In *MODA6: Model Oriented Data Analysis*, (A. Atkinson, P. Hackl, W. Müller, eds.), Physica Verlag, 2001, pp. 127–134.

Optimal Adaptive Designs for Delayed Response Models: Exponential Case

Janis Hardwick
Robert Oehmke
Quentin F. Stout

ABSTRACT: We propose a delayed response model for a Bernoulli 2-armed bandit. Patients arrive according to a Poisson process and their response times are exponential. We develop optimal solutions, and compare to previously suggested designs.

KEYWORDS: multi-arm bandit; sequential sampling; design of experiments; clinical trial; ethics; algorithms; parallel processing

1 Introduction

Adaptive designs are effective mechanisms for flexibly allocating experimental resources – particularly in clinical trials. Unfortunately, optimal *fully* sequential designs require immediate responses and cannot be applied when responses are delayed. In this paper, we seek to optimize an objective function for a problem in which there are two populations and the responses, which may be delayed, are independent Bernoulli random variables.

Perhaps the simplest model to consider is one in which observations are delayed a fixed amount of time. Such models have been considered by several researchers, including Bandyopadhyay and Biswas (1996), Douke (1994), Ivanova and Rosenberger (2000), although the optimal design was only recently obtained in Hardwick, Oehmke and Stout (2001). Far more complex, however, is the problem in which the response times follow arbitrary distributions. Such models are too difficult to optimize exactly.

Taking a less general approach, here we consider the model in which patients arrive via a Poisson process and their response times follow independent exponential distributions. We assume that the arrival rate and the mean response times are known, and the goal is to optimize total patient successes during the experiment. We can model this problem as a 2-armed bandit (2AB) with delayed response. Recall that the objective of a bandit problem is to allocate resources to different experimental "arms" in such a way that the total return from the experiment is optimized.

There has been some work done on the related problem of maximizing patient survival times in a 1-armed bandit (1AB) model. In the 1AB there are actually two arms, but the attributes of one of them are completely known. In Eick (1988), the author addresses the extent to which geometric response delays affect standard behavioral characteristics of the 1AB, where the survival rate of one arm is known and the goal is to maximize total survival time by allocating patients to either the "known" or unknown therapy. Some of these results have been extended in Wang (2000).

In the next section, we develop models for the delayed response bandit and present the requisite dynamic programming equations. In Section 3, we present a delayed version of the randomized play-the-winner rule (RPW). In Section 4, we compare the delayed bandit and RPW rules with each other and to the optimal non-delayed solution generated by the 2-armed bandit algorithm. The last section, Section 5, is a discussion.

2 Models with Exponential Delay

Suppose that patients arrive according to a Poisson process with rate λ_s. As they arrive, they are assigned either to arm (treatment) 1 or 2. Patient responses are Bernoulli with success rates π_1 and π_2. Prior distributions on the π_i are $\mathrm{Be}(a_i, b_i)$, $i = 1, 2$, respectively. The response time for a patient on arm i is exponential with mean λ_i, $i = 1, 2$. Response times are independent among themselves and independent of arrival times and of actual responses. The experiment will allocate a total of n patients.

If a patient arrival occurs at time t, the patient is allocated to arm 1 or 2 based on data up until t. This includes past arrival times, response times and the responses, as well as the priors. A sufficient statistic is $\langle s_1(t), f_1(t), u_1(t); s_2(t), f_2(t), u_2(t)\rangle$, where $s_i(t)$, $f_i(t)$ are the number of successes and failures on arm i and $u_i(t)$ is the number outstanding on arm i at time t, $i = 1, 2$. Because the problem is stationary in time, we can drop the time notation. Thus a policy is a function that depends on the priors and n and maps $\langle s_1, f_1, u_1; s_2, f_2, u_2\rangle$ to $\{1, 2\}$. Optimal solutions are policies that are optimized for a given objective function. As noted, the objective here is to maximize total patient successes during the experiment, and the problem thus has the form of a two armed bandit with delay. We call this optimization problem the *delayed 2-armed bandit*, D2AB. However, our approach also works for general objective functions.

It is well-known that such optimization problems can be solved via dynamic programming. However, computational space and time grow exponentially in the number of arms, and the delay complicates this further. The state space involves all possible variations of its components, as long as all are nonnegative and their sum is no greater than n. I.e., the state space corresponds to all possible sufficient statistics. There are $\binom{n+6}{6} = \Theta(n^6)$ states in the D2AB, and the delayed k-arm bandit will have $\binom{n+3k}{3k}$ states. This is in contrast to the $\Theta(n^4)$ states in the standard 2AB, and $\binom{n+2k}{2k}$ states in the standard k-arm Bernoulli bandit.

To apply dynamic programming, one needs to know the value of each terminal state, i.e., those states which can be directly evaluated without recourse to recursion. These are the states for which $s_1+f_1+s_2+f_2 = n$. Ultimately, the goal is to determine the value, V, of the initial state $\langle 0,0,0;0,0,0\rangle$.

There are various ways to tackle this problem, but finding one that is computationally feasible is a keystone of the solution. Perhaps the most natural approach is the one in which time is marked by patient arrivals, because these are the only times when action is taken and decisions are needed. Unfortunately, this formulation is too hard to solve computationally, taking $\Theta(n^{10})$ time. For further details, see Hardwick et al. (2001).

A second approach marks time by *events*, where an event is either a subject arrival or a response from one of the arms. Because we are using continuous time, we can assume that only one event occurs at a time. Let $P_1(u_1,u_2)$, $P_2(u_1,u_2)$, $P_s(u_1,u_2)$ represent the probability that the next event is an observation on arm 1, an observation on arm 2, or a subject arrival, respectively. Fortunately, P_1, P_2 and P_s have a simple form:

$$P_s(u_1,u_2) = \frac{\lambda_s}{\lambda_s + u_1\cdot\lambda_1 + u_2\cdot\lambda_2} \quad \text{and} \quad P_i(u_1,u_2) = \frac{u_i\cdot\lambda_i}{\lambda_s + u_1\cdot\lambda_1 + u_2\cdot\lambda_2}.$$

Let $\pi_i(s_i,f_i)$ denote the probability that an observation on arm i will be a success, given that s_i successes and f_i failures have been previously observed on the arm. Also, let $\widehat{y}$ represent component y increased by one and $\sigma+\widehat{y}$ be state σ with component y increased by one. Then the dynamic programming equation for determining the value of state $\sigma = \langle s_1,f_1,u_1;s_2,f_2,u_2\rangle$ is:

$$\begin{aligned} V(\sigma) \quad = \quad & P_1(u_1,u_2) * \Big[\pi_1(s_1,f_1)\cdot V(\sigma+\widehat{s_1}-\widehat{u_1}) \\ & \qquad\qquad +(1-\pi_1(s_1,f_1))\cdot V(\sigma+\widehat{f_1}-\widehat{u_1})\Big] \\ & +P_2(u_1,u_2) * \Big[\pi_2(s_2,f_2)\cdot V(\sigma+\widehat{s_2}-\widehat{u_2}) \\ & \qquad\qquad +(1-\pi_2(s_2,f_2))\cdot V(\sigma+\widehat{f_2}-\widehat{u_2})\Big] \\ & +P_s(u_1,u_2) * \max\{V(\sigma+\widehat{u_1}),\ V(\sigma+\widehat{u_2})\} \end{aligned}$$

Here, the allocation choice is handled in the last term, where if there is a subject arrival then we just determine to which arm we allocate. Initially this just means that the arm has one more unobserved allocation. The advantage of this approach is that it requires only $\Theta(n^6)$ time. While still formidable, this can be achieved for useful sample sizes. For example, problems of size $n = 200$ have been optimized using a parallel computer. See Oehmke, Hardwick and Stout (2001) for a discussion of the parallelization process and optimizations to improve performance.

3 A Randomized Play-the-Winner Rule

Exact evaluations of arbitrary, sub-optimal allocation designs are possible via slight modifications to the algorithm in Oehmke et al. (2001). One popular such rule is the randomized play the winner (RPW) rule which first appeared in Wei and Durham (1978). In this urn model, there are initial balls representing the treatment options. Patients are assigned to arms according to the type of ball drawn at random from the urn. Sampling is with replacement, and balls are added to the urn according to the last patient's response. Using RPW, the proportion of allocations to the better arm converges to one.

One advantage of urn models like RPW is the natural way in which delayed observations can be incorporated into the allocation process. When a delayed response eventually comes in, balls of the appropriate type are added to the urn. Since sampling is with replacement, any delay pattern can be accommodated. We call this design the *delayed RPW rule* (DRPW). The same approach was used in Ivanova and Rosenberger (2000), in which responses occurred with a fixed delay. In Bandyopadhyay and Biswas (1996) the authors consider a slightly altered version of this rule for a related best selection problem.

4 Results of Comparisons

We have carried out exact analyses of the exponential delay model for both the D2AB and DRPW. In these preliminary analyses, we take $n = 100$. For the DRPW we initialize the urn with one ball for each treatment. If a success is observed on treatment i then another ball of type i is added to the urn, while if a failure is observed then another ball of type $3 - i$ is added.

For comparative purposes, we look at base and best case scenarios. The best fixed in advance allocation procedure is the base case, i.e., the optimal solution when no responses will be available until after all n patients have been allocated. To maximize successes one should allocate all patients to the treatment with the higher expected success rate. We denote the expected number of successes in the base case by $\mathrm{E}_b[\mathrm{S}]$. Here, we consider only uniform priors on the treatment success rates π_1 and π_2, in which case any fixed allocation is best. For these priors, $\mathrm{E}_b[\mathrm{S}] = n/2$.

We encounter the best possible case when all responses are observed immediately (full information). In this situation, DRPW is simply the regular RPW and the D2AB is the regular 2-armed bandit. Recall that the regular 2-armed bandit optimizes the problem of allocating to maximize total successes. Letting $\mathrm{E}_{opt}[\mathrm{S}]$ represent expected successes in the best case, we have $\mathrm{E}_{opt}[\mathrm{S}] = 64.9$ for our example. Using the difference $\mathrm{E}_{opt}[\mathrm{S}] - \mathrm{E}_b[\mathrm{S}]$ as a scale for improvement, we can think of the values on this scale, (0,

λ_1	λ_2						
$\downarrow$	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}
10^{-5}	50.1						
10^{-4}	51.2	51.2					
10^{-3}	55.4	55.4	55.8				
10^{-2}	59.3	59.4	59.9	61.5			
10^{-1}	60.9	61.0	61.6	63.1	64.1		
10^{0}	61.3	61.3	61.9	63.5	64.5	64.8	
10^{1}	61.3	61.3	62.0	63.5	64.6	64.8	64.9

TABLE 1.1. Bandit: E[S] as (λ_1, λ_2) vary, $n = 100$, $\lambda_s = 1$, uniform priors

λ_1	λ_2						
$\downarrow$	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}
10^{-5}	50.0						
10^{-4}	50.2	50.4					
10^{-3}	51.6	51.7	52.6				
10^{-2}	54.8	54.8	54.9	55.7			
10^{-1}	56.5	56.5	56.5	56.7	57.3		
10^{0}	56.9	56.9	56.9	57.1	57.6	57.8	
10^{1}	57.0	57.0	57.0	57.2	57.6	57.8	57.9

TABLE 1.2. RPW: E[S] as (λ_1, λ_2) vary, $n = 100$, $\lambda_s = 1$, uniform priors

14.9), as representing the “extra” successes over the best fixed allocation of 100 observations. We take $\mathrm{R}(\delta) = \big(\mathrm{E}_\delta[\mathrm{S}] - \mathrm{E}_b[\mathrm{S}]\big)/\big(\mathrm{E}_{opt}[\mathrm{S}] - \mathrm{E}_b[\mathrm{S}]\big)$ to be the *relative improvement* over the base case for any allocation rule δ. While $\mathrm{R}(\delta)$ also depends on n and the prior parameters, these are omitted from the notation.

Note that $\mathrm{R(D2AB)} \to 1$ and $\mathrm{R(DRPW)} \to 1$ as $n \to \infty$. However, this asymptotic behavior gives little information about the values for practical sample sizes. Hence, their behavior must be determined computationally.

Tables 1.1 and 1.2 contain the expected successes for the D2AB and the DRPW rules, respectively. Patient response rates, λ_1 and λ_2, vary over a grid of values between 10^{-5} and 10^1, and the patient arrival rate is fixed at 1. Note that, for both rules, when $\lambda_1 = \lambda_2 = 10^{-5}$, $\mathrm{E[S]} \approx 50$. When $\lambda_1 = \lambda_2 = 10$, the delayed bandit rule gives E[S]=64.9 as one would expect. Note that in the best case scenario for the DRPW, E[S] = 57.9, which gives an R of 0.53. With the RPW, we can expect to gain only 7.9 successes as compared to the 14.9 for the optimal bandit.

Moving away from the extreme points, consider the case when λ_1, λ_2 and λ_s are all the same order of magnitude. The D2AB rule is virtually unaffected,

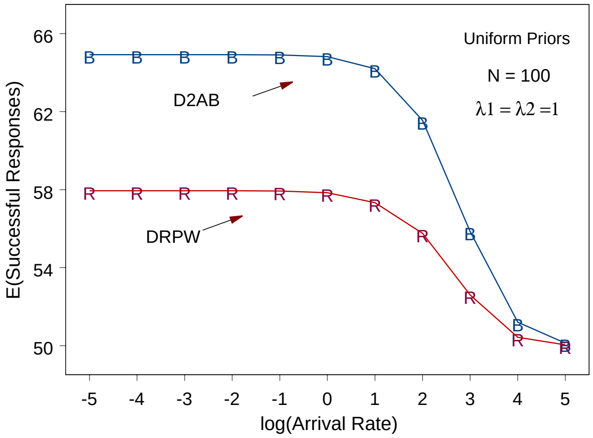

FIGURE 1. Expected successes for D2AB and DRPW, $\lambda_1 = \lambda_2 = 1$

with an R value of 0.99. This is true because, on average, there is only one patient unobserved (but allocated) throughout the trial. For the DRPW, also, R(DRPW) is only slightly smaller than R(RPW) = 0.52. Both rules seem quite robust to mild to moderate delays in adaptation. It is only when *both* response rates are at least three orders of magnitude below the arrival rate that results begin to degrade seriously. When $\lambda_1 = \lambda_2 = 10^{-3}$, for example, R(D2AB) is only 0.40, and R(DRPW) is a dismal 0.17. It is also interesting to note that even when the response rate is only $1/100^{\text{th}}$ the arrival rate, the D2AB does better than the RPW with immediate responses. Figure 1 illustrates R(D2AB) and R(DRPW) when the response rates are both one but the arrival rate varies between 10^{-5} and 10^{5}.

When we consider scenarios in which only one treatment arm supplies information to the system, we see an interesting result. For example, using uniform priors, when $\lambda_1 = \lambda_s = 1$ but $\lambda_2 = 10^{-5}$, the relative improvement is 0.76 for the D2AB and 0.47 for the DRPW. This is an intriguing result for the DRPW since its R-value is 89% of the best possible RPW value. Still, one clearly prefers the D2AB since we only get a 24% loss over the optimal solution while excluding half the information.

One way to view this problem independently from the allocation rules is to examine the expected number of allocated but unobserved patients when a new patient allocation decision must be made. As noted, when the response delay rate is 1, at any point in time one expects only a single observation to be delayed, and the impact on performance is minimal. When $\lambda = 0.1$, once approximately 20 patients have been allocated there is a consistent lag of about 10 patients. Connecting this value to the results in Tables 1.1 and 1.2, one finds that a loss of roughly 10% of the total information at the time of allocation of the last patient (and a significantly higher loss rate for earlier decisions), corresponds to a loss of only about 5% in terms of the improvement available from each rule.

When the response rate is about 100 times slower that the arrival rate, asymptotically there will be approximately 100 unobserved patients at any point in time. Fortunately, for a sample size of 100, one is quite far from this asymptotic behavior, and approximately 37% of the responses have been observed by the time the last allocation decision must be made. This allows the D2AB to achieve 77% of the relative improvement possible, while the DRPW rule attains only 38%.

While for space reasons this paper has only analyzed problems in which both treatments have uniform priors, similar results hold for more general priors.

5 Conclusions

Because there has been scant research addressing optimal adaptive designs with delayed responses, there are numerous outstanding problems in the area. One might argue that fully optimal designs aren't necessary in practice if good ad hoc options are available. However, without a basis of comparison it is difficult to know how good an ad hoc option is, since asymptotic analyses give only vague information about their behavior for practical sample sizes. Examining the properties of optimal designs can also lead to the development and selection of superior sub-optimal alternatives.

An important concern is the design's robustness. For example, one can evaluate robustness with respect to departures from prior specifications and from the assumption of exponential response times. One way to improve robustness might be to use prior distributions on the response rate parameters. We are also interested in operating characteristics such as the distribution of the objective function, number of allocations to each arm, how the allocations vary with increasingly delay, etc. Some of these issues are examined in Hardwick et al. (2001).

Recall that the goal of this paper is to develop exactly optimal delayed response designs that allow for the use of *any* objective function, not just the bandit objective of maximizing reward (successes). The algorithm presented in Oehmke et al. (2001) has this capability, and in future work we will examine its performance for other objectives. For example, some researchers have considered two-stage models in which the first stage is adaptive and in the second stage all patients are assigned to the arm judged to be best at the end of the first stage. In this situation, the optimal first stage allocation will be nudged closer to equal allocation to insure a better decision for the second stage.

To summarize our findings, we have developed optimal designs for a clinical trial model with Bernoulli observations and exponentially delayed response and patient arrival times. We found that under fairly broad circumstances, the delayed response design performed extremely well compared with the optimal non-delayed algorithm. We also found that the most commonly proposed ad hoc rule for such problems, the DRPW rule, performed significantly less well than the optimal delayed design, which suggests that there is need for better ad hoc strategies.

Acknowledgments: This work was partially supported by National Science Foundation grants DMS-9504980 and DMS-0072910. Parallel computing facilities were provided by the University of Michigan's Center for Parallel Computing.

References

Bandyopadhyay, U. and Biwas, A. (1996), Delayed response in randomized play-the-winner rule: a decision theoretic outlook, *Calcutta Statist. Assoc. Bul.* **46**, 69–88.

Douke, H. (1994), On sequential design based on Markov chains for selecting one of two treatments in clinical trials with delayed observations, *J. Japanese Soc. Comput. Statist.* **7**, 89–103.

Eick, S. (1988), The two-armed bandit with delayed responses, *Ann. Statist.* **16**, 254–264.

Hardwick, J., Oehmke, R. and Stout, Q.F. (2001), Optimal adaptive designs for delayed response models, EECS Technical Report, University of Michigan.

Ivanova, A and Rosenberger, W. (2000), A comparison of urn designs for randomized clinical trials of $k > 2$ treatments, *J. Biopharm. Statist.* **10**, 93–107.

Oehmke, R., Hardwick, J. and Stout, Q.F. (2001), Scalable algorithms for adaptive statistical designs. To appear in *Scientific Programming*.

Wang, X. (2000) A Bandit Process with Delayed Responses, *Statistics and Probability Letters* **48**, 303–307.

Wei, L.J. and Durham, S. (1978), The randomized play-the-winner rule in medical trials, *J. Amer. Statist. Assoc.* **73**, 830–843.