}\Big(\log p^{\epsilon}_{\theta}(\hat{Y}_1, \ldots, \hat{Y}_n|X_0 = x) - \log p^{\epsilon}_{\theta^*}(\hat{Y}_1, \ldots, \hat{Y}_n)|X_0 = x\Big) &\to \nabla^2_{\theta} l^{\epsilon}(\theta)
\end{aligned}$$

uniformly in θ.

Furthermore $l^{\epsilon}(\theta), \nabla_{\theta} l^{\epsilon}, \nabla^2_{\theta} l^{\epsilon} \to l^{0}(\theta), \nabla_{\theta} l^{0}, \nabla^2_{\theta} l^{0}$ as $\epsilon \to 0$, where the convergence is again uniform in θ.

We can now use Theorem 1 to analyse ABC based parameter estimators by comparing their the asymptotic behaviour (encapsulated in the surfaces $l^{\epsilon}(\theta)$) to the asymptotic behaviour of estimators based on using the true value of the likelihood (which is encapsulated in the surface $l^{0}(\theta)$). we shall start by analysing the behaviour of the ABC MLE which we formally define below.

PROCEDURE 1 (ABC MLE). *Given $\epsilon > 0$ and data $\hat{Y}_1, \ldots, \hat{Y}_n$, estimate θ^* with*

$$\hat{\theta}^{\epsilon}_{n} = \arg\max_{\theta \in \Theta} \mathbb{P}_{\theta}\left(Y_1 \in B^{\epsilon}_{\hat{Y}_1}, \ldots, Y_n \in B^{\epsilon}_{\hat{Y}_n}\right). \tag{12}$$

Using Theorem 1 we can now establish the following biased asymptotic consistency and normality type properties of the ABC MLE whose proofs are deferred to Appendix C.

THEOREM 2. *Suppose that one has a collection of HMMs parameterized by some parameter vector $\theta \in \Theta$ that satisfy assumptions (A1)-(A6). Let data $\hat{Y}_1, \ldots, \hat{Y}_n$ generated by the HMM corresponding to an unknown parameter vector θ^* be given and suppose that we use the ABC MLE to estimate the value of θ^*. Then for every $\epsilon > 0$ there exists a collection of sets*

$\mathcal{T}^{\epsilon}$ such that for all initial conditions X_0 the set of accumulation points of the ABC MLE $\hat{\theta}^{\epsilon}_n$ lies $\bar{\mathbb{P}}_{\theta^*}$ a.s. in $\mathcal{T}^{\epsilon}$ and

$$\lim_{\epsilon \to 0} \sup_{\theta \in \mathcal{T}^{\epsilon}} |\theta - \theta^*| = 0. \tag{13}$$

Furthermore let $l^0(\theta)$ be as in Theorem 1. If $\nabla^2_{\theta} l^0(\theta^*)$ is strictly negative definite then for sufficiently small values of ϵ the set $\mathcal{T}^{\epsilon}$ consists of a singleton $\theta^{*,\epsilon}$.

Remark 3. *The quantity $-\nabla^2_{\theta} l^0(\theta^*)$ is equal to the asymptotic Fisher information I of the HMM. For more details see [Douc et al., 2004].*

Theorem 3. *Suppose that one has a collection of HMMs parameterized by some parameter vector $\theta \in \Theta$ that satisfy assumptions (A1)-(A6) and that $\nabla^2_{\theta} l^0(\theta^*)$ is strictly negative definite where $l^0(\theta)$ is as in Theorem 1. Let data $\hat{Y}_1, \ldots, \hat{Y}_n$ generated by the HMM corresponding to an unknown parameter vector θ^* be given and suppose that we use the ABC MLE to estimate the value of θ^*. Then for sufficiently small values of ϵ there exists strictly positive definite matricies $J_{\epsilon}, I_{\epsilon}$ such that $\bar{\mathbb{P}}_{\theta^*}$ a.s.*

$$\sqrt{n}\left(\hat{\theta}_{n,\epsilon} - \theta^{*,\epsilon}\right) \to N(0, I_{\epsilon}^{-1} J_{\epsilon} I_{\epsilon}^{-1}). \tag{14}$$

Furthermore $J_{\epsilon}, I_{\epsilon} \to I$ as $\epsilon \to 0$ where I is as in Remark 3.

Next we consider the properties of the ABC Bayesian parameter estimator which we define below.

Procedure 2 (ABC Bayesian Estimator). *Given $\epsilon > 0$ a prior distribution π_0 and data $\hat{Y}_1, \ldots, \hat{Y}_n$ estimate θ^* via the ABC posterior*

$$\pi^{\epsilon}_n \propto \mathbb{P}_{\theta}\left(Y_1 \in B^{\epsilon}_{\hat{Y}_1}, \ldots, Y_n \in B^{\epsilon}_{\hat{Y}_n}\right) \pi_0. \tag{15}$$

Given Theorem 1 we can easily see that the ABC Bayesian estimator satisfies the following Bernstein-Von Mises type theorem, see [Borwanker et al., 1971] whose proof is again deferred to Appendix C.

Theorem 4. *Suppose that the assumptions of Theorem 3 hold and that one tries to infer the true value of θ^* using the ABC approximate Bayesian posterior (15). Suppose further that the prior distribution has a continuous density w.r.t. Lebesgue measure, then for sufficiently small values of ϵ one has that $\bar{\mathbb{P}}_{\theta^*}$ a.s.*

$$\pi^{\epsilon}_n\left(\sqrt{n}(\theta - \hat{\theta}_{n,\epsilon})\right) \to N\left(0, I_{\epsilon}^{-1}\right) \tag{16}$$

where I_{ϵ} is as in Theorem 3.

3.3. *Asymptotic Rates of Convergence.* Theorems 2, 3 and 4 show that asymptotically ABC based parameter estimators concentrate around a point $\theta^{*,\epsilon} \neq \theta^*$ and thus that the asymptotic bias will be of order $|\theta^{*,\epsilon} - \theta^*|$. It is natural to ask at what rate does $\theta^{*,\epsilon} \to \theta^*$ as $\epsilon \to 0$. We begin our answer to this question with the following example.

Example 1. *Let π_1 be the distribution on the set of diadic numbers of the form $\frac{1}{4^k}$; $k = 0, 1, \ldots$ given by $\pi_1(\frac{1}{4^k}) = \frac{3}{4^{k+1}}$ for all k and let π_2 be the distribution on the set of diadic numbers of the form $\frac{1}{2\cdot 4^k}$ given by $\pi_2(\frac{1}{2\cdot 4^k}) = \frac{3}{4^{k+1}}$ for all $k = 0, 1, \ldots$. Furthermore let $\{\pi_\theta\}_{\theta\in[0.25,0.75]}$ be the set of distributions defined such that for all θ, $\pi_\theta = \theta\pi_1 + (1-\theta)\pi_2$.*

It is clear that the distributions π_θ satisfy the conditions of Theorem 1 and hence that for any ϵ the limiting approximate mean log likelihood surface $l^\epsilon(\theta)$ exists and is well defined. Further if we assume that the true value of the parameter is equal to $\theta^* = \frac{1}{2}$ then it is easy to show that $\nabla^2_\theta l^0(\theta^*) \neq 0$ and that for all $k \geq 0$ that $\nabla_\theta l^{\frac{1}{4^{k+1}}}(\theta^*) = \frac{3}{4^{k+2}}$ from which it follows that

$$\theta^{*,\frac{1}{4^{k+1}}} - \theta^* = \frac{1}{\nabla^2_\theta l^0(\theta^*)}\frac{3}{4^{k+2}} + o\left(\frac{1}{4^{k+1}}\right).$$

The above example shows that in the general case one should expect that the size of the asymptotic bias will be at least $O(\epsilon)$. The next theorem shows that the behaviour of the asymptotic bias will be no worse than this. In order for it to hold we need to make the following Lipschitz assumptions.

(**A7**) There exists some $R > 0$ such that for all $\epsilon \leq R$.

$$\begin{gathered} E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B^0_\epsilon}\left|\frac{\nabla_z g_\theta\left(Y+z|x\right)}{g_\theta\left(Y|x\right)}\right|^2\right], \\ E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B^0_\epsilon}\left|\frac{\nabla_z\left(\nabla_\theta g_\theta\left(Y+z|x\right)\right)}{g_\theta\left(Y|x\right)}\right|^2\right] < \infty. \end{gathered} \tag{17}$$

Theorem 5. *Suppose that in addition to all of the assumptions of Theorem 4 one has that assumption (A7) above also holds. Then*

$$|\theta^{\epsilon,*} - \theta^*| = O(\epsilon). \tag{18}$$

Moreover, if the dominating measure ν is Lebesgue measure then one can show, under slightly stronger Lipschitz assumptions, that the asymptotic error in the ABC parameter estimate is of order $O(\epsilon)^2$.

(**A8**) There exists some $R > 0$ such that for all $\epsilon \leq R$.

$$E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B_\epsilon^0}\left|\frac{\nabla_z^2 g_\theta\left(Y+z|x\right)}{g_\theta\left(Y|x\right)}\right|^2\right],$$
$$E_{\theta^*}\left[\sup_{x\in\mathcal{X}}\sup_{\theta\in\Theta}\sup_{z\in B_\epsilon^0}\left|\frac{\nabla_z^2\left(\nabla_\theta g_\theta\left(Y+z|x\right)\right)}{g_\theta\left(Y|x\right)}\right|^2\right] < \infty. \tag{19}$$

THEOREM 6. *Suppose that ν is Lebesgue measure and that in addition to all of the assumptions of Theorem 5 one has that assumption (A8) above holds also. Then*

$$|\theta^{\epsilon,*} - \theta^*| = O(\epsilon^2). \tag{20}$$

The proofs of Theorems 5 and 6 are deferred to Appendix D. Finally we note that in the case that ν is Lebesgue measure we have from Theorems 3 and 4 that the variance of ABC based based estimators is of order $O(1/\sqrt{n})$ while their bias is of order $O(\epsilon^2)$. It follows that (at least in theory) it is optimal to scale ϵ as $O(1/\sqrt[4]{n})$ as n goes to infinity. Intriguingly this is the same rate as the optimal bandwidth in kernel density estimation (see for example [Wand and Jones, 1995]). This suggests an alternative interpretation of ABC as approximating the likelihood via a kind of kernel density based estimate.

4. Summary. In this paper we have shown that the framework developed in [Dean et al., 2010] to analyse the behaviour of the the ABC MLE can be extended to provide a rigourous analysis of the behaviour of ABC based estimators in both the Bayesian and frequentist contexts. In particular we have shown that ABC based parameter estimators satisfy results analogous to the asymptotic consistency, asymptotic normality and Bernstein-von Mises theorems for standard parameter estimators but that the ABC estimators are asymptotically biased. Furthermore we have shown that this asymptotic bias can be made arbitrarily small by choosing a sufficiently small value of the parameter ϵ. Together these theoretical resultshelp to solidify and extend existing intuition and provide a rigourous theoretical justification for ABC based parameter estimation procedures.

Appendix A: Auxillary Results. In this section we present without proof some well known results that will be needed in the proofs of Theorems 1, 2, 3, 4 and 5. The first two lemmas are standard result from real analysis.

Lemma 1. *Let a connected compact set $G \subset \mathbb{R}^u$ and some constant $K > 0$ 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 Lipschitz-K 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.*

Lemma 2. *Let a connected compact set $G \subset \mathbb{R}^u$ be given and suppose that there exists a continuous function $g : G \to \mathbb{R}^v$ and sequence of continuously differentiable functions $f_n : G \to \mathbb{R}^v$, $n \geq 1$, such that $\nabla f_n(z) \to g(z)$ uniformly in z and $f_n(z^*)$ is Cauchy for some $z^* \in G$. Then there exists a uniformly bounded and continuously differentiable function f such that $f_n(z) \to f(z)$ uniformly in z and $\nabla f(z) = g(z)$.*

Lemmas 3, 4 and 5 are essentially corollaries and extensions of Propositions 4 and 5 in [Douc et al., 2004] and may be proved in exactly the same manner. We leave the details to the reader.

Lemma 3. *Suppose that one has a collection of HMMs parameterised some vectors $\theta \in \Theta$ that satisfy assumption (A2). Furthermore suppose that one has a HMM $\{X_k, Y_k\}_{k \geq 1}$, defined on the same state spaces as the parameterised collection of HMMs, which satisfies assumption (A2) with the same values of $\underline{c}$ and $\overline{c}$.*

Given measurable functions $\phi_1, \phi_2, \phi_3 : \Theta \times \mathcal{X}^2 \times \mathcal{Y} \to \mathbb{R}$ and $y \in \mathcal{Y}$, $k < l$ and $s \in \{1, 2, 3\}$ define the following functions of the HMM $\{X_k, Y_k\}_{k \geq 1}$

$$\phi_{s;k:l}(\theta) \triangleq \sum_{i=k+1}^{l} \phi_s\left(\theta, X_{i-1}, X_i, Y_i\right)$$

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

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

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

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

and

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

Then there exist $\sigma(Y_{-\infty:0})$ measurable random variables $\Delta_{0,\infty}(\theta)$, $\Gamma_{0,\infty}(\theta)$, $\Psi_{0,\infty}(\theta)$ and $\Omega_{0,\infty}(\theta)$ and constants $C < \infty$ and $0 < \rho < 1$ which depend only on $\underline{c}$ and $\overline{c}$ such that for any initial condition on the collection of parameterised HMMs

$$\begin{aligned}\bar{E}\left[\sup_{\theta\in\Theta}|\Delta_{0,n}(\theta)-\Delta_{0,\infty}(\theta)|\right] &\leq C\rho^n\bar{E}\Big[\,\|\phi_1\|_\infty\Big] \\ \bar{E}\left[\sup_{\theta\in\Theta}|\Gamma_{0,n}(\theta)-\Gamma_{0,\infty}(\theta)|\right] &\leq C\rho^n\sup_{s\in\{1,2\}}\bar{E}\Big[\,\|\phi_s\|_\infty^2\Big] \\ \bar{E}\left[\sup_{\theta\in\Theta}|\Psi_{0,n}(\theta)-\Psi_{0,\infty}(\theta)|\right] &\leq C\rho^n\sup_{s\in\{1,2,3\}}\bar{E}\Big[\,\|\phi_s\|_\infty^3\Big] \\ \bar{E}\left[\sup_{\theta\in\Theta}|\Omega_{0,n}(\theta)-\Omega_{0,\infty}(\theta)|\right] &\leq C\rho^n\sup_{s\in\{1,2,3\}}\bar{E}\Big[\,\|\phi_s\|_\infty^3\Big]\end{aligned}\tag{A-21}$$

where for all $s \in \{1,2,3\}$

$$\|\phi_s\|_\infty(y) \triangleq \sup_{\theta\in\Theta}\sup_{x,x'\in\mathcal{X}}\left|\phi_s\left(\theta,x,x',y\right)\right|$$

$\bar{E}[\cdot]$ denotes expectation w.r.t. the law and stationary law respectively of the process $\{X_k,Y_k\}_{k\geq1}$.

Lemma 4. *Suppose that the assumptions of Lemma 3 all hold. Then there exist constants $C < \infty$ and $0 < \rho < 1$ such that for any initial condition on the collection of parameterised HMMs*

$$\bar{E}\left[\sup_{\theta\in\Theta}|\Delta_{0,n}(\theta)-\Delta_{0,\infty}(\theta)|^2\right] \leq C\rho^n\bar{E}\Big[\,\|\phi_1\|_\infty^2\Big].\tag{A-22}$$

Lemma 5. *Let the same assumptions and notation as Lemma 3 be given. Then there exist constants $C < \infty$ and $0 < \rho < 1$ such that for any k, n*

$$\bar{E}\left[\left|\bar{E}\left[\Delta_{0,n}(\theta)|Y_{-\infty:-k}\right] - \bar{E}\left[\Delta_{0,n}(\theta)|Y_{-\infty:-k-1}\right]\right|^2\right] \leq C\rho^k \bar{E}\left[\,\|\phi_1\|_\infty^2\,\right] \tag{A-23}$$

where $\bar{E}\left[\cdot|\cdot\right]$ denotes conditional expectation w.r.t. the law of the process $\{X_k, Y_k\}_{k\geq 1}$.

The last 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 results to third order derivatives of the log likelihood function. Given assumptions (A2)-(A6) it follows from a simple application of the dominated convergence theorem.

Lemma 6. *Suppose that assumptions (A2)-(A6) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$ where for each $\theta \in \Theta$ we let $g_\theta\left(y|x\right)$ and $q_\theta\left(x', x\right)$ denote the densities of the conditional law and transition kernel of the corresponding HMM. For any $\epsilon \geq 0$ let $g_\theta^\epsilon\left(y|x\right)$ denote the density of the conditional law of the corresponding perturbed HMM (10). By convention we let $g_\theta^0\left(y|x\right) = g_\theta\left(y|x\right)$.*

For any $\theta \in \Theta$, $\epsilon \geq 0$ and $n > 0$ let $\psi(\theta, x, x', y) = \log g_\theta^\epsilon\left(y|x'\right) q_\theta\left(x, x'\right)$ and following the notation of Lemma 3 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 $\log p_\theta^\epsilon(\cdots)$ is three times differentiable and

$$\nabla_\theta \log p_\theta^\epsilon(Y_1, \ldots, Y_n) = E_{\theta^\epsilon}\left[\nabla_\theta \psi_n(\theta)\middle|Y_{1:n}\right], \tag{A-24}$$

$$\begin{aligned}\nabla_\theta^2 \frac{1}{n} &\log p_\theta^\epsilon(Y_1, \ldots, Y_n)\\ &= E_{\theta^\epsilon}\left[\nabla_\theta^2 \psi_n\middle|Y_{1:n}\right] + E_{\theta^\epsilon}\left[\left(\nabla_\theta \psi_n\right)^{*2}\middle|Y_{1:n}\right] - E_{\theta^\epsilon}\left[\nabla_\theta \psi_n\middle|Y_{1:n}\right]^{*2},\end{aligned} \tag{A-25}$$

and

$$\begin{aligned}\nabla_\theta^3 \frac{1}{n} \log p_\theta^\epsilon(Y_1, \ldots, Y_n) &= E_{\theta^\epsilon}\left[\nabla_\theta^3 \psi_n\middle|Y_{1:n}\right]\\ &+ 3E_{\theta^\epsilon}\left[\nabla_\theta^2 * \psi_n \nabla_\theta \psi_n\middle|Y_{1:n}\right] - 3E_{\theta^\epsilon}\left[\nabla_\theta^2 \psi_n\middle|Y_{1:n}\right] * E_{\theta^\epsilon}\left[\nabla_\theta \psi_n\middle|Y_{1:n}\right]\\ &- 3E_{\theta^\epsilon}\left[\left(\nabla_\theta \psi_n\right)^{*2}\middle|Y_{1:n}\right] * E_{\theta^\epsilon}\left[\nabla_\theta \psi_n\middle|Y_{1:n}\right] + E_{\theta^\epsilon}\left[\left(\nabla_\theta \psi_n\right)^{*3}\middle|Y_{1:n}\right]\\ &+ 2E_{\theta^\epsilon}\left[\nabla_\theta \psi_n\middle|Y_{1:n}\right]^{*3}\end{aligned} \tag{A-26}$$

where $E_{\theta^\epsilon}\left[\cdot|\cdot\right]$ denotes conditional expectation w.r.t. the law of the perturbed HMM (10)*.*

Appendix B: Proof of Theorem 1. Theorem 1 is an immediate corollary of the following three lemmas.

LEMMA 7. *Suppose that assumptions (A1)-(A6) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Then for any $\epsilon \geq 0$ there exists a twice continuously differentiable function $l^\epsilon(\theta)$ such that*

(B-27)

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

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

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\nabla_\theta^2\left(\log p_\theta^\epsilon(Y_1,\ldots,Y_n)-\log p_{\theta^*}^\epsilon(Y_1,\ldots,Y_n)\right)-\nabla_\theta^2 l^\epsilon\left(\theta\right)\right|=0$$

$\bar{\mathbb{P}}_{\theta^*}$ a.s. and in $L_1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ where for all θ and ϵ, $p_\theta^\epsilon(\cdots)$ denotes the likelihood function of the perturbed HMM (10)*. By convention we define $p_\theta^0(\cdots)$ 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 $\epsilon\geq 0$*

$$l^\epsilon\left(\theta\right),\nabla_\theta l^\epsilon\left(\theta\right),\nabla_\theta^2 l^\epsilon\left(\theta\right)\leq K \tag{B-28}$$

and $l^\epsilon\left(\theta\right),\nabla_\theta l^\epsilon\left(\theta\right),\nabla_\theta^2 l^\epsilon\left(\theta\right)$ are K-Lipschitz (as functions of θ).

LEMMA 8. *Suppose that assumptions (A1)-(A6) 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 7. Then for all $\theta\in\Theta$ one has that*

$$\lim_{\epsilon\to0}\nabla_\theta l^\epsilon\left(\theta\right)=\nabla_\theta l^0\left(\theta\right).$$

LEMMA 9. *Suppose that assumptions (A1)-(A6) 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 7. Then for all $\theta\in\Theta$ one has that*

$$\lim_{\epsilon\to0}\nabla_\theta^2 l^\epsilon\left(\theta\right)=\nabla_\theta^2 l^0\left(\theta\right).$$

In order to complete this section we need to provide the proofs of Lemmas 7, 8 and 9. We start by stating some properties of the perturbed conditional likelihood (9) that will be needed in the sequel. First note that it follows from assumptions (A2) and (A5) and a simple application of the dominated convergence theorem that

$$\nabla_\theta g_\theta^\epsilon\left(y|x\right) \triangleq \frac{\int_{B_y^\epsilon} \nabla_\theta g_\theta\left(z|x\right)\nu(dz)}{\int_{B_y^\epsilon} \nu\left(dz\right)} \tag{B-29}$$

and that $\nabla_\theta g_\theta^\epsilon\left(y|x\right)$ is continuous w.r.t. θ for all ϵ, x and y. Furthermore since

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

it follows from (B-29) and assumption (A5) that for any $\epsilon > 0$

$$E_{\theta^*}\left[\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left\|\nabla_\theta \log g_\theta^\epsilon\left(Y|x\right)\right\|\right] < \infty. \tag{B-30}$$

Finally we note that analogous comments hold for $g_\theta^\epsilon\left(y|x\right)$ and $\nabla_\theta^2 g_\theta^\epsilon\left(y|x\right)$.

We now proceed to the proof of Lemma 7.

Proof of Lemma 7. First note that for any n the gradient of the mean log ABC likelihood may be decomposed into the following telescoping sum

$$\nabla_\theta \frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n) = \frac{1}{n}\sum_{i=1}^n h_\theta^\epsilon(Y_{1:i}). \tag{B-31}$$

where for any $k < n$

$$h_\theta^\epsilon(Y_{k:n}) := \nabla_\theta \log p_\theta^\epsilon(Y_k,\ldots,Y_n) - \nabla_\theta \log p_\theta^\epsilon(Y_k,\ldots,Y_{n-1}). \tag{B-32}$$

It then follows from (A-21) and (A-24) that there exist constants $K < \infty$ and $0 < \rho < 1$ such that for all $\theta \in \Theta$, $\epsilon \geq 0$ and $n > 0$ there exists some $\sigma(Y_{-\infty:0})$ measurable random variable $R_\theta^\epsilon(Y_{-\infty:0})$ such that

$$\bar{E}_{\theta^*}\left[\sup_{k\geq n}\left|h_\theta^\epsilon(Y_{-n:0}) - R_\theta^\epsilon(Y_{-\infty:0})\right|\right] \leq K\rho^n. \tag{B-33}$$

We note that by (B-29) and (B-30) and the accompanying comments and the dominated convergence theorem that $E_{\theta^*}\left[h_\theta^\epsilon(Y_{-n:0})\right]$ is continuous for all n and hence by (B-33) that $E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right]$ is continuous. Further it then follows from (B-33) and two applications of the ergodic theorem that for any $m > 0$

$$\begin{aligned}
&\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} h_\theta^\epsilon(Y_{i:i}) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right]\right| \\
&\leq \limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{m} h_\theta^\epsilon(Y_{1:i}) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right]\right| \\
&+ \limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=m+1}^{n} R_\theta^\epsilon(Y_{-\infty:i}) - E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right]\right| \\
&+ \limsup_{n\to\infty}\frac{1}{n}\sum_{i=m+1}^{n}\sup_{k\geq 0}\left|h_\theta^\epsilon(Y_{1-k:i}) - R_\theta^\epsilon(Y_{-\infty:i})\right| \\
&\leq K\rho^m \qquad \text{(B-34)}
\end{aligned}$$

and

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} h_\theta^\epsilon(Y_{1:i})\right| \leq \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}\sup_{k\leq i-1}\left|h_\theta^\epsilon(Y_{k:i})\right| \leq K. \qquad \text{(B-35)}$$

Thus we have by (B-31), (B-34) and (B-35) that $\overline{\mathbb{P}}_{\theta^*}$ a.s.

$$\nabla_\theta\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n) \to E_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right] \qquad \text{(B-36)}$$

pointwise in θ for some continuous in θ function $\bar{E}_{\theta^*}\left[R_\theta^\epsilon(Y_{-\infty:0})\right]$ and that $|\nabla_\theta\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ is eventually uniformly bounded above by K.

Moreover it follows from (A-21), (A-25) and (A-26) and a similar argument as above that for any $\theta\in\Theta$ and $\epsilon\geq 0$ there exist $\sigma(Y_{-\infty:0})$ measurable random variables $S_\theta^\epsilon(Y_{-\infty:0})$ and $T_\theta^\epsilon(Y_{-\infty:0})$ such that $E_{\theta^*}\left[S_\theta^\epsilon(Y_{-\infty:0})\right]$ and $E_{\theta^*}\left[T_\theta^\epsilon(Y_{-\infty:0})\right]$ are continuous functions of θ, that

$$\begin{aligned}
&\nabla_\theta^2\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n) \to \bar{E}_{\theta^*}\left[S_\theta^\epsilon(Y_{-\infty:0})\right], \\
&\nabla_\theta^3\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n) \to \bar{E}_{\theta^*}\left[T_\theta^\epsilon(Y_{-\infty:0})\right]
\end{aligned} \qquad \text{(B-37)}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. and in $L^1\left(\overline{\mathbb{P}}_{\theta^*}\right)$ and that $\overline{\mathbb{P}}_{\theta^*}$ a.s. eventually $|\nabla_\theta^2\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ and $|\nabla_\theta^3\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ are both uniformly bounded above by K. Since the fact that $|\nabla_\theta^2\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ and $|\nabla_\theta^3\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ are both uniformly bounded above implies that both $|\nabla_\theta\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ and $|\nabla_\theta^2\frac{1}{n}\log p_\theta^\epsilon(Y_1,\ldots,Y_n)|$ are Lipschitz the result now follows from Lemmas 1 and 2. □

In remains to prove Lemmas 8 and 9. Since the proofs of these two lemmas are almost identical we prove only Lemma 8 and leave the details of the proof of Lemma 9 to the reader.

Proof of Lemma 8. It follows from (B-31) and (B-33) that in order to prove the result it is sufficient to show that

$$\lim_{\epsilon\to 0}\bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_\theta^\epsilon(Y_1,\dots,Y_n)\right]=\bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_\theta(Y_1,\dots,Y_n)\right] \tag{B-38}$$

for all n and θ and hence by (A-24), (B-29) and (B-30) and the accompanying comments and the dominated convergence theorem that

$$\begin{aligned} &E_{\theta^\epsilon}\left[\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\\ &\qquad= E_\theta\left[\nabla_\theta\left(\log g_\theta\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right]\end{aligned} \tag{B-39}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s. for all θ and $1\le k\le n$. Recall that

(B-40)

$$E_{\theta^\epsilon}\left[\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right] = \frac{\int_{\mathcal{X}^n}\nabla_\theta\left(\log g_\theta^\epsilon\left(Y_k|x_k\right)q_\theta\left(x_{k-1},x_k\right)\right)\prod_{i=1}^n\left(g_\theta^\epsilon\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_{1:n})}{\int_{\mathcal{X}^n}\prod_{i=1}^n\left(g_\theta^\epsilon\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_{1:n})}$$

and

(B-41)

$$E_{\theta}\left[\nabla_\theta\left(\log g_\theta\left(Y_k|X_k\right)q_\theta\left(X_{k-1},X_k\right)\right)|Y_{1:n}\right] = \frac{\int_{\mathcal{X}^n}\nabla_\theta\left(\log g_\theta\left(Y_k|x_k\right)q_\theta\left(x_{k-1},x_k\right)\right)\prod_{i=1}^n\left(g_\theta\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_{1:n})}{\int_{\mathcal{X}^n}\prod_{i=1}^n\left(g_\theta\left(Y_i|x_i\right)q_\theta\left(x_{i-1},x_i\right)\right)\mu(dx_{1:n})}.$$

Further we have by (B-30) and the accompanying comments that we can use the Lebesgue differentiation theorem (see for example [Wheeden and Zygmund, 1977]) to deduce that for all $x\in\mathcal{X}$ that

$$\nabla_\theta g_\theta^\epsilon\left(Y_k|x\right)\to\nabla_\theta g_\theta\left(Y_k|x\right),\; g_\theta^\epsilon\left(Y_k|x\right)\to g_\theta\left(Y_k|x\right) \tag{B-42}$$

$\overline{\mathbb{P}}_{\theta^*}$ a.s.. It now follows from assumptions (A2) and (A5), (B-30) etc. and (B-42) and the dominated convergence theorem that the numerator and denominator of the quantity in (B-40) converge to respectively the numerator and denominator of the quantity in (B-41). Since by assumption (A4) we have that

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

$\overline{\mathbb{P}}_{\theta^*}$ a.s. we obtain (B-39). □

Appendix C: Proofs of Theorems 2, 3 and 4.

Proof of Theorem 2. It follows immediately from Theorem 1 that the first part of Theorem 2 will hold with the set $\mathcal{T}^\epsilon$ equal to the set of maximisers of $l^\epsilon(\theta)$. Note that since $l^\epsilon(\theta)$ is continuous and Θ compact $\mathcal{T}^\epsilon$ will always be well defined and non-empty. Further, (13) follows from the uniform convergence of $l^\epsilon(\theta)$ to $l^0(\theta)$ and the continuity of the surfaces.

It remains to prove the second part of the theorem. Suppose now that $\nabla^2_\theta l^0(\theta^*)$ is strictly negative definite. We have from the last part of Theorem 1 that

$$\lim_{\delta\to 0}\lim_{\epsilon\to 0}\sup_{|\theta-\theta^*|\le\delta}\left\|\nabla^2_\theta l^\epsilon(\theta) - \nabla^2_\theta l^0(\theta^*)\right\| = 0. \tag{C-43}$$

Equation (C-43) implies that there exists some $\delta > 0$ such that for sufficiently small ϵ the surface $l^\epsilon(\theta)$ has at most one local maximum in the δ neighbourhood of θ^*. The result now follows from (13). □

Proof of Theorem 3. Letting the matrix I_ϵ be equal to $\nabla^2_\theta l^\epsilon(\theta^{*,\epsilon})$ it follows from Theorem 1 and standard results on the asymptotic normality of the MLE (see for example [Douc et al., 2004]) that in order to prove Theorem 3 it is sufficient to show that for ϵ sufficiently small there exists some strictly positive definite matrix J_ϵ such that

$$\frac{1}{\sqrt{n}}\nabla_\theta \log p^\epsilon_{\theta^{*,\epsilon}}(Y_1,\ldots,Y_n) \to N(0, J_\epsilon) \tag{C-44}$$

and

$$J_\epsilon \to I \tag{C-45}$$

as $\epsilon \to 0$ where $I = \nabla^2_\theta l^0(\theta^*)$.

We begin by proving (C-44). We have by (B-31) and (B-33) that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\nabla_\theta \log p^\epsilon_{\theta^*,\epsilon}(Y_1,\ldots,Y_n)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i}) + \frac{1}{\sqrt{n}}\sum_{i=1}^n h^\epsilon_{\theta^*,\epsilon}(Y_{1:i}) - R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})\end{aligned} \tag{C-46}$$

where $h^\epsilon_{\theta^*,\epsilon}(Y_{1:i})$ is as defined in (B-32). We note that it follows from (A-22) that one can use similar arguments to those used to deduce (B-31) to show that

$$\bar{E}_{\theta^*}\left[\sup_{k\geq n}\left|h^\epsilon_\theta(Y_{-n:0}) - R^\epsilon_\theta(Y_{-\infty:0})\right|^2\right] \leq K\rho^n. \tag{C-47}$$

It then follows from (C-47) that

$$\bar{E}_{\theta^*}\left[h^\epsilon_{\theta^*,\epsilon}(Y_{-n:i})|Y_{-\infty:0}\right] \xrightarrow{L_2} \bar{E}_{\theta^*}\left[R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})|Y_{-\infty:0}\right]$$

and likewise for conditional expectations w.r.t. $\sigma(Y_{-\infty:-1})$ and hence by (A-23) that there exists some K such that

$$\bar{E}_{\theta^*}\left[\left|\bar{E}_{\theta^*}\left[R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})|Y_{-\infty:0}\right] - \bar{E}_{\theta^*}\left[R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})|Y_{-\infty:-1}\right]\right|^2\right] \leq K\rho^i \tag{C-48}$$

for all i. Equation (C-48) immediately implies that the sequence of random variables $R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:0}), R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:1}),\ldots$ satisfies the conditions of Theorem 5 in [Volný, 1993] and hence we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i}) \xrightarrow{\text{weakly}} N(0, J_\epsilon) \tag{C-49}$$

where

$$J_\epsilon = \lim_{n\to\infty} \bar{E}_{\theta^*}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})\right)^2\right]. \tag{C-50}$$

Finally we note that it follows from (B-33), Markov's inequality and the Borel-Cantelli lemma that

$$\bar{\mathbb{P}}_{\theta^*}\left(|h^\epsilon_{\theta^*,\epsilon}(Y_{1:i}) - R^\epsilon_{\theta^*,\epsilon}(Y_{-\infty:i})| > \frac{1}{i^2}\text{ i.o.}\right) = 0. \tag{C-51}$$

Equation (C-44) now follows from (C-49) and (C-51).

To complete the proof of the theorem it remains to prove (C-45). It immediately follows from (B-32), (C-47) and (C-50) that for all ϵ

$$J_\epsilon = \lim_{n\to\infty} \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p^\epsilon_{\theta^*,\epsilon}(Y_1,\dots,Y_n)^2\right] \tag{C-52}$$

where the convergence is uniform in n. Next we note that by a simple application of the Fisher identity (see for example [Douc et al., 2004]) that

$$\bar{E}_{\theta^*}\left[\nabla^2_\theta \log p_{\theta^*}(Y_1,\dots,Y_n)\right] = \bar{E}_{\theta^*}\left[\nabla_\theta \log p_{\theta^*}(Y_1,\dots,Y_n)^2\right]$$

and thus by (B-37) and Lemma 7 that

$$I = \lim_{n\to\infty} \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p^\epsilon_{\theta^*}(Y_1,\dots,Y_n)^2\right] \tag{C-53}$$

In order to complete the proof of (C-45) it is thus sufficient, by (C-52) and (C-53), to show that for all n

(C-54)
$$\lim_{\epsilon\to\infty} \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p^\epsilon_{\theta^*,\epsilon}(Y_1,\dots,Y_n)^2\right] = \bar{E}_{\theta^*}\left[\frac{1}{n}\nabla_\theta \log p_{\theta^*}(Y_1,\dots,Y_n)^2\right].$$

Finally we note that (C-54) can be proved in exactly the same way as (B-42) in the proof of Lemma 8. In order to this we need to show that

$$\nabla_\theta g^\epsilon_{\theta^*,\epsilon}\left(Y_k|x\right) \to \nabla_\theta g_{\theta^*}\left(Y_k|x\right),\ g^\epsilon_{\theta^*,\epsilon}\left(Y_k|x\right) \to g_{\theta^*}\left(Y_k|x\right) \tag{C-55}$$

as $\epsilon \to 0$. However (C-55) follows from (B-42) and the fact that by assumptions (A2) and (A6) we have that $\mathbb{P}_{\theta^*}$ a.s. the functions $\nabla_\theta g_\theta\left(Y_k+z|x\right)$ and $g_\theta\left(Y_k+z|x\right)$ are uniformly Lipschitz (as functions of θ) for all $z \in B^\epsilon_0$. □

Proof of Theorem 4. The proof of this result follows from standard Bernstein-Von Mises type arguments, see for example [Borwanker et al., 1971]. □

Appendix D: Proofs of Theorems 5 and 6. A central role in the proof of Theorem 5 will be played by the following time inhomogeneous versions of the perturbed HMM (10).

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 $\epsilon > 0$ define the HMM $\left\{X_k^{\epsilon,+}, Y_k^{\epsilon,+}\right\}_{k\in\mathbb{Z}}$ by

(D-56)
$$X_k^{\epsilon,+}, Y_k^{\epsilon,+} = X_k, Y_k \text{ for all } k \le 0; \quad X_k^{\epsilon,+}, Y_k^{\epsilon,+} = X_k, Y_k + \epsilon Z_k \text{ o.w.}$$

where $\{X_k, Y_k\}_{k\in\{\mathbb{Z}\}}$ is the original HMM and $\{Z_k\}_{k\ge 0}$ is a collection of i.i.d. $\mathcal{U}_{B_0^1}$ random variables. Similarly define the HMM $\left\{X_k^{\epsilon,-}, Y_k^{\epsilon,-}\right\}_{k\in\mathbb{Z}}$ by

(D-57)
$$X_k^{\epsilon,-}, Y_k^{\epsilon,-} = X_k, Y_k \text{ for all } k < 0; \quad X_k^{\epsilon,-}, Y_k^{\epsilon,-} = X_k, Y_k + \epsilon Z_k \text{ o.w..}$$

Clearly the transition kernels of the HMMs (D-56) and (D-57) are equal to $q_\theta\left(x,x'\right)$ and the conditional densities of the observed state are equal to

$$g_{\theta,k}^{\epsilon,+}\left(y|x\right) = \begin{cases} g_\theta^\epsilon\left(y|x\right) & \text{if } k > 0 \\ g_\theta\left(y|x\right) & \text{otherwise} \end{cases} \tag{D-58}$$

and

$$g_{\theta,k}^{\epsilon,-}\left(y|x\right) = \begin{cases} g_\theta^\epsilon\left(y|x\right) & \text{if } k \ge 0 \\ g_\theta\left(y|x\right) & \text{otherwise} \end{cases} \tag{D-59}$$

respectively.

Let $p_{\theta^\epsilon,+}(\cdots)$, $\mathbb{P}_{\theta^\epsilon,+}(\cdot)$, $E_{\theta^\epsilon,+}[\cdot]$, $E_{\theta^\epsilon,+}[\cdot|\cdot]$ and $\mathbb{P}_{\theta^\epsilon,+}(\cdot|\cdot)$ and $p_{\theta^\epsilon,-}(\cdots)$, $\mathbb{P}_{\theta^\epsilon,-}(\cdot)$, $E_{\theta^\epsilon,-}[\cdot]$, $E_{\theta^\epsilon,-}[\cdot|\cdot]$ and $\mathbb{P}_{\theta^\epsilon,-}(\cdot|\cdot)$ denote the likelihood functions, laws and expectation, conditional expectation and conditional probability operators w.r.t. to the laws of (D-56) and (D-57). It follows by definition that

(D-60)
$$\begin{aligned} p_\theta(y_1,\ldots,y_n) &= p_{\theta^\epsilon,+}(Y_{-n+1}=y_1,\ldots,Y_0=y_n) \\ p_\theta^\epsilon(y_1,\ldots,y_n) &= p_{\theta^\epsilon,-}(Y_0=y_1,\ldots,Y_{n-1}=y_n) \\ p_{\theta^\epsilon,+}(Y_{-k+1}=y_{-k+1},\ldots,Y_n=y_n) &= p_{\theta^\epsilon,-}(Y_{-k}=y_{-k+1},\ldots,Y_{n-1}=y_n). \end{aligned}$$

Recall that by (B-31) and (B-33) we have that

(D-61)
$$\begin{aligned} &\nabla_\theta l\left(\theta\right) - \nabla_\theta l^\epsilon\left(\theta\right) \\ &\quad = \lim_{n\to\infty}\frac{1}{n}\left(\bar{E}_{\theta^*}\left[\nabla_\theta \log p_\theta(Y_1,\ldots,Y_n)\right] - \bar{E}_{\theta^*}\left[\nabla_\theta \log p_\theta^\epsilon(Y_1,\ldots,Y_n)\right]\right) \end{aligned}$$

and thus by (D-60) and (D-61) that we have the telescoping sum

$$\nabla_\theta l(\theta) - \nabla_\theta l^\epsilon(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big(\bar{E}_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_{-n+i},\ldots,Y_{i-1})\right] \\ -\bar{E}_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{-n+i},\ldots,Y_{i-1})\right]\Big). \tag{D-62}$$

We now note that Theorems 5 and 6 follow immediately from (D-62) and the following lemma

Lemma 10. *Suppose that assumptions (A2)-(A7) hold for a collection of HMMs parametrised by some vector $\theta \in \Theta$. Then there exists a finite constant K such that for all $\epsilon > 0$ and integers k, n*

$$\Big| E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_{-k},\ldots,Y_n)\right] - E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{-k},\ldots,Y_n)\right]\Big| \\ \le K\epsilon. \tag{D-63}$$

Furthermore suppose that ν is Lebesgue measure and that assumption (A8) also holds. Then

$$\Big| E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,+}}(Y_{-k},\ldots,Y_n)\right] - E_{\theta^*}\left[\nabla_\theta \log p_{\theta^{\epsilon,-}}(Y_{-k},\ldots,Y_n)\right]\Big| \\ \le K\epsilon^2. \tag{D-64}$$

Proof. We shall prove only the first part of the lemma, the proof of the second part being almost identical. Clearly analogous expressions to (A-24) hold for the HMMs (D-56) and (D-57) and thus in particular we have that the term on the left hand side of (D-64) is bounded by

$$\Bigg| \bar{E}_{\theta^*}\Bigg[\sum_{i=-k}^{n} E_{\theta^{\epsilon,+}}\left[\nabla_\theta \log g^{\epsilon,+}_{\theta,i}(Y_i|X_i)\, q_\theta(X_{i-1},X_i)\,\middle|\, Y_{-k:n}\right] - \\ \sum_{i=-k}^{n} E_{\theta^{\epsilon,-}}\left[\nabla_\theta \log g^{\epsilon,-}_{\theta,i}(Y_i|X_i)\, q_\theta(X_{i-1},X_i)\,\middle|\, Y_{-k:n}\right]\Bigg]\Bigg| \tag{D-65}$$

where $g^{\epsilon,+}_{\theta,i}$ and $g^{\epsilon,-}_{\theta,i}$ are as in (D-58) and (D-59). Using the identity

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

it is clear from (B-29) and assumptions (A6) and (A7) that there exists some K' such that

$$\begin{aligned} &\left|\bar{E}_{\theta^*}\left[E_{\theta^{\epsilon},+}\left[\nabla_\theta \log g^{\epsilon,+}_{\theta,0}\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\middle|Y_{-k:n}\right] - \right.\right. \\ &\qquad \left.\left. E_{\theta^{\epsilon},+}\left[\nabla_\theta \log g^{\epsilon,-}_{\theta,0}\left(Y_i|X_i\right) q_\theta\left(X_{i-1},X_i\right)\middle|Y_{-k:n}\right]\right]\right| \leq K'\epsilon. \end{aligned} \tag{D-66}$$

It then follows from the definitions of $g^{\epsilon,+}_{\theta,i}$ and $g^{\epsilon,-}_{\theta,i}$ and from assumptions (A2)-(A6) that in order to derive (D-64) from (D-65) it is sufficient to show that for all i,k,n

$$\bar{E}_{\theta^*}\left[\left\|\mathbb{P}_{\theta^{\epsilon},+}\left(X_{i-1},X_i\middle|Y_{-k:n}\right) - \mathbb{P}_{\theta^{\epsilon},-}\left(X_{i-1},X_i\middle|Y_{-k:n}\right)\right\|_{TV}\right] \leq K''\epsilon\rho^{|i|} \tag{D-67}$$

for some K''.

We first note that it follows from standard results concerning uniformly mixing Markov chains, see for example [Cappé et al., 2005, Del Moral, 2004] that there exist some K and $0<\rho<1$ such that for all i and $x,x' \in \mathcal{X}$ one has that

$$\left\|\mathbb{P}_{\theta^{\epsilon},+}\left(X_{i-1},X_i|X_0=x\right) - \mathbb{P}_{\theta^{\epsilon},+}\left(X_{i-1},X_i|X_0=x'\right)\right\|_{TV} \leq K\rho^{|i|-1}. \tag{D-68}$$

Since by definition we have that the marginal laws of $\mathbb{P}_{\theta^{\epsilon},+}\left(Y_{-\infty:-1;1:\infty}\right)$ and $\mathbb{P}_{\theta^{\epsilon},-}\left(Y_{-\infty:-1;1:\infty}\right)$ are equal it follows that in order to prove (D-67) it is sufficient to only prove it for the case $i=0$.

To prove (D-67) for $i=0$ we shall make use of the following simple identities. For any $\phi \in L_\infty$

$$\begin{aligned} E_{\theta^{\epsilon},+}\left[\phi(X_0)|Y_{\infty:\infty}\right] &= \frac{E_{\theta^{\epsilon},+}\left[\phi(X_0)g_\theta\left(Y_0|X_0\right)|Y_{\infty:-1;1:\infty}\right]}{E_{\theta^{\epsilon},+}\left[g_\theta\left(Y_0|X_0\right)|Y_{\infty:-1;1:\infty}\right]} \\ E_{\theta^{\epsilon},-}\left[\phi(X_0)|Y_{\infty:\infty}\right] &= \frac{E_{\theta^{\epsilon},-}\left[\phi(X_0)g^{\epsilon}_\theta\left(Y_0|X_0\right)|Y_{\infty:-1;1:\infty}\right]}{E_{\theta^{\epsilon},-}\left[g^{\epsilon}_\theta\left(Y_0|X_0\right)|Y_{\infty:-1;1:\infty}\right]}. \end{aligned} \tag{D-69}$$

It then follows from (D-69) using basic algebra that

$$\begin{aligned} &\sup_{\phi:\|\phi\|_\infty\leq 1} \bar{E}_{\theta^*}\left[\left|E_{\theta^{\epsilon},+}\left[\phi(X_0)|Y_{\infty:\infty}\right] - E_{\theta^{\epsilon},-}\left[\phi(X_0)|Y_{\infty:\infty}\right]\right|\right] \\ &\leq \bar{E}_{\theta^*}\left[\frac{E_{\theta^{\epsilon},+}\left[|g_\theta - g^{\epsilon}_\theta|\,|Y_{\infty:-1;1:\infty}\right]}{E_{\theta^{\epsilon},+}\left[g_\theta|Y_{\infty:-1;1:\infty}\right]} + \frac{E_{\theta^{\epsilon},-}\left[|g_\theta - g^{\epsilon}_\theta|\,|Y_{\infty:-1;1:\infty}\right]}{E_{\theta^{\epsilon},-}\left[g_\theta|Y_{\infty:-1;1:\infty}\right]}\right]. \end{aligned} \tag{D-70}$$

The result now follows follows immediately from (D-70) and assumption (A7). □

References.

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. Technical report, Cambridge University Engineering Department., 2010.

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. Technical, University of Lancaster., 2010.

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.

A. Jasra, S. S. Singh, J. S. Martin, and E. McCoy. Filtering via approximate bayesian computation. *Stat. Comput,*, 2010.

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

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. Sisson and Y. Fan. Likelihood-free markov chain monte carlo. In *Brooks, S. P., Gelman, A., Jones, G. and Meng X.L. (Eds); Handbook of Markov Chain Monte Carlo.* Chapman & Hall: London, to be published.

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.

T.A.Dean, S.S. Singh,
Department of Engineering,
University of Cambridge,
Cambridge,
CB2 1PZ, UK
E-mail: tad36@cam.ac.uk; sss40@cam.ac.uk