

Bayesian Optimization for Intrinsically Noisy Response Surfaces

Anton van Beek

Professor, School of Mechanical and Materials Engineering, University College Dublin, Dublin, Ireland

ABSTRACT: While many advanced statistical methods for the design of experiments exist, it is still typical for physical experiments to be performed adaptively based on human intuition. As a consequence, experimental resources are wasted on sub-optimal experimental designs. Conversely, in the simulation-based design community, Bayesian optimization (BO) is often used to adaptively and efficiently identify the global optimum of a response surface. However, adopting these methods directly for the optimization of physical experiments is problematic due to the existence of experimental noise and the typically more stringent constraints on the experimental budget. Consequently, many simplifying assumptions need to be made in the BO framework, and it is currently not fully understood how these assumptions influence the performance of the method and the optimality of the final design. In this paper, we present an experimental study to investigate the influence of the controllable (e.g., number of samples, acquisition function, and covariance function) and noise factors (e.g., problem dimensionality, experimental noise magnitude, and experimental noise form) on the efficiency of the BO framework. The findings in this study include, that the Matér covariance function shows superior performance over all test problems and that the available experimental budget is most consequential when selecting the other settings of the BO scheme. With this study, we enable designers to make more efficient use of their physical experiments and provide insight into the use of BO with intrinsically noisy training data.

1. INTRODUCTION

While Bayesian optimization (BO) is a well-established method for data efficient optimization of deterministic and time intensive response surfaces [1], its application for the optimization of stochastic response surfaces is still an elusive objective. Stochastic response surfaces are functions that manifest intrinsic uncertainty so that when they are evaluated for the same inputs a variation in the output is observed. These types of response surfaces are encountered when doing physical experiments and some forms of simulation experiments (e.g., molecular dynamics [2], and agent-based models). One primary advantage of BO is that it provides a systematic approach to sequentially identify the next most appropriate input conditions to evaluate,

thus saving experimental resources. This approach is shown in Figure 1, where it can observe that the process starts with a small number of uniformly distributed samples. Subsequently, we have an iterative process that involves response surface approximation, identification of a new input condition, and then experimentation. Once a stopping condition has been reached this process is terminated. Moreover, the sampling decisions are made by maximizing an acquisition function that balances the mean of a posterior predictive distribution (i.e., a Bayesian-based response surface approximation conditioned on the available training data set) with the interpolation uncertainty (i.e., exploitation versus exploration).

The typical approach for the exploration of unknown response surfaces is the use of an experimental design that involves a set of a uniformly

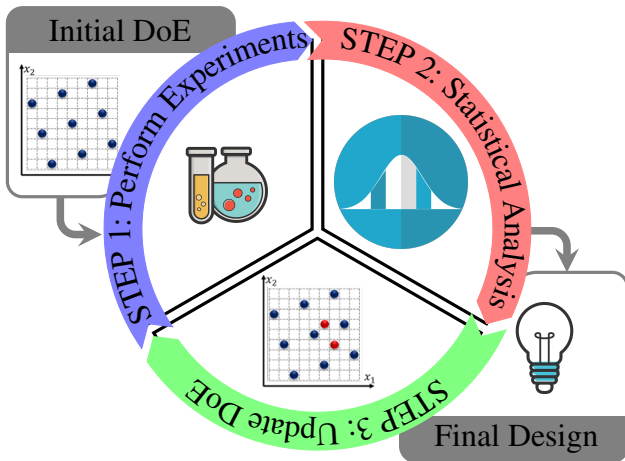


Figure 1: BO for the adaptive design of physical experiments, a generalized framework. In the iterative process between the acquisition of data and statistical analysis.

distributed set of input conditions [3, 4]. Examples of such experimental designs are factorial designs [5], Latin hypercube samples [6], and Sobol sequences [7]. These approaches have been applied for the design of experiments (DoEs) in the chemical sciences [8], civil engineering [9], and psychology [10]. While these approaches have been shown to be more data efficient than studying one factor at a time [11], they do not enable the designer to leverage the information of previous experiments to inform subsequent experiments. Consequently, BO provides a compelling, and potentially more data-efficient, alternative to one-shot DoEs.

The challenge when using BO to optimize stochastic response surfaces lies in the need to obtain a posterior predictive distribution, of the unknown response surface, that can quantify the intrinsic data uncertainty from the interpolation uncertainty. However, establishing such a posterior predictive distribution is a data-intensive task, and thus simplifying assumptions often need to be made. Examples of such methods include Practical Kriging [12], Stochastic Kriging [13], and Gaussian process (GP)-based quantile regression [14]. However, stochastic Kriging and GP-based quantile regression require many replicates to learn the form of the experimental uncertainty, whereas practical Kriging involves placing an additional GP on

the experimental variance that is data-intensive to learn.

In this paper, we present an empirical investigation into the effect that decisions in the construction of the posterior predictive distribution (e.g., choice of covariance function, acquisition function, the use of replicates, and initial batch size) and the properties of the response surface (e.g., noise magnitude, problem dimension, and noise form) have on the efficiency of the BO process. Through this effort we can make the following two knowledge claims about BO in the context of stochastic response surfaces:

1. The obtained insight enables engineers and scientists to use prior knowledge of the properties of the response surface to inform the construction of the posterior predictive distribution in the BO framework.
2. It highlights under what conditions it is appropriate to use the BO framework for the optimization of stochastic functions.

Finally, we will use the presented study to explore how the BO framework can be improved to be more appropriate for the optimization of stochastic functions. Through this effort, we hope to empower designers of physical experiments (e.g., engineers and scientists) to benefit from advanced statistical tools and get a better understanding of the knowledge embedded in the processes that they study.

2. STUDY BACKGROUND

In this section, we will introduce the methods used to approximate stochastic response surfaces, available acquisition functions, and the configuration of the experimental study presented in this work.

2.1. Gaussian Process Modeling

While multiple forms of response surfaces have been used for adaptive optimization of costly to evaluate objective functions (e.g., neural networks [15]), in this work we will only be using GPs. The reason is their ease of implementation and generalization to a plethora of different problems. Assume that we have a set of n noisy observations

$\mathbf{Y} = \{y_1, \dots, y_n\}^T$ for a set of d -dimensional input conditions $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}^T$. Consequently, we would like to establish an emulator on the function $f: \mathbb{R}^d \rightarrow \mathbb{R}$. Under the assumption that the observations are jointly normally distributed, we can place a GP prior on the unknown response surface f and characterize it through a mean function and a covariance function $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ [16]. Consequently, an observation model can be established as $y(\mathbf{x}_i) = f(\mathbf{x}_i) + \varepsilon_i$ where the noise is normally distributed as $\varepsilon_i \sim \mathcal{N}(0, \mathcal{S}(\mathbf{x}_i))$. Alternatively, we can write $Y \sim \mathcal{N}(\mathbf{M}\hat{\boldsymbol{\beta}}^T, \mathbf{K}_n + \boldsymbol{\Sigma}_n)$, where \mathbf{K}_n is a $n \times n$ matrix with the ij coordinates given as $k(\mathbf{x}_i, \mathbf{x}_j)$, $\boldsymbol{\Sigma}_n = \text{diag}(\mathcal{S}(\mathbf{x}_1), \dots, \mathcal{S}(\mathbf{x}_n))$ is a diagonal matrix that accounts for the experimental uncertainty, \mathbf{M} is a $n \times p$ matrix where the i^{th} row is a vector of p basis vectors given as $\mathbf{m}(\mathbf{x}) = \{m_1(\mathbf{x}), \dots, m_p(\mathbf{x})\}^T$.

Concerning the covariance structure of the GP, under the assumption of *homoscedastic* noise (i.e., $\mathcal{S}(\mathbf{x}) = \tau$), the experimental uncertainty can be defined by a single variable such that $\boldsymbol{\Sigma}_n = \tau \mathbf{I}_n$ where \mathbf{I}_n is an n -dimensional identity matrix. While GPs have been extended to be applicable to data with *heteroscedastic* noise (e.g., Practical Kriging [12], Stochastic Kriging [13], and GPs based quantile regression [14]), they are impractical for the optimization of physical experiments with small computational budgets (e.g., less than $10d$ samples [2]). In addition, the correlation between observations is accounted for through the covariance function $k(\cdot, \cdot)$, which is often selected as the squared exponential that is defined as

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \sigma^2 \exp\left(\sum_{i=1}^d -10^{\omega_i} (x_i - x'_i)^2\right), \\ &= \sigma^2 r(\mathbf{x}, \mathbf{x}'), \end{aligned} \quad (1)$$

where σ^2 is the prior variance and $\boldsymbol{\omega} = \{\omega_1, \dots, \omega_d\}^T$ is the roughness of the response surface.

Under these conditions, we can approximate the model parameters by maximizing their log-likelihood profile as

$$\hat{\boldsymbol{\omega}}, \hat{\tau} = \underset{\boldsymbol{\omega}, \tau \in \Omega \times T}{\text{argmax}} -n \log(\hat{\sigma}^2) - \log(|\mathbf{V}|), \quad (2)$$

where $\mathbf{V} = \mathbf{R}_n + \tau \mathbf{I}_n$, the i, j^{th} element of \mathbf{R}_n is $r(\mathbf{x}_i, \mathbf{x}_j)$, and the constant terms have been dropped. In addition, the search space has been defined as $\Omega \in [-10, 10]^d$ and $T \in [0, 1]$ (this is reasonable when normalizing the training data). Note that the consideration of experimental noise adds only a single additional hyperparameter to infer. Moreover, taking the derivative of the likelihood we can solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ as

$$\hat{\boldsymbol{\beta}} = (\mathbf{M}^T \mathbf{V}^{-1} \mathbf{M}) \mathbf{M}^T \mathbf{V}^{-1} \mathbf{Y}, \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}). \quad (4)$$

As an alternative to maximum likelihood estimation, the designer can use the maximum a posteriori probability, and cross-validation to get a points estimate of the hyperparameters or use a Bayesian approach to account for the uncertainty in the hyperparameters [16]. While the Bayesian approach has shown to be superior in performance [17], we rely on the maximum likelihood approximations of the parameters for its numerical stability and computational efficiency [18].

Having a point estimate of the hyperparameters enables a designer to condition the prior distribution of the unknown response surface on the observed data $\mathbf{D} = \{\mathbf{X}, \mathbf{Y}\}$. Specifically, the posterior approximation $Y(\mathbf{x})|\mathbf{D}$ for input conditions \mathbf{x} are fully defined through its mean and variance as

$$\mu(\mathbf{x}) = \mathbf{m}(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{k}(\mathbf{x})^T \Lambda^{-1} (\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}), \quad (5)$$

$$\begin{aligned} s^2(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{x}) \Lambda^{-1} \mathbf{k}(\mathbf{x}) \\ &\quad + \mathbf{W}^T (\mathbf{M}^T \Lambda^{-1} \mathbf{M})^{-1} \mathbf{W} + \hat{\sigma}^2 \hat{\tau}, \end{aligned} \quad (6)$$

respectively. Moreover, $\mathbf{W} = \mathbf{m}(\mathbf{x}) - \mathbf{M}^T \Lambda^{-1} \mathbf{k}(\mathbf{x})$, $\Lambda = \hat{\sigma}^2 \mathbf{V}$, and $\mathbf{k}(\mathbf{x})$ is an $n \times 1$ -dimensional vector whose i^{th} element is given as $k(\mathbf{X}_i, \mathbf{x})$. This posterior approximation has the advantage that it provides a quantification of the prediction uncertainty; however, it is not readily salient how it can be used to identify new input conditions to test.

2.2. Acquisition Functions

A wide variety of acquisition functions have been proposed in the literature; however, none of them have proven to universally outperform the others

[1]. These acquisition functions are used to identify what input conditions \mathbf{x}_t^{new} should be tested next at step t of the optimization process according to

$$\mathbf{x}_t^{new} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \alpha(\mathbf{x}|\mathbf{D}), \quad (7)$$

where \mathcal{X} is space of admissible input conditions, and $\alpha(\cdot)$ is the selected acquisition function. In this subsection, we will introduce a set of five alternative acquisition functions that are investigated in this study.

The first acquisition function that we will introduce is the statistical lower bound [19]. This is considered a nonrigorous branch-and-bound algorithm that involves minimizing $\mu(\mathbf{x}) - \pi s(\mathbf{x})$ where for any value of $\pi > 0$ we have the desired property of balancing exploration with exploitation [20]. To be consistent with the optimization formulate given in Equation 7, we reformulate this expression as

$$\alpha_{UC}(\mathbf{x}) = \pi s(\mathbf{x}) + \mu(\mathbf{x}), \quad (8)$$

and refer to this as the upper confidence (UC) acquisition function throughout the remainder of this paper. While the UC objective is intuitively appealing, it is known to exclude regions of the space of admissible input conditions, and thus does not guarantee the necessary sample density requirement to ensure convergence to a global optimum [21]. In panel A of Figure 2 we show the UC for four different values of π , from which it can be observed that more emphasis is placed on exploration for larger values of π .

The second acquisition function that we will investigate is the *probability of improvement* (PI)[20]. Improvement in this method is defined as $I(\mathbf{x}) = \max(\{y_t^* - Y(\mathbf{x}), 0\})$, where y_t^* is the best observed sample at the t^{th} iteration (i.e., $y_t^* = \min(\mathbf{Y})$). Given that $y(\mathbf{x})$ is a random variable, it becomes possible to calculate the PI through integration of the improvement from negative infinity up to the current best observed sample y_t^* . However, using this directly as an acquisition function has shown to place too much emphasis on exploitation, making it susceptible to waste experimental resources on improving the response surface accuracy around local optima. Consequently, the typical implementation

of the PI is defined as

$$\alpha_{PI}(\mathbf{x}) = \Phi\left(\frac{y_t^{(tar)} - \mu(\mathbf{x})}{s(\mathbf{x})}\right), \quad (9)$$

where $y_t^{(tar)} = y_t^* - \lambda (\max(\mathbf{Y}) - y_t^*)$ is the target response surface value, and $\Phi(\cdot)$ is the standard normal cumulative density function. In this case, designers can place more emphasis on exploration by selecting larger values for λ . For visualization of the PI, in panel B of Figure 2 we have plotted the PI for a test function with respect to five different values of $\lambda = \{0, 0.05, 0.10, 0.2, 0.5\}$ (scaled to be in a range of 0 to 3). Observe how the new sampling locations change more towards exploring unobserved regions for higher values of λ .

An alternative to the PI function is the *expected improvement* (EI) function. As the name suggests, it involves taking the expectation of the improvement as $\mathbb{E}(\max(\{y_t^* - Y(\mathbf{x}), 0\}))$. However, in the case of the stochastic response surface, the existence of experimental uncertainty places too much emphasis on exploitation. Consequently, [22] proposed a modified version of the EI function as

$$\begin{aligned} \alpha_{EI}(\mathbf{x}) &= \mathbb{E}(\max(\{y_t^* - Y(\mathbf{x}), 0\})) \left(1 - \frac{\hat{\sigma}^2 \hat{\tau}}{s(\mathbf{x})}\right), \\ &= s(\mathbf{x}) (u\Phi(u) + \phi(u)) \left(1 - \frac{\hat{\sigma}^2 \hat{\tau}}{s(\mathbf{x})}\right), \end{aligned} \quad (10)$$

where $\phi(\cdot)$ is the standard normal probability density function and $u = \frac{y_t^* - \mu(\mathbf{x})}{s(\mathbf{x})}$. Note that $s(\mathbf{x})$ equals $\hat{\sigma}^2 \hat{\tau}$ for any $\mathbf{x} \in \mathbf{X}$. For visualization purposes, we have plotted the modified EI function Equation 10 in panel C of Figure 2. Observe how the conventional EI is nonzero for observed samples and as such new samples are more likely to be allocated close to the current observed best sample y_t^* .

The next acquisition function that we will test is the *knowledge gradient* (KG) that was first introduced in [23]. The KG is an acquisition function that involves a one-step look-ahead policy that aims to maximize the difference between the optimal response $\min(Y(\mathbf{x})|\mathbf{D}_t)$ at step t and the optimal response after observing $\mathbf{x}_t^{(new)}$ [24]. Specifically, the

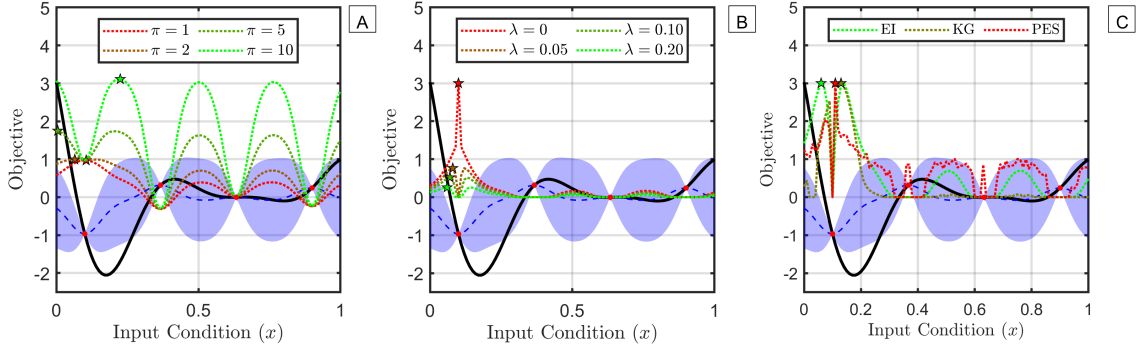


Figure 2: Visualization of acquisition functions for an arbitrary objective function (black line) that has been approximated with a GP (blue dashed line shaded region), trained on four samples (red dots). A) scaled UC for four values of $\pi = \{1, 2, 5, 10\}$, B) normalized PI for four different values of $\lambda = \{0, 0.05, 0.10, 0.20\}$, and C is the normalized EI, KG, and PES.) The stars indicate the optimal and new sampling locations for each acquisition function.

KG is defined as

$$\alpha_{KG}(\mathbf{x}) = \mathbb{E}(Y(\mathbf{x}_t^*)|\mathbf{D}_t - Y(\mathbf{x}_{t+1}^*)|\mathbf{D}_{t+1}, \mathbf{x}_t^{new} = \mathbf{x}), \quad (11)$$

where the subscripts on \mathbf{D}_t have been used to indicate that new observations have been added to the initial training data set (i.e., $\mathbf{D}_t = \bigcup_{i=1}^t \{x_i^{new}, y_i^{new}\} \cup \mathbf{D}$). However, it should be noted that for \mathbf{D}_{t+1} we have not yet observed the response y_{t+1}^{new} , and thus we need to take the expectation with respect to its predicted value [25]. In panel C of Figure 2 we have plotted the KG for a test function, and show its slight difference from the other acquisition functions.

The final accustoming function that we will introduce and investigate is the *predictive entropy search* (PES) that was first introduced in [26]. Instead of trying to maximize the improvement of the KG, the PES uses information theory to maximize learning about the spatial location of the globally optimal response. Specifically, the acquisition function is defined as

$$\alpha_{KG}(\mathbf{x}) = H(Y(\mathbf{x})|\mathbf{D}_t) - \mathbb{E}_{p(\mathbf{x}^*|\mathbf{D}_t)}(H(Y(\mathbf{x}_{t+1}^*)|\mathbf{D}_{t+1}, \mathbf{x}_t^{new} = \mathbf{x})), \quad (12)$$

where $H(p(\mathbf{x})) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$, and $\mathbb{E}_{p(\mathbf{x}^*|\mathbf{D}_t)}$ is the probability that the global optimum is found at input condition \mathbf{x} after observing observations \mathbf{D}_t . The first term on the left-hand side of

Equation 12 has a closed-form expression, whereas the right-hand side must be approximated through, for example, expectation propagation [27]. Similar to previous acquisition functions, we have plotted the PES in panel C of Figure 2.

2.3. Experimental Setup

In this subsection, we will discuss the setup of the experimental setup used to investigate the importance of alternative optimization conditions. For this purpose, we have made a distinction between factors that are controllable and a set of uncontrollable noise factors as shown in Table 1. The purpose of this distinction is that prior knowledge of the noise factors could inform what levels to choose for the controllable factors.

We considered four controllable factors, the initial number of replicates, the initial number of samples, the acquisition function, and the selected covariance function. The initial number of samples is the unique sampling location in the initial DoE. However, for small sample sizes, it might be found that no accurate approximation of the experimental variance $\sigma^2\tau$ can be obtained. Consequently, we also considered the scenario of having an initial number of replicates (i.e., for the same unique initial inputs the response surface function is evaluated 1, 2, or 3 times). In addition, we considered the use of all five previously introduced acquisition functions where for the UC and PI, we set $\pi = 5$ and

Table 1: Selected factors and their associated levels under which the experiments have been performed.

Level	Controllable Factors				Noise Factors		
	Initial replicates	Initial samples	Acquisition function	Covariance function	Problems	Noise magnitude	Noise form
1	1	2d	UCB	Gaussian	#1 = $f_1(\cdot)$	0.01 Δ_f	Constant
2	2	5d	PI	Power	#2 = $f_2(\cdot)$	0.05 Δ_f	Bad
3	3	10d	EI	Matérn	#3 = $f_3(\cdot)$	0.20 Δ_f	Good
4			KG				
4			PES				

$\lambda = 0.1$, respectively. Finally, we considered three different types of covariance functions, the previously introduced Gaussian covariance, the power exponential covariance function that are defined as

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(\sum_{i=1}^d -10^{\omega_i} (x_i - x'_i)^p\right), \quad (13)$$

and the Matérn covariance function that is defined as

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho}\right)^\nu K_\nu\left(\sqrt{2\nu} \frac{d}{\rho}\right), \quad (14)$$

where p, ρ, ν are additional hyperparameters, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, and $\Gamma(\cdot)$ is the gamma function. The consideration of additional power exponential and Matérn covariance is of interest as they allow more freedom in the form of the covariance but involves additional parameters. Consequently, considering their choice in this study will help determine under what conditions using more complex covariance functions will be beneficial.

Concerning the noise factors, we considered the magnitude of the noise and form of the noise. Specifically, we considered three noise scenarios where the maximum standard deviation of the noise is either $\{2, 5, 20\}$ % of the range of the function Δ_f . In addition, we considered three types of experimental uncertainty, one where the variance of the response surface is homoscedastic, one where the noise form is disadvantaged as the variance is maximized at the minimum of the response surface (referred to as bad), and one advantage where the variance is minimized at the global objective (referred to as good). This has been achieved through the formulation $\varepsilon(\mathbf{x}) \sim \mathcal{N}(0, \mathbb{E}(f(\mathbf{x}) + b)a)$, where parameter a and b have been selected to ensure that

the standard deviation ranges from 0.25 to 1.6 times the selected noise magnitude [2]. Finally, we considered the following set of noisy objective functions

$$f_1(x) = (3x - 2)^2 \sin(12x - 4) + \varepsilon, \quad (15)$$

$$f_2(\mathbf{x}) = \frac{1}{51.95} \left(15x_2 - \frac{5.1(15x_1 - 5)^2}{4\pi^2} + \frac{5(15x_1 - 5)}{\pi} - 6\right)^2 + \frac{1}{51.95} \left(\left(10 - \frac{10}{8\pi}\right) \cos(15x_1 - 5) - 44.81\right) + \varepsilon, \quad (16)$$

$$f_3(\mathbf{x}) = 4x_1^2 - 2.1x_1^4 + \frac{x_1^6}{3} + x_1x_2 - 4x_2^2 + 4x_2^4 + \varepsilon, \quad (17)$$

where $\varepsilon \sim \mathcal{N}(0, \mathcal{S}(\mathbf{x}))$. In addition, the spaces of admissible input conditions are $x \in (0, 1)$, $\mathbf{x} \in (0, 1)^2$, and $\mathbf{x} \in (-2, 2) \times (-1, 1)$ for $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$, respectively. Note that we have used only low-dimensional problems as optimizing high-dimensional objective functions will be practically infeasible with small experimental budgets.

From the above set of factors, we are able to identify a total of $3^6 \times 5 = 3645$ unique experiments. Finally, to account for the potential variability associated with random initial conditions, we have repeated all experiments five times for a total of 18225 experiments.

3. RESULTS AND INTERPRETATION

In this section, we will study the data obtained from the experiments delineated in the previous section. Specifically, we aim to find what the main effects are of each individual factor and then try to identify the interaction effects between the controllable and noise factors to help guide modeling decisions.

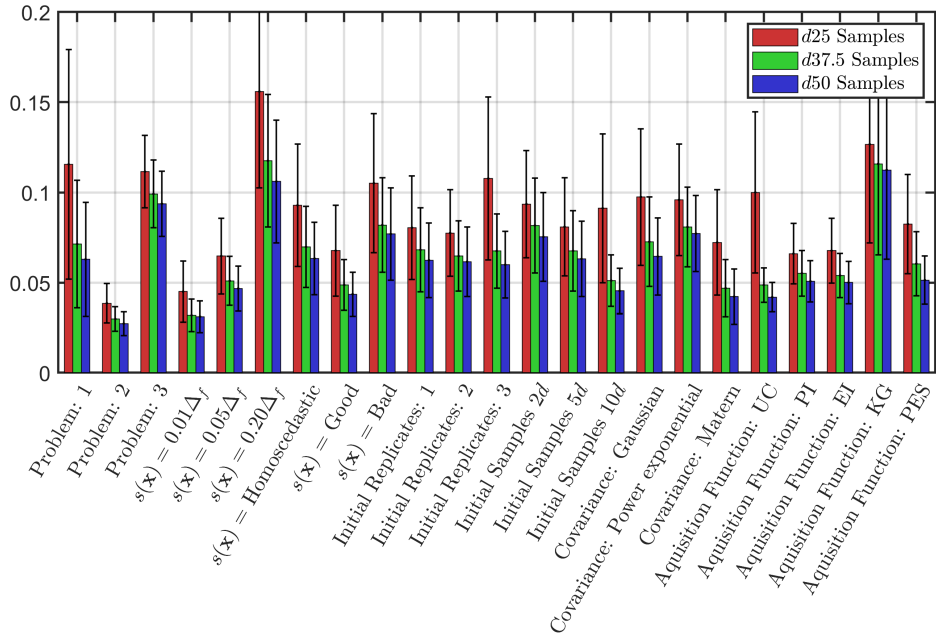


Figure 3: The main effects of the controllable and noise factors as measured for the GAP obtained after observing 25d (red bars), 37.5d (green bars), and 50d samples (blue bars).

3.1. Main Factor Effects

The main effects in this study have been measured by taking the average difference between the identified global optimum and the true global optimum for each factor. This metric is referred to as the GAP [28]. A bar chart of these results has been plotted in Figure 3 where the red bars indicate the GAP after observing the first 25d samples, the green bars indicate the GAP after observing the first 37.5d samples, and the blue bars indicate the gap after observing all 50d samples.

What can be observed from Figure 3 is that the initial number of replicates has only a minor effect on the optimality of the result obtained from the optimization process. Except, as fewer total samples have been observed, we find that the GAP becomes larger. This is intuitively sensible because when only a small number of samples have been observed then replication of experiments results in less coverage of the admissible design space. Conversely, when looking at the number of initial samples, we find that a large set of initial experiments has a positive effect on the GAP of the final result. What this implies, is that the tendency of the acquisition functions studied in this paper is that they empha-

size the exploitation over exploration. Concerning the choice of covariance function, we find that the power exponential performs worst, while the matér covariance performs significantly better. This could suggest that the freedom offered by the matér covariance is beneficial to the optimization process. However, it is interesting to note that this is not the case for the power exponential. The reason for this might be that the Matér covariance provides significantly more modeling freedom compared to the power exponential. Finally, concerning the acquisition function, it is interesting to observe that the UC acquisition function works best for large data scenarios (e.g., 50d) whereas, in low data scenarios, (e.g., 25d) the PI and EI appear to have similar performance. The reason for this might be that we used a relatively large value for $\pi = 5$ that significantly emphasizes exploration. Consequently, the main effects of the controllable factors suggest that conventional acquisition place too much emphasis on exploitation and too little on exploration.

Concerning the main effects of the noise factors, we observed intuitively sensible results. Specifically, we find that the relatively linear problem 2 has a small average GAP, whereas the opposite is

observed for problem 2 which has six local minima. In addition, functions with higher magnitudes of noise perform worst in the observed final GAP. Finally, functions with heteroscedastic noise that are minimum at the global optimum perform best. While these insights provide little novelty in terms of insight, they do provide validation of the performed study.

3.2. Interaction Effects

Next, we might be interested in investigating the interaction effects between the controllable and noise factors, as this could guide designers to make modeling decisions based on prior knowledge that they have of their experimental setup. In Figure 4 we have plotted the GAP averaged over all experiments that have the same controllable and noise factors for the different number of observed total samples (i.e., $d = \{25d, 37.5d, 50d\}$).

What we can observe from Figure 4 is that there are few interaction effects between the controllable and noise factors. This can be concluded from the observation that the optimal levels for each controllable factor are the same. The exception to this are the following two interaction effects.

1. The total number of observed samples has an interaction effect with the initial number of replicates. Specifically, as the experimental budget increases, it becomes advantageous to increase the initial number of replicates.
2. The total number of observed samples has an interaction effect with the selected acquisition function. Specifically, for a small experimental budget, the PI is most appropriate, whereas for larger experimental budgets the UC bound becomes more appropriate.

The interaction effect between the total number of samples and the number of initial replicates suggests that replication is beneficial for approximating the stochastic response surfaces. Specifically, when a larger experimental budget is available. Concerning the second interaction effect, we find that the acquisition function that places more emphasis on exploitation performs better with a larger

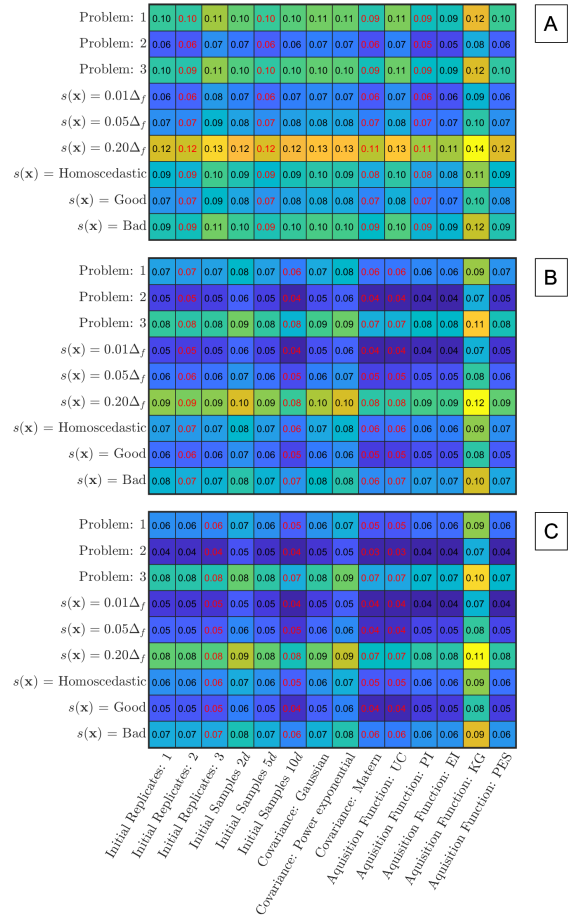


Figure 4: The GAP averaged (numbers in each block) over the simulations for all combinations of controllable (horizontal axis) and noise factors (vertical axis), where the red numbers are the optimal level for each controllable factor. A) interaction effects for 25d observed samples, B) interaction effects for 37.5d observed samples, and C) interaction effects for 50d observed samples,

experimental budget. This suggests, that most acquisition functions place too much emphasis on exploitation causing the sampling path to spend too many resources exploring the trivial region of the space of admissible input conditions. This is caused by the inflated uncertainty of the posterior predictive response as a consequence of the training data uncertainty.

3.3. Recommendations

From the previous results, we can make the following recommendations.

1. If your experimental budget is in the range of

15d to 40d, then it is recommended to replicate the initial number of samples twice. In addition, for a larger experimental budget, more replicates might be required. Conversely, for an experimental budget of less than 15d, it could be better to have no replicates. Although, this cannot be confirmed definitively from the performed study.

2. For larger experimental budgets, it is recommended to start with a larger number of initial samples. Specifically, for less than 30d it is recommended to have an initial sample size of 5d, whereas for larger experimental budgets 10d or more is recommended.
 3. For an experimental budget of less than 20d it is recommended to use the PI covariance function. For larger experimental budgets, it is best to use the UC interval with a relatively large value for π (e.g., $\pi > 5$).
 4. The Matér covariance has a superior performance on all test problems compared to other covariance functions.
 5. If there is no explicit limitation on the experimental budget, then it is advised to use two initial replicates for all 10d initial sampling locations and use the Matér covariance in combination with either the PI or EI acquisition function.
- #### 4. DISCUSSION

In this section, we will briefly discuss some of the considerations that went into the performed study and the aforementioned observations.

1. In this study we have assumed a functional form for the experimental uncertainty, either the variance is homoscedastic, or it is proportional to the functional response. While in practice, more functional forms of the variance could exist, the set studied in the paper is significantly broad. For example, experimental uncertainty, a form of intrinsic uncertainty, is often constant or proportional to the response surface.

2. We focused our study on normally distributed experimental uncertainty. Consequently, a designer should be careful of this when adopting the conclusion presented in this study when the experimental noise might be normally distributed.
3. We did not study the acquisition of batches of samples [2] and this could be important to the overall performance of the data acquisition process. While this is certainly a limitation of the presented study, it should be noted that our aim was to analyze low experimental setups with small experimental budgets. Consequently, sampling batches in this context is not practical.
4. One assumption in the presented approach is that we have used the maximum likelihood approach to approximate the GP hyperparameters. As a consequence, the uncertainty in the hyperparameters is not considered when evaluating the posterior predictive distribution (this can be observed from the true function poorly fitting the 95% confidence intervals in Figure 2). In addition, this also explains why most acquisition functions place more emphasis on exploitation. Consequently, it could be the case that using a Bayesian approach in the approximation of GP hyperparameters would be a more appropriate approach for the optimization of the stochastic response surface, especially with small experimental budgets.

5. CONCLUSIONS

In this paper, we have presented an experimental study in the modeling conditions associated with Bayesian optimization for noisy training data (e.g., physical experiments). In this study, we emphasized the scenario where only a small experimental budget is available, in which case stringent simplifying assumptions need to be made. Moreover, with this study, we aimed to identify what modeling conditions are most appropriate based on a priori knowledge on the nature of the problem (e.g., the magnitude of the experimental uncertainty, the functional form of the experimental uncertainty,

and the form of the objective function). From this study, we have discovered that the Matér covariance function performed better than all studied alternatives (i.e., Gaussian, and power exponential). In addition, there is a strong correlation between the experimental budgets and many of the controllable modeling considerations. Specifically, the experimental budget has an interaction effect with the initial number of replicates, the initial number of samples, and the choice of acquisition function. Alternatively, the form of the objective function, the magnitude of the experimental uncertainty, and the form of the experimental uncertainty do have no interaction effects with the controllable modeling decisions.

The work in this paper provides inspiration for future research directions. Specifically, it was found that conventional acquisition functions place too much emphasis on exploitation. This could be remedied by a Bayesian approach that accounts for the uncertainty in the hyperparameters of the emulator or through the development of problem-specific acquisition functions. More importantly, while many advanced methods for experimental design are available, physical experiments are often performed on input conditions that are selected based on the intuition of the designer. To make this process more systematic we explored the potential of adaptive sampling strategies, typically used in conjunction with simulation-based design, for physical experiments. However, as this involved simplifying assumptions, future research should be directed toward the discovery of new statistical methods for the optimization of noisy response surfaces in sparse data scenarios.

ACKNOWLEDGEMENTS

The financial support from the School of Mechanical and Materials Engineering at University College Dublin is greatly appreciated.

REFERENCES

- [1] Alexander I.J. Forrester and Andy J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1):50–79, 2009.
- [2] Anton van Beek, Umar Farooq Ghumman, Joydeep Munshi, Siyu Tao, TeYu Chien, Ganesh Balasubramanian, Matthew Plumlee, Daniel Apley, and Wei Chen. Scalable adaptive batch sampling in simulation-based design with heteroscedastic noise. *Journal of Mechanical Design*, 143(3), 2021.
- [3] Thomas J Santner, Brian J Williams, William I Notz, and Brian J Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.
- [4] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.
- [5] Rahul Mukerjee and Chien-Fu Wu. *A modern theory of factorial design*. Springer, 2006.
- [6] Ruichen Jin, Wei Chen, and Agus Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1):268–287, 2005.
- [7] I.M. Sobol. Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics*, 16(5):236–242, May 1976.
- [8] AARA Igder, Ali Akbar Rahmani, Ali Fazlavi, Mohammad Hossein Ahmadi, Mohammad Hossein Ahmadi Azqhandi, and Mohammad Hassan Omidi. Box-behnken design of experiments investigation for adsorption of Cd^{2+} onto carboxymethyl chitosan magnetic nanoparticles. *Journal of Mining and Environment*, 3(1):51–59, 2012.
- [9] Benjamin Durakovic and Muris Torlak. Experimental and numerical study of a pcm window model as a thermal energy storage unit. *International Journal of Low-Carbon Technologies*, 12(3):272–280, 2017.
- [10] Everet Franklin Lindquist. *Design and analysis of experiments in psychology and education*. Houghton Mifflin, 1953.
- [11] Ronald A Fisher. The arrangement of field experiments. In *Breakthroughs in statistics*, pages 82–91. Springer, 1992.

- [12] Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.
- [13] Bruce Ankenman, Barry L Nelson, and Jeremy Staum. Stochastic kriging for simulation meta-modeling. In *2008 Winter Simulation Conference*, pages 362–370. IEEE, 2008.
- [14] Matthew Plumlee and Rui Tuo. Building accurate emulators for stochastic simulations via quantile kriging. *Technometrics*, 56(4):466–473, 2014.
- [15] Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in neural information processing systems*, 29, 2016.
- [16] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
- [17] George De Ath, Richard M Everson, and Jonathan E Fieldsend. How bayesian should bayesian optimisation be? In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1860–1869, 2021.
- [18] Jay D Martin and Timothy W Simpson. Use of kriging models to approximate deterministic computer models. *AIAA journal*, 43(4):853–863, 2005.
- [19] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.
- [20] Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *International Conference on Machine Learning*, 21(4):345–383, 2001.
- [21] Aimo Törn and Antanas Zilinskas. *Global optimization*, volume 350. Springer, 1989.
- [22] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466, 2006.
- [23] Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.
- [24] Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.
- [25] Siyu Tao, Anton Van Beek, Daniel W Apley, and Wei Chen. Multi-model bayesian optimization for simulation-based design. *Journal of Mechanical Design*, 143(11), 2021.
- [26] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.
- [27] Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001.
- [28] Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and multi-disciplinary optimization*, 48:607–626, 2013.