Bayesian Analysis (2016) **11**, Number 1, pp. 191–213

Computational Enhancements to Bayesian Design of Experiments Using Gaussian Processes

Brian P. Weaver*, Brian J. Williams*, Christine M. Anderson-Cook*, and David M. Higdon*

Abstract. Bayesian design of experiments is a methodology for incorporating prior information into the design phase of an experiment. Unfortunately, the typical Bayesian approach to designing experiments is both numerically and analytically intractable without additional assumptions or approximations. In this paper, we discuss how Gaussian processes can be used to help alleviate the numerical issues associated with Bayesian design of experiments. We provide an example based on accelerated life tests and compare our results with large-sample methods.

Keywords: Bayesian design of experiments, Gaussian processes, accelerated life tests, preposterior expectation, expected quantile improvement.

1 Introduction

Bayesian design of experiments is a framework developed to incorporate some form of available prior information in the designing stage of an experiment. Such prior information could come from either a previous data analysis of the same items or from experimentation on similar items. Typically, before performing the experiment, practitioners have in mind some utility function that they wish to maximize. For example, in accelerated life tests (ALT), the engineers may want to estimate a quantile of the failure-time distribution as precisely as possible and their criterion function could be the posterior variance of the quantile, denoted as $U(\eta) = \text{var}(t_p|\mathbf{t}, \eta)$ where η is a given design, t_p is the p-quantile of the failure-time distribution, and $\mathbf{t}$ is a vector of observed data (failure times) collected according to η. Notice that improving the precision of the estimator of t_p means minimizing its posterior variance, i.e., minimizing $U(\eta)$ with respect to η. Because selecting the design must come before observing the data, in most cases the utility function cannot be calculated directly since it depends on the unobserved data. Instead, a preposterior expectation of the criterion function with respect to the unobserved data is performed. A desired experiment design would be one that minimizes the preposterior expectation of the criterion function:

$$\Lambda(\eta) = \int \text{var}(t_p|\mathbf{t}, \eta) dF_{\mathbf{t}} = \int U(\eta) dF_{\mathbf{t}}, \tag{1}$$

where $F_{\mathbf{t}}$ is the joint marginal distribution of the failure times. It is at this point that the difficulty arises in typical Bayesian design of experiments; how does one, in general,

*Statistical Sciences Group, CCS-6, Los Alamos National Laboratory, theguz@lanl.gov

 DOI: 10.1214/15-BA945

find a design that minimizes (or maximizes) a computationally difficult quantity (both numerically and analytically)?

Traditionally, normal approximations to the posterior distribution have been used to help alleviate the computational difficulties (Chaloner and Larntz, 1992). For example, the posterior variance of the failure-time quantile can be approximated by a function of the unknown model parameters θ and the Fisher information matrix (possibly through the delta method), denoted as $\mathbf{I}_\theta$ (notice that we have θ in the subscript to highlight its dependence on unknown parameter values),

$$\Lambda(\eta) \approx \int \mathbf{c}(\theta)' \mathbf{I}_\theta^{-1}(\eta) \mathbf{c}(\theta) \; dp(\theta), \tag{2}$$

where $p(\theta)$ is the assigned prior distribution for θ and $\mathbf{c}$ is a vector of partial derivatives of t_p with respect to the model parameters. A consequence of this approximation is that the integration in (2) is performed over the parameter space instead of both the sample space and the parameter space. A problem with this approach is that, in general, this may not be a good approximation to the true quantity the designers are interested in, and perhaps the quality of the approximation is not easily quantified.

An alternative solution to optimizing (1) is to use simulation to estimate $\Lambda(\eta)$. Then the surface of $\Lambda(\eta)$ can be adequately populated throughout the η space and the optimum (i.e., minimum) can be found, similar to the methods presented in Clyde et al. (1995). This approach, however, assumes that the computing time needed to calculate the integrand in (1) is small. In general, this may not be the case. For example, Picard and Williams (2013) present applications where researchers are interested in estimating rare event probabilities using computer experiments where a single code run requires a significant amount of time. In such situations it is not practical to run the computer code at enough η locations to adequately populate $\Lambda(\eta)$ due to the massive amount of time needed for the computations.

For physical or computer experiments that are restricted to smaller sample sizes (i.e., where large-sample approximations are not plausible) or where computation time is significant, a sequential approach to experiment design is preferred because it allows designers to be more efficient in optimizing the design and to leverage learning from early runs. The methods proposed in this paper are sequential in nature and are relevant to small-sample and large compute-time applications.

This sequential approach to designed experiments relates to the broader problem of resource allocation, where the experimenter must decide what new data to collect conditional on current understanding of the problem. Anderson-Cook et al. (2008) and Anderson-Cook et al. (2009) provide an overview of how balancing different proportions of data types can be considered as part of sequential optimization with flexible objectives.

The purpose of this paper is to propose an emulation approach for Bayesian design of experiments. In particular, we illustrate how a Gaussian spatial process (GP) can be used to mimic (1). The benefits of a GP are that it is quicker to calculate than (1) and requires fewer assumptions than the normal approximations to the posterior in (2). Additionally, this approach allows us to incorporate uncertainty from both the Monte Carlo estimates of the integrals in (1) and from the GP itself.

This paper and those cited below follow the framework developed by Raiffa and Schlaifer (1961), Lindley (1972), and Chaloner and Verdinelli (1995). Zhang and Meeker (2006) employ the methods from Chaloner and Verdinelli (1995) and replace the posterior variance by the variance of a normal approximation to the posterior distribution. For cases where the design space (i.e., the space that contains all possible experiments) is finite, Drovandi et al. (2013) and Drovandi et al. (2014) present efficient methods using sequential Monte Carlo and particle filters for selecting an optimum design in low-dimensional input spaces. Hamada et al. (2001) illustrates how genetic algorithms can be used to identify promising candidates for Bayesian design of experiments. Response surface methods for Bayesian design of experiments have been previously considered. For example, Clyde et al. (1995) and Müller (1999) use response surface representations of the criterion function for Bayesian design of experiments. In these papers, random designs were selected and Monte Carlo methods were used to estimate (1) for each design. A density estimate was then fit to these points (i.e., on $U(\eta)$ and η space) for each randomly generated design. The design yielding the maximum value of this surface was selected as optimum. In addition, Müller (1999) presents a Metropolis–Hastings approach for calculating (1). Recently, Huan and Marzouk (2013) showed how an emulator using polynomial chaos can be used to mimic (1), whereas in this paper we focus on building emulators of stochastic design criterion functions using GPs, taking advantage of readily available optimization software developed specifically for GPs (see Jones et al. (1998) and Picheny et al. (2013), for example).

The methods proposed in this paper can be used in two different design scenarios. The first problem, which our example represents, involves using the approach to determine the conditions for which a new physical experiment or test is to be performed. The second problem involves sequential design, such as iterative criterion-based design of computer experiments. In particular, these methods can be used to identify the next run of a computer code in a computer experiment. When this calculation is complete, the methods can be used again with the new information from the just completed run to determine the next run of the computer code.

The rest of this paper is organized as follows. Section 2 discusses the motivating application for this work and sets up the design problem. Section 3 introduces GPs and describes the algorithm that we use for finding an optimal ALT. Section 4 illustrates how to use a GP for finding the optimal design. Section 5 presents a simulation study used to illustrate the performance of our method. Lastly, Section 6 provides some conclusions.

2 Illustrative Example: Planning for Accelerated Life Tests

2.1 The Device-A Data

We consider the Device-A data in Table C.10 of the Appendix of Meeker and Escobar (1998) as our illustrative application. In this example, 165 units were tested at various levels of temperature, and their time of failure was recorded if the unit failed or alternatively the time the test ended if the observation was censored.

2.2 Accelerated Life Test Model

In Chapter 19 of Meeker and Escobar (1998), a lognormal distribution was fit to these data using maximum likelihood methods where the log-location parameter was modeled as a linear regression with the Arrhenius transformation of temperature, i.e.,

$$\mu = \gamma_0 + \gamma_1 x$$

where

$$x = \frac{11605}{\text{Temp}^{\circ}\text{C} + 273.15}.$$

For the device-A data, we follow the model of Meeker and Escobar (1998) and continue to use a lognormal distribution for the failure-times.

The lognormal distribution is a member of the log-location–scale family of distributions and is commonly used in reliability applications. A random variable $T > 0$ is a lognormal random variable if for $t > 0$

$$\Pr(T \leq t) = F(t) = \Phi_{\text{Norm}}\left(\frac{\log(t) - \mu}{\sigma}\right),$$

where Φ_{Norm} is the cumulative distribution function (cdf) of a standard normal distribution, $\mu \in \mathbb{R}$ is the log-location parameter, and $\sigma > 0$ is the log-scale parameter. We assume that σ is constant across temperatures.

The proposed accelerated model that describes the underlying failure mechanisms is only valid for a range of x, namely $[x_L, x_H]$. Testing at or below x_L does not produce any failures and testing above x_H tends to induce a new failure mechanism that is not seen in the field. Common practice is to standardize x to a new variable ξ so that $\xi \in [0, 1]$. One such standardization is $\xi(x) = (x - x_L)/(x_H - x_L)$. Adopting this standardization reparameterizes our accelerated model to $\mu = \beta_0 + \beta_1 \xi$ where $\beta_0 = \gamma_0 + \gamma_1 x_L$ and $\beta_1 = \gamma_1(x_H - x_L)$. For the lognormal distribution, the p quantile is

$$t_p = \exp(\mu + \Phi^{-1}_{\text{Norm}}(p)\sigma) = \exp(\beta_0 + \beta_1 \xi + \Phi^{-1}_{\text{Norm}}(p)\sigma). \tag{3}$$

In particular, we are interested in the p quantile at use conditions, which is specified when $\xi = 0$.

Lastly, for purposes of improving the convergence of our Markov chain Monte Carlo (MCMC) algorithm, we center ξ about 0.5 (its midpoint). This further reparameterizes our model to

$$\mu = \beta_0^* + \beta_1(\xi - 0.5), \tag{4}$$

where $\beta_0^* = \beta_0 + \beta_1 0.5$.

2.3 Accelerated Life Test Specification

We use the variable η to denote a given accelerated life test (design, for short). A design η tells us the factor-level combinations, $\boldsymbol{\xi}$, to be tested, the number of levels for testing,

k, and the number of units or proportion of units to be tested at those levels, π_i, $i = 1, \ldots, k$ where $\sum_{i=1}^{k} \pi_i = 1$. In general, $\boldsymbol{\xi}$ is a vector with length equal to the number of factors varying in the experiment (in our example, we only have 1 factor, temperature). Here,

$$\eta = \begin{bmatrix} \boldsymbol{\xi}_1, & \pi_1 \\ \vdots & \vdots \\ \boldsymbol{\xi}_k, & \pi_k \end{bmatrix}.$$

In general, there is some ambiguity to this notation. For example, consider the two designs

$$\eta_1 = \begin{bmatrix} \boldsymbol{\xi}_1, & \pi_1 \\ \boldsymbol{\xi}_2, & \pi_2 \end{bmatrix}, \tag{5}$$
$$\eta_2 = \begin{bmatrix} \boldsymbol{\xi}_2, & \pi_2 \\ \boldsymbol{\xi}_1, & \pi_1 \end{bmatrix}.$$

Notice that $\eta_1 = \eta_2$. To avoid such ambiguities, we recommend sorting the design by increasing π_i so that if $\pi_1 < \pi_2$, then both η_1 and η_2 can be represented by η_1 as defined in (5). In general, we define a design η as

$$\eta = \begin{bmatrix} \boldsymbol{\xi}_1, & \pi_1 \\ \vdots & \vdots \\ \boldsymbol{\xi}_k, & \pi_k \end{bmatrix},$$

where $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_k$.

2.4 Prior Distribution

Our accelerated model has three unknown parameters, $\theta = (\beta_0, \sigma, \beta_1)$. The available information about θ is quantified in a prior distribution denoted as $p(\theta)$. For the device-A example, our prior information $p(\theta)$ is based on the posterior distribution from an initial Bayesian model fit to the original device-A data, where diffuse priors were selected for θ. Specifically, a joint lognormal distribution $p(\theta)$ for the parameters was selected that matched the summary statistics from the initial Bayesian fit. The parameters for our prior distribution are given in Table 1. Note that this prior is given in terms of the centered parameterization used for the MCMC. Additionally, the posterior distribution obtained from analyzing the original device-A data assigned essentially all its probability to negative values of β_1. Therefore, we assigned a lognormal distribution to $-\beta_1$.

Model Parameter	Prior Log-location parameter	Prior Log-scale parameter	Prior Mean	Prior SD
β_0^*	2.282	0.027	9.80	0.265
σ	0.033	0.140	1.044	0.147
$-\beta_1$	1.667	0.135	5.345	0.725

Table 1: Lognormal hyperparameters for the prior distribution of θ.

2.5 Bayesian Planning Criterion

Suppose that we are in a situation where we either have a new sample of the device-A units or we have new devices that have similar properties to the device-A units such that the prior distribution given in Section 2.4 is applicable. The engineers are interested in planning an ALT that provides the best estimation precision for t_p at the nominal usage conditions. More specifically, the engineers want to know to which accelerated temperatures should the units be allocated in order to provide the most information for t_p at some nominal temperature. This is similar to the example in Zhang and Meeker (2006) where they designed plans to minimize the posterior variance of the log 0.10 quantile of the failure-time distribution at nominal conditions. The range of temperatures that the engineers are willing to test are $[10°\text{C}, 80°\text{C}]$, where $10°\text{C}$ is the nominal-use temperature and $80°\text{C}$ is the maximum temperature that can be used before an additional unwanted failure mechanism is induced. Note this range will be standardized to $[0, 1]$ where 0 is the nominal-use condition.

In our example, we wish to find the design η that minimizes the posterior variance of t_p (or equivalently, maximize the negative of the variance) at $\text{Temp}°\text{C} = 10°\text{C}$, i.e.,

$$\text{var}(t_p|\mathbf{t}, \eta) = \int [t_p - \text{E}(t_p|\mathbf{t}, \eta)]^2 f(t_p|\mathbf{t}, \eta) dt_p, \tag{6}$$

where $f(t_p|\mathbf{t}, \eta) = \int f(t_p|\mathbf{t}, \theta, \eta) p(\theta|\mathbf{t}, \eta) d\theta = \int f(t_p|\theta) p(\theta|\mathbf{t}, \eta) d\theta$ is the posterior distribution of t_p with expected value $\text{E}(t_p|\mathbf{t}, \eta)$ and $f(t_p|\theta)$ is a Dirac delta function. In general, $f(t_p|\mathbf{t}, \eta)$ is analytically intractable and we must therefore use simulation to estimate $\text{var}(t_p|\mathbf{t}, \eta)$. Because we are in the design phase of the ALT, the vector of item failure times $\mathbf{t}$ have not been observed. Instead, a preposterior expectation over the marginal distribution of the data will be used:

$$\Lambda(\eta) = \int \text{var}(t_p|\mathbf{t}, \eta) f(\mathbf{t}|\eta) d\mathbf{t}, \tag{7}$$

where $f(\mathbf{t}|\eta) = \int f(\mathbf{t}|\theta, \eta) p(\theta) d\theta$ and η gives the test temperatures and their corresponding proportions. Our simulation algorithm for estimating (7) is provided in Appendix A.1.

Equation (7) represents our test plan criterion. The goal at this stage is to find the test plan η^* such that

$$\eta^* = \text{argmin}_{\eta \in H} \Lambda(\eta)$$

where H is the class of suitable design measures.

Following the results of Zhang and Meeker (2006), we suppose for this application that the optimal design is a two-point plan, so we restrict our search to the class of plans structured as:

$$\eta = \begin{bmatrix} \xi_1, & \pi_1 \\ \xi_2, & \pi_2 \end{bmatrix} \tag{8}$$

with $\xi_1, \xi_2 \in [0, 1] \times [0, 1]$, $\pi_1 + \pi_2 = 1$, and $\pi_1 \in (0, 1)$. Notice that we require an open interval for π_1 to prevent the designs from collapsing to singular designs (designs

that do not allow estimability of all the parameters). These assumptions reduce design optimization to just a three-dimensional search over (ξ_1, ξ_2, π_1) because $\pi_2 = 1 - \pi_1$.

We assume we know how many items (or code runs if designing a computer experiment) will be used in the accelerated test (i.e., we know the length of $\mathbf{t}$). For our example, we select the sample size of 165 used in the original motivating experiment. In a general setting, this sample size may be dictated by the available budget for the physical experiment or the amount of processing time available on a supercomputer.

3 GP Based Search for the Optimal Design

3.1 Introduction of GPs

Criterion functions that are computationally intensive to compute, like $\Lambda(\eta)$, are prime candidates for statistical surrogate modeling. The criterion function $\Lambda(\eta)$ is evaluated directly on a limited set of candidate designs and a statistical surrogate is developed from the results that allows prediction of $\Lambda(\eta)$ with uncertainty quantification for arbitrary design η. This surrogate is chosen to be orders of magnitude faster to evaluate than the criterion function itself. We utilize a general approach to surrogate construction based on Gaussian spatial process models, which allow substantial flexibility in fitting smooth functions to available observations and lend themselves to automation in iterative algorithms such as optimization, which is essential to our primary goal of efficiently identifying an optimal design η^*.

Let $\mathcal{X}$ be a fixed subset of $\mathbb{R}^d$. We say that the random function $Y(\mathbf{x})$ is a GP for $\mathbf{x} \in \mathcal{X}$ if for any fixed L, and $\mathbf{x}_1, \ldots, \mathbf{x}_L \in \mathcal{X}$, the vector $(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_L))$ is jointly Gaussian distributed (Cramér and Leadbetter, 1967; Adler, 1981; Santner et al., 2003). A GP is defined in terms of its mean function $\mu(\mathbf{x}) = E[Y(\mathbf{x})]$ and covariance function $C(\mathbf{x}, \mathbf{x}') = Cov[Y(\mathbf{x}), Y(\mathbf{x}')]$. In the following, we assume further that the GP is second-order stationary, so that (i) its mean is constant, $\mu(\mathbf{x}) = \mu$ for all $\mathbf{x} \in \mathcal{X}$, and (ii) covariances of responses evaluated at two inputs $\mathbf{x}$ and $\mathbf{x}'$ only depend on their difference in $\mathcal{X}$, $C(\mathbf{x}, \mathbf{x}') = c(\mathbf{x} - \mathbf{x}')$ for an appropriate function c. Note this restriction implies that the process variance is constant, $\sigma^2 = C(\mathbf{x}, \mathbf{x}) = c(\mathbf{0})$ for all $\mathbf{x} \in \mathcal{X}$.

Often, $c(\cdot)$ is chosen from a parametric family of covariance (symmetric and positive semidefinite) functions to obtain a desired level of smoothness in process realizations. In our application, a product of one-dimensional Matérn covariance functions each having smoothness 5/2 (Matérn (1986)) is selected,

$$c(\mathbf{h}) = \sigma^2 \prod_{i=1}^{d} c_i(h_i) \text{ for } c_i(h) = \left(1 + \sqrt{5}\,\frac{|h|}{\alpha_i} + \frac{5}{3}\left(\frac{h}{\alpha_i}\right)^2\right) \exp\left(-\sqrt{5}\,\frac{|h|}{\alpha_i}\right),$$

where the range parameters α_i are to be estimated along with the constant mean μ and variance σ^2 as explained in the next paragraph. The resulting GP admits functional realizations that are twice continuously differentiable throughout $\mathcal{X}$.

In the following, we identify the isomorphism between the design $\eta = (\xi_1, \xi_2, \pi_1, \pi_2 = 1 - \pi_1)$ and the equivalent three parameters $\mathbf{x} = (\xi_1, \xi_2, \pi_1)$, noting that $\Lambda(\eta) = \Lambda(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ with input space $\mathcal{X}$ having bounds as previously specified. To minimize $\Lambda(\eta)$ in a computationally efficient manner, we first model the random function $\Lambda(\mathbf{x})$ as a GP. This process is evaluated at r selected initial designs $\eta_1, \ldots, \eta_r$ as described in the next section. Because evaluation of $\Lambda(\eta)$ is stochastic rather than deterministic, we include a nugget effect (see Chapter 4 of Santner et al. (2003) for more on the nugget effect) in the GP covariance function,

$$c_\varepsilon(\mathbf{h}) = c(\mathbf{h}) + \sigma_\varepsilon^2 \mathcal{I}(\mathbf{h} = \mathbf{0}),$$

where $\mathcal{I}(A)$ is the indicator function of the event A. The log-likelihood function of $\Lambda_r = (\Lambda(\eta_1), \ldots, \Lambda(\eta_r))$, up to a constant, is given by

$$-\frac{1}{2}\left[\log\det\left(\mathbf{C} + \sigma_\varepsilon^2 \mathbf{I}_r\right) + (\Lambda_r - \mu\mathbf{1}_r)^T \left(\mathbf{C} + \sigma_\varepsilon^2 \mathbf{I}_r\right)^{-1} (\Lambda_r - \mu\mathbf{1}_r)\right]$$

where $\mathbf{1}_r$ and $\mathbf{I}_r$ are the r-vector of ones and the $r \times r$ identity matrix, respectively, and the (i, j)-entry of $\mathbf{C}$ is given by $c(\mathbf{x}_i - \mathbf{x}_j)$. In the development to follow, the statistical model parameters $(\mu, \sigma^2, \sigma_\varepsilon^2, \alpha_1, \alpha_2, \alpha_3)$ are estimated by maximum likelihood using the `km` function distributed with the R package `DiceOptim` (Roustant et al., 2012).

3.2 GP Search for Optimal ALT

As mentioned in the introduction, our Bayesian design criterion presented in (7) is difficult to calculate because of the two imbedded integrals; the outer integral is with respect to the marginal distribution of the data and the inner integral is with respect to the posterior distribution of t_p. In general, these different integrals are analytically intractable and numerically intensive. Because (7) involves complicated integration, the inner integral is estimated using MCMC, in particular a Metropolis algorithm, and the outer integral is estimated using Monte Carlo integration.

Therefore, we minimize how often we perform the MCMC and Monte Carlo integration and instead use the GP to search the design space. This introduces the complication that as a random process, the GP can only predict the criterion value for any given design η up to a level of uncertainty determined by the available design evaluations. For deterministic criteria, the efficient global optimization (EGO) algorithm of Jones et al. (1998) introduces the notion of *expected improvement* criteria for optimization of computationally intensive computer experiments. EGO provides a criterion that balances local optimization (small predicted $\Lambda(\eta)$ and low prediction uncertainty) with uncertainty reduction (large predicted uncertainty in $\Lambda(\eta)$). Designs in the latter category are potentially promising and thus are investigated if the level of uncertainty in $\Lambda(\eta)$ could make them competitive with designs in the former category. These investigations reduce the prediction uncertainty in $\Lambda(\eta)$, allowing a more precise determination of the quality of the associated designs. Since this seminal work, the expected improvement concept has formed the basis of a wide variety of sequential experiment design criteria. Unfortunately, we cannot use EGO directly to optimize $\Lambda(\eta)$, as this criterion function is not deterministic upon evaluation at design η

because of the simulation approach used in its calculation. Huang et al. (2006) proposed heuristic modifications to the EGO framework for optimizing a uniformly noisy response. Gramacy and Lee (2010) point out that the EGO expected improvement criterion can be tweaked to satisfy a monotonicity constraint for use with nondeterministic output. Instead, we utilize the *expected quantile improvement* (EQI) framework recently developed by Picheny et al. (2013) to optimize our stochastic design criterion. EQI provides an expected improvement criterion consistent with the figure of merit adopted for selecting design candidates while facilitating user influence over the global vs. local nature of sequential iterates through specification of a future variance parameter.

We suppose that $\Lambda(\eta)$ has been evaluated for r initial designs $\mathbf{x}_1, \ldots, \mathbf{x}_r$ and a GP fit to the results $\Lambda_r = (\Lambda(\mathbf{x}_1), \ldots, \Lambda(\mathbf{x}_r))$, where $\mathbf{x} = (\xi_1, \xi_2, \pi_1)$ denotes the inputs defining designs η. With the GP parameters $(\mu, \sigma^2, \sigma_\varepsilon^2, \alpha_1, \alpha_2, \alpha_3)$ fixed at their maximum-likelihood values, we denote the mean and variance of the predictive GP (i.e., the process for predicting $\Lambda(\mathbf{x})$ with uncertainty quantification at arbitrary $\mathbf{x}$ based on the available data Λ_r) by $\mu_r(\mathbf{x})$ and $s_r^2(\mathbf{x})$ respectively. The β-quantile of this predictive GP is given by $q_r(\mathbf{x}) = \mu_r(\mathbf{x}) + \Phi_{\text{Norm}}^{-1}(\beta) s_r(\mathbf{x})$. The minimum β-quantile value from the evaluated designs is denoted $q_r^{\min} = \min\{q_r(\mathbf{x}_1), \ldots, q_r(\mathbf{x}_r)\}$. We replace q by Q to designate random versions of these quantities, necessary when the data vector Λ_r is viewed as random rather than fixed.

With these preliminaries, we define the *quantile improvement* by

$$\text{QI}_r(\mathbf{x}) = \max\{0, Q_r^{\min} - Q_{r+1}(\mathbf{x})\},$$

where $Q_{r+1}(\mathbf{x})$ refers to the (random) β-quantile based on Λ_r and the unobserved value $\Lambda(\mathbf{x})$. This quantity represents the random amount by which the β-quantile would be reduced if the design criterion were evaluated at new input $\mathbf{x}$ in addition to $\mathbf{x}_1, \ldots, \mathbf{x}_r$, which can never be known precisely. Instead, we compute the *expected* quantile improvement, which is the expected value of $\text{QI}_r(\mathbf{x})$ given the available data Λ_r,

$$\text{EQI}_r(\mathbf{x}) = \left(q_r^{\min} - \mu_{Q_{r+1}}\right) \Phi_{\text{Norm}}\left(\frac{q_r^{\min} - \mu_{Q_{r+1}}}{s_{Q_{r+1}}}\right) + s_{Q_{r+1}} \phi_{\text{Norm}}\left(\frac{q_r^{\min} - \mu_{Q_{r+1}}}{s_{Q_{r+1}}}\right),$$

where ϕ_{Norm} is the density function for the standard Gaussian distribution. Here, $\mu_{Q_{r+1}}$ is the conditional expected value and $s_{Q_{r+1}}^2$ the conditional variance of $Q_{r+1}(\mathbf{x})$ given Λ_r, which can be computed in closed form. We choose to then evaluate the design criterion at the value $\mathbf{x}_{r+1}$ that maximizes $\text{EQI}_r(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$. This process continues iteratively by augmenting the sets of available designs and evaluated design criterion values by $\mathbf{x}_{r+1}$ and $\Lambda(\mathbf{x}_{r+1})$, respectively, and updating the maximum likelihood values of the GP parameters until a budget of R allowable design criterion evaluations is expended. The EQI criterion is optimized using the `max_EQI` function distributed with the R package `DiceOptim` (Roustant et al., 2012), which utilizes an evolutionary search algorithm with a derivative-based (i.e., Newton or quasi-Newton) optimization method.

As with EGO, we note that the first term of $\text{EQI}_r(\mathbf{x})$ emphasizes local search ($\mu_{Q_{r+1}} \ll q_r^{\min}$ favored) while the second term emphasizes global search ($s_{Q_{r+1}}$ large). In EQI we are minimizing a specified upper bound ($\beta \geq 0.5$) of a stochastic process to account for noise in evaluation of the design criterion. If each such evaluation produced a deterministic result, EQI collapses to EGO. In order to compute $\text{EQI}_r(\mathbf{x})$, the degree of stochastic variability in $\Lambda(\cdot)$ associated with the new observation at $\mathbf{x}$ must be specified. We use a suggestion of Picheny et al. (2013) to set this *future variance* at $\sigma_\varepsilon^2/(R-j)$ where j is the number of runs in the current design. This allows EQI to globally explore the space of designs at the outset while transitioning to a more local search as the evaluation budget is expended.

We examine two approaches for generating the starting designs $\eta_1, \ldots, \eta_r$. In the first, the initial designs are chosen to be a random sample from the input space $\mathcal{X}$, while in the second, the initial designs are selected as a space-filling Latin hypercube sample (McKay et al., 1979) from $\mathcal{X}$. For budget R, we set $r = R/2$ and iterate EQI until the budget is expended as explained in the next section. The choice of r balances having adequate initial information to approximate the surface with reserving sufficient runs to take advantage of the understanding gained to optimize the solution.

4 Finding an Appropriate Design for the ALT Example

In this section, we find an optimal design based on the EQI algorithm and the Monte Carlo algorithms given in Appendix A.1. Figure 1 displays $\sqrt{\Lambda}$ as a function of η (i.e., each point in this space corresponds to a single η). The x and y axes give the temperatures (ξ_1 and ξ_2) for the test and the area of the point corresponds to π_1 (i.e., the proportion of units allocated to temperature 1). The gray points correspond to the initial surface generated for the EQI algorithm. In this case, these designs were generated using a Latin hypercube of size 30. The black points correspond to the next 30 (for a total budget of 60) candidate points selected by the EQI algorithm.

Notice that the EQI algorithm concentrated on two regions in the design space which are indicated by the large open ovals on the plot. These two regions yielded designs with the lowest values of Λ. These regions can be considered equivalent in the following sense: Recall the discussion in Section 2.3 where we describe how a given design has an equivalent counterpart by simply rearranging the rows of the design. This symmetry in the design space occurs because we did not impose any additional constraints on the optimization (for example, requiring $\pi_2 > \pi_1$), meaning that a design in region one, $\eta^{(1)}$, has an equivalent counterpart in region two, $\eta^{(2)}$, which is found by interchanging rows in its representation (8). The third row in Table 2 gives the accelerated test selected from this example. This test was selected by taking the design corresponding to the minimum estimated criterion value observed on the final design.

Figure 2 displays the expected-quantile improvement (Figure 2A) and the log of the nugget-effect estimate (Figure 2B), both as a function of the iteration. Figure 2C shows where the design was selected in the Temperature 1–Temperature 2 space where the symbol denotes the iteration. Notice that initially the nugget effect has not been discovered and the GP-inferred prediction uncertainty is low, so the greatest improvement

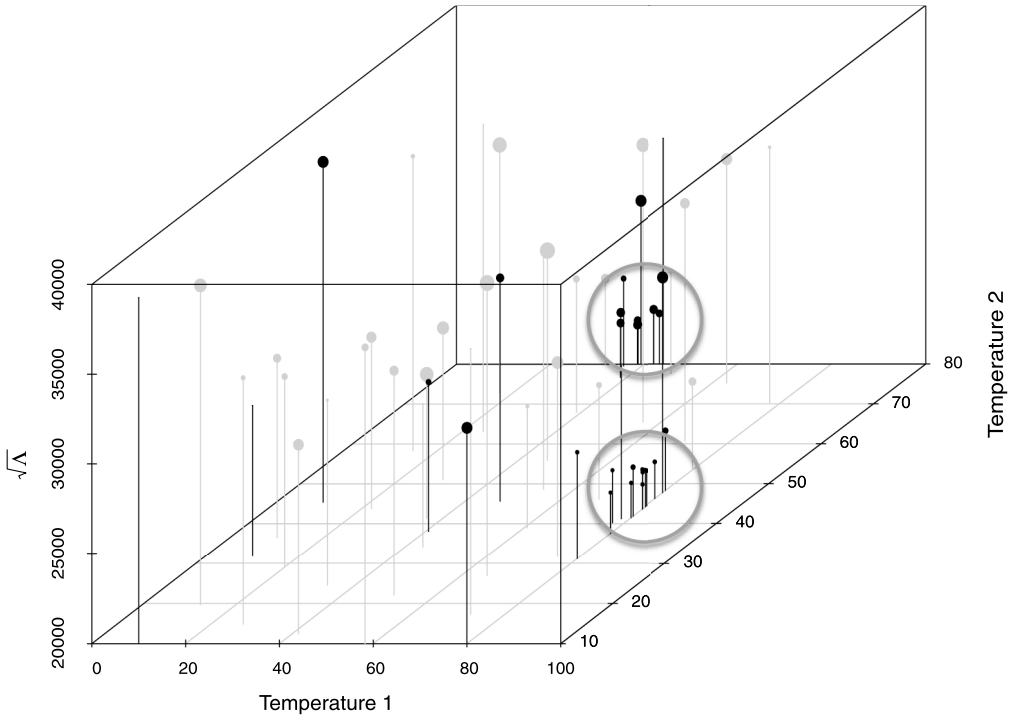


Figure 1: Plot of $\sqrt{\Lambda}$ as a function of the designs. The size of the dots is proportional to π_1 and the ovals are promising regions for locating the optimal design.

comes by selecting points in the regions where the GP predicts low criterion values. After iteration 50, the EQI algorithm has discovered the nugget effect and begins to select points that are at the boundaries of the experimental regions (iterations 55–60) where kriging models tend to have the most uncertainty. Note that although iteration 59 is not located at a boundary region in the Temperature 1–Temperature 2 space, it is on the boundary of the π_1 space (for this iteration, $\pi_1 = 0$). Eventually EQI would return to myopic searches near the optima with a larger budget of additional runs, as seen for example in Section 5.

In the next section, we compare how the GP method selects accelerated tests relative to other approaches currently in the literature.

4.1 Comparison with Traditional Methods

We now compare the results given in Section 3 with methods already present in the statistical literature. We use an approximation to (7) similar to (8) in Zhang and Meeker (2006), namely

$$\Lambda(\eta) \approx \int \mathbf{c}(\theta)'(\mathbf{I}_\theta(\eta) + \mathbf{S}^{-1})^{-1}\mathbf{c}(\theta)p(\theta)d\theta \tag{9}$$

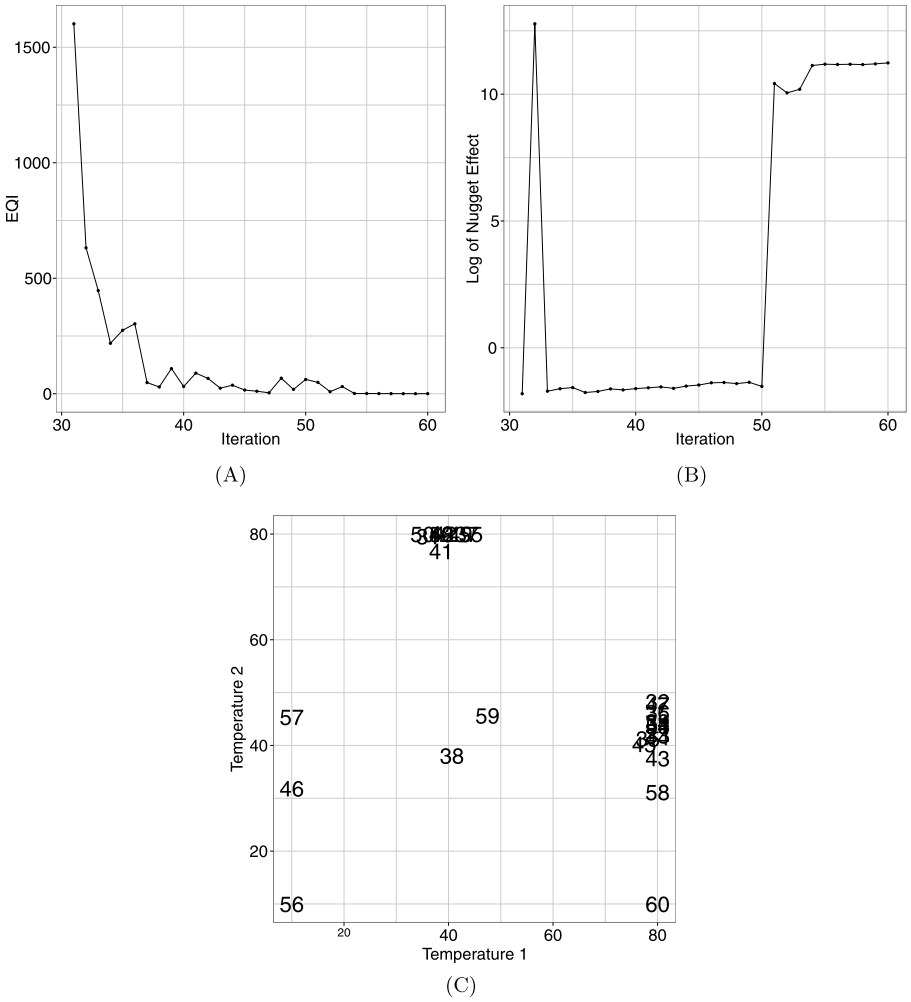


Figure 2: EQI (Figure 2A) and the log of the nugget-effect estimate (Figure 2B) as a function of the iteration after the initial design. Figure 2C shows where each design was selected in the Temperature 1–Temperature 2 space where the symbol denotes the iteration.

where again $p(\theta)$ is the prior of θ, $\mathbf{c}$ is the gradient vector of t_p with respect to θ (which is given in Appendix A.2), and $\mathbf{I}_\theta(\eta)$ is the Fisher information matrix evaluated at the design η and parameter values θ. The form of $\mathbf{I}_\theta(\eta)$ is

$$\mathbf{I}_\theta(\eta) = \frac{n}{\sigma^2} \sum \pi_i \mathcal{F}_i$$

where n is the total sample size for the test, π_i are the proportion of units allocated to the level ξ_i,

$$\mathcal{F}_i = \begin{bmatrix} f_{11}(\zeta_i) & f_{12}(\zeta_i) & f_{11}(\zeta_i)\xi_i \\ f_{12}(\zeta_i) & f_{22}(\zeta_i) & f_{12}(\zeta_i)\xi_i \\ f_{11}(\zeta_i)\xi_i & f_{12}(\zeta_i)\xi_i & f_{11}(\zeta_i)\xi_i^2 \end{bmatrix} \tag{10}$$

is the scaled Fisher information matrix for a single unit at ξ_i, and ζ_i is the standardized log-censoring time which for type I censoring is

$$\zeta_i = \frac{\log(t_c) - \mu_i}{\sigma} = \frac{\log(t_c) - \beta_0 - \beta_1\xi_i}{\sigma}.$$

The basic elements f_{11}, f_{12}, f_{22} are calculated using the algorithm provided in Escobar and Meeker (1994) at the standardized log-censoring times. In their paper, Zhang and Meeker (2006) automatically set the upper-standardized temperature to be 1 in their two-point plan and so the optimization is over the lower-standardized temperature, ξ_L, and the lower allocation proportion, π_L.

Zhang and Meeker (2006) require specification of a prior precision matrix $\mathbf{S}^{-1}$. Based on their discussion, this matrix represents the prior information for θ associated with estimation of these parameters, i.e., the prior information specifically used for estimating θ. This can be different than the prior information used for designing the test, $p(\theta)$, because, in general, those who design the test may have different goals than those who will perform the parameter estimation. For more on this distinction, see Tsutakawa (1972) and Etzioni and Kadane (1993). In our example, we choose a diffuse prior as no specific information is available to us from the Device-A study, which amounts to setting $\mathbf{S}^{-1} = \mathbf{0}$ in (9). Note that because we are setting $\mathbf{S}^{-1} = \mathbf{0}$, we are also able to factor n out of $\mathbf{I}_\theta(\eta)$ in estimation of (9).

4.2 Prior Information

We consider two prior distributions for the following calculations. As before, we use the prior distributions given in Section 2.4, which we denote as $p_1(\theta)$. In addition to these priors, we also assign a point-mass prior to θ (i.e., a prior that assigns all mass to a single point), denoted as $p_2(\theta)$. Note that this is equivalent to the frequentist methods given in Chapter 20 of Meeker and Escobar (1998) (for which the authors call $p_2(\theta)$ planning information). For $p_2(\theta)$, we use the maximum likelihood estimates (mle) given in Chapter 19 of Meeker and Escobar (1998) which are then converted to our parameterization of the model. $p_2(\theta)$ is given as

$$p_2(\theta) = \begin{cases} 1 & \text{if } \theta = (\beta_0, \sigma, \beta_1) = (12.321, 0.98, -5.118), \\ 0 & \text{otherwise.} \end{cases}$$

Table 2 gives the design using the methods in Zhang and Meeker (2006) and using priors p_1 and p_2. For our approach, we used a Latin hypercube design for generating the initial 30 design points and used a total budget of 60 runs. Notice that the three different approaches yield optimal designs that are quite similar and that the approach using the GP found the best design of the three.

Approach		ξ_L	π_L	$\sqrt{\Lambda(\eta)}$
Frequentist	$p_2(\theta)$	44	0.69	24886.4
Zhang and Meeker (2006)	$p_1(\theta)$	46.026	0.685	24939.6
GP	$p_1(\theta)$	41	0.65	24484.0

Table 2: Comparison of alternative methods for finding designs with the GP based method.

5 Examination of Different Starting Designs and Budgets

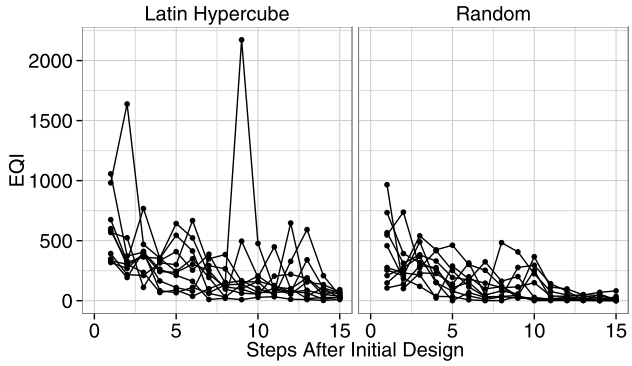
Recall in Section 3.2 that the EQI algorithm requires an initial collection of designs $\mathbf{x}_1, \ldots, \mathbf{x}_r$ with corresponding results $\Lambda_r = (\Lambda(\mathbf{x}_1), \ldots, \Lambda(\mathbf{x}_r))$ where $\mathbf{x} = (\xi_1, \xi_2, \pi_1)$. In this section, we consider two different ways of generating the set $\mathbf{x}_1, \ldots, \mathbf{x}_r$, by either randomly selecting this set or by using a Latin hypercube design. Furthermore, we consider different sizes of budgets R that we can commit to finding the Bayesian design and set the initial set size as $r = R/2$. For our study, we use $R = 30, 60, 90$.

The randomly-selected designs are generated by drawing three numbers from a uniform$(0, 1)$ distribution, two numbers for the accelerating variable (recall in Section 2.2 we standardized the accelerating variable to the unit interval $[0, 1]$) and one variable for the proportion allocated. We generate Latin hypercube designs by applying the algorithm described in Chapter 5 of Santner et al. (2003). First we divide each axis of the unit cube $[0, 1]^3$ into equally spaced intervals defined by the cut points $[0, 1/r, \ldots, [(r-1)/r, 1]$. Then for each input we sort these intervals according to a random permutation of the integers $1, \ldots, r$. Cartesian products of the sorted intervals across inputs identify r cells in the input domain. Points are then sampled from uniform distributions on each cell, resulting in a Latin hypercube design of size r.

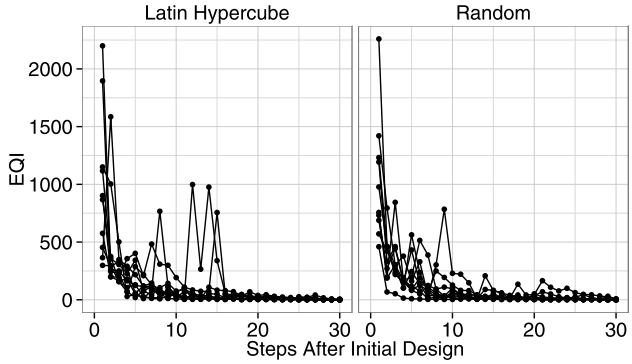
We generated 10 designs for each design type and budget (e.g., 10 randomly generated initial designs of size $r = 15$ and total budget $R = 30$, etc.). Then for a given generated design, we apply the algorithm given in Picheny et al. (2013) and search for the optimal design. These results are given in Figures 3–5. In each of these figures, the panels corresponding to A are for a budget of 30, the panels corresponding to B are for a budget of 60, and C has a budget of 90.

Figure 3 illustrates the EQI for both design types and for $R = 30, 60, 90$. Across these three plots, it is not obvious which of the two starting-design type is preferred. Notice that when $R = 30$ (Figure 3A), the initial random designs seem to have captured the important points relative to the EQI, which is not the case for the Latin hypercube design. This is not the case, however, for $R = 60$ (Figure 3B) and $R = 90$ (Figure 3C), where it appears both initial-design types perform equally well; both design types add a few important points early on and then the EQI gradually approaches zero, after about 15 iterations.

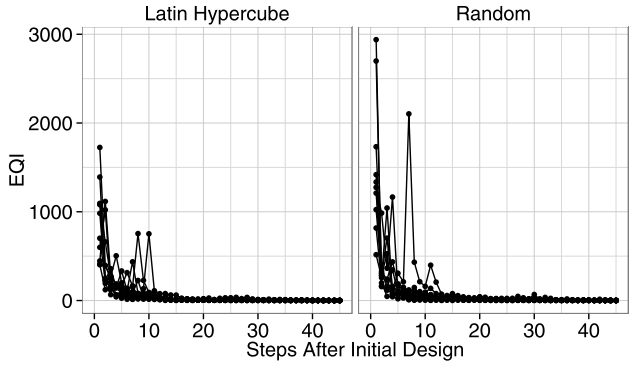
Figure 4 displays the (log) nugget effect for both design types and for $R = 30, 60, 90$. Again, as in Figure 3, there is no clear initial-design type that is preferred as both have



(A)

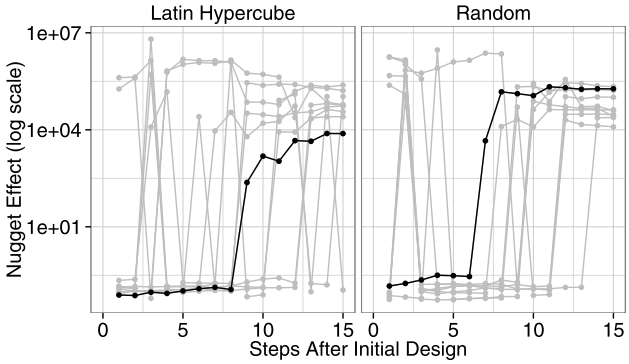


(B)

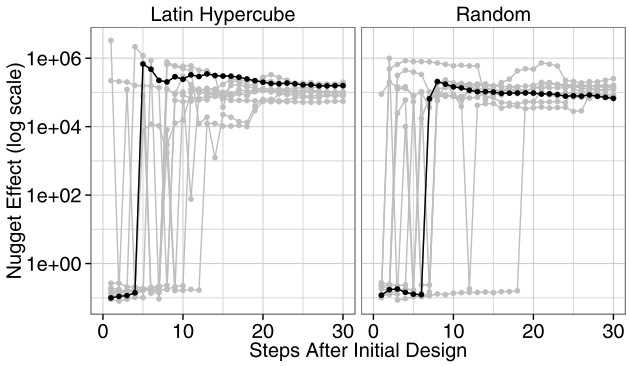


(C)

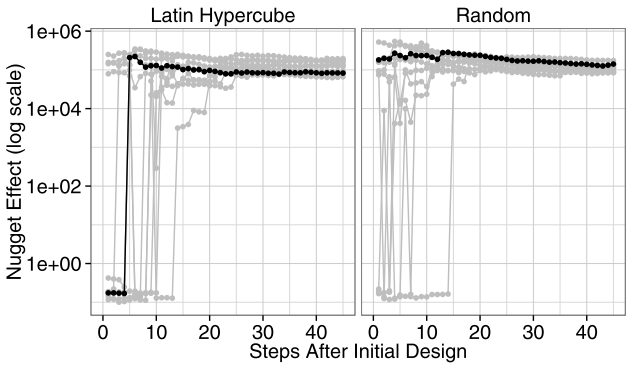
Figure 3: EQI as a function of steps after initial design and for $R = 30$ (A), 60 (B), and 90 runs (C).



(A)



(B)



(C)

Figure 4: Log-nugget effect as a function of steps after initial design and for $R = 30$ (A), 60 (B), and 90 runs (C). The black lines show the path of one of the simulation cases.

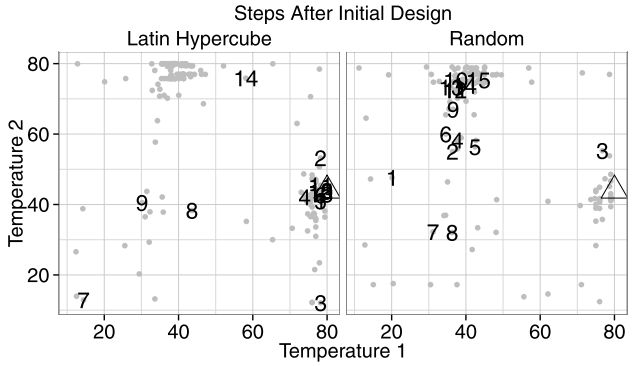
similar behaviors across R. Notice though that the nugget effect is discovered quite early in all the subpanels within Figure 4. Indeed, for the black paths highlighted for $R = 30$ (Figure 4A), the algorithm discovered the nugget effect after the eighth iteration for the Latin hypercube design and after about 6 iterations for the random design. These paths then appear to be converging to the same nugget level as represented in the plots for $R = 60$ (Figure 4B) and $R = 90$ (Figure 4C).

Lastly, Figure 5 shows the steps chosen by the EQI algorithm after the initial design within the Temperature 1–Temperature 2 space. The gray points in this plot display all points chosen in the simulations and the numbers correspond to a single realization from the simulation study. Notice that across all plots the points follow a similar pattern: they either focus on the two clusters of points corresponding to the region containing the minimum or they explore the boundaries of the space to reduce the uncertainty in the kriging estimator. The points that are roughly in the center of the plot have π_1 values at either 1 or 0, though this is not obvious from the figures. The sequence numbers illustrate that if less of the budget is allocated to the initial design, the algorithm will spend more of the earlier runs on globally searching the response surface to improve the fit of the GP. Once the response surface has been adequately resolved, the algorithm then focuses on identifying the minimum value of the response surface. If the initial design is large enough, resulting in the GP providing a good fit, then the algorithm immediately focuses on finding the minimum of the response surface. Finally, the triangle in the plot corresponds to the frequentist solution to this design problem. Notice that across all plots, the triangle is within the regions of the space where the EQI algorithm identified the minimum value.

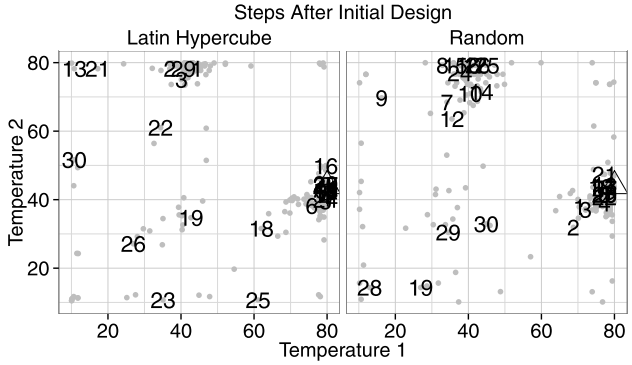
To conclude this section, we note that for our situation, there is minimal difference between results depending on which initial design is used. For the budget size, a value of $R = 30$ performed almost as well as the larger values of R. Of course, this is conditional on having half the budget dedicated to the initial design. More work is needed to understand the effects of different proportions of the entire budget used for the initial design size.

6 Conclusion

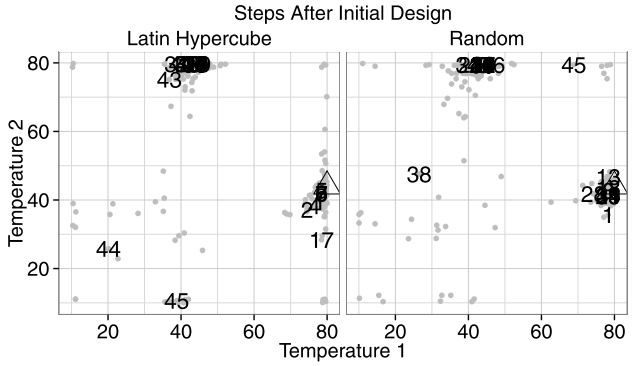
This paper provides an approach to reduce the computational burden of traditional Bayesian design of experiments by utilizing a Gaussian process as a surrogate for the design criterion to be optimized. For our accelerated testing application, this approach performed as well as some of the more established methods for solving this design problem. However, we emphasize that these methods apply more generally to settings in which existing methods are more likely to fail. For example, as mentioned in the introduction, these methods can be used in small sample applications and in situations involving computer codes that require extensive processor time. In particular, the Bayesian approach will work effectively when the joint posterior distribution is non-Gaussian, a scenario in which frequentist approaches that assume a Gaussian approximation to this distribution may fail.



(A)



(B)



(C)

Figure 5: Points selected by a single realization of the EQI in the temperature 1 and temperature 2 space for $R = 30$ (A), 60 (B), and 90 runs (C). The light gray points are all points chosen across simulations and the triangle denotes the region of the optimum.

Our initial designs did not contain any replicates, leading the EQI algorithm to proceed several iterations before discovering a stable estimate of the nugget effect. An alternate approach would design some replicates into the initial design, allowing the nugget effect to be better estimated going into EQI iterations at the expense of initial design space coverage.

Though not discussed in this paper, these methods are also applicable in uncertainty quantification applications. Bayesian design of experiments allows designers to incorporate parameter uncertainty in the form of a prior distribution. Using our proposed methods, experiments can then be designed to provide the most information about a flexibly defined metric, including selected quantiles of the predictive distribution for the quantity of interest.

Appendix

A.1 Simulation Algorithm for Estimating Equation (7)

For a given $\mathbf{t}$ assume that a collection of samples $(\theta^1, \ldots, \theta^B)$ of $p(\theta|\mathbf{t}, \eta)$ can be obtained (for example, we used a Metropolis algorithm as in Chib and Greenberg (1995)) for B being a large integer. Then for each i,

(a) Sample θ^i from $p(\theta|\mathbf{t}, \eta)$.

(b) Calculate $t_p^i = \exp(\beta_0^i + \Phi_{\text{Norm}}^{-1}(p)\sigma^i)$.

The $\text{var}(t_p|\mathbf{t}, \eta)$ is then approximated using Monte Carlo integration (see Chapter 3 of Robert and Casella (2004)),

$$\text{var}(t_p|\mathbf{t}, \eta) \approx \frac{\sum_{i=1}^{B}(t_p^i - \bar{t}_p)^2}{B-1} \quad \text{where } \bar{t}_p = \frac{\sum_{i=1}^{B} t_p^i}{B}. \tag{11}$$

We use the following steps to simulate a vector $\mathbf{t}$ needed for the outer integral of (7):

(i) Sample θ from $p(\theta)$.

(ii) Sample $\mathbf{t}$ from $f(\mathbf{t}|\theta, \eta)$.

Iterating between these steps (i) and (ii) a large number of times yields a sampling of $\mathbf{t}$ vectors.

We collect the above components into the following algorithm to estimate $\Lambda(\eta)$ for a given design η,

1. Sample θ from $p(\theta)$
2. Sample $\mathbf{t}$ from $f(\mathbf{t}|\theta, \eta)$.

(a) Using $\mathbf{t}$ obtain θ^i from $p(\theta|\mathbf{t},\eta)$, the posterior distribution.

(b) Calculate $t_p^i = \exp(\beta_0^i + \Phi_{\text{Norm}}^{-1}(p)\sigma^i)$.

3. Estimate $\text{var}(t_p|\mathbf{t},\eta)$ using the chain for t_p^i, $i = 1,\ldots,B$.

4. Repeat steps (1)–(3) A times where A is a large integer, giving a sample of variances $\text{var}(t_p|\mathbf{t},\eta)_i$, $i = 1,\ldots,A$.

5. Estimate $\Lambda(\eta)$ using Monte Carlo integration

$$\Lambda(\eta) \approx \frac{\sum_{i=1}^{A} \text{var}(t_p|\mathbf{t},\eta)_i}{A}.$$

A.2 Derivation of c, the Gradient Vector of t_p

When $\xi = 0$, (3) reduces to

$$t_p = \exp(\beta_0 + \Phi_{\text{Norm}}^{-1}(p)\sigma). \tag{12}$$

The gradient vector $\mathbf{c}$ is found by taking the partial derivatives of (12) with respect to θ. Here,

$$\begin{aligned}
\frac{\partial t_p}{\partial \beta_0} &= t_p, \\
\frac{\partial t_p}{\partial \beta_1} &= 0, \\
\frac{\partial t_p}{\partial \sigma} &= t_p \Phi_{\text{Norm}}^{-1}(p).
\end{aligned}$$

And therefore, $\mathbf{c} = (t_p, 0, t_p\Phi_{\text{Norm}}^{-1}(p))$.

References

Adler, R. J. (1981). *The Geometry of Random Fields*. New York: J. Wiley. MR0611857. 197

Anderson-Cook, C. M., Graves, T. L., and Hamada, M. S. (2008). "Resource allocation: Sequential design for analyses involving several types of data." In: *IEEE International Conference on Industrial Engineering and Engineering Management, 2008. IEEM 2008*, 571–575. IEEE. 192

— (2009). "Resource allocation for reliability of a complex system with aging components." *Quality and Reliability Engineering International*, 25(4): 481–494. 192

Chaloner, K. and Larntz, K. (1992). "Bayesian design for accelerated life testing." *Journal of Statistical Planning and Inference*, 33(2): 245–259. 192

Chaloner, K. and Verdinelli, I. (1995). "Bayesian Experimental Design: A Review." *Statistical Science*, 10(3): pp. 273–304. URL: http://www.jstor.org/stable/2246015 MR1390519. 193

Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis–Hastings algorithm." *The American Statistician*, 327–335. 209

Clyde, M., Müller, P., and Parmigiani, G. (1995). "Optimal Design for Heart Defibrillators." In: Gatsonis, C., Hodges, J., Kass, R., and Singpurwalla, N. (eds.), *Case Studies in Bayesian Statistics 2*, 278–292. Springer. 192, 193

Cramér, H. and Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes*. New York: J. Wiley. MR0217860. 197

Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2013). "Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data." *Computational Statistics and Data Analysis*, 57(1): 320–335. MR2981091. doi: http://dx.doi.org/10.1016/j.csda.2012.05.014. 193

— (2014). "A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design." *Journal of Computational and Graphical Statistics*, 23(1): 3–24. MR3173758. doi: http://dx.doi.org/10.1080/10618600.2012.730083. 193

Escobar, L. and Meeker, W. (1994). "Algorithm AS 292: Fisher information matrix for the extreme value, normal and logistic distributions and censored data." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3): 533–540. 203

Etzioni, R. and Kadane, J. B. (1993). "Optimal experimental design for another's analysis." *Journal of the American Statistical Association*, 88(424): 1404–1411. MR1245377. 203

Gramacy, R. and Lee, H. (2010). "Optimization under Unknown Constraints." In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., P, D. A., Heckerman, A. F. M., and Smith, M. W. (eds.), *Bayesian Statistics*. Oxford: Oxford University Press, 9th edition. MR3204008. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0008. 199

Hamada, M., Martz, H., Reese, C., and Wilson, A. (2001). "Finding near-optimal Bayesian experimental designs via genetic algorithms." *The American Statistician*, 55(3): 175–181. MR1947263. doi: http://dx.doi.org/10.1198/000313001317098121. 193

Huan, X. and Marzoukm Y. M. (2013). "Simulation-based optimal Bayesian experimental design for nonlinear systems." *Journal of Computational Physics*, 232(1): 288–317. MR2994301. doi: http://dx.doi.org/10.1016/j.jcp.2012.08.013. 193

Huang, D., Allen, T. T., Notz, W. I., and Zeng, N. (2006). "Global optimization of stochastic black-box systems via sequential Kriging meta-models." *Journal of Global Optimization*, 34(3): 441–466. MR2222283. doi: http://dx.doi.org/10.1007/s10898-005-2454-3. 199

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). "Efficient Global Optimization of Expensive Black-Box Functions." *Journal of Global Optimization*, 13: 455–492. MR1673460. doi: http://dx.doi.org/10.1023/A:1008306431147. 193, 198

Lindley, D. V. (1972). *Bayesian Statistics – A Review*. SIAM. 193

Matérn, B. (1986). *Spatial Variation*. Springer Verlag, second edition. MR0867886. doi: http://dx.doi.org/10.1007/978-1-4615-7892-5. 197

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code." *Technometrics*, 21(2): 239–245. MR0533252. doi: http://dx.doi.org/10.2307/1268522. 200

Meeker, W. and Escobar, L. (1998). *Statistical methods for reliability data*. John Wiley and Sons, INC. 193, 194, 203

Müller, P. (1999). "Simulation Based Optimal Design." In: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 459–474. Oxford University Press. MR1723509. 193

Picard, R. and Williams, B. (2013). "Rare event estimation for computer models." *The American Statistician*, 67(1): 22–32. MR3044816. doi: http://dx.doi.org/10.1080/00031305.2012.751879. 192

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013). "Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision." *Technometrics*, 55: 2–13. MR3038476. doi: http://dx.doi.org/10.1080/00401706.2012.707580. 193, 199, 200, 204

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration. MR0117844. 193

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag. MR2080278. doi: http://dx.doi.org/10.1007/978-1-4757-4145-2. 209

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization." *Journal of Statistical Software*, 51: 1–55. 198, 199

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer. MR2160708. doi: http://dx.doi.org/10.1007/978-1-4757-3799-8. 197, 198, 204

Tsutakawa, R. K. (1972). "Design of experiment for bioassay." *Journal of the American Statistical Association*, 67(339): 584–590. 203

Zhang, Y. and Meeker, W. (2006). "Bayesian methods for planning accelerated life tests." *Technometrics*, 48(1): 49–60. MR2236527. doi: http://dx.doi.org/10.1198/004017005000000373. 193, 196, 201, 203, 204

Acknowledgments

The authors wish to thank the editor, associate editor, and two reviewers for their helpful comments on an earlier version of this paper. We also thank C.C. Essix for her encouragement and support.