Sequential Bayesian Quadrature

by

Thomas D. Cook

and

Murray K. Clayton
Abstract

Many numerical problems involving properties of unknown functions can be understood as statistical problems by viewing the function as a realization of a random process and viewing function evaluations as collecting “data” about the function. Procedures of this type have long history, although very few have found their way into common usage. Essentially Bayesian in nature, a typical scheme is as follows. Start with a prior distribution on a space of functions of which the given function is a member. Evaluate the function at some set of points and generate a posterior distribution. Inference about properties of the function such as its value(s) at some additional point(s) (interpolation), its maxima or minima, or its integral can be made from the posterior. Often additional function evaluations can be made sequentially. Since the posterior gives a natural way of assessing our estimates of these properties, by choosing designs which optimize some criterion (like posterior expected loss) we can make maximal use of each function evaluation. For “expensive” functions this may be highly desirable. A flexible class of hierarchical Gaussian priors is proposed and subsequently used in a myopic sequential procedure. The result is an efficient quadrature rule which is shown to outperform fixed-sized rules for functions generated under the proposed model.
1 Introduction

Suppose that we are faced with a numerical analysis problem such as interpolating \( f(x) \), or finding \( \int_a^b f(x) \, dx \) for some function \( f \) and known constants \( a \) and \( b \). If \( f \) is well-behaved, we might be able to determine its integral analytically, or, in a statistical setting, \( f \) might correspond to a parametric distribution for which such integrals have been tabulated. In general, however, we must rely on traditional numerical techniques. In this paper we will be concerned with Bayesian numerical analysis, a method that serves as an alternative to these more common approaches. Our focus will be on the integration problem, although much of the paper can also be applied to interpolation.

The thing that distinguishes Bayesian numerical analysis from more standard approaches is the view that the function \( f \) represents a quantity about which we are uncertain. In particular, for complicated functions (that might require evaluation on a computer) we do not know the function in the sense that, until we actually evaluate it at particular points, we do not know its precise values. On the other hand, we typically do have some initial understanding about it. For example, we may know that it is continuous, or that it possesses some number of continuous derivatives, or that it takes some particular values at special points such as the origin. Following the Bayesian paradigm, we assume that our knowledge and our uncertainties about the function can be summarized by a prior distribution on a space of functions in which \( f \) lives. We then gather data regarding \( f \) by evaluating it for various values of \( x \). Based on these observations, we can determine the posterior distribution of \( f \) and use this to make inferences about properties of \( f \).

Bayesian numerical analysis, therefore, turns a numerical analysis problem into a problem of statistical inference. Interesting issues then become those of modeling (what is an appropriate method for constructing a prior for \( f \)?), inference (can we actually calculate the posterior for \( f(x) \)?) and design (what is the best way to allocate \( x \) values to obtain a good estimate of the property in question?). This paper addresses these three issues by introducing a hierarchical Gaussian process model as the prior for the unknown \( f \). The model has strengths that we find valuable, but inference is not trivial. We therefore discuss a number of approximations that are useful in computing posteriors. In addition, we take a decision-theoretic, sequential approach to the design problem. We assume that the goal is to estimate the integral of \( f \) using squared error loss, with a cost \( c \) per observation. A sequential rule then must dictate where to locate the \( x \) values and when to stop taking observations. Finding optimal sequential rules in this context is extraordinarily difficult. We will therefore focus on "myopic" rules and modifications of myopic rules; in myopic rules, at each stage we minimize the risk for a single additional observation. (See Berger (1985) for further discussion of myopic rules.) Although myopic rules are suboptimal, we will see that they have good properties for the problem at hand.

We know of no literature that takes a Bayes sequential decision theoretic approach to numerical analysis problems. Indeed, the literature on Bayesian quadrature in general is relatively limited. Diaconis (1988) has written an interesting historical account of various kinds of Bayesian numerical methods; he provides further discussion of the rationale for taking a Bayesian viewpoint in numerical analysis. He notes that a number of standard approaches can be viewed as being optimal rules from a Bayesian perspective. For example,
the trapezoid rule emerges as a Bayes rule for squared error loss if the prior on \( f \) is Brownian motion. If the prior is once-integrated Brownian motion, the Bayes rule dictates use of a cubic spline interpolant. Although not strictly a Bayesian result, it is of note that the usual Monte Carlo approach is a minimax rule (Diaconis, 1988). O’Hagan (1988) also discusses Monte Carlo quadrature, and shows it to arise as a limit of Bayes rules.

O’Hagan (1988, 1992) considers optimal fixed sample designs for Bayesian quadrature for some simple cases. Currin, Mitchell, Morris and Ylvisaker (1991) use Gaussian-process priors to find fixed-sample designs for interpolation problems in several dimensions which are optimal with respect to a posterior entropy criterion.

Of course, some non-Bayesian approaches to quadrature also treat the integration problem as a statistical one. Monte Carlo methods are prominent in this regard. Monte Carlo methods are most useful when it is relatively inexpensive to generate large numbers of samples from the distributions of interest. By contrast, this paper focuses on the case where function evaluations are expensive and therefore optimal designs and efficient use of function evaluations are critical.

The model we develop in this paper is based on an extension of the work of O’Hagan (1992). We now review O’Hagan’s approach in some detail; this will allow us to develop the background and notation for our approach.

O’Hagan (1992) considers a hierarchical model that incorporates a random regression component into the mean of \( f \). To simplify the following discussion of his model, we will assume that \( f \) has mean 0. His model then becomes:

\[
f(\cdot) | \sigma^2 \sim N(0, \sigma^2 v(\cdot, \cdot))
\]

and

\[
\frac{s_0}{\sigma^2} \sim \chi^2_{a_0}.
\]

That is, conditional on the parameter \( \sigma^2 \), \( f(\cdot) \) is a Gaussian process defined on some space \( \mathcal{X} \subset \mathcal{R}^a \), with mean 0 and covariance function \( \sigma^2 v(x, x') \), and \( \sigma^2 \) has an an inverse chi-squared distribution. Hence, the marginal distribution of \( f(x) \) for any \( x \) will be a t-distribution with mean 0 and variance \( s_0 v(x, x)/(a_0 - 2) \).

We will assume throughout that \( v(x, x') \) is positive definite. The covariance function \( v \) should be chosen to reflect the degree of smoothness we believe \( f \) to possess, and yet be mathematically tractable. In particular (see Equation (5) below) for the quadrature problem, we will need to be able to analytically evaluate integrals of \( v(\cdot, x) \).

Suppose that we have observed \( f(x) \) at \( x_1, x_2, \ldots, x_n \). Let \( f = (f(x_1), f(x_2), \ldots, f(x_n))^T \), \( t(x) = (v(x, x_1), v(x, x_2), \ldots, v(x, x_n))^T \), \( a = a_0 + n \) and define the matrix of covariances:

\[
A = \begin{pmatrix}
  v(x_1, x_1) & \cdots & v(x_1, x_n) \\
  \vdots & \ddots & \vdots \\
  v(x_n, x_1) & \cdots & v(x_n, x_n)
\end{pmatrix}
\]

Based on these quantities, let \( m(x) = t(x)^T A^{-1} f \), \( w(x, x') = v(x, x') - t(x)^T A^{-1} t(x') \), and \( s = s_0 + f^T A^{-1} f \).

Then from O’Hagan (1992), the posterior distribution for \( f \) will be

\[
f(\cdot) | \sigma^2, f \sim N(m(x), \sigma^2 w(\cdot, \cdot))
\]
with
\[ \frac{s}{\sigma^2} f \sim \chi^2. \]

Hence, at any particular \( x \) the posterior distribution for \( f(x) \) is a t-distribution with mean \( m(x) \) and variance \( sw(x, x)/(a - 2) \).

In the quadrature interest centers on the posterior distribution of \( I = \int_x f(x)dx \). Conditional on \( f \) and \( \sigma^2 \), the distribution of \( I \) is given by O’Hagan as follows.

\[ I|f, \sigma^2 \sim N(m^T, \sigma^2 w^T) \]

where
\[ m^T = \int_x m(x)dx = \int_x t^T(x)dx A^{-1} f, \]
and
\[ w^T = w^T(x_1, \ldots, x_n) = \int_x \int_x w(x, x')dx'dx' - \int_x t^T(x)dx A^{-1} \int_x t(x)dx. \]

Thus, the posterior distribution for \( I \) is a t-distribution with mean \( E[I|f] = m^T \) and variance
\[ \text{var}(I|f) = \frac{sw^T}{a - 2}. \]

O’Hagan (1988, 1992) generates designs which a priori minimize (7) for special cases. Note that the right hand side of (7) factors into \( s \), which is a \( \chi^2 \) random variable (given \( \sigma^2 \)) and therefore depends on the design only via the sample size, and \( w^T \), which does not depend on the observed values of \( f \). Thus, a limitation of models of the form in Equation (1) is that any variation in the smoothness of sample functions over the range \( \mathcal{X} \) is determined by the prespecified (and fixed) function \( v(\cdot, \cdot) \). Consequently, this model fails to "learn" about any changes in the smoothness of \( f \) as observations are taken. The model in effect averages all observed variability into the posterior distribution of \( \sigma^2 \) and can therefore lead to incorrect estimates of the posterior variance of \( f \).

Second, and probably more important, the model does not allow us to generate sequential designs that allocate observations more generously to regions of \( \mathcal{X} \) where we detect increased variability. The model that we develop in this paper is a hierarchical model that incorporates such variability into the prior distribution. We describe that model and discuss the posterior of \( f \) in the next section. In Section 3 we discuss the computation of the posterior distribution of the variance of \( f \) and in Section 4 we will show how our model can be used to derive a sequential (adaptive) procedure that allocates observations according to posterior information about local variability. Section 4.3 contains a series of technical results supporting the myopic rule. Finally, in Section 5 we provide some numerical results that demonstrate the efficiency of the proposed allocation rule.

2 A More General Model

In developing a hierarchical model we must address three principal issues. First, since it is a Bayesian model, it must agree with our prior opinion about the nature of the function we wish to integrate. Second, our approach replaces the original integration problem with
the integral of an interpolant; thus, we must be able to analytically perform the second integration. Hence, the final fitted model must be of a sufficiently simple form. Third, we must be able to perform the necessary model fitting efficiently, otherwise the cost of fitting the model will exceed any savings in the number of function evaluations. With these issues in mind, we propose the following model. We will assume throughout that the interval of integration is \([0, 1]\).

We will use a model based on once-integrated Brownian motion. This allows us to model continuous functions that have at least one continuous derivative (or are at least piecewise so). While we may believe that our function possesses additional derivatives or other properties, models based on m-fold integrated Brownian motion or other Gaussian processes rapidly become increasingly complex and it is not clear that the additional cost of model fitting can be offset by the improvement in the model.

The primary difference between our model and O’Hagan’s is that we replace the scalar random variable \(\sigma^2\) by a random process \(g^2(x)\). First, let \(z(x)\) be a Gaussian process defined on the interval \([0, 1]\) with mean \(\tau\) and covariance

\[
\text{cov}(z(x), z(y)) = \zeta^2 + \frac{\eta^2}{4}(1 - 2|x - y|) \tag{8}
\]

for some known \(\zeta\) and \(\eta\). The process \(z\) is a stationary version of Brownian motion. We then define the process \(g(x)\) by

\[
g(x) = \exp(z(x)). \tag{9}
\]

Note that since \(z\) is stationary, \(g\) (and \(g^2\)) will be stationary as well. As we will see, the local contribution to the overall posterior variance of \(\int f\) is proportional to \(g^2\). Thus, an important feature of \(g\) is how the ratio \(g(x)/g(y)\) behaves. Because of the way \(g\) is defined, \(g(x)/g(y)\) is independent of both \(g(y)\) and \(g(x)\) and depends on \(x\) and \(y\) only through the difference \(|x - y|\).

The assumption of stationarity also imposes a restriction: it is desirable when \(a \text{ priori}\) our expectations about the local behavior of \(f\) are independent of \(x\). If \(a \text{ priori}\) we expect that \(f(x)\) is more variable in one particular region than another, we may wish to use a model for \(z(x)\) that incorporates this information. (We will not pursue this idea here.)

Fixing a realization of a standard Brownian motion path, we now construct the process \(f(x)\) using:

\[
f(x) = \alpha + \beta x + \int_0^1 (x - u)^+ g(u) dB(u) \tag{10}
\]

where \(dB\) can be viewed as an Itô integral with respect to Brownian motion \(B(x)\) (see, for example Ash and Gardner (1975)). As usual, \((x)^+ = \max(x, 0)\). Also, we will assume that \(\alpha\) and \(\beta\) have non-informative, improper prior distributions on \((-, \infty)\): \(\pi(\alpha, \beta) \propto 1\). This completes the model description. Our next step is to calculate the various conditional distributions needed to proceed with a Bayesian numerical analysis.

We first note the following facts, which will be used several times. If \(f(x)\) is a Gaussian process with finite mean and variance, then by Fubini’s theorem,

\[
E \left[ \int_a^b f(x) \, dx \right] = \int_a^b E[f(x)] \, dx, \tag{11}
\]
\[ \text{cov} \left( f(x), \int_a^b f(y) \, dy \right) = \int_a^b \text{cov} \left( f(x), f(y) \right) \, dy \] (12)

and,
\[ \text{var} \left( \int_a^b f(x) \, dx \right) = \int_a^b \int_a^b \text{cov} \left( f(x), f(y) \right) \, dx \, dy. \] (13)

It is straightforward to show that, conditional on \( g, \alpha \) and \( \beta \), \( f \) will have mean \( \alpha + \beta x \) (since \( B(x) \) has mean zero) and the following simple covariance structure:
\[ v_g(x, y) = \int_0^1 (x - u)^+ (y - u)^+ g^2(u) \, du \]

Hence,
\[ f(\cdot) | g, \alpha, \beta \sim N(\alpha + \beta x, v_g(\cdot, \cdot)). \]

Note that, since \( \alpha \) and \( \beta \) have improper prior distributions, so will \( f \). This fact complicates the computations slightly. However, as we next show, (see Equation (15)), the distribution of \( f \) given observations at two distinct values of \( x \) (and given \( g \)) will be proper and Gaussian with known mean and known covariance which will be a slightly modified version of \( v_g \) as given above. The results of the previous section will apply with only minor modifications to the conditional process.

Now suppose that \( 0 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n < x_{n+1} = 1 \) and let \( f = (f(x_1), f(x_2), \ldots, f(x_n))^T \) and \( f' = (f(x_1), f(x_n))^T \). (The quantities \( x_0 \) and \( x_{n+1} \) are introduced for notational convenience; we do not assume that observations are taken there. Also, although the bounds between \( x_1 \) and \( x_0 \) and between \( x_n \) and \( x_{n+1} \) need not be strict, in practice making them strict is advantageous, and so we impose that condition at the outset.) We begin by computing the posterior distribution of \( f(x) \) given \( f' \) and \( g \). Conditional on \( g \), but marginally with respect to \( \alpha \) and \( \beta \), this distribution will be Gaussian with known mean. Let \( h(x) = (1, x)^T \) and let

\[ H = \begin{pmatrix} 1 & x_1 \\ 1 & x_n \end{pmatrix}. \]

Now define \( \alpha'(u) = ((x_1 - u)^+, (x_n - u)^+)^T \), \( r(x) = (H^T)^{-1} h(x) = \frac{1}{x_n - x_1} (x_n - x, x - x_1)^T \), and
\[ P(u, x) = (x - u)^+ - \alpha'(u)^T r(x). \]

Then, marginally with respect to \( \alpha \) and \( \beta \), we have
\[ f(x) | f', g \sim N(m'(x), v'_g(x, y)) \] (15)

where \( m'(x) = h(x)^T H^{-1} f' \) and
\[ v'_g(x, y) = \int_0^1 P(u, x) P(u, y) g^2(u) \, du. \]

Clearly the mean \( m'(x) \) is a linear function of \( x \) (since \( h(x) \) is) and since it must agree with \( f \) at \( x_1 \) and \( x_n \), it must be the unique straight line interpolant to the data. In particular,
\(m'(x)\) is independent of \(g\). Furthermore, the conditional covariance \(v'_g\) has the same form as the unconditional covariance \(v_g\) if we replace the functions \((x - u)^+\) and \((y - u)^+\) by the functions \(P(u, x)\) and \(P(u, y)\). Interestingly, this covariance depends only on \(x_1\) and \(x_n\), and not on the observed data \(f'\). Since, as noted above, the mean depends only on \(f'\) and not on \(g\), there is no information about \(g\) in the observations \(f(x_1), f(x_n)\). Hence, when we write the likelihood for \(g\), it will depend only on the deviations of the remaining observations from their conditional expectations, \(E[f(x)|g, f']\).

Following Section 1, we can now compute the conditional distribution of \(f\) given \(f\). First, let \(f'' = (f(x_2), f(x_3), \ldots, f(x_{n-1}))^T\), \(P(u) = (P(u, x_2), P(u, x_3), \ldots, P(u, x_{n-1}))^T\) and \(m' = (m'(x_2), m'(x_3), \ldots, m'(x_{n-1}))^T\). Also, define \(t'_g(x) = \int_0^1 P(u) P(u, x) g^2(u) \, du\) and

\[
A'_g = \int_0^1 P(u) P(u)^T g^2(u) \, du.
\]

Then,

\[
f(\cdot)|g, f \sim N(m_g(\cdot), w_g(\cdot, \cdot))
\]

where

\[
m_g(x) = m'(x) + t'_g(x)^T A'^{-1}_g (f'' - m')
\]

and

\[
w_g(x, x') = v'_g(x, x') - t'_g(x)^T A'^{-1}_g t'_g(x').
\]

Finally, we want to write down the posterior distribution for the integral \(\mathcal{I}\) given the data and \(g\). Using (4), we may write the posterior distribution of \(\mathcal{I}\) as

\[
\mathcal{I}|f, g \sim N(m^\mathcal{I}_g, w^\mathcal{I}_g)
\]

where

\[
m^\mathcal{I}_g = \int_0^1 m_g \, dx.
\]

and

\[
w^\mathcal{I}_g = \int_0^1 \int_0^1 w_g(x, x') \, dx \, dx'
\]

(The similarity between Equation (22) and Equation (6) should be noted.) Note, furthermore, that if \(\hat{g}(x)\) is an estimate of \(g(x)\) (for example, the posterior mode) which depends on \(f_n\), and if \(\mathcal{I}\) is estimated by (21) with \(g\) replaced by \(\hat{g}\), then under squared-error loss with cost \(c\) per observation, the posterior expected loss incurred by taking these \(n\) observations can be written as

\[
ro(f_n) = E[(\mathcal{I} - m^\mathcal{I}_g)^2|f_n] + nc
\]

\[
= E[(\mathcal{I} - m^\mathcal{I}_g)^2|f_n] + E[(m^\mathcal{I}_g - m^\mathcal{I}_g)^2|f_n] + nc
\]

\[
= E[w^\mathcal{I}_g] + E[(m^\mathcal{I}_g - m^\mathcal{I}_g)^2|f_n] + nc
\]

(23)

In Section 4.4, we argue that the middle term of (23) can be ignored. Hence we can use an approximation to (23), \(\hat{ro}(f_n) = w^\mathcal{I}_g + nc\), where expectation with respect to \(g\) is replaced by evaluation at \(\hat{g}\).
The only piece that remains to be computed is the posterior distribution of $z$. We do so by computing the likelihood for its transform, $g$. This is facilitated by reformulating the model. Specifically, suppose that we have observed $f(x_1), f(x_2), \ldots, f(x_n)$, where $x_1 < x_2 < \ldots < x_n$. We can construct a version of second divided differences of the $f$’s by

$$f_i^* = \frac{f(x_{i+2}) - f(x_{i+1})}{x_{i+2} - x_{i+1}} - \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \text{ for } i = 1, 2, \ldots, n - 2.$$ 

This set of $n - 2$ random variables does not depend on $\alpha$ and $\beta$ (or the conditional mean $E[f(x)|f']$) since any linear terms are annihilated by the above transformations, and, along with any two of the original $f(x_i)$’s, is a one-to-one linear transformation of the original data. Furthermore, this set of $n - 2$ observations is sufficient for $g$ since, as we stated previously, there is no information about $g(x)$ in the observations $f(x_1), f(x_n)$. Thus, we may write the likelihood for $g$ as a function of $f^* = f_n^* = (f_1^*, f_2^*, \ldots, f_{n-2}^*)$.

Again, we have that $E[f^*] = 0$. Next, to find the covariance of $f^*$, we have the following useful facts. If $\theta_1, \theta_2 \in \mathbb{R}^{n-2}$ then

$$\text{cov}(\theta_1^T f''(u), \theta_2^T f''(u) | g, f') = \int_0^1 \left( \theta_1^T P(u) \right) \left( \theta_2^T P(u) \right) g^2(u) du$$

and

$$\text{cov}(\theta_1^T f''(u), f(x) | g, f') = \int_0^1 \left( \theta_1^T P_n(u) \right) P(u,x) g^2(u) du.$$ 

In particular, to compute covariances of linearly transformed data, we simply apply the corresponding transformations to the vector $P(u)$ and perform the integration. In the case of the $f_i^*$ above, we simply compute the functions $Q_i(u), i = 1, 2, \ldots, n - 2$ as follows:

$$Q_i(u) = \frac{P(u, x_{i+2}) - P(u, x_{i+1})}{x_{i+2} - x_{i+1}} - \frac{P(u, x_{i+1}) - P(u, x_i)}{x_{i+1} - x_i}.$$ 

(Note that $P(u, x) = 0$ if $x = x_1$ or $x_n$.) These functions $Q_i(u)$ are tent functions with support $[x_i, x_{i+2}]$, and have value $Q_i(x_{i+1}) = 1$. Therefore, $Q_i(u)Q_j(u) = 0$ if $|i - j| > 2$ and, similar to Equation (16), we have that

$$A_g^* = \text{cov}(f^* | g, f') = \int_0^1 Q(u)Q^T(u)g^2(u) du,$$ 

and $t_g^*(x) = \text{cov}(f^*, f(x)) = \int_0^1 Q(u)P(u, x)g^2(u) du$ where $Q(u) = (Q_1(u), Q_2(u), \ldots, Q_{n-2}(u))^T$.

Thus, the covariance matrix for $f^*$, $A_g^*$, is tridiagonal, so computations involving this matrix can be performed quite efficiently.

As a result, we can write the likelihood for $g$ simply as follows:

$$p(f^* | g) = \frac{1}{(2\pi)^{n/2} |A_g^*|^{1/2}} \exp \left( -\frac{1}{2} f^{*T} A_g^{-1} f^* \right),$$

where dependence on $g$ is through the matrix $A_g^*$. In particular, the likelihood function for $g$ is a function of the quantities $\int_0^1 Q_i(u)Q_j(u)g^2(u) du$. Henceforth, we will consider the vector $f^*$ to be the “data” and will focus on the posterior distribution of $z$. 

7
3 The Posterior Distribution for $z(x)$

Our ultimate goal is to find the posterior distribution of $I$ given the data. In theory, we need only marginalize over $g$ in Equation (20) to do this. In practice, however, this is very difficult. Our approach, instead, will be to use the posterior mode of $z(x)$ which we then use to estimate the posterior distribution of $I$. Even the determination of the posterior mode of $z(x)$ is difficult, since it requires an optimization over a function space. We therefore use an approximation to the likelihood of $g$ that reduces the problem to a finite dimensional optimization problem.

For $x_k \leq x < x_{k+1}$, let

$$\bar{z}(x) = z_k = \int_{x_k}^{x_{k+1}} z(u) du / (x_{k+1} - x_k)$$

and define $\tilde{g}(x) = \exp(\bar{z}(x))$, $g_k = \exp(z_k)$ and $a_k = g_k^2(x_{k+1} - x_k)$. Then, for $j \geq i$ and $k = 1, 2, \ldots, n-1$,

$$\int_{x_k}^{x_{k+1}} Q_i(u) Q_j(u) g^2(u) du \approx \int_{x_k}^{x_{k+1}} Q_i(u) Q_j(u) \tilde{g}^2(u) du = g_k^2 \int_{x_k}^{x_{k+1}} Q_i(u) Q_j(u) du$$

$$= \begin{cases} a_k / 3 & \text{if } i = j = k \\
                               a_k / 6 & \text{if } i = j = k - 1 \\
                               0 & \text{otherwise} \end{cases}$$

This provides an approximation, $A^*_g$, for $A^*_g$ in (24). (The approximation \( \approx \) could be replaced by a more formal limit result, but the above statement is adequate for our purposes.)

In addition to providing obvious computational advantages over the true likelihood, this approximation also allows us to consider the posterior distribution of the vector $z = (z_1, z_2, \ldots, z_{n-1})^T$ rather than the entire process $z$, or the set of $\int_{x_k}^{x_{k+1}} Q_i(u) Q_j(u) g^2(u) du$ whose joint distribution is mathematically complicated. Clearly, a priori, $z$ is multivariate normal, its mean is $\tau = (\tau, \tau, \ldots, \tau)^T$ and it has covariance

$$\text{cov}(\bar{z}_i, \bar{z}_j) = \begin{cases} \sigma^2 + \frac{\sigma^4}{4} (1 - |(x_{j+1} + x_j) - (x_{i+1} + x_i)|) & \text{if } i \neq j \\
\sigma^2 + \frac{\sigma^4}{4} (1 - \frac{3}{2}(x_{i+1} - x_i)) & \text{if } i = j \end{cases}$$

Letting $C = \text{cov}(z)$, we can write the approximate marginal posterior density for $z$ as:

$$\frac{1}{|C|^{1/2}} \exp \left( -\frac{1}{2} (z - \tau)^T C^{-1} (z - \tau) \right) \frac{1}{|A^*_g|^{1/2}} \exp \left( -\frac{1}{2} f^* T A^*_g^{-1} f^* \right)$$

A further simplification to this likelihood which makes the computations much simpler is to ignore the correlations in $A^*_g$. In this case the log-likelihood becomes:

$$\log(p(f|g)) \approx K - \frac{1}{2} \sum_{i=1}^{n-2} \left[ \log(a_i + a_{i+1}) + \frac{3 f_i^*}{a_i + a_{i+1}} \right]$$
While it is computationally intractable to compute features such as the posterior mean and variance of this distribution, it is straightforward to compute derivatives and implement the Newton-Raphson algorithm to compute the posterior mode.

It is interesting to note that if we consider the log of the posterior density,

$$
\log(p(z|f_n)) = K - \frac{1}{2}(z - \tau)^T C^{-1}(z - \tau) + \log(p(f|g))
$$

we see that it has a log-prior term and a log-likelihood term. We can view the former term as a roughness penalty and view $1/\eta^2$ and $1/\zeta^2$ as smoothing parameters. Note that this term is the Hilbert space norm of $\int_{\mathcal{U}} z(u) du$ in a suitably defined reproducing kernel Hilbert space. (See Wahba (1990).) If $\eta^2$ and $\zeta^2$ are large, then the “penalty” for deviating from the prior mean will be small, and $z$ will be allowed to follow the data. Consequently the posterior mode will be near the maximum likelihood estimate if it exists. If $\eta^2$ is small, then the penalty term will tend to dominate, forcing $z$ to be nearly constant (and so the estimate of $g(x)$ will be flat), and if $\zeta^2$ is also small, the posterior mode will be near $\tau$, in spite of the data. In any event, because of the correlation structure of the prior process, $z$ will not be allowed to change too radically from one interval to the next, and so the prior distribution is forcing the estimate of $z$ to be relatively smooth. This interpretation of the prior distribution on $z(x)$ has appeal even to those who reject the Bayesian interpretation.

4 Allocation Rules

In this section we will describe how to use our model to allocate observations using a sequential rule. As noted in the introduction, we will focus on myopic rules that minimize squared error loss with a cost $c$ per observation.

4.1 Approximate Myopic Rule

From (21), our estimate of $\mathcal{I}$ after $n$ observations is given by $m_{n,g}^T$, obtained by replacing $g$ with our estimate of its posterior mode $\hat{g}_n$. The subscript $n$ is added to $m_0$ to emphasize that the estimate is based on $n$ observations. Suppose that we have observed $f_n$ and wish to employ an allocation rule, $\rho$, which prescribes that we sequentially sample at $x_{n+1}, x_{n+2}, \ldots$ until we have $N \geq n$ total observations ($N$ and $x_{n+2}, x_{n+3}, \ldots$ may be random variables). Then, generalizing (23), the expected posterior loss for $\rho$ will be

$$
r_\rho(f_n) = E[w_{N,g}^T f_n] + E[(m_{N,g}^T - m_{N,\hat{g}_n}^T)^2 f_n] + cE[N f_n]. \tag{26}
$$

The posterior variance, $w_{N,g}^T$, is given by (22), with the subscript $N$ again emphasizing that it is evaluated after $N$ observations. Note that the expectation is with respect to the posterior distribution of $f(x)$ (and implicitly $g(x)$) and depends on $\rho$ through $N$ and the allocation of future values of $x$. The optimal (Bayes) sequential rule is the rule that minimizes $r_\rho(f_n)$ among all allowable allocation rules, $\rho$. The Bayes rule would prescribe that we either take an additional observation at the value of $x_{n+1}$ that minimizes $E[r_\rho(f_{n+1})|f_n]$ or stop sampling in the event that $r_\rho(f_n) < E[r_\rho(f_{n+1})|f_n]$ for all possible values of $x_{n+1}$.
Unfortunately, the computation of \( E[r_T(f_{n+1})|f_n] \) is complicated by two factors. As noted, \( N \) is, in general, a random variable that will depend in a complicated way on the (as yet unobserved) values \( f(x_{n+1}), f(x_{n+2}), \ldots \) making the computation of expectations extremely difficult, if not impossible. Second, the posterior distribution of \( g(x) \) is complicated and expectations with respect to \( g(x) \) are equally intractable. We address the first problem by restricting attention to myopic rules so that at any stage the only values of \( N \) which we consider are \( n \) and \( n+1 \). We address the second difficulty by replacing expectation with respect to \( g(x) \) by evaluation at our estimate of its posterior mode. Furthermore, in Section 4.4 we provide a heuristic argument showing that the middle term in (26) is negligible. Our experience suggests that these approximations are adequate to ensure that the sequential rule which results will have good properties.

Since, at each stage under the myopic rule, \( N \) is no longer considered random (the only possibilities that we need to consider are that we might stop or take exactly one additional observation) and given \( g, w_N^T \) (for either \( N = n \) or \( N = n+1 \)) does not depend on \( f_n \), we may remove the expectation from the first term of (26). As stated above, we also ignore the second term of (26). Therefore, rather than minimizing \( r_T(f) \), we compute

\[
\hat{r}_1(f_n) = \min_{N=n,n+1} \left\{ w_{N,\hat{g}_n}^T + cN \right\}
\]

(27)

To apply a myopic rule, we compute \( w_{n,\hat{g}_n}^T \) and the minimum of \( w_{n+1,\hat{g}_n}^T + c \) over all possible values of \( x_{n+1} \). If the former is smaller, we stop sampling and \( n \) is the final sample size. If the latter is smaller, we take an additional observation at the optimal \( x_{n+1} \), and continue myopically. Equivalently, we chose to compute the decrease in posterior variance, \( w_{n,\hat{g}_n}^T - w_{n+1,\hat{g}_n}^T \), resulting from an optimally selected value of \( x_{n+1} \), deciding to continue or stop depending on whether the magnitude of the change is greater or less than \( c \). Easy computation shows that this difference is given by

\[
w_{n,\hat{g}_n}^T - w_{n+1,\hat{g}_n}^T = \frac{\text{cov}(T, f(x)|f_n, g = \hat{g}_n)^2}{\text{var}(f(x)|f_n, g = \hat{g}_n)\cdot \var(f(x)|f_n, g = \hat{g}_n)}. \]

(28)

To proceed, we make use of the fact that:

\[
\text{cov}(T, f(x)|f, g) = \bar{v}_g(x) - t_g^*(x)^T A_g^* \bar{t}^*
\]

(29)

and

\[
\text{var}(f(x)|f, g) = \bar{v}_g'(x, x') dx' = \int_0^1 P(u, x)^2 du, \bar{t}^* = \int_0^1 t_g^*(x) \bar{d} = \int_0^1 Q(u) \bar{P}(u) g^2(u) du \]

(30)

where \( \bar{v}_g(x) = \int_0^1 v_g'(x, x') dx' = \int_0^1 P(u, x) \bar{P}(u) g^2(u) du \)

and \( \bar{P} = \int_0^1 P(u, x) \bar{d} \).

Since our approximation \( \hat{g}_n \) is constant between data points, both numerator and denominator of the righthand side of Equation (28) are piecewise polynomials in \( x \). By computing the coefficients of these polynomials it will be a straightforward matter to find the maximum in each interval, and ultimately to find the global maximum.

Before moving on, we note that it is possible to generalize this rule to a group sequential rule. That is, we allocate observations in groups of possibly varying sizes at each stage. In this case (28) becomes:

\[
\text{cov}(\mathcal{I}, f|f_n, g)^T \text{cov}(f|f_n, g)^{-1} \text{cov}(\mathcal{I}, f|f_n, g)
\]
where $\mathbf{f}^t$ is a vector of future observations of $f$. In this case, the maximization needs to be done over a higher dimensional space and could be quite difficult. We do not address this issue here, but rather leave it for future work.

4.2 Computing the Reduction in Variance

As we shall see momentarily, the second term of the right-hand side of Equation (29), depends on $x$ through $t^*(x)$ which, if $g(x)$ is constant between observations, has degree at most 3 for $x_1 < x < x_n$ and degree 1 for $x < x_1$ and $x > x_n$. (If $x = x_i$ for some $i$, then the reduction in variance is 0, and thus we ignore such cases.) Furthermore, the first term of the right-hand side of Equation (29) will have degree 4. Notice, however that since $\text{var}(f(x_i)|f_n) = 0$, $\text{cov}(f(x), \mathcal{I}|f_n, g)$ must be zero when $x = x_i$ for all $i$, so the resulting fourth-degree polynomial will factor into a quadratic and two linear polynomials when $x$ lies between two previous observations, and into a cubic and a linear when $x$ lies outside the range of the previous observations. Similarly, by examining Equation (30) we will see that the first term on the right-hand side will be of degree 3, while the second term will have degree 6 for $x_1 < x < x_n$ (since as above, $t^*(x)$ has degree 3 there), and degree 2 for $x < x_1$ and $x > x_n$ where $t^*(x)$ is linear. Again, however, $\text{var}(f(x)|f_n, g)$ must be zero when $x = x_i$, and nonnegative otherwise, and since the functions $t(x)$ and $\tilde{u}_g(x)$ are continuous with one continuous derivative, we must have that the non-negative function $\text{var}(f(x)|f_n, g)$ has a double root at each data point. So, within the range of the data, it must be a quadratic polynomial times the square of the same two linear polynomials that appear in $\text{cov}(f(x), \mathcal{I}|f_n, g)$ and a linear times the square of a single linear polynomial outside the range of the data. Hence, we must have that the conditional decrease in variance from the next observation is

$$\frac{\text{cov}(f(x), \mathcal{I}|f_n, g)^2}{\text{var}(f(x)|f_n, g)} = \frac{q_1(x)^2}{q_2(x)}$$

where $q_1$ and $q_2$ are both quadratic polynomials when $x_1 < x < x_n$, and $q_1$ will be cubic and $q_2$ linear when $x < x_1$ or $x > x_n$. In either case, taking the derivative with respect to $x$ we get

$$\frac{\partial}{\partial x} q_1^2 = \frac{q_1(x)[2q_1'(x)q_2(x) - q_1(x)q_2'(x)]}{q_2(x)}.$$

The expression in square brackets must be zero at any interior maximum, and this is simply a piecewise cubic equation. Note that any point at which the first factor, $q_1(x)$, is zero is a point where we will realize no decrease in variance (and therefore, is a global minimum), and hence is a point of no interest to us. It is now a matter of solving $n + 1$ cubic equations, finding the decrease in variance associated with each of the roots as well as each endpoint (this involves only evaluating the ratio $q_1(x)^2/q_2(x)$). It is interesting to note that it is possible for the maximum to occur exactly at a previously observed data point. This is because of the way the linear factors cancel to produce the ratio $q_1(x)^2/q_2(x)$. This ratio will have a removable discontinuity at each data point. Of course if we sample exactly at a previous data point we cannot gain any additional information about $f(x)$, and thus any decrease in variance is not possible. However, if we sample an arbitrarily small distance away from a previous data point we are in effect gaining information about the derivative.
of \( f \) at that point. Thus, a maximum occurring at a previous data point suggests that the most useful information that we can collect is the value of the derivative at this point. In our implementation we do not actually test the endpoints of each interval for maxima, but always shrink the interval away from the ends slightly and evaluate the function there.

In the next section, we outline the computations involved with explicitly computing the coefficients of \( q_1 \) and \( q_2 \).

### 4.3 Computing \( q_1 \) and \( q_2 \)

As we have seen, to compute the coefficients of the piecewise polynomial functions \( q_1(x) \) and \( q_2(x) \) we need to compute the following integrals separately for each interval \((x_i, x_{i+1})\):

\[
\bar{v}_g(x) = \int_0^1 P(u, x) \bar{P}(u) g^2(u) du,
\]

\[
v'_g(x, x) = \int_0^1 P(u, x)^2 g^2(u) du,
\]

and

\[
t^*(x) = \int_0^1 Q(u) P(u, x) g^2(u) du.
\]

We also need the following integral which is independent of \( x \):

\[
\bar{t}^* = \int_0^1 Q(u) \bar{P}(u) g^2(u) du.
\]

All integrals and subsequent algebraic manipulation were done using Mathematica (Wolfram, 1991).

Since the last integral \( \bar{t}^* \) does not depend on \( x \), we will compute it first. Considering the \( j^{th} \) component and using the fact that the support of the function \( Q_j(u) \) is the interval \([x_j, x_{j+2}]\), we have

\[
\bar{t}^*_j = \int_0^1 Q_j(u) \bar{P}(u) g^2(u) du
\]

\[
= g_j^2 \int_{x_j}^{x_{j+1}} Q_j(u) \bar{P}(u) du + g_{j+1}^2 \int_{x_{j+1}}^{x_{j+2}} Q_j(u) \bar{P}(u) du
\]

where \( g_i = \exp(z_i) \). We will take each of these integrals separately.

\[
\int_{x_j}^{x_{j+1}} Q_j(u) \bar{P}(u) du
\]

\[
= \frac{(x_{j+1} - x_j)}{24} \left( 3x_{j+1}^2 + 2x_j x_{j+1} + x_j^2 + \frac{2(1 - 2x_n)(2x_{j+1} + x_j - 3x_1)}{x_n - x_1} \right).
\]

\[
\int_{x_{j+1}}^{x_{j+2}} Q_j(u) \bar{P}(u) du
\]

\[
= \frac{(x_{j+2} - x_{j+1})}{24} \times
\]

\[
\left( 3x_{j+1}^2 + 2x_{j+1} x_{j+2} + x_{j+2}^2 + \frac{2(1 - 2x_n)(2x_{j+1} + x_{j+2} - 3x_1)}{x_n - x_1} \right).
\]
For the remaining three integrals, first suppose that \( x < x_1 \). In this case \( q_1(x) \) will be cubic so we need to compute the first 4 coefficients of the fourth degree polynomial \( \text{cov}(\mathcal{I}, f(x) | f_n) \). (The constant term will be determined since \( q_1(x) \) must be divisible by \( x - x_1 \).) Similarly, we do not need to compute the linear and constant terms for \( q_2(x) \) (although as we see below, we will compute them automatically) since \( \text{var}(f(x) | f_n) \) must be divisible by \( (x - x_1)^2 \). The computations are as follows.

\[
\int_0^1 P(u,x)\tilde{P}(u)g^2(u)du = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u,x)\tilde{P}(u)g^2_i du.
\]

Now \( P(u,x) = 0 \) if \( u \geq x_n \) so the last term in the summation will be zero. Furthermore,

\[
\int_0^{x_1} P(u,x)\tilde{P}(u)g^2_0 du = g^2_0 \left( \frac{1}{24} x^4 - \frac{x^3}{6} x + \text{constant} \right),
\]

and

\[
\int_{x_i}^{x_{i+1}} P(u,x)\tilde{P}(u)g^2_i du = g^2_i \frac{x - x_1}{(x_n - x_1)^2} (S(x_{i+1}) - S(x_i))
\]

where

\[
S(x) = \frac{x^4}{8} + \frac{x^3(1 - 2x_n + x_1x_n - x^2_n)}{6} + \frac{x^2(x_1 + x_n)(2x_n - 1)}{4} - \frac{xx_1x_n(2x_n - 1)}{2}.
\]

We also have

\[
\tilde{\nu}(x,x) = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u,x)^2g^2_i du,
\]

where

\[
\int_0^{x_1} P(u,x)^2 g^2_0 du = g^2_0 (x_1 - x)^3 / 3
\]

and

\[
\int_{x_i}^{x_{i+1}} P(u,x)^2 g^2_i du = g^2_i (x - x_1)^2 \frac{(x_{i+1} - x_n)^3 - (x_i - x_n)^3}{3(x_n - x_1)^2}.
\]

To compute \( t^*(x) \), since the support of the function \( Q_j(u) \) is the interval \([x_j, x_{j+2}]\), and considering the \( j^{th} \) component of \( t^*(x) \), we have that

\[
t^*_j(x) = \int_0^1 Q_j(u)P(u,x)g^2 du = g^2_i (x - x_1)(x_{i+1} - x_i)(2x_{i+1} + x_i - 3x_n) \frac{6(x_n - x_1)}{6(x_n - x_1)} + g^2_{i+1} (x - x_1)(x_{i+2} - x_{i+1})(2x_{i+1} + x_{i+1} - 3x_n) \frac{6(x_n - x_1)}{6(x_n - x_1)},
\]

and this completes the case \( x < x_1 \).
The case $x > x_n$ is similar:

$$\int_0^1 P(u, x) \bar{P}(u) g_i^2(u) du = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u, x) \bar{P}(u) g_i^2 du,$$

$P(u, x) = 0$ if $u \leq x_1$, and

$$\int_{x_i}^{x_{i+1}} P(u, x) \bar{P}(u) g_i^2 du = g_i^2 \frac{x - x_1}{(x_n - x_1)^2} (T(x_{i+1}) - T(x_i))$$

where

$$T(x) = \frac{(x_n - x_1)^2 x^4}{8} + \frac{x^3(1 - 2x_n - x_1 x_n + x_1^2)}{6} + \frac{x x_1 (2x_n - 1)}{4} - \frac{x x_1^2 (2x_n - 1)}{2}$$

and

$$\int_{x_n}^{1} P(u, x) \bar{P}(u) g_i^2 du = g_n^2 \left( \frac{1}{24} x^4 - \frac{1}{6} x^3 + \frac{1}{4} x^2 - \frac{x_n (x_n^2 - 3x_n + 3)}{6} x + \text{constant} \right).$$

We also have

$$\bar{v}_g(x, x) = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u, x)^2 g_i^2 du,$$

where

$$\int_{x_i}^{x_{i+1}} P(u, x)^2 g_i^2 du = g_i^2 (x - x_n)^2 [(x_{i+1} - x_1)^3 - (x_i - x_1)^3]/3(x_n - x_1)^2.$$ 

and

$$\int_{x_n}^{1} P(u, x)^2 g_n^2 du = g_n^2 (x - x_n)^3 / 3$$

Furthermore, as above,

$$t_j^*(x) = g_j^2 \frac{(x - x_n)(x_{j+1} - x_j)(2x_{j+1} + x_j - 3x_1)}{6(x_n - x_1)} + g_{j+1}^2 \frac{(x - x_n)(x_{j+2} - x_{j+1})(2x_{j+1} + x_{j+1} - 3x_1)}{6(x_n - x_1)}.$$ 

We now treat the case where $x_j < x < x_{j+1}$. Here we have a larger number of sub-cases. First, note that for this case and for $q_1$, we only need to compute the coefficients of degree at least 2 in $\text{cov}(I, f(x) | f_n)$, since this covariance must have a factor of $(x - x_j)(x_{j+1} - x)$, and the coefficients of terms of degree at least 4 in $\text{var}(f(x) | f_n)$, since it must be divisible by $(x - x_j)^2(x - x_{j+1})^2$. 

14
To this end we have
\[ \int_{0}^{1} P(u, x) \bar{P}(u) g^2(u) du = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u, x) \bar{P}(u) g_i^2 du. \]

Since the support of \( P(u, x) \) is \( u \in (x_1, x_n) \), both the first and last terms of the sum will be zero. Now fix \( j \) such that \( x_i < x < x_{i+1} \) and note that, for any \( x_i \), one of three cases; \( i < j \), \( i > j \) and \( j = i \) will occur. If \( i < j \), we have \( x_i < x_{i+1} < x \) so that
\[ \int_{x_i}^{x_{i+1}} P(u, x) \bar{P}(u) g_i^2 du = (x - x_n) \int_{x_i}^{x_{i+1}} \frac{(u - x_1)}{x_n - x_1} \bar{P}(u) g_i^2 du. \]

And since the integral in the last expression does not depend on \( x \), this term will be linear and can be ignored. Similarly, we may ignore the case \( i > j \). For the case \( i = j \), we have
\[ \int_{x_i}^{x_{i+1}} P(u, x) \bar{P}(u) g_i^2 du \]
\[ = g_j^2 \left( \frac{1}{24} x^4 - \frac{2x_n - 1}{12(x_n - x_1)} x^3 + \frac{x_1(2x_n - 1)}{4(x_n - x_1)} x^2 + \text{lower degree terms} \right) \]

Similarly,
\[ \bar{v}_g(x, x) = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} P(u, x) g_i^2 du. \]

If \( i < j \),
\[ \int_{x_i}^{x_{i+1}} P(u, x)^2 g_i^2 du = (x - x_n)^2 \int_{x_i}^{x_{i+1}} \left( \frac{u - x_1}{x_n - x_1} \right)^2 g_i^2 du \]

and again the integral in the last expression is independent of \( x \) so this term will have degree 2 and we can ignore it. The case \( i > j \) is similar. Now suppose that \( i = j \).
\[ \int_{x_i}^{x_{j+1}} P(u, x)^2 g_j^2 du \]
\[ = g_j^2 \left( \frac{1}{3(x_n - x_1)} x^4 - \frac{2(x_1 + x_n)}{3(x_n - x_1)} x^3 + \text{lower degree terms} \right). \]

Considering the \( i^{th} \) component of \( t^*(x) \), we have that
\[ t_i^*(x) = \int_{x_i}^{x_{i+1}} Q_i(u) P(u, x) g_i^2 du + \int_{x_{i+1}}^{x_{i+2}} Q_i(u) P(u, x) g_{i+1}^2 du. \]

Here we have two integrals with three sub-cases for each one, \( i < j \), \( i = j \), and \( i > j \) for the first, and \( i < j - 1 \), \( i = j - 1 \), and \( i > j \) for the second. Furthermore, since \( t^*(x) \) appears in the quadratic form \( t^*(x)^T A^{*^{-1}} t^*(x) \), we will need to compute all terms of degree at least one. To begin, consider the first integral and take \( i < j \).
\[ \int_{x_i}^{x_{i+1}} Q_i(u) P(u, x) g_i^2 du = g_i^2 \frac{(x_{i+1} - x_i)(2x_{i+1} + x_i - 3x_1)}{6(x_n - x_1)} (x - x_n). \]
If $i = j$,
\[
\int_{x_j}^{x_{j+1}} Q_j(u) P(u, x) g_j^2 \, du = g_j^2 \left( \frac{1}{6(x_n - x_1)} x^3 - \frac{x_j}{2(x_{j+1} - x_j)} x^2 \right. \\
- \left. \frac{x_{j+1}^2}{2(x_{j+1} - x_j)} - \frac{(x_{j+1} - x_j)(2x_{j+1} + x_j - 3x_j)}{6(x_n - x_1)} \right) x + \text{constant} \right).
\]
And, if $i > j$,
\[
\int_{x_i}^{x_{i+1}} Q_i(u) P(u, x) g_i^2 \, du = g_i^2 \left( \frac{x_{i+1} - x_i}{6(x_n - x_1)} (x - x_1) \right).
\]
For the second integral, suppose that $i < j - 1$,
\[
\int_{x_{i+1}}^{x_{i+2}} Q_{i+1}(u) P(u, x) g_{i+1}^2 \, du = g_{i+1}^2 \left( \frac{x_{i+2} - x_{i+1}}{6(x_n - x_1)} (x - x_n) \right).
\]
And for $i = j - 1$,
\[
\int_{x_j}^{x_{j+1}} Q_{j-1}(u) P(u, x) g_j^2 \, du = g_j^2 \left( \frac{x_{j+1} - x_j}{6(x_n - x_1)} x^2 \right. \\
- \left. \frac{x_{j+1}^2}{2(x_{j+1} - x_j)} - \frac{(x_{j+1} - x_j)(2x_{j+1} + x_j - 3x_j)}{6(x_n - x_1)} \right) x + \text{constant} \right).
\]
And finally, if $i \geq j$,
\[
\int_{x_{i+1}}^{x_{i+2}} Q_i(u) P(u, x) g_i^2 \, du = g_{i+1}^2 \left( \frac{x_{i+2} - x_{i+1}}{6(x_n - x_1)} (x - x_1) \right).
\]

Now, for each interval, $[x_i, x_{i+1}]$, we may express the relevant quantities as polynomials in $x$ by using the results above. We begin with $q_i(x)$.

\[
\bar{v}_g(x) = b_4 x^4 + b_3 x^3 + b_2 x^2 + b_1 x + b_0.
\]

If $x_1 < x < x_n$, we will not need to compute the coefficients $b_1$ and $b_0$. If $x < x_1$ or $x > x_n$, we will need to compute $b_1$. We may also write

\[
t^*(x) = t_3 x^3 + t_2 x^2 + t_1 x + t_0.
\]

where the vectors $t_j$ have components given by the appropriate integrals above. Note that for $x < x_1$ or $x > x_n$, the coefficients $t_3$ and $t_2$ will both be zero and in any case, we need
not compute the vector $t_0$. Thus we may write $q_1(x)$ as
\[
q_1(x) = \frac{\tilde{v}_9(x) - t^*(x)^T A^{*-1} \tilde{t}^*}{x - x_1}
\]
\[
= \frac{b_4 x^4 + b_3 x^3 + b_2 x^2 + b_1 x + b_0 - (t_3 x^3 + t_2 x^2 + t_1 x + t_0)^T A^{*-1} \tilde{t}^*}{x - x_1}
\]
\[
= \frac{b_4 x^4 + b_3 x^3 + b_2 x^2 + (b_1 - t_1^T A^{*-1} \tilde{t}^*) x + (b_0 - t_0^T A^{*-1} \tilde{t}^*)}{x - x_1}
\]
when $x < x_1$ and
\[
q_1(x) = \frac{b_4 x^4 + b_3 x^3 + b_2 x^2 + (b_1 - t_1^T A^{*-1} \tilde{t}^*) x + (b_0 - t_0^T A^{*-1} \tilde{t}^*)}{x - x_n}
\]
when $x > x_n$. For $x_i < x < x_{i+1}$, we have
\[
q_1(x) = \frac{1}{(x - x_1)(x_n - x)} \left( b_4 x^4 + (b_3 - t_3^T A^{*-1} \tilde{t}^*) x^3 + (b_2 - t_2^T A^{*-1} \tilde{t}^*) x^2 
+ (b_1 - t_1^T A^{*-1} \tilde{t}^*) x + b_0 - t_0^T A^{*-1} \tilde{t}^* \right).
\]
The coefficients of $q_1(x)$ may now be computed by long division separately for each interval. As we stated earlier, $\text{cov}(T, f(x)|f_n)$ must be divisible by $x - x_i$ when $x$ and $x_i$ are adjacent, so the corresponding denominators will divide evenly into the numerator and $q_1(x)$ will reduce to a quadratic or cubic polynomial.

For $q_2(x)$, we have
\[
\tilde{v}_9(x, x) = c_4 x^4 + c_3 x^3 + c_2 x^2 + c_1 x + c_0.
\]
For the cases $x < x_1$ and $x > x_n$, $c_4$ will be zero and the only necessary coefficients will be $c_3$ and $c_2$. For $x_1 < x < x_n$, $c_4$ will not be zero, and it is the only necessary coefficient. We also have $t^*(x)$ as above, so that if $x < x_1$,
\[
q_2(x) = \frac{\tilde{v}_9(x) - t^*(x)^T A^{*-1} t^*(x)}{(x - x_1)^2}
\]
\[
= \frac{1}{(x - x_1)^2} [c_4 x^4 + c_3 x^3 + (c_2 - t_1^T A^{*-1} t_1) x^2 + \text{lower degree terms)].
\]
Similarly, for $x > x_n$,
\[
q_2(x) = \frac{1}{(x - x_n)^2} [c_4 x^4 + c_3 x^3 + (c_2 - t_1^T A^{*-1} t_1) x^2 + \text{lower degree terms}].
\]
For the case $x_i < x < x_{i+1}$, we have
\[
q_2(x) = \frac{1}{(x - x_1)^2(x_i + 1 - x)} \left(-t_3^T A^{*-1} t_3 x^6 - 2 t_3^T A^{*-1} t_2 x^5 
+ (c_4 - 2 t_3^T A^{*-1} t_1 - t_2^T A^{*-1} t_2) x^4 + \text{lower degree terms} \right).
\]
Again, using long division, $q_2(x)$ reduces to a quadratic or linear polynomial. Once these coefficients have been computed for $x$ in each interval $[x_i, x_{i+1}]$, it is simply a matter of maximizing the ratio $q_1(x)^2/q_2(x)$, which we can do with a Newton-Raphson algorithm and checking the ends of each interval. The myopic rule then dictates that the next observation be taken at the global maximum of this ratio.
4.4 Insensitivity of Bayes risk to changes in $g(x)$

Here we give a heuristic argument demonstrating the relative insensitivity of Bayes risk given in (26) to errors in our estimate of $g$. Conditional on some fixed set of $x_i$ fix $g_0(x)$ and write

$$t'_g = \int_0^1 P(u) \bar{P}(u) g_2(u) du.$$ 

and

$$A'_g = \int_0^1 P(u) P(u)^T g_2(u) du.$$ 

where $P(u) = (P(u, x_2), P(u, x_3), \ldots, P(u, x_{n-1})$ with $P(u, x)$ defined by (14). If we let $\epsilon^2(u) = g^2(u) - g_0^2(u)$, we see that $A_g = A_{g_0} + A_\epsilon$ and $t_g = t_{g_0} + t_\epsilon$. Now consider the difference between the coefficients of $f_n$ in the quadrature rules generated by these two variance functions. From (18), since $m'$ is independent of $g(x)$, this difference is

$$A^{-1}_g t_g - A^{-1}_{g_0} t_{g_0} = A^{-1}_{g_0} (A_{g_0} A^{-1}_g t_g - t_{g_0})$$

$$= A^{-1}_{g_0} (A_{g_0} - A_\epsilon) A^{-1}_g t_g - (t_g - t_\epsilon)$$

$$= A^{-1}_{g_0} (t_g - A_\epsilon A^{-1}_g t_g)$$

$$= A^{-1}_{g_0} \int_0^1 P(u) (\bar{P}(u) - P(u)^T A^{-1}_g t_g) \epsilon^2(u) du.$$ 

We have two cases for the integrand. If $u < x_i$ or $u > x_n$, by (14) $P(u) = 0$ since $x_1 \leq x_i \leq x_n$ for all $i$. On the other hand, if $x_1 < u < x_n$, recall that $\bar{P}(u) = \int_0^1 P(u, x) dx$ and observe that by (18), $P(u)^T A^{-1}_g t_g$ is a quadrature rule of this integral since, for fixed $u$, $P(u)$ is a vector of evaluations of the function $P(u, x)$ at the $x_i$, $i = 2, 3, \ldots, n - 1$, in our design and because $P(u, x_1) = P(u, x_n) = 0$, we have $m' = 0$. Thus, the factor $P(u) - P(u)^T A^{-1}_g t_g$ will be small, say $\delta(u)$. We will have that

$$|A^{-1}_g t_g - A^{-1}_{g_0} t_{g_0}| = \left| A^{-1}_{g_0} \int_{x_1}^{x_n} P(u) \delta(u) \epsilon^2(u) du \right|$$

$$\leq \int_{x_1}^{x_n} |A^{-1}_{g_0} P(u)| |\delta(u)| \epsilon^2(u) du$$

$$\leq \sup \{|\delta(u)\epsilon^2(u)|\} \int_{x_1}^{x_n} |A^{-1}_{g_0} P(u)| du$$

$$\leq \sup \{|\delta(u)\epsilon^2(u)|\} C$$

where $C = \int_{x_1}^{x_n} |A^{-1}_{g_0} P(u)| du$. Therefore, the coefficients of the quadrature rule using $g^2$ will be close to those using $g_0^2$. We conclude that $E[(m^2_{N, g} - m^2_{N, g_0})^2] \leq \sup \{\epsilon^2(u) \delta(u)\}^2 ||Cf^*||^2$, so the middle term in (26) may be ignored.

5 Evaluating the Myopic Rule

In this section we discuss Monte Carlo simulations used to evaluate the risk of the sequential rule developed in the previous sections under the Bayesian model from which it was derived.
We compare its performance to that of a fixed rule which is (asymptotically) optimal for particular value of the cost $c$. We begin by discussing how the Monte Carlo evaluation was performed.

5.1 Monte Carlo Evaluation

Recall that Bayes risk is the expected posterior loss, given by $E[(I - m_N,\hat{\phi}_y)^2 + Nc]$ under the prior distribution for $f$. Thus, our Monte Carlo integration will consist of generating a sample, $f$, from its prior distribution and sequentially evaluating it according to the rule in question. We evaluate the posterior expected loss at the point at which we stop, and by averaging this over sufficiently many replications, we get an estimate of Bayes risk. There are a number of refinements we can make over a naive implementation of this procedure.

5.1.1 Generating $f$

The naive procedure is as follows. Given a sufficiently dense grid of points, $(\xi_0, \xi_1, ..., \xi_M)$, in $[0,1]$, generate a sample of standard Brownian motion, $B(x)$, and generate $z(x)$ by $z(x) = a_0 + \eta(B(x) - B(1)/2)$ where $a_0 \sim N(\tau, \zeta^2)$. We may generate realizations of Brownian motion processes by taking partial sums of independent normal random variables with mean zero and variances, $\xi_{i+1} - \xi_i$. We view this realization as a step function approximation of a true realization of $z(x)$. While it might be more appealing to use a linear interpolant as our approximation to $z$, the subsequent computations will become more complex. If the sampling points are sufficiently dense in $[0,1]$, the error introduced at this stage will be negligible compared to the rest of the random variation in the model. We only need to ensure that the subintervals are small enough that $z(x)$ does not change very much from interval to interval. We will expand on this a bit more later. Once we have a realization of $z(x)$, we may compute $g(x)$ from (9). Next, rather than applying (10) directly to generate $f(x)$, we choose a slightly different approach.

First, since we have assumed that $\alpha$ and $\beta$ have improper prior distributions, we clearly cannot generate them. However, since the allocation and stopping rules as well as the posterior variance of $f$ do not depend on their values, we can choose them in any way we like without affecting the performance or evaluation of the rules. Hence, in our simulations we will always take them to be zero. Second, rather than numerically evaluating the last term of (10) directly, we evaluate it using the iterated integral,

$$f(x) = \int_0^x \int_0^t g(u)dB(u)dt.$$  \hspace{1cm} (31)

In this way we can store both $f(x)$ and $f'(x)$ for future use. The reason for doing this is that under the sequential allocation rule, it is impossible to predict in advance all the points at which we will need to evaluate $f(x)$. Since the ordered pair $(f'(x), f(x))$ is Markov we can generate a new pair, $(f'(x), f(x))$, at $x$ conditional only on the previous observations immediately on either side of $x$.

Therefore, we begin by generating a sample of observations from $f'$ by first generating a set of normal random variables with mean zero and variance $(\xi_i - \xi_{i-1})g^2(\xi_i)$, (taking
\( \xi_0 = 0 \) and \( f'(\xi_0) = 0 \) and computing the appropriate partial sums to approximate the inner integral in (31) for each point in our grid. Given this sample, \( \{ f'(\xi_i) \} \), \( f(\xi_i) \) is still not completely determined at the grid points. Between the grid points, the expected value of \( f' \) will correspond to the linear interpolant and assuming that \( g^2 \) does not change much between the grid points, deviations of \( f' \) from its expected value will approximately follow a Brownian bridge process. If we again assume that the initial set of data points is sufficiently dense, we may integrate the expected value of \( f' \) and ignore the contribution from the error processes which will be negligible. Alternatively, we could perturb \( f(\xi_i) \) by normal deviates with mean zero and the appropriate variance. For simplicity, we computed the integral of \( E[f'(x)] \) by simply applying the trapezoid rule to the set of values of \( f'(\xi_i) \). Interestingly, this gives us the Bayes estimate (posterior mean) of \( f \) given \( \{ f'(\xi_i) \} \). We can show that the contributions to variance of \( f(\xi_k) \) from each subinterval \((\xi_i, \xi_{i+1})\) are independent and equal the cube of the length of the subinterval times \( g^2(\xi_i) \), divided by twelve. Thus the posterior variance of our estimate of \( f(\xi_k) \) is

\[
\text{var}\left( \int_0^{x_k} f'(x)dx \mid f'(\xi_0), f'(\xi_1), \ldots, f'(\xi_k) \right) = \sum_{i=1}^{k} g^2(\xi_i)(\xi_i - \xi_{i-1})^3/12. \tag{32}
\]

In the case where the initial set of \( \xi_i \) are equally spaced, and we have a sample size of \( M \), this variance will be of order \( 1/M^2 \). If we choose \( M \) so that this variance is small, then the error introduced by our approximation will in fact be negligible.

Now suppose that we have samples \( \{ f'(\xi_i), f(\xi_i) \} \) and \( \{ f'(\xi_{i+1}), f(\xi_{i+1}) \} \) and we wish to generate a sample \( \{ f'(\xi), f(\xi) \} \) for some \( \xi \in (\xi_i, \xi_{i+1}) \). Since, given \( \{ f'(\xi_i), f(\xi_i) \} \) and \( \{ f'(\xi_{i+1}), f(\xi_{i+1}) \} \), \( \{ f'(\xi), f(\xi) \} \) is independent of \( f'(y), f(y) \), and \( g(y) \) for \( y \) outside this interval, and since we are treating \( g(\xi) \) as a step function, and hence constant in this interval, we can apply standard results for ordinary Brownian motion to the distribution of \( \{ f'(\xi), f(\xi) \} \). For simplicity, we treat the case \( \xi_0 = 0 \) and \( f'(\xi_i) = f(\xi_i) = 0 \). Recall that since, in this subinterval, \( f' \) is Brownian motion with diffusion rate \( g^2(0) \), \( \text{cov}(f'(x), f'(y)) = g^2(0) \min(x,y) \). Furthermore, by (12),

\[
\text{cov}(f'(x), f(y))/g^2(0) = \int_0^y \min(x,u)du = \begin{cases} 
\frac{y^2}{2} & \text{if } y < x \\
xy - \frac{x^2}{2} & \text{if } y > x
\end{cases}
\]

and by (13),

\[
\text{cov}(f(x), f(y))/g^2(0) = \int_0^x \int_0^y \min(t,u)dudt = \frac{y^2(3x-y)}{6} \text{ if } y < x.
\]

Therefore, we have the following three covariance matrices:

\[
\Sigma_1 = \text{cov}[f'(\xi), f(\xi)] = g^2(0) \begin{pmatrix}
\xi & \xi^2/2 \\
\xi^2/2 & \xi^3/3
\end{pmatrix},
\]

\[
\Sigma_2 = \text{cov}[f'(\xi_{i+1}), f(\xi_{i+1})] = g^2(0) \begin{pmatrix}
\xi_{i+1} & \xi_{i+1}^2/2 \\
\xi_{i+1}^2/2 & \xi_{i+1}^3/3
\end{pmatrix}
\]
\[ \Sigma_3 = \text{cov}[\langle f'(\xi), f(\xi) \rangle, \langle f'(\xi_{i+1}), f(\xi_{i+1}) \rangle] \]
\[ = g^2(0) \left( \begin{array}{c} \xi \\ \xi^2/2 \\ \xi^2(3\xi_{i+1} - \xi)/6 \end{array} \right). \]

Since both \( E[(f'(\xi), f(\xi))] \) and \( E[(f'(\xi_{i+1}), f(\xi_{i+1}))] \) equal zero, standard multivariate normal theory shows that

\[ E[(f'(\xi), f(\xi))|(f'(\xi_{i+1}), f(\xi_{i+1}))] = \Sigma_3 \Sigma_2^{-1} (f'(\xi_{i+1}), f(\xi_{i+1}))^T \tag{33} \]

and

\[ \text{cov}((f'(\xi), f(\xi)|f'(\xi_{i+1}), f(\xi_{i+1})) = \Sigma_1 - \Sigma_3 \Sigma_2^{-1} \Sigma_2^T \tag{34} \]

We may now simulate the vector \((f'(\xi), f(\xi))\) by computing the conditional expectation (33), generating a random vector with mean zero and covariance matrix (34), and adding them together. It is interesting to note that the expectation in (33) is the unique quadratic function with the values of both the function and its derivative matching the corresponding values previously generated. In the general case, we will simply replace \(g^2(0)\) with \(g^2(\xi_i)\), \(\xi\) with \(\xi - \xi_i\) and \(\xi_{i+1}\) with \(\xi_{i+1} - \xi_i\) in the covariance matrices above, add \(f'(\xi_i)\) to the sample of \(f'(\xi)\) and add the linear interpolant, \((f'(\xi_i), f(\xi_i) + f'(\xi_i)(\xi - \xi_i))\), to \(f(\xi)\). The computer program will update the array of sampled points of \((f'(\xi), f(\xi))\) to include the new observations so that future evaluations will condition on these as well. Once we are able to generate and sample from the processes described above, we may use all the machinery developed previously to apply the sequential rule to these realizations.

### 5.1.2 Posterior Variance

Once we have generated a random function \(f\) and evaluated it according to a sequential rule, \(\rho\), we need to evaluate its contribution to the Bayes risk given by \(r_\rho(f_0)\) where \(f_0\) indicates that \(r_\rho(\cdot)\) is evaluated prior to the collection of any data. As in Section 4, we will approximate \(r_\rho(f_0)\) by

\[ \tilde{r}_\rho = E[w_{N,\sigma}^T + cN] \tag{35} \]

where the expectation is estimated by averaging the integrand over the sample of randomly generated functions \(f(x)\). \(\tilde{r}_\rho\) is an approximation to \(r_\rho(f_0)\) in which the middle term of (26) is ignored. Note that the evaluation of the integrand in (35) does not require knowledge of \(f\), and in particular it does not require that we actually compute \(m_f^2\). The sole use of the simulated \(f\) in this context is to generate the random sequence \(x_1, x_2, \ldots, x_N\) (and implicitly the sample size \(N\)).

We may compute the integrand of (35) using (22), or the more numerically stable version,

\[ w_{N,\sigma}^T = \int_0^1 \left( \bar{P}(u) - \int_0^1 Q(v)^T \bar{P}(v) g^2(v)dv \right) A^{-1} Q(u) \bar{P}(u) g^2(u) du. \]

The integration above can be performed numerically using the trapezoid rule which, appropriately, is very close to the Bayes rule for such integrals.
5.1.3 Importance Sampling

There is an additional enhancement that we can make in the Monte Carlo integration process. From (9), we see that $g^2(x)$ is log-normally distributed for fixed $x$. Consequently, the distribution of $g^2(x)$ is skewed and we find that a few large observations have considerable influence on the observed mean of $g^2(x)$. While difficult to quantify, we may expect (and empirical evidence supports) that a few large observations can have a considerable effect on our Monte Carlo estimation of the Bayes risk as well. We address this difficulty using importance sampling.

In general, under importance sampling, we sample not from the distribution of interest, but rather from some other distribution which increases the frequency of “important” observations. Each observation is weighted according to its likelihood ratio under the the desired distribution versus the actual sampling distribution. Importance sampling relies on that fact that for a random variable $X$, a measurable function $\theta(x)$ and distributions $P_0$ and $P_1$,

$$E_{P_0}[\theta(X)] = E_{P_1}\left[\theta(X) \frac{dP_0}{dP_1}(X)\right]$$

provided $dP_0/dP_1$ exists. One reason to use importance sampling is that if we select $P_1$ carefully, we can coerce $\theta(X)dP_0(X)/dP_1$ to have smaller variance under $P_1$ than $\theta(X)$ does under $P_0$. Since in our case we can achieve this goal by realizing extreme observations with greater frequency under an alternative sampling distribution than under the true model (and subsequently giving each less weight), a naive approach might be to sample from a process like $g$, but which, for example, has a larger scale parameter, $\eta$, thus increasing the frequency of these large observations. The problem with this approach is that the support for $g$ under the true $\eta$ and under a different value of $\eta$ are disjoint so that the Radon-Nikodym derivative, $dP_0/dP_1$ does not exist. The fact that we are not observing true Brownian motion paths but rather finite dimensional approximations does not alleviate the difficulty. Specifically, if we have a realization of $g$ sampled at $M$ points, we in effect have $M - 1$ independent samples (the independent increments of $z$) whose variance is proportional to $\eta^2$, and hence we have $M - 1$ bits of information about $\eta$. If $M$ is large, we can distinguish with high probability between these two parameter values. Consequently, a sample drawn from $P_\eta$ will have likelihood ratio near zero most of the time. A few observations will have extremely large weights (since the expected value of the weights must be one), and so a very small proportion of our samples will still contribute significantly to the our estimate of variance.

Instead, in the scheme we chose to implement we expand $z(x)$ as a Fourier series and increase the variances of certain coefficients in such a way that extreme observations are generated with higher frequency, but the likelihood ratios are under control. In particular, if $b_0 \sim N(\tau, \zeta^2)$, and $a_i, b_i \sim N(0, 2\eta^2/(i^2\pi^2))$, $i = 1, 3, 5, \ldots$, then

$$z(x) = b_0 + \sum_{i=1,3,5,\ldots} a_i \sin(i\pi x) + b_i \cos(i\pi x) \quad (36)$$

will have mean $\tau$ and covariance given by (8). In general, one can approximate a Brownian motion process by generating a large number of terms of the expansion (36). An importance sample can be obtained by generating $z(x)$ according to (36) using values $\tau'$, $\zeta'$ and $\eta'_i$ for
the prior parameters, where we can use different values of \( \eta \) for each term of the expansion. Alternatively, if each of \( \tau' \), \( \zeta' \) and \( \eta_i' \) are at least as large as the corresponding original parameter, one can generate a sample \( z(x) \) according to the scheme described in Section 5.1.1 and add to it a sample obtained using (36) for some specified values of \( \tau' - \tau \), \( \zeta' - \zeta \) and \( \eta_i'^2 - \eta^2 \). In the latter case one can extract the actual values of the random coefficients, \( a_i \) and \( b_i \), via numerical integration of the resulting \( z(x) \). The likelihood ratios are

\[
\frac{\zeta'}{\zeta} \prod \frac{\eta_i'^2}{\eta^2} \exp \left( -\frac{1}{2} \left( \frac{(a_0 - \tau')^2}{\zeta'} - \frac{(a_0 - \tau)^2}{\zeta} + \sum_{i=1,3,5,\ldots} (a_i^2 + b_i^2) \left( \frac{1}{\eta_i'^2} - \frac{1}{\eta^2} \right) \right) \right)
\]

Note that if \( \tau' = \tau \) then the likelihood ratio is bounded by \( \zeta'/\zeta \prod \eta_i'^2/\eta^2 \) so by carefully choosing the \( \eta_i' \) and \( \zeta' \) we can balance the number of realizations with large posterior loss and low weight with the number of observations with high weight. For our simulations we used \( \tau' = \tau \), \( \zeta' = \zeta \), \( \eta_i'^2/\eta^2 = 2 \) and \( \eta_i = \eta \) for \( i = 3, 5, \ldots \). These values seemed to work quite well.

5.2 Fixed Rules

In this section we describe some of the analysis done for fixed rules and used for comparison with the sequential rule.

For purposes of this discussion, a fixed rule is a rule under which we select a set of points \( x_1, x_2, \ldots, x_n \) in advance, evaluate \( f(x) \) at each point and compute the posterior distribution of \( f \). No attempt is made to make additional function evaluations in the event that the posterior variance is larger than expected. The optimal fixed rule is the rule for which the posterior expected loss, \( E[(I - \tilde{m}_{n,g}^2)] + nc \) is minimized. Again we approximate this using \( \tilde{r}_n \) from (35). The minimization is performed by first fixing \( n \) and finding the design \( x_1, x_2, \ldots, x_n \), that minimizes expected posterior variance. We then find the sample size, \( n \), for which the overall posterior expected loss is minimized. Thus, the first problem will be to find optimal designs for fixed \( n \).

As we saw in Section 4.1, for known \( g \), the problem is reduced to finding the design which maximizes the reduction in posterior variance

\[
\text{cov}(\int f, f_n|g)^T \text{cov}(f_n|g)^{-1} \text{cov}(\int f, f_n|g).
\]

Unconditionally, we need to find the design which maximizes the expected value of this expression with respect to \( g \). Finding such a design for \( n \) larger than 3 or 4 is computationally prohibitive. We therefore resort to an approximation. For fixed \( g \) satisfying certain smoothness conditions, Wahba (see, for example, Wahba (1976) and Athavale and Wahba (1979)) and others (see Sacks and Ylvisaker (1966, 1968, 1970)) have shown that asymptotically, the optimal design will choose points \( x_1, x_2, \ldots, x_n \) so that the integrals \( \int_{x_i}^{x_{i+1}} (g(x)^{2/5}) \, dx \) will all have the same value. To allocate \( n \) points, Wahba recommends selecting \( 0 = x_1, x_2, \ldots, x_n = 1 \) so that

\[
\int_0^{x_i} g(x)^{2/5} \, dx = \frac{i - 1}{n - 1} \int_0^1 g(x)^{2/5} \, dx
\]

23
for $i = 2, 3, \ldots, n - 1$. Alternatively, it is simple to show that the (fixed sample) Bayes rule for a stationary Brownian motion prior takes observations at $(2i - 1)/2n$ (see Diaconis (1988)) so it seems reasonable in our case not to sample exactly at the ends of the interval, but to choose $x_1, x_2, \ldots, x_n$ so that
\[
\int_0^{x_i} g(x)^{2/5} dx = \frac{2i - 1}{2n} \int_0^{x_i} g(x)^{2/5} dx.
\]
For example, if $g$ were constant and $n = 5$, then we would sample at $1/10$, $3/10$, $5/10$, $7/10$ and $9/10$. To generalize this result to the expected value of posterior variance, a reasonable asymptotic approximation to the optimal design would be to replace $g(x)^{2/5}$ in the above expression with its expected value. Since $g(x)$ is defined to be stationary, this expectation is constant and so the above reduces to $x_i = (2i - 1)/2n$ regardless of the particular value of $E[g(x)^{2/5}]$. We will show that for the models we have discussed here, the asymptotic results are reasonably good approximations. Again, from Wahba (1976), we have that the posterior variance is approximately
\[
\text{var}(\int f | g, f_n) \approx \frac{1}{720n^3} \int_0^1 g^2(u) du.
\]
So using (9), the expected variance will be
\[
\text{var}(\int f | f_n) \approx E\left[\frac{1}{720n^3} \int_0^1 g^2(u) du\right]
= \frac{1}{720n^4} \exp\left\{2\tau + 2\zeta^2 + \frac{\eta^2}{2}\right\}.
\]
To find the value, $N$, of $n$ that gives the lowest posterior expected loss, we need to compute the expected posterior variance for a number of $n$ and choose the one with the smallest overall risk. Since, using the asymptotic approximation, the incremental changes in expected posterior variance are monotonically decreasing with $n$, we expect that the true changes in expected posterior variance will be as well so this is equivalent to choosing $N$ such that
\[
\inf\left\{E[\text{var}(\int f | f_{N-1}, g)]\right\} - \inf\left\{E[\text{var}(\int f | f_N, g)]\right\} > c
> \inf\left\{E[\text{var}(\int f | f_N, g)]\right\} - \inf\left\{E[\text{var}(\int f | f_{N+1}, g)]\right\}.
\]
These expectations can be derived from the Monte-Carlo simulations or we may use (37) to find the approximate $N$ which produces minimum risk. In the latter case, differentiating (37) with respect to $n$ and setting the derivative equal to $c$, we obtain an asymptotic estimate of the optimal $N$,
\[
N = \left(\frac{1}{180c} \exp\left\{2\tau + 2\zeta^2 + \frac{\eta^2}{2}\right\}\right)^{1/5}.
\]
5.3 Numerical Results

The goal of the Monte Carlo analysis is to compare the performance of the sequential rule described in Section 4 to an optimal fixed rule for particular values of the prior parameters and sampling cost. We began by evaluating the risk of the optimal fixed rule. Rather than choosing a value of sampling cost, $c$, and finding the optimal sample size $N$, we picked the $N$ and found the value of $c$ for which a sample size is optimal. For convenience we let $\gamma^2 = \exp(2b_0)$ where $b_0$ is given in (36). $\gamma^2$ can be thought of as a scale parameter that multiplies the covariance structure created by $g^2/\gamma^2$ and does not affect the allocation of observations. We chose $N = 15$ and $\eta^2 = 10$ and, since variance under the fixed rule is proportional to $\gamma^2$, the expected fixed sample risk depends on $\gamma^2$ only through its mean. Consequently, we chose to simply let $\gamma^2 = 1$ ($\tau = \zeta^2 = 0$). Note, however, that risk for the sequential rule will depend on the distribution of $\gamma^2$ since small values of $\gamma^2$ will generally lead to earlier stopping than large values. To find the value of $c$ corresponding to these parameters, we generated 5000 samples of $g(x)$ according the sampling scheme described in Section 5.1.3. We chose a grid of size 1001 with equally spaced points for the initial sample. In this case, (32) is on the order of $10^{-5}$ so our sampling distribution will be virtually indistinguishable from the desired distribution. Furthermore, when we are dealing with the fixed rule, the allocation of observations does not depend on $f(x)$ so we do not need to generate samples of $f(x)$ at this point. For each sample we then evaluated the posterior variance for samples of size 13, 14, 15, 16 and 17 to get a sense of how variance depends on sample size near the desired size of 15.

<table>
<thead>
<tr>
<th>Number of Observations</th>
<th>Asymptotic</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Variance</td>
<td>Total Risk</td>
</tr>
<tr>
<td>13</td>
<td>7.22</td>
<td>20.22</td>
</tr>
<tr>
<td>14</td>
<td>5.37</td>
<td>19.37</td>
</tr>
<tr>
<td>15</td>
<td>4.07</td>
<td>19.07</td>
</tr>
<tr>
<td>16</td>
<td>3.15</td>
<td>19.15</td>
</tr>
<tr>
<td>17</td>
<td>2.47</td>
<td>19.47</td>
</tr>
</tbody>
</table>

Table 1: Asymptotic and Monte Carlo estimates of posterior variance and of risk for the fixed sample rule. Risk is computed using $c = 10^{-6}$. SEM is the standard error for the total risk based on $N = 5000$ Monte Carlo simulations. All estimates of variance, risk, and SEM are times $10^6$.

We see in Table 1 that there is strong agreement between the asymptotic approximation (37) and the posterior variances obtained from the asymptotic approximation. Thus we may use the empirical results or the asymptotic results to estimate the cost $c$. By inverting (38) we see that

$$c = \frac{4\gamma^2 \exp\{\eta/2\}}{720N^5} = 1.09 \times 10^{-6}.$$ 

The estimated differences between the posterior variances at $N = 14$ and 15 and $N = 15$ and
16 are $-1.22 \times 10^{-6}$ and $-8.79 \times 10^{-7}$. The methods are in strong agreement, and since there is a wide range of values for $c$ for which 15 is the optimal fixed sample size (approximately $9 \times 10^{-7}$ through $1.2 \times 10^{-6}$), we chose $c = 1 \times 10^{-6}$.

To apply the sequential (myopic) rule, we generated realizations of the process $g(x)$ according to the scheme described in Section 5.1.3 and from this, generated $f(x)$ according to Section 5.1.1. After experimenting with various initial designs, we decided to begin with a sample of size 7, sampling initially at $x = 0.05, 0.256, 0.388, 0.500, 0.611, 0.744$ and 0.95. We chose this initial sample size because it seemed like a reasonable compromise between the minimum (informative) sample size of 4, for which we risk making early decisions based on tenuous information, and a larger sample, which while providing good information about $g$ may already have allocated too many observations poorly for the particular $g$ in question. We chose to allocate these seven observations this way for several reasons. We noticed empirically that if we did not initially observe $f(x)$ close to the endpoints 0 and 1, the first sequential observation was almost always taken near an endpoint. Thus it made sense to allocate observations relatively near the ends, and we arbitrarily chose 0.05 and 0.95. Furthermore, given the construction of the function $g(x)$, it is more likely that extreme values of $g(x)$, both large and small, will occur near the ends. In particular, we expect that most of the time we will need to allocate few observations to one end of the interval, more to the middle, and still more to the other end. Of course which end is which is unknown a priori. Thus, we can argue that the “minimum expected point density” will be larger in the middle than on the ends. Therefore, we begin with fewer observations at the ends than in the middle. Typically, the allocation rule will probably decide to allocate additional observations to one end and maybe a few to the middle while possibly not allocating any additional observations to the other end. Note that if our prior belief about $g$ led us to a nonstationary prior distribution on $g$, our initial allocation of observations could be chosen to also reflect our belief.

Given this initial sample, we can apply the myopic rule described in Section 4 to produce an observation of approximate posterior risk, $w^2_0 + nc$. A typical realization of this process is shown in Appendix A. There are a number of interesting features of the procedure which can be noted by studying these plots. For example, we can see how the estimates of $g(x)$ change as we make additional observations. When an observation is taken at a point where the previous interpolant was far from the actual function, $\hat{g}$ increases at that point. We can also see that as we add data points, $\hat{g}$ begins to follow the general shape of the true function, although in this sample, it never fully converges. Rates and type of convergence of the posterior mode to the true $g$ have not been studied and are left for future work.

It can be difficult to guess where the next observation will be taken. Often, one interval may be preferred by a large margin over another even when they have approximately the same length and value of $g$. The reduction in variance achieved by sampling at a particular point is a complicated combination of all the values of $g$ and the locations of all points in the sample. In fact, as we can see, it is possible for the reduction in variance from certain interior points to be nearly, if not exactly, zero. Since the function $q_1(x)$ as defined in Section 4.2 is quadratic, there can be as many as two such interior zeros in any interval. Apparently, while it may change the interpolant, sampling at such points does not change the estimate of the integral. There is no obvious way of predicting from the plot where such points might
<table>
<thead>
<tr>
<th>Rule</th>
<th>Posterior Variance</th>
<th>SEM</th>
<th>Total Risk</th>
<th>SEM</th>
<th>Mean N</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>3.93</td>
<td>0.42</td>
<td>18.93</td>
<td>0.42</td>
<td>15.00</td>
<td></td>
</tr>
<tr>
<td>Myopic:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(γ²) = 100</td>
<td>1.23</td>
<td>0.12</td>
<td>11.60</td>
<td>0.12</td>
<td>10.37</td>
<td>0.03</td>
</tr>
<tr>
<td>var(γ²) = 10</td>
<td>2.87</td>
<td>0.29</td>
<td>13.42</td>
<td>0.30</td>
<td>10.55</td>
<td>0.03</td>
</tr>
<tr>
<td>var(γ²) = 0</td>
<td>2.61</td>
<td>0.12</td>
<td>13.32</td>
<td>1.28</td>
<td>10.71</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 2: Mean posterior variance, risk, and sample size for Monte Carlo simulations for fixed rule and for myopic rule for various priors on γ². SEM refers to standard error of the corresponding estimated variance, risk, or sample size based on 2000 Monte Carlo simulations for the sequential rules, and 5000 simulations for the fixed rule. All estimates of variance, risk, and their SEMs are times 10⁶.

lie.

While examining results from our Monte Carlo studies, we observed that there were occasional realizations in which the estimate of q(x) derived from the initial sample was inadequate and the procedure stopped prematurely. The simplest remedy for this problem was to require that the procedure take a minimum number of observations before being allowed to stop. In the Monte-Carlo results given in Table 5.3 below, we required that three observations beyond the initial seven be taken before sampling could be stopped.

For our Monte Carlo study, we generated 2000 observations each for three different priors on γ². Since the fixed rule was evaluated at γ² = 1, we require that E[γ²] = 1 which is achieved by τ = −ζ². In this case var(γ²) = exp(4ζ²) − 1. We used var(γ²) = 100, 10 and 0. We expect that if the variance of γ² is large, we will benefit from early stopping a large portion of the time. When the variance of γ² is small, we will probably not gain as much from early stopping. However, since the allocation rule is independent of γ², we expect that the portion of the gain resulting from optimal allocation will be about the same. The values for posterior risk are shown in Table 5.3. In all cases, the risk for the myopic rule is significantly less than the values of 1.89 × 10⁻⁵ obtained empirically from the fixed rule, or the value of 1.907 × 10⁻⁵ obtained from the asymptotic approximation to the fixed rule. We also see that this reduction in risk is achieved by reducing both posterior variance and average sample size. The mean sample sizes for the myopic rule are a substantial reduction from that of the fixed rule. Given that a minimum of ten observations is required for each realization, at least half of all realizations did not take additional observations beyond those required. We also see that as var(γ²) decreases, our expected sample size increases only slightly, .37 observations in total. This appears to confirm our expectations that we benefit more from early stopping in the case where var(γ²) is large, although not by very much. The myopic rule, at least in this setting is performing as intended.
5.4 Some Deterministic Functions

In the previous subsection we examined the performance of the sequential rule by applying it to random functions generated according to the model underlying our procedure. Although the performance was good, a possible objection is that, in real applications, the function that we wish to integrate is not (with probability 1) in the family of random functions used in the simulation. It is nearly impossible to address this concern fully, since to do so would require a random sampling from the set of all “typical” functions. In lieu of such a sampling, we instead demonstrate the performance of the sequential rule by testing it on several deterministic functions with different kinds of behaviors.

Plots for these functions are shown in Appendix B. In these plots, we show the final estimated interpolant, the estimated variance function \( g(x) \) and the empirical distribution function of the sampling points. This function provides a useful summary of the allocation of these points. Where the EDF is steep, sampling points are more densely concentrated.

First, we consider the function

\[
 f(x) = \begin{cases} 
 10x/7 & \text{if } x \leq .7 \\
 10(1-x)/3 & \text{if } x > .7.
\end{cases}
\]

This function is a tent function with \( f(0) = f(1) = 0 \) and the peak at \( f(.7) = 1 \). We hope in this case that since the derivative is discontinuous at \( x = .7 \), the procedure will identify this point as one where \( f(x) \) is not smooth and thus concentrate sampling at this point. The accompanying plot shows this to be the case. The second function is a chi-square density with 1 degree of freedom. This function has a singularity at \( x = 0 \), so again the procedure should concentrate sampling near the origin. Since the procedure is not allowed to sample exactly at the endpoints of the interval, we will not have a problem with undefined function evaluations. Again, the observations are indeed more highly concentrated near the origin.

The third example is the function \( f(x) = \sin(30x^4) \) which is smooth for small \( x \) but begins to oscillate more frequently near \( x = 1 \). Again, in this case we hope that sampling will be concentrated toward the right end of the interval. As before, the procedure does quite well, although because of the nature of this function, more evaluations are needed to achieve a good estimate of the integral. Any other numerical procedure would have the same difficulty.

6 Concluding Remarks

In this paper we have focused on the development of a sequential rule for numerical quadrature. The rule is a modification of a myopic rule applicable when the function of interest is sampled from once-integrated Brownian motion. The modification entails sampling at least three observations sequentially before stopping — based on our observations this modification seems to enhance the risk behavior of the rule. To actually implement the rule we must also approximate the likelihood function — we do so with a piecewise linear approximation whose probabilistic properties we can describe. Based on simulations and the examination of specific examples, the rule as we have constructed it seems to perform quite well.
There remain a number of open problems and further issues to consider. For example, a fully sequential approach to this problem seems impractical. We have approached this by taking 7 observations chosen in an *ad hoc* fashion, and proceeding sequentially from there. An issue then is whether these 7 observations might be better placed, or if, indeed, 7 is a suitable number to start with. Although a fully sequential approach is nearly impossible, it might be possible in some cases to implement a “two-step ahead” procedure which behaves as if, at any point, there are at most two observations to be taken in a sequential manner.

Another issue of concern might be that, in most cases, the function we actually want to integrate does not lie in the space with which we have been working (paths of once-integrated Brownian motion). As noted in the examples above, the performance of the procedure seems good nonetheless. Still, this raises the possibility of using some other probability space of functions for modeling, an issue that we plan to address in the future.

**References**


A  A Sample Application of the Myopic Rule to a Randomly Generated Function

Here we show a series of plots showing how the myopic rule works when applied to a function generated by the process described in Chapter 6. Each figure is composed of two plots. The top plot shows the true random function $f(x)$ along with the interpolant (posterior mean conditional on the posterior mode of $g$) for the observations taken at the stage in question. The initial set of observations are shown with open circles while points sampled sequentially are filled. Also shown is a step function version of the true function $g(x)$, $\tilde{g}$, computed by averaging the process $z(x)$ over each interval, and the posterior mode of $g$ given the data. It can easily be seen that when $g(x)$ is large, there is more variation in the function $f(x)$. The bottom plot shows the expected reduction in posterior variance of the integral of $f$ were we to sample at an arbitrary point, $x$, in $[0,1]$. The point at which the maximum reduction is achieved is indicated by a solid diamond. The sample in the following figure in each case includes this additional point so we can see how all the relevant functions are updated from stage to stage. The discussion in Section 5.3 points out some features of these plots.

B  The Myopic Rule Applied to Some Deterministic Functions

Here we show a series of plots showing how the myopic rule works when applied to several deterministic functions. See Section 5.3 for a description of these functions. Each plot shows the true function $f(x)$ along with the interpolant (posterior mean conditional on the posterior mode of $g$) for the observations taken. Also shown is the posterior mode of $g$ given the data and the empirical distribution function of the data. The EDF is simply the proportion of observations falling to the left of a given point. In these examples five initial observations were taken at $x = .05, .235, .5, .765$ and .95, and 25 observations were taken in total.
Figure 1: The top plot shows the true function $f(x)$, the interpolant, the true (step function version) function $g(x)$ and its posterior mean after the initial sample of 7 observations. The bottom plot show the reduction in variance of the integral of $f$ as a function of a future observation at $x$. The point at which the maximum is achieved is indicated by a solid diamond.
Figure 2: The same as Figure 1 after 8 observations.
Figure 3: The same as Figure 1 after 9 observations.
Figure 4: The same as Figure 1 after 10 observations.
Figure 5: The same as Figure 1 after 11 observations.
Figure 6: The same as Figure 1 after 12 observations.
Figure 7: The same as Figure 1 after 13 observations.
Figure 8: The same as Figure 1 after 14 observations.
Figure 9: The same as Figure 1 after 15 observations.
Figure 10: The same as Figure 1 after 16 observations.
Figure 11: A plot of the function $f(x) = \sin(30x^4)$ along with the estimated (posterior mean) $f(x)$, the estimated (posterior mode) $g(x)$ and the EDF of the sampling points. The filled points were sampled sequentially and the unfilled points were the initial sample. The total sample size is 25.
Figure 12: A plot of the function $f(x) = 10x/7$ when $x \leq .7$ and $f(x) = 10(1 - x)/3$ when $x > .7$ along with the estimated (posterior mean) $f(x)$, the estimated (posterior mode) $g(x)$ and the EDF of the sampling points. The filled points were sampled sequentially and the unfilled points were the initial sample. The total sample size is 25.
Figure 13: A plot of the chi-squared density function with 1 degree of freedom along with the estimated (posterior mean) $f(x)$, the estimated (posterior mode) $g(x)$ and the EDF of 15 sampling points. The filled points were sampled sequentially and the unfilled points were the initial sample. The total sample size is 25.