Convergence Properties of Approximate Bayesian Computation

Mark Graham Moody Webster

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Mathematics

July 2016

Declaration

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The results in Sections 3.1, 3.2, 3.3, and 3.4.1, with the exception of Lemmas 3.10 and 3.11 and Theorems 3.12 and 3.13, were published in S. Barber, J. Voss, and M. Webster, The rate of convergence for approximate Bayesian computation, Electronic Journal of Statistics, 9:80–105, 2015. The mentioned exceptions are an alternative proof for the results in Section 4 of Barber et al. [2015].

The results in Section 3.1 were proven, and written, by JV. The numerical experiments in Section 3.3 were done by MW. All other results from the sections mentioned above were drafted by MW, and include contributions and edits from all three authors.

Abstract

Approximate Bayesian Computation is a family of Monte Carlo methods used for likelihood-free Bayesian inference, where calculating the likelihood is intractable, but it is possible to generate simulated data, and calculate summary statistics. While these methods are easy to describe and implement, it is not trivial to optimise the mean square error of the resulting estimate.

This thesis focuses on asymptotic results for the rate of convergence of ABC to the true posterior expectation as the expected computational cost increases. Firstly, we examine the asymptotic efficiency of the “basic” versions of ABC, which consists of proposal generation, followed by a simple accept-reject step. We then look at several simple extensions, including the use of a random accept-reject step, and the use of ABC to make kernel density estimates.

The asymptotic convergence rate of the basic versions of ABC decreases as the summary statistic dimension increases. A naïve conclusion from this result would be that, for an infinite-dimensional summary statistic, the ABC estimate would not converge. To show this need not be the case, we look at the asymptotic behaviour of ABC in the case of an observation that consists of a stochastic process over a fixed time interval. We find partial results for two different criteria for accepting proposals.

We also introduce a new variant of ABC, referred to in the thesis as the ABCLOC estimate. This belongs to a family of variants, in which the parameter proposals are adjusted, to reduce the difference between the distribution of the accepted proposals and the true posterior distribution. The ABCLOC estimate does this using kernel regression. We give preliminary results for the asymptotic behaviour of the ABCLOC estimate, showing that it potentially has a faster asymptotic rate of convergence than the basic versions for high-dimensional summary statistics.

Acknowledgements

Many thanks to my supervisors, Jochen Voss and Stuart Barber, for all of their help and advice over the last three and a half years. Both have provided valuable knowledge in Probability and Statistics, as well as helpful advice on programming with R and typesetting with LaTeX. They have both been gracious and patient with my sometimes elementary mistakes, and helped ease the realisation that doing research can end with you feeling more stupid than you did when you started.

Many thanks to my family: my wife Julie-Anne, my daughter Eva, and my parents Clive and Philippa, for their support and patience while I have been doing this work. Thanks also go to friends and colleagues both inside and outside the university, who have made the time pass much more enjoyably.

This thesis would not have been possible without funding from EPSRC.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Monte Carlo methods, where a large set of random samples is used rather than an exact calculation, have become increasingly popular with the increase in computer processing power, and with the desire to use more complex models. This allows us to approach problems that are infeasible to solve exactly.

Often we have some observation from a process, and some model of the process with unknown model parameters. Ideally, we start with a Bayesian prior over possible values for those parameters, and use the observation to update our posterior belief in the values. This requires use of the likelihood function, which is not available, or not tractable, in complicated problems. Instead, we can take random data samples from model simulations, and use these to inform our updating.

One family of such Monte Carlo methods has become known as *Approximate Bayesian Computation*, or ABC, and contains many variants with a simple common idea. Put briefly, we draw random proposed values for the model parameters from our prior parameter distribution, and use these in the model to generate new data samples. We then compare these new data samples to the original observation, and use this comparison to inform how we use their associated parameter value proposals to form our estimate for the property of interest for the posterior distribution.

The simplicity – or naïveté – of the approach results in easy implementation, even in problems with complex models, but efficient implementation has several significant difficulties. In addition to issues common to all of Bayesian statistics, such as choice of prior, the most prominent difficulties are the following.

- The algorithm is highly computationally inefficient, especially if the data used is high-dimensional. The latter fact is an instance of a problem referred to as the *curse of dimensionality*, and is commonly mitigated by using a lower-dimension summary statistic, rather than the data, to make the comparisons between the samples and the observation. However, usually the chosen summary statistic is not sufficient, so information is lost in the process. This affects the error of the algorithm in a way that cannot easily be analysed.

- The algorithm usually requires the choice of a "tolerance" parameter, or similar, that determines the algorithm's tolerance for discrepancies between the samples and the observation. We would like to choose the tolerance value that minimises the expected error of the estimate. However, unless we consider some specific problem, finding this value is difficult. Some variants use a k-nearest neighbours approach to determine which parameter proposals are used, rather than a tolerance parameter, but this raises a similar question regarding the choice of k.

This text solely addresses the latter problem, assuming any summary statistics used are sufficient. In Chapter 2, we define the basic form of the ABC algorithm, and the class of properties of the posterior parameter distribution that we will be interested in estimating. We will be interested in minimising the error of the ABC estimate for these properties by choosing a tolerance value, so we also define the measure of error. This measure, the mean square error, has a natural decomposition into two sources of error. Changing the tolerance value will decrease the error from one source, but increase the error from the other, so choosing a tolerance value is a problem of balancing the error from these two sources. We also define some variants on the ABC algorithm, which we return to later, and discuss previous results for the error.

In Chapter 3, we consider whether, and how quickly, the ABC estimate converges to the correct answer as the available computational running time increases. Finding the expected error for an ABC estimate is difficult, because it is very problem-specific. However, we can more easily consider asymptotic results, where the estimate is shown to converge to the truth at a certain rate, at worst, as the running time increases to infinity. We find conditions under which the estimate converges to the correct value, and find the asymptotic rate of the convergence. We also find the asymptotic rate of convergence for the variants given in Chapter 2. In practice, the running time might be small enough that these asymptotic rates are not dominant, but they are useful when considering which aspects of the problem are important. For example, the rates of convergence cast light on the curse of dimensionality, by showing how we expect the dimensionality of the summary statistic to affect the estimate's rate of convergence.

When a parameter proposal is used in an ABC estimate, error is introduced, due to the fact that the associated summary statistic does not have to be equal to the observation. One class of ABC variants, the first of which was proposed by Beaumont et al. [2002], attempts to mitigate this error, by adjusting the proposals, according to the difference between the sample and the observation, before using them in an estimate. In Chapter 4, we propose a new variant in this class. Where Beaumont et al. determined the proposal adjustments using a single kernel regression, the new variant does a kernel regression centred at the observation, plus another regression for each of the accepted proposals. This substantially increases the computational cost. However, we also present some early theoretical results, which indicate that this variant might have a faster rate of convergence than the basic ABC algorithm for summary statistics with a high dimension. It is also hoped that the new variant will be more robust when the posterior expectation function is non-linear.

When using asymptotic rates of convergence, it must be remembered that the "big O" notation, that is usually used to denote asymptotic rates, does not conserve the relevant proportionality constants, which can depend on variables that are kept constant in the asymptotic analysis. Therefore, when using asymptotic results, it must be kept in mind what is tending to a limit, and what is kept constant. For example, the asymptotic rates in Chapter 3 assume that the dimensionality of the summary statistic is kept constant. Ignoring this leads to the naïve suggestion that, for infinite-dimensional summary statistics, the algorithm would not converge at all. In Chapter 5, we look at an example problem where the statistic is infinite-dimensional, but the algorithm still converges. Specifically, we look at the case where the observed statistic is a

Brownian motion, plus a linear trend with unknown rate. This rate is then the parameter of interest. While we do not find the asymptotic rate of convergence for this problem, we do find bounds for the rate, so it is shown that the ABC algorithm still converges.

Chapter 2

Introduction to ABC

In this chapter, we introduce the basic version of the ABC algorithm, the type of problems it is used to approach, and the issues considered when optimising its use.

2.1 Method and Variants

2.1.1 Basic Method

We suppose that we are interested in some process that has produced the observed data x^*. Using some summary statistic function s, we describe this data with the observed summary statistic $s^* := s(x^*) \in \mathbb{R}^q$. We have some model function $f_{S|\theta}$ for the probability density of observed data, given some model parameter $\theta \in \mathbb{R}^p$. However, θ is unknown. Instead, we have some prior density f_θ that describes our belief in possible values for the parameter. We wish to obtain the posterior density $f_{\theta|S}(\cdot \,|\, s^*)$ for the parameter condition on the observed summary statistic. Alternatively, instead of the entire posterior density, we may be interested in certain properties of the posterior density, such as the expectation.

Before we proceed, we define some notation for concepts that we will refer to regularly.

Definition 2.1. *We refer to the summary statistic as the* statistic*, eschewing the word "summary". For example, we refer to s^* as the observed statistic.*

Notation 2.2. *Any expectation $\mathbb{E}\left(\cdot \,|\, S = s\right)$ that is conditional on a value s of the statistic will be written as $\mathbb{E}\left(\cdot \,|\, s\right)$. If some parameter θ has density f*

conditional on another parameter S being equal to s, we write $\theta \,|\, s \sim f$, rather than $\theta \,|\, S = s \sim f$. An event is said to happen for f_S-almost all s if the set of values of s for which the event happens has measure one with respect to the marginal distribution function of S.

Definition 2.3. *A* parameter proposal *is a sample value for the parameter θ sampled from the prior parameter distribution. A* data sample *is a set of data generated with a model simulation. A* statistic sample *is the summary statistic of a data sample. The word* statistic *will be used to mean summary statistic, and will be assumed to refer to a sufficient statistic, unless stated otherwise. A* sample *consists of a parameter proposal, and the statistic sample for the data sample generated using the parameter proposal for parameter values.*

We assume, unless stated otherwise, that we are interested in the posterior mean

$$m(s^*) := \mathbb{E}\left(h(\theta) \,|\, s^*\right) = \mathbb{E}\left(h(\theta) \,|\, X = x^*\right) \tag{2.1}$$

of some function $h\colon \mathbb{R}^p \to \mathbb{R}$ of interest on the parameter θ. We also assume that calculating the conditional density $f_{S|\theta}$ is infeasible or impossible. Therefore, it is also infeasible to calculate the true posterior distribution, which satisfies, by Bayes's Theorem,

$$f_{\theta|S}(t \,|\, s) \propto f_\theta(t) f_{S|\theta}(s \,|\, t).$$

Instead, we would like an estimate Z for the value of $m(s^*)$. Finally, we assume that we assess the estimate Z according to its mean square error (MSE),

$$\begin{aligned}
\text{MSE}(Z) &:= \mathbb{E}\left((Z - m(s^*))^2\right) \\
&= \mathbb{E}\left((Z - \mathbb{E}(Z) + \mathbb{E}(Z) - m(s^*))^2\right) \\
&= \mathbb{E}\left((Z - \mathbb{E}(Z))^2\right) + (\mathbb{E}(Z) - m(s^*))^2 \\
&= \text{Var}(Z) + \text{bias}(Z)^2 .
\end{aligned}$$

For example, if we can take n proposals from the true posterior distribution $f_{\theta|S}(\cdot \,|\, s^*)$, and the estimate Z is their mean, then Z has expectation

$$\mathbb{E}(Z) = \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} h(\theta_k) \,\middle|\, s^*\right) = \mathbb{E}\left(h(\theta) \,|\, s^*\right) = m(s^*).$$

Therefore, Z is unbiased, and has MSE equal to the variance

$$\text{Var}(Z) = \text{Var}\left(\frac{1}{n}\sum_{k=1}^{n} h(\theta_k) \,\middle|\, s^*\right) = \frac{1}{n}\text{Var}\left(h(\theta) \,|\, s^*\right) = \frac{v(s^*)}{n},$$

```
Estimate with sampling from true posterior

Input is data $x^* \in \mathbb{R}^d$, summary function $s\colon \mathbb{R}^d \to \mathbb{R}^q$, prior parameter density $f_\theta$ for $\theta \in \mathbb{R}^p$, conditional statistic density $f_{S|\theta}$, tolerance parameter $\delta > 0$, and a required number $n$ of accepted proposals.

  1. Set $k = 1$.
  2. Generate parameter proposal $\theta_k \sim f_\theta$, and data sample $X_k \mid \theta_k \sim f_{X|\theta}(\cdot \mid \theta_k)$.
  3. $s^* = S(x^*)$, $s_k = s(X_k)$, accept $\theta_k$ if $s_k = s^*$.
  4. If less than $n$ proposals have been accepted, increase $k$ by one, and return to Step 2.

Output is estimate $Z = \frac{1}{n}\sum_{j=1}^{n} h(\theta_{k_j})$, the mean of the function values for the accepted proposals.
```

Algorithm 2.1.1: Example algorithm for exact sampling from the true posterior distribution. This assumes that, instead of sampling directly from the distribution for $\theta \mid S$, we can only generate samples from the distributions for θ and $S \mid \theta$.

where

$$v(s) := \operatorname{Var}\left(h(\theta) \mid s\right) \tag{2.2}$$

is the posterior variance. The MSE is therefore inversely proportional to the number of samples used.

One possible algorithm for this estimate is given in Algorithm 2.1.1: we generate a proposed value for θ, and then use this to generate a value for s. If this is equal to s^*, we accept the proposal.

While this samples exactly from the true posterior, the algorithm will take a long time to accept a proposal, even in small discrete sample spaces for S. In our problem, acceptance will almost never occur. To deal with this, we can relax the acceptance condition, so that s need only be within a small distance from s^*. This is the most basic version of ABC.

It is common practice to measure the distance between a statistic sample s

and the observed statistic s^* using the Euclidean norm. Specifically, we choose a tolerance parameter δ, then accept a sample if

$$\|s - s^*\| \leq \delta.$$

Alternatively, we can write the acceptance condition with respect to the ball around s^*, as defined below.

Definition 2.4. *The* ball of radius δ with centre s^* *is equal to the set*

$$B_\delta(s^*) := \{s : \|s - s^*\| \leq \delta\},$$

and has volume $|B_\delta|$.

Therefore, the acceptance condition $\|s - s^*\| \leq \delta$ can also be written as

$$s \in B_\delta(s^*).$$

In addition to the tolerance δ, we must also decide on a stopping condition for the algorithm. One simple choice is to stop after accepting n samples. This prevents cases where our estimate is based on a very low number of samples. In particular, it prevents the case where no samples are accepted. However, the variance of the computational cost can be large: in the case where the computational cost of each proposal is fixed, the cost follows a negative binomial distribution, usually with a small success probability. This can make the running time of the algorithm unreliable. We refer to the estimate with this stopping condition as the ABCACC estimate Y_n, whose algorithm is given in Algorithm 2.1.2.

Alternatively, we can decide to stop after generating N samples. The variance of the computational cost then depends only on the variance of the cost of generating a single sample, so this stopping condition makes the algorithm's running time more predictable. However, the number of accepted samples is binomially distributed, and the number of accepted samples may be small. In particular, it is possible to accept no samples, so, to use this stopping condition, we must also choose a default value for the estimate if no samples are accepted. We refer to the estimate with this stopping condition as the ABCBAS estimate Z_N, whose algorithm is given in Algorithm 2.1.3.

ABCACC estimate Y_n

Input is data $x^* \in \mathbb{R}^d$, summary function $s\colon \mathbb{R}^d \to \mathbb{R}^q$, prior parameter density f_θ for $\theta \in \mathbb{R}^p$, conditional statistic density $f_{S|\theta}$, tolerance parameter $\delta > 0$, and a required number n of accepted proposals.

1. Set $k = 1$.
2. Generate parameter proposal $\theta_k \sim f_\theta$ and data sample $X_k \,|\, \theta_k \sim f_{X|\theta}(\cdot \,|\, \theta_k)$.
3. $s^* = S(x^*)$, $s_k = s(X_k)$, accept θ_k if $\|s_k - s^*\| \leq \delta$.
4. If less than n proposals have been accepted, increase k by one, and return to Step 2.

Output is estimate $Y_n = \frac{1}{n}\sum_{j=1}^{n} h(\theta_{k_j})$, the mean of the function values for the accepted proposals.

Algorithm 2.1.2: Algorithm for basic ABC estimate with a fixed number n of accepted samples.

```
ABCBAS estimate $Z_N$

Input is data $x^* \in \mathbb{R}^d$, summary function $s\colon \mathbb{R}^d \to \mathbb{R}^q$, prior parameter
density $f_\theta$ for $\theta \in \mathbb{R}^p$, conditional statistic density $f_{S|\theta}$, tolerance parameter
$\delta > 0$, default estimate $c$, and a required number $N$ of proposals.

    1. Set $k = 1$.
    2. Generate parameter proposal $\theta_k \sim f_\theta$ and data sample
       $X_k \,|\, \theta_k \sim f_{X|\theta}(\cdot \,|\, \theta_k)$.
    3. $s^* = S(x^*)$, $s_k = s(X_k)$, accept $\theta_k$ if $\|s_k - s^*\| \leq \delta$.
    4. If $k < N$, increase $k$ by one, and return to Step 2.

If $n > 0$, output is estimate $Z_N = \frac{1}{n}\sum_{j=1}^{n} h(\theta_{k_j})$, the mean of the function
values for the $n$ accepted proposals. If $n = 0$, output is estimate $Z_N = c$.
```

Algorithm 2.1.3: Algorithm for basic ABC estimate with a fixed number N of samples, both accepted and rejected.

For both of these ABC estimates, the accepted proposals are no longer sampled from the true posterior distribution, but are instead drawn from the distribution defined below.

Definition 2.5. *The* ABC posterior distribution *for tolerance parameter δ is the conditional distribution of $\theta \,|\, S \in B_\delta(s^*)$.*

For small values of the tolerance δ, we can expect this distribution to be similar to the true posterior. However, this also results in a small probability of samples being accepted. As δ increases, the two posterior distributions will become less alike, but the acceptance probability will increase. In particular, as δ tends to infinity, the ABC posterior distribution will tend towards the prior distribution, since all proposals will be accepted. Thus, choosing a value for δ involves a balance between the acceptance probability and the similarity of the two posterior distributions.

In terms of the MSE, the bias is equal to the difference between the posterior expectations,

$$\mathbb{E}\left(h(\theta)\,|\,S \in B_\delta(s^*)\right) - m(s^*),$$

which will, in general, increase with δ. The variance is equal to the ABC posterior variance

$$\frac{1}{n}\mathrm{Var}\left(h(\theta)\,|\,S \in B_\delta(s^*)\right).$$

For the ABCACC estimate, where n is fixed, this will also generally increase with δ, as the information in the acceptance condition decreases. Therefore, the MSE will increase as δ increases, but the expected computational cost will decrease, and vice versa. The behaviour of the variance of the ABCBAS estimate is more complicated: since the expectation of n will increase with δ, the variance may change in either direction.

Example 2.6. *To illustrate the challenge of choosing a tolerance, we look at the simple case where the parameter θ has a* $\mathrm{N}(0,1)$ *prior distribution, h is the identity function, and the data observation is a vector of q independent and identically distributed* (IID) *variables $(X_k)_{k=1}^q$, where $X_k\,|\,\theta \sim$* $\mathrm{N}(\theta,1)$*. In this case, we can find the posterior for θ exactly, since, by Lemma A.1, this has density*

$$\begin{aligned} f_{\theta|X}(t\,|\,x) &\propto \exp\left(-\frac{1}{2}t^2\right)\exp\left(-\frac{1}{2}\sum_{i=1}^q (x_i - t)^2\right) \\ &\propto \exp\left(-\frac{q+1}{2}\left(t - \frac{q}{q+1}\bar{x}\right)^2\right), \end{aligned}$$

and so the posterior distribution is $\theta\,|\,X \sim$ $\mathrm{N}(\frac{q}{q+1}\bar{X}, \frac{1}{q+1})$*, where $\bar{X}$ is the data mean, and is therefore a minimal sufficient statistic for X.*

Now consider an ABC *estimate that uses the sufficient statistic $s(X) = \bar{X}$. For positive δ, we accept samples whose statistic samples are in the ball $B_\delta(s^*)$, and the* ABC *posterior has a density proportional to*

$$\begin{aligned} f_{\mathrm{ABC}}(t\,|\,s^*) &\propto \int_{B_\delta(s^*)} f_{S|\theta}(s\,|\,t)\,\mathrm{d}s\, f_\theta(t) \\ &\propto \left(\Phi\left(\sqrt{q}\left(s^* + \delta - t\right)\right) - \Phi\left(\sqrt{q}\left(s^* - \delta - t\right)\right)\right)\phi\left(t\right), \end{aligned}$$

where Φ is the standard normal distribution function. By Taylor expansion, *using Theorem A.2, this is equal to*

$$\begin{aligned}
f_{\mathrm{ABC}}(t \mid s^*) &\propto \left(\phi\left(\sqrt{q}(s^*-t)\right) + \frac{1}{6}\delta^2\phi''\left(\sqrt{q}(s^*-t)\right) + \mathcal{O}\left(\delta^4\right)\right)\phi(t) \\
&= \left(\phi\left(\sqrt{q}(s^*-t)\right) + \frac{q(s^*-t)^2-1}{6}\phi\left(\sqrt{q}(s^*-t)\right)\delta^2 + \mathcal{O}\left(\delta^4\right)\right)\phi(t) \\
&= \phi\left(\sqrt{q}(s^*-t)\right)\phi(t)\left(1+\mathcal{O}\left(\delta^2\right)\right) \\
&\propto \exp\left(-\frac{q+1}{2}\left(t-\frac{q}{q+1}s^*\right)^2\right)\left(1+\mathcal{O}\left(\delta^2\right)\right) \\
&\propto f_{\theta|S}(t \mid s^*)\left(1+\mathcal{O}\left(\delta^2\right)\right)
\end{aligned}$$

as $\delta \downarrow 0$, where $\phi(x)$ is the standard normal density at x, and $\mathcal{O}(\cdot)$ is defined in Definition A.4. The expected value of the estimate will therefore have a bias of order $\mathcal{O}\left(\delta^2\right)$. We show two examples of the resulting ABC *posterior distribution in Figures 2.1.1 and 2.1.2.*

On the other hand, the probability of accepting a proposal is equal to

$$\begin{aligned}
\mathbb{P}\left(S \in B_\delta(s^*)\right) &= \int \mathbb{P}\left(S \in B_\delta(s^*) \mid t\right) f_\theta(t)\,\mathrm{d}t \\
&= \int \left(\Phi\left(\sqrt{q}(s^*+\delta-t)\right) - \Phi\left(\sqrt{q}(s^*-\delta-t)\right)\right)\phi(t)\,\mathrm{d}t \\
&= \sqrt{q}\int \left(2\delta\phi(\sqrt{q}(s^*-t)) + \mathcal{O}\left(\delta^3\right)\right)\phi(t)\,\mathrm{d}t \\
&= \mathcal{O}(\delta),
\end{aligned}$$

as $\delta \downarrow 0$. Therefore, lowering the tolerance will lower the acceptance probability, which will increase the expected computational cost for the ABCACC *estimate Y_n, and will increase the variance for the* ABCBAS *estimate Z_N. Therefore, lowering the tolerance will decrease the bias, but increase either the computational cost or the variance, and vice versa. Choosing a tolerance value requires finding a balance between these two issues.*

For the rest of this chapter, we look at other variants of the ABC algorithm.

2.1.2 Generalisations of the Acceptance-Rejection Step

This section addresses what can be considered as a generalisation, rather than a variant. It covers two changes we can make to the basic algorithms:

1. Instead of a proposal whose statistic s is in a ball around s^*, we can accept in a differently-shaped region around s^*. For example, instead of

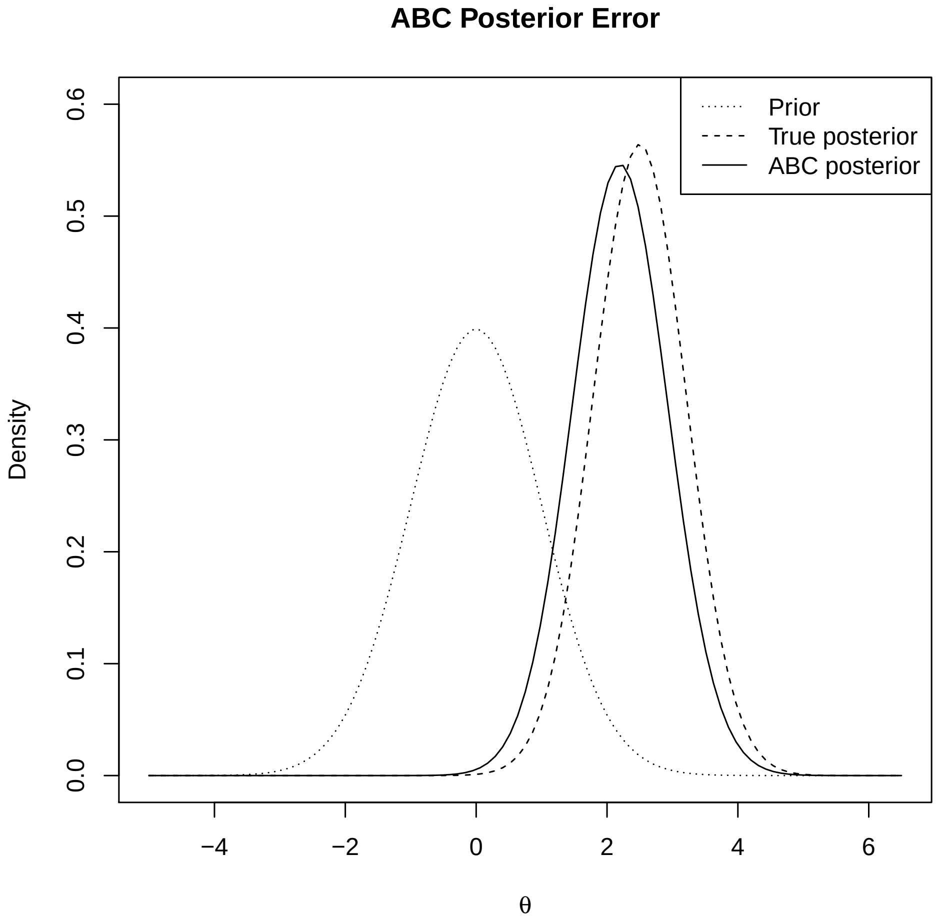

Figure 2.1.1: Plot of prior, true posterior, and ABC posterior densities, for $h(t) = t$, $s^* = 5$, and $\delta = 1$, in Example 2.6

accepting when $\|s - s^*\| = (s - s^*)^T(s - s^*) \leq \delta^2$, we can accept when

$$(s - s^*)^T H^{-1}(s - s^*) \leq 1,$$

for some positive definite, symmetric matrix H. The matrix H then describes an ellipsoidal acceptance region, and H has an equivalent rôle to δ^2. More generally, we can define a region such that some acceptance function is equal to one inside the region, and zero outside it.

2. Instead of the accept-reject scheme used so far, we can use random acceptance, where the distance between the statistic sample and s^* is used to determine an acceptance probability for the sample.

3. Instead of using the distance between the statistic sample and s^* to determine whether a sample is accepted, we can use the distance to weight

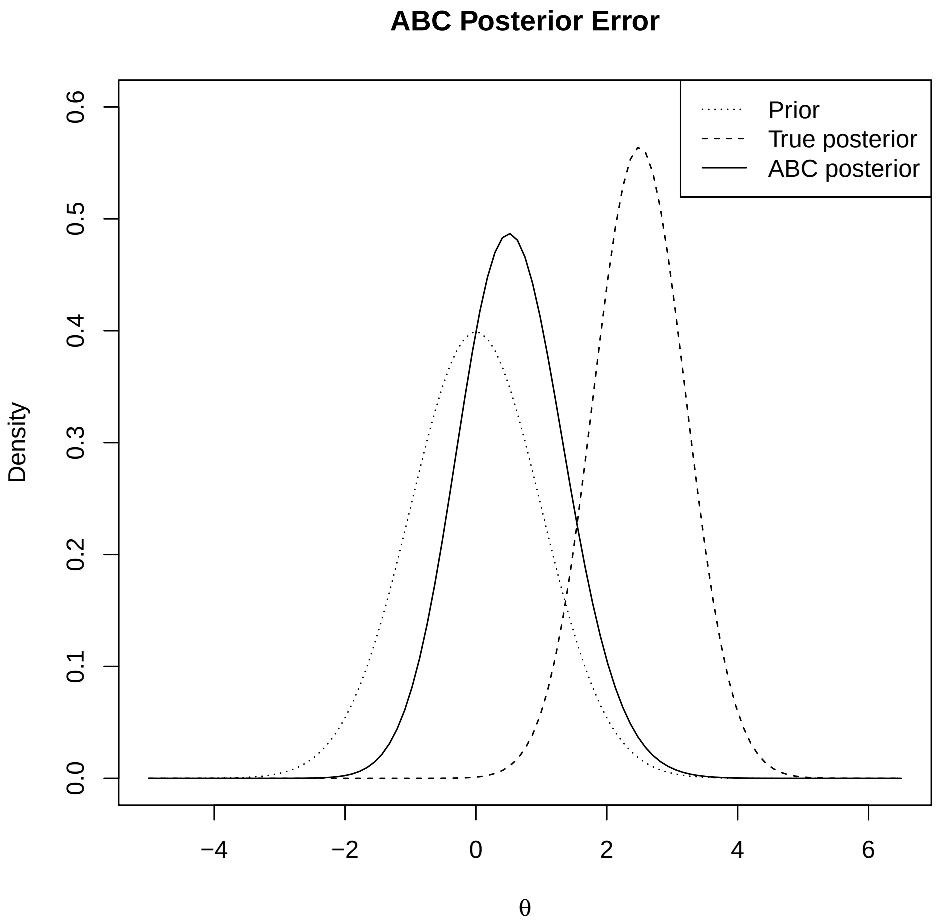

Figure 2.1.2: Plot of prior, posterior, and ABC posterior densities, for $h(t) = t$, $s^* = 5$, and $\delta = 5$, Example 2.6.

the samples in the estimate.

We can express cases 1 and 2 with the same notation.

Definition 2.7. *The function* $K\colon \mathbb{R}^q \to \mathbb{R}$ *is a* kernel function *if the following conditions hold:*

1. $K(\cdot)$ *is non-negative;*

2. K *is symmetric;*

3. $\int K(u)\,\mathrm{d}u = 1$.

The bandwidth matrix H *determines the* scaled kernel function K_H, *where*

$$K_H(u) := |H|^{-1/2} K(H^{-1/2}u).$$

The matrix H *is the* square-bandwidth matrix.

We then accept a proposal with probability $K_H(S - s^*)/\max_u K_H(u)$. For accept-reject algorithms, K is proportional to a discrete indicator function, so that the acceptance probability is either zero or one.

Notation 2.8. *We write indicator functions in Iverson bracket notation:*

$$[A] := \begin{cases} 1 & A \text{ true}, \\ 0 & A \text{ false}. \end{cases}$$

Example 2.9 (Acceptance on a ball)**.** *Let K be the kernel function with domain $\mathbb{R}^q$,*

$$K(u) = \frac{1}{|B_1|}[u \in B_1(0)].$$

Additionally, let $H = \delta^2 I$ for some δ. Then the scaled kernel function is equal to

$$K_H(u) = \frac{1}{|B_1|\,\delta^q}[\delta^{-1}u \in B_1(0)] = \frac{1}{|B_\delta(0)|}[u \in B_\delta(0).]$$

Samples are then accepted if the statistic sample is in $B_\delta(s^*)$, and rejected otherwise, giving the original acceptance condition. Therefore, the square-bandwidth matrix H can be considered as a generalisation of the square of the tolerance, δ^2.

If H is diagonal, and the diagonal elements are not all equal, then the acceptance region becomes an ellipsoid, whose principal axes correspond to elements of the statistic vector. If H is not diagonal, then the acceptance region is an ellipsoid, where the principal axes correspond to a different basis.

We will be interested in several properties of kernel functions, which we define now.

Definition 2.10. *The* second moment matrix *of the kernel function K is equal to*

$$\int_{\mathbb{R}^q} K(u)uu^T \,\mathrm{d}u.$$

The roughness *of the kernel function K is defined as*

$$R(K) := \int_{\mathbb{R}^q} K(u)^2 \,\mathrm{d}u.$$

Lemma 2.11. *Let the kernel function K have second moment matrix $M(K)$ and roughness $R(K)$. Then the scaled kernel function K_H has second moment matrix*

$$M(K_H) = H^{1/2}M(K)H^{1/2}.$$

In particular, if $M(K) = \mu_2(K)I$ for some scalar $\mu_2(K) > 0$, then K has second moment matrix $M(K_H) = \mu_2(K)H$. The roughness of the scaled kernel function is equal to

$$R(K_H) = |H|^{-1/2} R(K).$$

Later, we will be interested in kernel functions with second moment matrix of the form $\mu_2(K)I$ mentioned above. This includes kernel functions that are spherically symmetric, or products of one-dimensional symmetric kernel functions. Many commonly-used kernels are in this category, including the uniform kernel, as described in Example 2.9, the Gaussian kernel, equivalent to the simple normal density function, and the Epanechnikov kernel. The second moment matrix for such a kernel function has a convenient form when accounting for the introduction of the bandwidth, as shown below.

Proof. The scaled kernel function has second moment matrix

$$M(K_H) = \int_{\mathbb{R}^q} K_H(u)uu^T \,\mathrm{d}u = \int_{\mathbb{R}^q} |H|^{-1/2} K(H^{-1/2}u)uu^T \,\mathrm{d}u.$$

Changing variables to $v = H^{-1/2}u$,

$$M(K_H) = \int_{\mathbb{R}^q} K(v) \left(H^{1/2}v\right) \left(H^{1/2}v\right)^T \mathrm{d}v = H^{1/2} \int_{\mathbb{R}^q} K(v)vv^T \,\mathrm{d}v \, H^{1/2},$$

as required. The general scaled kernel function has roughness

$$R(K_H) = \int_{\mathbb{R}^q} K_H(u)^2 \,\mathrm{d}u = |H|^{-1} \int_{\mathbb{R}^q} K(H^{-1/2}u)^2 \,\mathrm{d}u.$$

By the same change of variables,

$$R(K_H) = |H|^{-1/2} \int_{\mathbb{R}^q} K(v)^2 \,\mathrm{d}v = |H|^{-1/2} R(K). \qquad \square$$

Example 2.12. *The uniform kernel in Example 2.9 has second moment matrix*

$$\frac{1}{|B_1|} \int_{B_1(0)} uu^T \,\mathrm{d}u.$$

Since $B_1(0)$ is spherically symmetric, non-diagonal elements will be equal to zero, and the diagonal elements are equal to

$$\mu_2(K) = \frac{1}{|B_1|} \int_{B_1(0)} u_1^2 \,\mathrm{d}u = \frac{1}{q\,|B_1|} \int_{B_1(0)} \|u\|^2 \,\mathrm{d}u.$$

The general uniform kernel function has roughness

$$R(K) = \int_{B_1(0)} 1/\,|B_1|^2 \,\mathrm{d}u = 1/\,|B_1|\,.$$

We further evaluate these in Section 3.2.1, when calculating the asymptotic bias for the ABCACC *estimate.*

Since we define kernel functions to be non-negative, $\mu_2(K)$, if it exists, is positive, and is zero only if K is a Dirac delta function. This is equivalent to only accepting if $s = s^*$, giving the true posterior sampling estimate described in Algorithm 2.1.1.

Kernel functions can be defined without the non-negativity condition. In this case, $M(K)$ and $\mu_2(K)$ can be zero for kernels other than the Dirac delta function. In kernel density estimation, such kernels are called *higher-order* kernels, or *bias-reducing* kernels. For example, the Silverman kernel contains a sine function that allows it to take negative values, and the resulting second moment matrix is equal to zero.

While it is not possible to use such kernels to give acceptance probabilities, they can be used in other cases. For example, in case 3, where we weight samples in the estimate, instead of either accepting or rejecting them, the weight "kernel" function J can take negative values, giving negative weights. This can give implausible estimates: for example, if the parameters θ can only be positive, the weights can result in a strictly-negative estimate. We briefly discuss why using such a kernel here might be useful in Section 3.4.3.

It is also possible to use such a kernel in kernel regression.However, we do not consider this possibility when using kernel regression in Chapter 4, because the asymptotic analysis indicates that it is not necessary for improving the rate of convergence.

2.1.3 Generalised Ball Acceptance Regions

We will be measuring distances on the statistic space $\mathbb{R}^q$ using the Euclidean norm $\|\cdot\|_2$, but there are other possible choices: for example, we can use the Manhattan norm $\|\cdot\|_1$, or the supremum norm $\|\cdot\|_\infty$. More generally, we can use generalised balls as our acceptance regions, where we can use a different norm for each dimension of the statistic space.

Definition 2.13. *Let l be the q-dimensional vector $l := (l_1, \ldots, l_q)$, and let δ be a similar vector with elements δ_k, where $l_k, \delta_k > 0$ for all k. Then we define the associated* generalised ball *at s^* to be*

$$B_\delta^{(l)}(s^*) := \left\{ s : \sum_{k=1}^{q} \left| \frac{s_k - s_k^*}{\delta_k} \right|^{l_k} \leq 1 \right\}.$$

In the case $l_k \uparrow \infty$, the relevant summand is equal to

$$\lim_{l_k \uparrow \infty} |u|^{l_k} = \begin{cases} 0 & |u| < 1, \\ 1 & |u| = 1, \\ \infty & |u| > 1. \end{cases}$$

Note that, if $0 < l_k < 1$ for some k, the resulting ball is not convex. For example, if $l = (2/3, 2/3)$, then the acceptance region is the astroid for the circle of radius δ. On the other hand, if all l_k tend to infinity, the ball will tend towards the q-dimensional hypercube, minus the vertices. Figures showing other examples are given in Wang [2005].

2.1.4 Kernel Density Estimates

There is often interest in some property of the true posterior distribution that can not be expressed as $\mathbb{E}\left(h(\theta) \mid s^*\right)$ for some one-dimensional function h. For example, there is no such expression for the posterior quantiles, and there is also no such expression for points on the posterior density function. A common alternative approach to estimate such properties is to use an estimate of the posterior density function $f_{\theta|S}$. This is done using a kernel density estimate: we choose a kernel function $\tilde{K}$, and a square-bandwidth matrix $\tilde{H}$. The estimate for $f_{\theta|S}(\theta_0 \mid s^*)$ is then

$$Z(\theta_0) = \frac{1}{n} \sum_{j=1}^{n} \tilde{K}_{\tilde{H}}\left(\theta_{k_j} - \theta_0\right).$$

For a fixed $\tilde{K}$ and $\tilde{H}$, this is the ABC estimate for

$$\mathbb{E}\left(\tilde{K}_{\tilde{H}}(\theta - \theta_0) \mid s^*\right).$$

This can be thought of either as the ABC estimate with parameter function $h(t) = \tilde{K}_{\tilde{H}}(t - \theta_0)$, or as the ABC estimate for an infinite-sample kernel density estimate. However, $\tilde{H}$ is chosen when setting up the algorithm, so can be dependent on the choice of n in the ABCACC estimate, or on the choice of N in the ABCBAS estimate.

2.1.5 Discrete Data

All of the above variants, and all of the results that follow, assume that the statistic space is continuous. However, we can consider the case where the statistics are in $\mathbb{N}^q$, or some subset, resulting in a discrete statistic space. We can then accept proposals in the same region as before. Alternatively, since the probability of generating a statistic equal to s^* can now be non-zero, we can use the exact posterior sampler in Algorithm 2.1.1, where $f_{S|\theta}$ is now a probability mass function.

2.2 Previous Results

While there were previous papers that had a similar approach to ABC, such as Diggle and Gratton [1984] and Rubin [1984], the paper that began interest in the topic is generally considered to be Tavaré et al. [1997], which introduced a rejection algorithm for inference on coalescence times in phylogenetics, as described in Section 2.3.7. Here, the algorithm generating parameter proposals θ also generates an intermediate data sample L. While the conditional data distribution $X \mid \theta$ is not known, the distribution $X \mid L$ is a simple Poisson distribution. Once θ and L have been generated, this intermediate likelihood can be used to give an acceptance probability. The algorithm, therefore, samples from the true posterior distribution, but has an intermediate data sampling step. This algorithm is given in Algorithm 2.3.1.

Pritchard et al. [1999] later expanded on this algorithm to do inference on human coalescence times, introducing the first example of ABC. This involved full data sampling, with a summary statistic s, and acceptance condition

$$\left\| \frac{s - s^*}{s^*} \right\|_\infty < \delta,$$

where the division on the left hand side is element-wise. Beaumont et al. [2002] established the name and definition of the ABC approach. Since then, it has spread into areas outside of population genetics. Reviews can be found in Beaumont [2010], Bertorelle et al. [2010], Csilléry et al. [2010], and Marin et al. [2012].

There are many papers introducing new variants, some of which we describe in this section. The variants with parameter adjustment in Section 2.2.3, and the results in Section 2.2.4, have special importance in this text, as they are referred to in later chapters.

2.2.1 ABC Within Other Methods

One way of viewing the ABC is as a likelihood estimator. For this reason, and because ABC is simple to understand and implement, ABC is often inserted into other Monte Carlo methods that otherwise require knowledge of the likelihood. The sampling is then split between the methods, with ABC generating statistic samples, and the other method generating parameter proposals.

A simple example is ABC within rejection sampling or importance sampling, where the parameters are sampled from a distribution different to the prior, and accepted proposals are weighted to account for the difference between the two distributions. This is useful if the prior distribution is not trivial to draw samples from. For an example of an algorithm for ABC with importance sampling, see Fearnhead and Prangle [2012].

The most common example is ABC within Markov Chain Monte Carlo (MCMC), where ABC is used to give an approximate acceptance probability. MCMC methods generate parameter proposals from a Markov chain, rather than independently from the prior distribution. This approach can not make as much use of parallel computing as simpler variants, and shares the usual weakness of MCMC methods, such as highly-autocorrelated proposals, and a tendency to get stuck in low-probability regions of the parameter space. However, MCMC methods have the advantage of more efficient sampling over high-dimensional parameter spaces: see, for example, MacKay [2002] and Voss [2013].

ABC within MCMC usually refers to ABC within the Metropolis-Hastings algorithm, first proposed by Marjoram et al. [2003]. Each time a new proposal is generated, an acceptance probability is calculated. Specifically, if the old proposal is equal to t, and the new proposal is equal to t', then the Metropolis-Hastings acceptance probability for sampling for the posterior distribution is equal to

$$\alpha(t' \,|\, t) := \min\left\{1, \frac{f_{\theta|S}(t' \,|\, s^*)q(t' \,|\, t)}{f_{\theta|S}(t \,|\, s^*)q(t \,|\, t')}\right\} = \min\left\{1, \frac{f_\theta(t')f_{S\,|\,\theta}(s^* \,|\, t')q(t' \,|\, t)}{f_\theta(t)f_{S|\theta}(s^* \,|\, t)q(t \,|\, t')}\right\},$$

where $q(t' \,|\, t)$ is the probability density for a Markov Chain currently at t to move to t'. If t' is rejected, then t is used again instead.

Since $\alpha(t' \,|\, t)$ contains two unknown likelihoods, ABC can be used to estimate the acceptance probability. Marjoram et al. [2003] use the approximation

$$\hat{\alpha}(t' \,|\, t) := \min\left\{1, \frac{f_\theta(t')q(t' \,|\, t)}{f_\theta(t)q(t \,|\, t')}\,[\|s - s^*\| \leq \delta]\right\}.$$

This is equivalent to using the Metropolis-Hastings algorithm to sample from the prior distribution, and requiring an accepted new proposal to also pass an accept-reject step on its statistic sample. For the case where $\delta = 0$, Marjoram et al. [2003] show that the resulting stationary distribution for proposals is the true posterior distribution.

Other examples of usage of ABC within MCMC include Bortot et al. [2007], Wegmann et al. [2009], and Meeds and Welling [2014], where the latter two also do proposal adjustment, described in Section 2.2.3. A simple example of ABC within MCMC is ABC within the Gibbs sampler, a type of MCMC method whose Markov chain traverses over one dimension of the parameter space at a time.

Another common variant is ABC within sequential Monte Carlo, proposed by [Sisson et al., 2007]. Here, the algorithm generates T populations of N samples, where the population distributions are intended to tend gradually toward the target distribution. In this case, the target distribution is the ABC posterior.

The first population is drawn from some initial density $q(\cdot)$, and accepted with an accept-reject step, using tolerance δ_1. These initial proposals $\theta_{k,1}$ are assigned weights $w_{k,1} = f_\theta(\theta_{k,1})/q(\theta_{k,1})$, as in importance sampling. For the remaining populations, a proposal $\theta_{k,t}$ for population t consists of drawing a sample $\theta^*_{k,t}$ from the previous population, with weights $w_{\cdot,t-1}$, and generating the new proposal $\theta_{k,t} \sim K_t(\theta_{k,t} \,|\, \theta^*_{k,t})$ for some transition kernel K_t. We then generate a statistic sample, and run an accept-reject step with tolerance δ_t. If accepted, the new proposal is assigned, according to the correction in Sisson et al. [2009] , the weight

$$w_{i,t} = f_\theta(\theta_{i,t}) / \sum_{j=1}^{N} w_{j,t-1} K_t\left(\theta_{k,t} \,|\, \theta_{j,t-1}\right).$$

There is another round of sampling within a population if the weights $w_{\cdot,t}$ are dominated by a small number of large weights. This is done to keep the effective sampling size of each population above a specified threshold.

One important feature of this variant is that there are now T tolerances δ_t to specify, with the condition that the tolerances decrease as the population number t increases. The final population is expected to be sampled from the ABC posterior distribution for tolerance δ_T.

ABC within SMC is also used by Del Moral et al. [2012], by Toni et al. [2009] on both parameter inference and model selection, and by Jasra et al. [2012] on hidden Markov models.

2.2.2 Approximate Data and Statistic Samples

Most of the computational cost of an ABC estimate, and therefore most of its inefficiency, is due to the need for a large number of model simulations, often from highly complex models. This has led to several proposals to replace the original model with an approximate model that is faster to simulate. This is commonly done by forming an approximate model from a pilot sample, and using it to generate all the samples. While this introduces bias, the high cost of the original model often means that the variance is a larger part of the error, so introducing bias to reduce the variance is an acceptable trade-off.

A common approach is to approximate the likelihood as being normally distributed, and is comparable to the indirect inference approach proposed in Wood [2010], where $f_{S|\theta}$ is modelled as a normal distribution $\mathrm{N}\left(\mu_\theta, \Sigma_\theta\right)$, and μ_θ and Σ_θ are determined from statistic samples for θ using the method of moments.

Variants that use this approach include Fan et al. [2013], where $f_{S|\theta}$ is approximated as a mixture of normal distributions, and Wilkinson [2014], where the log-likelihood is approximated as a normal distribution.

A more complex example of such a variant is GPS-ABC, proposed by Meeds and Welling [2014], where generated samples are stored in a training set, and used to estimate the acceptance probability for future proposals. For each new proposal, the algorithm makes M estimates of the acceptance probability $\alpha(t' \mid t)$, using M pairs of estimates for $f_{S|\theta}(s^* \mid t)$ and $f_{S|\theta}(s^* \mid t')$. The two densities $f_{S|\theta}(\cdot \mid t)$ and $f_{S|\theta}(\cdot \mid t')$ are estimated as products of q independent normal distributions, where the distributions for element k of the statistic have means

$$\mu_{k,t'} \sim \mathrm{N}\left(\bar{\mu}_{k,t'}, \sigma^2_{k,t'}\right) \text{ and } \mu_{k,t} \sim \mathrm{N}\left(\bar{\mu}_{k,t}, \sigma^2_{k,t}\right),$$

and variance $\sigma_k^2 + \delta^2$. The likelihood estimate parameters $\bar{\mu}_{k,\cdot}$, σ_k^2, and $\sigma^2_{k,\cdot}$ are estimated from all of the samples generated so far, assuming that each dimension k is a Gaussian process.

Once the M estimates of $\alpha(t' \mid t)$ are obtained, the algorithm looks for the acceptance probability τ that would minimise the probability of making the wrong accept-reject choice, in comparison to the unknown exact value of α. This is equal to the empirical median of α. Additionally, an estimate is made for the resulting probability $\mathcal{E}$ of a wrong accept-reject decision. If this is higher than some fixed threshold ξ, then another sample is generated for some informative value of θ, and added to the training set. The likelihood estimate parameters are re-calculated, and a new set of M estimates of $\alpha(t' \mid t)$ are generated. This continues until $\mathcal{E} < \xi$. The new proposal is then accepted with probability τ.

While this algorithm makes many assumptions on the statistic distribution, it has the advantage samples are only generated when needed, and that less new samples need to be generated over time, so the computational cost of obtaining new accepted proposals decreases. This behaviour is clearly seen in the numerical experiments in Meeds and Welling [2014], where the MSE of several ABC estimates is compared against both the number of proposals and the number of samples generated. The GPS-ABC algorithm is comparable to other ABC estimates with respect to error against number of proposals. However, when the error is compared against number of samples generated, GPS-ABC reaches a point at no more samples are required, and the error rapidly reaches its minimum.

A non-parametric approach is taken by Buzbas and Rosenberg [2015], in a variant called Approximate Approximate Bayesian Computation, that has similarities to the Bayesian bootstrap proposed by Rubin [1981]. In the case where the data consists of J IID observations $x_1^*, \ldots, x_J^*$, AABC begins by generating a test set of M proposals $(\hat{t}_1, \ldots, \hat{t}_M)$, and, for each test proposal $\hat{t}_m$, a test set of data $(\hat{x}_{m,1}, \ldots, \hat{x}_{m,J})$. This test set is used to generate data samples, rather than using the original model.

When a new proposal t is generated, the distances $\left\|t-\hat{t}_m\right\|$ are calculated for all m, and are used to assign weights to the test proposals. Specifically, the k^{th} nearest neighbour, $\hat{t}_{m_k}$, is assigned the weight

$$w(m_k) = K_H\left(\left\|t-\hat{t}_{m_k}\right\|\right)[k \le K],$$

where K_H is the Epanechnikov kernel with bandwidth $K = \left\|t-\hat{t}_{m_{K+1}}\right\|$. Each of the K nearest neighbours is then assigned a drawing probability for the bootstrap step, where the probabilities are sampled from the Dirichlet distribution, $p \sim \text{Dir}\left(w(m_1),\ldots,w(m_K)\right)$. Next, each of the M data samples for t is then drawn independently from the data samples $\hat{x}_{m,j}$, in two steps. First, the test set m to draw from is chosen from the K nearest neighbours, with probabilities p. Second, the data sample is drawn uniformly from the J data samples in test set m. Once the M data samples for t_k, these are summarised with a statistic s_k, which is used for accept-reject as usual.

AABC is shown by Buzbas and Rosenberg [2015] to give similar results to both ABCBAS and MCMC with respect to number of proposals. As with GPS-ABC, however, it is shown to be more efficient with respect to number of samples generated from the original model.

2.2.3 Other Variants

Proposal Adjustments and Density Estimation

If a sample has statistic s_k, the parameter proposal θ_k has distribution $\theta \,|\, s_k$, rather than $\theta \,|\, s^*$, and expectation $m(s_k)$, rather than $m(s^*)$. Beaumont et al. [2002] proposed adjusting the proposal values before using them in the estimate, attempting to account for this discrepancy. In particular, they assumed the linear model

$$\theta_k = \hat{\alpha} + \hat{\beta}^T(s_k - s^*) + \epsilon_k, \quad \mathbb{E}\left(\epsilon_k\right) = 0, \quad \text{Var}\left(\epsilon_k\right) = \sigma^2.$$

where the residual ϵ_k is independent of s_k. For one-dimensional parameters, $\hat{\alpha}$ and $\hat{\beta}$ are chosen to minimise

$$\sum_{k=1}^{N}\left(\theta_k - \hat{\alpha} - \hat{\beta}^T(s_k - s^*)\right)^2 \hat{K}_{\hat{H}}(s_k - s^*),$$

for some kernel function $\hat{K}$, and some square-bandwidth matrix $\hat{H}$. Then the estimated conditional expectation function is $\hat{m}(s_k) := \hat{\alpha} + \hat{\beta}^T(s_k - s^*)$. Rather than use the regression point estimate $\hat{m}(s^*) = \hat{\alpha}$, the adjusted proposals

$$\hat{\theta}_k = \theta_k - \hat{m}(s_k) + \hat{m}(s^*) = \hat{m}(s^*) + \epsilon_k$$

are used to form the estimate

$$Z = \frac{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)\hat{\theta}_k}{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)} = \hat{m}(s^*) + \frac{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)\epsilon_k}{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)}.$$

Alternatively, the adjusted proposals can be used to estimate the posterior density, using kernel density estimation. In this case, the estimate density at θ_0 is equal to

$$Z(\theta_0) = \frac{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)\tilde{K}_{\tilde{H}}(\hat{\theta} - \theta_0)}{\sum_{k=1}^N \hat{K}_{\hat{H}}(s_k - s^*)},$$

for some kernel function $\tilde{K}$, and some square-bandwidth matrix $\tilde{H}$. In both cases, using $\hat{H} = \delta^2 I$ and $\hat{K}(u) = \frac{1}{2}[|u| \leq 1]$ gives an accept-reject method, as in the ABCBAS estimate.

Blum [2010] proposes using quadratic adjustment, which uses quadratic polynomial regression, rather than simple linear regression, to estimate m.

Biau et al. [2015] propose a similar estimate to Blum [2010], that uses an accept-reject method instead of weights. Rather than using a square-tolerance $\hat{H}$ to define an acceptance region, both N and n are fixed, so the algorithm accepts the n nearest neighbours to s^*.

Blum and François [2010] propose adjustments using non-linear regression, via neural networks, that uses the model

$$\theta_k = \hat{m}(s_k) + \hat{\sigma}(s_k)\epsilon_k,$$

where $\mathrm{Var}(\epsilon_k) = 1$, and $\hat{\sigma}(s_k)$ is the estimate of $v(s_k)$. The adjusted proposals are then equal to

$$\hat{\theta}_k = \hat{m}(s^*) + \frac{\hat{\sigma}(s^*)}{\hat{\sigma}(s_k)}\epsilon_k.$$

The authors claim increased computational efficiency compared to the linear adjustment variant, as well as a decreased sensitivity to the value of $\hat{H}$.

Proposal adjustment is also used by Wegmann et al. [2009] on proposals from ABC within MCMC, described in Section 2.2.1. A variation is used by Leuenberger and Wegmann [2010], where a general linear model is used on the accepted proposals and the observed statistic to estimate the true likelihood.

Early Stopping of Simulations

Another approach for reducing the computational cost of model simulations is to stop the generation of a set of data that is unlikely to be accepted. Lazy ABC proposed by Prangle [2016] and summarised in Prangle [2015], adds an intermediate step to ABC with importance sampling. For some intermediate data r, the sample generation is stopped, and the proposal t rejected, with probability $1 - \alpha(t, r)$. If the sample is fully generated, and the proposal is accepted, then the weight is adjusted by the factor $1/\alpha(t, r)$, to target the same distribution. This increases the variance, but reduces the expected computational cost.

Theoretical results are given for the asymptotically optimal choice of α, expressed in terms of optimising the effective sample size of the estimate, but it is noted that a heuristic choice may result in significantly less computational time spent on unpromising samples. It is also possible to have multiple stopping decision points during data generation.

Summary Selection

A large amount of ABC research has focused on good choice of summary statistics. Since this is highly problem-specific, there has been some research into taking a starting set of statistics, chosen by hand, and choosing an efficient transformation to a lower-dimension statistic set. This includes Fearnhead and Prangle [2012], where the new statistic is an estimate of $m(s)$.

More recently, there has been research into choosing summary statistics, or accepting proposals, in an fully non-parametric approach, removing the need for the initial statistics used in Fearnhead and Prangle [2012]. Some of this research brings ideas from the field of sufficient dimension reduction SDR: this includes Park et al. [2015], Mitrovic et al. [2016], and Zhong and Ghosh [2016]. More specifically, Zhong and Ghosh note the similarity in goals. In ABC summary selection, the aim is to find a summary function s such that $\theta \perp X \mid s(X)$. By comparison, SDR, a topic usually associated more with classical statistics, aims to find a transformation ϕ on the predictors X for a response Y such that $Y \perp X \mid \phi(X)$. It therefore seems reasonable to use SDR methods in ABC.

In Zhong and Ghosh [2016] the q-dimensional data is reduced by passing the samples into a support vector machine. The reduced data then consists of the first q principal components, where q is any value between 1 and d.

Other non-parametric approaches include Zhou and Fukumizu [2016].

2.2.4 Theoretical Results

Results regarding good practices when running ABC generally split into two types: theoretical asymptotic convergence results, for both the tolerance value and the resulting error, and methods to choose good summary statistics. We address the former below. For the latter, Blum et al. [2013] is a recent review of dimension reduction techniques in general. Also of note is Wilkinson [2013], where an ABC algorithm with proposal acceptance probability proportional to $K_H(s - s^*)$ is shown to be equivalent to exact inference in the case where the observations are subjected to noise with density function K_H. This is described as ABC giving exact inference for the wrong model.

Since choosing a good set of summary statistics and tolerance value for ABC is highly problem-specific, most theoretical results are asymptotic rates of convergence for the tolerance and the MSE. These are usually given as asymptotic rates as the number of samples N tends to infinity, since the latter is usually roughly proportional to the computational cost. For the ABCACC and ABCBAS estimates, Barber et al. [2015] show the optimal rate of convergence for the MSE to be of order $\mathcal{O}\left(N^{-\frac{4}{q+4}}\right)$, as N tends to infinity, as defined in Definition A.4. For comparison, the MSE for unbiased Monte Carlo methods is of order $\mathcal{O}\left(N^{-1}\right)$. A similar result is given in [Prangle, 2011, Appendix A.4], for the case where the summary statistic is equal to the true posterior $m(x)$:

$$s = S(x) := m(x).$$

Blum [2010] examines three variants of ABC for kernel density estimation on a one-dimensional parameter space, with differing methods for parameter adjustment, as described in Section 2.2.3. In all three variants, the kernel density estimation requires a bandwidth parameter, in addition to the ABC tolerance. This causes a decrease in the rate of convergence: all three are shown to have an pointwise MSE convergence rate of order $\mathcal{O}\left(N^{-\frac{4}{q+5}}\right)$. The three proportionality constants have no fixed order, so whether the variants with adjustment decrease the error is problem-specific, and depends on the non-linearity of the statistic density function at s^*.

Biau et al. [2015] consider the convergence rate for the n-nearest-neighbours ABC density estimate for general parameter dimension p, as is described in Section 2.2.3, and show it to be equal to

$$\text{MSE}(Z) = \begin{cases} \mathcal{O}\left(N^{-\frac{4}{p+8}}\right) & q < 4, \\ \mathcal{O}\left(N^{-\frac{4}{p+8}} \log(N)\right) & q = 4, \\ \mathcal{O}\left(N^{-\frac{4}{p+q+4}}\right) & q > 4. \end{cases}$$

This is the same rate as that in Blum [2010] for $q > 4$, and worse otherwise.

Fearnhead and Prangle [2012] propose a variant called Noisy ABC, where noise is added to the original observations before running the algorithm. This ensures the estimate converges to the correct answer as the observational information tends to infinity, for a fixed tolerance, including estimates for properties of the posterior not expressible as $\mathbb{E}\left(h(\theta) \,|\, s^*\right)$ for some function h, such as the posterior variance. However, the noise reduces the convergence rate to order $\mathcal{O}\left(N^{-\frac{2}{q+2}}\right)$. This assumes the statistic form $s(x) = \mathbb{E}\left(\theta \,|\, x\right)$, as in Prangle [2011], and is given as the motivation for using some conditional expectation estimate $\hat{m}$ as the summary function. They suggest obtaining such an estimate by doing quartic polynomial regression on a test set of summary statistics.

More recently, there have been results on the asymptotic behaviour of the ABC posterior distribution, as $q \uparrow \infty$, rather than a point estimate. These include Dean and Singh [2011], Frazier et al. [2015], and Li and Fearnhead [2016]. These are conditioned on the statistic distribution being $\mathcal{O}\left(f(q)\right)$ for some function f. Specifically, the limit of $f(q)\left(S - s^* \,|\, \theta\right)$, as $q \uparrow \infty$, is normally distributed, with zero mean and a covariance matrix dependent on θ.

2.3 Applications

We now give some example problems to which ABC can be applied. We begin with some simple theoretical examples, useful for comparing ABC estimates to the known exact answer. We then progress to more complicated problems, where ABC has been used in practice. We also give some examples of summary statistics used for these problems by other authors.

2.3.1 Conjugate Normal Inference

Since any problem with a normal prior distribution and a normal conditional statistic density has a normal posterior distribution, a simple inference problem to test ABC with is where $\theta \sim \mathrm{N}(0,1)$ and q IID data elements $X_i \sim \mathrm{N}(\theta,1)$. This problem is described in Example 2.6.

More complicated is the case where we wish to estimate the variance θ of a normal distribution, where the mean is known to be zero, based on q samples. In this case, we can take a conjugate prior, so that we have densities

$$f_\theta(t) \propto t^{-1}, \quad f_{X\,|\,\theta}((x_1,\dots,x_q),t)) \propto t^{-q/2} \exp\left(-\frac{\sum_{k=1}^q x_k}{2} t^{-1}\right),$$

resulting in sufficient statistic $s^* = \sum_{k=1}^q x_k^2$, and posterior density

$$f_{\theta|S}(t\,|\,s^*) \propto t^{-q/2-1} \exp\left(-\frac{s^*}{2} t^{-1}\right).$$

Therefore, the true posterior distribution is scaled inverse chi-squared [Lee, 2012],

$$\theta\,|\,s^* \sim s^* \chi_q^{-2}.$$

Fearnhead and Prangle [2012] remark that, if we use a normal acceptance kernel with variance $\delta < \sigma^2$, the basic ABC estimates tend to a point estimate at $\sigma^2 - \delta$ as the q tends to infinity.

We can further consider the case where the mean μ is unknown, with normal prior distribution $\mu \sim \mathrm{N}(0,\sigma^2)$, and the variance has prior $\sigma^2 \sim \chi_1^2$. One sufficient summary statistic is the empirical mean and variance of the sample [Blum, 2010]. An example is the Iris dataset, of petal lengths for the virginica species, with statistic elements $\bar{x} = 5.552$ and $s^2 = 0.304$.

2.3.2 g-and-k Distribution

The g-and-k distribution can be used to accurately approximate many common distributions [Haynes et al., 1997]. It has no closed-form density, and is instead defined by its inverse distribution,

$$F^{-1}(x\,|\,A,B,c,g,k) = A + B\left(1 + c\frac{1-e^{-gz(x)}}{1+e^{-gz(x)}}\right)(1+z(x)^2)^k z(x),$$

where $z(x) := \Phi^{-1}(x)$ is the x^{th} quantile of a standard normal distribution. A and $B > 0$ are location and scale parameters. The parameters g and $k > -1/2$ are related to skewness and kurtosis; setting them to zero results in a normal distribution. The parameter c is usually set to 0.8. It is possible to calculate likelihoods numerically, but this is computationally expensive. However, since we have the inverse distribution, drawing from it is straightforward, so we can use ABC to do inference on the distribution parameters A, B, g and k. This example is used in Fearnhead and Prangle [2012].

2.3.3 Ricker Model

Here, we consider the ecological model where the population N_t changes over time according to the equation

$$N_{t+1} = N_t r e^{-N_t + \epsilon_t},$$

where $\epsilon_t \sim N(0, \sigma_e^2)$ are independent, and $N_0 = 1$. The parameter is equal to $\theta = (\log r, \sigma_e, \phi)$, and the data consists of Poisson observations $x_t \sim \text{Po}(\phi N_t)$ at time-points 50 to 100 [Wood, 2010]. This example is used in Fearnhead and Prangle [2012].

2.3.4 M/G/1 Queue

Here we consider a single queue that is initially empty, where the times between two people joining the queue are exponentially distributed with rate θ_3. The service time for each person is uniformly distributed in the interval $[\theta_1, \theta_2]$, and the observation is the vector of times between people leaving. This example is used in Blum [2010] and Fearnhead and Prangle [2012].

2.3.5 Stochastic Kinetic Networks

Here we begin with a certain amount of two types of molecule, and consider how the amounts change over time as the molecules interact. The Lotka-Volterra model from Boys et al. [2008] contains two types of molecules, with counts y_1 and y_2. The possible events are birth of a type-1 molecule, death of a type-2 molecule, and interaction between one of each molecule, that turns the type-1 molecule into a type-2 molecule. These events occur as independent Poisson processes, with respective transition rates $\theta_1 y_1$, $\theta_2 y_2$, and $\theta_3 y_1 y_2$.

Cluster size	1	2	3	4	5	8	10	15	23	30
Frequency	282	20	13	4	2	1	1	1	1	1

Table 2.3.1: Observed cluster sizes in tuberculosis outbreak in San Francisco, 1991-1992.

This process can be described with a stochastic Petri net, as described by Baez and Fong [2014], and has a simple equilibrium distribution, and expected rate of change, by the Anderson–Craciun–Kurtz Theorem [Anderson et al., 2010]. However, we do not observe the general equilibrium distribution. Instead, we observe the number of molecules at given time points, and these observations are used to do inference on the transition rate constants θ_k. This is far more complicated, as the likelihood for the observed counts is intractable. Fearnhead and Prangle [2012] consider the case where the number of type-1 molecules is observed at discrete time points, and the number of type-2 molecules is only observed at the initial time point.

2.3.6 Tuberculosis Transmission Example

From Tanaka et al. [2006], this is an SIR model with mutation. The observation is the tuberculosis genotype present in 473 infected individuals, categorised into cluster sizes, as shown in Table 2.3.1. To model the spread of the disease, we begin with a single infected individual. In continuous time, we then have three possible events, distributed according to three independent Poisson processes, similarly to Section 2.3.5. Possible events are:

1. An infection spreading to a new individual, with the same genotype. This has rate constant α, called the *birth rate.* We assume a large, roughly-constant susceptible population, whose effect on the total birth rate is included in α, so the total birth rate for an infected population of size I is equal to αI.

2. An individual recovering from the disease. This has rate constant δ, called the *death rate.*

3. An infection in one individual mutating into a new genotype. We assume a genotype almost never has more than one origin. This has rate constant

μ, called the *mutation rate.*

All three rate constants are assumed to be the same for all genotypes. This process is continued until the total population is 10000. A sample of the genotype present in 473 infected individuals is taken. These are sorted into their respective clusters, and the data sample is the number of clusters of each size, as in the observation.

There are some improvements we can make to speed up the simulations. Since the time to process completion is not of importance, we are only interested in the order of events, so we can simulate the above process using a discrete-time Markov chain, with an event at each time step [Tanaka et al., 2006]. We can re-parametrise the rate constants to given event probabilities $a = \frac{\alpha}{\alpha+\delta+\theta}$, $d = \frac{\delta}{\alpha+\delta+\theta}$, and $m = 1 - a - d$. The likelihood then only depends on a and d.

Possible parameter properties of interest are the *net transmission rate* $\alpha - \delta$, the *doubling time* $\log(2)/(\alpha - \delta)$, and the *reproduction number* α/δ. The latter is the expected number of people an individual will infect before they recover, and is often used to measure the difficulty of controlling an epidemic.

Tanaka et al. [2006] choose the summary statistics to be the number of genotypes, and the gene diversity $1 - \sum_i (n_i/n)^2$, where n_i is the number of organisms in genotype i and $n = 473$ is the total number of organisms.

2.3.7 Phylogenetic Tree Example

This is the problem considered in Tavaré et al. [1997]. We begin with n contemporary individuals from a large population of size N, from each of which we have sampled some set of characteristics, which they will only inherit from one parent. These characteristics are most commonly taken to be disjoint genetic sequences. For example, we can observe only the male population, and look at the male-specific part of the Y chromosome [Tavaré et al., 1997], or observe the female population, and look at mitochondrial DNA. Alternatively, we can observe some collection of phenotypes.

We consider the genealogical tree of these individuals. For simplicity, we assume that each generation will have the same size N : these generations are discrete, and do not overlap [Kingman, 1982]. We begin with the generation that includes the observed individuals, then recursively add parent generations.

For each new parent generation, individuals in the child generation are assumed to be equally likely to descend from each parent, independently of the parentage for the other children. Consequently, as we add parent generations, the n observed individuals will begin to share ancestors, and eventually there will be only one common ancestor. The parameter of interest is how long ago this occurred, which is known as the *time to most recent common ancestor* (TMRCA).

If we now begin at the common ancestor, and move forwards in time, the observed individuals' ancestors' characteristics can be subject to mutations, which will then be carried forward in time. The observation is the set of characteristics for each of the n observed individuals, and the summary statistic is the number of *segregating sites* where mutations have resulted in differences between the samples.

For simulations, we make the following assumptions.

1. There are infinitely many characteristics. Each site will, therefore, almost never be subject to more than one mutation. This makes the number of segregating sites sufficient for the TMRCA.

2. Each characteristic is equally likely to be the target of a mutation.

3. Characteristics do not overlap. In the case of DNA, this requires observed sequences to be distinct.

4. Characteristics are independent with respect to mutations.

5. The mutation rate is the same between all individuals.

6. The general population size N is assumed to tend to infinity. This model is known as Kingman's *n-coalescent.*

First, we decide on a prior. We measure time by number of generations, divided by N, which we refer to as the *scaled time.* The TMRCA is equal to $T_n = \sum_{k=2}^{n} W_k$, a sum of independent variables W_k, where W_k is the length of scaled time in which the observed individuals have k distinct ancestors. The distribution for W_2 is simple to calculate: for finite N, the number of generations in W_2 is negative binomial with success probability $1/N$. The probability density for W_2, in scaled time, is therefore equal to

$$\mathbb{P}\left(W_2 = k\right) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{Nk-1}.$$

Since we assume that N tends to infinity, the resulting probability density for W_2 is equal to

$$\begin{aligned} f_{W_2}(t) &= \lim_{N\uparrow\infty} \frac{\mathbb{P}\left(W_2 \le t + 1/N\right) - \mathbb{P}\left(W_2 \le t\right)}{1/N} \\ &= \lim_{N\uparrow\infty} N\mathbb{P}\left(W_2 = t + 1/N\right) \\ &= \lim_{N\uparrow\infty} \left(1 - \frac{1}{N}\right)^{Nt} \\ &= e^{-t}, \end{aligned}$$

so $W_2 \sim \text{Exp}(1)$. By more complicated reasoning, the distribution for W_k at the limit is $W_k \sim \text{Exp}\left(\binom{k}{2}\right)$. Therefore, T_n is the sum of exponential variables with rate parameters $\binom{k}{2}$ for $k \in \{2, \dots, n\}$.

For statistic generation, we can again use the intervals W_k. If we assume a mutation rate of μ per generation, then the number of segregating sites can be simulated as a Poisson process with rate $N\mu$ on each branch. Consequently, if we let $L_n = \sum_{k=1}^{n} kW_k$ be the total length of the tree, then the number of segregating sites follows a Poisson distribution with rate $N\mu L_n$.

From the results above, we can see that a simple rejection algorithm for inference on T_n, given s^* segregating sites, is to simulate W_k, calculate the TMRCA $T_n = \sum_{k=1}^{n} W_k$ and total tree length $L_n = \sum_{k=1}^{n} kW_k$, and accept T_n with probability equal to

$$u = \frac{(N\mu L_n)^{s^*}/s^*!}{\max_\lambda \lambda^{s^*}/s^*!} = (N\mu L_n/s^*)^{s^*}.$$

This is given as Algorithm 1 in Tavaré et al. [1997]. This is followed by explicit equations for the posterior expectation of T_n. The final algorithm, reproduced in Algorithm 2.3.1, has additional steps to account for two complications.

1. We usually do not know the values of N and μ, so we must generate sample values for them.

2. We are usually interested in the TMRCA of all N of the current generation, rather than the n individuals we observe. This requires us to track the remaining number of ancestors for both the observed individuals and the general population. Once the observed individuals have a common ancestor, we check for sample acceptance. If a sample is accepted, we then simulate the remaining time to the TMRCA of the general population.

Rejection-sampling Method for Infinite-Site Phylogenetic Trees

Input is statistic $s^* \in \mathbb{R}$ of number of segregating sites, prior density f_N for population size $N \in \mathbb{R}^+$, prior density f_μ for mutation rate $\mu \in \mathbb{R}^+$.

1. Generate proposals $N_k \sim f_N$ and $\mu_k \sim f_\mu$ for $k \in C \subset \mathbb{N}$.

2. For each proposal k, set $N = N_k, \mu = \mu_k, n' = n$. generate coalescence time $W_{N,k} \sim \text{Exp}\left(\binom{N}{2}\right)$. Reduce n' by one with probability $n'(n'-1)/N(N-1)$. Reduce N by one. Repeat until $n' = 1$. Calculate T_n and L_n.

3. Accept (N, μ, T_n) with probability $\frac{e^{-\theta L_n/2}(\theta L_n/2)^k/k!}{e^{-k}k^k/k!}$.

4. For accepted samples, if $N = 1$, return $T_N = T_n$. Else, generate coalescence times $W_j \sim \text{Exp}\left(j(j-1)/2\right)$ for $j = 2$ to N, and set $T_N = T_n + W_N + W_{N-1} + \ldots + W_2$.

5. Repeat until the required number of proposals or accepted proposals is reached.

Output is accepted proposals $\hat{N}_j = N_{k_j}$ and $\hat{\mu}_j = \mu_{k_j}$, and TMRCA T_N.

Algorithm 2.3.1: Rejection-sampling method from Tavaré et al. [1997], for the problem described in Section 2.3.7

For the prior distribution of N and μ, Tavaré et al. [1997] give several examples, where N and μ are independently distributed, and each has either a Gamma or a log-normal distribution.

Tavaré et al. [1997] also note that W_k are no longer independent in the case where the general population varies in size over time. This, in addition to complications with regard to modelling mutations, is the motivation for the ABC approach used in Pritchard et al. [1999].

Chapter 3

Convergence of Basic ABC

In this chapter, we look at the ABCACC estimate Y_n described in Algorithm 2.1.2, where the algorithm stops after accepting a fixed number n of samples. We analyse the effect of the tolerance δ, and the number n of accepted proposals, on the MSE and expected cost of the estimate. In particular, we look at the case where our expected computational cost is fixed, and we would like to minimise the MSE by our choice of δ and n. We later consider how variants affect this optimisation.

Finding the exact optimum values for δ and n is highly problem-specific. Instead, we look for optimal rates of change for δ and n as the cost increases, which leads to more general, asymptotic results. Firstly, we look for conditions under which the estimate converges to the true posterior expectation as $\delta \downarrow 0$ and $n \uparrow \infty$. Secondly, we find asymptotic expressions for the MSE and the expected computational cost. Finally, we use these asymptotic expressions to find the optimal asymptotic rates of convergence for δ, n, and the MSE.

3.1 Convergence Conditions

Before we look at the asymptotic convergence rate of Y_n, we would like to know whether the estimate converges to the correct value, the true posterior expectation $m(s^*) = \mathbb{E}\left(h(\theta) \,|\, s^*\right)$, as the expected computational cost tends to infinity.

First, we introduce the functions ϕ_h and $\phi_h^{(\delta)}$, which are useful for looking at convergence conditions and the asymptotic bias. Later, we look at their equivalents for other ABC variants, which is a convenient way to examine how choice of variant affects the asymptotic bias.

Definition 3.1. *Let $\phi_h(s) := \int h(t) f_{S,\theta}(s,t)\,\mathrm{d}t = m(s) f_S(s)$, where m is defined in Equation 2.1, and*

$$\phi_h^{(\delta)}(s^*) := \frac{1}{|B_\delta(s^*)|} \int_{B_\delta(s^*)} \phi_h(s)\,\mathrm{d}s = \mathbb{E}\left(h(\theta)[S \in B_\delta(s^*)]\right),$$

where the ball B is defined in Definition 2.4. In particular, ϕ_h and $\phi_h^{(\delta)}$ are such that, for the estimates Y_n and Z_N,

$$m(s^*) = \frac{\phi_h(s^*)}{\phi_1(s^*)}, \quad \mathbb{E}\left(h(\theta) \mid S \in B_\delta(s^*)\right) = \frac{\phi_h^{(\delta)}(s^*)}{\phi_1^{(\delta)}(s^*)},$$

where $\phi_1(s) = f_S(s)$.

Theorem 3.2. *Let the function $h\colon \mathbb{R}^p \to \mathbb{R}$ be such that $\mathbb{E}\left(|h(\theta)|\right) < \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^p$, the ABCACC estimate Y_n satisfies*

1. $\lim_{n\uparrow\infty} Y_n = \mathbb{E}\left(Y_n\right)$ *almost surely for all $\delta > 0$; and*
2. $\lim_{\delta\downarrow 0} \mathbb{E}\left(Y_n\right) = m(s^*)$ *for all $n \in \mathbb{N}$.*

Proof. Since $\mathbb{E}\left(|h(\theta)|\right) < \infty$, we have

$$\mathbb{E}\left(|Y_n|\right) \leq \mathbb{E}\left(|h(\theta)| \mid s^*\right) = \frac{\phi_{|h|}(s^*)}{\phi_1(s^*)} < \infty$$

whenever $\phi_1(s^*) = f_S(s^*) > 0$, and, by the law of large numbers, Y_n converges to $\mathbb{E}\left(Y_n\right)$ almost surely.

For the second statement, since $\phi_1 \in \mathcal{L}^1(\mathbb{R}^q)$, we can use the Lebesgue differentiation theorem (Theorem A.6) to conclude that $\phi_1^{(\delta)}(s^*) \to \phi_1(s^*)$ as $\delta \downarrow 0$ for almost all $s^* \in \mathbb{R}^q$. Similarly, since

$$\int_{\mathbb{R}^q} |\phi_h(s)|\,\mathrm{d}s \leq \int_{\mathbb{R}^p} |h(t)| \int_{\mathbb{R}^q} f_{S,\theta}(s,t)\,\mathrm{d}s\,\mathrm{d}t = \int_{\mathbb{R}^p} |h(t)|\, f_\theta(t)\,\mathrm{d}t < \infty,$$

and thus $\phi_h \in \mathcal{L}^1(\mathbb{R}^q)$, we have $\phi_h^{(\delta)}(s^*) \to \phi_h(s^*)$ as $\delta \downarrow 0$ for almost all $s^* \in \mathbb{R}^q$. Using Definition 3.1, we get

$$\lim_{\delta\downarrow 0} \mathbb{E}\left(Y_n\right) = \lim_{\delta\downarrow 0} \frac{\phi_h^{(\delta)}(s^*)}{\phi_1^{(\delta)}(s^*)} = \frac{\phi_h(s^*)}{\phi_1(s^*)} = m(s^*)$$

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

Note that if the support of f_θ for a one-dimensional parameter covers all of $\mathbb{R}$, then the above theorem might not guarantee convergence in the case where h is the identity function. For example, if θ has a prior Cauchy distribution, then $\mathbb{E}\left(|\theta|\right)$ is unbounded.

In addition to the conditional expectation $m(s)$ of $h(\theta)$, we will also be interested in the conditional variance $v(s) = \operatorname{Var}\left(h(\theta) \mid s\right)$, since it will appear frequently when considering the estimate's asymptotic variance.

Corollary 3.3. *Assume that $\mathbb{E}\left(h(\theta)^2\right) < \infty$. Then, for almost all s^*, the* ABCACC *estimate Y_n satisfies*

$$\lim_{\delta \downarrow 0} n \operatorname{Var}\left(Y_n\right) = v(s^*),$$

uniformly in n, where v is defined in Equation 2.2.

Proof. From the definition of the variance, we know that

$$\begin{aligned}\operatorname{Var}\left(Y_n\right) &= \frac{1}{n} \operatorname{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right) \\ &= \frac{1}{n}\left(\mathbb{E}\left(g(\theta) \mid S \in B_\delta(s^*)\right) - \mathbb{E}\left(h(\theta) \mid S \in B_\delta(s^*)\right)^2\right).\end{aligned}$$

where g is the function such that $g(\cdot) = h(\cdot)^2$. Applying Theorem 3.2 to the function g, we see that

$$\lim_{\delta \downarrow 0} \mathbb{E}\left(h(\theta)^2 \mid S \in B_\delta(s^*)\right) = \mathbb{E}\left(h(\theta)^2 \mid s^*\right).$$

Since $\mathbb{E}\left(h(\theta)^2\right) < \infty$ implies $\mathbb{E}\left(|h(\theta)|\right) < \infty$, h satisfies

$$\lim_{\delta \downarrow 0} \mathbb{E}\left(h(\theta) \mid S \in B_\delta(s^*)\right) = m(s^*),$$

and thus

$$\begin{aligned}\lim_{\delta \downarrow 0} n \operatorname{Var}\left(Y_n\right) &= \mathbb{E}\left(h(\theta)^2 \mid s^*\right) - m(s^*)^2 \\ &= v(s^*).\end{aligned}$$

This completes the proof. □

3.2 Asymptotic Rates of Convergence

Now that we have a condition for the ABCACC estimate to correctly converge, we examine the rate at which it converges. In Chapter 2, we showed that the mean square error (MSE) of an estimate consists of a variance term and a square bias term. In this section, we will analyse the asymptotic behaviour of the bias and the variance separately, and then go on to look at the implications for the rate of convergence.

3.2.1 Asymptotic Bias

We first look for an expression for the asymptotic bias of Y_n.

Theorem 3.4. *Let the function $h\colon \mathbb{R}^p \to \mathbb{R}$ be bounded, $m(s)$ have continuous third-order derivatives with respect to s, and $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B^q_\delta(s)$ for all possible θ and s. Then, for almost all s^*, there is a constant c such that*

$$\operatorname{bias}(Y_n) = \mathbb{E}\left(h\left(\theta_{i_1}\right) \mid s^*\right) - m(s^*) = c(s^*)\delta^2 + \mathcal{O}\left(\delta^3\right)$$

as $\delta \downarrow 0$.

Proof. By Definition 3.1, the bias is equal to

$$\operatorname{bias}(Y_n) = \operatorname{bias}(Y_1) = \frac{\phi_h^{(\delta)}(s^*)}{\phi_1^{(\delta)}(s^*)} - \frac{\phi_h(s^*)}{\phi_1(s^*)}.$$

Expanding $\phi_h(s)$ around s^*, using Theorem A.2, we get

$$\phi_h(s) = \phi_h(s^*) + \nabla\phi_h(s^*)(s-s^*) + \frac{1}{2}(s-s^*)^T\mathcal{H}_{\phi_h}(s^*)(s-s^*) + \dots.$$

Substituting this expansion into $\phi_h^{(\delta)}(s^*)$, the first-order term vanishes, since its integral is equal to $\int_{B_\delta(s^*)}(s-s^*)\,\mathrm{d}s = 0$, and we are left with

$$\phi_h^{(\delta)}(s^*) = \frac{1}{\left|B^q_\delta\right|}\left(\phi_h(s^*)\left|B^q_\delta\right| + \frac{1}{2}\int_{B^q_\delta(s^*)}(s-s^*)^T\mathcal{H}_{\phi_h}(s^*)(s-s^*)\,\mathrm{d}s + \dots\right).$$

The second-order integrand is scalar, so we use the fact that the trace is invariant under cyclic permutations, we obtain

$$\begin{aligned}\int_{B^q_\delta(s^*)}(s-s^*)^T\mathcal{H}_{\phi_h}(s^*)(s-s^*)\,\mathrm{d}s &= \int_{B^q_1(0)} x^T\mathcal{H}_{\phi_h}(s^*)x\,\mathrm{d}x\,\delta^{q+2}\\ &= \int_{B^q_1(0)}\operatorname{Tr}\left(xx^T\mathcal{H}_{\phi_h}(s^*)\right)\,\mathrm{d}x\,\delta^{q+2}\\ &= \operatorname{Tr}\left(\int_{B^q_1(0)} xx^T\,\mathrm{d}x\,\mathcal{H}_{\phi_h}(s^*)\right)\delta^{q+2}.\end{aligned}$$

Since $B_1(0)$ is symmetric on each axis, $\int_{B_1(0)} xx^T \, \mathrm{d}x$ is diagonal. Furthermore, since $B_1(0)$ is spherically symmetric, all the diagonal elements are identical, so that

$$\int_{B_1(0)} xx^T \, \mathrm{d}x = \int_{B_1(0)} x_1^2 \, \mathrm{d}x I.$$

Therefore, the integrand is equal to

$$\begin{aligned}\int_{B_\delta^q(s^*)} (s-s^*)^T \mathcal{H}_{\phi_h}(s^*)(s-s^*) \, \mathrm{d}s &= \int_{B_1^q(0)} x_1^2 \, \mathrm{d}x \, \Delta\phi_h(s^*)\delta^{q+2} \\ &= \frac{\Delta\phi_h(s^*)}{q} \int_{B_1^q(0)} \|x\|^2 \, \mathrm{d}x \, \delta^{q+2}.\end{aligned}$$

The integral $\int_{B_1^q(0)} \|x\|^2 \, \mathrm{d}x$ is equal to $\int_0^1 r^2 \mathrm{SA}_{q-1}(r) \, \mathrm{d}r$, where $\mathrm{SA}_{q-1}(r)$ is the surface area of the q-dimensional ball with radius r. We now use the formulae from Huber [1982], for the volume and surface area of a ball:

$$|B_1^q| = \frac{\pi^{\frac{q}{2}}}{\Gamma(1+\frac{q}{2})}, \quad \mathrm{SA}_{q-1}(r) = \frac{2\pi^{\frac{q}{2}}}{\Gamma(\frac{q}{2})} r^{q-1}.$$

Therefore,

$$\frac{\mathrm{SA}_{q-1}(r)}{|B_1^q|} = qr^{q-1},$$

and

$$\begin{aligned}\phi_h^{(\delta)}(s^*) &= \phi_h(s^*) + \frac{\Delta\phi_h(s^*)}{2q\,|B_1^q|} \int_{B_1^q(0)} \|x\|^2 \, \mathrm{d}x \, \delta^2 + \mathcal{O}\left(\delta^3\right) \\ &= \phi_h(s^*) + \frac{\Delta\phi_h(s^*)}{2} \int_0^1 r^{q+1} \, \mathrm{d}r \, \delta^2 + \mathcal{O}\left(\delta^3\right) \\ &= \phi_h(s^*) + \frac{\Delta\phi_h(s^*)}{2(q+2)} \delta^2 + \mathcal{O}\left(\delta^3\right).\end{aligned}$$

Substituting this result into the bias, we find that

$$\begin{aligned}\mathrm{bias}(Y_n) &= \frac{\phi_h^{(\delta)}(s^*)}{\phi_1^{(\delta)}(s^*)} - \frac{\phi_h(s^*)}{\phi_1(s^*)} \\ &= \frac{\phi_h^{(\delta)}(s^*)\phi_1(s^*) - \phi_1^{(\delta)}(s^*)\phi_h(s^*)}{\phi_1^{(\delta)}(s^*)\phi_1(s^*)} \\ &= \frac{\Delta\phi_h(s^*)\phi_1(s^*) - \phi_h(s^*)\Delta\phi_1(s^*)}{2(q+2)\phi_1(s^*)^2}\delta^2 + \mathcal{O}\left(\delta^3\right) \\ &= \frac{\Delta\phi_h(s^*) - m(s^*)\Delta\phi_1(s^*)}{2(q+2)\phi_1(s^*)}\delta^2 + \mathcal{O}\left(\delta^3\right).\end{aligned}$$

We therefore have the desired result, with constant

$$c(s^*) = \frac{\Delta\phi_h(s^*) - m(s^*)\Delta\phi_1(s^*)}{2(q+2)\phi_1(s^*)} = \frac{\Delta m(s^*) + 2\nabla m(s^*)\nabla \log f_S(s^*)^T}{2(q+2)}. \quad (3.1)$$

□

If $m(s^*)$ has continuous fourth-order derivatives, then it is straightforward to show that

$$\text{bias}(Y_n) = c(s^*)\delta^2 + \mathcal{O}\left(\delta^4\right),$$

since all the odd-order terms in the Taylor expansion of ϕ_h will vanish, due to the symmetry of the acceptance region $B_\delta(s^*)$.

Example 3.5. *We continue the problem from Example 2.6, for $q = 1$, which has true posterior $\theta \mid s^* \sim \mathrm{N}\left(\frac{1}{2}s^*, \frac{1}{2}\right)$, and data samples $S \sim \mathrm{N}(0, 2)$. We look at the case where $h(\theta) = [\theta \in [-1/2, 1/2]]$, and would like to know the asymptotic bias for the* ABCACC *estimate of*

$$\begin{aligned} m(s^*) &= \mathbb{P}\left(\theta \in [-1/2, 1/2] \mid s^*\right) \\ &= \Phi\left(\sqrt{2}\left(\frac{1}{2} - \frac{s^*}{2}\right)\right) - \Phi\left(\sqrt{2}\left(-\frac{1}{2} - \frac{s^*}{2}\right)\right) \\ &= \Phi\left(\frac{1}{\sqrt{2}}\left(1 - s^*\right)\right) - \Phi\left(-\frac{1}{\sqrt{2}}\left(1 + s^*\right)\right). \end{aligned}$$

Taking derivatives, we find that

$$m'(s^*) = \frac{1}{\sqrt{2}}\phi\left(-\frac{1}{\sqrt{2}}\left(1 + s^*\right)\right) - \frac{1}{\sqrt{2}}\phi\left(\frac{1}{\sqrt{2}}\left(1 - s^*\right)\right),$$

$$\begin{aligned} m'(s^*)\log\left(f_S(s^*)\right)' &= -\frac{s^*}{2}m'(s^*) \\ &= -\frac{s^*}{2\sqrt{2}}\phi\left(-\frac{1}{\sqrt{2}}(1 + s^*)\right) + \frac{s^*}{2\sqrt{2}}\phi\left(\frac{1}{\sqrt{2}}(1 - s^*)\right), \end{aligned}$$

and

$$\begin{aligned} m''(s^*) &= -\frac{1}{2}\phi'\left(-\frac{1}{\sqrt{2}}\left(1 + s^*\right)\right) + \frac{1}{2}\phi'\left(\frac{1}{\sqrt{2}}\left(1 - s^*\right)\right) \\ &= -\frac{1}{2\sqrt{2}}\left(1 + s^*\right)\phi\left(-\frac{1}{\sqrt{2}}\left(1 + s^*\right)\right) - \frac{1}{2\sqrt{2}}\left(1 - s^*\right)\phi\left(\frac{1}{\sqrt{2}}\left(1 - s^*\right)\right). \end{aligned}$$

Substituting these into Equation 3.1, we find that the bias coefficient is equal to

$$c(s^*) = \frac{\left(-1 - 3s^*\right)\phi\left(\frac{1}{\sqrt{2}}\left(1 + s^*\right)\right) + \left(-1 + 3s^*\right)\phi\left(\frac{1}{\sqrt{2}}\left(1 - s^*\right)\right)}{12\sqrt{2}}.$$

The behaviour of this constant, with respect to the value of s^*, is not trivial, but it is easy to show that c is symmetric, and takes a negative value when $s^* = 0$. This is expected: since $m(s)$ is largest when $s = 0$, a non-zero tolerance will introduce ABC *samples with a reduced chance of being in the relevant interval.*

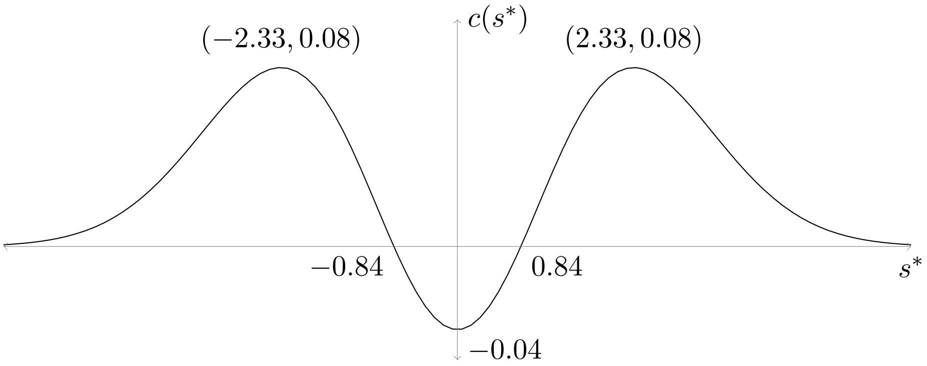

Figure 3.2.1: Plot of bias parameter $c(s^*)$ against s^* for Example 3.5.

Figure 3.2.1 shows some other points of interest. Notably, c becomes positive before it tends to zero, so there are two values of s^* where $c(s^*) = 0$. In this case, the bias is asymptotically proportional to a higher order of δ.

3.2.2 Asymptotic Variance

Theorem 3.6. *Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, and $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s. For almost all s^*, the variance of the estimate Y_n is equal to*

$$\operatorname{Var}(Y_n) = \frac{v(s^*)}{n}(1 + \mathcal{O}\left(\delta^2\right))$$

as $\delta \downarrow 0$, where $v(s^*)$ is defined in Equation 2.2.

Proof. The variance is clearly equal to

$$\operatorname{Var}(Y_n) = \frac{\operatorname{Var}(h(\theta) \mid S \in B_\delta(s^*))}{n}.$$

Since

$$\begin{aligned}\mathbb{E}(m(S) \mid S \in B_\delta(s^*)) &= \frac{\int_{B_\delta(s^*)} m(s)f(s)\,\mathrm{d}s}{\int_{B_\delta(s^*)} f(s)\,\mathrm{d}s} \\ &= \frac{m(s^*)f(s^*) + \mathcal{O}\left(\delta^2\right)}{f(s^*) + \mathcal{O}\left(\delta^2\right)} \\ &= m(s^*) + \mathcal{O}\left(\delta^2\right)\end{aligned}$$

as $\delta \downarrow 0$, then, by the law of total variance, the numerator is equal to

$$\begin{aligned}
\operatorname{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right) &= \mathbb{E}\left(v(S) \mid S \in B_\delta(s^*)\right) + \operatorname{Var}\left(m(S) \mid S \in B_\delta(s^*)\right) \\
&= \frac{\int_{B_\delta(s^*)} v(s) f_S(s) \,\mathrm{d}s}{\int_{B_\delta(s^*)} f_S(s^*) \,\mathrm{d}s} \\
&\quad + \frac{\int_{B_\delta(s^*)} \left(m(s) - m(s^*) + \mathcal{O}\left(\delta^2\right)\right)^2 f_S(s) \,\mathrm{d}s}{\int_{B_\delta(s^*)} f_S(s^*) \,\mathrm{d}s} \\
&= \frac{\int_{B_1(0)} v(s^* + \delta\epsilon) f_S(s^* + \delta\epsilon) \,\mathrm{d}\epsilon}{\int_{B_1(0)} f_S(s^* + \delta\epsilon) \,\mathrm{d}\epsilon} \\
&\quad + \frac{\int_{B_1(0)} \left(m(s^* + \delta\epsilon) - m(s^*) + \mathcal{O}\left(\delta^2\right)\right)^2 f_S(s^* + \delta\epsilon) \,\mathrm{d}\epsilon}{\int_{B_1(0)} f_S(s^* + \delta\epsilon) \,\mathrm{d}\epsilon} \\
&= v(s^*) + \mathcal{O}\left(\delta^2\right)
\end{aligned}$$

as $\delta \downarrow 0$. The result follows. □

3.2.3 Optimising the Error and Cost

Combining the bias and variance from Theorems 3.4 and 3.6 gives the mean square error

$$\text{MSE}(Y_n) = \frac{v(s^*)}{n}\left(1 + \mathcal{O}\left(\delta^2\right)\right) + c(s^*)^2\delta^4 + \mathcal{O}\left(\delta^5\right), \tag{3.2}$$

as $\delta \downarrow 0$. This is asymptotically decreasing in δ, but must be balanced against the expected computational cost of generating the samples. The cost is negative binomial, with n required success, and success probability

$$p(s^*) = \int_{B_\delta(s^*)} f_S(s) \,\mathrm{d}s = |B_1| f_S(s^*)\delta^q(1 + \mathcal{O}\left(\delta^2\right)) \tag{3.3}$$

as $\delta \downarrow 0$, so the expected cost takes the form [Voss, 2013, Lemma 5.9]

$$\begin{aligned}
C &= \mathbb{E}\left(\text{Cost of } n \text{ accepted proposals} \mid s^*\right) \qquad (3.4) \\
&= \frac{n\mathbb{E}\left(\text{Cost of one proposal}\right)}{p(s^*)} \\
&= \frac{\mathbb{E}\left(\text{Cost of one proposal}\right)}{|B_1|\, f_S(s^*)} \frac{n}{\delta^q}(1 + \mathcal{O}\left(\delta^2\right)) \qquad (3.5) \\
&= k(s^*) n \delta^{-q}(1 + \mathcal{O}\left(\delta^2\right)).
\end{aligned}$$

We want to know, given a desired value C for the expected computational cost, which values of δ and n will minimise the MSE of Y_n. We do this in two steps. First, in Theorem 3.7, we show that δ and n can be chosen so that

$\textsc{mse}(Y_n) = \mathcal{O}\left(C^{-\frac{4}{q+4}}\right)$ as C tends to infinity. Second, in Theorem 3.8, we show that this rate cannot be improved upon, only the proportionality constant.

Theorem 3.7. *Let $\delta \downarrow 0$ and $n \uparrow \infty$ as $C \uparrow \infty$, such that $D := \lim_{C\uparrow\infty} n\delta^4$ exists, and is strictly positive. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*) > 0$. Then the error and expected cost are such that*

$$\lim_{C\uparrow\infty} n\,\textsc{mse}(Y_n),\ \lim_{C\uparrow\infty} n^{-(1+q/4)}C,\ \textit{and}\ \lim_{C\uparrow\infty} C^{4/(q+4)}\textsc{mse}(Y_n)$$

exist, and are non-zero and finite.

Before we prove Theorem 3.7, we give a rough approach that justifies the introduction of D. If we wish to show that $\textsc{mse}(Y_n) = cC^\alpha + \mathcal{O}\left(C^{\alpha+1}\right)$ as $C \uparrow \infty$, for some constant $c > 0$ and some $\alpha < 0$, then we require the limit of

$$C^{-\alpha}\textsc{mse}(Y_n) = \left(k(s^*)n\delta^{-q}\right)^{-\alpha}\left(1+\mathcal{O}\left(\delta^2\right)\right)^{-\alpha}\left(\frac{v(s^*)}{n} + c(s^*)^2\delta^4 + \mathcal{O}\left(\delta^5\right)\right),$$

as $C \uparrow \infty$, to be non-zero and finite. By rearranging, and noting the fact that $\left(1+\mathcal{O}\left(\delta^2\right)\right)^k = 1 + \mathcal{O}\left(\delta^2\right)$ as $\delta \downarrow 0$ for all $k > 0$, we see that we require the limit of

$$C^{-\alpha}\textsc{mse}(Y_n) = k(s^*)^{-\alpha}(n^{-1-\alpha}\delta^{\alpha q})\left(v(s^*) + c(s^*)^2 n\delta^4 + \mathcal{O}\left(\delta^5\right)\right)\left(1+\mathcal{O}\left(\delta^2\right)\right),$$

as $C \uparrow \infty$, to be non-zero and finite. This holds if the limits for $n^{-1-\alpha}\delta^{\alpha q}$ and $n\delta^4$ exist, and are finite. If we assume that the latter limit $D := \lim_{C\uparrow\infty} n\delta^4 > 0$ is finite, then the former limit exists if $\alpha \geq -4/(q+4)$, and is equal to

$$D_2 := \lim_{C\uparrow\infty} n^{-1-\alpha}\delta^{\alpha q} = D^{\alpha q/4} \lim_{C\uparrow\infty} n^{-1-\alpha(1+q/4)}.$$

Since we wish to maximise the rate of convergence, we minimise α by setting it equal to $-4/(q+4)$. Then $D_2 = D^{-q/(q+4)}$, and $\textsc{mse}(Y_n) = \mathcal{O}\left(C^{-4/(q+4)}\right)$ as $C \uparrow \infty$, with proportionality constant

$$\lim_{C\uparrow\infty} C^{4/(q+4)}\textsc{mse}(Y_n) = k(s^*)^{4/(q+4)}D^{-q/(q+4)}\left(v(s^*) + c(s^*)^2 D\right) > 0.$$

Proof. Using Corollary 3.3 and Theorem 3.4, we find that

$$\begin{aligned}
\lim_{K\uparrow\infty} n\,\textsc{mse}(Y_n) &= \lim_{C\uparrow\infty} n\left(\operatorname{Var}(Y_n) + \operatorname{bias}(Y_n)^2\right) \\
&= v(s^*) + \lim_{C\uparrow\infty} n\left(c(s^*)\delta^2 + \mathcal{O}\left(\delta^3\right)\right)^2 \\
&= v(s^*) + \lim_{C\uparrow\infty} (c(s^*) + \mathcal{O}(\delta))^2 n\delta^4 \\
&= v(s^*) + c(s^*)^2 D.
\end{aligned}$$

For the expected cost, using Equation 3.4, we find that

$$\begin{aligned}\lim_{C\uparrow\infty} n^{-(1+q/4)}C &= k(s^*) \lim_{C\uparrow\infty} n^{-(1+q/4)} n\delta^{-q}(1+\mathcal{O}\left(\delta^2\right)) \\ &= k(s^*) \lim_{C\uparrow\infty} n^{-q/4}\delta^{-q}(1+\mathcal{O}\left(\delta^2\right)) \\ &= k(s^*)D^{-q/4}.\end{aligned}$$

Finally, combining the above results for cost and error, we get the result

$$\begin{aligned}\lim_{C\uparrow\infty} C^{4/(q+4)}\mathrm{MSE}(Y_n) &= \lim_{C\uparrow\infty}\left(n^{-(1+q/4)}C\right)^{4/(q+4)} n\,\mathrm{MSE}(Y_n\mid s^*) \\ &= k(s^*)^{4/(q+4)}D^{-q/(q+4)}\left(v(s^*)+c(s^*)^2D\right),\end{aligned} \tag{3.6}$$

which is non-zero and finite. □

Some remarks:

1. Since we can choose the value of D, we can choose its value to minimise the proportionality constant. This value is $D = qv(s^*)/4c(s^*)^2$, which gives

$$\lim_{C\uparrow\infty} C^{4/(q+4)}\mathrm{MSE}(Y_n) = \left(\frac{qv(s^*)k(s^*)}{4c(s^*)^2}\right)^{-q/(q+4)}(1+q/4)v(s^*)k(s^*).$$

   Theoretically, we can also choose the summary statistic function s that minimises this leading term for a fixed q. This is equivalent to minimising $(v(s^*)k(s^*))^2/c(s^*)^q$ However, since this requires knowledge of the values of $k(s^*)$, $v(s^*)$, and $c(s^*)$, finding the optimal D and s is rarely feasible.

2. If $c(s^*) = 0$, the bias has a faster rate of convergence, and this allows a faster rate for the mean square error. Specifically, if $\mathrm{bias}(Y_n) = \mathcal{O}\left(\delta^r\right)$, it can be shown that $\mathrm{MSE} = \mathcal{O}\left(C^{-\frac{2r}{q+2r}}\right)$.

3. If considered naïvely, the statement $\mathrm{MSE}(Y_n) = \mathcal{O}\left(C^{-4/(q+4)}\right)$ would suggest that there is no convergence when we use a minimal statistic that is infinite-dimensional. However, the proportionality constant depends on $c(s^*)$, and therefore on q, and this can prevent the convergence rate from vanishing. An example is examined in Chapter 5

We now show that, in the case where $c(s^*)$ is non-zero, no other choice of δ can lead to a better asymptotic convergence rate.

Theorem 3.8. *Let $\delta \downarrow 0$ and $n \uparrow \infty$ as $C \uparrow \infty$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*)$ and $c(s^*)$ be non-zero. Then, for almost all $s^* \in \mathbb{R}^q$,*

$$\liminf_{n\uparrow\infty} C^{4/(q+4)} \textsc{mse}(Y_n) > 0.$$

Proof. We know that

$$\begin{aligned} \textsc{mse}(Y_n) &= \mathrm{Var}\,(Y_n) + \mathrm{bias}(Y_n)^2 \\ &= \frac{\mathrm{Var}\,(h(\theta)\,|\,S \in B_\delta(s^*))}{n} + \left(c(s^*)\delta^2 + \mathcal{O}\left(\delta^3\right)\right)^2 \\ &= \frac{\mathrm{Var}\,(h(\theta)\,|\,S \in B_\delta(s^*))}{n} + c(s^*)^2\delta^4 + \mathcal{O}\left(\delta^5\right) \\ &\geq \frac{\mathrm{Var}\,(h(\theta)\,|\,S \in B_\delta(s^*))}{n} + \frac{c(s^*)^2}{2}\delta^4 \end{aligned} \tag{3.7}$$

for all sufficiently large n, and thus all sufficiently small δ. By Lemma A.3, this is bounded by

$$\begin{aligned} \textsc{mse}(Y_n) \geq &\left(\frac{4}{q+4}\frac{\mathrm{Var}\,(h(\theta)\,|\,S \in B_\delta(s^*))}{n}\right)^{4/(q+4)} \left(\frac{q}{q+4}\frac{c(s^*)^2}{2}\delta^4\right)^{q/(q+4)} \\ = &\,A(\delta)^{4/(q+4)} B^{q/(q+4)} \left(n\delta^{-q}\right)^{-4/(q+4)}, \end{aligned}$$

where

$$A(\delta) := \frac{4}{q+4}\mathrm{Var}\,(h(\theta)\,|\,S \in B_\delta(s^*)), \quad B := \frac{q}{q+4}\frac{c(s^*)^2}{2}.$$

For the expected cost, we have

$$(n\delta^{-q})^{-1} C \geq \frac{1}{2}k(s^*),$$

for all sufficiently large n, and so

$$(n\delta^{-q})^{-4/(q+4)} C^{4/(q+4)} \geq \left(\frac{1}{2}k(s^*)\right)^{4/(q+4)}.$$

Therefore, for sufficiently large n, we have

$$C^{4/(q+4)} \textsc{mse}(Y_n) \geq A(\delta)^{4/(q+4)} B^{q/(q+4)} \left(\frac{1}{2}k(s^*)\right)^{4/(q+4)}.$$

Since the right hand side is greater than zero, we have the required result. □

3.3 Numerical Experiments

To demonstrate the above results, we consider the following toy problem.

1. We choose $p = 1$, and assume that our prior belief in the value of the single parameter θ has a standard normal distribution.
2. We assume that the data X consists of two IID samples, X_1 and X_2, each with conditional distribution $\mathrm{N}(\theta, 1)$.
3. We choose $q = 2$, and the (non-minimal) sufficient statistic to be $S(x) = x$ for all $x \in \mathbb{R}^2$.
4. 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 $\mathbb{P}\left(\theta \in [-1/2, 1/2] \,|\, s^*\right)$.
5. We fix the observed data to be $s^* = (1, 1)$.

This problem is simple enough that all the quantities of interest can be determined explicitly. In particular, we have the conditional distributions $\theta \,|\, S \sim \mathrm{N}\left((s_1 + s_2)/3, 1/3\right)$ and $\theta \,|\, s^* \sim \mathrm{N}(2/3, 1/3)$, and S is bivariate normally distributed with mean 0 and covariance matrix

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

To find the value of the bias coefficient $c(s^*)$, we proceed as in Example 3.5. From the above, we can see that

$$m(s^*) = \Phi\left(\frac{1}{\sqrt{3}}\left(3/2 - 1^T s^*\right)\right) - \Phi\left(-\frac{1}{\sqrt{3}}\left(3/2 + 1^T s^*\right)\right),$$

with derivatives

$$\begin{aligned}\nabla m(s^*) &= \frac{1}{\sqrt{3}} 1^T \phi\left(-\frac{1}{\sqrt{3}}\left(3/2 + 1^T s^*\right)\right) - \frac{1}{\sqrt{3}} 1^T \phi\left(\frac{1}{\sqrt{3}}\left(3/2 - 1^T s^*\right)\right) \\ &= \frac{1}{\sqrt{3}} 1^T \phi\left(-7/2\sqrt{3}\right) - \frac{1}{\sqrt{3}} 1^T \phi\left(-1/2\sqrt{3}\right),\end{aligned}$$

$$\begin{aligned}\Delta m(s^*) =& \frac{2}{3\sqrt{3}}(3/2 + 1^T s^*)\phi\left(-\frac{1}{\sqrt{3}}\left(3/2 + 1^T s^*\right)\right) \\ &- \frac{2}{3\sqrt{3}}(3/2 - 1^T s^*)\phi\left(\frac{1}{\sqrt{3}}\left(3/2 - 1^T s^*\right)\right) \\ =& \frac{7}{3\sqrt{3}}\phi\left(-7/2\sqrt{3}\right) + \frac{1}{3\sqrt{3}}\phi\left(-1/2\sqrt{3}\right).\end{aligned}$$

Additionally, we can show the statistic has marginal log density

$$\log(f_S(s^*)) = -\frac{1}{2}\log(2\pi\,|\Sigma|) - \frac{1}{2} s^{*T} \Sigma^{-1} s^*,$$

with derivative

$$\nabla \log(f_S(s^*)) = -s^{*T}\Sigma^{-1} = -\frac{1}{3}\begin{pmatrix}1 & 1\end{pmatrix}.$$

The bias coefficient $c(s^*)$ is thus equal to

$$c(s^*) = \frac{3\phi\left(-7/2\sqrt{3}\right) + 5\phi\left(-1/2\sqrt{3}\right)}{24\sqrt{3}}.$$

It can be shown numerically that the prior and posterior expectation for $h(\theta)$ are $\mathbb{E}\left(h(\theta)\right) = 0.3829$ and $m(s^*) = 0.3648$, respectively, and that the bias coefficient is $c(s^*) = 0.0323$.

The function

$$\phi_1(s) = f_S(s) = \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{3}(s_1^2 - s_1 s_2 + s_2^2)}$$

is multivariate normal, so its third derivatives exist, and are bounded and continuous, as do those for the function

$$\phi_h(s) = \int_{-1/2}^{1/2} f_{\theta,S}(t,s)dt \leq \phi_1(s).$$

Thus, the assumptions hold.

The figures in this section were plotted using scripts in the R programming language, which are given in Appendix B.1.1.

Experiment 1

This experiment demonstrates the statement of Theorem 3.4. For fixed δ, we generate k independent ABC estimates, each based on n accepted proposals. For each of the k estimates, we calculate its discrepancy from the true posterior expectation. We then calculate the discrepancies' mean and standard error, 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 3.3.1 shows the result of a simulation, using $n = 500$ accepted proposals for each ABC estimate, and using $k = 5000$ ABC estimates for each value of δ. For comparison, the figure includes the theoretically predicted asymptotic bias $c(s^*)\delta^2$, using the value $c(s^*) = 0.0323$. The plot shows that the theoretical curve is indeed a good fit to the numerical estimates of the bias for small values of δ. For larger values of δ, the bias tends towards the difference between the prior and true posterior expectations.

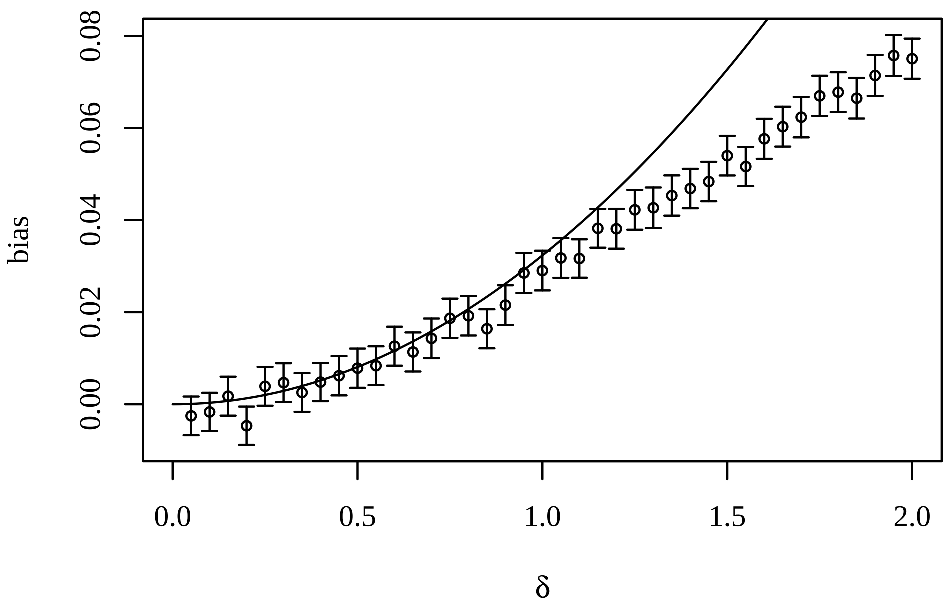

Figure 3.3.1: Simulation results for Experiment 1. The error bars indicate mean ±1.96 standard errors, estimated from using 5000 samples for each value of δ. The parabola uses the theoretical constant from Theorem 3.4.

Experiment 2

This experiment demonstrates the statement of Theorems 3.7 and 3.8, by numerically estimating the optimal choice of δ and the corresponding ABC.

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 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 [Voss, 2013, Lemma 5.9], we know that the expected cost is $\mathcal{O}\left(n\delta^{-2}\right)$, as $n \uparrow \infty$ and $\delta \downarrow 0$, and thus we choose $n = \mathcal{O}\left(\delta^2\right)$ in this example. From Theorem 3.4, we know that $\text{bias}(Y_n) = \mathcal{O}\left(\delta^2\right)$. Thus, we expect the MSE for constant expected cost to be of the form

$$\text{MSE}(Y_n) = \frac{\text{Var}(Y_n)}{n} + \text{bias}(Y_n)^2 \simeq a\delta^{-2} + b\delta^4$$

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 $k = 500$

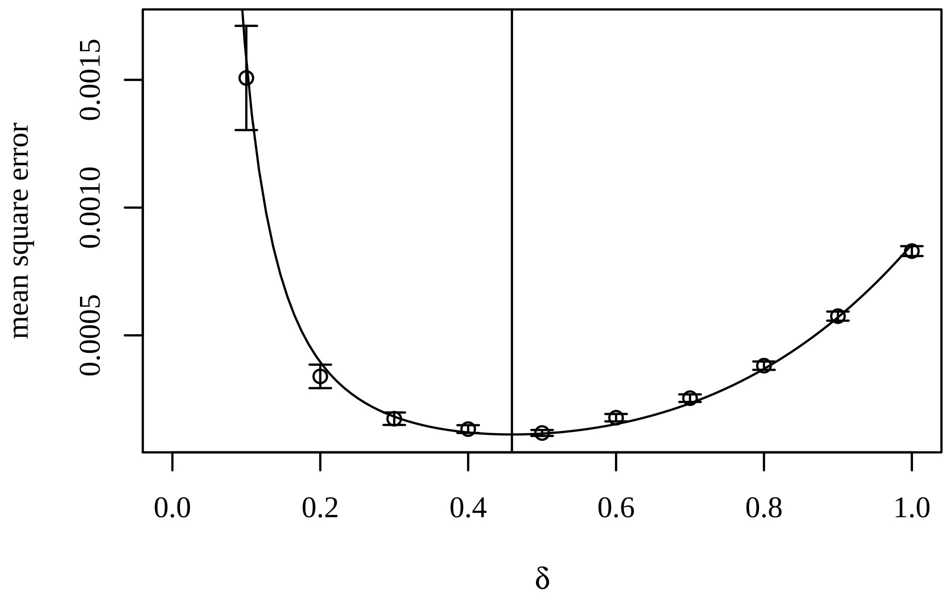

Figure 3.3.2: The estimated MSE as a function of δ (for fixed expected cost), together with the fitted curve and the location of the optimal δ. This figure is for the time constant 16000. We set $k = 500$ instead of 5000. However, the error bars are already relatively small.

samples for each δ, is shown in figure 3.3.2, and shows the curve to be a good fit.

This good fit between the fitted curve and the empirical MSE justifies estimating the optimal values of δ and MSE, given the expected computational cost, as those at the minimum of the fitted curve.

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 dependency of the optimal δ and the MSE on the cost to take the form $x = A \cdot \text{cost}^B$. To demonstrate the statements of Theorem 3.8 we numerically estimate the exponent B from simulated data. The result of such a simulation is shown in figure 3.3.3. The data are roughly on straight lines, as expected, and 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.

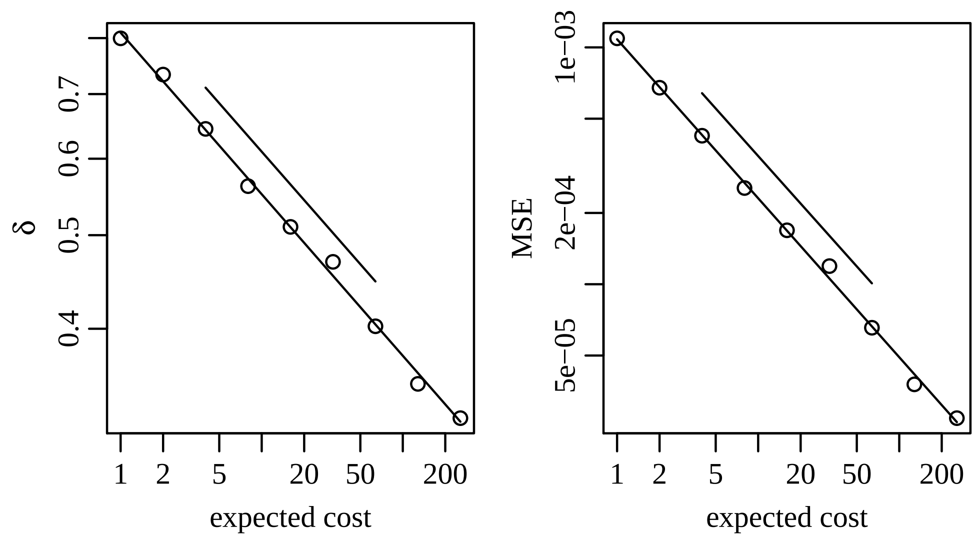

Figure 3.3.3: Numerically found dependency of the optimal δ and the corresponding MSE on the computational cost. The computational cost is given in arbitrary units, chosen such that the smallest sample size under consideration has cost 1. For comparison, the additional line above the fit has the gradient expected from the theoretical results.

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 the empirical values and the theoretically predicted values.

3.4 Convergence of ABC Variants

In this section, we look at some simple variants on the algorithm for the ABC estimate Y_n, as given in Algorithm 2.1.2, and examine their effect on the estimate' optimal rate of convergence, as given in Theorem 3.8.

3.4.1 Constant Number of Proposals

We first consider the estimate Z_N, as described in Algorithm 2.1.3: instead of stopping the algorithm after a fixed number of acceptances, we stop it after a fixed number of proposals. For cases where the cost of one sample is roughly constant, this is roughly equivalent to fixing the computational cost.

To consider this estimate, we need to consider the distribution of the number of accepted proposals, and what to do when none of the proposals are accepted.

For the fixed number of proposals N, the number n of accepted proposals is binomial, with parameters

$$n \sim \mathrm{Bin}(N, p(s^*)).$$

For the case where no proposals are accepted, we can use the expectation of the prior density, $\mathbb{E}\left(h\left(\theta\right)\right).$ In this case, the mean square error is equal to $\left(\mathbb{E}\left(h\left(\theta\right)\right) - m(s^*)\right)^2.$

Theorem 3.9. *Let the function $h\colon \mathbb{R}^p \to \mathbb{R}$ be such that $\mathbb{E}\left(|h(\theta)|\right) < \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^p$, the* ABCBAS *estimate Z_N satisfies*

1. $\lim_{N\uparrow\infty} Z_N = \mathbb{E}\left(Y_n\right)$ *almost surely for all $\delta > 0$; and*
2. $\lim_{\delta\downarrow 0} \mathbb{E}\left(Y_n\right) = m(s^*)$ *for all $n \in \mathbb{N}$.*

Proof. The probability of Z_N converging to $\lim_{n\uparrow\infty} Y_n$ satisfies

$$\begin{aligned}\mathbb{P}\left(\lim_{N\uparrow\infty} Z_N = \lim_{n\uparrow\infty} Y_n\right) &\geq \mathbb{P}\left(\lim_{N\uparrow\infty} n = \infty\right)\\ &= \mathbb{P}\left(\text{accept infinitely often}\right).\end{aligned}$$

Since each sample is accepted with the same non-zero probability, it follows that $\sum_k \mathbb{P}\left(S_k\,\text{accepted}\right) = \infty$, and therefore $\mathbb{P}\left(\lim_{N\uparrow\infty} Z_N = \lim_{n\uparrow\infty} Y_n\right) = 1$ by the second Borel-Cantelli Lemma.

The proof of the second statement is the same as for Theorem 3.2. □

If $n > 0$ proposals are accepted, the bias is the same as for Y_n, and the expected cost is $N\mathbb{E}\left(\text{Cost of 1 proposal}\right).$ For the variance, we know that

$$\mathrm{Var}\left(Z_N \,|\, n\right) = \frac{\mathrm{Var}\left(h(\theta)\,|\, S \in B_\delta(s^*)\right)}{n}.$$

We now show that Z_N has the same order of convergence as Y_n. This is proved slightly differently in Barber et al. [2015]

Lemma 3.10. *Let $n \sim Bin(N, p(s^*))$. Then*

$$\lim_{Np(s^*)\uparrow\infty} Np(s^*)\mathbb{E}\left(\frac{1}{n}[n > 0]\,\middle|\, s^*\right) = 1.$$

Proof. For any $0<\epsilon<1$, we split the expectation into three parts:

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}[n>0]\,\middle|\,s^*\right)=&\,\mathbb{E}\left(\frac{1}{n}[0<n\leq(1-\epsilon)Np(s^*)]\,\middle|\,s^*\right)\\&+\mathbb{E}\left(\frac{1}{n}[n\geq(1+\epsilon)Np(s^*)]\,\middle|\,s^*\right)\\&+\mathbb{E}\left(\frac{1}{n}[(1-\epsilon)Np(s^*)<n<(1+\epsilon)Np(s^*)]\,\middle|\,s^*\right).\end{aligned}$$

The first term satisfies the bound

$$\mathbb{E}\left(\frac{1}{n}[0<n\leq(1-\epsilon)Np(s^*)]\,\middle|\,s^*\right)\leq\mathbb{P}\left(0<n\leq(1-\epsilon)Np(s^*)\,|\,s^*\right),$$

and, by Lemma A.10, therefore satisfies

$$\mathbb{E}\left(\frac{1}{n}[0<n\leq(1-\epsilon)Np(s^*)]\,\middle|\,s^*\right)\leq\exp\left(-\epsilon^2Np(s^*)/2\right),$$

which tends to zero exponentially quickly. The second term satisfies the bound

$$\mathbb{E}\left(\frac{1}{n}[n\geq(1+\epsilon)Np(s^*)]\,\middle|\,s^*\right)\leq\frac{1}{(1+\epsilon)Np(s^*)}\mathbb{P}\left(n\geq(1+\epsilon)Np(s^*)\,|\,s^*\right),$$

and, by Lemma A.11, therefore satisfies

$$\mathbb{E}\left(\frac{1}{n}[n\geq(1+\epsilon)Np(s^*)]\,\middle|\,s^*\right)\leq\frac{1}{(1+\epsilon)Np(s^*)}\exp\left(-\epsilon^2Np(s^*)/3\right),$$

and vanishes exponentially quickly. We are left with the final term, which is

$$\mathbb{E}\left(\frac{1}{n}\left[(1-\epsilon)Np(s^*)<n<(1+\epsilon)Np(s^*)\right]\,\middle|\,s^*\right)=\mathbb{E}\left(\frac{1}{n}\left[\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\right]\,\middle|\,s^*\right).$$

This satisfies the lower bound

$$\mathbb{E}\left(\frac{1}{n}\left[\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\right]\,\middle|\,s^*\right)>\frac{1}{(1+\epsilon)Np(s^*)}\mathbb{P}\left(\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\,\middle|\,s^*\right),$$

and the upper bound

$$\mathbb{E}\left(\frac{1}{n}\left[\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\right]\,\middle|\,s^*\right)<\frac{1}{(1-\epsilon)Np(s^*)}\mathbb{P}\left(\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\,\middle|\,s^*\right).$$

The probability $\mathbb{P}\left(\left|\frac{n}{Np(s^*)}-1\right|<\epsilon\,\middle|\,s^*\right)$ tends to one, so the final term in the limit is bounded in $[1/(1+\epsilon),1/(1-\epsilon)]$. Letting ϵ tend to zero gives the result.

□

Lemma 3.11. *The variance of the estimate Z_N satisfies*

$$\lim_{N\delta^q\uparrow\infty}Np(s^*)\mathrm{Var}\left(Z_N\right)=v(s^*).$$

Proof. The variance is equal to

$$\begin{aligned}\mathrm{Var}\left(Z_N\right) =& \mathbb{E}\left(\mathrm{Var}\left(Z_N \mid n\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(Z_N \mid n\right)\right)\\ =& \mathbb{E}\left(\frac{\mathrm{Var}\left(h(\theta)[n>0] \mid S \in B_\delta(s^*)\right)}{n}\right)\\ &+ \mathrm{Var}\left(\mathbb{E}\left(h(\theta) \mid S \in B_\delta(s^*)\right)[n>0] + \mathbb{E}\left(h(\theta)\right)[n=0]\right)\\ =& \mathrm{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right)\mathbb{E}\left(\frac{[n>0]}{n}\right)\\ &+ \mathrm{Var}\left(\mathbb{E}\left(h(\theta)\right) + \left(\mathbb{E}\left(h(\theta) \mid S \in B_\delta(s^*)\right) - \mathbb{E}\left(h(\theta)\right)\right)B\right),\end{aligned}$$

where B is a Bernoulli variable with success probability $p_B(s^*)$, such that

$$1 - p_B(s^*) = \mathbb{P}\left(n = 0 \mid s^*\right) \leq \exp\left(-Np(s^*)/2\right),$$

by Lemma A.10. The variance of B satisfies

$$\lim_{Np(s^*)\uparrow\infty} Np(s^*)p_B(1-p_B) \leq \lim_{Np(s^*)\uparrow\infty} Np(s^*)\exp\left(-Np(s^*)/2\right) = 0,$$

so the second term vanishes exponentially quickly. By Lemma 3.10, the first term satisfies

$$\lim_{Np(s^*)\uparrow\infty} Np(s^*)\mathrm{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right)\mathbb{E}\left(\frac{[n>0]}{n} \,\middle|\, s^*\right) = v(s^*),$$

as required. □

The MSE of the estimate Z_N is, therefore, equal to

$$\mathrm{MSE}(Z_N) = \left(\frac{\hat{v}(s^*)}{N\delta^q} + c(s^*)^2\delta^4\right)\left(1 + \mathrm{o}\left(1\right)\right), \tag{3.8}$$

as $N\delta^q \uparrow \infty$, where $\hat{v}(s^*) := v(s^*)/f_S(s^*)|B_1|$, and the computational cost is clearly proportional to the number of proposals. Therefore, we can prove the following theorem.

Theorem 3.12. *Let $\delta \downarrow 0$ and $N \uparrow \infty$ as $C \uparrow \infty$, such that $D := \lim_{C\uparrow\infty} N\delta^{q+4}$ exists, and is strictly positive. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*) > 0$. Then the error $\mathrm{MSE}(Z_N)$ and expected cost C are such that*

$$\lim_{C\uparrow\infty} N\delta^q\,\mathrm{MSE}(Z_N),\ \lim_{C\uparrow\infty} N^{-(1+q/4)}\delta^{-q(1+q/4)}C,\ \textit{and}\ \lim_{C\uparrow\infty} C^{4/(q+4)}\mathrm{MSE}(Z_N)$$

exist, and are non-zero and finite.

Proof. Using Lemma 3.11, we find that

$$\begin{aligned}
\lim_{C\uparrow\infty} N\delta^q \,\textsc{mse}(Z_N) &= \lim_{C\uparrow\infty} N\delta^q \left(\mathrm{Var}\,(Z_N) + \mathrm{bias}(Z_N)^2\right) \\
&= \hat{v}(s^*) + \lim_{C\uparrow\infty} N\delta^q \left(c(s^*)^2\delta^4 + \mathcal{O}\left(\delta^5\right)\right) \\
&= \hat{v}(s^*) + c(s^*)^2 \lim_{C\uparrow\infty} N\delta^{q+4} \\
&= \hat{v}(s^*) + c(s^*)^2 D.
\end{aligned}$$

For the expected cost $C = k_2(s^*)N$, we find that

$$\begin{aligned}
\lim_{C\uparrow\infty} N^{-(1+q/4)}\delta^{-q(1+q/4)}C &= k_2(s^*) \lim_{C\uparrow\infty} N^{-q/4}\delta^{-q(1+q/4)} \\
&= k_2(s^*)D^{-q/4}.
\end{aligned}$$

Finally, combining the above results for cost and error, we get the result

$$\begin{aligned}
\lim_{C\uparrow\infty} C^{4/(q+4)}\textsc{mse}(Z_N) = &\lim_{C\uparrow\infty} \left(N^{-(1+q/4)}\delta^{-q(1+q/4)}C\right)^{4/(q+4)} \\
&\times N\delta^q \,\textsc{mse}(Z_N) \\
= &k_2(s^*)^{4/(q+4)} D^{-q/(q+4)} \left(\hat{v}(s^*) + c(s^*)^2 D\right),
\end{aligned}$$

which is non-zero and finite. □

Theorem 3.13. *Let $\delta \downarrow 0$ and $N \uparrow \infty$ as $C \uparrow \infty$, in such a way that $N\delta^q \uparrow \infty$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*)$ and $c(s^*)$ be non-zero. Then, for f_S-almost all $s^* \in \mathbb{R}^q$,*

$$\liminf_{N\uparrow\infty} C^{4/(q+4)}\textsc{mse}(Z_N) > 0.$$

Proof. We know that

$$\begin{aligned}
\textsc{mse}(Z_N) &= \mathrm{Var}\,(Z_N) + \mathrm{bias}(Z_N)^2 \\
&= \frac{\mathrm{Var}\,(h(\theta) \mid S \in B_\delta(s^*))}{N\delta^q f_S(s^*)|B_1|} + \left(c(s^*)\delta^2 + \mathcal{O}\left(\delta^3\right)\right)^2 \\
&= \frac{\mathrm{Var}\,(h(\theta) \mid S \in B_\delta(s^*))}{N\delta^q f_S(s^*)|B_1|} + c(s^*)^2\delta^4 + \mathcal{O}\left(\delta^5\right) \\
&\geq \frac{\mathrm{Var}\,(h(\theta) \mid S \in B_\delta(s^*))}{N\delta^q f_S(s^*)|B_1|} + \frac{c(s^*)^2}{2}\delta^4,
\end{aligned}$$

for all sufficiently large N and $N\delta^q$. By Lemma A.3, this is bounded by

$$\begin{aligned}
\textsc{mse}(Z_N) \geq &\left(\frac{4}{q+4}\frac{\mathrm{Var}\,(h(\theta) \mid S \in B_\delta(s^*))}{N\delta^q f_S(s^*)|B_1|}\right)^{4/(q+4)} \left(\frac{q}{q+4}\frac{c(s^*)^2}{2}\delta^4\right)^{q/(q+4)} \\
= &A(\delta)^{4/(q+4)} B^{q/(q+4)} N^{-4/(q+4)},
\end{aligned}$$

where

$$A(\delta) := \frac{4}{q+4}\frac{\operatorname{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right)}{f_S(s^*)|B_1|}, \quad B := \frac{q}{q+4}\frac{c(s^*)^2}{2}.$$

For the expected cost, we have

$$N^{-1}C = k(s^*),$$

and so

$$N^{-4/(q+4)}C^{4/(q+4)} = k(s^*)^{4/(q+4)}.$$

Therefore, for sufficiently large N and $N\delta^q$, we have

$$C^{4/(q+4)}\mathrm{MSE}(Y_n) \geq A(\delta)^{4/(q+4)}B^{q/(q+4)}k(s^*)^{4/(q+4)}.$$

Since the right hand side is greater than zero, we have the required result. □

3.4.2 Non-Ball Norms and Random Acceptance

If we use the generalised algorithm in Section 2.1.2, the proof for the asymptotic convergence rate is similar to before, with small changes. Instead of using $\phi_h^{(\delta)}$ from Definition 3.1, we define the following:

Definition 3.14.

$$\begin{aligned}\phi_h^{(\delta,H)}(s^*) :=& \frac{\int K_H(s-s^*)\phi_h(s)\,\mathrm{d}s}{\int K_H(s-s^*)\,\mathrm{d}s} \\ =& \frac{\int K(u)\phi_h(s^*+H^{1/2}u)\,\mathrm{d}u}{\int K(u)\,\mathrm{d}u} \\ =& \int K(u)\phi_h(s^*+H^{1/2}u)\,\mathrm{d}u.\end{aligned}$$

In particular, $\phi_h^{(\delta,H)}$ *is such that, for the estimate with kernel function* K *and square-bandwidth matrix* H,

$$\mathbb{E}\left(h(\theta) \mid S \,\text{accepted}\ \textsc{wrt}\ s^*\right) = \frac{\phi_h^{(\delta,H)}(s^*)}{\phi_1^{(\delta,H)}(s^*)}.$$

We then expand as before, using the known values for the lower-order moments of K from Section 2.1.2:

$$\begin{aligned}
\phi_h^{(\delta,H)}(s^*) =& \int K(u)\phi_h(s^*)\,\mathrm{d}u + \int K(u)\Delta\phi_h(s^*)H^{1/2}u\,\mathrm{d}u \\
&+ \int K(u)\left(\frac{1}{2}uH^{1/2}\mathcal{H}_{\phi_h}(s^*)H^{1/2}u\right)\,\mathrm{d}u + \mathcal{O}\left(\mathrm{Tr}(H)^2\right) \\
=&\phi_h(s^*) + \frac{1}{2}\mathrm{Tr}\left(\int K(u)uu^T\,\mathrm{d}uH^{1/2}\mathcal{H}_{\phi_h}(s^*)H^{1/2}\right) + \mathcal{O}\left(\mathrm{Tr}(H)^2\right) \\
=&\phi_h(s^*) + \frac{1}{2}\mu_2(K)\mathrm{Tr}\left(H\mathcal{H}_{\phi_h}(s^*)\right) + \mathcal{O}\left(\mathrm{Tr}(H)^2\right).
\end{aligned}$$

We can now follow the same steps as in the proof for Theorem 3.4 to find that the bias for the random acceptance version is equal to

$$\begin{aligned}
\mathrm{bias}(Z) &= \frac{\mu_2(K)\left(\mathrm{Tr}(H\mathcal{H}_{\phi_h}(s^*)) - m(s^*)\mathrm{Tr}(H\mathcal{H}_{\phi_1}(s^*))\right)}{2\phi_1(s^*)} + \mathcal{O}\left(\mathrm{Tr}(H)^2\right) \\
&= \frac{1}{2}\mu_2(K)\mathrm{Tr}\left(H\left[\mathcal{H}_m(s^*) + 2\nabla m(s^*)\nabla \log f_S(s^*)^T\right]\right) + \mathcal{O}\left(\mathrm{Tr}(H)^2\right).
\end{aligned}$$

The ABCACC and ABCBAS estimates are then the case where $H = \delta^2 I_q$, and where $K(u) = [u \in B_1(0)]/|B_1|$, with $\mu_2(K) = 1/(q+2)$.

For the acceptance probability, we take account of the kernel's scaling:

$$\begin{aligned}
p(s^*) &= \frac{\int K_H(s-s^*)f_S(s)\,\mathrm{d}s}{\max_u K_H(u)} = \frac{\int K(u)f_S(s^*+H^{1/2}u)\,\mathrm{d}u}{|H|^{-1/2}\max_u K(u)} \\
&= \frac{f_S(s^*)}{\max_u K(u)}|H|^{1/2}(1+\mathcal{O}\left(\mathrm{Tr}(H)\right)).
\end{aligned} \tag{3.9}$$

If $H = \delta^2 I$, then $p(s^*) = \mathcal{O}\left(\delta^q\right)$, as before. For the ABCBAS estimate, we now require $N|H|^{1/2} \uparrow \infty$ as $N \uparrow \infty$, where before we required $N\delta^q \uparrow \infty$, and the variance is of order $\mathcal{O}\left(\frac{1}{N|H|^{1/2}}\right)$ as $N|H|^{1/2} \uparrow \infty$.

For either basic estimate, optimising the asymptotic rate of convergence of the estimate now involves an equation that includes both $\mathrm{Tr}(H)$ and $|H|^{1/2}$. This is not trivial to solve for general sequences of values for H. However, if H is of the form

$$H = f(\delta)H_0,$$

for some non-zero constant matrix H_0, where $f(\delta) \downarrow 0$ as $\delta \downarrow 0$, then the two H terms are $\mathrm{Tr}(H) = f(\delta)\mathrm{Tr}(H_0)$ and $|H|^{1/2} = f(\delta)^{q/2}|H_0|$, and the optimal rate can be found with a similar argument to that in Theorems 3.7 and 3.8. Furthermore, since the leading term of the asymptotic error depends on the choice of K and H_0, we can consider the choice of one, or both, to minimise the leading term.

Example 3.15. *If we suppose that the sequence of square-tolerance matrices H is such that $H = \delta^2 I$, then we find that the variant has bias coefficient*

$$\hat{c}(s^*) = \frac{1}{2}\mu_2(K)\left(\triangle m(s^*) + 2\nabla m(s^*)\nabla \log f_S(s^*)^T\right),$$

and acceptance probability

$$p(s^*) = \frac{f_S(s^*)}{\max_u K(u)}\delta^q(1 + \mathcal{O}\left(\delta^2\right)),$$

where $\mu_2(K) = 1/(q+2)$ and $\max_u K(u) = 1/|B_1|$ in the basic estimates. By comparison to the proof for Theorem 3.12, we can show that the mean square error satisfies

$$\begin{aligned}\lim_{C\uparrow\infty} C^{4/(q+4)}\mathrm{MSE}(Z_N) &= \lim_{K\uparrow\infty}\left(N^{-(1+q/4)}\delta^{-q(1+q/4)}C\right)^{4/(q+4)} \\ &\quad\times N\delta^q\,\mathrm{MSE}(Z_N) \\ &= k_2(s^*)^{4/(q+4)}D^{-q/(q+4)}\left(\hat{v}(s^*) + \hat{c}(s^*)^2 D\right),\end{aligned}$$

where $\hat{v}(s^*) := \frac{v(s^*)}{f_S(s^*)}\max_u K(u)$. The value of D that minimises this expression is $D = q\hat{v}(s^*)/4c(s^*)^2$, giving a minimal value of

$$\begin{aligned}\lim_{C\uparrow\infty} C^{4/(q+4)}\mathrm{MSE}(Z_N) &= \frac{1+q/4}{(q/4)^{q/(q+4)}}k_2(s^*)^{4/(q+4)}\hat{v}(s^*)^{4/(q+4)}\hat{c}(s^*)^{2q/(q+4)} \\ &\propto \hat{v}(s^*)^{4/(q+4)}\hat{c}(s^*)^{2q/(q+4)}.\end{aligned}$$

While this still requires knowledge of the values of $\hat{v}(s^*)$ and $\hat{c}(s^*)$, we can consider the choice of kernel that further minimises this optimal leading MSE *term. We therefore want the kernel K that minimises the* kernel inefficiency

$$In(K) := \mu_2(K)^{q/2}\max_u K(u).$$

For example, we compare the kernel inefficiency for three q-dimensional distributions: the uniform kernel $K_1(u) = [u^T u \le 1]/|B_1|$, the normal kernel $K_2(u) = (2\pi)^{-q/2}\exp\left(-u^T u/2\right)$, and the q-dimensional Epanechnikov kernel $K_3(u) := \frac{q+2}{2|B_1|}(1 - u^T u)[u^T u \le 1]$.

For the uniform kernel, in Example 2.12 and the proof of Theorem 3.4, we showed that $\mu_2(K_1) = \frac{1}{q+2}$. Since the kernel is uniform over a region of size $|B_1|$, we also know that $\max_u K_1(u) = 1/|B_1|$. The uniform kernel therefore has kernel inefficiency

$$In(K_1) = \frac{1}{(q+2)^{q/2}|B_1|} = \frac{\Gamma(1+q/2)}{\pi^{q/2}(q+2)^{q/2}}.$$

The normal kernel can be shown to have $\mu_2(K_2) = 1$ *and* $\max_u K_2(u) = (2\pi)^{-q/2}$, *giving a kernel inefficiency of*

$$In(K_2) = 1/(2\pi)^{q/2},$$

and the Epanechnikov kernel satisfies

$$\mu_2(K_3) = \frac{q+2}{2|B_1|}\int_{\|u\|\leq 1} \|u\|^2 \left(1 - \|u\|^2\right) \mathrm{d}u = \frac{q}{2}\left(1 - \frac{q+2}{q+4}\right) = \frac{q}{q+4},$$

and $\max_u K_3(u) = (q+2)/2|B_1|$, *giving a kernel inefficiency of*

$$In(K_3) = \frac{q^{q/2}(q+2)}{2(q+4)^{q/2}|B_1|} = \frac{q^{q/2}(q+2)\Gamma(1+q/2)}{2(q+4)^{q/2}\pi^{q/2}}.$$

For $q = 1$, *the inefficiencies given above are equal to* $1/\sqrt{12}$, $1/\sqrt{2\pi}$, *and* $3/4\sqrt{5}$, *respectively, so the uniform kernel is the more efficient kernel. More generally, we can calculate the inefficiency ratios*

$$\frac{In(K_2)}{In(K_1)} = \frac{(1+q/2)^{q/2}}{\Gamma(1+q/2)}, \quad \frac{In(K_3)}{In(K_1)} = \frac{q^{q/2}(q+2)^{q/2}(1+q/2)}{(q+4)^{q/2}},$$

which are monotonically increasing, and tend to infinity, as q *tends to infinity. Therefore, the uniform kernel has the better asymptotic efficiency for any* q, *and its relative performance increases with* q.

Some remarks:

1. The choice of kernel to minimise some measure of kernel inefficiency also occurs in kernel density estimation, as proposed by Parzen [1962]. However, the resulting choice of kernels is very different: in kernel density estimation, for a one-dimensional density, the kernel inefficiency is defined to be $\mu_2(K)^{1/2}R(K)$, and the Epenechnikov kernel is optimal. For ABC posterior mean estimation, the $\max_u K(u)$ is used instead of $R(K)$.

2. It should be noted that this choice of kernel is for the case where we are only interested in estimating the posterior expectation. If we are also interested in other properties of the posterior distribution, or in estimating the entire distribution, then another choice of kernel may be preferable. Even in the case of posterior mean expectation, random acceptance is still useful for other reasons, such as the use of importance sampling over the prior.

3. It is expected that the uniform kernel can be shown to minimise the kernel inefficiency. Since the kernel inefficiency is invariant with respect to the bandwidth, we can constrain $\mu_2(K)$ to be equal to one, as done by Epanechnikov [1969] for kernel density estimation, and seek to minimise $\max_u K(u)$.

3.4.3 Weighted Proposals

If we weight the proposals by $J_H(S - s^*)$for some kernel function J, we then replace $\phi_h^{(\delta)}$ with

$$\phi_h^{(\delta,J)}(s^*) := \int J_H(s - s^*)\phi_h(s)\,\mathrm{d}s = \int J(u)\phi_h(s^* + H^{1/2}u)\,\mathrm{d}u.$$

Proposals are then considered as accepted if $S - s^* \in \operatorname{supp}(J_H)$. The effect on the bias is similar to that for K, except that $J(\cdot)$ can now be negative. This means that the second moment matrix $M(J)$ of J can be zero, which allows for a higher-order rate of convergence for the bias.

3.4.4 Generalised Ball Acceptance Regions

Suppose that, instead of accepting proposals inside the 2-ball

$$B_\delta^{(2)}(s^*) := B_\delta(s^*) = \left\{ s : \sum_{i=1}^{q} |s_i - s_i^*|^2/\delta^2 \leq 1 \right\},$$

we accept proposals inside the generalised ball $B_\delta^{(l)}$, as defined in Definition 2.13, where $l_k, \delta_k > 0$ for all k. We then use the modified $\phi_h^{(\delta)}$ function

$$\phi_h^{(\delta,l)}(s^*) := \frac{1}{|B_\delta^{(l)}|} \int_{B_\delta^{(l)}(s^*)} \phi_h(s)\,\mathrm{d}s.$$

Since there is a straight line from s^* to each point in $B_\delta^{(l)}(s^*)$, we can use the Taylor expansion for ϕ_h around s^* to show that

$$\begin{aligned}
\phi_h^{(\delta,l)}(s^*) &= \phi_h(s^*) + \frac{1}{2|B_\delta^{(l)}|} \int_{B_\delta^{(l)}(s^*)} (s - s^*)^T \mathcal{H}_{\phi_h}(s^*)(s - s^*)\,\mathrm{d}s + R(s^*, \delta, l) \\
&= \phi_h(s^*) + \frac{1}{2|B_\delta^{(l)}|} \int_{B_\delta^l(0)} u^T \mathcal{H}_{\phi_h}(s^*) u\,\mathrm{d}u + R(s^*, \delta, l) \\
&= \phi_h(s^*) + \frac{1}{2|B_\delta^{(l)}|} \mathrm{Tr}\left(\int_{B_\delta^l(0)} uu^T\,\mathrm{d}u \mathcal{H}_{\phi_h}(s^*) \right) + R(s^*, \delta, l),
\end{aligned}$$

where $R(s^*, \delta, l)$ is the remainder term. Note that the first-order term still disappears, because generalised balls are symmetric on each axis.

We now require evaluations of $|B_\delta^{(l)}|$ and $\int_{B_\delta^l(0)} uu^T\,\mathrm{d}u$.

Lemma 3.16. *Let δ be a positive vector of elliptic radii, and l be a strictly positive vector, with elements $0 < l_k \leq \infty$. Then the generalised ball $B_\delta^{(l)}(0)$ has volume*

$$|B_\delta^{(l)}| = 2^q \frac{\prod_{k=1}^q \Gamma(1+1/l_k)}{\Gamma(1+\sum_{k=1}^q 1/l_k)} \prod_{k=1}^q \delta_k.$$

Furthermore, $\int_{B_\delta^{(l)}(0)} uu^T \,\mathrm{d}u$ is diagonal, with diagonal elements

$$\int_{B_\delta^{(l)}(0)} |u_k|^2 \,\mathrm{d}u = \frac{1}{3}|B_\delta^{(l)}| \frac{\Gamma(1+3/l_k)}{\Gamma(1+1/l_k)} \frac{\Gamma\left(1+\sum_{j=1}^q 1/l_j\right)}{\Gamma\left(1+\sum_{j=1}^q 1/l_j + 2/l_k\right)} \delta_k^2.$$

Proof. The proof for the volume follows Wang [2005]. Firstly, suppose that $p_k < \infty$ for all k. The volume of $|B_\delta^{(l)}|$ is equal to

$$|B_\delta^{(l)}| = \int_{\sum_{k=1}^q |u_k/\delta_k|^{l_k} \leq 1} \mathrm{d}u.$$

We now change variables from the vector u to the vector r, with elements equal to $r_k := (u_k/\delta_k)^{l_k/2}$ for all k [Wang, 2005]. Using the resulting substitution $u_k = r_k^{2/l_k}\delta_k$, we find that

$$|B_\delta^{(l)}| = 2^q \int_{\sum_{k=1}^q |r_k|^2 \leq 1} \prod_{k=1}^q r_k^{2/l_k - 1} \,\mathrm{d}r \prod_{k=1}^q \delta_k/l_k,$$

where the integral is now over the regular ball of radius 1. If we define the function I, where $I(a_1,\ldots,a_q) := \int_{B_1(0)} \prod_{k=1}^q |r_k|^{a_k} \,\mathrm{d}r_k$, then the volume is equal to

$$|B_\delta^{(l)}| = 2^q I(2/l_1 - 1, \ldots, 2/l_q - 1) \prod_{k=1}^q \delta_k/l_k.$$

By recursion [Wang, 2005], we can show that

$$I(a_1,\ldots,a_q) = \frac{\prod_{k=1}^q \Gamma(b_k)}{\Gamma\left(1+\sum_{k=1}^q b_k\right)},$$

where $b_k := (a_k+1)/2$. Therefore, the volume is equal to

$$|B_\delta^{(l)}| = 2^q \frac{\prod_{k=1}^q \Gamma(1/l_k)}{\Gamma(1+\sum_{k=1}^q 1/l_k)} \prod_{k=1}^q \delta_k/l_k = 2^q \frac{\prod_{k=1}^q \Gamma(1+1/l_k)}{\Gamma(1+\sum_{k=1}^q 1/l_k)} \prod_{k=1}^q \delta_k. \quad (3.10)$$

For the term $\int_{B_\delta^{(l)}(0)} |u_k|^2 \,\mathrm{d}u$, we can again change variables to see that

$$\begin{aligned}\int_{B_\delta^{(l)}(0)} |u_k|^2 \,\mathrm{d}u &= 2^q \left(\prod_{j=1}^q l_j^{-1}\right) \int_{B_1(0)} |r_k|^{4/l_i} \prod_{j=1}^q |r_j|^{2/l_j - 1} \,\mathrm{d}r \prod_{j=1}^q \delta_j \,\delta_k^2 \\ &= 2^q \frac{\prod_{j=1}^q \delta_j}{\prod_{j=1}^q l_j} I\left(\frac{2+4[k=1]}{l_1} - 1, \ldots, \frac{2+4[k=q]}{l_q} - 1\right) \delta_k^2.\end{aligned}$$

We again use the explicit form of $I(a_1, \ldots, a_q)$ to see that

$$\begin{aligned}\int_{B_\delta^{(l)}(0)} |u_k|^2 \, \mathrm{d}u &= 2^q \left(\prod_{j=1}^q l_j^{-1} \right) \frac{\prod_{j\neq k}^q \Gamma\left(1/l_j\right) \Gamma\left(3/l_k\right)}{\Gamma\left(1 + \sum_{j=1}^q 1/l_j + 2/l_k\right)} \prod_{j=1}^q \delta_j \, \delta_k^2 \\ &= \frac{2^q}{3} \frac{\prod_{j\neq k} \Gamma\left(1 + 1/l_j\right) \Gamma(1 + 3/l_k)}{\Gamma\left(1 + \sum_{j=1}^q 1/l_j + 2/l_k\right)} \prod_{j=1}^q \delta_j \, \delta_k^2 \qquad (3.11) \\ &= \frac{1}{3} |B_\delta^{(l)}| \frac{\Gamma(1 + 3/l_k)}{\Gamma(1 + 1/l_k)} \frac{\Gamma\left(1 + \sum_{j=1}^q 1/l_j\right)}{\Gamma\left(1 + \sum_{j=1}^q 1/l_j + 2/l_k\right)} \, \delta_k^2,\end{aligned}$$

as required.

Now, suppose that $l_k = \infty$ for some k. Then the generalised ball is the set

$$B_\delta^{(l)}(0) = \left\{ u : \left(|u_k| = \delta \cap \sum_{j\neq k} |u_j| = 0 \right) \cup \left(|u_k| < \delta_k \cap \sum_{k\neq j} |u_j/\delta_j|^{l_j} \leq 1 \right) \right\}.$$

Therefore, the volume is equal to

$$|B_\delta^{(l)}| = \int_{-\delta_k}^{\delta_k} \int_{B_{\delta_{-k}}^{(l_{-k})}} \mathrm{d}u_{-k} \, \mathrm{d}u_k = 2 \, |B_{\delta_{-k}}^{(l_{-k})}| \, \delta_k. \qquad (3.12)$$

If $l_k = \infty$ for all k, the volume is equal to $|B_\delta^{(l)}| = 2^q \prod_{k=1}^q \delta_k$. Similarly, the integral $\int_{B_\delta^{(l)}(0)} uu^T \, \mathrm{d}u$ has diagonal elements

$$\int_{B_\delta^{(l)}(0)} |u_k|^2 \, \mathrm{d}u = \int_{-\delta_k}^{\delta_k} |u_k|^2 \, \mathrm{d}u_k \, |B_{\delta_{-k}}^{(l_{-k})}| = \frac{2}{3} \, |B_{\delta_{-k}}^{(l_{-k})}| \, \delta_k^3, \qquad (3.13)$$

and

$$\int_{B_\delta^{(l)}(0)} |u_j|^2 \, \mathrm{d}u = \int_{-\delta_k}^{\delta_k} \int_{B_{\delta_{-k}}^{(l_{-k})}(0)} |u_j|^2 \, \mathrm{d}u_{-k} \, \mathrm{d}u_k = 2 \int_{B_{\delta_{-k}}^{(l_{-k})}(0)} |u_j|^2 \, \mathrm{d}u_{-k} \, \delta_k. \qquad (3.14)$$

for all $j \neq k$. If $l_k = \infty$ for all k, the integral has diagonal elements

$$\int_{B_\delta^{(l)}(0)} |u_k|^2 \, \mathrm{d}u = \frac{2^q}{3} \prod_{j=1}^q \delta_j \, \delta_k^2. \qquad (3.15)$$

Comparing Equations (3.12)–(3.15) to the results for finite l in Equations 3.10 and 3.11, the results for l with infinite-valued elements can be treated as the limit of the results for finite l. Therefore, we have the result for $0 < l_k \leq \infty$. □

Lemma 3.17. *Let the function $h\colon \mathbb{R}^p \to \mathbb{R}$ be bounded, $m(s)$ have continuous third derivatives with respect to s, and $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta^{(l)}(s)$ for all possible θ and s. Then, for almost all*

s^*, there is a constant vector c_k, such that the ABCBAS *estimate Z_N that accepts samples (θ_k, s_k) whose statistic sample s_k is in the generalised ball $B_\delta^{(l)}(s^*)$ satisfies*

$$\mathrm{bias}(Z_N) = \sum_{k=1}^{q} c_k(s^*)\delta_k^2 + \mathcal{O}\left(\|\delta\|^4\right),$$

as $\delta \downarrow 0$.

Proof. Let $\gamma_k(l) := \frac{\Gamma(1+3/l_k)}{\Gamma(1+1/l_k)} \frac{\Gamma(1+\sum_{j=1}^{q} 1/l_j)}{\Gamma(1+\sum_{j=1}^{q} 1/l_j+2/l_k)}$. Then, by Lemma 3.16,

$$\begin{aligned}\phi_h^{(\delta)}(s^*) &= \phi_h(s^*) + \frac{1}{2|B_\delta^{(l)}|}\mathrm{Tr}\left(\int_{B_\delta^{(l)}} uu^T \,\mathrm{d}u \mathcal{H}_{\phi_h}(s^*)\right) + \mathcal{O}\left(\sum_{k=1}^{q}\delta_k^4\right)\\ &= \phi_h(s^*) + \frac{1}{6}\sum_{k=1}^{q}\gamma_k(l)\frac{\partial^2}{\partial s_k^2}\phi_h(s^*)\delta_k^2 + \mathcal{O}\left(\|\delta\|^4\right),\end{aligned}$$

as $\delta \downarrow 0$. Substituting this into the expression for the bias gives

$$\begin{aligned}\mathrm{bias}(Z_N) &= \frac{\phi_h^{(\delta)}(s^*)}{\phi_1^{(\delta)}(s^*)} - \frac{\phi_h(s^*)}{\phi_1(s^*)}\\ &= \frac{\phi_h^{(\delta)}(s^*)\phi_1(s^*) - \phi_h(s^*)\phi_1^{(\delta)}(s^*)}{\phi_1(s^*)\phi_1^{(\delta)}(s^*)}\\ &= \frac{\frac{1}{6}\sum_{k=1}^{q}\gamma_k(l)\left(\frac{\partial^2}{\partial s_k^2}\phi_h(s^*)\phi_1(s^*) - \phi_h(s^*)\frac{\partial^2}{\partial s_k^2}\phi_1(s^*)\right)\delta_k^2 + \mathcal{O}\left(\|\delta\|^4\right)}{\phi_1(s^*)\left(\phi_1(s^*) + \mathcal{O}\left(\|\delta\|^2\right)\right)}\\ &= \frac{\sum_{k=1}^{q}\gamma_k(l)\left(\frac{\partial^2}{\partial s_k^2}\phi_h(s^*)\phi_1(s^*) - \phi_h(s^*)\frac{\partial^2}{\partial s_k^2}\phi_1(s^*)\right)\delta_k^2}{6\phi_1(s^*)^2} + \mathcal{O}\left(\|\delta\|^4\right)\\ &= \frac{\sum_{k=1}^{q}\gamma_k(l)\left(\frac{\partial^2}{\partial s_k^2}\left(m(s^*)f_S(s^*)\right) - m(s^*)\frac{\partial^2}{\partial s_k^2}f_S(s^*)\right)\delta_k^2}{6f_S(s^*)} + \mathcal{O}\left(\|\delta\|^4\right)\\ &= \frac{1}{6}\sum_{k=1}^{q}\gamma_k(l)\left(\frac{\partial^2}{\partial s_k^2}m(s^*) + 2\frac{\partial}{\partial s_k}m(s^*)\frac{\partial}{\partial s_k}\log\left(f_S(s^*)\right)\right)\delta_k^2 + \mathcal{O}\left(\|\delta\|^4\right),\end{aligned}$$

as $\delta \downarrow 0$. This is the required result, with bias constant

$$c_k(s^*) = \frac{1}{6}\gamma_k(l)\left(\frac{\partial^2}{\partial s_k^2}m(s^*) + 2\frac{\partial}{\partial s_k}m(s^*)\frac{\partial}{\partial s_k}\log\left(f_S(s^*)\right)\right). \qquad \square$$

Example 3.18. *We can consider the case where the elements of δ are identical, and the elements of l are also identical. The resulting acceptance region is an l-ball. In this case, if we instead write δ and l for the equivalent scalars, the asymptotic bias is equal to*

$$\begin{aligned}\mathrm{bias}(Z_N) =& \frac{1}{6}\gamma_k(l)\left(\triangle m(s^*) + 2\nabla m(s^*)\nabla\log\left(f_S(s^*)\right)^T\right)\delta^2 + \mathcal{O}\left(\delta^4\right)\\ =& \frac{1}{6}\frac{\Gamma(1+3/l)}{\Gamma(1+1/l)}\frac{\Gamma(1+q/l)}{\Gamma(1+(q+2)/l)}\\ &\times\left(\triangle m(s^*) + 2\nabla m(s^*)\nabla\log\left(f_S(s^*)\right)^T\right)\delta^2 + \mathcal{O}\left(\delta^4\right),\end{aligned}$$

as $\delta \downarrow 0$. For the case where $l = 2$, this is equal to

$$\mathrm{bias}(Z_N) = \frac{\triangle m(s^*) + 2\nabla m(s^*)\nabla \log\left(f_S(s^*)\right)^T}{2(q+2)}\,\delta^2 + \mathcal{O}\left(\delta^4\right).$$

This is the case where we accept on a 2-ball, and agrees with the result in Theorem 3.4. Two other simple choices are $l = 1$, which gives

$$\mathrm{bias}(Z_N) = \frac{\triangle m(s^*) + 2\nabla m(s^*)\nabla \log\left(f_S(s^*)\right)^T}{(q+1)(q+2)}\,\delta^2 + \mathcal{O}\left(\delta^4\right),$$

and $l = \infty$, which gives

$$\mathrm{bias}(Z_N) = \frac{\triangle m(s^*) + 2\nabla m(s^*)\nabla \log\left(f_S(s^*)\right)^T}{6}\,\delta^2 + \mathcal{O}\left(\delta^4\right).$$

The latter, where the acceptance region is a hypercube centred on s^*, has a bias that is especially sensitive to the statistic dimension q, since increasing q does not increase the denominator in the bias constant.

The bias for other l-balls, where l is an integer, is more difficult to evaluate, because the gamma function ratios do not simplify easily. However, any version of the special case where $l = 2/m$, and m is an integer, has a bias easily expressible in terms of falling powers. Specifically, the resulting bias is equal to

$$\begin{aligned}\mathrm{bias}(Z_N) =& \frac{1}{6}\frac{\Gamma(1+3m/2)}{\Gamma(1+m/2)}\frac{\Gamma(1+mq/2)}{\Gamma(1+m(q+2)/2)} \\ &\times \left(\triangle m(s^*) + 2\nabla m(s^*)\nabla \log\left(f_S(s^*)\right)^T\right)\delta^2 + \mathcal{O}\left(\delta^4\right) \\ =& \frac{(3m/2)^{\underline{m}}}{6(1+mq/2)}\left(\triangle m(s^*) + 2\nabla m(s^*)\nabla \log\left(f_S(s^*)\right)^T\right)\delta^2 + \mathcal{O}\left(\delta^4\right),\end{aligned}$$

as $\delta \downarrow 0$, where $u^{\underline{m}} := \prod_{k=0}^{m-1}(u-k)$ is the m^{th} falling power of u. This case includes the astroid mentioned in Section 2.1.3, which has $q = 2$ and $m = 3$.

There are some other simple acceptance regions that have a simple bias expression. For example, the estimate with a cylindrical acceptance region has $l = (2, 2, \infty)$. If the elliptic radii are all equal to δ, the asymptotic bias is equal to

$$\begin{aligned}\mathrm{bias}(Z_N) =& \frac{1}{8}\sum_{k=1}^{2}\left(\frac{\partial^2}{\partial s_k^2}m(s^*) + 2\frac{\partial}{\partial s_k}m(s^*)\frac{\partial}{\partial s_k}\log\left(f_S(s^*)\right)^T\right)\delta^2 \\ &+ \frac{1}{6}\left(\frac{\partial^2}{\partial s_3^2}m(s^*) + 2\frac{\partial}{\partial s_3}m(s^*)\frac{\partial}{\partial s_3}\log\left(f_S(s^*)\right)^T\right)\delta^2 + \mathcal{O}\left(\delta^4\right).\end{aligned}$$

For the asymptotic variance, we can use the result from Lemma 3.11, where the success probability is now equal to

$$p(s^*) = \int_{B_\delta^{(l)}(s^*)} f_S(s)\,\mathrm{d}s = |B^{(l)}_{(1,\ldots,1)}|\prod_{k=1}^{q}\delta_k\left(1 + \mathcal{O}\left(\|\delta\|^2\right)\right)$$

as $\delta \downarrow 0$. Therefore, the generalised form of the MSE in Equation (3.8) is

$$\text{MSE}(Z_N) = \left(\frac{v(s^*)}{N \prod_{k=1}^{q} \delta_k} + \sum_{k=1}^{q} c_k(s^*) \delta_k^2 \right) (1 + \mathrm{o}(1)) \tag{3.16}$$

as $N \prod_{k=1}^{q} \delta_k \uparrow \infty$. The expected computational cost is linear with respect to N, as before.

Optimising the rate of convergence for Z_N, using Equation (3.16), now involves optimising an expression that includes both $\prod_{k=1}^{q} \delta_k$ and $\sum_{k=1}^{q} \delta_k$. This optimisation problem is similar to that in Section 3.4.2, where we have a square-tolerance matrix H, and wish to optimise an expression that includes both $|H|^{1/2}$ and $\text{Tr}(H)$. For the same reasons as in Section 3.4.2, the optimisation problem is not trivial to solve for general sequences of values for the tolerance vector δ.

3.4.5 Discrete Data

If the data is discrete, or a mixture of discrete and continuous, then we cannot use Taylor expansions. One possible alternative, if the data space is one-dimensional, is to use the discrete equivalent to the Taylor series, which is the Newton series. This expresses a function in terms of difference operators, and is outlined in Lemma A.7. However, this leaves the issue of summing over the points inside the acceptance region. Specifically, we can consider the discrete equivalent of the ϕ functions,

$$\tilde{\phi}_h(s) := m(s)\mathbb{P}\left(S = s\right), \quad \tilde{\phi}_h^{(\delta)}(s^*) := \frac{1}{\#\{s : s \in B_\delta(s^*)\}} \sum_{s \in B_\delta(s^*)} \tilde{\phi}_h(s).$$

If the possible values of s are on a uniform grid, then the acceptance region will, again, be symmetric, and we have, for $k > 0$, the expansions

$$\tilde{\phi}_h(s^* + k) = \sum_{i=0}^{k} \binom{k}{i} \Delta^i \tilde{\phi}_h(s^*)$$

and

$$\tilde{\phi}_h(s^* - k) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \nabla^i \tilde{\phi}_h(s^*),$$

so that

$$\tilde{\phi}_h(s^* + k) + \tilde{\phi}_h(s^* - k) = \sum_{i=0}^{k} \binom{k}{i} \left(\Delta^i + (-1)^i \nabla^i\right) \tilde{\phi}_h(s^*).$$

While this is not useful in the general case, it can be useful if values for $\tilde{\phi}_h(\cdot)$ and $\tilde{\phi}_1(\cdot)$ follow some known difference equation.

In practice, if the data space is discrete, then we will often be in either a very small space – in which case we can set δ to zero without much loss of efficiency – or a very large space, in which case the space can be approximated as a continuous one.

3.5 Discussion and Comparison to Previous Results

In Sections 3.2 and 3.4.1, we showed that the basic ABC estimates have an asymptotic convergence rate either of order $\mathcal{O}\left(n^{-4/(q+4)}\right)$ or of $\mathcal{O}\left(N^{-4/(q+4)}\right)$, depending on whether n or N is fixed.

We can compare this to the convergence rate of exact, unbiased Monte Carlo methods, which is $\mathcal{O}\left(N^{-1}\right)$. For one-dimensional statistics, where $q = 1$, ABC has asymptotic rate $\mathcal{O}\left(N^{-4/5}\right)$, already a noticeable reduction in efficiency. If the statistic dimension is increased to $q = 4$, then the rate is reduced to $\mathcal{O}\left(N^{-1/2}\right)$ at best, already half the rate of convergence of exact Monte Carlo methods. The rate of convergence therefore decreases rapidly as q is increased, so much effort is put into dimension reduction when choosing summary statistics.

In Section 3.4.2, we considered the effect on the asymptotic convergence rate from using random acceptance: the asymptotic rate is the same, but the choice of kernel affects the magnitude of the leading term. However, Example 3.18 demonstrates that the uniform kernel is likely to minimise the leading term, so we gain no advantage from using random acceptance, if we are only interested in estimating the posterior expectation. In practice, there are still reasons to use random acceptance, even when estimating the posterior expectation. These reasons include making use of rejection and importance sampling on the prior, as mentioned in Section 2.2.1.

The asymptotic rates of convergence found in Sections 3.2 and 3.4.1 are similar to the rates in Section 2.2.4 for other ABC variants, to which we now compare them. Most of the variants mentioned in the following fix the total number N of proposals, so we will compare them to the rate for the ABCBAS estimate Z_N. Since all the variants that will be discussed have an expected cost

C that is linear with respect to N, the comparisons given below in terms of N are equivalent to comparisons in terms of C.

It should be remembered that these convergence rates are asymptotic. While they can be used to compare long-term performance between different algorithms, a variant with slower asymptotic convergence can be more accurate for practical computational running times. In particular, many variants are designed to be used on specific types of models, so will perform better in those cases.

Prangle [2011] found the same asymptotic rate of $\mathcal{O}\left(N^{-q/q+4}\right)$, in the case where the summary statistic is $S(x) = m(x)$, the true conditional expectation. Although this is a similar result, there are two differences to note:

1. This summary statistic is not sufficient, since it is not necessarily true that $f_{\theta|S}(t \mid S(x)) = f_{\theta|X}(t \mid x)$, for all x, so Prangle's results are in a different setting to the one in this text. However, this summary statistic is sufficient in the more limited sense that $\mathbb{E}\left(\theta \mid S(x)\right) = \mathbb{E}\left(\theta \mid x\right)$, which is sufficient for estimating the posterior mean.

2. The parameter θ is not reduced by some one-dimensional function h. Instead, the estimate is for $\mathbb{E}\left(\theta \mid s^*\right)$, and the error is minimised with respect to the expectation of the quadratic loss function

$$L(\theta, Z; A) := (Z - \theta)^T A (Z - \theta),$$

   where Z is the estimate, and A is a positive-definite matrix of full rank. This has posterior expectation

$$\begin{aligned}\mathbb{E}\left(L(\theta, Z; A) \mid s^*\right) = &\operatorname{Tr}(A \operatorname{Cov}(Z \mid s^*)) + \operatorname{Tr}(A \operatorname{Cov}(\theta \mid s^*)) \\ &+ \operatorname{bias}(Z \mid s^*)^T A \operatorname{bias}(Z \mid s^*).\end{aligned}$$

   Although similar to the mean square error, this can not, in general, be rewritten as minimising the mean square error with respect to some one-dimensional function h, so these results also have a different goal for optimisation.

Blum [2010] found an asymptotic rate of $\mathcal{O}\left(N^{-q/(q+5)}\right)$ when using ABC for kernel density estimation on a one-dimensional parameter θ. This is a more general estimate, at the cost of a slightly slower asymptotic rate. Note that this is the asymptotic rate for pointwise convergence, rather than uniform convergence. More specifically, for tolerance δ and density estimate bandwidth b, the bias is shown to be $\mathcal{O}\left(\delta^2 + b^2\right)$, and the variance is shown to be $\mathcal{O}\left(\frac{1}{N\delta^q b}\right)$. The asymptotic bias is roughly equivalent to that of the basic ABC estimates, but the asymptotic variance is affected by the addition of the bandwidth constant. In effect, the bandwidth constant has the same effect as increasing the statistic dimension q by one.

The asymptotic rate in Blum [2010] was found to be the same when using either local-linear regression or local-quadratic regression to adjust proposals. However, these methods of proposal adjustment do affect the bias coefficient. Which method has the smaller coefficient, and thus a smaller asymptotic upper bound on the bias, depends on the shape of the bandwidth matrix, and on the behaviour of the function $m(s) = \mathbb{E}\left(\theta \mid s\right)$ in the neighbourhood of s. For example, using local-linear or local-quadratic adjustment is better if m is linear, because their bias coefficients are zero, and so their asymptotic bias is higher-order. Furthermore, if m is non-linear, but the distribution of the residual $\theta - m(s)$ is independent of s, then the bias coefficient for the variant with local-quadratic adjustment is still zero. This leads Blum to consider transformations of the summary statistics that reduce the dependence of the residual on s. We will return to the results from this paper in the discussion section of Chapter 4.

Biau et al. [2015] also considered using ABC for kernel density estimation, but let the parameter θ be multi-dimensional. Additionally, they accepted samples using a nearest-neighbours approach, rather than a fixed acceptance region. For direct comparison to other results, we can consider the case where the parameter dimension is $p = 1$. The estimate has three different asymptotic rates of convergence, depending on the statistic dimension q. At best, when $q < 4$, the estimate convergence rate is $\mathcal{O}\left(N^{-4/9}\right)$, a much slower rate than any other variant we consider. At best, when $q > 4$, the convergence rate is the same as that in Blum [2010]. However, this rate is not directly comparable to the others, since Biau et al. define the error as the mean integrated square error $\mathbb{E}\left(\int (Z(t) - f_{\theta|S}(t \mid s^*)^2 \,\mathrm{d}t \mid t\right)$, where $Z(t)$ is the ABC estimate of $f_{\theta|S}(t \mid s^*)$, rather than the pointwise mean square error considered in Blum [2010] and elsewhere.

Finally, Fearnhead and Prangle [2012] found that use of Noisy ABC results in an asymptotic rate of $\mathcal{O}\left(N^{-2/(q+2)}\right)$, because the asymptotic bias is now $\mathcal{O}(\delta)$, rather than $\mathcal{O}\left(\delta^2\right)$. This is slower than the basic variant, about the equivalent of doubling the statistic dimension. Again, although this is a similar result to that in Section 3.4.1, there are two differences to consider:

1. As in Prangle [2011], the algorithm estimates the posterior expectation of parameter θ, not of some one-dimensional function of θ, and minimises error with respect to a quadratic loss function.

2. Noisy ABC is designed to be calibrated: specifically, if the ABC posterior assigns a probability p to being in a certain region, then proposals from the prior will be within this region with the same probability. While useful, in the sense that it guarantees Noisy ABC will converge to the correct value of θ, this calibration is done by adding noise to the original observations before using them. This reduces the information available from the observations, so Noisy ABC gives less accurate estimates then basic ABC for small δ. Fearnhead and Prangle therefore suggest using Noisy ABC when the statistic dimension, and hence δ, are very large, particularly when combining ABC analyses of large datasets.

Chapter 4

ABC with Local-Linear Regression

The previous chapter looked at the asymptotic error for the basic ABC estimates Y_n and Z_N. We now introduce, and analyse, a new estimate, that we will refer to as the ABCLOC estimate $\hat{Z}_N$. The algorithm includes both an accept-reject step and a proposal adjustment step, and involves two kernels: one is used for random acceptance, as described in Section 2.1.2, and the other is used to weight samples in the linear regression used for proposal adjustment, as described in Section 2.2.3.

4.1 Algorithm

Instead of using the accepted ABC samples directly in the estimate, we can first adjust them, to account for the difference between their generated statistic and the observed statistic. These adjustments are commonly made using simple or local polynomial regression (Beaumont et al. [2002], Fearnhead and Prangle [2012]). We first describe the general approach, look at the version proposed by Beaumont et al. [2002], and then describe the version we propose in this chapter.

4.1.1 General Proposal Adjustment

To do proposal adjustment we begin by defining an estimate $\hat{m}(s; \hat{H})$ for the conditional expectation function $m(s)$, as defined in Equation (2.1), where $\hat{H}$

denotes the estimate parameters. Then, as in Section 2.1.2, we can express the proposals θ_k as $\theta_k = \hat{m}(s_k; \hat{H}) + \epsilon_k$, so that ϵ_k is the empirical residual. We then use the adjusted proposals $\hat{\theta}_k := \hat{m}(s^*; \hat{H}) + \epsilon_k$, which will roughly be sampled from the true conditional distribution. Substituting in the value of the residuals gives

$$\hat{\theta}_k = \theta_k - \hat{m}(s_k; \hat{H}) + \hat{m}(s^*; \hat{H}).$$

Using these adjusted proposals results in the estimate

$$\hat{Z}_N := \frac{1}{n} \sum_{j=1}^{n} h(\hat{\theta}_{k_j}) = \frac{1}{n} \sum_{j=1}^{n} h(\theta_{k_j} - \hat{m}(s_{k_j}; \hat{H}) + \hat{m}(s^*; \hat{H})),$$

where θ_{k_j} are the n proposals that are accepted. In this chapter, we will consider the special case $h(x) = x$, where the above equation simplifies to

$$\hat{Z}_N = Z_N + A_N, \quad A_N := \hat{m}(s^*; \hat{H}) - \frac{1}{n} \sum_{j=1}^{n} \hat{m}(s_{k_j}; \hat{H}),$$

so that A_N describes the adjustment to the original estimate.

4.1.2 Local-Linear Adjustment

Beaumont et al. [2002] defined $\hat{m}(s; \hat{H})$ to be the value at s of a local-linear regression centred at s^*. Specifically, for some kernel function $\hat{K}$, and some square-tolerance matrix $\hat{H}$, both as defined in Definition 2.7, they found the coefficients $\hat{\alpha}$ and $\hat{\beta}$ such that

$$\begin{pmatrix} \hat{\alpha} \\ \beta \end{pmatrix} := \operatorname*{argmin}_{\alpha, \beta} \sum_{k=1}^{N} (\theta_k - \alpha - \beta^{(} s_k - s^*))^2 \hat{K}_{\hat{H}}(s_k - s^*).$$

They then let $\hat{m}(s; \hat{H}) = \hat{\alpha} + \hat{\beta}^T(s_k - s^*)$, resuliting in adjusted proposals

$$\tilde{\theta}_k = \theta_k - \hat{\beta}^T(s_k - s^*).$$

An example of the adjustment is given in Figure 4.1.1.

The kernel function $\hat{K}$ and square-bandwidth matrix $\hat{H}$ were also used to weight the adjusted proposals in the ABC estimate, as in case 3 in Section 2.1.2. The full algorithm is given in Algorithm 4.1.1.

4.1.3 Multiple Regression Adjustment

In this chapter, for the proposed ABCLOC estimate, we define $\hat{m}(s; \hat{H})$ to be the value at s of a kernel linear regression centred at s. Note this now uses a separate regression for s^*, and for each accepted sample.

Local-linear ABC estimate from Beaumont et al. [2002]

Input is data $x^* \in \mathbb{R}^d$, summary function $s\colon \mathbb{R}^d \to \mathbb{R}^q$, prior density f_θ for $\theta \in \mathbb{R}^p$, conditional statistic density $f_{S|\theta}$, weight kernel function $\hat{K}$, square-bandwidth matrix $\hat{H}$, and a required number N of proposals. Let $s^* = S(x^*)$.

1. Set $k = 1$.
2. Generate proposal $\theta_k \sim f_\theta$ and data $X_k \sim f_{X|\theta}(\cdot \mid \theta_k)$.
3. Let $s_k = S(X_k)$, calculate weight $w_k = \hat{K}_{\hat{H}}(s_k - s^*)$.
4. If $k < N$, increase k by one, and return to Step 2.
5. Form a local-linear regression with predictors s_k, responses θ_k, and weights w_k, for $k \in \{1, \dots, N\}$, to get estimate regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ in the expression
$$\theta_k = \hat{\alpha} + (s_k - s^*)^T \hat{\beta} + \epsilon_k.$$
6. Calculate adjusted proposals $\hat{\theta}_k = \theta_k - (s_k - s^*)^T \hat{\beta}$.

Output is estimate $Z_N = \sum_{k=1}^N w_k h(\hat{\theta}_k) / \sum_{k=1}^N w_k$, the weighted mean of the function values for the adjusted proposals.

Algorithm 4.1.1: Algorithm for the estimate used in Beaumont et al. [2002].

Specifically, we choose two kernel functions, K and $\hat{K}$, and two square-bandwidth matrices, H and $\hat{H}$. We will use K and H for random acceptance, as described in Section 3.4.2, and $\hat{K}$ and $\hat{H}$ to adjust the accepted proposals. In practice, we will set $K = \hat{K}$ and $H = \hat{H}$, but for now we let them be different.

We then do a local-linear regression centred on s^*, as in Section 4.1.2. However, we are only interested in the estimated conditional mean at the centre point s^*, which is equal to $\hat{m}(s^*; \hat{H}) = \alpha$. By Lemma A.8, this is equal to

$$\hat{m}(s^*; \hat{H}) = e_1^T \left(X_{s^*}^T \hat{W}_{s^*} X_{s^*} \right)^{-1} X_{s^*}^T \hat{W}_{s^*} \Theta,$$

where

$$X_s = \begin{pmatrix} 1 & (s_1 - s)^T \\ \vdots & \vdots \\ 1 & (s_N - s)^T \end{pmatrix}, \quad \hat{W}_s = \operatorname{diag} \left(\hat{K}_{\hat{H}}(s_1 - s), \ldots, \hat{K}_{\hat{H}}(s_N - s) \right),$$

and Θ is the vector of proposals.

It now remains to determine the estimated conditional mean $\hat{m}(s_{k_j}; \hat{H})$ for each of the accepted statistic samples. Therefore, for each of the accepted statistics s_{k_j}, we will do an additional local-linear regression, centred at s_{k_j} rather than s^*. By Lemma A.8, the resulting conditional mean estimates are equal to

$$\hat{m}(s; \hat{H}) = e_1^T \left(X_s^T \hat{W}_s X_s \right)^{-1} X_s^T \hat{W}_s \Theta, \tag{4.1}$$

where $s = s_{k_j}$.

If we accept n proposals, then we do a total of $n + 1$ separate local-linear regressions: one centred on each accepted proposal, and one centred on s^*. The full algorithm for the ABCLOC estimate is given in Algorithm 4.1.2. An example of the proposal adjustment is given in Figure 4.1.2.

Before we proceed to the asymptotic results, we make some remarks:

1. The ABCLOC estimate uses a fixed number N of proposals, rather than a fixed number n of accepted proposals. The latter would require us to consider two sets of conditional statistic densities: one for accepted proposals, and one for rejected proposals.

2. The number of accepted samples can be zero. In this case, we define $\hat{Z}_N$ to be equal to some predetermined value c, as we do for Z_N. The adjustment term A_N is zero in this case.

ABCLOC estimate $\hat{Z}_N$

Input is data $x^* \in \mathbb{R}^d$, summary function $s\colon \mathbb{R}^d \to \mathbb{R}^q$, prior density f_θ for $\theta \in \mathbb{R}^p$, conditional statistic density $f_{S|\theta}$, acceptance kernel function K_H, weight kernel function $\hat{K}$, square-bandwidth matrix $\hat{H}$, and a required number N of proposals. Let $s^* = S(x^*)$.

1. Set $k = 1$.
2. Generate proposal $\theta_k \sim f_\theta$ and data $X_k \sim f_{X|\theta}(\cdot \mid \theta_k)$.
3. Let $s_k = S(X_k)$, and calculate weight $w_k = \hat{K}_{\hat{H}}(s_k - s^*)$. Accept θ_k with probability $K_H(s_k - s^*)/\max_u K_H(u)$.
4. If $k < N$, increase k by one, and return to Step 2.
5. Form a local-linear regression with inputs s_k, responses θ_k, and weights w_k, for $k \in \{1, \ldots, N\}$, to get estimated regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ in the expression
$$\theta_k = \hat{\alpha} + (s_k - s^*)^T \hat{\beta} + \epsilon_k.$$
6. For each accepted statistic s_{k_j}, form a local-linear regression j with inputs s_k, responses θ_k, and weights $w_{k,j} = \hat{K}_{\hat{H}}(s_k - s_{k_j})$, for all $k \in \{1, \ldots, N\}$, to get estimated regression coefficients $\hat{\alpha}_j$ and $\hat{\beta}_j$ in the expression
$$\theta_k = \hat{\alpha}_j + (s_k - s_{k_j})^T \hat{\beta}_j + \epsilon_{k,j}.$$
7. Calculate adjusted proposals $\hat{\theta}_{k_j} = \theta_{k_j} - \hat{\alpha}_j + \hat{\alpha}$.

Output is estimate $Z_N = \frac{1}{n}\sum_{j=1}^n h(\hat{\theta}_{k_j})$, the mean of the function values for the accepted, adjusted proposals.

Algorithm 4.1.2: Algorithm for the ABCLOC estimate, with an accept-reject step and a proposal adjustment step.

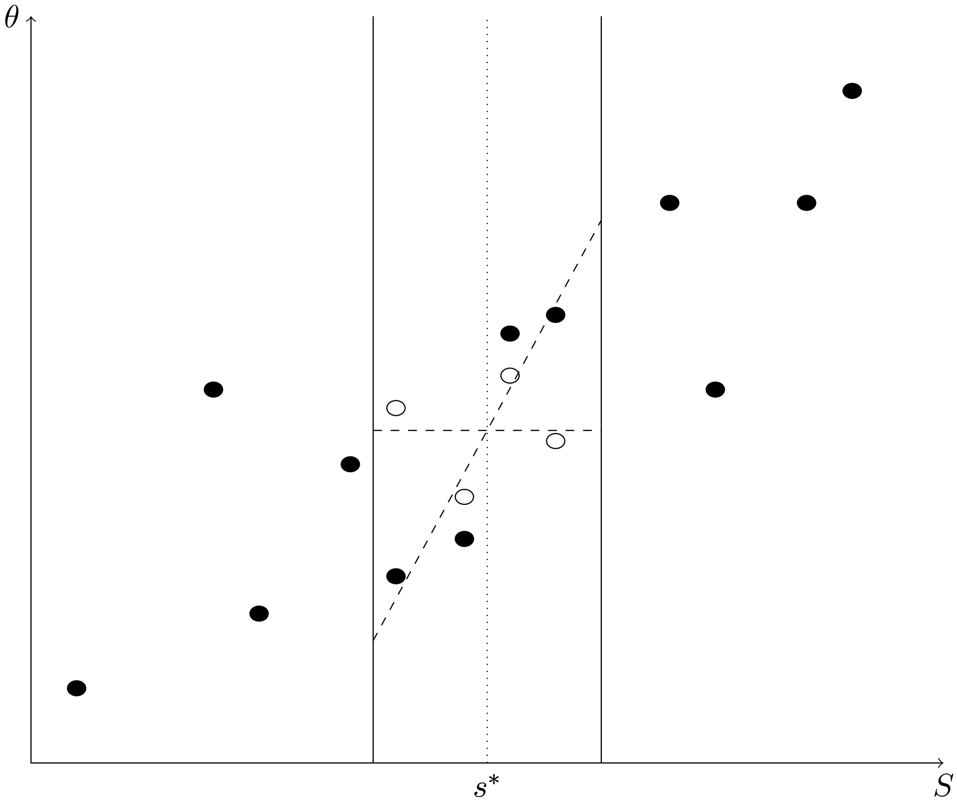

Figure 4.1.1: Example of the proposal adjustments in the estimate used in Beaumont et al. [2002]. The larger black dots are the original samples. For the local-linear regression, we use the kernel function $\hat{K}(u) = [u \in [-1, 1]] /2$, and choose $\hat{H}$ so that $\hat{m}(s, \hat{H})$ is the value at s of a linear regression, which uses the samples in the interval marked by the two solid lines. The adjustment then consists of subtracting the values predicted from the regression, shown as the sloped dashed line, and adding the predicted value at s^*, shown as the horizontal dashed line. The resulting adjusted samples are shown as white dots.

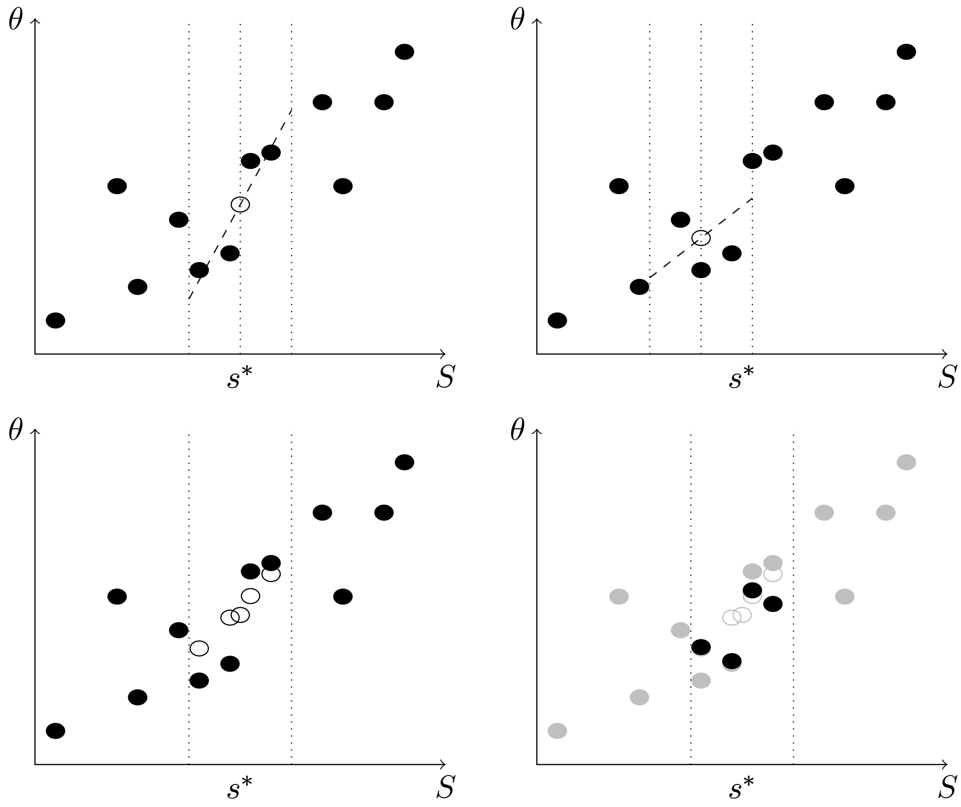

Figure 4.1.2: Example of the proposal adjustments in the ABCLOC estimate, using kernel functions $K(u) = \hat{K}(u) = [u \in [-1, 1]]/2$, and square-bandwidth matrices $H = \hat{H}$.

Top-left: Local-linear regression centred on s^*. The samples are shown as black dots. The observation s^*, and the boundaries of the region in which the regression includes samples, are shown as dotted lines. This region includes four of the samples. The regression line is the dashed line, with the centre point $(s^*, \hat{m}(s^*; \hat{H}))$ marked by the white dot. This is the only part of the regression that will be used.

Top-right: Local-linear regression on one of the accepted samples, s'. The regression region, enclosed by dotted lines, is now centred on s'. This affects which samples are used, so the regression line is different. The point $(s', \hat{m}(s; \hat{H}))$ is marked by the white dot.

Bottom-left: All required regression centre points are shown as white dots. The dotted lines represent the acceptance region. Since $K = \hat{K}$ and $H = \hat{H}$, this is the same as the region used for the regression in the top-left panel.

Bottom-right: Samples and regression centre points are now greyed out. The new black dots represent the adjusted samples. Each proposal is shifted according to the vertical difference between its white dot and that of the observation. For example, the first proposal to the left from s^* has little difference between the respective white dots, so its adjustment is small, compared to that of the others.

3. We must account for the randomness introduced by the acceptance kernel function K. In what follows, we define W_s to be the random diagonal matrix whose k^{th} diagonal element is a Bernoulli variable w_k with success probability $K_H(s_k - s)/\max_u K_H(u)$. Then the basic ABC estimate is written as $Z_N = (1^T W_s 1)^{-1} 1^T W_s \Theta$, and the adjustment term is equal to
$$A_N = \hat{m}(s^*; \hat{H}) - (1^T W_s 1)^{-1} 1^T W_s \begin{pmatrix} \hat{m}(s_1; \hat{H}) \\ \vdots \\ \hat{m}(s_N; \hat{H}) \end{pmatrix}.$$

4.2 Asymptotics

We begin with an asymptotic result for local-linear regression. Our estimate $\hat{m}(s; \hat{H})$ for $m(s)$ was shown by Ruppert and Wand [1994] to have conditional bias
$$\mathbb{E}\left(\hat{m}(s; \hat{H}) - m(s) \,\middle|\, s_1, \ldots, s_N\right) = \frac{1}{2}\mu_2(\hat{K})\operatorname{Tr}(\hat{H}\mathcal{H}_m(s)) + \mathrm{o}_P(\operatorname{Tr}(\hat{H})), \quad (4.2)$$
and conditional variance
$$\operatorname{Var}\left(\hat{m}(s; \hat{H}) \,\middle|\, s_1, \ldots, s_N\right) = \frac{R(\hat{K})}{N|\hat{H}|^{1/2}} \frac{v(s)}{f_S(s)}(1 + \mathrm{o}_P(1)), \quad (4.3)$$
where $\mathcal{H}_m(s)$ is the Hessian matrix of m at the point s, $v(s) = \operatorname{Var}(\theta \mid s)$ is the conditional variance of θ, and $\mu_2(\hat{K})$ and $R(\hat{K})$ are as defined in Definition 2.10. The term $\mathrm{o}_P(\cdot)$ is defined in Definition A.9.

We now give a set of assumptions, based on those for the results in Ruppert and Wand [1994], that will be sufficient for our following results to hold.

Definition 4.1. *The* condition number $\kappa(A)$ *of a matrix A is equal to the ratio between its largest eigenvalue and its smallest eigenvalue.*

Assumption 4.2 (Based on A-1 to A-4 in Ruppert and Wand [1994])**.** *All of the following conditions hold:*

1. *The kernel function $\hat{K}$ is compactly-supported and bounded, and is such that $\mu_2(\hat{K})$ is non-zero, and such that all odd-order moments of $\hat{K}$ vanish.*

2. *All regression centre points s are in* $\operatorname{supp}(f_S)$*. For all such s, $v(s)$ is strictly positive and continuous, f_S is continuously differentiable, all of the fourth-order moments of m are continuous, and there is a convex subset of* $\operatorname{supp}(f_S)$ *around s with non-null interior.*

3. *The sequence of square-bandwidth matrices $\hat{H}$ is such that $N^{-1}|\hat{H}|$, and all elements of $\hat{H}$, tend to zero as $N \uparrow \infty$, and $\hat{H}$ remains symmetric and positive definite. Addtionally, $\kappa(\hat{H})$ is bounded by some fixed constant L for all N.*

Equations 4.2 and 4.3 also raise the following points:

- The form of the conditional bias and variance given in Equations 4.2 and 4.3 assumes that each regression centre point s is an interior point of f_S, in the sense that there is some M such that $\text{supp}(\hat{K}_{\hat{H}}(\cdot - s)) \subset \text{supp}(f_S)$ for all $N \geq M$. If this is not the case, then $\text{supp}(\hat{K}_{\hat{H}}(\cdot - s)) \cap \text{supp}(f_S)$, the space in which statistic proposals are used for the regression, need not be symmetric. Ruppert and Wand [1994] note that this increases the asymptotic order of the bias of $\hat{m}(s; \hat{H})$, and the smaller size of the usable space increases the asymptotic variance.

- If only the accepted proposals are used for each regression, we use points in the support of the density function for accepted samples. In the simple case where $\text{supp}(K)$ is finite, $\hat{K} = K$, and $\hat{H} = H$, this result in s^* being the only interior point. To prevent this, we allow each regression to also use the rejected samples. In this case, all accepted statistics are interior points if the set of all points that could be used in a regression,

$$\{u : \exists s \in \text{supp}(K_H) \text{ such that } u \in \text{supp}(\hat{K}_{\hat{H}}(\cdot - s))\}, \tag{4.4}$$

is in $\text{supp}(f_S)$. In the simple case where $K = \hat{K}$ and $H = \hat{H}$, this condition simplifies to $\text{supp}(K_{4H}(\cdot - s^*)) \subset \text{supp}(f_S)$.

- Ruppert and Wand [1994] note that the unconditional expectation of the $o_P(1)$ term in Equations 4.2 and 4.3 has an undefined absolute value, because, in the case where no points contribute to the regression, so that $\hat{W}_s = 0$, the authors leave the estimate $\hat{m}(s; \hat{H})$ undefined. However, in the ABCLOC estimate, the regression centred at s_k always includes the sample (θ_k, s_k), and the regression centred at s^* always includes the accepted samples if $\text{supp}(K_H) \subset \text{supp}(\hat{K}_{\hat{H}})$, so this is not an issue if there are any accepted samples. If there are no accepted samples, then $\hat{Z}_N$ is set to a pre-determined value c, and no regression is done.

- The results from Ruppert and Wand [1994] are for when the regression centre point s is a fixed value. In the ABCLOC algorithm, this holds for the regression at s^*, but the other regressions use random centre points in supp $\left(\hat{K}_{\hat{H}}(\cdot - s^*)\right)$. Additionally, each centre point may also be used in other regressions. We must determine the effect this has on the asymptotic behaviour.

The goals of this section are to make a rigorous statement about the asymptotic bias, variance, and expected cost of $\hat{Z}_N$, and to find the new optimal rate for the MSE. Before we begin, we give a brief sketch of why the asymptotic bias for $\hat{Z}_N$ might have a higher order than that for the ABCBAS estimate Z_N. This suggests $\hat{Z}_N$ might have a faster rate of convergence, which motivates asymptotic analysis of $\hat{Z}_N$.

The bias for Z_N can be expressed as

$$\text{bias}(Z_N) = \mathbb{E}\left(m(S) \mid S \,\text{accepted}\right) - m(s^*),$$

and the expectation for the adjustment term can be expressed as

$$\mathbb{E}\left(A_N\right) = \mathbb{E}\left(\hat{m}(s^*; \hat{H})\right) - \mathbb{E}\left(\hat{m}(S; \hat{H}) \,\middle|\, S \,\text{accepted}\right).$$

Since $\hat{Z}_N = Z_N + A_N$, this means that

$$\text{bias}(\hat{Z}_N) = \text{bias}(\hat{m}(s^*; \hat{H})) - \mathbb{E}\left(\text{bias}(\hat{m}(S; \hat{H})) \,\middle|\, S \,\text{accepted}\right).$$

Using Equation 4.2, this is roughly equal to

$$\text{bias}(\hat{Z}_N) \simeq \frac{1}{2}\mu_2(\hat{K})\text{Tr}\left(\hat{H}\left[\mathcal{H}_m(s^*) - \mathbb{E}\left(\mathcal{H}_m(S) \mid S \,\text{accepted}\right)\right]\right).$$

The expression in the square brackets is proportional to the asymptotic bias of the ABCBAS estimate for some variable with conditional expectation $\mathcal{H}_m(\cdot)$. Therefore, the bias is $\mathcal{O}(\text{Tr}(\hat{H})\text{Tr}(H))$ as H and $\hat{H}$ tend to zero element-wise, a clear improvement on the bias of the ABCBAS estimate, which is $\mathcal{O}\left(\text{Tr}(H)\right)$, and of the local regression estimate $\hat{m}(s^*; \hat{H})$,, which is $\mathcal{O}(\text{Tr}(\hat{H}))$. However, the computational cost will rise at a faster rate due to the regressions, and we do not yet know how the variance is affected.

4.2.1 Bias

We suppose that we use kernel function K and square-bandwidth matrix H for acceptance, and $\bar{K}$ and $\bar{H}$ for regression weighting. Then we would like to know bias$(\hat{Z}_N)$. To do this, we write the three components of $\hat{Z}_N$ in matrix notation. In Equation 4.1, we expressed the component for the regression at s^* as

$$\hat{m}(s^*; \hat{H}) = e_1^T \left(X_{s^*}^T \hat{W}_{s^*} X_{s^*} \right)^{-1} X_{s^*}^T \hat{W}_{s^*} \Theta,$$

and, in Section 4.1, we expressed the ABCBAS estimate Z_N as

$$Z_N = \left(1^T \hat{W}_{s^*} 1 \right)^{-1} 1^T \hat{W}_{s^*} \Theta, \tag{4.5}$$

and the remaining term as

$$-(A_N - \hat{m}(s^*; \hat{H})) = \frac{1}{n} \sum_{j=1}^{n} \hat{m}(s_{k_j}) = \left(1^T \hat{W}_{s^*} 1 \right)^{-1} 1^T \hat{W}_{s^*} \begin{pmatrix} \hat{m}(s_1; \hat{H}) \\ \vdots \\ \hat{m}(s_N; \hat{H}) \end{pmatrix},$$

where

$$\hat{m}(s_k; \hat{H}) = e_1^T \left(X_{s_k}^T \hat{W}_{s_k} X_{s_k} \right)^{-1} X_{s_k}^T \hat{W}_{s_k} \Theta.$$

These components include many terms of the form

$$Z = \begin{cases} Y^{-1} X & |Y| \neq 0, \\ 0 & |Y| = 0, \end{cases}$$

where the denominator Y is a square matrix, and where the numerator X is either a matrix or a vector. Additionally, X and Y have expectations with simple Taylor expansions, so we would like to reduce the asymptotic analysis of $\mathbb{E}(Z)$ to that of the ratio of expectations $\mathbb{E}(Y)^{-1} \mathbb{E}(X)$. The naïve approach would be to use the Taylor expansion

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}\left(\frac{X}{Y} \right) \\ &= \mathbb{E}\left(\frac{\mathbb{E}(X)}{\mathbb{E}(Y)} + \frac{X - \mathbb{E}(X)}{\mathbb{E}(Y)} - \frac{\mathbb{E}(X)(Y - \mathbb{E}(Y))}{\mathbb{E}(Y)^2} + \dots \right) \\ &= \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} + \mathcal{O}\left(\mathrm{Cov}(X, Y) + \mathrm{Var}(Y) \right). \end{aligned}$$

However, Z has a special form when $|Y| = 0$, so this Taylor expansion does not hold. We therefore require an approach that accounts for this special case, but allows us to use the ratio of expectations.

For this section, we approach the problem by showing that $\mathbb{E}\left(Y^{-1}X\right)$ tends to $\mathbb{E}(Y)^{-1} \mathbb{E}(X)$, as N tends to infinity. It then follows that

$$\lim_{N \uparrow \infty} \mathbb{E}\left(Y^{-1}X\right) = \left(\lim_{N \uparrow \infty} \mathbb{E}(Y) \right)^{-1} \lim_{N \uparrow \infty} \mathbb{E}(X),$$

if the limits of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ are finite, and the latter is non-zero.

ABCBAS **Estimate**

Using the approach described above, the expectation for the ABCBAS estimate Z_N has a numerator X and a denominator Y that are scalar. This makes the ABCBAS term the easiest to evaluate under this approach, so we begin by re-finding the asymptotic bias of Z_N.

First, we find conditions under which $\mathbb{E}\left(Z\right)$ tends to $\mathbb{E}\left(X\right)/\mathbb{E}\left(Y\right)$.

Lemma 4.3. *Let X_N be the sum of N IID scalar random variables, and Y_N be a binomial variable $Y_N \sim \mathrm{Bin}(N, p_N)$, such that*

$$\frac{\mathbb{E}\left(X_N^2\right)}{N^2 p_N^2} = \mathcal{O}\left(1\right), \ \textit{and} \ \frac{\mathbb{E}\left(Y_N\right)}{N p_N} \to 1,$$

as $N \uparrow \infty$. Let $g(X_N, Y_N) = c$, for some constant c, in the case where the binomial variable Y_N is zero, and $g(X_N, Y_N) = X_N/Y_N$ otherwise. Let p_N tend to zero, and Np_N tend to infinity, as $N \uparrow \infty$. Then

$$\lim_{N\uparrow\infty} \mathbb{E}\left(g(X_N, Y_N)\right) = \lim_{N\uparrow\infty} \frac{\mathbb{E}\left(X_N\right)}{\mathbb{E}\left(Y_N\right)}.$$

Proof. Similar to the proof for Lemma 3.10. We first note that

$$\begin{aligned}\mathbb{E}\left(g\left(X_N, Y_N\right)\right) =& \mathbb{E}\left(\frac{X_N}{Y_N}[Y_N > 0]\right) + c\mathbb{P}\left(Y_N = 0\right) \\ =& \mathbb{E}\left(\frac{X_N}{Y_N}[Y_N \geq (1+\epsilon)Np_N]\right) \\ &+ \mathbb{E}\left(\frac{X_N}{Y_N}[0 < Y_N \leq (1-\epsilon)Np_N]\right) \\ &+ \mathbb{E}\left(\frac{X_N}{Y_N}[(1-\epsilon)Np_N < Y_N < (1+\epsilon)Np_N]\right) \\ &+ c\mathbb{P}\left(Y_N = 0\right).\end{aligned}$$

Since Y_N is binomial,

$$\log \mathbb{P}\left(Y_N = 0\right) = N \log(1 - p_N) \leq -Np_N,$$

so $c\,\mathbb{P}\left(Y_N = 0\right) \leq c \exp^{-Np_N}$, which vanishes exponentially quickly as $N \uparrow \infty$. By Hölder's inequality and Lemma A.11, the first term satisfies the bound

$$\begin{aligned}\mathbb{E}\left(\frac{X_N}{Y_N}[Y_N \geq (1+\epsilon)Np_N]\right) &\leq \frac{1}{(1+\epsilon)Np_N}\mathbb{E}\left(X_N[Y_N > (1+\epsilon)Np_N]\right) \\ &\leq \frac{\mathbb{E}\left(X_N^2\right)^{1/2}}{(1+\epsilon)Np_N}\mathbb{P}\left(Y_N > (1+\epsilon)Np_N\right)^{1/2} \\ &\leq \frac{\mathbb{E}\left(X_N^2\right)^{1/2}}{(1+\epsilon)Np_N}\exp\left(-\frac{\epsilon^2}{6}Np_N\right).\end{aligned}$$

Since $\mathbb{E}\left(X_N^2\right)^{1/2} = \mathcal{O}\left(Np_N\right)$, this term vanishes exponentially quickly.

Similarly, by Lemma A.10, the second term satisfies the bound

$$\begin{aligned}\mathbb{E}\left(\frac{X_N}{Y_N}[0 < Y_N < (1-\epsilon)Np_N]\right) &\leq \mathbb{E}\left(X_N[0 < Y_N < (1-\epsilon)Np_N]\right) \\ &\leq \mathbb{E}\left(X_N^2\right)^{1/2} \mathbb{P}\left(0 < Y_N < (1-\epsilon)Np_N\right)^{1/2} \\ &\leq \mathbb{E}\left(X_N^2\right)^{1/2} \exp\left(-\frac{\epsilon^2}{4}Np_N\right),\end{aligned}$$

which also vanishes exponentially quickly as $N \uparrow \infty$. Finally, the middle term satisfies the bounds

$$\begin{aligned}\mathbb{E}\left(\frac{X_N}{Y_N}[(1-\epsilon)Np < Y_N < (1+\epsilon)Np]\right) &\leq \frac{1}{(1-\epsilon)Np}\mathbb{E}\left(X_N\left[\left|\frac{Y_N}{Np}-1\right| < \epsilon\right]\right) \\ &\to \frac{\mathbb{E}\left(X_N\right)}{(1-\epsilon)Np} \\ &= \frac{1}{(1-\epsilon)}\frac{\mathbb{E}\left(X_N\right)}{\mathbb{E}\left(Y_N\right)},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(\frac{X_N}{Y_N}[(1-\epsilon)Np < Y_N < (1+\epsilon)Np]\right) &\geq \frac{1}{(1+\epsilon)Np}\mathbb{E}\left(X_N\left[\left|\frac{Y_N}{Np}-1\right| < \epsilon\right]\right) \\ &\to \frac{\mathbb{E}\left(X_N\right)}{(1+\epsilon)Np} \\ &= \frac{1}{(1+\epsilon)}\frac{\mathbb{E}\left(X_N\right)}{\mathbb{E}\left(Y_N\right)},\end{aligned}$$

since $\mathbb{P}\left(\left|\frac{Y_N}{Np}-1\right| < \epsilon\right) \uparrow 1$ as $N \uparrow \infty$. We get the result by letting $\epsilon \downarrow 0$. □

We can now derive the asymptotic bias of Z_N in the new matrix notation.

Lemma 4.4. *Let Assumption 4.2 hold, and the sequence of matrices H be such that $N|H|^{1/2} \uparrow \infty$ as $N \uparrow \infty$, and such that $\lim_{N\uparrow\infty} N|H|^{1/2}T_H^4 > 0$, where*

$$T_H := \operatorname{Tr}\left(H\left[\mathcal{H}_m(s^*) + 2\nabla m(s^*)\frac{\nabla f_S(s^*)}{f_S(s^*)}\right]\right).$$

Then the ABCBAS *estimate Z_N satisfies*

$$\operatorname{bias}(Z_N) = \frac{1}{2}\mu_2(K)T_H + \mathcal{O}\left(\operatorname{Tr}(H)^2\right),$$

as N tends to infinity.

Proof. Let

$$\begin{aligned}X_N &:= \frac{n(Z_N - m(s^*) - \frac{1}{2}\mu_2(K)T_H)}{T_H^2} \\ &= \sum_{k=1}^{N}\frac{\theta_k - m(s^*) - \frac{1}{2}\mu_2(K)T_H}{T_H^2}\left[S_k \,\text{accepted}\right].\end{aligned}$$

We then observe that

$$\begin{aligned}\frac{\mathbb{E}\left(X_N^2\right)}{N^2 p(s^*)^2} &= \frac{\mathbb{E}\left(\left(\sum_{k=1}^N (\theta_k - m(s^*) - \frac{1}{2}\mu_2(K)T_H)\,[S_k \,\text{accepted}]\right)^2\right)}{N^2 p(s^*)^2 T_H^4} \\ &= \frac{\mathbb{E}\left((\theta - m(s^*) - \frac{1}{2}\mu_2(K)T_H)^2 \mid S\,\text{accepted}\right)}{N p(s^*) T_H^4} \\ &\quad + \binom{N}{2} \frac{\left(\mathbb{E}\left(\theta \mid S\,\text{accepted}\right) - m(s^*) - \frac{1}{2}\mu_2(K)T_H\right)^2}{N^2 T_H^4}.\end{aligned}$$

This tends to a finite limit, since

$$\mathbb{E}\left(\theta \mid S\,\text{accepted}\right) = m(s^*) + \frac{1}{2}\mu_2(K)T_H + \mathcal{O}\left(\text{Tr}(H)^2\right),$$

and $\mathbb{E}\left((\theta - m(s^*) - \frac{1}{2}\mu_2(K)T_H)^2 \mid S\,\text{accepted}\right)$ tends to $v(s^*)$, as N tends to infinity. Therefore, the conditions for Lemma 4.3 hold, and so

$$\mathbb{E}\left(Z_N\right) \to \frac{\mathbb{E}\left(X_N\right)}{\mathbb{E}\left(n\right)}, \tag{4.6}$$

as $N \uparrow \infty$. The expectation for X_N is equal to

$$\mathbb{E}\left(X_N\right) = \mathbb{E}\left(n\right)\left(\mathbb{E}\left(\theta \mid S\,\text{accepted}\right) - m(s^*) - \frac{1}{2}\mu_2(K)T_H\right)/T_H^2,$$

which is equal to

$$\mathbb{E}\left(X_N\right) = \mathbb{E}\left(n\right)\mathcal{O}\left(T_H^2\right)/T_H^2 = \mathbb{E}\left(n\right)\mathcal{O}\left(1\right), \tag{4.7}$$

as $N \uparrow \infty$. Substituting Equation 4.7 into Equation 4.6 gives the result

$$\lim_{N\uparrow\infty} \frac{\text{bias}(Z_N) - \frac{1}{2}\mu_2(K)T_H}{T_H^2} = \lim_{N\uparrow\infty} \mathcal{O}\left(1\right),$$

or

$$\lim_{N\uparrow\infty} \text{bias}(Z_N) = \frac{1}{2}\mu_2(K)T_H + \mathcal{O}\left(T_H^2\right).$$

Since $T_H = \mathcal{O}\left(\text{Tr}(H)\right)$ as $N \uparrow \infty$, this completes the proof. □

This marks the end of the finished part of the proof. We now give an outline of the rest of the proof.

Partial Result for Regression Term

We next look at the regression term $\hat{m}(s^*;\hat{H}) = e_1^T \left(X_{s^*}^T \hat{W}_{s^*} X_{s^*}\right)^{-1} X_{s^*}^T \hat{W}_{s^*} \Theta$. We first aim to find conditions under which

$$\mathbb{E}\left(\hat{m}(s^*;\hat{H})\right) \to e_1^T \mathbb{E}\left(X_{s^*}^T \hat{W}_{s^*} X_{s^*}\right)^{-1} \mathbb{E}\left(X_{s^*}^T \hat{W}_{s^*} \Theta\right),$$

as $N \uparrow \infty$. This requires a matrix version of Lemma 4.3.

The proof of Lemma 4.3 involves decomposing the space of possible values for Z_N into several subspaces, based on the proximity of n to its expectation. This raises the question of whether we can define "proximity" so as to obtain a similar decomposition for the space of possible values for $\hat{m}(s;H)$, based on the value of $X_{s^*}^T \hat{W}_{s^*} X_{s^*}$. In this section, we consider defining proximity in terms of the induced matrix norm.

Definition 4.5. *The* induced matrix norm *for the vector norm* $\|\cdot\|$, *for the matrix* A, *is*

$$\|A\| := \sup_{x:\|x\|=1} \{\|Ax\|\} = \sup_{x:\|x\|\neq 0} \{\|Ax\| \,/\, \|x\|\}.$$

Example 4.6. *For the Euclidean norm* $\|\cdot\|_2$, *the induced norm for the real matrix* A *is the square root of the solution to the minimisation problem*

$$\text{Minimise}\, x^T A^T A x \,\text{such that}\, x^T x = 1.$$

By Lagrange multipliers, this solution satisfies

$$A^T A x = \lambda x,\; x^T x = 1,$$

for some Lagrange multiplier λ. *Therefore,* λ *is equal to some eigenvalue of* $A^T A$, *and*

$$x^T A^T A x = \lambda x^T x = \lambda,$$

which is maximised if λ *is equal to the largest eigenvalue of* $A^T A$. *Therefore,* $\|A\|_2$ *is the square root of the largest eigenvalue of* $A^T A$.

Definition 4.7. *The* condition number $\kappa(A) := \|A\|_2 \left\|A^{-1}\right\|_2$ *of the matrix* A *is equal to the square root of the ratio between the largest eigenvalue of* $A^T A$ *and the smallest eigenvalue of* $A^T A$.

Lemma 4.8. *Let* $\|\cdot\|$ *be the induced matrix norm for the vector norm* $\|\cdot\|$, *and* $\kappa(A)$ *be the condition number of the matrix* A. *Let the random matrix* Y *have expectation* μ *and*

$$1_\epsilon(Y) := \left[\|Y - \mu\| \leq \epsilon \frac{\|\mu\|}{\kappa(\mu)}\right].$$

Then

$$\left\|\mathbb{E}\left(\left(Y^{-1}X - \mu^{-1}X\right) 1_\epsilon(Y)\right)\right\| \leq \frac{\epsilon}{1-\epsilon} \mathbb{E}\left(\left\|\mu^{-1}X\right\|\right)$$

for all $0 < \epsilon < 1$.

Proof. Since the vector norm is convex,

$$\left\|\mathbb{E}\left(\left(Y^{-1}X-\mu^{-1}X\right)1_{\epsilon}(Y)\right)\right\| \leq \mathbb{E}\left(\left\|Y^{-1}X-\mu^{-1}X\right\|1_{\epsilon}(Y)\right).$$

If $\frac{\|Y-\mu\|}{\|\mu\|}\kappa(\mu) \leq \epsilon < 1$, it follows that

$$\frac{\left\|Y^{-1}X-\mu^{-1}X\right\|}{\left\|\mu^{-1}X\right\|} \leq \frac{\kappa(\mu)\left\|Y-\mu\right\|/\left\|\mu\right\|}{1-\kappa(\mu)\left\|Y-\mu\right\|/\left\|\mu\right\|} \leq \frac{\epsilon}{1-\epsilon},$$

and therefore [Stoer and Bulirsch, 2002, Theorem 4.4.15]

$$\mathbb{E}\left(\left\|Y^{-1}X-\mu^{-1}X\right\|1_{\epsilon}(Y)\right) \leq \frac{\epsilon}{1-\epsilon}\mathbb{E}\left(\left\|\mu^{-1}X\right\|1_{\epsilon}(Y)\right) \leq \frac{\epsilon}{1-\epsilon}\mathbb{E}\left(\left\|\mu^{-1}X\right\|\right),$$

as required. □

Lemma 4.9. *Let the random matrix Y have expectation μ, and $\|\cdot\|$ be some induced matrix norm. Then Y is invertible if*

$$\|Y-\mu\| \leq \epsilon\frac{\|\mu\|}{\kappa(\mu)},$$

for some $0 \leq \epsilon < 1$.

Proof. Let $F = \mu^{-1}(Y-\mu)$. Then

$$\|F\| \leq \left\|\mu^{-1}\right\|\|Y-\mu\| \leq \epsilon\frac{\left\|\mu^{-1}\right\|\|\mu\|}{\kappa(\mu)} = \epsilon < 1.$$

Therefore [Stoer and Bulirsch, 2002, Theorem 4.4.14], $I+F$ is invertible, and so is $Y = \mu(I+F)$. □

We can now begin a proof of the matrix equivalent of Lemma 4.3.

Conjecture 4.10. *Let the $(q+1)$-length random vector $X_N := A^TBv$, and the random $(q+1)$-square matrix $Y_N := A^TBA$ with expectation μ_N, be such that the $N \times (q+1)$ matrix A has* IID *rows, and almost surely has full rank, such that B is an N-square diagonal matrix with* IID *positive diagonal elements, that are non-zero with probability $\hat{p}(s^*)$, and such that the vector v has* IID *elements. Additionally, let X_N and Y_N be such that*

$$\frac{\mathbb{E}\left(X_N^2\right)}{N^2p_N^2} = \mathcal{O}(1), \text{ and } \frac{\mathbb{E}\left(Y_N\right)}{Np_N} = \mathcal{O}(1),$$

as $N \uparrow \infty$. Let g be the function

$$g(X_N, Y_N) := \begin{cases} Y_N^{-1}X_N & Y_N \textit{ has full rank,} \\ ce_1 & Y_N \textit{ is rank deficient, for some } c. \end{cases}$$

Additionally, let $\hat{p}(s^*)$ tend to zero, and $N\hat{p}(s^*)$ tend to infinity, as $N \uparrow \infty$. Then

$$\lim_{N\uparrow\infty} e_1^T \mathbb{E}\left(g(X_N, Y_N)\right) = \lim_{N\uparrow\infty} e_1^T \mu_N^{-1} \mathbb{E}\left(X_N\right).$$

The conditions given above for Conjecture 4.10 to hold are taken from the conditions for Lemma 4.3 to hold, and may require alteration once the proof is completed.

Partial proof. The desired result is equivalent to

$$d_\infty := \lim_{N\uparrow\infty} \left\| e_1^T \left(\mathbb{E}\left(g\left(X_N, Y_N\right)\right) - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right) \right\| = 0. \tag{4.8}$$

Let $1_F(Y_N) := [Y_N \text{ invertible}]$, and $1_F^C(Y_N) := [Y_N \text{ rank-deficient}] = 1 - 1_F(Y_N)$ be its complement. Then d_∞ satisfies the bound

$$\begin{aligned} d_\infty &\le \|e_1\| \lim_{N\uparrow\infty} \left\| \mathbb{E}\left(ce_1 1_F^C(Y_N) + Y_N^{-1} X_N 1_F(Y_N) \right) - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right\| \\ &\le \lim_{N\uparrow\infty} \left\| \mathbb{E}\left(\left(ce_1 - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right) 1_F^C(Y_N) \right) \right\| \\ &\quad + \lim_{N\uparrow\infty} \left\| \mathbb{E}\left(\left(Y_N^{-1} X_N - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right) 1_F(Y_N) \right) \right\| \qquad (4.9) \\ &\le \lim_{N\uparrow\infty} \left\| ce_1 - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right\| \mathbb{P}\left(Y_N \text{ rank-deficient}\right) \\ &\quad + \lim_{N\uparrow\infty} \left\| \mathbb{E}\left(\left(Y_N^{-1} X_N - \mu_N^{-1} \mathbb{E}\left(X_N\right) \right) 1_F(Y_N) \right) \right\|. \end{aligned}$$

The matrix Y_N is rank-deficient in the case where the number of non-zero diagonal elements in B is no more than q, and in the zero-probability case where used rows of A are not linearly independent. Therefore, if we denote the number of non-zero diagonal elements in B by $\hat{n}$, the probability of Y_N being rank-deficient satisfies the bound

$$\mathbb{P}\left(Y_N \text{ rank-deficient}\right) = \mathbb{P}\left(\hat{n} \le q\right) \le \exp\left(-\frac{\left(N\hat{p}(s^*) - q\right)^2}{2N\hat{p}(s^*)} \right),$$

by Lemma A.10. Since $N\hat{p}(s^*)$ tends to infinity as $N \uparrow \infty$,

$$\mathbb{P}\left(\hat{n} \le q\right) = \mathcal{O}\left(\exp\left(-\frac{1}{2} N\hat{p}(s^*) \right) \right),$$

and vanishes exponentially, as $N \uparrow \infty$.

We now decompose the invertible case according to the deviation of Y_N

from μ_N. Let ϵ be some constant such that $0 < \epsilon < 1$. Then

$$\begin{aligned} d_\infty &\leq \lim_{N\uparrow\infty} \left\|\mathbb{E}\left(\left(Y_N^{-1}X_N - \mu_N^{-1}X_N\right)1_F(Y_N)\right)\right\| \\ &\leq \lim_{N\uparrow\infty} \left\|\mathbb{E}\left(\left(Y_N^{-1}X_N - \mu_N^{-1}X_N\right)1_F(Y_N)\left[\|Y_N - \mu_N\| \leq \epsilon\frac{\|\mu_N\|}{\kappa(\mu_N)}\right]\right)\right\| \\ &\quad + \lim_{N\uparrow\infty} \left\|\mathbb{E}\left(\left(Y_N^{-1}X_N - \mu_N^{-1}X_N\right)1_F(Y_N)\left[\|Y_N - \mu_N\| > \epsilon\frac{\|\mu_N\|}{\kappa(\mu_N)}\right]\right)\right\| \\ &\leq \frac{\epsilon}{1-\epsilon}\lim_{N\uparrow\infty}\mathbb{E}\left(\left\|\mu_N^{-1}X_N\right\|\right) \\ &\quad + \lim_{N\uparrow\infty} \left\|\mathbb{E}\left(\left(Y_N^{-1}X_N - \mu_N^{-1}X_N\right)1_F(Y_N)\left[\|Y_N - \mu_N\| > \epsilon\frac{\|\mu_N\|}{\kappa(\mu_N)}\right]\right)\right\|, \end{aligned}$$

where $1_F(Y_N)\left[\|Y_N - \mu_N\| \leq \epsilon\frac{\|\mu_N\|}{\kappa(\mu_N)}\right] = \left[\|Y_N - \mu_N\| \leq \epsilon\frac{\|\mu_N\|}{\kappa(\mu_N)}\right]$, and the last inequality follows from Lemma 4.8.

This marks the end of current progress on the proof, with two terms for which it remains to prove vanishment. For $\frac{\epsilon}{1-\epsilon}\lim_{N\uparrow\infty}\mathbb{E}\left(\left\|\mu^{-1}X_N\right\|\right)$, the first remaining term, ϵ will later be taken to zero, so it will suffice to show that $\lim_{N\uparrow\infty}\mathbb{E}\left(\left\|\mu^{-1}X_N\right\|\right)$ is finite. □

If Conjecture 4.10 holds, then we can consider the asymptotic behaviour of $\hat{m}(s^*; H)$ in a similar way to that of Z_N.

Conjecture 4.11. *Let Assumption 4.2 hold, and $N|\hat{H}|^{1/2} \uparrow \infty$ as $N \uparrow \infty$. Then the expression*

$$\hat{m}(s^*; \hat{H}) = e_1^T\left(X_{s^*}^T\hat{W}_{s^*}X_{s^*}\right)^{-1}X_{s^*}^T\hat{W}_{s^*}\Theta.$$

has expectation

$$\mathbb{E}\left(\hat{m}(s^*; \hat{H})\right) = m(s^*) + \frac{1}{2}\mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\mathcal{H}_m(s^*)\right) + \mathcal{O}\left(\mathrm{Tr}(\hat{H})^2\right)$$

as $N \uparrow \infty$.

Partial proof. Given that Conjecture 4.10 is true, $\mathbb{E}\left(\hat{m}(s^*; H)\right)$ satisfies

$$\lim_{N\uparrow\infty}\mathbb{E}\left(\hat{m}(s^*; H)\right) = \lim_{N\uparrow\infty} e_1^T D_{s^*}^{-1}M_{s^*}, \tag{4.10}$$

where $D_{s^*} := \mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}X_{s^*}\right)$, and $M_{s^*} := \mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}\Theta\right)$. We proceed in a way similar to that used in [Ruppert and Wand, 1994, Theorem 2.1]. The denominator is equal to

$$\begin{aligned} D_{s^*} &= \begin{pmatrix} \mathbb{E}\left(\hat{K}_{\hat{H}}(S-s^*)\right) & \mathbb{E}\left(\hat{K}_{\hat{H}}(S-s^*)(S-s^*)^T\right) \\ \mathbb{E}\left(\hat{K}_{\hat{H}}(S-s^*)(S-s^*)\right) & \mathbb{E}\left(\hat{K}_{\hat{H}}(S-s^*)(S-s^*)(S-s^*)^T\right) \end{pmatrix} \\ &= \begin{pmatrix} \int \hat{K}(u) f_S(s^* + \hat{H}^{1/2}u)\,\mathrm{d}u & \int \hat{K}(u) u^T \hat{H}^{1/2} f_S(s^* + \hat{H}^{1/2}u)\,\mathrm{d}u \\ \int \hat{K}(u) \hat{H}^{1/2} u f_S(s^* + \hat{H}^{1/2}u)\,\mathrm{d}u & \int \hat{K}(u) \hat{H}^{1/2} u u^T \hat{H}^{1/2} f_S(s^* + \hat{H}^{1/2}u)\,\mathrm{d}u \end{pmatrix} \\ &= \begin{pmatrix} f_S(s^*) & \mu_2(\hat{K}) \nabla f_S(s^*) \hat{H} \\ \mu_2(\hat{K}) \hat{H} \nabla f_S(s^*)^T & \mu_2(\hat{K}) f_S(s^*) \hat{H} \end{pmatrix} \left(1 + \mathcal{O}\left(\operatorname{Tr}(\hat{H})\right)\right), \end{aligned}$$

as $N \uparrow \infty$, and so, by Lemma A.12,

$$D_{s^*}^{-1} = \begin{pmatrix} f_S(s^*)^{-1} & -\nabla f_S(s^*) f_S(s^*)^{-2} \\ -\nabla f_S(s^*)^T f_S(s^*)^{-2} & \mu_2(\hat{K})^{-1} f_S(s^*)^{-1} \hat{H}^{-1}. \end{pmatrix} \left(1 + \mathcal{O}\left(\operatorname{Tr}(\hat{H})\right)\right). \tag{4.11}$$

For the numerator, we first note that the conditional expectation is equal to

$$\mathbb{E}\left(X_{s^*}^T \hat{W}_{s^*} \Theta \,\middle|\, s_1, \ldots, s_N\right) = X_{s^*}^T \hat{W}_{s^*} \mathbb{E}\left(\Theta \,|\, s_1, \ldots, s_N\right).$$

The conditional expectation of Θ is equal to

$$\begin{aligned} \mathbb{E}\left(\Theta \,|\, s_1, \ldots, s_N\right) &= \begin{pmatrix} m(s_1) & \ldots & m(s_N) \end{pmatrix}^T \\ &= X_{s^*} \begin{pmatrix} m(s^*) \\ \nabla m(s^*)^T \end{pmatrix} + \frac{1}{2} d_m^{(2)}(s^*) + R_m(s^*), \end{aligned}$$

where $R_m(s^*)$ is a remainder term, and

$$d_m^{(2)}(s^*) := \begin{pmatrix} (s_1 - s^*)^T \mathcal{H}_m(s^*)(s_1 - s^*) \\ \vdots \\ (s_N - s^*)^T \mathcal{H}_m(s^*)(s_N - s^*) \end{pmatrix}. \tag{4.12}$$

The unconditional expectation of M_{s^*} is therefore equal to

$$\begin{aligned}
M_{s^*} =& \mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}X_{s^*}\begin{pmatrix} m(s^*) \\ \nabla m(s^*)^T\end{pmatrix}\right) + \frac{1}{2}\mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}d_m^{(2)}(s^*)\right) \\
&+ \mathbb{E}\left(R_m(s^*)\right) \\
=& D_{s^*}\begin{pmatrix} m(s^*) \\ \nabla m(s^*)^T\end{pmatrix} \\
&+ \frac{1}{2}\mathbb{E}\left(\frac{1}{N}\begin{pmatrix} \sum_{k=1}^q \hat{K}_{\hat{H}}(s_k - s^*)(s_k - s^*)^T\mathcal{H}_m(s^*)(s_k - s^*) \\ \sum_{k=1}^q (s_k - s^*)\hat{K}_{\hat{H}}(s_k - s^*)(s_k - s^*)^T\mathcal{H}_m(s^*)(s_k - s^*)\end{pmatrix}\right) \\
&+ \mathbb{E}\left(R_m(s^*)\right) \\
=& D_{s^*}\begin{pmatrix} m(s^*) \\ \nabla m(s^*)^T\end{pmatrix} \\
&+ \frac{1}{2}\begin{pmatrix} \mathbb{E}\left(\hat{K}_{\hat{H}}(S - s^*)(S - s^*)^T\mathcal{H}_m(s^*)(S - s^*)\right) \\ \mathbb{E}\left(\hat{K}_{\hat{H}}(S - s^*)(S - s^*)(S - s^*)^T\mathcal{H}_m(s^*)(S - s^*)\right)\end{pmatrix} \\
&+ \mathbb{E}\left(R_m(s^*)\right),
\end{aligned}$$

where the middle term is proportional to

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}d_m^{(2)}(s^*)\right) &= \begin{pmatrix} \mathbb{E}\left(\hat{K}_{\hat{H}}(S - s^*)(S - s^*)^T\mathcal{H}_m(s^*)(S - s^*)\right) \\ \mathbb{E}\left(\hat{K}_{\hat{H}}(S - s^*)(S - s^*)(S - s^*)^T\mathcal{H}_m(s^*)(S - s^*)\right)\end{pmatrix} \\
&= \begin{pmatrix} \int \hat{K}(u)u^T\hat{H}^{1/2}\mathcal{H}_m(s^*)\hat{H}^{1/2}uf\left(s^* + \hat{H}^{1/2}u\right)\,\mathrm{d}u \\ \int \hat{K}(u)\hat{H}^{1/2}uu^T\hat{H}^{1/2}\mathcal{H}_m(s^*)\hat{H}^{1/2}uf\left(s^* + \hat{H}^{1/2}u\right)\,\mathrm{d}u\end{pmatrix} \\
&= \begin{pmatrix} \mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\mathcal{H}_m(s^*)\right)f(s^*)\left(1 + \mathcal{O}\left(\mathrm{Tr}(\hat{H})\right)\right) \\ \mathcal{O}\left(\hat{H}^{3/2}\right)\end{pmatrix},
\end{aligned}$$

as $N \uparrow \infty$. Therefore, M_{s^*} is equal to

$$\begin{aligned}
M_{s^*} =& D_{s^*}\begin{pmatrix} m(s^*) \\ \nabla m(s^*)^T\end{pmatrix} \\
&+ \frac{1}{2}\begin{pmatrix} \mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\mathcal{H}_m(s^*)\right)f(s^*)\left(1 + \mathcal{O}\left(\mathrm{Tr}(\hat{H})\right)\right) \\ \mathcal{O}\left(\hat{H}^{3/2}\right)\end{pmatrix} \\
&+ \mathcal{O}\left(\mathrm{Tr}(\hat{H})^2\right).
\end{aligned} \tag{4.13}$$

Substituting Equations 4.11 and 4.13 into Equation 4.10 gives

$$\mathbb{E}\left(\hat{m}(s^*;\hat{H})\right) \to m(s^*) + \frac{1}{2}\mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\mathcal{H}_m(s^*)\right) + \mathcal{O}\left(\mathrm{Tr}(\hat{H})^2\right),$$

as $N \uparrow \infty$, as required. □

We note that this partial proof currently only shows that $\mathbb{E}\left(\hat{m}(s^*;\hat{H})\right)$ tends to $e_1^T D_{s^*}^{-1} M_{s^*}$ as $N \uparrow \infty$, but we require

$$\mathbb{E}\left(\hat{m}(s^*;\hat{H})\right) = e_1^T D_{s^*}^{-1} M_{s^*} + \mathcal{O}\left(\mathrm{Tr}(\hat{H})^2\right).$$

This is also the case for partial proofs in following sections.

Partial Result for Remainder Term

Conjecture 4.12. *Let Assumption 4.2 hold, and $N|\hat{H}|^{1/2}$ tend to infinity as N does. Then the expression*

$$B_N := \frac{1}{n}\sum_{j=1}^{n} \hat{m}(s_{k_j};\hat{H}) = \left(1^T W_{s^*} 1\right)^{-1} 1^T W_{s^*} \begin{pmatrix} \hat{m}(s_1;\hat{H}) \\ \vdots \\ \hat{m}(s_N;\hat{H}) \end{pmatrix}$$

has expectation

$$\begin{aligned}\mathbb{E}(B_N) =& \mathbb{E}(Z_N) + \frac{1}{2}\mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\mathcal{H}_m(s^*)\right) \\ &+ \mathcal{O}\left(\mathrm{Tr}(H)\mathrm{Tr}(\hat{H})\right) + \mathcal{O}\left(\mathrm{Tr}(\hat{H})^2\right) \\ =& m(s^*) + \frac{1}{2}\mu_2(K)\mathrm{Tr}\left(H\mathcal{H}_m(s^*)\right) \\ &+ \frac{1}{2}\mu_2(\hat{K})\mathrm{Tr}\left(\hat{H}\left[\nabla m(s^*)^T \frac{\nabla f(s^*)}{f(s^*)} + \mathcal{H}_m(s^*)\right]\right) \\ &+ \mathcal{O}\left(\left(\mathrm{Tr}(H) + \mathrm{Tr}(\hat{H})\right)^2\right),\end{aligned}$$

as $N \uparrow \infty$.

Partial proof. By Lemma 4.3,

$$\mathbb{E}(B_N) \to \frac{\mathbb{E}\left(1^T W_{s^*} \begin{pmatrix} \hat{m}(s_1;\hat{H}) \\ \vdots \\ \hat{m}(s_N;\hat{H}) \end{pmatrix}\right)}{\mathbb{E}\left(1^T W_{s^*} 1\right)},$$

as $N \uparrow \infty$, where $\mathbb{E}\left(1^T W_{s^*} 1\right) = \mathbb{E}(n)$ and

$$\mathbb{E}\left(1^T W_{s^*} \begin{pmatrix} \hat{m}(s_1;\hat{H}) \\ \vdots \\ \hat{m}(s_N;\hat{H}) \end{pmatrix}\right) = \mathbb{E}(n)\,\mathbb{E}\left(e_1^T (X_{s_1}^T \hat{W}_{s_1} X_{s_1})^{-1} X_{s_1}^T \hat{W}_{s_1} \Theta \,\middle|\, s_1 \text{ accepted}\right),$$

so that $\mathbb{E}\left(B_N\right) \to \mathbb{E}\left(\hat{m}(s_1;\hat{H})\,\middle|\, s_1\,\text{accepted}\right)$ as $N \uparrow \infty$. Again, we assume Conjecture 4.10 is true. The individual regression estimate

$$\hat{m}(s_1;\hat{H}) = e_1^T\left(X_{s_1}^T\hat{W}_{s_1}X_{s_1}\right)^{-1}X_{s_1}^T\hat{W}_{s_1}\Theta,$$

then has expectation

$$\mathbb{E}\left(\hat{m}(s_1;\hat{H})\,\middle|\, s_1\,\text{accepted}\right) \to e_1^T D^{-1}M,$$

as $N \uparrow \infty$, where

$$D := \frac{1}{N}\mathbb{E}\left(X_{s_1}^T\hat{W}_{s_1}X_{s_1}\,\middle|\, s_1\,\text{accepted}\right), \quad M := \frac{1}{N}\mathbb{E}\left(X_{s_1}^T\hat{W}_{s_1}\Theta\,\middle|\, s_1\,\text{accepted}\right).$$

The denominator D is equal to

$$D = \frac{N-1}{N}\begin{pmatrix} \frac{\hat{K}_{\hat{H}}(0)}{N-1} + \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)\right) & \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)(s_k-s_1)^T\right) \\ \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)(s_k-s_1)\right) & \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)(s_k-s_1)(s_k-s_1)^T\right)\end{pmatrix}$$

where all the expectations are conditional on s_1 being accepted. The numerator M is equal to

$$\begin{aligned} M &= \mathbb{E}\left(\frac{1}{N}X_{s_1}^T\hat{W}_{s_1}\Theta\,\middle|\, s_1\,\text{accepted}\right) \\ &= \frac{N-1}{N}\begin{pmatrix} \frac{1}{N-1}\hat{K}_{\hat{H}}(0) + \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)\theta\,\middle|\, s_1\,\text{accepted}\right) \\ \mathbb{E}\left(\hat{K}_{\hat{H}}(s_k - s_1)\theta(s_k-s_1)\,\middle|\, s_1\,\text{accepted}\right)\end{pmatrix}, \end{aligned}$$

The proof will then expand D and M, in a similar way to the expansions for D_{s^*} and M_{s^*} in the proof for Conjecture 4.11. This is complicated by the $\frac{1}{N-1}\hat{K}_{\hat{H}}(0)$ term, which is $\mathcal{O}\left(\frac{1}{N|\hat{H}|^{1/2}}\right)$ as $N \uparrow \infty$. □

4.2.2 Variance

We now give an outline of the intended approach for finding the asymptotic variance. The variance can be decomposed into variance terms for the separate components of the estimate $\hat{Z}_N$, and their covariance terms:

$$\begin{aligned} \text{Var}\left(\hat{Z}_N\right) &= \text{Var}\left(Z_N + \hat{m}(s^*;\hat{H}) - \frac{1}{n}\sum_{j=1}^{n}\hat{m}(s_{k_j};\hat{H})\right) \\ &= \text{Var}\left(Z_N\right) + \text{Var}\left(\hat{m}(s^*;\hat{H})\right) + \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n}\hat{m}(s_{k_j};\hat{H})\right) \\ &\quad + 2\text{Cov}\left(Z_N, \hat{m}(s^*;\hat{H})\right) - 2\text{Cov}\left(Z_N, \hat{m}(s_1;\hat{H})\,|\, s_1\,\text{accepted}\right). \end{aligned}$$

Variance Terms

The variance terms for Z_N and $\hat{m}(s^*;\hat{H})$ are more straightforward, as the approach is similar to that used for the bias terms in Section 4.2.1. The initial motivation is the result, from Ruppert and Wand [1994], that

$$\begin{aligned}\operatorname{Var}\left(\hat{m}(s^*;\hat{H})\,\middle|\, s_1,\ldots,s_N\right) &= \operatorname{Var}\left(\left(X_{s^*}^T\hat{W}_{s^*}X_{s^*}\right)^{-1}X_{s^*}^T\hat{W}_{s^*}\Theta\,\middle|\, s_1,\ldots,s_N\right)\\ &= \left(X_{s^*}^T\hat{W}_{s^*}X_{s^*}\right)^{-1}X_{s^*}^T\hat{W}_{s^*}V\hat{W}_{s^*}X_{s^*}\left(X_{s^*}^T\hat{W}_{s^*}X_{s^*}\right)^{-1},\end{aligned}$$

where $V := \operatorname{Var}(\Theta\,|\,s_1,\ldots,s_N) = \operatorname{diag}(v(s_1),\ldots,v(s_N))$. The asymptotic expansion for $\mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}V\hat{W}_{s^*}X_{s^*}\right)$, as $N\uparrow\infty$, can be found with the same approach as used in Section 4.2.1. It is then hoped that, for some term

$$Z = \begin{cases} W^{-1}XY^{-1} & |Y| \neq 0,\\ 0 & |Y| = 0,\end{cases}$$

where W,X, and Y are matrices, we can find conditions under which

$$\mathbb{E}(Z) \to \mathbb{E}(W)^{-1}\,\mathbb{E}(X)\,\mathbb{E}(Y)^{-1},$$

as $N\uparrow\infty$, similar to those in Lemma 4.3. This would allow us to find an asymptotic expansion for $\mathbb{E}(\operatorname{Var}(Z\,|\,s_1,\ldots,s_N))$, where Z can be equal to any of the components: Z_N, $\hat{m}(s^*;\hat{H})$, or B_N. The asymptotic expansion for $\operatorname{Var}(Z)$ could then be found using the law of total variance,

$$\operatorname{Var}(Z) = \mathbb{E}(\operatorname{Var}(Z\,|\,s_1,\ldots,s_N)) + \operatorname{Var}(\mathbb{E}(Z\,|\,s_1,\ldots,s_N)).$$

As an example, we now give a proof for the variance of the ABCBAS estimate Z_N.

Lemma 4.13. *Let Assumption 4.2 hold, and the sequence of matrices H be such that $N|H|^{1/2}\uparrow\infty$ as $N\uparrow\infty$. Then Z_N satisfies*

$$\operatorname{Var}(Z_N) \to \frac{R(K)}{N|H|^{1/2}}\frac{v(s^*)}{f(s^*)}\left(1+\mathcal{O}(\operatorname{Tr}(H))\right),$$

as $N\uparrow\infty$.

Proof. We note that

$$\operatorname{Var}(Z_N) = \mathbb{E}(\operatorname{Var}(Z_N\,|\,s_1,\ldots,s_N,W_{s^*})) + \operatorname{Var}(\mathbb{E}(Z_N\,|\,s_1,\ldots,s_N,W_{s^*})),$$

by the law of total variance. If we let $V = \operatorname{diag}(v(s_1), \dots, v(s_N))$, then [Ruppert and Wand, 1994]

$$\operatorname{Var}(Z_N \mid s_1, \dots, s_N, W_{s^*}) = \left(1^T W_{s^*} 1\right)^{-1} 1^T W_{s^*} V W_{s^*} 1 \left(1^T W_{s^*} 1\right)^{-1},$$

where

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N} 1^T W_{s^*} 1\right) &= \int \frac{K_H(s - s^*)}{\max_u K_H(u)} f_S(s)\, \mathrm{d}s \\ &= |H|^{1/2} \frac{f_S(s^*)}{\max_u K(u)} \left(1 + \mathcal{O}\left(\operatorname{Tr}(H)\right)\right),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N} 1^T W_{s^*} V W_{s^*} 1\right) &= \int \frac{K_H(s - s^*)^2}{\left(\max_u K_H(u)\right)^2} v(s) f_S(s)\, \mathrm{d}s \\ &= |H|^{1/2} \int \frac{K(u)^2}{\left(\max_u K(u)\right)^2} v(s^* + H^{1/2} u) f_S(s^* + H^{1/2} u)\, \mathrm{d}u \\ &= |H|^{1/2} R(K) \frac{v(s^*) f_S(s^*)}{\left(\max_u K(u)\right)^2} \left(1 + \mathcal{O}\left(\operatorname{Tr}(H)\right)\right),\end{aligned}$$

as $N \uparrow \infty$. Therefore, the expectation of the conditional variance tends to

$$\frac{\mathbb{E}\left(\frac{1}{N} 1^T \hat{W}_{s^*} V \hat{W}_{s^*} 1\right)}{N \mathbb{E}\left(\frac{1}{N} 1^T \hat{W}_{s^*} 1\right)^2} = \frac{R(K)}{|H|^{1/2}} \frac{v(s^*)}{f_S(s^*)} \left(1 + \mathcal{O}\left(\operatorname{Tr}(H)\right)\right), \tag{4.14}$$

as $N \uparrow \infty$. For the variance of the conditional expectation, we know that [Ruppert and Wand, 1994]

$$\begin{aligned}\mathbb{E}(Z_N \mid s_1, \dots, s_N, W_{s^*}) &= \left(1^T W_{s^*} 1\right)^{-1} 1^T W_{s^*} \begin{pmatrix} m(s_1) \\ \vdots \\ m(s_N) \end{pmatrix} \\ &= m(s^*) + \frac{1}{2} d_m^{(2)}(s^*) + \mathcal{O}\left(\operatorname{Tr}(H)^2\right),\end{aligned}$$

where $d_m^{(2)}(s^*)$ is defined in Equation 4.12. The variance of this conditional expectation tends to

$$\begin{aligned}\operatorname{Var}\left(\mathbb{E}(Z_N \mid s_1, \dots, s_N, W_{s^*})\right) &= \frac{\operatorname{Var}(\theta \mid S \,\text{accepted})}{N |H|^{1/2}} (1 + \mathrm{o}(1)) \\ &= \frac{v(s^*)}{N |H|^{1/2}} (1 + \mathrm{o}(1)),\end{aligned} \tag{4.15}$$

as $N \uparrow \infty$. Adding Equations 4.14 and 4.15 gives the required result. □

Covariance Terms

The covariance terms are less trivial, as the resulting kernel integrals require more knowledge about the kernel functions K and $\hat{K}$.

For example, finding the value of $\operatorname{Cov}\left(\hat{m}(s^*;\hat{H}), \hat{m}(s_1;\hat{H}) \mid s_1 \text{ accepted}\right)$, by a similar approach to the other results in this chapter, requires finding the value of $\mathbb{E}\left(\frac{1}{N}X_{s^*}^T\hat{W}_{s^*}V\hat{W}_{s_1}X_{s_1}\,[s_1 \text{ accepted}]\right)$, the top-left element of which is equal to

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{N}\sum_{k=1}^{N}\hat{K}_{\hat{H}}(s_k - s^*)\hat{K}_{\hat{H}}(s_k - s_1)v(s_k)\,[s_1 \text{ accepted}]\right)\\
&= \frac{1}{N}\hat{K}_{\hat{H}}(0)\int K_H(s_1 - s^*)\hat{K}_{\hat{H}}(s_1 - s^*)v(s_1)f_S(s_1)\,\mathrm{d}s_1\\
&\quad + \left(1 - \frac{1}{N}\right)\iint K_H(s_1 - s^*)\hat{K}_{\hat{H}}(s - s^*)\hat{K}_{\hat{H}}(s - s_1)v(s)f(s)\,\mathrm{d}s_1\,\mathrm{d}s\\
&= \frac{1}{N|\hat{H}|}\hat{K}(0)\int K(u)\hat{K}(\hat{H}^{-1/2}H^{1/2}u)v(s^* + H^{1/2}u)f_S(s^* + H^{1/2}u)\,\mathrm{d}u\\
&\quad + \frac{N-1}{N|\hat{H}|}\iint K(u)\hat{K}(v)\hat{K}(\hat{H}^{-1/2}H^{1/2}u + v)v(s^* + H^{1/2}u + \hat{H}^{1/2}v)\\
&\qquad\qquad \times f(s^* + H^{1/2}u + \hat{H}^{1/2}v)\,\mathrm{d}u\,\mathrm{d}v.
\end{aligned}$$

In the case that $H = \hat{H}$, this simplifies to

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{N}\sum_{k=1}^{N}\hat{K}_{\hat{H}}(s_k - s^*)\hat{K}_{\hat{H}}(s_k - s_1)v(s_k)\,[s_1 \text{ accepted}]\right)\\
&= \frac{1}{N|H|}\hat{K}(0)\int K(u)\hat{K}(u)\,\mathrm{d}u\,v(s^*)f_S(s^*)\left(1 + \mathcal{O}\left(\operatorname{Tr}(H)\right)\right) \qquad (4.16)\\
&\quad + \frac{N-1}{N|H|}\iint K(u)\hat{K}(v)\hat{K}(u+v)\,\mathrm{d}u\,\mathrm{d}v\,v(s^*)f(s^*)\left(1 + \mathcal{O}\left(\operatorname{Tr}(H)\right)\right),
\end{aligned}$$

since the matrix

$$\iint K(u)\hat{K}(v)\hat{K}(u+v)\begin{pmatrix}1\\u\end{pmatrix}\begin{pmatrix}1\\v\end{pmatrix}^T\mathrm{d}u\,\mathrm{d}v$$

is block-diagonal. The evaluation of Equation 4.16 requires knowledge of the value of $\iint K(u)\hat{K}(v)\hat{K}(u+v)\,\mathrm{d}u\,\mathrm{d}v$, an integral of higher order than the roughness.

4.2.3 Optimising the Error and Cost

For the expected computational cost C, the cost of generating and accepting samples is linear in N. This results in $n = \mathcal{O}\left(N\,|H|^{1/2}\right)$ accepted proposals.

Each of the $n+1$ resulting linear regressions then has an expected cost that is proportional to the number of samples used, the magnitude of which depends on the support of the regression kernel function $\hat{K}$. In the case where $\hat{K}$ has finite support, each regression uses $\mathcal{O}\left(N|\hat{H}|^{1/2}\right)$ samples. In the case where $\hat{K}$ has infinite support, each regression uses all N samples. The computational cost is therefore equal to

$$C = c_1(s^*)N + c_2(s^*)N^2|H|^{1/2}|\hat{H}|^{1/2}(1+\mathrm{o}(1)),$$

or

$$C = c_1(s^*)N + c_2(s^*)N^2|H|^{1/2}(1+\mathrm{o}(1)),$$

respectively, as $N \uparrow \infty$.

To optimise the rate of convergence, we must optimise an expression that includes both $|H|^{1/2}$ and $\mathrm{Tr}(H)$, and that includes both $|\hat{H}|^{1/2}$ and $\mathrm{Tr}(\hat{H})$. This presents the same difficulties as in Section 3.4.2. We therefore assume that the square-bandwidth matrices have the forms $H = \hat{H} = \delta^2 I$, so that

$$\mathrm{MSE}(\hat{Z}_N) = \left(\frac{v(s^*)}{N\delta^q} + d(s^*)\delta^8\right)(1+\mathrm{o}(1)), \tag{4.17}$$

and, if the regression kernel function $\hat{K}$ has finite support,

$$C = c_1(s^*)N + c_2(s^*)N^2\delta^{2q}(1+\mathrm{o}(1)),$$

as $C \uparrow \infty$. For this simple case, we can prove the following two theorems for the rate of convergence.

Theorem 4.14. *Let Assumption 4.2 hold, $\hat{K}$ have finite support, and the square-bandwidth matrices be $H = \hat{H} = \delta^2 I$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*) > 0$. Then the following statements hold:*

1. *Let $q < 8$, and $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that the limit $D_2 := \lim_{C\uparrow\infty} N\delta^{2q}$ is strictly positive and finite. Then the error and expected cost are such that*

$$\lim_{C\uparrow\infty} N\delta^q\,\mathrm{MSE}(\hat{Z}_N),\ \lim_{C\uparrow\infty} N^{-2}\delta^{-2q}C,\ \textit{and}\ \lim_{C\uparrow\infty} C^{1/2}\mathrm{MSE}(\hat{Z}_N)$$

are strictly positive and finite.

2. *Let $q \geq 8$, and $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that the limit $D := \lim_{C\uparrow\infty} N\delta^{q+8}$ is strictly positive and finite. Then the error and expected cost are such that*

$$\lim_{C\uparrow\infty} N\delta^q \, \mathrm{MSE}(\hat{Z}_N), \; \lim_{C\uparrow\infty} N^{-(1+q/4)}\delta^{-q(1+q/4)}C, \; \lim_{C\uparrow\infty} C^{8/(q+8)}\mathrm{MSE}(\hat{Z}_N)$$

are strictly positive and finite.

As for Theorem 3.7, we first give a rough approach. We require the limit of

$$\begin{aligned} C^{-\alpha}\mathrm{MSE}(\hat{Z}_N) &= \left(c_1(s^*)N + c_2(s^*)N^2\delta^{2q}\right)^{-\alpha}\left(\frac{v(s^*)}{N\delta^q} + d(s^*)\delta^8\right)(1+\mathrm{o}\,(1)) \\ &= c_1(s^*)^{-\alpha}(N^{-1-\alpha}\delta^{-q})\left(1+\frac{c_2(s^*)}{c_1(s^*)}N\delta^{2q}\right)^{-\alpha}\left(v(s^*)+d(s^*)N\delta^{q+8}\right), \end{aligned}$$

as $C \uparrow \infty$, to be non-zero and finite, for some negative α. This holds if the limits for $N^{-1-\alpha}\delta^{-q}$, $N\delta^{q+8}$, and $N\delta^{2q}$ exist. The order of the latter two limits depends on the value of q, so we must consider two cases.

First, we rewrite the limit as

$$\lim_{C\uparrow\infty} C^{-\alpha}\mathrm{MSE}(\hat{Z}_N) = c_1(s^*)^{-\alpha}D_1\left(1+\frac{c_2(s^*)}{c_1(s^*)}D_2\right)^{-\alpha}\left(v(s^*)+d(s^*)D\right),$$

where the order of the limits $D := \lim_{C\uparrow\infty} N\delta^{q+8}$ and $D_2 := \lim_{C\uparrow\infty} N\delta^{2q}$ depends on the value of q. Additionally, the limit $D_1 := \lim_{C\uparrow\infty} N^{-1-\alpha}\delta^{-q}$ must exist. Since D_1 satisfies

$$D_1 = D^{-q/(q+8)} \lim_{C\uparrow\infty} N^{-8/(q+8)-\alpha} = D_2^{-1/2} \lim_{C\uparrow\infty} N^{-1/2-\alpha},$$

and must be non-zero, either D or D_2 must be non-zero.

If $q \geq 8$, D must be non-zero, so we can proceed as in Theorem 3.7. For D_1 to exist, we therefore require $\alpha \geq -8/(q+8)$. We therefore minimise α by setting it equal to $-8/(q+8)$. Then $D_1 = D^{-q/(q+8)}$ and $D_2 = D[q=8]$, and $\mathrm{MSE}(\hat{Z}_N) = \mathcal{O}\left(C^{-8/(q+8)}\right)$ as $C \uparrow \infty$.

If $q < 8$, D_2 must be non-zero, and $D = 0$. For D_1 to exist, we therefore require $\alpha \geq -1/2$. The minimal value of α is therefore $-1/2$, so that $D_1 = D_2^{-1/2}$, and the mean square error satisfies $\mathrm{MSE}(\hat{Z}_N) = \mathcal{O}\left(C^{-1/2}\right)$ as $C \uparrow \infty$.

Proof. Using Equation 4.17, we find that

$$\begin{aligned}\lim_{C\uparrow\infty} N\delta^q \,\mathrm{MSE}(\hat{Z}_N) &= \lim_{C\uparrow\infty} N\delta^q \left(\mathrm{Var}(\hat{Z}_N) + \mathrm{bias}(\hat{Z}_N)^2\right)\\ &= v(s^*) + d(s^*) \lim_{C\uparrow\infty} N\delta^{q+8}\\ &= v(s^*) + d(s^*)D.\end{aligned}$$

We first consider the case where $q \geq 8$. For the expected cost

$$C = c_1(s^*)N + c_2(s^*)N^2\delta^{2q}(1 + \mathrm{o}\,(1)),$$

we find that

$$\begin{aligned}\lim_{C\uparrow\infty} (N\delta^q)^{-1-q/8}\, C &= c_1(s^*) \lim_{C\uparrow\infty} N^{-q/8}\delta^{-q(1+q/8)} + c_2(s^*) \lim_{C\uparrow\infty} N^{1-q/8}\delta^{(1-q/8)q}\\ &= c_1(s^*)D^{-q/8} + c_2(s^*)D^{-q/8} \lim_{C\uparrow\infty} N\delta^{2q}\\ &= c_1(s^*)D^{-q/8} + c_2(s^*)D^{1-q/8}[q = 8].\end{aligned}$$

Finally, combining this result for the cost with that for the error, we get the result

$$\begin{aligned}\lim_{C\uparrow\infty} C^{8/(q+8)}\mathrm{MSE}(\hat{Z}_N) &= \lim_{C\uparrow\infty} \left((N\delta^q)^{-1-q/8}\, C\right)^{8/(q+8)} N\delta^q \,\mathrm{MSE}(\hat{Z}_N)\\ &= (c_1(s^*) + c_2(s^*)D[q = 8])^{8/(q+8)}\\ &\quad \times D^{-q/(q+8)} \left(v(s^*) + d(s^*)D\right),\end{aligned}$$

which is non-zero and finite.

For the case where $q < 8$, we observe that

$$\begin{aligned}\lim_{C\uparrow\infty} (N\delta^q)^{-2}\, C &= c_1(s^*) \lim_{C\uparrow\infty} N^{-1}\delta^{-2q} + c_2(s^*)\\ &= c_1(s^*)D_2^{-1} + c_2(s^*).\end{aligned}$$

Combining this result for the cost with that for the error, and noting that $D = 0$, we get the result

$$\begin{aligned}\lim_{C\uparrow\infty} C^{1/2}\mathrm{MSE}(\hat{Z}_N) &= \lim_{C\uparrow\infty} \left((N\delta^q)^{-2}\, C\right)^{1/2} N\delta^q \,\mathrm{MSE}(\hat{Z}_N)\\ &= \left(c_1(s^*)D_2^{-1} + c_2(s^*)\right)^{1/2} v(s^*),\end{aligned}$$

which is non-zero and finite. □

Theorem 4.15. *Let Assumption 4.2 hold, $\hat{K}$ have finite support, and the square-bandwidth matrices be $H = \hat{H} = \delta^2 I$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*)$ and $d(s^*)$ be non-zero. Then the following statements hold:*

1. *Let $q < 8$, and $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that $N\delta^q \uparrow \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^q$,*

$$\liminf_{C\uparrow\infty} C^{1/2}\text{MSE}(\hat{Z}_N) > 0.$$

2. *Let $q \geq 8$, and $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that $N\delta^q \uparrow \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^q$,*

$$\liminf_{C\uparrow\infty} C^{8/(q+8)}\text{MSE}(\hat{Z}_N) > 0.$$

Proof. We know that

$$\begin{aligned}\text{MSE}(\hat{Z}_N) &= \text{Var}\left(\hat{Z}_N\right) + \text{bias}(\hat{Z}_N)^2 \\ &= \frac{\text{Var}\left(h(\theta)\,|\, S \in B_\delta(s^*)\right)}{N\delta^q} + d(s^*)\delta^8 + \mathcal{O}\left(\delta^9\right) \\ &\geq \frac{\text{Var}\left(h(\theta)\,|\, S \in B_\delta(s^*)\right)}{N\delta^q} + \frac{d(s^*)}{2}\delta^8,\end{aligned}$$

for all sufficiently large N and $N\delta^q$.

First, we consider the case where $q \geq 8$. By Lemma A.3, the error is bounded by

$$\begin{aligned}\text{MSE}(\hat{Z}_N) &\geq \left(\frac{8}{q+8}\frac{\text{Var}\left(h(\theta)\,|\, S \in B_\delta(s^*)\right)}{N\delta^q}\right)^{8/(q+8)}\left(\frac{q}{q+8}\frac{d(s^*)}{2}\delta^8\right)^{q/(q+8)} \\ &= A(\delta)^{8/(q+8)}B^{q/(q+8)}N^{-8/(q+8)},\end{aligned}$$

where

$$A(\delta) := \frac{8}{q+8}\text{Var}\left(h(\theta)\,|\, S \in B_\delta(s^*)\right), \quad B := \frac{q}{q+8}\frac{d(s^*)}{2}.$$

For the expected cost, we have

$$N^{-1}C \geq \frac{1}{2}c_1(s^*),$$

for all sufficiently large N, and so

$$N^{-8/(q+8)}C^{8/(q+8)} \geq \left(\frac{1}{2}c_1(s^*)\right)^{8/(q+8)}.$$

Therefore, for sufficiently large N and $N\delta^q$, we have

$$C^{8/(q+8)}\text{MSE}(\hat{Z}_N) \geq A(\delta)^{8/(q+8)} B^{q/(q+8)} \left(c_1(s^*)/2\right)^{8/(q+8)}.$$

Since the right hand side is greater than zero, we have the required result.

For the case where $q < 8$, the error is bounded by

$$\text{MSE}(\hat{Z}_N) \geq \frac{\text{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right)}{N\delta^q}.$$

For the expected cost, we have

$$N^{-2}\delta^{-2q}C \geq \frac{1}{2}c_2(s^*),$$

for sufficiently large N and $N\delta^q$, and so

$$N^{-1}\delta^{-q}C^{1/2} \geq \left(\frac{1}{2}c_2(s^*)\right)^{1/2}.$$

Therefore, for sufficiently large N and $N\delta^q$, we have

$$C^{1/2}\text{MSE}(\hat{Z}_N) \geq \text{Var}\left(h(\theta) \mid S \in B_\delta(s^*)\right) \left(\frac{1}{2}c_2(s^*)\right)^{1/2}.$$

Since the right hand side is greater than zero, we have the required result. □

We now show what happens if the support of $\hat{K}$ is not finite.

Theorem 4.16. *Let Assumption 4.2 hold, $\hat{K}$ have infinite support, and the square-bandwidth matrices be $H = \hat{H} = \delta^2 I$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*) > 0$. Let $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that the limit $D := \lim_{C\uparrow\infty} N\delta^{q+8}$ is strictly positive and finite. Then the error and expected cost are such that*

$$\lim_{C\uparrow\infty} N\delta^q\,\text{MSE}(\hat{Z}_N),\ \lim_{C\uparrow\infty} N^{-(2+q/8)}\delta^{-q(2+q/8)}C,\ \textit{and}\ \lim_{C\uparrow\infty} C^{8/(q+16)}\text{MSE}(\hat{Z}_N)$$

are strictly positive and finite.

Proof. Using Equation 4.17, we find that

$$\begin{aligned}\lim_{C\uparrow\infty} N\delta^q\,\text{MSE}(\hat{Z}_N) &= \lim_{C\uparrow\infty} N\delta^q\left(\text{Var}(\hat{Z}_N) + \text{bias}(\hat{Z}_N)^2\right)\\ &= v(s^*) + d(s^*)\lim_{C\uparrow\infty} N\delta^{q+8}\\ &= v(s^*) + d(s^*)D.\end{aligned}$$

For the expected cost

$$C = c_1(s^*)N + c_2(s^*)N^2\delta^q(1 + \mathrm{o}\,(1)),$$

we find that

$$\begin{aligned}\lim_{C\uparrow\infty} (N\delta^q)^{-2-q/8}\, C &= c_1(s^*) \lim_{C\uparrow\infty} N^{-1-q/8}\delta^{-(2+q/8)q} + c_2(s^*) \lim_{C\uparrow\infty} N^{-q/8}\delta^{-(1+q/8)q} \\ &= c_1(s^*)D^{-q/8} \lim_{C\uparrow\infty} N^{-1}\delta^{-q} + c_2(s^*)D^{-q/8} \\ &= c_2(s^*)D^{-q/8}.\end{aligned}$$

Combining this result for the cost with that for the error, we get the result

$$\begin{aligned}\lim_{C\uparrow\infty} C^{8/(q+16)}\mathrm{MSE}(\hat{Z}_N) &= \lim_{C\uparrow\infty} \left((N\delta^q)^{-2-q/8}\, C\right)^{8/(q+16)} N\delta^q\, \mathrm{MSE}(\hat{Z}_N) \\ &= c_2(s^*)^{8/(q+16)} D^{-q/(q+16)} \left(v(s^*) + d(s^*)D\right),\end{aligned}$$

which is non-zero and finite. □

Theorem 4.17. *Let Assumption 4.2 hold, $\hat{K}$ have infinite support, and the square-bandwidth matrices be $H = \hat{H} = \delta^2 I$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, and $v(s^*)$ and $d(s^*)$ be non-zero. Let $N \uparrow \infty$ as the expected computational cost C tends to infinity, such that theN$\delta^q \uparrow \infty$. Then, for f_S-almost all $s^* \in \mathbb{R}^q$,*

$$\liminf_{C\uparrow\infty} C^{8/(q+16)}\mathrm{MSE}(\hat{Z}_N) > 0.$$

Proof. We know that

$$\begin{aligned}\mathrm{MSE}(\hat{Z}_N) &= \mathrm{Var}(\hat{Z}_N) + \mathrm{bias}(\hat{Z}_N)^2 \\ &= \frac{\mathrm{Var}\left(h(\theta) \,|\, S \in B_\delta(s^*)\right)}{N\delta^q} + d(s^*)\delta^8 + \mathcal{O}\left(\delta^9\right) \\ &\geq \frac{\mathrm{Var}\left(h(\theta) \,|\, S \in B_\delta(s^*)\right)}{N\delta^q} + \frac{d(s^*)}{2}\delta^8,\end{aligned}$$

for all sufficiently large N and $N\delta^q$.

By Lemma A.3, the error is bounded by

$$\begin{aligned}\mathrm{MSE}(\hat{Z}_N) &\geq \left(\frac{8}{q+16}\frac{\mathrm{Var}\left(h(\theta) \,|\, S \in B_\delta(s^*)\right)}{N\delta^q}\right)^{8/(q+16)} \left(\frac{q+8}{q+16}\frac{d(s^*)}{2}\delta^8\right)^{(q+8)/(q+16)} \\ &= A(\delta)^{8/(q+16)} B^{(q+8)/(q+16)} N^{-8/(q+16)} \delta^{64/(q+16)},\end{aligned}$$

where

$$A(\delta) := \frac{8}{q+16}\mathrm{Var}\left(h(\theta)\,|\,S \in B_\delta(s^*)\right), \quad B := \frac{q}{q+16}\frac{d(s^*)}{2}.$$

For the expected cost, we have

$$N^{-1}\delta^8 C \geq \frac{1}{2}c_2(s^*)D,$$

for all sufficiently large N, and so

$$N^{-8/(q+16)}\delta^{64/(q+16)}C^{8/(q+16)} \geq \left(\frac{1}{2}c_1(s^*)D\right)^{8/(q+16)}.$$

Therefore, for sufficiently large N and $N\delta^q$, we have

$$C^{8/(q+16)}\mathrm{MSE}(\hat{Z}_N) \geq A(\delta)^{8/(q+16)}B^{(q+8)/(q+16)}\left(c_2(s^*)D/2\right)^{8/(q+16)}.$$

Since the right hand side is greater than zero, we have the required result. □

We therefore have three sets of optimal rate of convergence for the ABCLOC estimate $\hat{Z}_N$. Firstly, suppose that $\hat{K}$ has finite support. For $q \geq 8$, the error is $\mathcal{O}\left(C^{-8/(q+8)}\right)$ as the expected cost C tends to infinity, so $\hat{Z}_N$ has a faster asymptotic rate of convergence for $q \geq 8$. For $q < 8$, the error is $\mathcal{O}\left(C^{-1/2}\right)$ as C tends to infinity. This is a faster asymptotic rate than that of the basic estimates Y_n and Z_N in the case where

$$\frac{4}{q+4} < \frac{1}{2}.$$

This holds if $q > 4$.

Now, suppose that $\hat{K}$ has infinite support. In this case, for any q, the error is $\mathrm{MSE}(\hat{Z}_N) = \mathcal{O}\left(-C^{8/(q+16)}\right)$. Since $\frac{4}{q+4} < \frac{8}{q+16}$ if $q > 8$, $\hat{Z}_N$ still has a faster rate of convergence for $q > 8$. A summary of the asymptotic rates is given in Figure 4.2.1.

If the bias and variance have the asymptotic forms that we suggest, then we can, therefore, make the following statement:

Proposition 4.18. *Let Assumption 4.2 hold, and let the square-bandwidth matrices satisfy $H = \hat{H} = \delta^2 I$. Let $\mathbb{E}\left(h(\theta)^2\right) < \infty$, $f_{S|\theta}$ have a Hessian matrix with a spectral radius bounded on the ball $B_\delta(s)$ for all possible θ and s, $v(s^*)$ be non-zero, and the bias coefficient $c(s^*)$ for the* ABCACC *and* ABCBAS *estimates be non-zero. Then the following statements hold:*

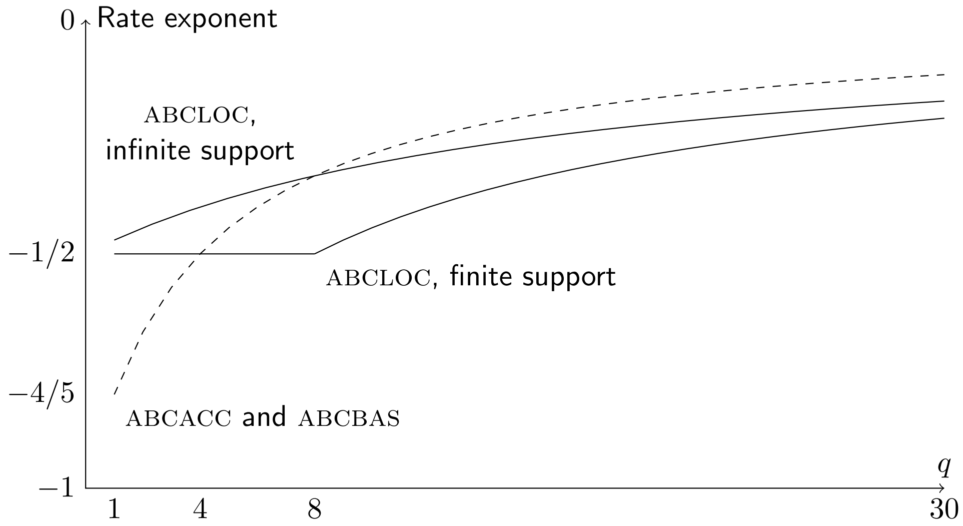

Figure 4.2.1: Rate exponents for the ABCACC and ABCBAS estimates from Section 3, and the ABCLOC estimate, against the statistic dimension q. The ABCLOC estimate has two sets of rate exponents, depending on whether the regression kernel function $\hat{K}$ has finite or infinite support. For finite support, we use the exponents from Theorems 4.14 and 4.15. For infinite support, we use the exponents from Theorems 4.16 and 4.17. A larger negative exponent indicates a faster asymptotic rate of decay for the mean square error. For comparison, exact, unbiased Monte Carlo methods have rate exponent -1. Note that the exponents for the basic estimates and ABCLOC are equal at 4 for a finite support, and at $q = 8$ for an infinite support.

1. *If $\hat{K}$ has finite support, and the limit of $N\delta^{q+\min\{q,8\}}$, as the expected cost C tends to infinity, be non-zero and finite, then the* ABCLOC *estimate has a superior asymptotic rate of convergence to the* ABCACC *and* ABCBAS *estimates if the statistic dimension is greater than four.*

2. *If $\hat{K}$ has infinite support, and the limit of $N\delta^{q+8}$, as the expected cost C tends to infinity, be non-zero and finite, then the* ABCLOC *estimate has a superior asymptotic rate of convergence to the* ABCACC *and* ABCBAS *estimates if the statistic dimension is greater than eight.*

4.2.4 Final Remarks

We finish by discussing the details of the asymptotic rates for the ABCLOC estimate: in particular, how the rate of convergence is affected by the regression kernel $\hat{K}$ and the statistic dimension q. Use of the term *dominated* in this section refers to being *asymptotically dominated*, for the sake of brevity.

First, we consider the case where the support of $\hat{K}$ is finite. In the rough solution given before the proof for Theorem 4.14, there are two asymptotic limits, $D = \lim_{C\uparrow\infty} N\delta^{q+8}$ and $D_2 = \lim_{C\uparrow\infty} N\delta^{2q}$. that must exist. The additional limit D_1 that prevents both D and D_2 from being zero.

Unless $q = 8$, only one of D and D_2 is zero. Since they are associated with the square-bias term and the regression adjustment term respectively, the value of q therefore determines which of these terms is dominated. If $q < 8$, the square-bias term is dominated, so the ABCLOC estimate functions as if it is unbiased when doing asymptotic analysis. If $q > 8$, the regression term is dominated: for asymptotic analysis, we can treat the ABCLOC estimate as an ABCBAS estimate that has a bias of order $\mathcal{O}\left(\delta^4\right)$, and therefore has a higher rate of convergence.

If the support of $\hat{K}$ is infinite, writing a rough solution for Theorem 4.16 would show that there is no equivalent to D_2: the sample generation term in the expected cost is dominated for all values of q, because the regression term is equal to the sample generation term times some factor of order $\mathcal{O}\left(N\delta^q\right)$, or $\mathcal{O}\left(n\right)$, which will tend to infinity. There is, therefore, no change in the rate of convergence dependent on the value of q. However, the cost now increases rapidly, so the result is an inferior asymptotic convergence rate.

Estimate	Kernel support	q	Dominant error term	Dominant cost term
ABCBAS	Either	Any	Both	Sample generation
ABCLOC	Finite	< 8	Variance	Both
ABCLOC	Finite	$= 8$	Both	Both
ABCLOC	Finite	> 8	Both	Sample generation
ABCLOC	Infinite	Any	Both	Regression

Table 4.2.1: Asymptotically dominant terms in the mean square error and the expected computational cost. The basic ABCBAS estimate, which has no regression cost term, is included for comparison. Note that the ABCLOC with a finite support and $q > 8$ has the same dominant terms as ABCBAS, so the only asymptotic difference is the order of the bias.

We can also consider the case where the kernel is used for weighting the proposals, as in Beaumont et al. [2002] and Blum [2010], rather than for random acceptance. The effect depends on the support of the kernel. If the support is finite, the number and cost of regressions performed are still $\mathcal{O}\left(N\delta^q\right)$, so this is equivalent to random acceptance with a finite support. However, if the support is infinite, all of the proposals are adjusted and used in the estimate. In this case, both the number of regressions and the number of samples used in each regression are $\mathcal{O}\left(N\right)$, rather than $N\delta^q$, and the expected cost is dominated by a $\mathcal{O}\left(N^2\right)$ regression term. The resulting asymptotic error is of order $\mathcal{O}\left(C^{-4/(q+8)}\right)$, which is asymptotically inferior to the ABCBAS estimate. This further motivates the use of a kernel with finite support, such as the uniform and Epanechnikov kernels, rather than one with infinite support, such as the Gaussian kernel, for estimating the posterior mean.

An overview of which terms are asymptotically dominant, under which circumstances, when random acceptance is used, is given in Table 4.2.1.

4.3 Discussion

4.3.1 Comparison to Previous Results

We now compare the results to the theoretical results from Section 4.2.3 to theoretical results for similar variants, described in Section 2.2.4.

Beaumont et al. [2002] and Blum [2010] used a similar approach to do proposal adjustment, with Blum using the adjusted proposals to do kernel posterior density estimation, as described in Section 2.1.4. For this section, we refer to the ABC estimate with this version of proposal adjustment as the ABCREG estimate.

The most significant methodological difference is that ABCREG uses a single regression, centred at s^*, for all the adjustments. As previously mentioned, from Ruppert and Wand [1994], the bias of regression points that are not at the centre of the regression, for $\hat{H} = \delta^2 I$, is $\mathcal{O}\left(\delta\right)$, rather than $\mathcal{O}\left(\delta^2\right)$. Therefore, the asymptotic bias of the adjusted proposals will not be higher-order, and might be worse, than the bias of the non-adjusted proposals if the posterior expectation function m is highly non-linear.

There are two other main differences between the ABCLOC variant and the variant used by Beaumont et al. [2002] and Blum [2010], which we expect to have no asymptotic effect:

1. Beaumont et al. [2002] and Blum [2010] use the kernel for regression and weighting, whereas we use the kernel for regression and random acceptance. In Section 4.2.4, we discussed the asymptotic effect on the ABCLOC estimate of this change in the use of the kernel. However, this is dependent on the use of multiple regressions, whereas the variants in Beaumont et al. [2002] and Blum [2010] only use one. In this case, weighting and random acceptance are asymptotically equivalent if done with the same kernel and the same square-bandwidth matrix, so this difference in usage is not expected to have a significant effect.

2. For small values of q, the rate of convergence for the ABCLOC estimate is different for a kernel with finite support, due to the computational cost added by the use of regression. The effect of the computational cost is not accounted for in the asymptotic analysis in Blum [2010]. However, accounting for the cost does not have any asymptotic effect: since there is only one regression, the regression term in the computation cost is either $\mathcal{O}\left(N\delta^2\right)$ or $\mathcal{O}\left(N\right)$, depending on whether the support of $\hat{K}$ is finite or infinite. In both cases, the expected cost is still $\mathcal{O}\left(N\right)$. Since Blum found the error to be $\mathcal{O}\left(N^{-4/(q+5)}\right)$, the error in terms of expected

cost is therefore $\mathcal{O}\left(C^{-4/(q+5)}\right)$, as expected.

Blum [2010] discusses how single-regression proposal adjustments can make an ABC estimate worse, rather than better. This comes from examining the leading term in the asymptotic expansion of the error. While it is possible that the estimate can also be worse from the adjustment increasing the cost, as it can be in ABCLOC, this would be due to differences in the leading term, rather than in the asymptotic order.

4.3.2 Practical Use

While the ABCLOC estimate has a higher asymptotic rate of convergence than the basic ABC variants for large q, it is expected that, for practical expected computational costs, ABCLOC's performance is likely to be much worse. This is because, before the estimate runs long enough for the error reduction to take effect, the main effect of using ABCLOC is the large increase in the computation cost due to the regressions. However, this needs to be investigated with numerical experiments.

Chapter 5

ABC for Infinite-Dimensional Statistics

In Chapter 3, we saw that the ABCBAS estimate Z_N from Algorithm 2.1.3, that uses a sufficient statistic of dimension q, and runs for time C, has a mean square error of order $\mathcal{O}\left(C^{-4/(q+4)}\right)$ as K tends to infinity. Considered naïvely, this might lead to the conclusion that using a statistic of infinite dimension prevents the algorithm from converging as the cost increases. However, the constant $\lim_{C\uparrow\infty} C^{4/(q+4)}\text{MSE}(Z_N)$ in Equation 3.6 depends on the bias coefficient $c(s^*)$, and therefore on q. At the limit $q \uparrow \infty$, there might, therefore, be cases where the asymptotic convergence rate is different to the rate given in Theorem 3.13. This chapter looks at such a problem.

5.1 Problem and Exact Inference

For finite-dimensional statistics, we consider a set of parameters $\theta \in \mathbb{R}^p$, and a set of observational data x^* with summary statistic $s^* = S(x^*) \in \mathbb{R}^q$. Here, we consider a set of observational data $x^* \in C([0,1], \mathbb{R})$, that is a continuous process on the time interval $[0,1]$. Specifically, we observe a diffusion process that satisfies the equations

$$\mathrm{d}X_t = \theta \mathrm{d}t + \mathrm{d}B_t, \quad X_0 = 0,$$

where the drift rate θ is unknown, with a normal prior distribution, and B_t is a standard Brownian motion. The inference for this problem can be done exactly, as described below.

Lemma 5.1. *If we last observe the process* $\mathrm{d}X_t = \theta\mathrm{d}t + \mathrm{d}B_t$ *at time* T, *and the drift rate* θ *of the linear trend has a simple normal prior distribution, then the trend* θ *has the normal posterior distribution*

$$\theta \mid X \sim \mathrm{N}\left(X_T/(T+1), 1/(T+1)\right).$$

This result holds for both a finite-dimensional observation, where the process is observed at certain time points, including time T, *and a infinite-dimensional observation, consisting of the entire process up to time* T.

Proof. For the finite-dimensional case, we have the linear filtering problem

$$X_{t_{n+1}} = X_{t_n} + \theta(t_{n+1} - t_n) + \epsilon_{n+1}, \quad \epsilon_{n+1} \sim \mathrm{N}\left(0, t_{n+1} - t_n\right),$$

where $t_1 < t_2 < \ldots < t_q = T$ are the q time points at which the process is observed. Thus, the sequence of posterior expectations $\tilde{\theta}_n = \mathbb{E}\left(\theta \mid t_1, \ldots, t_n\right)$ of the drift rate satisfies the recurrence relation [Williams, 1991]

$$\frac{\tilde{\theta}_n}{V_n} = \frac{\tilde{\theta}_{n-1}}{V_{n-1}} + (X_n - X_{n-1}), \quad \tilde{\theta}_0 = 0,$$

where the sequence of posterior varianes $V_n = \mathbb{E}\left((\theta - \tilde{\theta}_n)^2 \mid t_1, \ldots, t_n\right)$ satisfies the recurrence relation

$$\frac{1}{V_n} = \frac{1}{V_{n-1}} + t_n - t_{n-1}, \quad V_0 = 1.$$

For the infinite-dimensional case, we have the linear filtering problem

$$\mathrm{d}X_t = \theta\,\mathrm{d}t + \mathrm{d}B_t,$$

and our posterior expectation $\bar{\theta}$ of the drift rate obeys the stochastic differential equation [Øksendal, 2003, Example 6.2.9]

$$\mathrm{d}\bar{\theta}_t = -W_t\bar{\theta}_t\,\mathrm{d}t + W_t\,\mathrm{d}X_t, \quad \bar{\theta}_0 = 0,$$

where $W_t = \mathbb{E}\left((\theta - \bar{\theta}_t)^2\right)$ obeys the equation

$$W_t' = -W_t^2,$$

with initial condition $W_0 = 1$ determined from the prior variance.

The equation for V_n can be written as $\Delta\frac{1}{V_n} = \Delta t_n$, and applying the discrete anti-derivative gives the solution $\frac{1}{V_n} - \frac{1}{V_0} = t_n - t_0$. Similarly, the differential equation for W_t can be written as

$$-\frac{\mathrm{d}W_t}{W_t^2} = \mathrm{d}t,$$

and integrating gives the solution $\frac{1}{W_t} - \frac{1}{W_0} = t$. Therefore, the variance behaves as either $V_n = 1/(t_n + 1)$ or $W_t = 1/(t + 1)$, and the solution for $\tilde{\theta}$ is

$$\tilde{\theta}_n = \frac{X_{t_n}}{t_n + 1}, \quad \text{or} \quad \bar{\theta}_t = \frac{X_t}{t + 1}.$$

In either case, θ thus has posterior $\theta \,|\, X \sim \mathrm{N}\left(\frac{X_T}{T+1}, \frac{1}{T+1}\right)$. □

By Lemma 5.1, letting $T = 1$, the problem has a one-dimensional sufficient statistic X_1. We could also form a summary statistic of arbitrary dimension, by taking additional observations before time 1.

While, in practice, an infinite-dimensional observation is computationally infeasible, considering it is useful to see what to expect as the dimension of the summary statistic tends to infinity.

The structure of this chapter is as follows: we first discuss two common choices of norm that can be used to determine whether a proposal is accepted. We then present asymptotic results for each norm separately. For one of these choices, we construct a sequence of inference problems, such that the observed process $x^*_{(q)}$ for problem q has a q-dimensional summary statistic. This sequence converges to the original problem, with observation $x^* = \lim_{q\uparrow\infty} x^*_{(q)}$. We find the asymptotic bias for a problem with finite q, then let q tend to infinity.

5.1.1 Choice of Acceptance Criterion

To use ABC in an infinite-dimensional space, we must reconsider when to accept proposals. In the finite-dimensional case, we accept a proposal with a q-dimensional statistic s if $\|s - s^*\|_2 \leq \delta$, where $\|\cdot\|_2$ is the Euclidean norm. In Section 3.4.4, we showed the effect on the asymptotic bias when using a different norm: the effect was relatively small, since different $\mathcal{L}$-norms are equivalent in finite dimensions. For infinite dimensions, this no longer holds, so the choice has more effect.

We can again use the Euclidean norm on the interval $[0, 1]$, which, for the infinite-dimensional case, is equal to

$$\|X\|_2 = \left(\int_0^1 x_t^2 \,\mathrm{d}t\right)^{1/2},$$

where x_t is the value of the process X at time t, and then accept a path X if the integral of its square distance from the original motion x^* up to time t is no greater than δ^2 :

$$\|X - x^*\|_2 \leq \delta.$$

An alternative would be to use the supremum norm, and accept a proposal if its generated process stays within a ball around the observed process in the interval $[0, 1]$:

$$\|X - x^*\|_\infty := \max\{|x_t - x_t^*| : t \in [0, 1]\} \leq \delta.$$

These two norms have different advantages when calculating asymptotic results for the estimate. We will therefore consider the use of both norms separately.

5.2 Asymptotic Results for the Supremum Norm

In this section, we look at the asymptotic results we have obtained for the case where we accept using the supremum norm. Accepted processes are thus always within a certain distance of the observed process x^*, staying inside an envelope around it.

5.2.1 Convergence Conditions

We begin by looking for conditions for the ABCBAS estimate Z_N to converge as the computation cost $C \uparrow \infty$ tends to infinity, similarly to Theorem 3.2.

Theorem 5.2. *Let the function* $h \colon \mathbb{R}^p \to \mathbb{R}$ *be such that* $\mathbb{E}(|h(\theta)|) < \infty$, *and the likelihood function* $F_{S|\theta}(s \,|\, t)$ *be bounded over all* t *and all* $s \in B_\delta(s^*)$ *for sufficiently small* δ^*. *Then, for* f_S*-almost all* $s^* \in \mathbb{R}^p$, *the* ABCBAS *estimate* Z_N *satisfies*

1. $\lim_{N\to\infty} Z_N = \mathbb{E}(Z_N \,|\, n > 0)$ *almost surely for all* $\delta > 0$*; and*

2. $\lim_{\delta\downarrow 0} \mathbb{E}(Z_N \,|\, n > 0) = m(s^*)$ *for all* $N \in \mathbb{N}$,

where n *is the number of accepted proposals.*

Proof. As N tends to infinity, the probability of no proposals being accepted tends to zero. Therefore, since $\mathbb{E}(|h(\theta)|) < \infty$, we have

$$\mathbb{E}(|Z_N|) \leq \mathbb{E}(|h(\theta)| \mid s^*) = \frac{\phi_{|h|}(s^*)}{\phi_1(s^*)} < \infty$$

whenever $\phi_1(s^*) = f_S(s^*) > 0$, and, by the law of large numbers, Z_N converges to $\mathbb{E}\left(Z_N \,|\, n > 0\right)$ almost surely.

For the second statement, we first define

$$\hat{\phi}_h(s) := m(s) f_{S_1}(s_1)/f_{S_1|\theta}(s_1 \,|\, 0) = \int h(t) f_\theta(t) \exp\left(-\frac{1}{2}t^2 + s_1 t\right) \,\mathrm{d}t,$$

and

$$\hat{\phi}_h^{(\delta)}(s^*) := \int h(t) f_\theta(t) \frac{\mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = t\right)}{\mathbb{P}\left(\|B - s^*\|_\infty \le \delta\right)} \,\mathrm{d}t,$$

so that

$$m(s^*) = \frac{\hat{\phi}_h(s^*)}{\hat{\phi}_1(s^*)}, \quad \mathbb{E}\left(h(\theta) \mid \|S - s^*\|_\infty \le \delta\right) = \frac{\hat{\phi}_h^{(\delta)}(s^*)}{\hat{\phi}_1^{(\delta)}(s^*)}.$$

It is then sufficient to show that

$$\lim_{\delta \downarrow 0} \hat{\phi}_h^{(\delta)}(s^*) = \hat{\phi}_h(s^*).$$

To show this, we note that the absolute difference $d(s^*)$ between $\hat{\phi}_h^{(\delta)}(s^*)$ and $\hat{\phi}_h(s^*)$ satisfies

$$\begin{aligned} d(s^*) &= \left|\hat{\phi}_h^{(\delta)}(s^*) - \hat{\phi}_h(s^*)\right| \\ &= \left|\int h(t) f_\theta(t) \left(\frac{\mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = t\right)}{\mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = 0\right)} - \exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\right) \mathrm{d}t\right| \\ &\le \int |h(t)| f_\theta(t) \left|\frac{\mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = t\right)}{\mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = 0\right)} - \exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\right| \,\mathrm{d}t. \end{aligned}$$

Using Theorem A.13, we observe that

$$\begin{aligned} \mathbb{P}\left(\|S - s^*\|_\infty \le \delta \,|\, \theta = t\right) &= \mathbb{E}_{\mathbb{Q}}\left([\|S - s^*\|_\infty \le \delta] \exp\left(-\frac{1}{2}t^2 + S_1 t\right)\right) \\ &= \mathbb{E}\left([\|B - s^*\|_\infty \le \delta] \exp\left(-\frac{1}{2}t^2 + B_1 t\right)\right), \end{aligned}$$

where $\mathbb{Q}$ is the measure under which S is a martingale, and B is standard Brownian motion. Since $\|B - s^*\| \le \delta$ requires that $|B_1 - s_1^*| \le \delta$, it follows that

$$\left|\exp\left(-\frac{1}{2}t^2 + B_1 t\right) - \exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\right| \le \exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\left(e^{\delta|t|} - 1\right).$$

Specifically, there is some random variable

$$\eta(\delta, t) = \int_{C([0,1],\mathbb{R})} [\|B - s^*\|_\infty \le \delta] \left|e^{\delta|t|} - 1\right| \,\mathrm{d}\mathbb{W}(\omega),$$

that is bounded for small δ and tends to zero as δ tends to zero, such that

$$\left|\frac{\mathbb{P}\left(\|S-s^*\|_\infty \leq \delta \,|\, \theta = t\right)}{\mathbb{P}\left(\|S-s^*\|_\infty \leq \delta \,|\, \theta = 0\right)} - \exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\right| \leq \eta(\delta, t)\exp\left(-\frac{1}{2}t^2 + s_1^* t\right).$$

Therefore,

$$\lim_{\delta\downarrow 0}\left|\hat{\phi}_h^{(\delta)}(s^*) - \hat{\phi}_h(s^*)\right| \leq \lim_{\delta\downarrow 0}\int |h(t)| f_\theta(t)\eta(\delta,t)\exp\left(-\frac{1}{2}t^2 + s_1^* t\right)\,\mathrm{d}t$$

The integrand is bounded above by

$$g(t) := |h(t)| f_\theta(t)\exp\left(-\frac{1}{2}t^2 + (s_1^* + 1)t\right)$$

for all $\delta < 1$ and all t. Since $\int |g(t)|\,\mathrm{d}t < \infty$ for f_S-almost all s^*, the result follows from the dominated convergence theorem. □

5.2.2 Asymptotic Variance

For the asymptotic variance, we require the value of the acceptance probability $\mathbb{P}\left(\|X\|_\infty \leq \delta\right)$. In the case where the sample process is a simple Brownian motion B with no trend, this satisfies [Li and Shao, 2001, Theorem 6.3]

$$\lim_{\delta\downarrow 0}\delta^2 \log \mathbb{P}\left(\|B\|_\infty \leq \delta\right) = -\pi^2/8.$$

In particular, the acceptance probability in this case is equal to [Feller, 1968]

$$\mathbb{P}\left(\|B\|_\infty \leq \delta\right) = \frac{4}{\pi}\sum_{k\geq 0}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8\delta^2}\right). \tag{5.1}$$

For general X, we can use the following lemma.

Lemma 5.3. *Let X be the motion $X_t = B_t + \theta t$, a simple Brownian motion plus a linear trend. Then the probability that $\|X\| \leq \delta$ is equal to*

$$\mathbb{P}\left(\|X\|_\infty \leq \delta \,|\, \theta\right) = \exp\left(-\frac{1}{2}\theta^2\right)\mathbb{E}\left([\|B\|_\infty \leq \delta]\exp\left(\theta B_1\right)\right).$$

Proof. By the Girsanov theorem (Theorem A.13), X is a martingale with respect to the measure

$$\mathrm{d}\mathbb{Q} = \exp\left(-\theta B_1 - \frac{1}{2}\theta^2\right)\,\mathrm{d}\mathbb{P}.$$

Therefore, the acceptance probability is equal to

$$\begin{aligned}
\mathbb{P}\left(\|X\|_\infty \leq \delta \,|\, \theta\right) &= \mathbb{E}\left([\|X\|_\infty \leq \delta]\right)\\
&= \mathbb{E}_\mathbb{Q}\left([\|X\|_\infty \leq \delta]\exp\left(\theta B_1 + \frac{1}{2}\theta^2\right)\right)\\
&= \exp\left(\frac{1}{2}\theta^2\right)\mathbb{E}_\mathbb{Q}\left([\|X\|_\infty \leq \delta]\exp\left(\theta B_1\right)\right)\\
&= \exp\left(-\frac{1}{2}\theta^2\right)\mathbb{E}_\mathbb{Q}\left([\|X\|_\infty \leq \delta]\exp\left(\theta X_1\right)\right),
\end{aligned}$$

where $\mathbb{E}_{\mathbb{Q}}$ is the expectation with respect to $\mathbb{Q}$, rather than $\mathbb{P}$. Since X is a martingale with respect to $\mathbb{Q}$,

$$\mathbb{P}\left(\|X\|_\infty \leq \delta \mid \theta\right) = \exp\left(-\frac{1}{2}\theta^2\right) \mathbb{E}\left([\|B\|_\infty \leq \delta] \exp\left(\theta B_1\right)\right),$$

as required. □

Further analysis, that accounts for staying near s^* rather than 0, is difficult, as adjusting for an additional trend process with Girsanov requires the process to be differentiable. The process s^* will almost never be differentiable, due to its being a Brownian path, so the acceptance probability cannot account for s^* by use of Girsanov.

However, we can more easily make progress if we consider the asymptotic behaviour of the ABC estimate over all values of s^*. In this case, we consider the mean acceptance probability $\bar{p}$, and we can show the following results.

Corollary 5.4. *Let X_t be the motion $X_t = B_t + \theta t$, a simple Brownian motion plus a linear trend, and X_t^* be the motion $X_t^* = B_t^* + \theta^* t$, where θ θ^* have independent simple normal distributions, and B_t and B_t^* are independent Brownian motions. Then the probability that $\|X - X^*\|_\infty \leq \delta$ is equal to*

$$\begin{aligned}\bar{p} := \mathbb{P}\left(\|X - X^*\|_\infty \leq \delta\right) = \mathbb{P}\left(\|B\|_\infty \leq \frac{\delta}{2}\right) \\ = \frac{4}{\pi}\sum_{k\geq 0}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{2\delta^2}\right).\end{aligned}$$

Proof. The processes $X_t/\sqrt{2}$ and $X_t^*/\sqrt{2}$ are independent Brownian motions, and so the process $Y_t := \left(X_t - X_t^*\right)/2$ is a Brownian motion. Since the required probability is equal to $\mathbb{P}\left(\|Y\|_\infty \leq \delta/4\right)$, the result follows from Equation 5.1. □

Lemma 5.5. *The mean acceptance probability $\bar{p}$ has upper bound*

$$\bar{p} \leq \frac{4}{\pi}\exp\left(-\frac{\pi^2}{2\delta^2}\right) + \left(1 - \frac{4}{\pi}\right)\exp\left(-\frac{9\pi^2}{2\delta^2}\right).$$

Proof. The series in Corollary 5.4 has upper bound

$$\sum_{k\geq 0}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{2\delta^2}\right) \leq \exp\left(-\frac{\pi^2}{2\delta^2}\right) + \exp\left(-\frac{9\pi^2}{2\delta^2}\right)\sum_{k\geq 1}\frac{(-1)^k}{2k+1}.$$

Since $\sum_{k\geq 0}(-1)^k/(2k+1) = \pi/4$, the result follows. □

5.3 Asymptotic Results for the Euclidean Norm

In this section, we look at the asymptotic results we have obtained for the case where we accept using the Euclidean norm. Here, we approach the bias of the ABCBAS estimate Z_N by constructing a series of problems P_q, which use a q-dimensional summary statistic, finding the bias for P_q with finite q, and then letting $q \uparrow \infty$.

5.3.1 Choice of Summary Statistic

We will consider the asymptotic error for a sequence of finite-dimensional problems, which tend to the infinite-dimensional problem. This requires a choice of a sequence of summary statistics with increasing dimension.

The simple choice would be for each statistic to be a set of observations at evenly-distributed points. While this allows the use of the asymptotic results from Chapter 3, we would have to account for the finite-dimensional Euclidean norm not tending to the infinite-dimensional Euclidean norm. Specifically, if we let $r_{(q)} := (x_{1/q}, x_{2/q}, \dots, x_1)$ be the resulting q-dimensional statistic, then the limit of the sequence of Euclidean distances is

$$\lim_{q\uparrow\infty} \|r_{(q)}\|_2 = \lim_{q\uparrow\infty} \left(\sum_{k=1}^{q} x_{k/q}^2 \right)^{1/2},$$

which is not equal to

$$\|X\|_2 = \left(\int_0^1 x_t^2 \,\mathrm{d}t \right)^{1/2} = \lim_{q\uparrow\infty} \left(\sum_{k=1}^{q} \frac{x_{k/q}^2}{q} \right)^{1/2},$$

the infinite-dimensional Euclidean distance. In particular, at the limit $q \uparrow \infty$, $\|s_{(q)}\|_2$ would almost surely be infinitely large, so the ABC algorithm would almost always reject proposals. For the algorithm to scale properly to the infinite-dimensional case, we would therefore need to introduce an adjustment factor $1/q$ to the norm.

To avoid this adjustment factor, we instead choose the Karhunen-Loève decomposition, which we introduce now.

By the Mercer theorem, for any stochastic process X with an autocovariance function K, we can construct a decomposition

$$X_t = \sum_{k\geq 1} \alpha_k \psi_k(t),$$

where $(\psi_k(t))_{k\geq 0}$ is a sequence of orthogonal functions with respect to some inner product. If X has a covariance function $K(\cdot,\cdot)$, we can choose these to be the eigenfunctions of K, satisfying

$$\int_0^1 K(s,t)\psi_k(s)\,\mathrm{d}s = \lambda_k\psi_k(t).$$

If we define the inner product

$$\langle f, g\rangle := \int_0^1 f(s)g(s)\,\mathrm{d}s$$

on $\mathcal{L}^2[0,1]$, and let $V_t(s) := K(s,t)$, then we can rewrite $\int_0^1 K_(s,t)\psi_k(s)\,\mathrm{d}s$ as an inner product,

$$\langle V_t, \psi_k\rangle = \lambda_k\psi_k(t),$$

and $\lambda_k\psi_k(0) = 0$ in particular, since $V_0(\cdot) = 0$. The coefficients are then equal to

$$\alpha = \langle X, \psi_k\rangle.$$

Example 5.6. *In the case of the motion X being a Brownian motion plus a fixed linear trend θ, X has covariance function*

$$V_t(s) = \min(s,t) = s + (t-s)H_t(s),$$

where $H_t(s) = H(s-t)$ is the Heaviside step function. Differentiating once with respect to t, we then see that

$$\frac{\partial}{\partial t}V_t(s) = H_t(s) - (t-s)\delta(s-t).$$

Since the latter term disappears inside the inner product, we have

$$\lambda_k\psi_k'(t) = \langle H_t, \psi_k\rangle, \quad \lambda_k\psi_k'(1) = 0.$$

Differentiating again, and ignoring terms that disappear in the inner product, we obtain

$$\lambda_k\psi_k''(t) = \langle \delta_t, \psi_k\rangle = \psi_k(t),$$

where $\delta_t(s) = \delta(s-t)$. Therefore, $\psi_k(t) = A\sin\left(\frac{t}{\sqrt{\lambda_k}}\right)$, where

$$A\sqrt{\lambda_k}\cos\left(\frac{1}{\sqrt{\lambda_k}}\right) = 0,$$

by the boundary condition on the first derivative, and so

$$\lambda_k = \left(\frac{1}{(k-1/2)\pi}\right)^2 \quad \textit{for } k \geq 1. \tag{5.2}$$

To have $\langle \psi_k, \psi_k \rangle = 1$, *we then require that* $A = \sqrt{2}$. *Therefore, we have eigenfunctions*

$$\psi_k(t) = \sqrt{2} \sin\left((k - 1/2)\pi t\right), \quad k \geq 1. \tag{5.3}$$

Since the observations have a linear trend $\theta\xi(t)$, *where*

$$\xi(t) := t, \tag{5.4}$$

the coefficients α_k *have conditional expectation*

$$\begin{aligned} \mathbb{E}\left(\alpha_k \,|\, \theta\right) &= \mathbb{E}\left(\langle X, \phi_k \rangle \,|\, \theta\right) \\ &= \langle \mathbb{E}\left(X \,|\, \theta\right), \phi_k \rangle \\ &= \theta \langle \xi, \phi_k \rangle \\ &= (-1)^{k-1} \theta \sqrt{2} \lambda_k, \end{aligned}$$

and conditional variance

$$\begin{aligned} \mathbb{E}\left((\alpha_j - \mathbb{E}\left(\alpha_j\right))(\alpha_k - \mathbb{E}\left(\alpha_k\right)) \,|\, \theta\right) &= \mathbb{E}\left(\langle B, \phi_j \rangle \langle B, \phi_k \rangle\right) \\ &= \mathbb{E}\left(\int_0^1 \int_0^1 B_s B_t \phi_j(s) \phi_k(t) \,\mathrm{d}s \,\mathrm{d}t\right) \\ &= \int_0^1 \int_0^1 \mathbb{E}\left(B_s B_t\right) \phi_j(s) \phi_k(t) \,\mathrm{d}s \,\mathrm{d}t \\ &= \int_0^1 \int_0^1 K(s,t) \phi_j(s) \phi_k(t) \,\mathrm{d}s \,\mathrm{d}t \\ &= \int_0^1 \lambda_j \phi_j(t) \phi_k(t) \,\mathrm{d}t \\ &= \lambda_j [j = k]. \end{aligned}$$

Therefore, the coefficients are independent, given the parameter value θ, *and have the distributions*

$$\alpha_k \,|\, \theta \sim \mathrm{N}\left((-1)^{k-1} \sqrt{2} \theta \lambda_k, \lambda_k\right).$$

We can now form a sequence of problems, where the the q-dimensional problem has generated statistics consisting of the first q coefficients α_k.

Definition 5.7. *The* q-dimensional spectral problem P_q *is the problem of finding the posterior distribution* $\theta \,|\, X_{(q)} = x^*_{(q)}$, *given some prior distribution, where the observed process* $x^*_{(q)}$ *is the spectral approximation*

$$x^*_{(q)}(t) := \sum_{k=1}^{q} \alpha^*_k \psi_k(t)$$

of x^*, and the functions ψ_k are defined in Equation 5.3. This process has sufficient statistic

$$s^*_{(q)} := (\alpha^*_1, \dots, \alpha^*_q)^T,$$

whose elements α^*_k are independent, and have conditional distributions

$$\alpha^*_k \,|\, \theta \sim N((-1)^{k-1}\sqrt{2}\theta\lambda_k, \lambda_k),$$

where $\lambda_k = ((k-1/2)\pi)^{-2}$. Data samples $s_{(q)}$ are then generated from the same distribution. The original problem of finding the distribution for $\theta \,|\, X = x^*$ is written as $P = \lim_{q\uparrow\infty} P_q$.

Using this decomposition has two advantages.

1. Conditionally on θ, the coefficients α_k are independent of each other, and of q, and have a simple distribution. The observed statistic for problem P_{q+1} will thus be that for the previous problem P_q, plus a new element α^*_{q+1} that is independent of the previous ones.

2. The acceptance criterion, when using the 2-norm, has a simple definition in terms of the Karhunen-Loève coefficients, since the distance between the two truncated processes is equal to

$$\begin{aligned} \|X_{(q)} - x^*_{(q)}\|_2 &= \langle X_{(q)} - x^*_{(q)}, X_{(q)} - x^*_{(q)} \rangle^{1/2} \\ &= \left(\sum_{k=1}^{q} (\alpha_k - \alpha^*_k)^2 \right)^{1/2} \\ &= \|s_{(q)} - s^*_{(q)}\|_2. \end{aligned}$$

   Therefore, if we use the Karhunen-Loève coefficients as the summary statistic, the acceptance condition is the same as for the ABCBAS estimate Z_N.

We now look for the sequence of true posterior distributions for the sequence $(\theta \,|\, s^*_{(q)})_q$.

Lemma 5.8 (Exact inference for K-L approximation). *Let $\theta \sim \mathrm{N}\left(\mu_0, \sigma_0^2\right)$, and $x^*_{(q)}$ be a process on the time interval $[0,1]$ whose value at time t is equal to*

$$x^*_{(q)}(t) = \sum_{k=1}^{q} \alpha^*_k \psi_k(t),$$

where $\psi_k(t)$ are defined in Equation 5.3, and the α_k^* are independent conditional on θ, with conditional distributions

$$\alpha_k^* \,|\, \theta \sim \mathrm{N}\left((-1)^{k+1}\sqrt{2}\theta\lambda_k, \lambda_k\right), \tag{5.5}$$

where λ_k are defined in Equation 5.2, and have sum

$$L_q := \sum_{k=1}^{q} \lambda_k. \tag{5.6}$$

Then θ has posterior distribution

$$\theta \,|\, s_{(q)}^* \sim \mathrm{N}\left(\mu_q, \sigma_q^2\right),$$

where

$$\sigma_q^{-2} = \sigma_0^{-2} + 2L_q, \quad \mu_q\sigma_q^{-2} = \mu_0\sigma_0^{-2} + x_{(q)}^*(1). \tag{5.7}$$

Proof. The log-likelihood for the sufficient statistic $s_{(q)} = (\alpha_1, \dots, \alpha_q)^T$ is equal to

$$\begin{aligned}
\log f_{S|\theta}(s_{(q)} \,|\, t) &= c_1 - \frac{1}{2}\sum_{k=1}^{q} \frac{(\alpha_k - (-1)^{k+1}\sqrt{2}t\lambda_k)^2}{\lambda_k} \\
&= c_1 - \frac{1}{2}\sum_{k=1}^{q} 2\lambda_k(t - (-1)^{k+1}\alpha_k/\sqrt{2}\lambda_k)^2 \\
&= c_2 - \frac{1}{2}(2L_q)\left(t - (2L_q)^{-1}\sum_{k=1}^{q}(-1)^{k+1}\sqrt{2}\alpha_k\right)^2,
\end{aligned}$$

by Lemma A.1, for some c_1 and c_2 that are constant with respect to t. Since $\psi_k(1) = (-1)^{k+1}\sqrt{2}$, this is equal to

$$\log f_{S|\theta}(s_{(q)} \,|\, t) = c_2 - \frac{1}{2}(2L_q)\left(t - (2L_q)^{-1}x_{(q)}^*(1)\right)^2,$$

and so $S \,|\, \theta \sim \mathrm{N}\left(x_{(q)}^*(1)/2L_q, 1/2L_q\right)$. Therefore, θ has posterior distribution

$$\theta \,|\, s_{(q)}^* \sim \mathrm{N}\left(\mu_q, \sigma_q^2\right),$$

where

$$\sigma_q^{-2} := \sigma_0^{-2} + 2L_q, \quad \mu_q\sigma_q^{-2} := \mu_0\sigma_0^{-2} + \sqrt{2}\sum_{k=1}^{q}(-1)^{k+1}\alpha_k.$$

Since $\psi_k(1) = (-1)^{k+1}\sqrt{2}$, the result follows. □

Note that the endpoint value of the process $x_{(q)}$ is still a minimal sufficient statistic.

5.3.2 Bias

We consider the sequence of biases. The asymptotic bias for problem q is equal to

$$\begin{aligned}\mathrm{bias}(Z_N^{(q)}) &= \mathbb{E}\left(h(\theta)\,|\,s_{(q)} \in B_\delta(s^*_{(q)})\right) - \mathbb{E}\left(h(\theta)\,|\,s_{(q)} = s^*_{(q)}\right)\\ &= \frac{\phi_h^{(\delta)}(s^*_{(q)})}{\phi_1^{(\delta)}(s^*_{(q)})} - \frac{\phi_h(s^*_{(q)})}{\phi_1(s^*_{(q)})},\end{aligned}$$

where ϕ_h and $\phi_h^{(\delta)}$ are defined in Definition 3.1. By Theorem 3.4, for finite values of q, this satisfies

$$\lim_{\delta\downarrow 0} \delta^{-2}\mathrm{bias}(Z_N^{(q)}) = c(s^*_{(q)}),$$

with bias coefficient

$$c(s^*_{(q)}) = \frac{\Delta\phi_h(s^*_{(q)}) - m(s^*_{(q)})\Delta\phi_1(s^*_{(q)})}{2(q+2)\phi_1(s^*_{(q)})}.$$

To find the Laplacian for ϕ_h, we recall that

$$\alpha_k\,|\,\theta \sim \mathrm{N}\left((-1)^{k+1}\sqrt{2}\theta\lambda_k, \lambda_k\right).$$

Therefore, ϕ_h has second derivatives

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_k^2}\phi_h(s_{(q)}) &= \frac{\partial^2}{\partial\alpha_k^2}\int h(t)p_\theta(t)p_{S|\theta}(s_{(q)}\,|\,t)\,\mathrm{d}t\\ &= \int h(t)p_\theta(t)\left(\left(\frac{\alpha_k - (-1)^{k+1}\sqrt{2}\lambda_k t}{\lambda_k}\right)^2 - \frac{1}{\lambda_k}\right)p_{S|\theta}(s_{(q)}\,|\,t)\,\mathrm{d}t\\ &= \left(\frac{\alpha_k^2}{\lambda_k^2} - \frac{1}{\lambda_k}\right)\phi_h(s_{(q)}) - 2\sqrt{2}(-1)^{k+1}\frac{\alpha_k}{\lambda_k}\phi_g(s_{(q)}) + 2\phi_f(s_{(q)}),\end{aligned}$$

where $g(t) := th(t)$ and $f(t) := t^2h(t)$. The Laplacian for ϕ_h is then equal to

$$\begin{aligned}\triangle\phi_h(s_{(q)}) = &\sum_{k=1}^q \left(\frac{\alpha_k^2}{\lambda_k^2} - \frac{1}{\lambda_k}\right)\phi_h(s_{(q)}) - 2\sqrt{2}\sum_{k=1}^q (-1)^{k-1}\frac{\alpha_k}{\lambda_k}\phi_g(s_{(q)})\\ &+ 2q\phi_f(s_{(q)}),\end{aligned}$$

and so the bias coefficient $c(s^*_{(q)})$ has numerator

$$\begin{aligned}\triangle\phi_h(s^*_{(q)}) - m(s^*_{(q)})\triangle\phi_1(s^*_{(q)}) = &\sum_{k=1}^q \left(\frac{\alpha_k^2}{\lambda_k^2} - \frac{1}{\lambda_k}\right)\left(\phi_h(s^*_{(q)}) - m(s^*_{(q)})\phi_1(s^*_{(q)})\right)\\ &- 2\sqrt{2}\sum_{k=1}^q (-1)^{k-1}\frac{\alpha_k^*}{\lambda_k}\left(\phi_g(s^*_{(q)}) - m(s^*_{(q)})\phi_t(s^*_{(q)})\right)\\ &+ 2q\left(\phi_f(s^*_{(q)}) - m(s^*_{(q)})\phi_{t^2}(s^*_{(q)})\right).\end{aligned}$$

Since $\phi_h(s^*_{(q)}) = m(s^*_{(q)})\phi_1(s^*_{(q)})$ for general h, the first term above is equal to zero, and the bias coefficient is equal to

$$\begin{aligned} c(s^*_q) = &\left(\mathbb{E}(\theta^2 h(\theta) \,|\, s^*_{(q)}) - \mathbb{E}(\theta^2 \,|\, s^*_{(q)}) m(s^*_{(q)})\right) \frac{q}{q+2} \\ &- \sqrt{2}\left(\mathbb{E}(\theta h(\theta) \,|\, s^*_{(q)}) - \mathbb{E}(\theta \,|\, s^*_{(q)}) m(s^*_{(q)})\right) \frac{C_q}{q+2} \\ = &\frac{q}{q+2}\operatorname{Cov}(\theta^2, h(\theta) \,|\, s^*_{(q)}) - \sqrt{2}\frac{C_q}{q+2}\operatorname{Cov}(\theta, h(\theta) \,|\, s^*_{(q)}), \end{aligned} \tag{5.8}$$

where

$$C_q := \sum_{k=1}^{q} (-1)^{k-1} \alpha^*_k / \lambda_k. \tag{5.9}$$

Therefore, if the limit for the bias coefficient exists, it is equal to

$$\lim_{q\uparrow\infty} c(s^*_{(q)}) = \lim_{q\uparrow\infty} \operatorname{Cov}(\theta^2, h(\theta) \,|\, s^*_{(q)}) - \sqrt{2}\lim_{q\uparrow\infty} \frac{C_q}{q+2}\operatorname{Cov}(\theta, h(\theta) \,|\, s^*_{(q)}).$$

Whether this is bounded depends on the value of $\lim_{q\uparrow\infty} C_q/(q+2)$.

Lemma 5.9. *Let θ have a conjugate normal prior, $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$. Then the* ABCBAS *estimate $Z_N^{(q)}$ for $\mathbb{E}(\theta \,|\, s^*_{(q)})$, where q is finite, is such that*

$$\operatorname{bias}(Z_N^{(q)}) = c(s^*_{(q)})\delta^2 + \mathcal{O}\left(\delta^3\right),$$

as $\delta \downarrow 0$, for some constant c.

Proof. By Lemma 5.8,

$$\theta \,|\, s^*_{(q)} \sim \mathrm{N}\left(\mu_q, \sigma_q^2\right),$$

where

$$\sigma_q^{-2} = \sigma_0^{-2} + 2L_q,$$

L_q is defined in Equation (5.6), and

$$\mu_q \sigma_q^{-2} = \mu_0\sigma_0^{-2} + \sqrt{2}\sum_k (-1)^{k-1}\alpha^*_k = \mu_0\sigma_0^{-2} + x^*_{(q)}(1).$$

We observe that

$$\mathbb{E}(\theta \,|\, s^*_{(q)}) = \mu_q,\ \mathbb{E}(\theta^2 \,|\, s^*_{(q)}) = \mu_q^2 + \sigma_q^2,\ \mathbb{E}(\theta^3 \,|\, s^*_{(q)}) = \mu_q^3 + 3\mu_q\sigma_q^2.$$

Substituting these into Equation (5.8) shows the bias coefficient to be

$$\begin{aligned} c(s^*_{(q)}) &= 2\mu_q\sigma_q^2\frac{q}{q+2} - \sqrt{2}\sigma_q^2\frac{C_q}{q+2} \\ &= 2\sigma_q^4(\mu_0\sigma_0^{-2} + x^*_{1,(q)})\frac{q}{q+2} - \sqrt{2}\sigma_q^2\frac{C_q}{q+2}, \end{aligned}$$

where C_q is defined in Equation (5.9). □

If the limit for $c(s^*_{(q)})$ as $q \uparrow \infty$ exists, then it is equal to

$$\lim_{q \uparrow \infty} c(s^*_{(q)}) = 2\frac{\mu_0 {\sigma_0}^{-2} + x^*_1}{({\sigma_0}^{-2}+1)^2} - \frac{\sqrt{2}}{{\sigma_0}^{-2}+1} \lim_{q \uparrow \infty} \frac{C_q}{q+2}.$$

Example 5.10. *The observed process $X^*_t = t$ has Karhunen-Loève coefficients*

$$\alpha^*_k = (-1)^{k+1}\sqrt{2}\lambda_k,$$

and the resulting true posterior distribution is $\theta \mid s^*_{(q)} \sim \mathrm{N}\left(\frac{\mu{\sigma_0}^{-2}+1}{{\sigma_0}^{-1}+1}, \frac{1}{{\sigma_0}^{-1}+1}\right)$. We can show that

$$x^*_{(q)}(1) = 2L_q, \quad C_q = \sqrt{2}q,$$

and so the sequence of bias coefficients is equal to

$$\begin{aligned} c(s^*_{(q)}) &= 2\frac{\mu_0 {\sigma_0}^{-2} + 2L_q}{({\sigma_0}^{-2}+2L_q)^2}\frac{q}{q+2} - 2\frac{1}{{\sigma_0}^{-2}+2L_q}\frac{q}{q+2} \\ &= 2(\mu_0 - 1)\frac{{\sigma_0}^{-2}}{({\sigma_0}^{-2}+2L_q)^2}\frac{q}{q+2}. \end{aligned}$$

This sequence has the limit

$$c(s^*) = 2(\mu_0 - 1)\frac{\sigma_0^2}{(\sigma_0^2+1)^2},$$

which is absolutely bounded by $\frac{1}{2}\left|\mu_0 - 1\right|$.

For most generated processes, such as depicted in Figure 5.3.1, the resulting bias coefficient, plotted against the number of spectral coefficients used, as in Figure 5.3.2, resembles a stochastic process.

This raises the question of whether the bias coefficient will tend to a finite limit.

Lemma 5.11 (Non-Boundedness of bias coefficient)**.** *Let the sequence C_q be defined as in Equation* (5.9)*, and θ have a simple normal distribution. Then $C_q/(q+2)$ almost never converges to a finite value.*

Proof. For a fixed value of θ, the components of C_q are independent, with conditional distributions

$$(-1)^{k-1}\alpha_k\lambda_k^{-1} \mid \theta \sim \mathrm{N}\left(\sqrt{2}\theta, \lambda_k^{-1}\right),$$

by Equation (5.5), and so C_q has conditional distribution

$$C_q \mid \theta \sim \mathrm{N}\left(\sqrt{2}\theta q, \sum_{k=1}^{q} \lambda_k^{-1}\right) = \mathrm{N}\left(\sqrt{2}\theta q, \frac{\pi^2}{4}\sum_{k=1}^{q}(2k-1)^2\right).$$

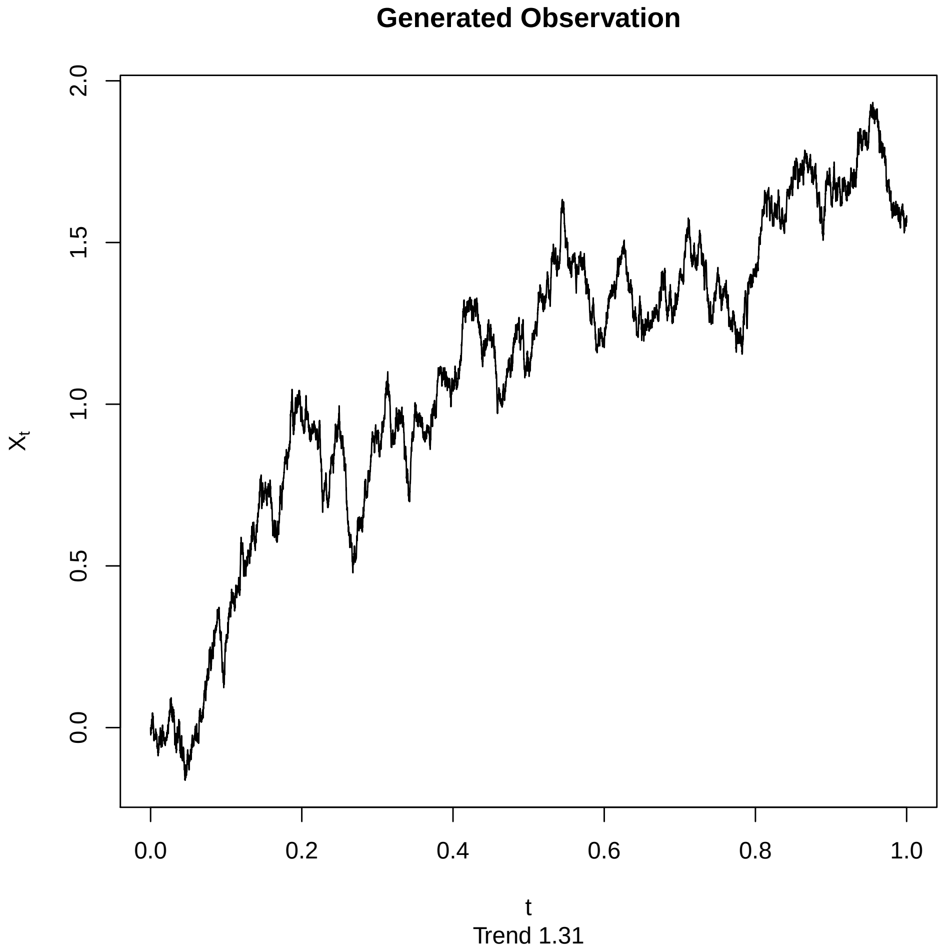

Figure 5.3.1: A spectral approximation of a generated process, using 1000 coefficients.

The sum on the right hand side is equal to

$$\sum_{k=1}^{q}(2k-1)^2 = 4\sum_{k=1}^{q} k^{\underline{2}} + q,$$

where $n^{\underline{2}} = n(n-1)$ is the second falling power of n. By the calculus of finite differences [Graham et al., 1994], this is equal to

$$\begin{aligned} 4\sum_{1}^{q+1} k^{\underline{2}}\,\mathrm{d}k + q &= \frac{4}{3}(q+1)^{\underline{3}} + q \\ &= q\left(\frac{4}{3}(q^2-1)+1\right) \\ &= \frac{4}{3}q^3 - \frac{1}{3}q, \end{aligned}$$

and so C_q has conditional distribution

$$C_q \mid \theta \sim \mathrm{N}\left(\sqrt{2}\theta q, \frac{\pi^2}{3}q^3 - \frac{\pi^2}{12}q\right).$$

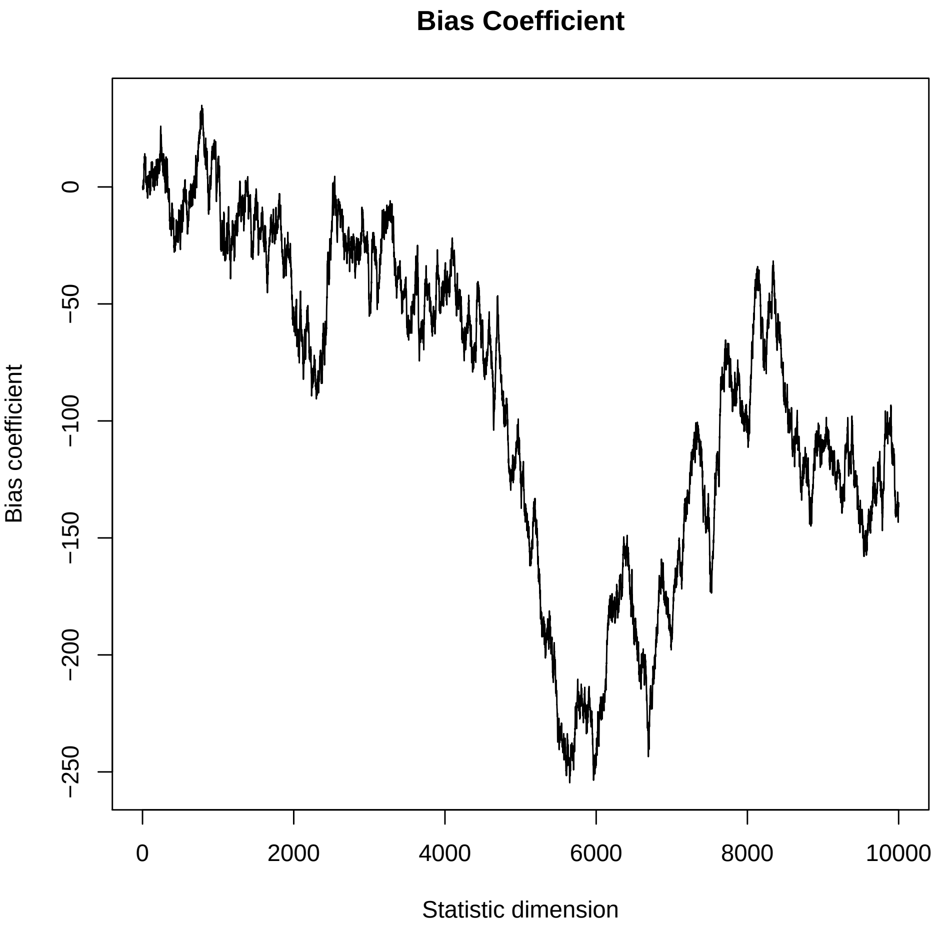

Figure 5.3.2: Plot depicting the value of $c(x^*_{(q)})$ for the process in Figure 5.3.1, as the number of Karhunen-Loève coefficients used increases.

Therefore, since $\theta \sim \mathrm{N}(0,1)$, C_q has unconditional distribution

$$C_q \sim \mathrm{N}\left(0, \frac{\pi^2}{3}q^3 + 2q^2 - \frac{\pi^2}{12}q\right),$$

and the expression $C_q/(q+2)$ has unconditional distribution

$$C_q/(q+2) \sim \mathrm{N}\left(0, \frac{\pi^2}{3}\frac{q^3}{(q+2)^2} + 2\frac{q^2}{(q+2)^2} - \frac{\pi^2}{12}\frac{q}{(q+2)^2}\right).$$

Since the variance diverges as $q \uparrow \infty$, and

$$\mathbb{P}\left(\left|\frac{C_q}{q+2} - x\right| \leq \epsilon\right) \leq \mathbb{P}\left(\frac{C_q}{q+2} \leq x + \epsilon\right)$$

for all x, and all $\epsilon > 0$, we can find, for all finite $x, \epsilon, \eta > 0$, a value Q, such that

$$\mathbb{P}\left(\left|\frac{C_q}{q+2} - x\right| \leq \epsilon\right) < \eta$$

for all $q > Q$. Therefore, $C_q/(q+2)$ diverges in probability, which implies almost sure divergence. □

Corollary 5.12. *The bias for the infinite-dimensional case is such that, for $\mathbb{P}$-almost all s^*,*

$$\lim_{\delta \downarrow 0} \delta^{-1} \text{bias}(Z_N) = 0,$$

and such that $\delta^{-2}\text{bias}(Z_N)$ diverges, as $\delta \downarrow 0$, for $\mathbb{P}$-almost all s^*.

Therefore, the bias is no longer $\mathcal{O}\left(\delta^2\right)$ as $\delta \downarrow 0$, so it vanishes more slowly than in the finite-dimensional case. However, it is at least $\mathrm{o}\left(\delta\right).$

5.3.3 Asymptotic Variance

For the asymptotic variance, we require the value of the acceptance probability $\mathbb{P}\left(\|X\|_2 \leq \delta\right).$ In the case where the sample process is a simple Brownian motion B with no trend, this satisfies [Li and Shao, 2001, Theorem 6.3]

$$\lim_{\delta \downarrow 0} \delta^2 \log \mathbb{P}\left(\|B\|_2 \leq \delta\right) = -1/8.$$

For general X, we can use the following lemma.

Lemma 5.13. *Let X be the motion $X_t = B_t + \theta t$, a simple Brownian motion plus a linear trend. Then the probability that $\|X\|_2 \leq \delta$ is equal to*

$$\mathbb{P}\left(\|X\|_2 \leq \delta \,|\, \theta\right) = \exp\left(-\frac{1}{2}\theta^2\right) \mathbb{E}\left([\|B\|_2 \leq \delta] \exp\left(\theta B_1\right)\right).$$

Proof. By the Girsanov theorem (Theorem A.13), X is a martingale with respect to the measure

$$\mathrm{d}\mathbb{Q} = \exp\left(-\theta B_1 - \frac{1}{2}\theta^2\right) \mathrm{d}\mathbb{P}.$$

Therefore, the acceptance probability is equal to

$$\begin{aligned}
\mathbb{P}\left(\|X\|_2 \leq \delta \,|\, \theta\right) &= \mathbb{E}\left([\|X\|_2 \leq \delta]\right) \\
&= \mathbb{E}_{\mathbb{Q}}\left([\|X\|_2 \leq \delta] \exp\left(\theta B_1 + \frac{1}{2}\theta^2\right)\right) \\
&= \exp\left(\frac{1}{2}\theta^2\right) \mathbb{E}_{\mathbb{Q}}\left([\|X\|_2 \leq \delta] \exp\left(\theta B_1\right)\right) \\
&= \exp\left(-\frac{1}{2}\theta^2\right) \mathbb{E}_{\mathbb{Q}}\left([\|X\|_2 \leq \delta] \exp\left(\theta X_1\right)\right),
\end{aligned}$$

where $\mathbb{E}_{\mathbb{Q}}$ is the expectation with respect to $\mathbb{Q}$, rather than $\mathbb{P}$. Since X is a martingale with respect to $\mathbb{Q}$,

$$\mathbb{P}\left(\|X\|_2 \leq \delta \,|\, \theta\right) = \exp\left(-\frac{1}{2}\theta^2\right) \mathbb{E}\left(\left[\|B\|_2 \leq \delta\right] \exp\left(\theta B_1\right)\right),$$

as required. □

As in Section 5.2.2, analysis that accounts for staying near s^* rather than 0 is difficult. However, we can again consider the asymptotic behaviour of the ABC estimate over all s^*, and consider the mean acceptance probability $\bar{p}$.

Lemma 5.14. *Let X_t be the motion $X_t = B_t + \theta t$, a simple Brownian motion plus a linear trend, and X_t^* be the motion $X_t^* = B_t^* + \theta^* t$, where θ and θ^* have independent simple normal distributions, and B_t and B_t^* are independent Brownian motions. Then the probability that $\|X - X^*\|_2 \leq \delta$ is equal to*

$$\bar{p} := \mathbb{P}\left(\|X - X^*\|_2 \leq \delta\right) = \mathbb{P}\left(\|B\|_2 \leq \frac{\delta}{2}\right),$$

and satisfies

$$\lim_{\delta \downarrow 0} \delta^2 \log(\bar{p}) = -1/2.$$

Proof. Similar to Corollary 5.4. □

Theorem 5.15. *Let X_t be the motion $X_t = B_t + \theta t$, a simple Brownian motion plus a linear trend, and X_t^* be the motion $X_t^* = B_t^* + \theta^* t$, where θ and θ^* have independent simple normal distributions, and B_t and B_t^* are independent Brownian motions. Let $N \uparrow$ and $\delta \downarrow 0$, such that $N \exp\left(-1/2\delta^2\right) \uparrow \infty$. Then the variance of the ABCBAS estimate Z_N satisfies*

$$\lim_{N \exp(-1/2\delta^2) \uparrow \infty} N\bar{p} \operatorname{Var}\left(Z_N\right) = v$$

$\mathbb{P}$-almost surely, where $v := \mathbb{E}\left(v(X^*)\right)$ is the prior variance.

Proof. By the proof of Lemma 3.11, we know that the mean variance of the ABCBAS estimate Z_N for the observation X^* satisfies

$$\lim_{N\bar{p} \uparrow \infty} N\bar{p} \operatorname{Var}\left(Z_N\right) = v$$

$\mathbb{P}$-almost surely, where $\bar{p} = \exp\left(-1/2\delta^2 + \mathrm{o}\left(1/\delta^2\right)\right)$, as $\delta \downarrow 0$, by Lemma 5.14. □

5.3.4 Optimising the Error

We have found the asymptotic order of the bias to be $\mathcal{O}(\delta^r)$, where $r > 1$, and $r < 2$ $\mathbb{P}$-almost surely, and the asymptotic variance to be $\mathcal{O}\left(e^{1/2\delta^2}/N\right)$, as δ tends to zero. However, this does not necessarily allow us to optimise the asymptotic error. In Theorem 3.12, we assume the stricter condition that the associated limits exist. Here, the associated limits are those for the expressions $\text{bias}(Z_N)/\delta^r$ and $\operatorname{Var}(Z_N)/N\exp\left(-1/2\delta^2\right)$, and we can not determine whether these limits exist. For example, whether the associated limit for the variance exists depends on the behaviour of

$$A(\delta) := \bar{p}/\exp\left(-1/2\delta^2\right) = \exp\left(\mathrm{o}\left(1/\delta^2\right)\right)$$

as δ tends to zero. Specifically, the condition $N\exp\left(-1/2\delta^2\right) \uparrow \infty$ that is sufficient for the variance to be $\mathcal{O}\left(1/N\exp\left(1/2\delta^2\right)\right)$ is only sufficient for the associated limit to exist if $A(\delta)$ converges as δ tends to zero.

To get an idea of how the statistic being a diffusion process affects the optimal asymptotic error, we suppose that $\text{bias}(Z_N) = c\delta^r(1+\mathrm{o}(1))$ as δ tends to zero, and that $A(\delta)$ tends to some constant $A > 0$ as δ tends to zero. In this case, if $D := \lim_{N\exp(-1/2\delta^2)\uparrow\infty} N\delta^4\exp\left(-1/2\delta^2\right)$ exists, the error satisfies

$$\lim_{N\exp(-1/2\delta^2)\uparrow\infty} N\exp\left(-1/2\delta^2\right)\textsc{mse}(Z_N) = v/A + c^2D.$$

To prove the equivalent of Theorem 3.12, since the expected cost is equal to $C = kN$ for some $k > 0$, we now require some function f such that the limit

$$\lim_{C\uparrow\infty} N\exp\left(-1/2\delta^2\right)f(kN)$$

exists, and is non-zero. This would give the result that

$$\lim_{C\uparrow\infty} f(C)^{-1}\textsc{mse}(Z_N) > 0,$$

and so that the error is $\mathcal{O}(f(C))$ as C tends to infinity. More generally, if $A(\delta)$ does not necessarily converge, then we require the limit $\bar{D} := \lim_{C\uparrow\infty} N\delta^4\bar{p}$ to exist, and look for a function $\bar{f}$ such that

$$\lim_{C\uparrow\infty} N\bar{p}\bar{f}(kN)$$

exists, and is non-zero.

Finding such a function f, or $\bar{f}$, and so finding the asymptotic order of the error, is not trivial. However, it is clear, from the presence of the exponential tolerance term, that the order of convergence is much slower than those found in Chapter 3.

5.4 Discussion

There has recently been research on the behaviour of ABC estimates as the statistic dimension tends to infinity. However, the focus has been on the consistency of the estimate, rather than its asymptotic convergence. More specifically, the statistic is taken to consist of q independent observations, and the estimate is coherent if the estimate converges to the true value as q tends to infinity, with N and δ fixed. Some examples are discussed in Section 2.2.4.

Under current computational limitations, observing a continuous process is only possible with the loss of information, or with the use of finite sufficient statistics, such as described in Section 5.1. However, the sequence P_q of spectral problems are feasible, so the results in Section 5.3 are of some use. In particular, they demonstrate that, even for small q, a change in q can greatly affect the bias coefficient. This further emphasises the point that the asymptotic results should not be used to directly compare estimates with different values of q : increasing q decreases the asymptotic rate, but can greatly decrease the leading term.

It should be noted that some of the results given in this chapter – in particular, those for the asymptotic bias under the Euclidean norm – make use of the problem having a simple one-dimensional minimal statistic. It would be much more complicated to consider problems where the lowest-dimension minimal statistic has a higher dimensional, or is even infinite-dimensional.

Chapter 6

Conclusion

This text has focused on the asymptotic behaviour of Approximate Bayesian Computation, when used to estimate the posterior expectation. In Chapter 3, we looked at the asymptotic order of the mean square error for basic variants of ABC, when estimating the posterior expectation. The basic variants were shown to have asymptotic error of optimal order $\mathcal{O}\left(C^{-4/(q+4)}\right)$, for a choice of tolerance with order $\mathcal{O}\left(C^{-1/(q+4)}\right)$, where C is the expected computational cost and tends to infinity, and q is the dimension of the sufficient summary statistic. By comparison, exact Monte Carlo methods have error of order $\mathcal{O}\left(C^{-1}\right)$. We extended the analysis to look at the asymptotic error when using a kernel K for random acceptance of proposals. The error is of the same asymptotic order, and the leading asymptotic term of the error is likely to be optimised by using a uniform kernel, which is the same as using the standard accept-reject scheme. We also looked at some other minor variations, which have the same asymptotic order with a different leading term. The dependence of the asymptotic error on q motivates the use of low-dimensional summary statistics in practice.

In Chapter 4, we proposed a new variant of ABC for posterior expectation estimation, called ABCLOC. This is a variation on the regression adjustment introduced by Beaumont et al. [2002], where we use a regression centred on the original observed statistic, and an additional regression centred on each of the accepted statistic samples, to adjust the accepted proposals. Additionally, the kernel is used for regression and random proposal acceptance, rather than regression and proposal weighting. While the asymptotic analysis of ABCLOC is not complete, the variant is thought to have two advantages. First, the use of multiple regressions improves the asymptotic order of the bias. Second, the resulting improvement in the order, with respect to the number of proposals, is enough to make up for the increased computational cost introduced by the regressions. The order is thought to be $\mathcal{O}\left(C^{-8/(8+\max\{q,8\})}\right)$ or $\mathcal{O}\left(C^{-8/(q+16)}\right)$, depending on whether the support of the kernel is finite.

In Chapter 5, we looked at the hypothetical case where the observation is a path of a Brownian motion with an unknown linear trend, and no summary statistic is used. Since the asymptotic rate of decay for the error slows as q increases, we look at this case to refute the naïve thought that using an infinite-dimensional statistic results in no convergence. Instead, we find that the bias is order $\mathcal{O}\left(\delta^r\right)$, for some r such that $1 < r < 2$, and that the variance is $\exp\left(-1/2\delta^2 + \mathrm{o}\left(1/\delta^2\right)\right)$, as the tolerance parameter δ tends to zero. While this would result in a very slow rate of convergence, the estimate still converges.

6.1 Brief Comment on Practical Usage

Little discussion has been given in this text on practical consequences of the results. This because, due to the asymptotic nature of the results, there are few such consequences: optimal usage for practical running times need not be that which is optimal as the computational running time tends to infinity. As mentioned in Chapter 1, it can also be highly problem-specific.

For example, while the main results of the text give asymptotic rates for the optimal choice of tolerance parameter δ, and the resulting optimal mean square error of the estimate, this gives no guidance on the choice of tolerance for a single ABC estimate. In practice, a common rule of thumb for choosing the tolerance value, mentioned by Beaumont et al. [2002], is to choose the tolerance so that a small fixed proportion – one percent, for example – of the proposals are accepted. This can be done either by generating a pilot run of samples and fixing the tolerance so that a fixed proportion of the pilot proposals would be accepted, or by choosing the tolerance after the samples have been generated, if the number N of proposals is fixed. The latter approach is equivalent to using n-nearest neighbours rather than a tolerance value, and the asymptotic behaviour of the error for this approach is discussed by Biau et al. [2015].

There are some possible practical consequences of the results, such as the use of ABCLOC from Chapter 4, or using accept-reject instead of random acceptance, based on Example 3.15. However, suggesting these for practical usage should be done based on more numerical experiments, conducted on more complicated example problems, than were done for this text.

6.2 Planned Extensions

The planned extensions focus on the new ABCLOC estimate, with the main priority being to complete the asymptotic analysis given in Chapter 4. This requires proving the remaining corollaries in that chapter, and finding the asymptotic behaviour of the variance. Running numerical experiments is also needed, to determine whether ABCLOC is likely to be useful in practice.

Since ABCLOC is a variation on the regression step used in Beaumont et al. [2002] and Blum [2010], it may then be possible to adapt the analysis of ABCLOC to variations of other ABC variants that make use of regression. For example, Fearnhead and Prangle [2012] consider the case where we begin with an initial summary statistic function, s. They then generate a preliminary set of samples, and use regression to determine a one-dimensional summary statistic $\hat{m}$ that is approximately sufficient for s. We can, therefore, consider a variant, where $\hat{m}$ is determined using multiple regressions on the preliminary samples, similarly to ABCLOC. If we assume the initial statistic s is sufficient, then it might be possible to determine the asymptotic effect of using $\hat{m}$ rather than s.

The asymptotic rates given for the mean square error of different variants of ABC are given in terms of the expected computational cost. Since some of the variants, such as ABCACC, and ABCLOC, have a large variance in their cost, one extension would be to find the asymptotic error in terms of the actual cost. As a more practical alternative, we could bring the algorithm closer to how ABC is used in practice: the algorithm is given a strict upper bound on its running time, or on the time taken to generate samples.

Appendix A

Miscellaneous Theorems

This appendix contains definitions, theorems, and lemmas that are not included in the main text. This has been done either because they are commonly-known, because they are trivial (Lemma A.1), or because they appear in reference to previous results, and are not used in this thesis (Definition A.9).

Lemma A.1.

$$\sum_{k=1}^{N} a_k \left(x - c_k\right)^2 = \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 + \frac{\sum_{k<j} a_k a_j (c_k - c_j)^2}{\sum_{k=1}^{N} a_k}.$$

Proof.

$$\begin{aligned}
\sum_{k=1}^{N} a_k \left(x - c_k\right)^2 &= \left(\sum_{k=1}^{N} a_k\right) x^2 - 2\left(\sum_{k=1}^{N} a_k c_k\right) x + \sum_{k=1}^{N} a_k c_k^2 \\
&= \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 - \frac{\left(\sum_{k=1}^{N} a_k c_k\right)^2}{\sum_{k=1}^{N} a_k} + \sum_{k=1}^{N} a_k c_k^2 \\
&= \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 - \frac{\left(\sum_{k=1}^{N} a_k c_k\right)^2 - \sum_{k=1}^{N} a_k \sum_{k=1}^{N} a_k c_k^2}{\sum_{k=1}^{N} a_k} \\
&= \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 - \frac{\sum_{k,j=1}^{N} a_k a_j c_k c_j - \sum_{k,j=1}^{N} a_k a_j c_j^2}{\sum_{k=1}^{N} a_k} \\
&= \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 - \frac{\sum_{k<j}\left(2a_k a_j c_k c_j - a_k a_j (c_k^2 + c_j^2)\right)}{\sum_{k=1}^{N} a_k} \\
&= \left(\sum_{k=1}^{N} a_k\right)\left(x - \frac{\sum_{k=1}^{N} a_k c_k}{\sum_{k=1}^{N} a_k}\right)^2 + \frac{\sum_{k<j} a_k a_j (c_k - c_j)^2}{\sum_{k=1}^{N} a_k}. \qquad \square
\end{aligned}$$

Theorem A.2 (Multiple-Dimensional Taylor's Theorem)**.** *[Burkill, 1962, an extension from Theorem 8.7] Let f be a function $f\colon \mathbb{R}^d \to \mathbb{R}$, whose partial derivatives are continuous up to order n in a region around $\mathbf{x}$ containing $\mathbf{x}+\mathbf{h}$. Then there exists a constant $0<\phi<1$ such that*

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) &= f(x_1+h_1,\ldots,x_d+h_d) \\ &= f(x_1,\ldots,x_d) + \left(h_1\frac{\partial}{\partial x_1}+\cdots+h_d\frac{\partial}{\partial x_d}\right) f(x_1,\ldots,x_d) \\ &\quad + \frac{1}{2}\left(h_1\frac{\partial}{\partial x_1}+\cdots+h_{d-1}\frac{\partial}{\partial x_{d-1}}+h_d\frac{\partial}{\partial x_d}\right)^2 f(x_1,\ldots,x_d) \\ &\quad +\cdots \\ &\quad + \frac{1}{(n-1)!}\left(h_1\frac{\partial}{\partial x_1}+\cdots+h_d\frac{\partial}{\partial x_d}\right)^{n-1} f(x_1,\ldots,x_d) \\ &\quad + \frac{1}{n!}\left(h_1\frac{\partial}{\partial x_1}+\cdots+h_d\frac{\partial}{\partial x_d}\right)^{n} f(x_1+\phi h_1,\ldots,x_d+\phi h_d) \\ &= f(\mathbf{x}) + \mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} f(\mathbf{x}) + \cdots \\ &\quad + \frac{1}{(n-1)!}\left(\mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\right)^{n-1} f(\mathbf{x}) + \frac{1}{n!}\left(\mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\right)^{n} f(\mathbf{x}+\phi\mathbf{h}). \end{aligned}$$

Remark: In the case $n=2$, this can also be written in matrix notation as

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T\nabla f(\mathbf{x}) + \mathbf{h}^T\mathcal{H}_f(\mathbf{x}+\phi\mathbf{h})\,\mathbf{h},$$

where $\mathcal{H}_f(\mathbf{x}+\phi\mathbf{h})$ is the Hessian matrix for $f(\mathbf{x}+\phi\mathbf{h})$.

Proof. We define $F(t)=f(\mathbf{x}+t\mathbf{h})$, parametrizing along the line between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$. Then, by the one-dimensional Taylor's theorem, for some $0<\phi<1$,

$$F(1) = F(0) + F'(0) + \cdots + \frac{1}{(n-1)!}F^{(n-1)}(0) + \frac{1}{n!}F^{(n)}(\phi).$$

In terms of the original notation,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}) = \sum_{i=1}^{n} h_i \frac{\partial}{\partial x_i} f(\mathbf{x}) = \mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} f(\mathbf{x}),$$

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} f(\mathbf{x}) + \cdots + \frac{1}{n!}\left(\mathbf{h}\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\right)^{n} f(\mathbf{x}+\phi\mathbf{h}). \qquad \square$$

Lemma A.3 (Young's Inequality)**.** *Let a and b be positive real numbers, and p and q be positive real numbers such that $1/p+1/q=1$. Then*

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}.$$

Definition A.4 (Big O notation)**.** *Functions f and g satisfy $f(x) = \mathcal{O}\left(g(x)\right)$ as $x \uparrow \infty$, with respect to function g, if, and only if, there is some positive constant ϵ such that $|f(x)| \leq \epsilon\, |g(x)|$ for all sufficiently large values of x.*

Functions f and g are such that $f(x) = \mathcal{O}\left(g(x)\right)$ as $x \to x_0$, for some finite x_0, if, and only if, there are some constants $\delta, \epsilon > 0$ such that $|f(x)| \leq \epsilon\, |g(x)|$ if $|x - x_0| \leq \delta$.

Definition A.5 (Little O notation)**.** *Functions f and g satisfy $f(x) = \mathrm{o}\left(g(x)\right)$ as $x \uparrow \infty$ if, for all $\epsilon > 0$, $|f(x)| \leq \epsilon\, |g(x)|$ for all sufficiently large x.*

Functions f and g are such that $f(x) = \mathrm{o}\left(g(x)\right)$ as $x \to x_0$ if, for all $\epsilon > 0$, there is some constant $\delta > 0$ such that $|f(x)| \leq \epsilon\, |g(x)|$ if $|x - x_0| > \delta$.

Lemma A.6 (Lebesgue Differentiation Theorem)**.** *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be some real-valued function. Then the derivative*

$$\lim_{B \to x} \frac{1}{\|B\|} \int_B f \,\mathrm{d}\mu,$$

where B is the ball centred at x, exists and is equal to $f(x)$ for almost all $x \in \mathbb{R}^n$.

Proof. See [Rudin, 1987, Theorem 7.7]. □

Lemma A.7 (Newton Series)**.** *Graham et al. [1994] Let f be some function $f \colon \mathbb{N} \to \mathbb{R}$, and the forward and backward differences*

$$\Delta^n f(x) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f(x+k), \quad \nabla^n f(x) = \sum_{k=0}^{n} (-1)^{k} \binom{n}{k} f(x-k)$$

exist up to order h. Then

$$f(x+h) = \begin{cases} \sum_{k=0}^{h} \binom{h}{k} \Delta^k f(x) & h \geq 0, \\ \sum_{k=0}^{h} (-1)^k \binom{h}{k} \nabla^k f(x) & h \leq 0. \end{cases}$$

Proof. By the summation inversion formula,

$$(-1)^h \Delta^h f(x) = \sum_{k=0}^{h} (-1)^k \binom{h}{k} f(x+k) \iff f(x+h) = \sum_{k=0}^{h} \binom{h}{k} \Delta^k f(x),$$

$$\nabla^h f(x) = \sum_{k=0}^{h} (-1)^k \binom{h}{k} f(x-k) \iff f(x-h) = \sum_{k=0}^{h} (-1)^k \binom{h}{k} \nabla^k f(x). \quad \square$$

Lemma A.8. *The minimisation problem*

$$\underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\theta_i - \alpha - \beta(s_i - s^*)\right)^2 w(s_i - s^*)$$

has solution

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \left(X_{s^*}^T W_{s^*} X_{s^*}\right)^{-1} X_{s^*}^T W_{s^*} \Theta,$$

where

$$X_{s^*} = \begin{pmatrix} 1 & \dots & 1 \\ (s_1 - s^*) & \dots & (s_N - s^*) \end{pmatrix}^T,$$

and $W_{s^*} = \operatorname{diag}\left(w(s_1 - s^*), \dots, w(s_N - s^*)\right)$.

Proof. If we let $\gamma := (\alpha\, \beta^T)^T$, and Θ be the vector such that $(\Theta)_i = \theta_i$, the minimisation problem can be written as

$$\underset{\gamma}{\operatorname{argmin}} \left(\Theta - X_{s^*}\gamma\right)^T W_{s^*} \left(\Theta - X_{s^*}\gamma\right).$$

Differentiating by γ, we get

$$-2X_{s^*}^T W_{s^*} \left(\Theta - X_{s^*}\gamma\right) = 0.$$

Rearranging, we get

$$X_{s^*}^T W_{s^*} X_{s^*} \gamma = X_{s^*}^T W_{s^*} \Theta.$$

Left-multiplying by the inverse of $X_{s^*}^T W_{s^*} X_{s^*}$ gives the required result. □

Definition A.9 (Little O_P notation)**.** *The functions f and g are such that $f(x) = \mathrm{o}_P\left(g(x)\right)$ as $x \uparrow \infty$ if, for all $\delta, \epsilon > 0$, there is some constant $\gamma > 0$, such that*

$$\mathbb{P}\left(f(x) \geq \delta g(x)\right) \leq \epsilon$$

for all $x \geq \gamma$. The sequence $f(x)/g(x)$ is said to converge in probability.

Lemma A.10 (Chernoff bound for lower tail)**.** *Let X be a Bin(N, p) variable. Then, for $\epsilon > 0$,*

$$\mathbb{P}\left(X \leq (1-\epsilon)Np\right) \leq \exp\left(-\frac{\epsilon^2 Np}{2}\right).$$

Proof. By the Markov inequality,

$$\mathbb{P}\left(X \leq (1-\epsilon)Np\right) = \mathbb{P}\left(e^{-tX} \geq e^{-t(1-\epsilon)Np}\right) \leq e^{t(1-\epsilon)Np}\mathbb{E}\left(e^{-tX}\right),$$

where $\mathbb{E}\left(e^{-tX}\right) = \left(1+p(e^{-t}-1)\right)^N \leq \exp\left(Np(e^{-t}-1)\right)$, so

$$\mathbb{P}\left(X \leq (1-\epsilon)Np\right) \leq \exp\left(t(1-\epsilon)Np + Np(e^{-t}-1)\right).$$

Minimising with regard to t gives $t = -\log(1-\epsilon)$ if $\epsilon > 0$, giving the upper bound

$$\begin{aligned}\mathbb{P}\left(X \leq (1-\epsilon)Np\right) &\leq \exp\left(-(1-\epsilon)\log(1-\epsilon)Np - \epsilon Np\right)\\ &= \exp\left(-Np\left((1-\epsilon)\log(1-\epsilon)+\epsilon\right)\right)\\ &\leq \exp\left(-\frac{1}{2}\epsilon^2 Np\right),\end{aligned}$$

by taking the Taylor expansion of the logarithm. □

Lemma A.11 (Chernoff bound for upper tail)**.** *Let X be a Bin(N,p) variable. Then, for $\epsilon > 0$,*

$$\mathbb{P}\left(X \geq (1+\epsilon)Np\right) \leq \exp\left(-\frac{\epsilon^2 Np}{3}\right).$$

Proof. By similar reasoning to that in the proof for Lemma A.10,

$$\mathbb{P}\left(X \geq (1+\epsilon) \leq \exp\left(-t(1+\epsilon)Np + Np\left(e^t-1\right)\right),\right.$$

Minimising with regard to t gives $t = \log(1+\epsilon)$ for $\epsilon > 0$, giving the upper bound

$$\begin{aligned}\mathbb{P}\left(X \leq (1-\epsilon)Np\right) &\leq \exp\left(-(1+\epsilon)\log(1+\epsilon)Np + \epsilon Np\right)\\ &= \exp\left(-Np\left((1+\epsilon)\log(1+\epsilon)-\epsilon\right)\right).\end{aligned}$$

Taking the Taylor expansion of the logarithm, we have, since $0 < \epsilon < 1$,

$$(1+\epsilon)\log(1+\epsilon) \leq (1+\epsilon)\left(\epsilon - \frac{1}{2}\epsilon^2 + \frac{1}{3}\epsilon^3\right) \leq \epsilon + \frac{1}{2}\epsilon^2 - \frac{1}{6}\epsilon^3 \leq \epsilon + \frac{1}{3}\epsilon^2.$$

Substituting this into the previous inequality gives the result. □

Lemma A.12 (Block matrix inversion)**.**

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

Proof. Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, and $M/A = D - CA^{-1}B$ and $M/D = A - BD^{-1}C$ be the Schur complement of A and D, respectively. Then it can be shown that

$$\begin{aligned} M &= \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & M/A \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} M/D & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}. \end{aligned} \tag{A.1}$$

Taking the inverse, we see that

$$M^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} M/D^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}.$$

The result easily follows. □

Theorem A.13 (Girsanov)**.** *[Øksendal, 2003] Let X be a stochastic process obeying the equation*

$$\mathrm{d}X_t = a(X_t, t)\,\mathrm{d}t + \mathrm{d}B_t, \quad t \leq T,\, X_0 = 0,$$

where B_t is a Brownian motion, for $t \leq T$ and $X_0 = 0$. Define the process M_t as

$$M_t = \exp\left(- \int_0^t a(s,\omega)\,\mathrm{d}B_s - \frac{1}{2}\int_0^t a(s,\omega)^2\,\mathrm{d}s \right), \quad t \leq T,$$

and define the measure $\mathbb{Q}$ such that

$$\mathrm{d}\mathbb{Q} = M_T\,\mathrm{d}\mathbb{P}.$$

Then, if a satisfies Novikov's condition

$$\mathbb{E}\left(\exp\left(\frac{1}{2}\int_0^T a(X_s, s)^2\,\mathrm{d}s \right) \right) < \infty,$$

X is a Brownian motion with respect to the probability law $\mathbb{Q}$ on $t \leq T$.

Appendix B

Code

This is a general repository for code used in the main text.

B.1 Problem and Error

B.1.1 Bias and Variance Plots

These scripts generate the data, and plot, Figures 1-4.

Figure 1

```
# fig1.R - generate data for figure 1.
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

set.seed(52228)

sigma.theta <- 1
sigma.x <- 1
ind.lower <- -0.5
ind.upper <- 0.5

q <- 2
s.star <- c(1, 1)

# Generate n ABC samples for the posterior distribution of theta.
GenerateABCSamples <- function(n, delta) {
    accepted <- 0
    samples <- numeric(n)
    while (accepted < n) {
        theta <- rnorm(1, sd = sigma.theta)
        X <- rnorm(q, mean = theta, sd = sigma.x)
        if (sum((X - s.star)^2) <= delta^2) {
            accepted <- accepted + 1
            samples[accepted] <- theta
        }
    }
    return(samples)
}

# Return one ABC estimate, using n ABC samples.
GetABCEstimate <- function(n, delta) {
    theta <- GenerateABCSamples(n, delta)
    Z <- ind.lower <= theta & theta <= ind.upper
    est <- mean(Z)
    return(est)
}

# Compute the exact result of the estimation problem.
# This is used to assess the quality of the ABC estimates.
GetTruePosterior <- function() {
    inter <- q * sigma.theta^2 + sigma.x^2
    inter2 <- q * mean(s.star) * sigma.theta^2/inter
    inter3 <- sigma.theta * sigma.x/sqrt(inter)
    lower <- pnorm(ind.lower, mean = inter2, sd = inter3)
    upper <- pnorm(ind.upper, mean = inter2, sd = inter3)
    return(upper - lower)
}

# Return a Monte Carlo estimate of the bias, using k ABC estimates.
# Returns the estimate and its standard error.
EstimateBias <- function(delta, k, n) {
    exact <- GetTruePosterior()
    biases <- replicate(k, GetABCEstimate(n, delta)) - exact
    return(c(mean(biases), sd(biases)/sqrt(k)))
}

deltas <- seq(0.05, 2, by = 0.05)
k <- 5000
n <- 10
biases <- t(sapply(deltas, EstimateBias, k, n))
write.table(cbind(deltas, biases), "fig1.dat", row.names = F,
            col.names = c("delta", "bias", "sd"))

# fig1-plot.R - draw figure 1, using data generated by fig1.R
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

data <- read.table("fig1.dat", header = T)
deltas <- data[, 1]
biases <- data[, 2]
ses <- data[, 3]

pdf("fig1.pdf", width = 4.6, height = 3, family = "serif",
    pointsize = 10)
par(oma = c(0, 0, 0, 0), mai = c(0.8, 0.8, 0.15, 0.1))

plot(deltas, biases, xlim = c(0, max(deltas)),
     ylim = range(biases + 1.96 * ses, biases - 1.96 * ses),
     xlab = expression(delta), ylab = "bias", cex = 0.7)
arrows(deltas, biases - 1.96 * ses, deltas, biases + 1.96 * ses,
       0.7 * 0.05, 90, 3)

C.parabola <- -0.018338 - 0.3647609 * (-0.138889)
t <- seq(0, max(deltas), length.out = 100)
lines(t, C.parabola * t^2)

invisible(dev.off())
```

Figure 2

```
# fig2.R - generate data for figure 2.
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

set.seed(52228)

sigma.theta <- 1
sigma.x <- 1
ind.lower <- -0.5
ind.upper <- 0.5

GetTruePosterior <- function(q, s.star) {
    inter <- q * sigma.theta^2 + sigma.x^2
    inter2 <- q * mean(s.star) * sigma.theta^2/inter
    inter3 <- sigma.theta * sigma.x/sqrt(inter)
    lower <- pnorm(ind.lower, mean = inter2, sd = inter3)
    upper <- pnorm(ind.upper, mean = inter2, sd = inter3)
    return(upper - lower)
}
```

```
# generate n ABC samples for the posterior distribution of theta
GenerateABCSamples <- function(n, q, delta, s.star) {
    accepted <- 0
    samples <- numeric(n)
    while (accepted < n) {
        theta <- rnorm(1, sd = sigma.theta)
        X <- rnorm(q, mean = theta, sd = sigma.x)
        if (sum((X - s.star)^2) <= delta^2) {
            accepted <- accepted + 1
            samples[accepted] <- theta
        }
    }
    return(samples)
}

ComputeABCEstimate <- function(delta, n, q, s.star) {
    theta <- GenerateABCSamples(n, q, delta, s.star)
    return(mean(ind.lower <= theta & theta <= ind.upper))
}

EstimateMSE <- function(delta, k, q, s.star, const) {
    exact <- GetTruePosterior(q, s.star)

    n <- round(const * delta^q)
    print(c(delta, n))
    estimates <- replicate(k, ComputeABCEstimate(delta, n, q, s.star))
    tmp <- (estimates - exact)^2
    return(c(mean(tmp), sd(tmp)/sqrt(k)))
}

deltas <- seq(0.1, 1, by = 0.1)
MSEs <- t(sapply(deltas, EstimateMSE, 500, 2, c(1, 1), 16000))
```

```
write.table(cbind(deltas, MSEs), "fig2.dat", row.names = F,
            col.names = c("delta", "MSE", "se"))

# fig2-plot.R - draw figure 2, using data generated by fig2.R
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

data <- read.table("fig2.dat", header = T)
deltas <- data[, 1]
errors <- data[, 2]
ses <- data[, 3]

q <- 2

pdf("fig2.pdf", width = 4.6, height = 3, family = "serif",
    pointsize = 10)
par(oma = c(0, 0, 0, 0), mai = c(0.8, 0.8, 0.15, 0.1))

plot(deltas, errors, xlim = c(0, max(deltas)),
     ylim = range(errors + 1.96 * ses, errors - 1.96 * ses),
     xlab = expression(delta), ylab = "mean square error")
```

```
arrows(deltas, errors - 1.96 * ses, deltas, errors + 1.96 * ses, 0.05,
       90, 3)

deltas.nq <- deltas^(-q)
deltas.4 <- deltas^4
fit <- lm(errors ~ deltas.nq + deltas.4 + 0, weights = 1/ses^2)

plot.deltas <- seq(min(deltas)/2, max(deltas),
                   length.out = length(deltas) * 10)
lines(plot.deltas, predict(fit,
                           data.frame(deltas.nq = plot.deltas^(-q),
                                      deltas.4 = plot.deltas^4)))
coef.nq <- fit$coefficients[1]
coef.4 <- fit$coefficients[2]

# curve is ad^(-q)+bd^4,
# so opt at -qad^(-q-1)+4bd^3=0 => d=(aq/4b)^(1/(q+4))
opt.delta <- (coef.nq * q/(4 * coef.4))^(1/(q + 4))
abline(v = opt.delta)

invisible(dev.off())

print(paste("Optimal delta is", opt.delta))
print(paste("Optimal MSE is",
            predict(fit,
                    data.frame(deltas.nq = opt.delta^(-q),
                               deltas.4 = opt.delta^4))))
```

Figure 3

```
# fig3.R - generate data for figure 3.
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
```

```
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

set.seed(52228)

sigma.theta <- 1
sigma.x <- 1
ind.lower <- -0.5
ind.upper <- 0.5

GetTruePosterior <- function(q, s.star) {
    inter <- q * sigma.theta^2 + sigma.x^2
    inter2 <- q * mean(s.star) * sigma.theta^2/inter
    inter3 <- sigma.theta * sigma.x/sqrt(inter)
    lower <- pnorm(ind.lower, mean = inter2, sd = inter3)
    upper <- pnorm(ind.upper, mean = inter2, sd = inter3)
    return(upper - lower)
}

# generate n ABC samples for the posterior distribution of theta
GenerateABCSamples <- function(n, q, delta, s.star) {
    accepted <- 0
    samples <- numeric(n)
    while (accepted < n) {
        theta <- rnorm(1, sd = sigma.theta)
        X <- rnorm(q, mean = theta, sd = sigma.x)
        if (sum((X - s.star)^2) <= delta^2) {
            accepted <- accepted + 1
            samples[accepted] <- theta
        }
    }
    return(samples)
}

ComputeABCEstimate <- function(delta, n, q, s.star) {
    theta <- GenerateABCSamples(n, q, delta, s.star)
    return(mean(ind.lower <= theta & theta <= ind.upper))
}

EstimateMSE <- function(delta, k, q, s.star, const) {
    exact <- GetTruePosterior(q, s.star)

    n <- round(const * delta^q)
    print(c(delta, n))
    estimates <- replicate(k, ComputeABCEstimate(delta, n, q, s.star))
    tmp <- (estimates - exact)^2
    return(c(mean(tmp), sd(tmp)/sqrt(k)))
}

OptimalDeltaAndMSE <- function(expected.cost, k, q, s.star) {
    deltas <- seq(0.1, 1, by = 0.1)
    MSEs <- sapply(deltas, EstimateMSE, k, q, s.star, expected.cost)
    deltas.nq <- deltas^(-q)
    deltas.4 <- deltas^4
    fit <- lm(MSEs[1, ] ~ deltas.nq + deltas.4 + 0,
              weights = MSEs[2, ])
    coef.nq <- fit$coefficients[1]
    coef.4 <- fit$coefficients[2]
    # curve is ad^(-q)+bd^4,
    # so opt at -qad^(-q-1)+4bd^3=0 => d=(aq/4b)^(1/(q+4))
    opt.delta <- (coef.nq * q/(4 * coef.4))^(1/(q + 4))
    opt.MSE <- coef.nq * opt.delta^(-q) + coef.4 * opt.delta^4
    return(c(expected.cost, opt.delta, opt.MSE))
}

costs <- 2^(0:8) * 500
data <- t(sapply(costs, OptimalDeltaAndMSE, 500, 2, c(1, 1)))
write.table(data, "fig3.dat", row.names = F,
            col.names = c("exp. cost", "opt. delta", "opt. MSE"))

# fig3-plot.R - draw figure 3, using data generated by fig3.R
#
# Copyright (C) 2013  S. Barber, J. Voss, M. Webster
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

data <- read.table("fig3.dat", header = T)
q <- 2
costs <- data[, 1]/min(data[, 1])
```

```
deltas <- data[, 2]
errors <- data[, 3]
num.c <- length(costs)
short.c <- costs[floor(1 + num.c/4):ceiling(num.c * 3/4)]

pdf("fig3.pdf", width = 4.6, height = 2.5, family = "serif",
    pointsize = 10)
par(oma = c(0, 0, 0, 0), mai = c(0.5, 0.5, 0.1, 0.1), mfrow = c(1, 2))

plot(costs, deltas, log = "xy", xlab = "expected cost",
     ylab = expression(delta), mgp = c(1.8, 0.6, 0))
ld <- log(deltas)
lc <- log(costs)
fit <- lm(ld ~ lc)
coefs <- summary(fit)$coefficients
lines(costs, exp(fitted.values(fit)))
expected.gradient <- -1/(q + 4)
lines(short.c,
      exp(coefs[1, 1] + expected.gradient * log(short.c) + 0.1))
print(paste("Delta order is", coefs[2, 1], "s.e.", coefs[2, 2]))
print(paste("Expected order is", expected.gradient))

plot(costs, errors, log = "xy", xlab = "expected cost", ylab = "MSE",
     mgp = c(1.8, 0.6, 0))
le <- log(errors)
lc <- log(costs)
fit <- lm(le ~ lc)
coefs <- summary(fit)$coefficients
lines(costs, exp(fitted.values(fit)))
expected.gradient <- -4/(q + 4)
lines(short.c,
      exp(coefs[1, 1] + expected.gradient * log(short.c) + 0.4))
print(paste("Error order is", coefs[2, 1], "s.e.", coefs[2, 2]))
print(paste("Expected order is", expected.gradient))

invisible(dev.off())
```

Figure 4

```
set.seed(52228)

sigma.theta <- 1
sigma.x <- 1
ind.lower <- -0.5
ind.upper <- 0.5

d <- 2
q <- 1
x.star <- c(1, 1)
s.star <- 1

# Generate n ABC samples for the posterior distribution of theta.
GenerateABCSamples <- function(n, delta) {
    accepted <- 0
    samples <- numeric(n)
    while (accepted < n) {
        theta <- rnorm(1, sd = sigma.theta)
        X <- rnorm(q, mean = theta, sd = sigma.x)
        if (sum((X - s.star)^2) <= delta^2) {
            accepted <- accepted + 1
            samples[accepted] <- theta
        }
    }
    return(samples)
}

# Return one ABC estimate, using n ABC samples.
```

```
GetABCEstimate <- function(n, delta) {
    theta <- GenerateABCSamples(n, delta)
    Z <- ind.lower <= theta & theta <= ind.upper
    est <- mean(Z)
    return(est)
}

# Compute the exact result of the estimation problem.
# This is used to assess the quality of the ABC estimates.
GetTruePosterior <- function() {
    inter <- d * sigma.theta^2 + sigma.x^2
    inter2 <- d * mean(x.star) * sigma.theta^2/inter
    inter3 <- sigma.theta * sigma.x/sqrt(inter)
    lower <- pnorm(ind.lower, mean = inter2, sd = inter3)
    upper <- pnorm(ind.upper, mean = inter2, sd = inter3)
    return(upper - lower)
}

# Return a Monte Carlo estimate of the bias, using k ABC estimates.
# Returns the estimate and its standard error.
EstimateBias <- function(delta, k, n) {
    exact <- GetTruePosterior()
    biases <- replicate(k, GetABCEstimate(n, delta)) - exact
    return(c(mean(biases), sd(biases)/sqrt(k)))
}

deltas <- seq(0.05, 2, by = 0.05)
k <- 5000
n <- 10
biases <- t(sapply(deltas, EstimateBias, k, n))
write.table(cbind(deltas, biases), "fig4.dat", row.names = F,
            col.names = c("delta", "bias", "sd"))

data <- read.table("fig4.dat", header = T)
```

```
deltas <- data[, 1]
biases <- data[, 2]
ses <- data[, 3]

pdf("fig4.pdf", width = 4.6, height = 3, family = "serif",
    pointsize = 10)
par(oma = c(0, 0, 0, 0), mai = c(0.8, 0.8, 0.15, 0.1))

plot(deltas, biases, xlim = c(0, max(deltas)),
     ylim = c(0, max(biases + 1.96 * ses)),
     xlab = expression(delta), ylab = "bias", cex = 0.7)
arrows(deltas, biases - 1.96 * ses, deltas, biases + 1.96 * ses,
       0.7 * 0.05, 90, 3)

C.parabola <- -0.018338 - 0.3647609 * (-0.138889)
t <- seq(0, max(deltas), length.out = 100)
lines(t, C.parabola * t^2)

invisible(dev.off())
```

B.1.2 ABC Posterior Plots

The code given below includes a TCL/TK section, allowing interactive sliders to change the observation and the tolerance. Compiling only the commands given before the TCLTK package is required will give functions to make static plots, like those in the main text. The TCL/TK section is modified from the eponymous package's demo code.

```
cdfs <- function(x.star,delta,cmean,cvar) {
# cdf of x*+delta-cmean/sqrt(cvar) minus that of
# x*-delta-cmean/sqrt(cvar)
    pnorm(x.star,
          mean=cmean-delta,
          sd=sqrt(cvar)) - pnorm(x.star,
                                 mean=cmean+delta,
                                sd=sqrt(cvar))
}

dabcpost <- function(x,x.star,pmean,pvar,datvar,delta) {
    dnorm(x, mean=pmean, sd=sqrt(pvar) ) *
    cdfs(x.star, delta, x, datvar) /
    cdfs(x.star, delta, pmean, pvar+datvar)
}

ABCplot <- function(x.star,delta,xlim,ylim) {
    curve(dnorm(x,mean=x.star/2,sd=1/sqrt(2) ) ,
          lty="dashed",xlab=expression(theta) ,
          ylab="Density",main="ABC Posterior Error",
          xlim=xlim,ylim=ylim)
    curve(dnorm,lty="dotted",add=T)
    curve(dabcpost(x,x.star,pmean=0,pvar=1,datvar=1,
                   delta=delta) ,
          add=T)
    legend(x="topright",
           c("Prior","True posterior","ABC posterior") ,
           lty=c("dotted","dashed","solid") )
}

require(tcltk) || stop("tcltk support is absent")
require(graphics); require(stats)
local({
    have_ttk <- as.character(tcl("info", "tclversion")) >= "8.5"
    if(have_ttk) {
        tkbutton <- ttkbutton
        tkframe <- ttkframe
        tklabel <- ttklabel
        tkradiobutton <- ttkradiobutton
    }
```

```
xlim <- c(-5,5)
ylim <- c(0,0.6)
x.star <- tclVar(3)
x.star.sav <- 3
bw    <- tclVar(1)
bw.sav <- 1 # in case replot.maybe is called too early

replot <- function(...) {
    bw.sav <<- b <- as.numeric(tclObj(bw))
    x.star.sav <<- xs <- as.numeric(tclObj(x.star))
    eval(substitute(ABCplot(xs,b,xlim,ylim)))
}

replot.maybe <- function(...)
{
    if (as.numeric(tclObj(bw)) != bw.sav ||
        as.numeric(tclObj(x.star)) != x.star.sav) replot()
}

regen <- function(...) {
    xlim <<- c(min(0,as.numeric(tclObj(x.star) ) /2) -5,
               max(0,as.numeric(tclObj(x.star) ) /2) +5)
    replot()
}

grDevices::devAskNewPage(FALSE) # override setting in demo()
tclServiceMode(FALSE)
base <- tktoplevel()
tkwm.title(base, "Density")

spec.frm <- tkframe(base,borderwidth=2)
right.frm <- tkframe(spec.frm)
```

```
    frame3 <-tkframe(right.frm,relief="groove",borderwidth=2)
    tkpack(tklabel(frame3,text="Observation") )
    tkpack(tkscale(frame3,command=replot.maybe,from=1,to=10,
                   showvalue=T,variable=x.star,
                   resolution=0.1,orient="horiz") )

    frame4 <-tkframe(right.frm, relief="groove", borderwidth=2)
    tkpack(tklabel (frame4, text="Tolerance"))
    tkpack(tkscale(frame4, command=replot.maybe, from=0.05, to=16.00,
                   showvalue=T, variable=bw,
                   resolution=0.05, orient="horiz"))

    tkpack(frame3,frame4, fill="x")
    tkpack(right.frm,side="left", anchor="n")

    ## ‘Bottom frame’ (on base):
    q.but <- tkbutton(base,text="Quit",
                      command=function() tkdestroy(base))

    tkpack(spec.frm, q.but)
    tclServiceMode(TRUE)

    regen()
})
```

Bibliography

D. F. Anderson, G. Craciun, and T. G. Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Bulletin of Mathematical Biology*, 72(8):1947–1970, 2010.

J. C. Baez and B. Fong. Quantum techniques for studying equilibrium in reaction networks. *Journal of Complex Networks*, 2014.

S. Barber, J. Voss, and M. Webster. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9:80–105, 2015.

M. A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

G. Bertorelle, A. Benazzo, and S. Mona. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, 19(13):2609–2625, 2010.

G. Biau, F. Cérou, and A. Guyader. New insights into approximate Bayesian computation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 51(1):376–403, 2015.

M. G. B. Blum. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, 2010.

M. G. B. Blum and O. François. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

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

R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18 (2):125–135, 2008.

J. C. Burkill. *A First Course in Mathematical Analysis*. Cambridge University Press, 1962.

E. O. Buzbas and N. A. Rosenberg. AABC: Approximate approximate Bayesian computation for inference in population-genetic models. *Theoretical Population Biology*, 99:31–42, 2015.

K. Csilléry, M. G. B. Blum, O. E. Gaggiotti, and O. Franois. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25: 410–418, 2010.

T. A. Dean and S. S. Singh. Asymptotic behaviour of approximate Bayesian estimators. *ArXiv e-prints*, 2011.

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B*, 46(2): 193–227, 1984.

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

Y. Fan, D. J. Nott, and S. A. Sisson. Approximate Bayesian computation via regression density estimation. *Stat*, 2(1):34–48, 2013.

P. Fearnhead and D. 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.

W. Feller. *An Introduction to Probability Theory and its Applications.* Wiley, 1968.

D. T. Frazier, G. M. Martin, and C. P. Robert. On consistency of approximate Bayesian computation. *ArXiv e-prints*, 2015.

R. A. Graham, D. E. Knuth, and O. Pateshnik. *Concrete Mathematics.* Addison-Wesley, 1994.

M. A. Haynes, H. L. MacGillivray, and K. L. Mengersen. Robustness of ranking and selection rules using generalised g-and-k distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66, 1997.

G. Huber. Gamma function derivation of n-sphere volumes. *The American Mathematical Monthly*, 89(5):301–302, May 1982.

A. Jasra, S. S. Singh, J. S. Martin, and E. McCoy. Filtering via approximate Bayesian computation. *Statistics and Computing*, 22(6):1223–1237, 2012.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

P. M. Lee. *Bayesian Statistics: An Introduction.* Wiley, 2012.

Christoph Leuenberger and Daniel Wegmann. Bayesian computation and model selection without likelihoods. *Genetics*, 184(1):243–252, 2010.

W. Li and P. Fearnhead. Improved convergence of regression adjusted approximate Bayesian computation. *ArXiv e-prints*, 2016.

W. V. Li and Q-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19:533–597, 2001.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms.* Cambridge University Press, 2002.

J-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100 (26):15324–15328, 2003.

E. Meeds and M. Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 593–602. AUAI Press, 2014.

J. Mitrovic, D. Sejdinovic, and Y. W. Teh. DR-ABC: Approximate Bayesian computation with kernel-based distribution regression. In *International Conference on Machine Learning (ICML)*, pages 1482–1491, 2016.

B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2003.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. 2015.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.

D. Prangle. *Summary Statistics and Sequential Methods for Approximate Bayesian Computation*. PhD thesis, April 2011.

D. Prangle. Lazier ABC. *ArXiv e-prints*, 2015.

D. Prangle. Lazy ABC. *Statistics and Computing*, 26(1):171–185, 2016.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, December 1984.

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

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, September 1994.

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

S. A. Sisson, Y. Fan, and M. M. Tanaka. Correction for Sisson et al., Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 106(39):16889, 2009.

J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer, 2002.

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

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

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009.

J. Voss. *An Introduction to Statistical Computing: A Simulation-based Approach*. Wiley, 2013.

X. Wang. Volumes of generalized unit balls. *Mathematics Magazine*, 78(5): 390–395, December 2005.

D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

R. 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, May 2013.

R. D. Wilkinson. Accelerating ABC methods using Gaussian processes, 2014.

D. Williams. *Probability With Martingales*. Cambridge University Press, 1991.

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104, 2010.

X. Zhong and M. Ghosh. Approximate Bayesian computation via sufficient dimension reduction. *ArXiv e-prints*, 2016.

J. Zhou and K. Fukumizu. Local kernel dimension reduction in approximate Bayesian computation. *ArXiv e-prints*, 2016.