TECHNICAL REPORT NO. 932
August 1994

THE RESPONSE SURFACE BANDIT

by

Josep Ginebra
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706
ginebra@tekamthi.stat.wisc.edu

Murray K. Clayton
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706
clayton@tekamthi.stat.wisc.edu
Response Surface Bandits

by

Josep Ginebra and Murray K. Clayton\textsuperscript{1}
Department of Statistics
University of Wisconsin-Madison

ABSTRACT

In this paper we define a response surface bandit as the sequential design problem that maximises an expected bandit utility but where the outcomes $y_n$ are continuous and can be related through a response surface to a set of controllable variables $x_n = (x_{1n}, x_{2n}, \ldots, x_{kn})$. We link this problem to other traditional optimisation problems from industrial engineering and to the traditional bandit problem.

We consider two approaches to the problem. The first is based on a myopic sequential design. The second approach uses the best design out of a family of designs related to upper bounds for the predicted surface; the family includes myopic and sequential versions of D-optimal designs. These approaches can be generalised to more broadly defined sequential problems.

KEY WORDS: Bandit problem, response surface methodology, two-stage problems, myopic, learning component, sequential optimisation.

\textsuperscript{1}Address for Correspondence: Department of Statistics, 1210 W. Dayton Street, Madison, Wisconsin 53706, USA
1 Introduction

The usual bandit problem as formulated in Berry and Fristedt (1985) and Gittins (1989) looks for the strategy or sequential design \( \tau \) that maximises the total expected sum of outcomes coming from two (or more) stochastic processes. That is, it maximises the worth, \( W_\tau = E_\tau(\sum_{n=1}^{\infty} \alpha_n y_n) \), where \( y_n \) is the outcome of the process selected at the \( n \)th stage.

In this paper we extend this problem to the case where the outcomes \( y_n \) can be related by some function to a set of controllable variables \( x_n \). We are interested in applications where the \( y_n \) can be observed only with noise and where the function modeling the expectation of \( y_n \) conditional on \( x_n \) will depend on unknown parameters \( \beta \) and will likely be empirical in nature.

1.1 Definition of the Response Surface Bandit

Assume that, at the \( n \)th stage, we can control a \( k \)-dimensional input \( x_n = (x_{1n}, x_{2n}, \ldots, x_{kn}) \) and we can observe an output \( y_n \) such that:

\[
y_n(x_n) = f(x_n; \beta) + e_n
\]

for \( n = 1, 2, \ldots \). We will assume that the \( e_n \) are independent, normally distributed with \( E(e_n) = 0 \) and \( \text{var}(e_n) = \sigma^2 \). In addition, assume that \( \beta \) represents a vector of parameters that governs the relationship between \( y_n \) and \( x_n \) for \( x_n \in C \subset \mathbb{R}^k \). Assume that a prior distribution \( G_0(\cdot) \) exists for the unknown \( \beta \). Finally, assume that, for each \( \beta \), \( f(x; \beta) \) has a unique maximum at \( x^* = (x_{1\beta}^*, x_{2\beta}^*, \ldots, x_{k\beta}^*) \) in \( C \), where \( C \) is a compact set.

Of particular interest is the linear model where \( f(x_n; \beta) = \beta_0 + \beta_1 x_{1n} + \cdots + \beta_k x_{kn} \). This includes quadratic response surfaces as a special case. Let \( X_n \) be the \( n \times (k+1) \) matrix whose \( j \)th row is the adjoining vector \( z_j = (1 | x_j) = (1, x_{1j}, x_{2j}, \ldots, x_{kj}) \) and let \( Y_n = (y_1, y_2, \ldots, y_n) \). Assuming that \( \sigma^2 \) is known and that the prior distribution at stage \( n = 0 \) is \( G_0(\beta | \sigma) \sim N(\mu, \sigma^2 M^{-1}) \), then the posterior distribution after stage \( n \), \( G_n(\beta | Y_n, X_n, G_0, \sigma) \) is (De Groot, 1970):

\[
\beta | Y_n, X_n, G_0, \sigma \sim N\{(M + X_n'X_n)^{-1}(M\mu' + X_n'Y_n), \sigma^2(M + X_n'X_n)^{-1}\}
\]

Of course, the usual Bayesian formulation dictates that this will also serve as the prior for stage \( n + 1 \).

We consider sequential designs for this problem: a sequential design is a rule that specifies, for any history of observations \( Y_{n-1} \) and \( X_{n-1} \), how to set the inputs for the \( n \)th stage — i.e. how to choose \( x_n = (x_{1n}, x_{2n}, \ldots, x_{kn}) \) \( C \subset \mathbb{R}^k \). We denote such a design by \( d \). As in the usual bandit paradigm we define the worth of a bandit to be:

\[
W(d, f, G_0, A, \sigma) = E_d[\sum_{n=1}^{\infty} \alpha_n g(y_n(x_n))].
\]

Here, \( A = (\alpha_1, \alpha_2, \alpha_3, \ldots) \) is a discount sequence of non-negative numbers with finite sum, \( N = \inf\{n : \alpha_m = 0 \text{ for } m > n\} \) is its horizon and \( g(y) \) is a continuous, monotone function of \( y \).
An optimal design is one that maximises the worth over the space \( \Delta \) of all possible sequential designs \( d \). We call the optimal worth the ‘value’ of the response surface bandit, and denote it by \( V(f, G_0, A, \sigma) = \sup_d W(d, f, G_0, A, \sigma) \). Any sequential designs that attain this supremum (if such designs exist) will be called optimal bandit designs. Of interest in this paper is the behaviour of strategies when the total number of observations is small (as opposed to a study of the asymptotic properties of these designs).

In the usual bandit terminology, an ‘arm’ is a process that can be sampled and yields, at the \( n \)th stage, an observation \( y_n \). In that sense, the number of arms \( x_i \) in the response surface bandit is the number of \( x \)'s in \( C \). If \( C \) has infinitely many points, then we are faced with an infinite-arm normal bandit problem. In this setting the distribution of the response for different arms will be highly dependent and so the outcome of pulling one of them has the potential to provide information about all the others. This stands in contrast to the independent arm bandit that is more commonly studied (Berry and Fristedt, 1985 and Gittins, 1989).

We note that when the \( x_n \)'s are random and thus uncontrollable, as is often the case in the clinical trials setting, we get extensions of the one-armed covariate bandit problem of Sarkar (1991) and Clayton (1989). We will restrict ourselves here to the assumption that the \( x_n \)'s are controllable and that our goal is to decide sequentially on their values so that we maximise our expected utility.

In practice, it is impossible to explicitly determine the optimal design for the response surface bandit. Ideally, for finite horizon problems, backward induction could be used to identify optimal designs (see De Groot, 1970). However, this requires a complicated \( N \)-dimensional integration with numerous function maximisations. The alternatives that we propose consist of designs that can be implemented in a forward direction and are best in a subclass of sequential designs \( \Delta' \subset \Delta \). They are similar to designs shown in Ginebra (1993) to have good properties for the Bernoulli case.

### 1.2 Some Response Surface Bandit Utilities

There are a few discount sequences of particular interest for the response surface problem. (A general discussion of discount sequences can be found in Berry and Fristedt, 1985.) For the finite horizon uniform discount sequence (Berry and Fristedt, 1985, p.150) the first \( N \) elements of \( A \) are 1 and the remainder are 0. Because of the overall goal associated with such a sequence, we refer to the corresponding bandit as a ‘transformed the larger the better’ response surface bandit — the problem requires the maximisation of the expected sum of \( N \) transformed responses \( g(y) \), i.e. \( W(d, f, G_0, \sigma) = E_d[\sum_{n=1}^{N} g(\{y_n(x_n)\})] \).

A second case of interest is the ‘two-stage’ response surface bandit where we want to maximise \( W(d, f, G_0, \sigma) = E_d(\sum_{n=1}^{N} y_n + S y_{N+1}) \), for some \( S > 0 \). This corresponds to the discount sequence where the first \( N \) elements of \( A \) are 1, the \( N + 1 \)st element is \( S \), and the remainder are 0; thus, after \( N \) stages, we have to choose a fixed \( x_{N+1} \) to be used for the next \( S \) stages. This utility structure can be used to represent many industrial statistics problems where we are willing to ‘fiddle with the knobs’ during the first phase \( (n = 1, 2, \ldots, N) \) and
then keep the knobs fixed for the production of the remaining $S$ products. One reference discussing this discount sequence for the Bernoulli bandit is Clayton and Witmer (1988).

When $S$ becomes much larger than $N$, the second stage observations become increasingly important. A discount sequence for which only second stage observations are important in terms of utility is the sequence $A = (0, 0, \ldots, 0, 1, 0, \ldots)$ with 0's everywhere except at position $n = N + 1$. We call this third problem, `the larger the endpoint the better' response surface bandit. In this case we want to end up with the $x_{N+1}$'s maximising $W(d, f, G_0, \sigma) = E_d\{y_{N+1}(x_{N+1})\}$.

Related to these maximisation problems is that where the goal is to sequentially choose the settings for $x_n$, $n = 1, 2, \ldots, N$ such that we minimise $E_d(\sum_{n=1}^N(y_n(x_n) - T)^2)$ for a known target value $T$. We call this the `target bandit problem'. This problem shows up in many areas of engineering where we want to keep a noisy performance characteristic $y$ close to a target value $T$ for every observation $y_n$. The target bandit can be viewed in the same maximisation framework of this paper if we define $g(y) = -(y - T)^2$. Kackar (1985) discusses how this kind of problem can be related to the quality of an industrial product.

The tools we will discuss herein could be applied to a very broad class of utility functions to help find sequential designs $d$ maximising:

$$W(d, f, G_0, A, \sigma) = E_d\{U(y_1, y_2, \ldots, y_N, \beta; A)\}$$

where $U$ is an arbitrary utility function of the $N$ observables $y_n$ and the parameters $\beta$ and where $A$ is a set of known constants. For example, such utility functions might model the trade-off between utility of observations $y_n$ and good estimation of the parameters $\beta$. We shall not explore such utilities in this paper.

### 1.3 Connection to Other Optimisation Problems

In the general experimental design setting, it is important to identify the reason for estimating $f(x; \beta)$. In the theory of optimal designs, the goal is to learn about $f$ as efficiently as possible. If we assume that the model is linear in the parameters, this translates into a range of optimality criteria based on functions of $X'X$, the covariance matrix of the estimated parameters in the model; see, for example, Fedorov (1972) and Silvey (1980). In response surface methodology, the goals are to approximately map the surface $f(x_n, \beta)$ within a limited region of the $X$'s in order to choose suitable operating conditions. References on this topic are Box and Draper (1969, 1987) and Mead and Pike (1975).

In a different setting, the stochastic approximation literature stemming from Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), looks at the problem of estimating the roots or optima of functions that can be observed only with noise and that might have unknown functional form. Of typical interest in this literature is the asymptotic convergence of a sequence of random variables $x_n$ in the space of input variables.

All these approaches to function optimisation with noise require a trade-off between learning about the surface, on one hand, and running as many experiments as possible close to the maximum, on the other. (In a stochastic approximation problem this trade-off is
driven by the desire to have a good rate of convergence.) When fitting low-degree response surfaces or estimating the effects of different factors on the responses, the theory of optimally designed experiments typically suggests the use of designs balanced and spread over the X-space, thus providing information about the curvatures of the surface of interest. In this case the designs are driven by the quest for knowledge about the parameters and not payoff in terms of observable responses.

On the other hand, the response surface bandit problem involves learning about the response surface and its model only in order to sequentially move the design towards the highest yielding set of X-conditions. Thus, this type of adaptive design is driven by payoff in terms of observed responses, not knowledge about model parameters. Consequently, the response surface bandit problem involves a loss structure that is opposite, in some sense, to that of the typical theory of optimally designed experiments. Methods such as steepest ascent, EVOP, simplex methods for process improvement and stochastic approximations are methods that might be viewed as lying somewhere in between these two extremes.

Also, Box (1966) and Box and Draper (1987) distinguish between two kinds of feedback we should try to get from EVOP techniques and, more generally, from any kind of industrial experimentation. They call empirical feedback the type where a particular response pattern leads more or less automatically to a particular action, like shifting to the X-conditions providing higher yields. They call scientific feedback that which comes from the interaction of the results with prior technical knowledge to produce actions which could not be taken on a purely automatic basis, like deciding that we have to change the model f for the y’s. In a certain sense, the bandit problem models the case where only empirical feedback matters, and it matters in a very narrow way.

2 Myopic Designs

Myopic designs choose at each step \( n \), the point \( x_n = (x_{1n}, x_{2n}, \ldots, x_{kn}) \in R^k \) that provides the highest expected payoff for the next run. In general, for any given \( f(x; \beta) \) this requires finding the maximum of the predicted mean value of the surface after stage \( n - 1 \), i.e. the \( x_n = x_n^* \) maximising (in \( x \)):

\[
\hat{y}_n(x) = E(y_n(x)|Y_{n-1}, X_{n-1}, G_0, \sigma).
\]

For the linear model this becomes the problem of finding the \( x_n^* \) that maximises (over \( x \)):

\[
z(M + X_{n-1}^\top X_{n-1})^{-1}(M\mu + X_{n-1}^\top Y_{n-1}^\top) \quad \text{where} \quad z = (1|x) = (1, x_1, x_2, \ldots, x_k). \]

Note that this might involve choosing \( x_n = x_n^* \) on the boundary of \( C \). For more general models, a numerical optimisation algorithm might be needed to find \( x_n^* \).

Myopic designs behave optimally for the case where there is only one observation left. However, as observed above, the arms of the response surface bandit are highly dependent, and thus the myopic design will provide information about the whole surface, not just the point where the run was made. From this perspective, the myopic design is more informed than it first appears, and thus has the potential for coming closer to the optimal strategy than might be expected. See Ginebra (1993) for further discussion of this issue.
3 Upper Bound Designs

3.1 The Family of Upper Bound Designs

The upper bound designs are constructed by noting that, as an approximation, the estimated fitted surface at stage $n$ as defined by $\hat{y}_n(x)$ is distributed like $N[\hat{y}_n(x), \text{var}\{\hat{y}_n(x)\}]$. This makes it natural to define an upper bound of the predicted mean value of the surface at $x = (x_1, x_2, \ldots, x_k)$ after observing $y_{n-1}$ as:

$$\text{ub}_n(x) = \hat{y}_n(x) + K \text{sd}\{\hat{y}_n(x)\}$$

where sd denotes standard deviation and where $K$ is some positive constant. For the linear model case with $z = (1|x) = (1, x_1, x_2, \ldots, x_k)$ this is:

$$\text{ub}_n(x) = z(M + X'_{n-1}X_{n-1})^{-1}(M\mu' + X'_{n-1}Y_{n-1}) + K\sigma z(M + X'_{n-1}X_{n-1})^{-1}z'.$$

In an upper bound design, at the $n$th stage we observe $y$ at the point $x_n$ that maximises the upper bound function $\text{ub}_n(x)$. Note that if $K = 0$, this reduces to the myopic design. For general $K$, this design tends to favour experimenting at points in $C$ where the information we have about the surface is the smallest and, in particular, given two points with the same predicted response, we favour experimenting at the point $x$ with the largest prediction error. This involves sacrificing some possible immediate payoff for the sake of improving our chances of doing better in later runs. A similar idea for independent arm bandits has been discussed by Lai (1987) and Ginebra (1993). Lai (1987) shows that, in certain situations, such an approach can be asymptotically optimal.

The constant $K$ regulates the relative weight of the ‘learning component’ (see Gittins, 1989) in the design with respect to the myopic immediate expected payoff. By varying this constant $K$ from zero to infinity we get a large family of designs. When $K$ grows towards infinity we get rules that experiment at the points in the $X$-space where the variance of the predicted value is largest, thus taking observations at those places where the amount of information about the response $y$ is small.

Silvey (1980) and Box and Draper (1987) describe a host of nonsequential optimality design criteria based on $\text{var}\{\hat{y}(x)\} = \sigma^2 z(X'_{n}X_{n})^{-1}z'$ and its inverse, the information function. They are the D, G, E, A and c-optimal designs. These criteria are optimal when the goal is to design the experiment so that the estimated model based on the $N$ runs available has the best predictive properties over the region of interest $C$. For the linear model and these criteria that only involve the variance of the predicted response, there is little advantage in using sequential designs since the variance depends only on the $x$’s and not on the response $y$’s. But in our upper bound indices we combine the variance of the predicted response with the predicted response itself; that is a function of the observations $y$, and thus it now becomes appropriate to proceed sequentially.
3.2 Selection of the Best Upper Bound Design

If the upper bound design is to be used, there remains the question of how \( K \) should be chosen. We propose choosing the ‘best’ \( K \), which we call \( K^* = K^*(f, G_0, A, \sigma) \), as follows. For a given utility, model, \( \sigma \), discount sequence and prior, we arbitrarily pick one \( K \) and simulate the use of the upper bound design, estimating its worth based on that choice of \( K \). Then, by searching among various values of \( K \), we find that choice \( K^* \) that is estimated to be the best. By implementing the sequential upper bound design with the best \( K^* \) we have the guarantee that we will do at least as well as the myopic design because it is a member of the family of upper bound strategies. (Indeed, when \( K^* \) is close to zero, myopic will be best.) In this way, we are able to tailor the design criterion to work best for the utility of the problem at hand. This best \( K \), \((K^*)\), is a measure of the amount of learning required by the problem.

These ideas are illustrated in Figures 1 and 2. In the examples, \( f(x; \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 \), \( \sigma = 1 \), and \( C = [-1, 1] \) and the prior \( G_0(\beta_0, \beta_1, \beta_2) \) is multivariate normal distribution with mean \( \mu = (1, 0, -1) \) and covariance matrix \( I \). That is, we are guessing that the expected surface is close to \( y = 1 - x^2 \). The upper bound designs will choose the \( x_n \in [-1, 1] \) that maximises in \( x \):

\[
\text{ub}_n(x) = \hat{\beta}_0_n + \hat{\beta}_1_n x + \hat{\beta}_2_n x^2 + K s_n \sqrt{z(X'_{n-1}X_{n-1})^{-1}z'}
\]

at each step \( n \) with \( z = (1, x, x^2) \). In practice we will not know \( \sigma \); in this application we use \( s_n \) which is the residual standard deviation from fitting the model after observing \( y_{n-1} \). Note also that our estimate of the upper bound surface ignores the specific prior used, and thus it may be viewed as representing a nonBayesian (but perhaps more typical) approach.

With these choices for the model, we investigated the worths as a function of \( K \) for six different utilities. Figure 1 shows the worth for ‘the larger the better’ utility and for the ‘two-stage’ utility. In each case we use \( N = 20 \); for the two-stage utility we consider both \( S = 20 \) and \( S = 200 \). Figure 2 shows the case where \( N = 20 \) and we want to maximise ‘the larger the endpoint the better’ utility and the expected sum of \( y_n^2 \) and \( \log[y_n] \). For each \( K \) the upper bound design is applied and its worth estimated; for the ‘larger the end-point the better’ we estimate \( E(|x_{N+1} - x_{opt}|) \) and seek its minimum. This is essentially equivalent to maximising the worth \( E(y_{N+1}) \). A total of 200,000 simulations were used, sampling first from the prior on the parameters and then using the parameters obtained to generate one response from the corresponding model. The large number of simulations were used to reduce the estimated standard error of estimates. An example that is typical is the case where \( K = 1 \); the estimated worth for ‘the larger the better’ utility in Figure 1 is 22.7152 with an estimated standard error of 0.0458.

From Figure 1 we see that for two-stage bandit problems, this optimal \( K^* \) increases slightly with increasing \( S \). The best \( K^* \) for ‘the larger the endpoint the better’ utility is larger than for these two-stage utilities, as would be expected; the amount of immediate payoff we are willing to sacrifice for the sake of learning how to obtain a larger \( y_{N+1} \) will increase as \( S \) increases.
What is striking from both Figure 1 and 2 is that the best choice of $K$ seems to be about 1, regardless of the specific utility used, i.e. regardless of the choice of discount sequence and transformation $g$. As mentioned, this points first of all to a sequential design that is superior to the myopic. Second, it suggests that the implementation of a nonsequential design based wholly on the variance of prediction will be inadequate. The patterns of dependence of the worth as a function of $K$ are similar to those for “the larger the better” Bernoulli bandit with independent arms as shown in Ginebra (1993). In the same reference it is also shown what happens for different priors on the parameters for this quadratic model. It seems that the higher the probability that there is a maximum centered in $C$, the larger the improvement on the myopic designs through these upper bound designs.

3.3 Adaptive Upper Bound Designs with $K^*_n$

The upper bound designs as described above should perform well, but will not in general be optimal. One of the reasons for this is that we compute $K^*$ at the beginning and keep it constant for the rest of the design. To address this, recall that the prior distribution available at stage $n$ is $G_{n-1}(\beta|Y_{n-1}, X_{n-1}, G_0, \sigma)$. We would propose, at each stage $n$, recomputing the optimal $K^* = K^*_n$ that is best for this posterior. Then, we would run the $n$th experiment at the $x_n \in C$ that maximises in $x$ the upper bound surface:

$$ ub_n(x) = \hat{y}_n(x) + K^*_n sd\{\hat{y}_n(x)\} $$

with $K^*_n$ changing with $n$. In this way we tailor the design criteria not only to be best for our utility but to adapt itself to the observed responses $y_n$ at the same time. We use the outcomes of the first $n$ stages to sharpen not only the estimate of the predicted surface but also to simultaneously adapt the index we use to drive our design. The design criterion itself would be driven sequentially by the outcomes obtained up to that stage.

An alternative way of implementing designs that look into the future would be to use the equivalent of the Gittins indexed strategies in this setting. (See Gittins, 1979 and Berry and Fristedt, 1985). In Ginebra (1993) it is argued that the break-even one-armed bandit indices used there can be thought of as upper bounds of $E(y|\beta)$ that adapt to the future discount sequence, and thus, using this type of adaptive upper bound strategies should give similar results to the use of strategies based on break-even values, which are much harder to compute. It is our conjecture that these designs that chose $K^*$ adaptively as well as designs that use truncated discount sequences for computing $K^*$ could be robust with respect to changes of the function $f$ with time. Work of Ginebra (1993) on the Bernoulli bandit suggests that such truncation leads to more robust strategies and in addition it makes them more easily computed.

4 Comparison of Myopic and Upper Bound Designs

In this section we compare the myopic and the upper bound family of designs with constant $K^*$. We consider four different functional forms relating $E(y|\beta)$ to $x$; these are summarised in
The larger the better bandit with $N = 20$.

Two stage bandit with $N = 20$ and $S = 20$.

Two stage bandit with $N = 20$ and $S = 200$.

Figure 1: Worth as a function of the learning coefficient $K$ of the upper bound design. We assume $y(x)|\beta \sim N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2 = 1), \beta \sim N\{(1, 0, -1), I\}$ and $x \in [-1, 1]$ with $N = 20$. 

Transformed the larger the better bandit; $N = 20$ and $g(y) = \log|y|$.  

Transformed the larger the better bandit; $N = 20$ and $g(y) = y^2$.  

The larger the endpoint the better bandit with $N = 20$.  

Figure 2: Worth as a function of the learning coefficient $K$ of the upper bound design. We assume $y(x)|\beta \sim N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2 = 1)$, $\beta \sim N\{(1, 0, -1), I\}$ and $x \in [-1, 1]$ with $N = 20$. 


<table>
<thead>
<tr>
<th>Case</th>
<th>$E(y)$</th>
<th>$x_{\text{max}}$</th>
<th>$y_{\text{max}}$</th>
<th>$x_{\text{min}}$</th>
<th>$y_{\text{min}}$</th>
<th>$W_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1 - 2x^2$</td>
<td>0</td>
<td>1</td>
<td>1, -1</td>
<td>-1</td>
<td>20/3</td>
</tr>
<tr>
<td>2</td>
<td>$1 - 4x^2$</td>
<td>0</td>
<td>1</td>
<td>1, -1</td>
<td>-3</td>
<td>-20/3</td>
</tr>
<tr>
<td>3</td>
<td>$7/16 + 3/2x - x^2$</td>
<td>.75</td>
<td>1</td>
<td>-1</td>
<td>-33/16</td>
<td>25/12</td>
</tr>
<tr>
<td>4</td>
<td>$3/4 + 1/2x - 1/4x^2$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>40/3</td>
</tr>
</tbody>
</table>

Table 1: Test cases used to compare the performance of the myopic, upper bound and stochastic approximation designs. ($x_{\text{max}}, x_{\text{min}}$) are the locations of the expected maximum and minimum, ($y_{\text{max}}, y_{\text{min}}$) the magnitude of those extreme expected values and $W_u$ is the worth for the nonsequential design that experiments uniformly at random in $[-1, 1]$.

Table 1. We will only deal with 'the larger the better' response surface bandit with $N = 20$. The models that we consider are of the form:

$$y(x) \sim N(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma^2).$$

The value of $\beta = (\beta_0, \beta_1, \beta_2)$ is fixed throughout a given simulation, but the sequential designs are constructed, in keeping with the general approach of this paper, as if $\beta$ were unknown. We take $\beta_2 < 0$ so that there is one absolute maximum and we will assume that $x \in [-1, 1]$. Table 1 gives the functional form of the expectation of $y$; all four of these cases have $y_{\text{max}} = 1$ and so the maximum possible worth will always be $N$. Most of the examples are likely sample realisations of the prior distributions for the parameters we used in the preceding section but case 2 where $\beta_2 = -4$ is three standard deviations away from $-1$, the mean of $\beta_2$.

To make the model estimable, we start the designs with $x_1 = -1, x_2 = 0, x_3 = 1$. This is equivalent to assuming a noninformative prior on the model. Since all four models have their minimum values at either $x = -1$ or both $x = 1$ and $x = -1$, by taking initial values there, we lose a lot of expected payoff in the first three observations; this will significantly affect the limit of the worth as $\sigma$ goes to 0. We saw in Figure 1 that for $\sigma = 1$, the optimum choice of $K$ is close to $K = 1$. We will use $K = 1$ and $K = 2$ fixed for all the stages of the design.

Figures 3 and 4 present the expected sum for these three designs as a function of $\sigma$ for our four test cases. Each worth has been computed by averaging 20,000 replications. The horizontal line corresponds to the worth we would get using the non-sequential uniform design. The fact that for all the four cases the worth falls below the nonsequential uniform worth when $\sigma$ is large is a bit deceptive. The maximum for $E(y)$ is 1, so when $\sigma/E(y)$ is as large as 10, it will be practically impossible to obtain any good estimate of the location of the maximum with $N = 20$ runs; the maximum of the predicted surface and its upper bound will most likely fall outside $[-1, 1]$, thus pushing the myopic and upper bound designs to experiment at the most unfavourable $x$'s of $C = [-1, 1]$. This problem might be less important if we had used the best $K^*$ for each given $\sigma$.

Figure 3 shows situations where the real maximum is centered at $x = 0$. In both cases we see that the upper bound designs perform considerably better than the myopic one for all the
$\sigma$'s. The flatter $E\{y(x)\}$ is as a function of $x$, the faster the worth of the designs decreases when $\sigma$ increases. This occurs because as $\sigma$ increases, it gets more difficult to locate the real maximum. It also happens that the flatter is $E\{y(x)\}$, the less we lose in absolute terms since the difference between the expected rewards for the best $x$'s and the worst $x$'s gets smaller. Upper bound designs with $K = 1, 2$ seem to do better than the myopic designs for these two expectations.

Figure 4 presents two cases where $E\{y(x)\}$ has its maximum at $x = 0.75$ and $x = 1.0$ respectively. These values are quite different from the optimal location suggested by the prior: as noted, the prior distribution on the model that we used to select $K = 1$ had $x = 0$ as the most likely position for the maximum of $E\{y(x)\}$. This figure shows how the myopic design outperforms the upper bound design for small $\sigma$, while the upper bound design does better for large $\sigma$.

In all the examples considered here, the upper bound design with $K = 1$ does slightly better than that design with $K = 2$ for small and medium $\sigma$, and does a bit worse for very large $\sigma$. The fact that the behaviour for the expected worth looks so similar for $K = 1$ and $K = 2$ indicates that those designs could be rather robust against the wrong choice of $K^*$.

One quantity that is especially meaningful is the smallest standard deviation $\sigma$ that makes these designs worse than the nonsequential design that experiments uniformly at random on the space of the $x$'s. For case 1 of Table 1, Figure 3 shows that this happens around $\sigma = 2$ with myopic designs while it happens close to $\sigma = 3.5$ for the upper bound design. For case 3 in Figure 4, the myopic and the upper bound designs improve upon the nonsequential uniform design for any $\sigma < 8.5$. Since those threshold $\sigma$s are larger than $y_{max}$, the new sequential designs we propose are doing much better than the nonsequential uniform designs for the cases that might typically arise in practice.

5 Conclusion

In this paper we have introduced a broad family of bandit problems where the expectation of the observed response $y_n$ can be modelled through a function of a set of controllable input variables $x_n$. The utility functions that we associate with these problems include the typical bandit-type utility, but can be extended more broadly. While the response surface literature has always emphasised the sequential nature of its methodology, this paper proposes a more direct approach to the use of sequential methods for response surface problems.

In considering the transformed 'larger the better' utilities, we constructed the strategies using untransformed data, even though the utility was in terms of the transformed values $g(y)$. Doing so is convenient, but raises the issue of whether it would be more effective to construct strategies directly using the transformed observations.

In general, any family of sequential optimisation methods suitable for treating data with noise could be a candidate family of designs $\Delta'$ from which we select the best design given the utility at hand. For example Ginebra (1993) shows how stochastic approximation sequences (Kiefer and Wolfowitz, 1952) can be adapted to this response surface bandit problem. Also,
Figure 3: Worths for the myopic and upper bound sequential designs with $K = 1, 2$, $x_1 = -1, x_2 = 0, x_3 = 1$ and $x \in [-1, 1]$, for 'the larger the better' bandit with $N = 20$ and with $y|x \sim N(b_0 + b_1 x + b_2 x^2, \sigma^2)$. 
Figure 4: Worths for the myopic and upper bound sequential designs with $K = 1, 2$, $x_1 = -1, x_2 = 0, x_3 = 1$ and $x \in [-1, 1]$, for ‘the larger the better’ bandit with $N = 20$ and with $y|x \sim N(b_0 + b_1 x + b_2 x^2, \sigma^2)$. 
finding the best step sizes for the steepest ascent and simplex methods through simulation for the utility at hand and the assumed model might work as well.

Our designs can be applied quite broadly, but some problems will allow for sequential designs that are tailored to the problem. That is the case presented in Ruppert et al. (1984). Their problem consists of finding optimal values for two parameters in a proposed class of Menhaden harvesting policies. The scientists provide a class of good strategies that fit their problem at hand, and they find the constants that maximise their worth defined as the expected sum of the yearly harvests for twenty five years. The purpose of their paper is to illustrate how to use stochastic approximation to optimise the worth of their family of strategies for a bandit problem, although they do not describe it as such.

One setting where the response surface bandits could be useful is with processes that have strong ‘day to day’ nonstationarities present but are rather stable during a given day, and where the number of runs per day is not too large. They could also be relevant in reliability where we want to maximise the aggregate life-lengths $y_n$ depending on controllable environmental conditions $x$ following an exponential regression model, or in more general problems where the relationship between the $y_n$’s and $x_n$’s could be modelled according to more general linear, non-linear and non-parametric models.

6 Acknowledgements

We are very grateful to Professors G.E.P. Box, S. Bisgaard, C. Ritter, D. Steinberg and one referee for helpful feedback on these subjects.

7 Bibliography


