The Rate of Convergence for Approximate Bayesian Computation

Stuart Barber, Jochen Voss and Mark Webster

November 29, 2013

Abstract

Approximate Bayesian Computation (ABC) is a popular computational method for likelihood-free Bayesian inference. The term "likelihood-free" refers to problems where the likelihood is intractable to compute or estimate directly, but where it is possible to generate simulated data X relatively easily given a candidate set of parameters θ simulated from a prior distribution. Parameters which generate simulated data within some tolerance δ of the observed data x^* are regarded as plausible, and a collection of such θ is used to estimate the posterior distribution $\theta \,|\, X = x^*$. Suitable choice of δ is vital for ABC methods to return good approximations to θ in reasonable computational time.

While ABC methods are widely used in practice, particularly in population genetics, study of the mathematical properties of ABC estimators is still in its infancy. We prove that ABC estimates converge to the exact solution under very weak assumptions and, under slightly stronger assumptions, quantify the rate of this convergence. Our results can be used to guide the choice of the tolerance parameter δ.

keywords: Approximate Bayesian Computation, likelihood-free inference, Monte Carlo methods, convergence of estimators, rate of convergence

MSC2010 classes:
62F12 (Asymptotic properties of estimators),
62F15 (Bayesian inference),
65C05 (Monte Carlo methods)

1 Introduction

Approximate Bayesian Computation (ABC) is a popular method for likelihood-free Bayesian inference. ABC methods were originally introduced in population genetics, but are now widely used in applications as diverse as epidemiology (Tanaka et al., 2006; Blum and Tran, 2010; Walker et al., 2010), materials science (Bortot et al., 2007), parasitology (Drovandi and Pettitt, 2011), genetic evolution (Thornton and Andolfatto, 2006; Fagundes et al., 2007; Ratmann et al., 2009; Wegmann and Excoffier, 2010; Beaumont, 2010; Wilkinson et al., 2011), population migration (Guillemaud et al., 2010) and conservation studies (Lopes and Boessenkool, 2010).

One of the earliest articles about the ABC approach, covering applications in population genetics, is Tavaré et al. (1997). Newer developments and extensions of the method include placing ABC in an MCMC context (Marjoram et al., 2003), sequential ABC (Sisson et al., 2007), enhancing ABC with nonlinear regression models (Blum and François, 2010), agent-based modelling (Sottoriva and Tavaré, 2010), and a Gibbs sampling ABC approach (Wilkinson et al., 2011). A survey of recent developments is given by Marin et al. (2012).

ABC methods allow for inference in a Bayesian setting where the parameter $\theta \in \mathbb{R}^p$ of a statistical model is assumed to be random with a given prior distribution f_θ, we have observed data $X \in \mathbb{R}^d$ from a given distribution $f_{X|\theta}$ depending on θ and we want to use these data to draw inference about θ. In the areas where ABC methods are used, the likelihood $f_{X|\theta}$ is typically not available in an explicit form. The term "likelihood-free" is used to indicate that ABC methods do not make use of the likelihood $f_{X|\theta}$, but only work with samples from the joint distribution of θ and X. In the context of ABC methods, the data are usually summarised using a statistic $S\colon \mathbb{R}^d \to \mathbb{R}^q$, and analysis is then based on $S(X)$ instead of X.

The basic idea of ABC methods is to replace samples from the exact posterior distribution $\theta \,|\, X = x^*$ or $\theta \,|\, S(X) = s^*$ with samples from an approximating distribution like $\theta \,|\, S(X) \approx s^*$. There are many variants of the basic ABC method available, including different implementations of the condition $S(X) \approx s^*$. All variants use a tolerance parameter δ which controls the trade-off between fast generation of samples (large values of δ) and accuracy of samples (small values of δ). The easiest approach to implement the condition $S(X) \approx s^*$, considered in this paper, is to use $\|S(X) - s^*\| \leq \delta$ where $\|\cdot\|$ is some norm on $\mathbb{R}^q$. Different approaches to choosing the statistic S are used; a semi-automatic approach is described in Fearnhead and Prangle (2012). In many cases considerable improvements can be achieved by choosing the norm for comparison of $S(X)$ to s^* in a problem-specific way.

Despite the popularity of ABC methods, theoretical analysis is still in its infancy. The aim of this article is to provide a foundation for such analysis by providing rigorous results about the convergence properties of the ABC method. Here, we restrict discussion to the most basic variant, to set a baseline to which different ABC variants can be compared. We consider Monte Carlo estimates of posterior expectations, using the ABC samples for the estimate. Theorem 3.1 shows that such ABC estimates converge to the true value; once this is established we investigate the rate of convergence in theorem 3.2. Similar results, but in the context of estimating posterior densities rather than posterior expectations can be found in Blum (2010) and Biau et al. (2013).

The choice of norm and of the tolerance parameter δ are major challenges in the practical application of ABC methods, but not many results are available in the literature. A numerical study of the trade-off between accuracy and computational cost, in the context of sequential ABC methods, can be found in Silk et al. (2013). Wilkinson (2013) establishes that an ABC method which accepts or rejects proposals with a probability based on the difference between the observed and proposed data converges to the correct solution under assumptions on model or measurement error.

The error of an ABC estimate is affected both by the bias of the ABC samples, controlled by the tolerance parameter δ, and by Monte Carlo error,

controlled by the number n of accepted ABC samples. Similarly, computational cost is proportional to n and also increases when the tolerance is decreased. It is well-known that for Monte Carlo methods the error decays, as a function of computational cost, proportional to $\text{cost}^{-1/2}$, where the exponent $-1/2$ is independent of dimension (see *e.g.* Voss, 2014, section 3.2.2). In contrast, the main result of this article, theorem 3.2, shows that, under optimal choice of δ, the basic ABC method satisfies

$$\text{error} \sim \text{cost}^{-2/(q+4)}.$$

Thus the rate of decay for the error gets worse as the dimension q increases and even in the one-dimensional case the exponent $-2/(1+4) = -2/5$ is worse than the exponent $-1/2$ for Monte Carlo methods. For comparison, for the problem of estimating the posterior density, Blum (2010) reports the slightly worse exponent $-2/(q+5)$.

We continue by giving a very short introduction to the basic ABC method in section 2. The main results, theorems 3.1 and 3.2 are presented in section 3, together with their proofs. Section 4 illustrates the results of this paper with the help of numerical experiments. Finally, in section 5, we consider the practical implications of our results.

2 Approximate Bayesian Computation

This section gives a short introduction to the basic ABC algorithm. A more complete description is, for example, given in Voss (2014, section 5.1). We describe the algorithm in the context of the following Bayesian inference problem:

- A parameter vector $\theta \in \mathbb{R}^p$ is assumed to be random. Before observing any data, our belief about its value is summarised by the prior distribution f_θ. The value of θ is unknown to us and our aim is to make inference about θ.
- The available data $X \in \mathbb{R}^d$ are assumed to be a sample from a distribution $f_{X|\theta}$, depending on the parameter θ. Inference about θ is based on a single observed sample x^* from this distribution; repeated observations can be assembled into a single vector if needed.
- In the context of ABC, the data are often summarised using a statistic $S\colon \mathbb{R}^d \to \mathbb{R}^q$. Since X is random, $S = S(X)$ is random with a distribution $f_{S|\theta}$ depending on θ. If a summary statistic is used, inference is based on the value $s^* = S(x^*)$ instead of on the full sample x^*.

Our aim is to explore the posterior distribution of θ, *i.e.* the conditional distribution of θ given $X = x^*$, using Monte Carlo methods. More specifically, we aim to estimate the posterior expectations

$$y = E\big(h(\theta)\,\big|\,X = x^*\big) \tag{1}$$

for given test functions $h\colon \mathbb{R}^p \to \mathbb{R}$. Expectations of this form allow us to study all relevant properties of the posterior distribution, including the posterior mean when $h(\theta) = \theta_i$ with $i = 1, \ldots, p$, posterior second moments when $h(\theta) = \theta_i\theta_j$

with $i, j = 1, \ldots, p$, and the posterior probability of hitting a given set $A \subseteq \mathbb{R}^p$ when $h(\theta) = 1_A(\theta)$.

The basic ABC method for generating approximate samples from the posterior distribution is given in the following algorithm.

Algorithm 2.1. For a given observation $s^* \in \mathbb{R}^q$ and given tolerance $\delta > 0$, *ABC samples* approximating the distribution $\theta \,|\, S = s^*$ are random samples $\theta_j^{(\delta)} \in \mathbb{R}^p$ computed by the following algorithm:

1: let $j \leftarrow 0$
2: **while** $j < n$ **do**
3: sample θ with density $f_\theta(\,\cdot\,)$
4: sample X with density $f_{X|\theta}(\,\cdot\,|\theta)$
5: **if** $\|S(X) - s^*\|_A \leq \delta$ **then**
6: let $j \leftarrow j + 1$
7: let $\theta_j^{(\delta)} \leftarrow \theta$
8: **end if**
9: **end while**

The norm $\|\cdot\|_A$ used in the acceptance criterion is defined by $\|s\|_A^2 = s^\top A^{-1} s$ for all $s \in \mathbb{R}^q$, where A is a positive definite symmetric matrix. This includes the case of the Euclidean distance $\|\cdot\| = \|\cdot\|_I$ for $A = I$. In practical applications, the matrix A is often chosen to be an approximation to the local covariance matrix of $S(X)$ in the vicinity of the point s^*.

Using the output of the algorithm, an estimate for the posterior expectation (1) can be obtained as

$$Y_n^{(\delta)} = \frac{1}{n} \sum_{j=1}^{n} h(\theta_j^{(\delta)}) \tag{2}$$

where $\theta_1^{(\delta)}, \ldots, \theta_n^{(\delta)}$ are computed by algorithm 2.1. Since the output of the ABC algorithm approximates the posterior distribution, the Monte Carlo estimate $Y_n^{(\delta)}$ can be used as an approximation to the posterior expectation.

The ABC samples are only approximately distributed according to the posterior distribution, and thus the estimate $Y_n^{(\delta)}$ will not exactly converge to the true value y as $n \to \infty$. The quality of the approximation can be improved by decreasing the tolerance parameter δ, but this leads at the same time to a lower acceptance probability in algorithm 2.1 and thus, ultimately, to higher computational cost for obtaining the estimate $Y_n^{(\delta)}$. Hence, a trade-off must be made between accuracy of the results and speed of computation.

Since the algorithm is not using x^* directly, but uses $s^* = S(x^*)$ instead, we require S to be a sufficient statistic, so that we have

$$y = E\big(h(\theta) \,\big|\, X = x^*\big) = E\big(h(\theta) \,\big|\, S = s^*\big).$$

If S is not sufficient, an additional error will be introduced.

For *application* of the ABC method, knowledge of the distributions of θ, X and $S = S(X)$ is not required; instead we assume that we can simulate large numbers of samples of these random variables. In contrast, in our *analysis* we will assume that the joint distribution of θ and S has a density $f_{S,\theta}$ and we will need to consider properties of this density in some detail.

To conclude this section, we remark that there are two different approaches to choosing the sample size used to compute the ABC estimate $Y_n^{(\delta)}$. If we denote the number of proposals required to generate n output samples by $N \geq n$, then one approach is to choose the number N of proposals as fixed; in this case the number $n \leq N$ of accepted proposals is random. Alternatively, for given n, the loop in the ABC algorithm could be executed until n samples are accepted, resulting in random N and fixed n. In order to avoid complications with the definition of $Y_n^{(\delta)}$ for $n = 0$, we follow the second approach here.

3 Results

This section presents the main results of the paper, in theorems 3.1 and 3.2, followed by proofs of these results.

Throughout, we assume that the joint distribution of θ and S has a density $f_{S,\theta}$, and we consider the marginal densities of S and θ given by

$$f_S(s) = \int_{\mathbb{R}^p} f_{S,\theta}(s,t)\,dt$$

for all $s \in \mathbb{R}^q$ and

$$f_\theta(t) = \int_{\mathbb{R}^q} f_{S,\theta}(s,t)\,ds$$

for all $t \in \mathbb{R}^p$, respectively. We also consider the conditional density of θ given $S = s$, defined by

$$f_{\theta|S}(t|s) = \begin{cases} \frac{f_{S,\theta}(s,t)}{f_S(s)}, & \text{if } f_S(s) > 0, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Our aim is to study the convergence of the estimate $Y_n^{(\delta)}$ to y as $n \to \infty$. The basic result for this convergence is stated in the following theorem: the ABC estimate converges to the true value under very weak conditions.

Theorem 3.1. *Let $h\colon \mathbb{R}^p \to \mathbb{R}$ be such that $\mathbb{E}\big(|h(\theta)|\big) < \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^q$, the ABC estimate $Y_n^{(\delta)}$ given by (2) satisfies*

1. $\lim_{n\to\infty} Y_n^{(\delta)} = \mathbb{E}\big(Y_n^{(\delta)}\big)$ *almost surely for all $\delta > 0$;*
2. $\lim_{\delta\downarrow 0} \mathbb{E}\big(Y_n^{(\delta)}\big) = E\big(h(\theta)\,\big|\,S = s^*\big)$ *for all $n \in \mathbb{N}$; and*

The phrase "for f_S-almost all $s^* \in \mathbb{R}^q$" in the theorem indicates that the result holds for all s^* in a set $A \subseteq \mathbb{R}^q$ with $P\big(S(X) \in A\big) = 1$. Since in practice s^* will be an observed sample of $S(X)$, this condition forms no restriction to the applicability of the theorem.

Our second result, theorem 3.2, quantifies the speed of convergence of $Y_n^{(\delta)}$ to y. We consider the mean squared error

$$\mathrm{MSE}(Y_n^{(\delta)}) = \mathbb{E}\big((Y_n^{(\delta)} - y)^2\big) = \mathrm{Var}(Y_n^{(\delta)}) + \mathrm{bias}(Y_n^{(\delta)})^2,$$

and relate this to the computational cost of computing the estimate $Y_n^{(\delta)}$. For the computational cost, rather than using a sophisticated model of computation, we restrict ourselves to the naïve approach of assuming that the time required to obtain the estimate $Y_n^{(\delta)}$ in algorithm 2.1, denoted $\operatorname{cost}(Y_n^{(\delta)})$, is proportional to $a + bN$, where a and b are constants and N is the number of iterations the while-loop in the algorithm has to perform until the condition $j = n$ is met. To describe the asymptotic behaviour of MSE and cost, we use the following notation.

Notation. For sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ of positive real numbers we write $a_n \sim b_n$ to indicate that the limit $c = \lim_{n\to\infty} a_n/b_n$ exists and satisfies $0 < c < \infty$.

Our result about the speed of convergence requires the density of $S(X)$ to have continuous third partial derivatives. More specifically, we will use the following technical assumptions on S.

Assumption A. The density f_S and the function $s \mapsto \int_{\mathbb{R}^p} h(t)\, f_{S,\theta}(s,t)\, dt$ are three times continuously differentiable in a neighbourhood of $s^* \in \mathbb{R}^q$.

Theorem 3.2. *Let $h\colon \mathbb{R}^p \to \mathbb{R}$ be such that $\mathbb{E}\big(h(\theta)^2\big) < \infty$ and let S satisfy assumption A. Then, for f_S-almost all s^*, the following statements hold:*

1. *Let $(\delta_n)_{n\in\mathbb{N}}$ be a sequence with $\delta_n \sim n^{-1/4}$. Then the mean squared error satisfies*
$$\operatorname{MSE}(Y_n^{(\delta_n)}) \sim \mathbb{E}\big(\operatorname{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}$$
as $n \to \infty$.

2. *The rate given in the preceding statement is optimal: for any sequence $(\delta_n)_{n\in\mathbb{N}}$ with $\delta_n \downarrow 0$ as $n \to \infty$ we have*
$$\liminf_{n\to\infty} \frac{\operatorname{MSE}(Y_n^{(\delta_n)})}{\mathbb{E}\big(\operatorname{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}} > 0.$$

The rate of convergence in the first part of theorem 3.2 should be compared to the corresponding rate for the usual Monte Carlo estimate. Since Monte Carlo estimates are unbiased, the root-mean squared error for a Monte Carlo estimate is proportional to $1/\sqrt{n}$, while the cost of generating n samples is proportional to n. Thus, for Monte Carlo estimates we have $\operatorname{RMSE} \sim \operatorname{cost}^{-1/2}$. The corresponding exponent from theorem 3.2, obtained by taking square roots, is $-2/(q+4)$, showing slower convergence for ABC estimates. This reduced efficiency is a consequence of the additional error introduced by the bias in the ABC estimates.

The result of theorem 3.2 describes the situation where, for a fixed problem, the tolerance δ is decreased. Caution is advisable when comparing ABC estimates for different q. Since the constant implied in the $\sim$-notation depends on the choice of problem, and thus on q, the dependence of the error on q for fixed δ is not in the scope of theorem 3.2. Indeed, Blum (2010) suggests that in practice ABC methods can be successfully applied for larger values of q, despite results like theorem 3.2.

Before we turn our attention to proving the results stated above, we first remark that, without loss of generality, we can assume that the acceptance criterion in the algorithm uses Euclidean distance $\|\cdot\|$ instead of $\|\cdot\|_A$. This can be achieved by considering the modified statistic $\tilde{S}(x) = A^{-1/2}S(x)$ and $\tilde{s}^* = \tilde{S}(x^*) = A^{-1/2}s^*$, where $A^{-1/2}$ is the positive definite symmetric square root of A^{-1}. Since

$$\begin{aligned}\|S(X) - s^*\|_A^2 &= \big(S(X) - s^*\big)^\top A^{-1}\big(S(X) - s^*\big) \\ &= \Big(A^{-1/2}\big(S(X) - s^*\big)\Big)^\top \Big(A^{-1/2}\big(S(X) - s^*\big)\Big) \\ &= \|\tilde{S}(X) - \tilde{s}^*\|^2,\end{aligned}$$

the ABC algorithm using S and $\|\cdot\|_A$ has identical output to the ABC algorithm using $\tilde{S}$ and the Euclidean norm. Finally, a simple change of variables shows that the assumptions of theorems 3.1 and 3.2 are satisfied for S and $\|\cdot\|_A$ if and only they are satisfied for $\tilde{S}$ and $\|\cdot\|$. Thus, for the proofs we will assume that Euclidean distance is used in the algorithm.

The rest of this section contains the proofs of theorems 3.1 and 3.2. Throughout, we will use the notation introduced in the following definition.

Definition 3.3. For $h\colon \mathbb{R}^p \to \mathbb{R}$ with $\mathbb{E}\big(|h(\theta)|\big) < \infty$ we define $\varphi_h\colon \mathbb{R}^q \to \mathbb{R}$ to be

$$\varphi_h(s) = \int_{\mathbb{R}^p} h(t) f_{S,\theta}(s,t)\,dt$$

for all $s \in \mathbb{R}^q$ and $\varphi_h^{(\delta)}\colon \mathbb{R}^q \to \mathbb{R}$ to be

$$\varphi_h^{(\delta)}(s^*) = \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} \varphi_h(s)\,ds, \tag{3}$$

for all $s^* \in \mathbb{R}^q$, where $\big|B(s^*,\delta)\big|$ denotes the volume of the ball $B(s^*,\delta)$.

Using the definition for $h \equiv 1$ we get $\varphi_1 \equiv f_S$. Both the exact value y from (1) and the mean of the estimator $Y_n^{(\delta)}$ from (2) can be conveniently expressed in the notation of definition 3.3. This is shown in the following lemma.

Lemma 3.4. *Let $h\colon \mathbb{R}^p \to \mathbb{R}$ be such that $\mathbb{E}\big(|h(\theta)|\big) < \infty$. Then*

$$\mathbb{E}\big(h(\theta)\,\big|\,S = s^*\big) = \frac{\varphi_h(s^*)}{\varphi_1(s^*)}$$

and

$$\mathbb{E}\big(Y_n^{(\delta)}\big) = \mathbb{E}\big(h(\theta_1^{(\delta)})\big) = \frac{\varphi_h^{(\delta)}(s^*)}{\varphi_1^{(\delta)}(s^*)}$$

for f_S-almost all $s^* \in \mathbb{R}^q$.

Proof. From the assumption $\mathbb{E}\big(|h(\theta)|\big) < \infty$ we can conclude

$$\begin{aligned}\int_{\mathbb{R}^q}\int_{\mathbb{R}^p} \big|h(t)\big| f_{S,\theta}(s,t)\,dt\,ds &= \int_{\mathbb{R}^p} \big|h(t)\big| \int_{\mathbb{R}^q} f_{S,\theta}(s,t)\,ds\,dt \\ &= \int_{\mathbb{R}^p} \big|h(t)\big| f_\theta(t)\,dt = \mathbb{E}\big(|h(\theta)|\big) < \infty,\end{aligned}$$

and thus we know that $\int_{\mathbb{R}^p} \big|h(t)\big| f_{S,\theta}(s,t)\,dt < \infty$ for almost all $s \in \mathbb{R}^q$. Consequently, the conditional distribution $\mathbb{E}\big(h(\theta)\,\big|\,S = s^*\big)$ exists for f_S-almost all $s^* \in \mathbb{R}^q$. Using Bayes' rule we get

$$\begin{aligned}\mathbb{E}\big(h(\theta)\,\big|\,S = s^*\big) &= \int_{\mathbb{R}^p} h(t) f_{\theta|S}(t|s^*)\,dt \\ &= \int_{\mathbb{R}^p} h(t) \frac{f_{S,\theta}(s^*,t)}{f_S(s^*)}\,dt = \frac{\varphi_h(s^*)}{\varphi_1(s^*)}.\end{aligned}$$

On the other hand, the samples $\theta_j^{(\delta)}$ are distributed according to the conditional distribution of θ given $S \in B(s^*,\delta)$. Thus, the density of the samples $\theta_j^{(\delta)}$ can be written as

$$f_{\theta_j^{(\delta)}}(t) = \frac{1}{Z} \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} f_{S,\theta}(s,t)\,ds,$$

where the normalising constant Z satisfies

$$Z = \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} f_S(s)\,ds = \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} \varphi_1(s)\,ds = \varphi_1^{(\delta)}(s^*),$$

and we get

$$\begin{aligned}\mathbb{E}\big(Y_n^{(\delta)}\big) &= \mathbb{E}\big(h(\theta_1^{(\delta)})\big) \\ &= \frac{1}{\varphi_1^{(\delta)}(s)} \int_{\mathbb{R}^p} h(t) \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} f_{S,\theta}(s,t)\,ds\,dt \\ &= \frac{1}{\varphi_1^{(\delta)}(s)} \frac{1}{\big|B(s^*,\delta)\big|} \int_{B(s^*,\delta)} \varphi_h(s)\,ds \\ &= \frac{\varphi_h^{(\delta)}(s^*)}{\varphi_1^{(\delta)}(s^*)}.\end{aligned}$$

This completes the proof. □

Using these preparations, we can now present a proof of theorem 3.1.

Proof of theorem 3.1. Since $\mathbb{E}\big(|h(\theta)|\big) < \infty$ we have

$$\mathbb{E}\big(|Y_n^{(\delta)}|\big) \le \mathbb{E}\big(|h(\theta_1)|\big) = \frac{\varphi_{|h|}(s^*)}{\varphi_1(s^*)} < \infty$$

whenever $\varphi_1(s^*) = f_S(s^*) > 0$, and by the law of large numbers $Y_n^{(\delta)}$ converges to $\mathbb{E}\big(Y_n^{(\delta)}\big)$ almost surely.

For the second statement, since $\varphi_1 \equiv f_S \in L^1(\mathbb{R}^q)$, we can use the Lebesgue differentiation theorem (Rudin, 1987, theorem 7.7) to conclude that $\varphi_1^{(\delta)}(s^*) \to \varphi_1(s^*)$ as $\delta \downarrow 0$ for almost all $s^* \in \mathbb{R}^q$. Similarly, since

$$\int_{\mathbb{R}^q} \big|\varphi_h(s)\big|\,ds \le \int_{\mathbb{R}^p} \big|h(t)\big| \int_{\mathbb{R}^q} f_{S,\theta}(s,t)\,ds\,dt = \int_{\mathbb{R}^p} \big|h(t)\big| f_\theta(t)\,dt < \infty$$

and thus $\varphi_h \in L^1(\mathbb{R}^q)$, we have $\varphi_h^{(\delta)}(s^*) \to \varphi_h(s^*)$ as $\delta \downarrow 0$ for almost all $s^* \in \mathbb{R}^q$ and using lemma 3.4 we get

$$\lim_{\delta \downarrow 0} \mathbb{E}\big(Y_n^{(\delta)}\big) = \lim_{\delta \downarrow 0} \frac{\varphi_h^{(\delta)}(s^*)}{\varphi_1^{(\delta)}(s^*)} = \frac{\varphi_h(s^*)}{\varphi_1(s^*)} = \mathbb{E}\big(h(\theta) \,\big|\, S = s^*\big)$$

for almost all $s^* \in \mathbb{R}^q$. This completes the proof. □

For later use we also state the following, simple consequence of theorem 3.1.

Corollary 3.5. *Assume that $\mathbb{E}\big(h(\theta)^2\big) < \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^q$ we have*

$$\lim_{\delta \downarrow 0} n \operatorname{Var}(Y_n^{(\delta)}) = \operatorname{Var}\big(h(\theta) \,\big|\, S = s^*\big),$$

uniformly in n.

Proof. From the definition of the variance we know

$$\operatorname{Var}\big(Y_n^{(\delta)}\big) = \frac{1}{n} \operatorname{Var}\big(h(\theta_j^{(\delta)})\big) = \frac{1}{n}\Big(\mathbb{E}\big(h(\theta_j^{(\delta)})^2\big) - \mathbb{E}\big(h(\theta_j^{(\delta)})\big)^2\Big).$$

Applying theorem 3.1 to the function h^2 first, we get $\lim_{\delta \downarrow 0} \mathbb{E}\big(h(\theta_j^{(\delta)})^2\big) = \mathbb{E}\big(h(\theta)^2 \,\big|\, S = s^*\big)$. Since $\mathbb{E}\big(h(\theta)^2\big) < \infty$ implies $\mathbb{E}\big(|h(\theta)|\big) < \infty$, we also get $\lim_{\delta \downarrow 0} \mathbb{E}\big(h(\theta_j^{(\delta)})\big) = \mathbb{E}\big(h(\theta) \,\big|\, S = s^*\big)$ and thus

$$\begin{aligned}\lim_{\delta \downarrow 0} n \operatorname{Var}\big(Y_n^{(\delta)}\big) &= \mathbb{E}\big(h(\theta)^2 \,\big|\, S = s^*\big) - \mathbb{E}\big(h(\theta) \,\big|\, S = s^*\big)^2 \\ &= \operatorname{Var}\big(h(\theta) \,\big|\, S = s^*\big).\end{aligned}$$

This completes the proof. □

The rest of this section is devoted to a proof of theorem 3.2. We first consider the bias of the estimator $Y_n^{(\delta)}$. As is the case for Monte Carlo estimates, the bias of the ABC estimate $Y_n^{(\delta)}$ does not depend on the sample size n. The dependence on the tolerance parameter δ is given in the following lemma. This lemma is the key ingredient in the proof of theorem 3.2.

Lemma 3.6. *Assume that $\mathbb{E}\big(|h(\theta)|\big) < \infty$ and that S satisfies assumption A. Then, for f_S-almost all $s^* \in \mathbb{R}^q$, we have*

$$\operatorname{bias}(Y_n^{(\delta)}) = C(s^*)\, \delta^2 + O(\delta^3)$$

as $\delta \downarrow 0$ where the constant $C(s^*)$ is given by

$$C(s^*) = \frac{\Delta\varphi_h(s^*) - y \cdot \Delta\varphi_1(s^*)}{2(q+2)\varphi_1(s^*)}$$

and Δ denotes the Laplace operator.

Proof. Using lemma 3.4 we can write the bias as

$$\operatorname{bias}(Y_n^{(\delta)}) = \frac{\varphi_h^{(\delta)}(s^*)}{\varphi_1^{(\delta)}(s^*)} - \frac{\varphi_h(s^*)}{\varphi_1(s^*)}. \tag{4}$$

To prove the lemma, we have to study the rate of convergence of the averages $\varphi_h^{(\delta)}(s^*)$ to the centre value $\varphi_h(s^*)$ as $\delta \downarrow 0$. Using Taylor's formula we find

$$\begin{aligned}\varphi_h(s) = \varphi_h(s^*) + \nabla\varphi_h(s^*)(s-s^*) + \frac{1}{2}(s-s^*)^\top H_{\varphi_h}(s^*)(s-s^*) \\ + r_3(s-s^*)\end{aligned}$$

where H_{φ_h} denotes the Hessian of φ_h and the error term r_3 satisfies

$$\big|r_3(v)\big| \le \max_{|\alpha|=3} \sup_{s\in B(s^*,\delta)} \big|\partial_s^\alpha \varphi_h(s)\big| \cdot \sum_{|\beta|=3} \frac{1}{\beta!}\big|v^\beta\big| \tag{5}$$

for all $s \in B(s^*,\delta)$, and ∂_s^α denotes the partial derivative corresponding to the multi-index α. Substituting the Taylor approximation into equation (3), we find

$$\begin{aligned}\varphi_h^{(\delta)}(s^*) &= \frac{1}{\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} \Big(\varphi_h(s^*) + \nabla\varphi_h(s^*)(s-s^*) \\ &\qquad + \frac{1}{2}(s-s^*)^\top H_{\varphi_h}(s^*)(s-s^*) + r_3(s-s^*)\Big)\,ds \\ &= \varphi_h(s^*) + 0 \\ &\quad + \frac{1}{2\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} (s-s^*)^\top H_{\varphi_h}(s^*)(s-s^*)\,ds \\ &\quad + \frac{1}{\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} r_3(s-s^*)\,ds.\end{aligned} \tag{6}$$

Since the Hessian $H_{\varphi_h}(s^*)$ is symmetric and since the domain of integration is invariant under rotations we can choose a basis in which $H_{\varphi_h}(s^*)$ is diagonal, such that the diagonal elements coincide with the eigenvalues $\lambda_1,\dots,\lambda_q$. Using this basis we can write the quadratic term in (6) as

$$\begin{aligned}&\frac{1}{2\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} (s-s^*)^\top H_{\varphi_h}(s^*)(s-s^*)\,ds \\ &\quad= \frac{1}{2\big|B(0,\delta)\big|}\int_{B(0,\delta)} \sum_{i=1}^q \lambda_i u_i^2\,du \\ &\quad= \frac{1}{2\big|B(0,\delta)\big|}\sum_{i=1}^q \lambda_i \frac{1}{q}\sum_{j=1}^q \int_{B(0,\delta)} u_j^2\,du \\ &\quad= \frac{1}{2\big|B(0,\delta)\big|}\sum_{i=1}^q \lambda_i \frac{1}{q}\int_{B(0,\delta)} |u|^2\,du \\ &\quad= \operatorname{tr} H_{\varphi_h}(s^*)\cdot\frac{1}{2q\big|B(0,\delta)\big|}\int_{B(0,\delta)} |u|^2\,du\end{aligned}$$

where $\operatorname{tr} H_{\varphi_h} = \Delta\varphi_h$ is the trace of the Hessian. Here we used the fact that the value $\int_{B(0,\delta)} u_i^2\,du$ does not depend on i and thus equals the average $\frac{1}{q}\sum_{j=1}^{q}\int_{B(0,\delta)} u_j^2\,du$. Rescaling space by a factor $1/\delta$ and using the relation $\int_{B(0,1)} |x|^2\,dx = \big|B(0,1)\big| \cdot q/(q+2)$ we find

$$\begin{aligned}
&\frac{1}{2\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} (s-s^*)^\top H_{\varphi_h}(s^*)(s-s^*)\,ds \\
&\quad = \Delta\varphi_h(s^*)\cdot\frac{1}{2q\delta^q\big|B(0,1)\big|}\int_{B(0,1)} |y|^2\delta^{q+2}\,dy \\
&\quad = \frac{\Delta\varphi_h(s^*)}{2(q+2)}\delta^2.
\end{aligned}$$

For the error term we can use a similar scaling argument in the bound (5) to get

$$\frac{1}{\big|B(s^*,\delta)\big|}\int_{B(s^*,\delta)} \big|r_3(s-s^*)\big|\,ds \le C\cdot\delta^3$$

for some constant C. Substituting these results back into equation (6) we find

$$\varphi_h^{(\delta)}(s^*) = \varphi_h(s^*) + a_h(s^*)\cdot\delta^2 + \mathcal{O}(\delta^3) \tag{7}$$

where

$$a_h(s^*) = \frac{\Delta\varphi_h(s^*)}{2(q+2)}.$$

Using formula (7) for $h \equiv 1$ we also get

$$\varphi_1^{(\delta)}(s^*) = \varphi_1(s^*) + a_1(s^*)\cdot\delta^2 + \mathcal{O}(\delta^3)$$

as $\delta \downarrow 0$.

Using representation (4) of the bias, and omitting the argument s^* for brevity, we can now express the bias in powers of δ:

$$\begin{aligned}
\operatorname{bias}(Y_n^{(\delta)}) &= \frac{\varphi_1\varphi_h^{(\delta)} - \varphi_1^{(\delta)}\varphi_h}{\varphi_1^{(\delta)}\varphi_1} \\
&= \frac{\varphi_1\Big(\varphi_h + a_h\delta^2 + \mathcal{O}(\delta^3)\Big) - \Big(\varphi_1 + a_1\delta^2 + \mathcal{O}(\delta^3)\Big)\varphi_h}{\varphi_1^{(\delta)}\varphi_1} \\
&= \frac{\varphi_1 a_h - \varphi_h a_1}{\varphi_1^{(\delta)}\varphi_1}\cdot\delta^2 + \mathcal{O}(\delta^3).
\end{aligned}$$

Since $1/\varphi_1^{(\delta)} = 1/\varphi_1 \cdot \big(1 + \mathcal{O}(\delta^2)\big)$ and $y = \varphi_h/\varphi_1$, the right-hand side can be simplified to

$$\operatorname{bias}(Y_n^{(\delta)}) = \frac{a_h(s^*) - y a_1(s^*)}{\varphi_1(s^*)}\cdot\delta^2 + \mathcal{O}(\delta^3).$$

This completes the proof. □

To prove the statement of theorem 3.2, the bias of $Y_n^{(\delta)}$ has to be balanced with the computational cost for computing $Y_n^{(\delta)}$. When δ decreases, fewer

samples will satisfy the acceptance condition $\|S(X) - s^*\| \leq \delta$ and the running time of the algorithm will increase. The following lemma makes this statement precise.

Lemma 3.7. *Let f_S be continuous at s^*. Then the expected computational cost for computing the estimate $Y_n^{(\delta)}$ satisfies*

$$\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta)})\big) = c_1 + c_2 n \delta^{-q}\big(1 + c_3(\delta)\big)$$

for all $n \in \mathbb{N}$, $\delta > 0$ where c_1 and c_2 are constants, and c_3 does not depend on n and satisfies $c_3(\delta) \to 0$ as $\delta \downarrow 0$.

Proof. The computational cost for algorithm 2.1 is of the form $a + bN$ where a and b are constants, and N is the random number of iterations of the loop required until n samples are accepted. The number of iterations required to generate one ABC sample is geometrically distributed with parameter

$$p = P\big(S \in B(s^*, \delta)\big)$$

and thus N, being the sum of n independent geometrically distributed values, has mean $\mathbb{E}(N) = n/p$. The probability p can be written as

$$p = \int_{B(s^*,\delta)} f_S(s)\, ds = \big|B(s^*, \delta)\big| \cdot \varphi_1^{(\delta)}(s^*) = \delta^q \big|B(s^*, 1)\big| \cdot \varphi_1^{(\delta)}(s^*).$$

Since $\varphi_1 = f_S$ is continuous at s^*, we have $\varphi_1^{(\delta)}(s^*) \to \varphi_1(s^*)$ as $\delta \downarrow 0$ and thus $p = c\delta^q\big(1 + o(1)\big)$ for some constant c. Thus, we find that the computational cost satisfies

$$\begin{aligned}
\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta)})\big) &= a + b \cdot \mathbb{E}(N) \\
&= a + b \cdot \frac{n}{c\delta^q\big(1 + o(1)\big)} \\
&= a + \frac{b}{c}\, n\delta^{-q}\, \frac{1}{1 + o(1)},
\end{aligned}$$

and the proof is complete. □

Finally, the following two lemmata give the relation between the approximation error caused by the bias on the one hand and the expected computational cost on the other hand.

Lemma 3.8. *Let $\delta_n \sim n^{-1/4}$ for $n \in \mathbb{N}$. Assume that $\mathbb{E}\big(h(\theta)^2\big) < \infty$ and that S satisfies assumption A. Then, for f_S-almost all $s^* \in \mathbb{R}^q$, the error satisfies $\mathrm{MSE}(Y_n^{(\delta_n)}) \sim n^{-1}$, the expected computational cost increases as $\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big) \sim n^{1+q/4}$ and we have*

$$\mathrm{MSE}(Y_n^{(\delta_n)}) \sim \mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}$$

as $n \to \infty$.

Proof. By assumption, the limit $D = \lim \delta_n n^{1/4}$ exists. Using lemma 3.6 and corollary 3.5, we find

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathrm{MSE}(Y_n^{(\delta_n)})}{n^{-1}} &= \lim_{n\to\infty} n\Big(\mathrm{Var}(Y_n^{(\delta_n)}) + \big(\mathrm{bias}(Y_n^{(\delta_n)})\big)^2\Big) \\
&= \lim_{n\to\infty} n\,\mathrm{Var}(Y_n^{(\delta_n)}) + \lim_{n\to\infty} n\big(C(s^*)\,\delta_n^2 + O(\delta_n^3)\big)^2 \\
&= \mathrm{Var}\big(h(\theta)\,\big|\,S = s^*\big) + \lim_{n\to\infty}\Big(C(s^*) + \frac{\mathcal{O}(\delta_n^3)}{\delta_n^2}\Big)^2 (\delta_n n^{1/4})^4 \\
&= \mathrm{Var}\big(h(\theta)\,\big|\,S = s^*\big) + C(s^*)^2 D^4.
\end{aligned}$$

On the other hand, using lemma 3.7, we find

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)}{n^{1+q/4}} &= \lim_{n\to\infty} \frac{c_1 + c_2 n \delta_n^{-q}\big(1 + c_3(\delta_n)\big)}{n^{1+q/4}} \\
&= 0 + c_2 \lim_{n\to\infty} \frac{1}{(\delta_n n^{1/4})^q}\big(1 + \lim_{n\to\infty} c_3(\delta_n)\big) \\
&= c_2 / D^q.
\end{aligned}$$

Finally, combining the rates for cost and error we get the result

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\mathrm{MSE}(Y_n^{(\delta_n)})}{\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}} \\
&\quad= \lim_{n\to\infty} \frac{\mathrm{MSE}(Y_n^{(\delta_n)})}{n^{-1}} \Big/ \lim_{n\to\infty} \Big(\frac{\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)}{n^{1+q/4}}\Big)^{-4/(q+4)} \\
&\quad= \Big(\mathrm{Var}\big(h(\theta)\,\big|\,S = s^*\big) + C(s^*)^2 D^4\Big) \cdot \Big(\frac{D^q}{c_2}\Big)^{-4/(q+4)}.
\end{aligned} \tag{8}$$

This completes the proof. □

Lemma 3.8 shows that the rate of decay of the error as given in the first part of theorem 3.2 is indeed achieved by the choice $\delta_n \sim n^{-1/4}$. The following result completes the proof of the theorem by showing that no other choice of δ_n can lead to a better rate for the error while retaining the same cost.

Lemma 3.9. *For $n \in \mathbb{N}$ let $\delta_n \downarrow 0$. Assume that $\mathbb{E}\big(h(\theta)^2\big) < \infty$ and that S satisfies assumption A. Then, for f_S-almost all $s^* \in \mathbb{R}^q$, we have*

$$\liminf_{n\to\infty} \frac{\mathrm{MSE}(Y_n^{(\delta_n)})}{\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}} > 0.$$

Proof. From lemma 3.6 we know that

$$\begin{aligned}
\mathrm{MSE}(Y_n^{(\delta_n)}) &= \mathrm{Var}(Y_n^{(\delta_n)}) + \big(\mathrm{bias}(Y_n^{(\delta_n)})\big)^2 \\
&= \frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n} + \big(C(s^*)\,\delta_n^2 + \mathcal{O}(\delta_n^3)\big)^2 \\
&= \frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n} + C(s^*)^2\,\delta_n^4 + \mathcal{O}(\delta_n^5) \\
&\geq \frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n} + \frac{C(s^*)^2}{2}\,\delta_n^4
\end{aligned} \tag{9}$$

for all sufficiently large n. By lemma 3.7, as $n \to \infty$, we have

$$\begin{aligned}
&\mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)} \\
&\quad \sim (n\delta^{-q})^{-4/(q+4)} \\
&\quad = \big(n^{-1}\big)^{4/(q+4)} \big(\delta^4\big)^{q/(q+4)} \\
&\quad \sim \Big(\frac{4}{q+4}\frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n}\Big)^{4/(q+4)} \Big(\frac{q}{q+4}\,\frac{C(s^*)^2}{2}\delta^4\Big)^{q/(q+4)}
\end{aligned} \tag{10}$$

where we were able to insert the constant factors into the last term because the $\sim$-relation does not see constants. Using Young's inequality we find

$$\begin{aligned}
&\frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n} + \frac{C(s^*)^2}{2}\delta^4 \\
&\quad \geq \Big(\frac{4}{q+4}\frac{\mathrm{Var}\big(h(\theta_j^{(\delta)})\big)}{n}\Big)^{4/(q+4)} \Big(\frac{q}{q+4}\,\frac{C(s^*)^2}{2}\delta^4\Big)^{q/(q+4)}
\end{aligned} \tag{11}$$

and combining equations (9), (10) and (11) we get

$$\mathrm{MSE}(Y_n^{(\delta_n)}) \geq c \cdot \mathbb{E}\big(\mathrm{cost}(Y_n^{(\delta_n)})\big)^{-4/(q+4)}$$

for some constant $c > 0$ and all sufficiently large n. This is the required result. □

The statement of theorem 3.2 coincides with the statements of lemmata 3.8 and 3.9 and thus we have completed the proof of theorem 3.2.

4 Numerical Experiments

To illustrate the results from section 3, we present a series of numerical experiments for the following toy problem.

- We choose $p = 1$, and assume that our prior belief in the value of the single parameter θ is standard normally distributed.
- We choose $d = 2$, and assume the data X to be composed of i.i.d. samples X_1, X_2 each with distribution $X_i \,|\, \theta \sim \mathcal{N}(\theta, 1)$.
- We choose $q = 2$, and the (non-minimal) sufficient statistic to be $S(x) = x$ for all $x \in \mathbb{R}^2$.
- We consider the test function $h(\theta) = 1_{[-1/2,1/2]}(\theta)$, *i.e.* the indicator function for the region $[-1/2, 1/2]$. The ABC estimate is thus an estimate for the posterior probability $P\big(\theta \in [-1/2, 1/2] \;\big|\; S = s^*\big)$.
- We fix the observed data to be $s^* = (1, 1)$.

This problem is simple enough that all quantities of interest can be determined explicitly. In particular, $\theta|S$ is $\mathcal{N}\big((s_1 + s_2)/3, 1/3\big)$ distributed, $\theta|S{=}s^*$

is $\mathcal{N}(2/3, 1/3)$ distributed, and S is bivariate normally distributed with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Therefore, the prior and posterior expectation for $h(\theta)$ are $\mathbb{E}\big(h(\theta)\big) = 0.3829$ and $\mathbb{E}\big(h(\theta) \,\big|\, S = s^*\big) = 0.3648$, respectively. Similarly, the constant from lemma 3.6 can be shown to be $C(s^*) = 0.0323$.

Assumption A can be shown to hold. The function

$$\varphi_1(s) = f_S(s) = \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{3}(s_1^2 - s_1 s_2 + s_2^2)}$$

is multivariate normal, so its third derivatives exist, and are bounded and continuous. Similarly, the function

$$\varphi_h(s) = \int_{-1/2}^{1/2} f_{\theta,S}(t, s)dt \leq \varphi_1(s)$$

also has bounded and continuous third derivatives. Thus, the assumptions hold.

The programs used to perform the simulations described in this section, written using the R environment (R Core Team, 2013) are available as supplementary material.

Experiment 1

Our first experiment validates the statement about the bias given in lemma 3.6. Since we know the exact posterior expectation, the bias can be determined experimentally. For fixed δ, we generate k independent ABC estimates, each based on n proposals. For each of the k estimates, we calculate its distance from the true posterior expectation. We then calculate the mean and standard error of these differences to obtain a Monte Carlo estimate of the bias.

Repeating this procedure for several values of δ, we can produce a plot of the estimated bias against δ, with 95% error bars. Figure 1 shows the result of a simulation, using $n = 500$ samples for each ABC estimate and $k = 5000$ ABC estimates for each value of δ. For comparison, the plot includes the theoretically predicted asymptotic bias $C(s^*)\delta^2$, using the value $C(s^*) = 0.0323$. The plot shows that for small values of δ the theoretical curve is indeed a good fit to the numerical estimates of the bias. The result of the lemma is only valid as $\delta \downarrow 0$ and indeed the plot shows a discrepancy for larger values. This discrepancy is caused by only a small fraction of the sample being rejected; as δ increases, the distribution of the ABC samples approaches the prior distribution.

Experiment 2

Our second experiment validates the statement of theorem 3.2, by numerically estimating the optimal choice of delta and the corresponding MSE.

For fixed values of expected computational cost and δ, we estimate the mean squared error by generating k different ABC estimates and taking the mean of their squared distance from the true posterior expectation. This reflects how

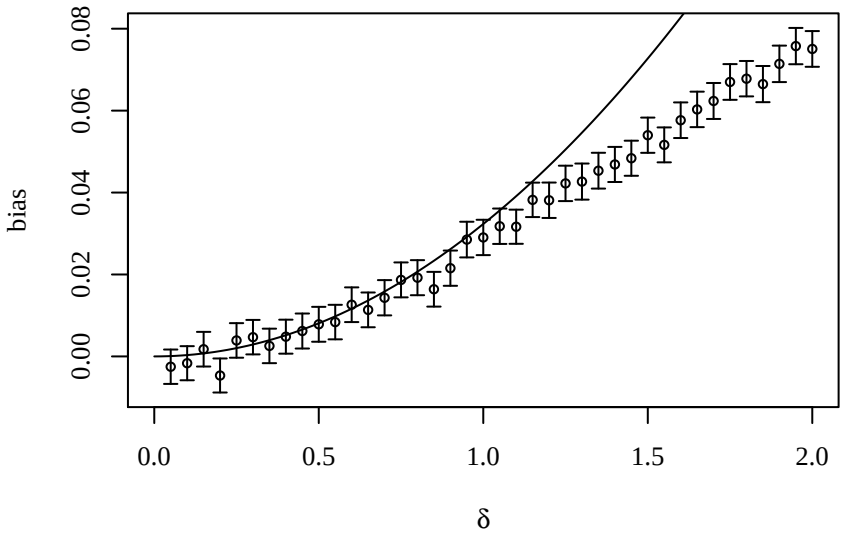

Figure 1: Simulation results illustrating the relationship between bias and δ. Circles give the mean empirical bias from 5000 simulations for each value of δ. The error bars indicate mean ± 1.96 standard errors. The solid line shows the theoretically predicted asymptotic bias from lemma 3.6.

the bias is estimated in experiment 1. Repeating this procedure for several values of δ, the estimates of the MSE are plotted against δ.

Our aim is to determine the optimal value of δ for fixed computational cost. From lemma 3.7 we know that the expected cost is of order $n\delta^{-q}$ and thus we choose $n \sim \delta^2$ in this example. From lemma 3.6 we know that bias $\sim \delta^2$. Thus, we expect the MSE for constant expected cost to be of the form

$$\mathrm{MSE}(\delta) = \frac{\mathrm{Var}}{n} + \mathrm{bias}^2 = a\delta^{-2} + b\delta^4 \tag{12}$$

for some constants a and b. Thus, we fit a curve of this form to the numerically estimated values of the MSE. The result of one such simulation, using $k = 500$ samples for each δ, is shown in figure 2. The curve fits the data well.

We estimate the optimal values of δ and MSE, given an expected computational cost, to be those at the minimum of the fitted curve. Given the good fit between the predicted form (12) of the curve and the empirical MSE values, this procedure promises to be a robust way to estimate the optimal value of δ. The direct approach would likely require a much larger number of samples to be accurate.

Repeating the above procedure for a range of values of expected cost gives corresponding estimates for the optimal values of δ and the MSE as a function of expected cost. We expect the optimal δ and the MSE to depend on the cost like $x = A \cdot \mathrm{cost}^B$. To validate the statements of theorem 3.2 we numerically estimate the exponent B from simulated data. The result of such a simulation is shown in figure 3. The data are roughly on straight lines, as expected, and

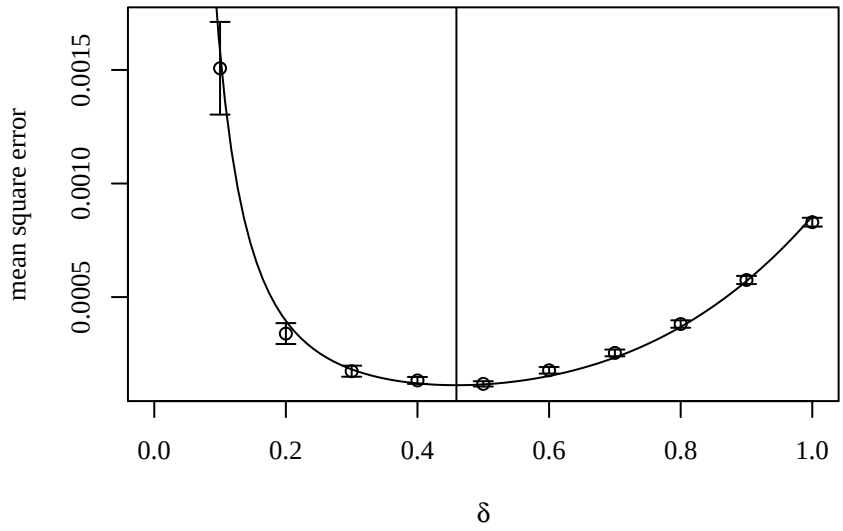

Figure 2: The estimated MSE as a function of the tolerance δ for fixed expected cost. The fitted curve has the form given in equation (12). The location of the optimal δ is marked by the vertical line.

the gradients are close to the theoretical gradients, shown as smaller lines. The numerical results for estimating the exponent B are given in the following table.

Plot	Gradient	Standard error	Theoretical gradient
δ	-0.167	0.0036	$-1/6 \approx -0.167$
MSE	-0.671	0.0119	$-2/3 \approx -0.667$

The table shows an an excellent fit between empirical and theoretically predicted values.

5 Discussion

While the work of this article is mostly theoretical in nature, the results can be used to guide the design of an analysis using ABC methods.

Lemma 3.8 only specifies the optimal *rate* for the decay of δ_n as a function of n. By inspecting the proof, we can derive the corresponding optimal constant: δ_n should be chosen as $\delta_n = Dn^{-1/4}$ where D minimises the expression in equation (8). The optimal value of D can analytically be found to be

$$D_{\text{opt}} = \Big(\frac{q \operatorname{Var}\big(h(\theta) \,\big|\, S = s^*\big)}{4C(s^*)^2} \Big)^{1/4},$$

where $C(s^*)$ is the constant from lemma 3.6. The variance $\operatorname{Var}\big(h(\theta) \,\big|\, S = s^*\big)$ is easily estimated, but it seems difficult to estimate $C(s^*)$ in a likelihood-free way.

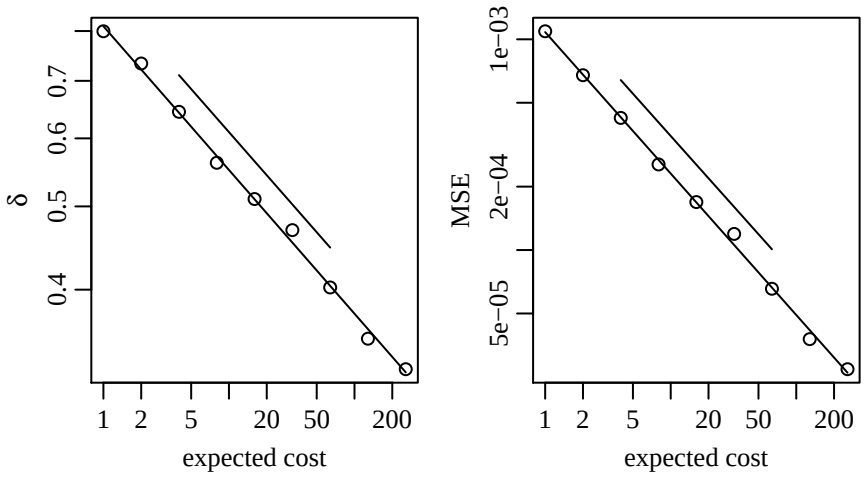

Figure 3: Estimated dependency of the optimal δ and of the corresponding MSE on the computational cost. The computational cost is given in arbitrary units, chosen such that the smallest sample size under consideration has cost 1. For comparison, the additional line above the fit has the gradient expected from the theoretical results.

However, our results can be applied directly in cases where a pilot run is used to tune the algorithm. The optimal rate, from theorem 3.2, is given by

$$\text{error} \sim \text{cost}^{-2/(q+4)}.$$

and optimality is achieved by choosing δ proportional to $n^{-1/4}$. Consequently, if the full run is performed by increasing the number of accepted ABC samples by a factor of k, then the tolerance δ for the full run should be divided by $k^{1/4}$.

The computational cost of the ABC algorithm depends both on the number n of samples accepted and on the tolerance δ. From lemma 3.7 we know that the expected computational cost is asymptotically proportional to $n\delta^{-q}$. This should be taken into account when changing n and δ simultaneously as suggested above. There are two possible scenarios:

- If we want to reduce the root mean-squared error by a factor of α, we should increase the number n of accepted ABC samples by α^2 and reduce the tolerance δ by a factor of $\sqrt{\alpha}$. These changes will increase the expected running time of the algorithm by a factor of $\alpha^{(q+4)/2}$.

- If we plan to increase the (expected) running time by a factor of β, we should increase the number of accepted samples by a factor of $\beta^{4/(q+4)}$ and divide δ by $\beta^{1/(q+4)}$. These changes will reduce the root mean squared error by a factor of $\beta^{2/(q+4)}$.

These guidelines will lead to a choice of parameters which, at least asymptotically, maximises the efficiency of the analysis.

Acknowledgements. The authors thank Alexander Veretennikov for pointing out that the Lebesgue differentiation theorem can be used in the proof of theorem 3.1. MW was supported by an EPSRC Doctoral Training Grant at the School of Mathematics, University of Leeds.

References

Mark A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, December 2010. doi: 10.1146/annurev-ecolsys-102209-144621.

Gerard Biau, Frédéric Cérou, and Arnaud Guyader. New insights into approximate Bayesian computation. *Annales de l'Institut Henri Poincaré*, 2013. In press.

Michael G. Blum and Olivier François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20:63–73, 2010. doi: 10.1007/s11222-009-9116-0.

Michael G. Blum and Viet Chi Tran. HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, 11(4):644–660, 2010.

Michael G. B. Blum. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, September 2010. doi: 10.1198/jasa.2010.tm09448.

P. Bortot, Stuart G. Coles, and Scott A. Sisson. Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92, 2007.

Christopher C. Drovandi and Anthony N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

Nelson J. R. Fagundes, Nicolas Ray, Mark Beaumont, Samuel Neuenschwander, Francisco M. Salzano, Sandro L. Bonatto, and Laurent Excoffier. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45):17614–17619, 2007.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74(3):419–474, 2012. doi: 10.1111/j.1467-9868.2011.01010.x.

T. Guillemaud, Mark A. Beaumont, M. Ciosi, J.-M. Cornuet, and A. Estoup. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, 104(1):88–99, 2010. doi: 10.1038/hdy.2009.92.

Joao S. Lopes and Sanne Boessenkool. The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conservation Genetics*, 11(2):421–433, 2010.

Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22 (6):1167–1180, 2012. doi: 10.1007/s11222-011-9288-2.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

Oliver Ratmann, Christophe Andrieu, Carsten Wiuf, and Sylvia Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581, 2009.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, third edition, 1987.

Daniel Silk, Sarah Filippi, and Michael P. H. Stumpf. Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology*, 12(5):603–618, September 2013. doi: 10.1515/sagmb-2012-0043.

S.A. Sisson, Y. Fan, and M. N. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007. doi: 10.1073/pnas.0607208104.

Andrea Sottoriva and Simon Tavaré. Integrating approximate Bayesian computation with complex agent-based models for cancer research. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 57–66. Springer, 2010. doi: 10.1007/978-3-7908-2604-3_5.

Mark M. Tanaka, Andrew R. Francis, Fabio Luciani, and S. A. Sisson. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3):1511–1520, 2006.

Simon Tavaré, David J. Balding, R. C. Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

Kevin Thornton and Peter Andolfatto. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics*, 172(3):1607–1619, 2006.

Jochen Voss. *An Introduction to Statistical Computing: A Simulation-Based Approach*. Wiley Series in Computational Statistics. Wiley, 2014. ISBN 978-1118357729.

David M. Walker, David Allingham, Heung Wing Joseph Lee, and Michael Small. Parameter inference in small world network disease models with approximate Bayesian computational methods. *Physica A*, 389(3):540–548, 2010.

Daniel Wegmann and Laurent Excoffier. Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution*, 27(6):1425–1435, 2010. doi: 10.1093/molbev/msq028.

Richard D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141, 2013.

Richard D. Wilkinson, Michael E. Steiper, Christophe Soligo, Robert D. Martin, Ziheng Yang, and Simon Tavaré. Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology*, 60(1):16–31, 2011.