In *Optimum Design 2000*, A. Atkinson, B. Bogacka, and A. Zhigljavsky, eds., Kluwer, 2001, pp. 195–208.

Chapter 1

OPTIMIZING A UNIMODAL RESPONSE FUNCTION FOR BINARY VARIABLES

Janis Hardwick
Purdue University
West Lafayette, IN 47907
jphard@umich.edu

Quentin F. Stout
University of Michigan
Ann Arbor, MI 48109
qstout@umich.edu

Abstract Several allocation rules are examined for the problem of optimizing a response function for a set of Bernoulli populations, where the population means are assumed to have a strict unimodal structure. This problem arises in dose response settings in clinical trials. The designs are evaluated both on their efficiency in identifying a good population at the end of the experiment, and in their efficiency in sampling from good populations during the trial. A new design, that adapts multi-arm bandit strategies to this unimodal structure, is shown to be superior to the designs previously proposed. The bandit design utilizes approximate Gittin's indices and shape constrained regression.

Keywords: nonparametric, adaptive, sequential sampling, experimental design, dose-response, clinical trial, multi-arm bandit, up and down, random walk, stochastic approximation, Polya urn, unimodal regression

Introduction

Consider a problem in which there are k linearly ordered populations or "arms". Associated with Arm i is a binary random variable, Y_i, the outcome of which is governed by an unknown distribution function $F(i)$. The expected return from Arm i, $p_i = \mathbf{E}(Y_i)$, is referred to as its *value*.

Of interest here is the problem of identifying the arm with the highest value, given a strict unimodal structure on the values $p_1, \ldots, p_k$. As unimodality is location invariant, the arms may be located at integers $\{1, \ldots, k\}$ without any loss of generality. There are n opportunities to sample and allocation rules or sampling algorithms are assessed according to two measures — an sampling error and a decision error.

We evaluate four nonparametric sampling designs — three from the literature and a new one proposed here. Since a common application motivated development of several of these designs, we briefly review the application here. See Durham et al. (1998) and Hardwick et al. (2000) for details.

An Application:

Classical dose response problems focus on locating specified quantiles of the relationship between drug dose and probability of toxic response to drug therapy. A common framework for these problems is to model a patient's response at dose s with a Bernoulli random variable with success probability (i.e., non-toxic outcome) $1 - Q(s)$, where $Q(s)$ is a continuous nondecreasing function.

We focus on a variation of this problem in which there are only k doses. We wish to maximize the probability that a patient being treated exhibits not only a *non-toxic* response but is also *cured* at a given dose. Let $R(i)$ be a non-decreasing response curve that models the probability that dose i is effective, and take $F(i) = R(i)(1 - Q(i))$ to be the product curve for efficacy and non-toxicity, $i = 1, \ldots, k$. Note that in many common parametric dose response settings, the curve F is unimodal.

The goal is to develop a sampling design that identifies the optimal dose, $i^* = \arg\max_i F(i)$. The problem may be formulated with R conditioned on Q, and here we take the special case of R and Q independent. Further, while in some scenarios we are provided with the individual outcomes from R and Q, in this case, we assume that only the outcome of F is observed (see Hardwick et al. (2000)).

In clinical trials there are typically two populations of patients to consider — the n trial subjects and the unknown number of patients in the future who will be affected by the terminal decision of the experiment. We consider a good design to be one that has a relatively high probability that both

- trial subjects will be cured, and
- the optimal dose is selected as best at the trial's termination.

Since these two criteria are opposing, a single optimal design for this problem doesn't exist. Instead, we seek designs that lie on or close to an optimal trade-off curve representing performance along these two measures. Note that it is not known how to optimize either measure in this setting.

In this paper, we compare a new shape constrained multiarm bandit sampling rule with three other designs that have been proposed for virtually the same problem. In Sections 1.1 and 1.2, respectively, we describe urn and up and down designs as proposed in Durham et al. (1998). In Section 1.3, we describe a stochastic approximation type of design delineated in Herkenrath (1983). The bandit design, adapted to the unimodal structure, is outlined in Section 1.4. In Section 2 we describe evaluation criteria and experimental results. In Section 3 we discuss asymptotic behavior, and in Section 4 we close with a short discussion.

1. SAMPLING DESIGNS

We begin with some notation. Recall that i^* denotes the best arm and let p^* be the value of Arm i^*. Arm i_m is sampled at stage m (the m^{th} observation). Let $\mathcal{I}_{im} = 1$ if $i = i_m$ and 0 otherwise. Then $n_{im} = \sum_{j=1}^{m} \mathcal{I}_{ij}$ is the number of observations sampled from Arm i by stage m for $i = 1, \ldots, k$; $m = 1, \ldots, n$. Thus $\sum_{i=1}^{k} n_{im} = m$.

Let Y_{im} represent the outcome of Arm i at trial m. For convenience we take Y_{im} to be 0 unless $i = i_m$, in which case, Y_{im} has a Bernoulli outcome with success rate p_i. Then, $r_m = \sum_{i=1}^{k} Y_{im}$ is the return or "reward" received at stage m. Two designs considered here take observations in pairs, (Y_1, Y_2) as opposed to individually. To keep notation consistent, we set $Y_1(ij) = Y_{im}$ and $Y_2(ij) = Y_{i(m+1)}$ for $j = 1, \ldots, n/2$, $m = 1, \ldots, n-1$.

The empirical mean $\hat{p}_{im}$ of p_i after stage m is given by $\sum_{j=1}^{m} Y_{ij}/n_{im}$. For Bayesian designs we assume that each p_i follows an independent beta distribution with parameters a_i, b_i. Thus the posterior mean $\bar{p}_{im}$ of p_i after stage m is given by $(a_i + \sum_{j=1}^{m} Y_{ij})/(a_i + b_i + n_{im})$.

To specify a design, one needs:

- A *sampling rule* that determines which arm to sample at each stage. This may involve special rules to handle startup, boundaries, ties, and observations near the end of the experiment.
- A *terminal decision rule*, to determine the arm declared best at the end of the experiment.

1.1 Randomized Polya Urn Design

A *randomized Polya urn for selecting optima* is proposed in Durham et al. (1998). In this design, arms are sampled according to a draw from the urn. For each i in $\{1, \ldots, k\}$, the urn initially contains $\alpha_i > 0$ balls labeled i. (In the present case, $\alpha_i = 1$). At stage m, $m = 1, \ldots, n$:

1. A ball is drawn at random from the urn (and replaced), and an observation is taken from the arm corresponding to the ball's label.

2. If the response is a success, then another ball with the same label is added to the urn.

3. If the response is a failure, then no new balls are added.

A stopping rule is associated with this urn process and the authors' urn contains some additional information that pertains only to the stopping time. Here, in order to compare the urn design with the others discussed, we assume a fixed sample size. The terminal decision rule is to select the arm with the highest number of balls in the urn as best.

1.2 Up and Down Design

Random walks or up and down designs are well known for the problem of locating quantiles of a non-decreasing dose response function (see Dixon (1965) and Flournoy et al. (1995)). For the case in which the response curve is strictly unimodal, Durham et al. (1998) propose an *up and down rule for targeting the optimum*, defined as follows.

At each stage j, $j = 1, \ldots, \frac{n}{2}$, a pair of arms is sampled. Let $M(j)$ represent the midpoint between the two arms sampled at stage j. Observations $Y_1(ij)$ and $Y_2(ij)$, are taken at $M(j)-c$ and $M(j)+c$, respectively, where $c = \frac{b}{2}$ and b is an odd positive integer. The midpoint for the next two observations at stage $j+1$ is given by $M(j+1) = M(j) + V_j$, where

$$V_j = \begin{cases} 1 & \text{if } Y_1(ij) = 0 \text{ and } Y_2(ij) = 1 \\ 0 & \begin{array}{l} \text{if } Y_1(ij) = 0 \text{ and } Y_2(ij) = 0 \\ \text{or } Y_1(ij) = 1 \text{ and } Y_2(ij) = 1 \end{array} \\ -1 & \text{if } Y_1(ij) = 1 \text{ and } Y_2(ij) = 0 \end{cases}$$

If $M(j+1)$ would fall outside the range, then $M(j+1) = M(j)$. In practice (and herein), b is typically 1, in which case the start-up rule for the process selects $M(1) = 1.5$, the midpoint of the leftmost two arms.

Durham et al. (1998) do not specify a terminal decision rule, but here we assume that the decision is to select the arm having the largest sample mean, i.e., $\arg\max_i \hat{p}_{in}$. Note that with a Markov chain such as this, one would ordinarily select the site most visited as best. For this design, however, such a decision process does not always converge to the optimal arm in the limit so the empirical mean is used instead.

1.3 Stochastic Approximation Design

Another approach for this problem is to use a Keifer-Wolfowitz type of stochastic approximation rule for locating a local maximum [Keifer and Wolfowitz (1952) and Sacks (1958)]. Typically a design of this sort samples at points $X(j) - c_j$ and $X(j) + c_j$, $c_j > 0$, where $\lim c_j \to 0$. The "decision point" $X(j+1)$ is given by

$$X(j+1) = X(j) + \tfrac{a_j}{c_j} V_j,$$

where the sequence $\{a_j > 0\}$ is such that $\Sigma a_j = \infty$ and $\Sigma a_j^2 c_j^{-2} < \infty$.

One difficulty with the above rule is that the sampling points $X(j) \pm c_j$ do not lie on a lattice. Herkenrath (1983) proposed a modified procedure adapted both to the discrete case and to Bernoulli observations. With this procedure, observations are only taken at arms representing the left and right neighbors of $X(j)$. The method is as follows. Let $0 < d < 1/2$ and $q(x) = x - x^l$ where $x^l = \lfloor x \rfloor$ (though k^l is set to $k-1$), and let $x^r = x^l + 1$.

According to Herkenrath (1983), the process begins with $X(1) = 1$. At stage j, allocation of 2 observations is randomized between left and right neighbors of $X(j)$ according to their relative distances. The sampling is guided by the position of $x = X(j)$ as specified in the first column of the following table. The succeeding three columns of the table display the probabilities of sampling pairs of observations as indicated at the top of the columns. Once sampling has taken place, the $j+1^{st}$ decision point is given by

$$X(j+1) = Proj_{[1,k]}\{X(j) + a_j S(Y_1(ij), Y_2(ij), X(j))\},$$

where $a_j \to 0$ and S is a time invariant function that depends on $X(j)$ and the outcomes of the observations at stage j (see Herkenrath (1983) for details).

For this design, the terminal decision is to select the arm closest to $X(n+1)$.

If $\downarrow$ then sample $\rightarrow$	x^l, x^l	x^l, x^r	x^r, x^r
$x < x^l + d$	$1 - q(x^l + d)$	$q(x^l + d)$	0
$x^l + d \le x < 1/2(x^l + x^r)$	$1 - q(x)$	$q(x)$	0
$x = 1/2(x^l + x^r)$	0	1	0
$1/2(x^l + x^r) < x \le x^r - d$	0	$1 - q(x)$	$q(x)$
$x^r - d < x$	0	$1 - q(x^r - d)$	$q(x^r - d)$

1.4 Unimodal Bandit Design

In general k-arm bandit problems, the k populations have unknown reward structures, *arm pulling* or sampling takes place sequentially, and decisions are made with the goal of optimizing a discounted sum of all returns, $\Sigma_1^n \beta_m r_m$ for discount sequence $\{\beta_m \geq 0\}$ (see Berry and Fristedt (1985) for details). This formulation covers a broad class of problems in learning theory, and optimal strategies balance the impulse to earn immediate rewards against the need to gather information for future decisions.

It is natural to model initial information about the arms of a bandit using a prior distribution on the "values", p_i, $i = 1, \ldots, k$, of the arms. As sampling take place, posterior distributions reflect the additional information acquired. Optimal strategies for Bayesian bandits with finite "horizon" or sample size can be determined via dynamic programming Bellman (1956). However, computing such solutions is a daunting task. For the simple case involving independent Bernoulli arms, the dynamic programming equations require computational space and time that grow as $n^{2k}/(2k-1)!$ for a problem of horizon n. (See Hardwick et al. (1999) for the largest problems yet solved.)

As the bandit model becomes more complex (e.g., the number of outcomes per arm increases or there is structure on the arms), the problem quickly outgrows available computer resources. In the present situation, we face k *dependent* arms, and thus can obtain optimal solutions in only trivial cases. One option in handling these more complex problems, however, is to exploit known characteristics of optimal solutions to the simpler bandit problems.

In particular, there is a bandit model that, appropriately parameterized, approximates the independent finite horizon model with discount sequence $\beta_j = 1$, $j = 1, \ldots, n$. This model, the *geometric* bandit in which $\beta_j = \beta^{j-1}$ for $0 < \beta < 1$ and $j = 1, 2, \ldots$, offers a surprising solution. Note first that the infinite horizon in this model makes dynamic programming impractical. However, with independent arms, optimal solutions for the geometric bandit may be defined in terms of index rules. Thus, at each stage of the experiment and for each arm, there exists an index depending only on the arm, such that sampling from the arm with the highest index yields the optimal solution, (Gittins and Jones (1974)). Known as Gittin's Indices, these quantities incorporate available "knowledge" about an arm along with the arm's value. The existence of an index rule greatly simplifies the solution to the geometric bandit problem because it reduces the complexity of the k-arm bandit from being exponential in k to being linear in k. However, despite the elegant solution presented in Gittins and Jones (1974), the indices themselves are extremely complicated. Except in very simple models, they cannot be calculated exactly. Still, the idea of using an index rule with traits similar to the Gittin's index has great appeal. Because the objective function of the finite horizon problem is to maximize return, it's reasonable to ask how well the bandit solution accommodates statistical goals of gathering information for inferential purposes.

Here, we use a lower bound for the Gittin's index proposed in Hardwick (1995). If the beta prior parameters for an arm are A and B, then a lower bound for the Gittins index for this arm is given by $\Lambda^* = \sup\{\Lambda_r : r = 1, 2, \ldots\}$, where

$$\Lambda_r = \frac{\frac{\Gamma(A+1)}{\Gamma(A+B+1)} - B\Sigma_1^r \beta^i \frac{\Gamma(A+i)}{\Gamma(A+B+i+1)}}{\frac{\Gamma(A)}{\Gamma(A+B)} - B\Sigma_1^r \beta^i \frac{\Gamma(A+i-1)}{\Gamma(A+B+i)}}.$$

Generally speaking, it is not difficult to compute Λ^* since Λ_r is a unimodal function of r.

Interestingly, geometric discounting can be viewed as an ethically equitable mechanism for balancing the well being of current and future patients. The parameter β represents the relative weight of the outcome of each subject as compared to the weight of all future patients. In the finite horizon setting, optimal decisions will be virtually identical if $n \sim 1/(1-\beta)$ when $n \to \infty$ and $\beta \to 1$. This relationship is quite accurate when n is as low as, say 30.

In the present problem, we do not focus on these interpretations, but instead wish to choose β to optimize our joint goals of minimizing loss and making good decisions. It's an open question how best to do this. For simplicity in our simulation experiments, a fixed value of $\beta = 0.95$ was used for all sample sizes. Naturally somewhat better results would be obtained if β changed somehow with n. Note that it may even make sense to adjust β during a given experiment. Since it is best to explore more early in the experiment, higher values may be more useful then, with lower values towards the end helping to diminish experimental losses.

The *unimodal bandit* sampling algorithm used to generate the results in the next section is constructed as follows. Let the prior distributions of the p_i be beta with parameters (a_i, b_i) (we use $a_i = b_i = 1$). At stage $m+1$,

1. Calculate the posterior means, $\bar{p}_{im}$, $i = 1, \ldots, k$.

2. Using least squares, fit the best unimodal curve to the posterior means using weights $n_{im} + a_i + b_i$ (Stout and Hardwick (2000)).

3. Adjust the posterior distributions so that their means lie on the curve from (2) by adding the smallest (fractional) number of successes, u_{im}, or failures, v_{im}, needed to bring the posterior means into alignment. For Arm i the *adjusted posterior parameters* are

$$A_{im} = a_i + \sum_{j=1}^{m} Y_{ij} + u_{ij} \quad \text{and} \quad B_{im} = b_i + n_{im} - \sum_{j=1}^{m} Y_{ij} + v_{ij}.$$

4. For each arm i, calculate $\Lambda^*_{i,m}$, based on the adjusted posterior distributions in (3), and let $j = \arg\max_i \Lambda^*_{i,m}$. If there are ties for the maximum, then pick j at random from the tied arms.

5. Determine where to take the next observation:

   (a) If observation m was a success, or was not on arm j, then sample next from arm j.

   (b) *Exploration rule:* If observation m was a failure and was on arm j, then pick j' uniformly from $\{j-1, j+1\}$ for $j' \in \{1, \ldots, k\}$. If the p-value for the one sided test of equality of j and j' is at least $1/(n-m+1)$ then sample from j', otherwise sample again from j.

At the end of the experiment, fit a unimodal curve to the weighted posterior means and select the arm corresponding to the mode.

Note the *exploration rule* in Step 5. This is used to avoid premature convergence to a suboptimal arm. While the allocation index also serves this purpose, this may not force as much exploration as needed. The exploration rule also lends robustness to prior misspecification.

2. EVALUATION CRITERIA AND RESULTS

As discussed, we seek designs that behave well along two performance measures – an experimental or sampling error to assess losses during the experiment and a decision error to assess future losses based on the terminal decision.

Given any decision rule and sampling design there exists a probability measure on the arms which reflects the chance, $\pi_n(i)$, that Arm i is selected as best at the end of an experiment of size n. One could take $\pi_n(i^*)$, the *probability of correct selection*, as a measure of decision efficiency. However, in clinical trials, a major goal is to minimize harm to patients. In such cases, selecting an arm with a success rate close to p^* is much preferred to selecting those with lower rates. This leads to the following definition:

$$\text{Decision Efficiency}: \quad \mathcal{D}_n = (\Sigma_{i=1}^{k} \pi_n(i) p_i)/p^*$$

The sampling error is the normalized expected loss incurred when sampling from arms other than i^*. Noting that $\mathbf{E}[r_m]$ is the expected return at stage m, we define

$$\text{Sampling Efficiency}: \quad \mathcal{S}_n = (\Sigma_{m=1}^n \mathbf{E}[r_m])/(n \cdot p^*)$$

This is closely related to the *expected successes lost*, $\Sigma_{m=1}^n p^* - \mathbf{E}[r_m]$.

An initial simulation study was conducted to examine the relative behavior of the designs in Section 1. To assess the impact of variation in the unimodal functions themselves, we selected four curves with diverse characteristics. These are shown in Figure 1. For each curve and each design, we consider sample sizes 25, 50, 100 and 200. At each configuration 5,000 iterations of the experiments were run and used to estimate the decision efficiency and sampling efficiency, shown in Figure 3. Figure 3 also includes results for equal allocation, in which there are $\lfloor k/n \rfloor$ rounds of sampling from each arm and the remaining observations are assigned at random and without replacement to $\{1, \ldots, k\}$.

Generally speaking, the bandit design performed the best along both measures, and no design was uniformly worst. For the urn, bandit and equal allocation rules, the efficiencies move from the southwest towards the northeast as the sample size increases, which is the behavior that one would expect. However, for the up and down design, this movement is reversed in Curves 2 and 3, and the stochastic approximation design shows no consistent behavior.

These anomalies point to the impact of design details that are often ignored. Start-up rules, for example, can have a grave impact on design performance. For the urn method, there is a random start, while for the bandit design, the start is dictated by the prior distribution, which in the present case yields a random start. For the up and down and stochastic approximation designs, however, sampling begins with the leftmost arm, as designated by the developers of the designs. If a curve has its mode to the left, these designs tend to initially stay near the mode. For example, with Curve 3, which is strictly decreasing, the up and down design samples *only* from Arms 1 and 2 nearly 9% of the time when $n = 25$. This is primarily due to observing repeated failures on these arms yet not moving away. Note, however, that as the sample sizes increase, this tendency is overcome and the design's performance stabilizes.

We reran the experiments with the ordering on the arms reversed, i.e., so that curves that originally were decreasing were now increasing and so forth. This switch caused the efficiencies of the up and down and stochastic approximation designs to plummet for Curves 2 and 3, whereas they improved for Curve 4.

To draw more direct comparisons among the rules, we also reran the experiments using a random start-ups for all designs. The results of these runs appear in Figure 2. For the up and down design, the unbiased start-up reordered the efficiencies upward and to the right with the sample size. Further, the results show an overall improvement over those obtained using the leftmost start-up. Note, however, that the preferred start-up rule may depend on the application. For the dose response example discussed in the Introduction, Durham et al. (1998) determine that it is preferable to have a patient succumb to disease than it is to be exposed to toxic levels of drug therapy, and thus it was deemed desirable to begin with low rather than arbitrary drug doses.

Figure 2 shows that even when utilizing a random start-up, the stochastic approximation design still behaved erratically. In Curve 1, for example, the value for $n = 25$ was wildly high on the E_n measure – so much so that it has been omitted from the figure. A less extreme but similar behavior can be observed in Curve 3. For neither of these curves did stochastic approximation show the desired trend of efficiency increasing with the sample size. We believe that this unusual behavior may be the result of the choice of the S function mentioned in Section 1.3. Herkenrath (1983) notes that the system of equations defining S are underdetermined and thus adds conditions to guarantee unique solutions. The specific form of S chosen by the author appears to be questionable, and it is likely that a different solution would improve performance and produce less erratic behavior.

While the bandit rule performs the best with regard to decision making, note the up and down and equal allocation rules have similar $\mathcal{D}_n$ when n is large. As expected, however, the bandit rule outmatches all rules on the $\mathcal{S}_n$ measure. Among the adaptive rules examined, the urn model seems to perform least well on our measures. However, the main strength of this model is that it is randomized and can better protect patients from selection bias.

3. CONVERGENCE

While in practice these designs are intended for use with small to moderate sample sizes, it is also important that they exhibit good asymptotic efficiency, i.e., that $\mathcal{D}_n$ and $\mathcal{S}_n$ converge almost surely ($a.s.$) to 1 as $n \to \infty$. If a design is asymptotically efficient with respect to both kinds of error, then the design is asymptotically (first order) optimal. In this section, we assume that efficiency is “asymptotic” and thus drop the adjective.

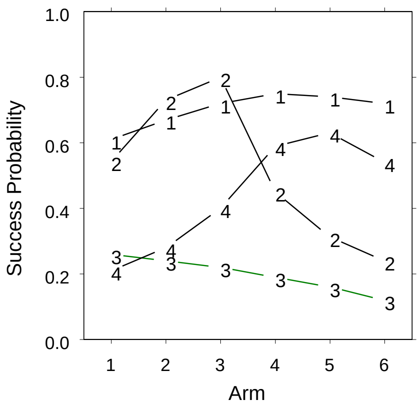

Figure 1. Unimodal curves used for evaluation.

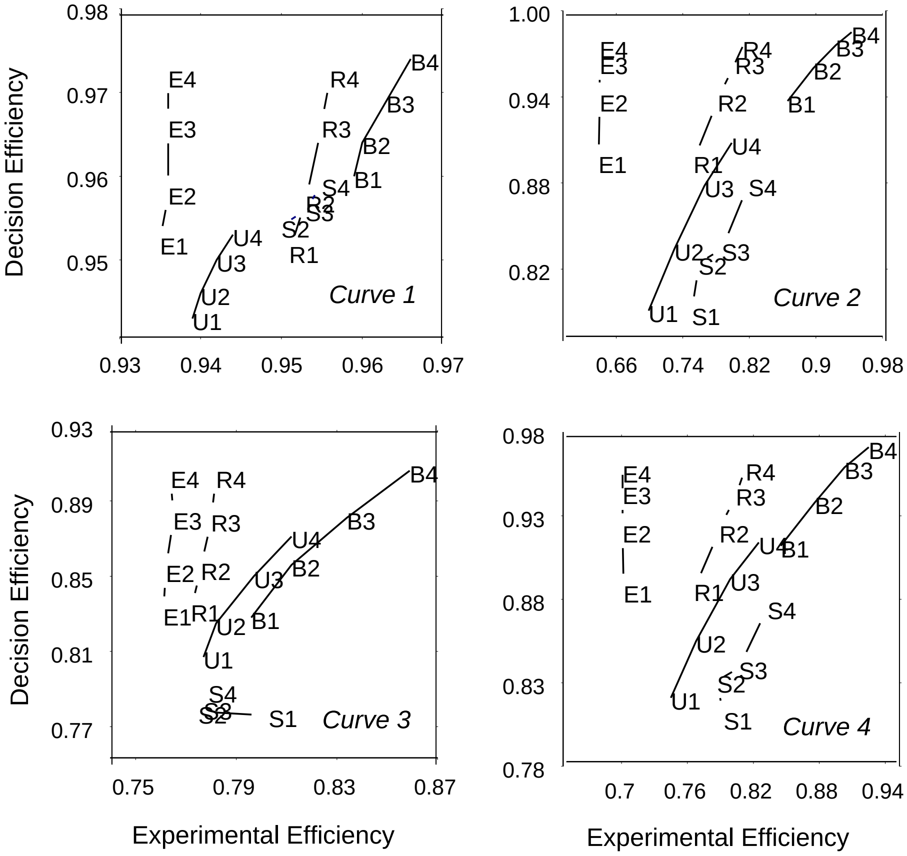

Figure 2. Efficiencies, for n=(1) 25, (2) 50, (3) 100, (4) 200.
B=Bandit, E=Equal, R=Up-Down, S=Sto. Approx., U=Urn
Random start-up for all methods

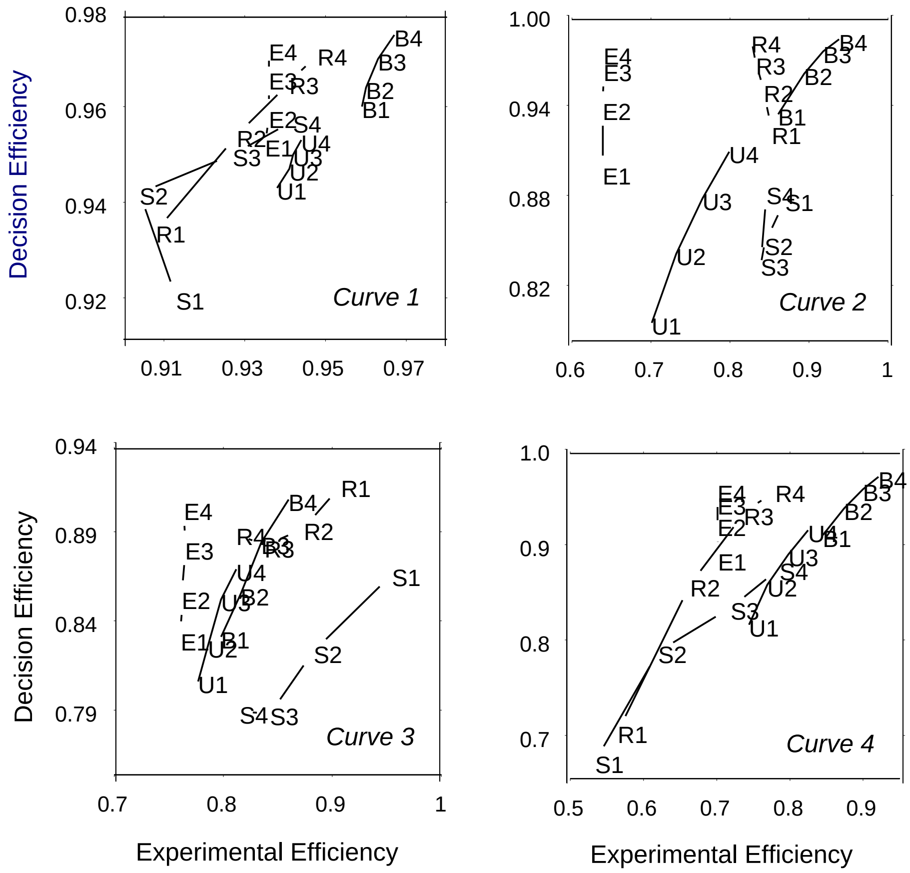

Figure 3. Efficiencies, for n=(1) 25, (2) 50, (3) 100, (4) 200.
B=Bandit, E=Equal, R=Up-Down, S=Sto. Approx., U=Urn
Start-up as proposed with methods.

To obtain sampling efficiency, it is necessary and sufficient that the rate at which arm i^* is sampled goes to 1 as $n \to \infty$, while decision efficiency requires that the probability of selecting arm i^* as best goes to 1 as $n \to \infty$. Note that for a design to insure that it has not prematurely converged in its selection of a good arm, with the unimodal assumption it is necessary that arms $i^* \pm 1$ be sampled infinitely often. Without such an assumption, all arms would have to be sampled infinitely often.

Equal Allocation: It is straightforward to show that this design is decision efficient, but not sampling efficient.

Up and Down Design: The decision rule for this design selects the arm corresponding to $\max_i \hat{p}_{in}$ as best. Since all arms are sampled $i.o.$, $\hat{p}_{in} \to$

p_i *a.s.* for each i. Thus, choosing the arm with the highest sample mean guarantees *a.s.* selection of Arm i^* in the limit, i.e., the up and down rule is decision efficient. However, because the asymptotic rate at which suboptimal arms are sampled is nonzero, the design is not sampling efficient. Note that unimodality is not required for decision efficiency.

Urn Design: Let pr_i be the asymptotic proportion of balls of type i in the urn as $n \to \infty$. In Theorem 2 of Section 4.2, Durham et al. (1998) show that $pr_{i^*} = 1$ and $pr_i = 0, i \neq i^*$ *a.s.* Since the decision rule chooses the arm with the maximum proportion of balls at stage n as best, the rule *a.s.* makes the correct selection as $n \to \infty$. Since $pr_{i^*} = 1$, the design is also sampling efficient. Note that unimodality is not required for the urn design to be both decision and sampling efficient.

Stochastic Approximation: Herkenrath (1983) shows that when the sequence $\{a_n\} \to 0$, $\sum_1^\infty a_n = \infty$ and $\sum_1^\infty a_n^2 < \infty$, then $X(n)$ converges to i^* *a.s.* Thus, in the limit, the decision to select the arm closest to $X(n)$ as best is *a.s.* the optimal decision, and hence this design is decision efficient. Unfortunately, with the probability of sampling adjacent pairs being fixed, this design cannot be sampling efficient.

Unimodal Bandit: The exploration rule, part 5b, ensures that the arm thought to be best and its neighbors are sampled *i.o.*, to avoid premature convergence. Under the unimodal assumption, it can be shown that this insures that asymptotically the bandit *a.s.* selects i^* as the best arm, and thus is decision efficient (see Hardwick and Stout (2000)). Further, since the rate of sampling other arms goes to zero, the design is also sampling efficient. Note that if β is bounded away from 1, then the arms not adjacent to i^* are *a.s.* sampled only finitely often. Also note that sampling efficiency can be lost if β goes to 1 too rapidly as n increases.

The conditions needed for first order sampling efficiency are not difficult to achieve. However, to obtain second order efficiency, more delicate control of the rates at which suboptimal arms are sampled is needed. For the unimodal bandit design, the discount parameter β and the exploration rule together dictate this rate, as ordinarily would the sequence $\{a_n\}$ for the stochastic approximation design. (As mentioned, it is the deterministic sampling of adjacent pairs in the Herkenrath (1983) design that precludes sampling efficiency.) One must be careful, however, in changing the convergence rates since they affect *both* efficiency types. In fact, it is precisely these rates that control the trade-off between the competing goals of gaining high reward ($\mathcal{S}_\infty$) and drawing good inferences ($\mathcal{D}_\infty$). Forcing the sampling of suboptimal arms to go to 0 too fast reduces the rate at which we arrive at an optimal decision.

In summary, for the unimodal problem, a design needs to sample the best arm and its neighbors *i.o.* to be decision efficient. However, sampling too much from suboptimal arms negatively impacts a design's experimental regret. Thus a good rule will be one that carries out enough exploration to determine the best arm but then samples mostly from it.

4. DISCUSSION

While asymptotic analytical results can give some guidelines, they don't appear to be able to determine which designs are best on useful sample sizes, and hence computer experiments are needed. The experiments reported here, and those in Hardwick and Stout (2000), show that while all of the designs considered are asymptotically decision efficient, for fixed sample sizes the unimodal bandit appears to do slightly better than the others, at least on the curves and sample sizes examined. It appears to achieve both a good sampling efficiency and good decision efficiency for a wide range of situations.

However, there is significant work needed to tune these basic approaches to produce better designs for this setting. We are conducting experiments to evaluate designs on large numbers of unimodal functions, and to evaluate a number of variations. For example, there are many alternatives possible for the exploration rule in the unimodal bandit, and it is unclear which is best. It is also not obvious how β should go to 1 as n goes to ∞, and whether β should be reduced as an experiment progresses, to optimize performance. Further, there is significant interaction between the choice of β and the exploration rule. This is a problem for future study. There are also many variations possible for random walks, urn designs, etc., and in initial experiments we have been able to achieve some improvement.

Returning to the motivating example of dose-response problems, we note that there is also the important case in which the competing failure modes, Q and R in the Introduction, are observed. As expected, one can do better by using designs which exploit the structure of the failure modes (see Hardwick et al. (2000)).

References

Bellman, R. (1956), "A problem in the sequential design of experiments", *Sankhya A* **16:** 221–229.

Berry, D.A. and Fristedt, B. (1985), *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall.

Dixon, W. (1965), "The up and down method for small samples", *J. Amer. Statist. Assoc.* **60:** 967–978.

Durham, S., Flournoy, N., Li, W. (1998), "A sequential design for maximizing the probability of a favourable response", *Can. J. Statist.* **26:** 479–495.

Flournoy, N., Durham, S., and Rosenberger, W. (1995), "Toxicity in sequential dose-response experiments", *Sequential Anal.* **14:** 217–227.

Gittins, J.C. and Jones, D.M. (1974), "A dynamic allocation index for the sequential design of experiments", *Progress in Statistics* (ed. by J. Gani et al.), 241–266, North Holland.

Hardwick, J. (1995), "A modified bandit as an approach to ethical allocation in clinical trials", *Adaptive Designs: Institute Math. Stat. Lecture Notes* **25** 223–237. (B. Rosenberger & N. Flournoy, ed.'s).

Hardwick, J., Oehmke, R. and Stout, Q.F. (1999), "A program for sequential allocation of three Bernoulli populations", *Comput. Stat. and Data Analysis* **31:** 397–416.

Hardwick, J., Meyer, M. and Stout, Q.F. (2000), "Adaptive designs for dose response problems with competing failure modes", submitted.

Hardwick, J. and Stout, Q.F. (2000), "Bandit designs for optimizing a unimodal response function", in preparation.

Herkenrath, U. (1983), "The N-armed bandit with unimodal structure", *Metrika* **30**: 195–210.

Keifer, J. and Wolfowitz, J. (1952), "Stochastic estimation of the maximum of a regression function", *Ann. Math. Statist.* **25:** 529–532.

Sacks, J. (1958), "Asymptotic distribution of stochastic approximation procedures", *Ann. Math. Statist.* **29:** 373–405.

Stout, Q.F. and Hardwick, J. (2000), "Optimal algorithms for univariate unimodal regression". To appear in *Computing Science and Statistics* **32**.