

DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 934

October 1994

**Improvements on the myopic strategy for the Bernoulli
two-armed bandit**

by

Josep Ginebra
Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
08028 Barcelona, Spain
ginebra@eio.upc.es

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

IMPROVEMENTS ON THE MYOPIC STRATEGY FOR THE BERNOULLI TWO-ARMED BANDIT

Josep Ginebra and Murray K. Clayton

Department of Statistics
University of Wisconsin-Madison
1210 W. Dayton Street
Madison, WI 53706
U.S.A.

Key words and Phrases: Bandit problem; Gittins index; sequential optimization; suboptimal strategy; one-armed bandit.

ABSTRACT

We investigate strategies for the Bernoulli two-armed bandit based on a one-armed bandit threshold value (an index analogous to the “Gittins index”), and on upper confidence bounds for θ_i . Using backward induction and the Bayesian viewpoint, we observe that for independent beta priors, these strategies improve on the myopic strategy and get very close to optimal in terms of total expected reward. We also show that the strategy proposed in Berry (1978) is roughly equivalent to the myopic strategy in terms of expected reward.

1. THE BERNOULLI TWO-ARMED BANDIT

Assume that we have two stochastic processes (arms) generating a sequence of 0 – 1 Bernoulli random variables with parameters θ_1 and θ_2 respectively. We are able to make an observation (pull) of one of the arms at each of a possibly infinite number of stages. After choosing one arm we may choose that arm again, or we may switch and choose the other arm. Our objective is to maximize the expected value of the payoff $\sum_{n=1}^{\infty} \alpha_n Z_n$ where

$A = (\alpha_1, \alpha_2, \dots, \alpha_n, \dots)$ is a discount sequence with $\alpha_n \geq 0$ and $0 < \sum_{n=1}^{\infty} \alpha_n < \infty$, and Z_n is the observed random variable at stage n . The discount sequences that are most frequently used are the finite horizon uniform discount sequence $A = (1, 1, 1, \dots, 1, 1, 0, 0, \dots)$ and the geometric discount sequence $A = (1, \alpha, \alpha^2, \alpha^3, \dots)$ with $0 < \alpha < 1$.

Following the Bayesian approach of Berry and Fristedt (1985), the problem is defined once we know the prior probability distribution $G(\theta_1, \theta_2)$ and the discount sequence A . We will address the case where θ_1 and θ_2 are independent and we will denote the prior distribution in terms of the marginal priors: $F_1(\theta_1), F_2(\theta_2)$. A strategies τ will assign to each partial history of observations the arm to be selected at the next stage. We define the worth of τ for the (F_1, F_2, A) -bandit to be $W(F_1, F_2, A; \tau) = E_{\tau}(\sum_{n=1}^{\infty} \alpha_n Z_n)$. The value of the problem is $V(F_1, F_2, A) = \sup_{\tau} W(F_1, F_2, A; \tau)$ and an optimal strategy is one attaining $V(F_1, F_2, A)$. We use the notation $\sigma^{s_i} \varphi^{f_i} F_i$ to denote the distribution of θ_i after having observed s_i successes and f_i failures on arm i , $i = 1, 2$.

The horizon of A is $N = \inf\{n : \alpha_m = 0 \text{ for } m > n\}$. When the horizon is finite, in principle we can use backward induction to determine the arm chosen first by an optimal strategy as well as the value of the problem; in practice, for large N this is too computer intensive both in terms of time and storage. Thus much research has focused on describing structural properties of optimal strategies. For example, when the priors on θ_1 and θ_2 are independent, Berry (1972), building on Bradt et al. (1956), gives conditions under which all Bayes optimal rules "stay on a winner." A second line of research has concentrated on constructing and evaluating good suboptimal strategies (Bather, 1981, Berry, 1978, Jones and Kandeel, 1985 and Wahrenberger et al., 1977).

Many sequential suboptimal strategies have been proposed. Some strategies base the decisions only upon the last outcome. For example, Robbins (1952) proposed the "play on a winner/switch from a loser" strategy and proved that it beats, uniformly in (θ_1, θ_2) , any rule that is not data dependent. The "myopic" one step look-ahead strategy, (τ_{myopic}) , chooses the arm i with the largest posterior expectation for θ_i given all the previous observations. As a generalization, " m -step look ahead" strategies are those that are optimal when we have m observations left. Specifically, Berry (1978) presents a strategy based on the fact that for two-point priors concentrated on (a, b) and (b, a) , the myopic strategy is optimal

(Feldman, 1962). Berry suggests mapping any prior $G(\theta_1, \theta_2)$ down to three parameters: $a = E(\theta_1 | \theta_1 > \theta_2)$, $b = E(\theta_2 | \theta_1 > \theta_2)$ and $r = P[(\theta_1, \theta_2) = (a, b)] = P(\theta_1 > \theta_2)$. He then suggests proceeding myopically for this new two-point prior. These and other suboptimal strategies are discussed in Berry and Fristedt (1985).

There are many criteria of optimality other than the worth used here. Robbins (1952) defined a rule to be asymptotically optimal if the proportion of successes converges almost surely to $\max(\theta_1, \theta_2)$, as N increases. Bayesian optimal and myopic strategies are not asymptotically optimal (Gittins, 1979). There exist a whole host of randomized strategies that are asymptotically optimal but for finite horizons they could be far from best; see, for example, Bather (1981).

2. LAMBDA STRATEGIES AND IMPROVEMENTS

To introduce what we will call “Lambda” strategies, we first have to present the one-armed bandit problem. This is a special case of a two-armed problem where we know $\theta_2 (= \lambda$, say). The problem is completely characterized by the prior distribution F of the unknown θ_1 , the mean λ , and the discount sequence A . Berry and Fristedt (1979, 1985) define a discount sequence to be regular if given $\gamma_m = \sum_{j=m}^{\infty} \alpha_j$ we have that, whenever $\gamma_{m+1} > 0$, $(\gamma_{m+2}/\gamma_{m+1}) \leq (\gamma_{m+1}/\gamma_m)$, for any $m = 1, 2, 3, \dots$. Finite horizon uniform and geometric discounting are regular. Regular discount sequences are important because for them the one-armed bandit problem becomes an optimal stopping problem; we only have to decide when to stop experimenting with arm 1 and switch to arm 2. Also when A is regular with $\alpha_1 > 0$, for every distribution F on $[0, 1]$, there exists a unique $\Lambda(F, A) \in [0, 1]$ such that arm 1 is optimal initially in the (F, λ, A) -bandit if and only if $\lambda \leq \Lambda(F, A)$.

If A is geometric $\Lambda(F, A)$ is called the “Dynamic Allocation Index” or the “Gittins Index”; see Gittins (1979, 1989) and Gittins and Jones (1974). For general regular discounting, where A is not necessarily geometric, we refer to the index $\Lambda(F, A)$ computed from A as the “Lambda index.” For finite horizon discount sequences, this break-even $\Lambda(F, A)$ can be computed to any degree of accuracy by finding the $\lambda = \Lambda(F, A)$ that makes both arms 1 and 2 optimal at the first stage. Since $V(F, \lambda, A)$ is increasing with λ , a bisection method can

be used to compute $\lambda = \Lambda(F, A)$ quite efficiently.

Suppose that instead of using $E(\theta_i|F_i)$ as the criterion of the attractiveness of arm i , the way the myopic strategy does, we use $\Lambda(F_i, A)$. Given the (F_1, F_2, A) -two armed bandit problem with a regular discount sequence, we define a Lambda strategy, (τ_{Lambda}) , to be the one that chooses the arm i such that:

$$\Lambda(F_i, A) = \max_{j=1,2} \Lambda(F_j, A).$$

Whenever the two indices are equal we choose the arm that has been used the least. When A is geometric, Gittins and Jones (1974), building on Bellman (1956), proved that this strategy is the optimal strategy. Berry and Fristedt (1985) prove that if A is regular with $\alpha_1 > 0$ and if this Lambda strategy is optimal for any set of (F_1, F_2) , then A has to be geometric. Given that these $\Lambda(F_i, A)$ indices are easy to compute for finite horizons (even easier than for geometric discounting), they have appeal even if they turn out to be suboptimal.

2.1 The Relationship Between $\Lambda(F, A)$ and $E(\theta|F)$

To compare the performance of the Lambda strategies and the myopic strategies, it is useful to know how $\Lambda(F, A)$ relates to $E(\theta|F)$. We can view $\Lambda(F, A)$ indices as posterior upper bounds of $E(\theta|F)$ that depend on A . Let:

$$\Lambda(F_i, A) = k(F_i, A)E(\theta_i|F_i) = E(\theta_i|F_i) + l(F_i, A).$$

Both $l(F_i, A)$ and $k(F_i, A)$ regulate the importance of sampling to obtain information instead of just looking for the immediate reward $E(X|F_i)$. Gittins (1989) calls $l(F, A)$ the “learning component”. When we have two independent Bernoulli arms with $F_i = \text{beta}(a_i, b_i)$ priors such that $E(X|F_1) = E(X|F_2)$ and A is geometric, Gittins and Wang (1992) prove that the learning component is larger for the arm i for which $a_i + b_i$ is smaller, (and thus when $\theta|F_i$ is more uncertain). In the case where the horizon is finite and uniform, no such general result has been found so far. However for $N = 2$ it is straightforward to show that for any F ,

$$\Lambda(F, A) = E(\theta|F) \frac{\alpha_1 + \alpha_2 E(\theta|\sigma F)}{\alpha_1 + \alpha_2 E(\theta|F)} = E(\theta|F) + \alpha_2 \frac{\text{Var}(\theta|F)}{\alpha_1 + \alpha_2 E(\theta|F)}.$$

Thus for the family of priors F with the same expectation and $N = 2$, $\Lambda(F, A)$ is a linear function of $\text{Var}(\theta|F)$. The smaller $\text{Var}(\theta|F)$, the more informative F is about θ , and the closer $\Lambda(F, A)$ is to $E(\theta|F)$ and thus the closer the Lambda strategy is to the myopic strategy. For larger N , it is possible to express $\Lambda(F, A)$ as a complex function of the central moments of F of order less than or equal to N . Ginebra (1993) gives $\Lambda(F, A)$ in a closed form for $N = 2, 3, 4$ and any regular discount sequence, generalizing the results of Bradt et al. (1956) for uniform discounting.

For the two-point prior distribution, Ginebra (1993) shows that for most of the cases tried there, the Lambda strategy yields total expected payoffs larger than the myopic strategy, but sometimes the myopic strategy does better than Lambda. When $N = 1$, both the myopic and Lambda strategies agree with the optimal strategy. When $N = 2$ and the priors are independent, Ginebra (1993) proves that: $W(F_1, F_2, A; \tau_{\text{lambda}}) \geq W(F_1, F_2, A; \tau_{\text{myopic}})$. We conjecture that for independent arms and regular discounting, the worth of the Lambda strategy is larger than the worth for the myopic one for any N . This conjecture is strengthened by the results in Section 4 below and in Ginebra and Clayton (1994b).

2.2 The $\Lambda^{(1)}(F, A)$ and $\Lambda^{(2)}(F, A)$ Indexed Strategies

In Section 4, we see that the amount of improvement we get by using the Lambda strategies instead of the myopic one is considerable. Looking for indices that improved on $\Lambda(F, A)$, we checked what was going wrong in the backward induction process when the Lambda strategy chose the arm that was not optimal. For all the priors and horizons we tried, whenever $\Lambda(F_i, A)$ made the wrong decision, $\Lambda(\sigma F_i, A^{(1)})$ would have made the right one, with $A^{(n)} = (\alpha_{n+1}, \alpha_{n+2}, \alpha_{n+3}, \dots)$. We combined these two indices into one index by mirroring the way $\Lambda(F, A)$ itself improves upon $E(\theta|F)$ when $N = 2$. We assumed that this should be done in a way such that if F collapses to a point mass at λ , the new index should coincide with $\Lambda(F, A)$. Two alternatives are:

$$\Lambda^{(1)}(F, A) = \Lambda(F, A) \frac{\alpha_1 + \alpha_2 \Lambda(\sigma F, A^{(1)})}{\alpha_1 + \alpha_2 \Lambda(F, A)}$$

$$\Lambda^{(2)}(F, A) = \Lambda(F, A^{(1)}) \frac{\alpha_1 + \alpha_2 \Lambda(\sigma F, A^{(1)})}{\alpha_1 + \alpha_2 \Lambda(F, A^{(1)})}$$

Actually $\Lambda^{(2)}(F, A)$ coincides with $\Lambda(F, A)$ for $N = 2$. In Section 4, when we test these new $\Lambda^{(i)}(F, A)$ indexed strategies, we use $\Lambda(\sigma^{s_i} \varphi^{f_i} F, A^{(n)})$ as the index at stages $n = N - 2, N - 1$ and we switch to either $\Lambda^{(1)}(\sigma^{s_i} \varphi^{f_i} F, A^{(n)})$ or $\Lambda^{(2)}(\sigma^{s_i} \varphi^{f_i} F, A^{(n)})$ for all the previous stages.

2.3 The Truncated $\Lambda(F, A')$ Indexed Strategies

One alternative simple strategy would be the one that chooses the arm i that has the largest $\Lambda(F_i, A')$, where A' is the truncated $N = 2$ horizon discount sequence that we get from taking α_1 and α_2 from the true discount sequence A , and setting the other values equal to 0. In general, suppose that we have taken n observations so far and we have observed s_i successes and f_i failures with arm i , ($s_i + f_i < n$). The idea is to improve on the myopic strategy without using backward induction, by pulling the arm i with the largest $\Lambda(\sigma^{s_i} \varphi^{f_i} F_i, A'_{(n)})$ index, defined as:

$$\Lambda(\sigma^{s_i} \varphi^{f_i} F_i, A'_{(n)}) = E(\theta | \sigma^{s_i} \varphi^{f_i} F_i) + \alpha_{n+2} \frac{\text{Var}(\theta | \sigma^{s_i} \varphi^{f_i} F_i)}{\alpha_{n+1} + \alpha_{n+2} E(\theta | \sigma^{s_i} \varphi^{f_i} F_i)},$$

with $A'_{(n)} = (\alpha_{n+1}, \alpha_{n+2}, 0, 0, \dots)$. This strategy is evaluated in Section 4. While we could do better using closed expressions for the Lambda index truncating at $N = 3, 4$ or using even better lower bounds for $\Lambda(F, A)$, the ease of computation of $\Lambda(F, A')$ makes it appealing.

3. UPPER BOUND STRATEGIES

The main advantage of myopic strategies over Lambda strategies is that they are easier to implement. We now define new strategies that are as simple as the myopic one and tend to get larger worths than myopic most of the time.

If we knew θ_1 and θ_2 , the optimal strategy would trivially select the arm i with the largest θ_i . The myopic rule estimates these two parameters through their posterior expectation and selects the arm with the largest estimate. Lai (1987) presents a class of strategies that experiment at the arm i with the largest upper confidence bound for the estimated parameter for that arm. He defines the upper bound in a way that allows him to prove the asymptotic optimality of those strategies when the distribution on the arms belongs to the exponential family and has its prior on the natural parameter space. This setting does not include the Bernoulli model in the current form since θ is not the natural parameter. Here we adapt his idea by proposing to use a simple normal approximation for the posterior confidence interval of θ_i , i.e.:

$$\text{UB}(F_i) = E(\theta_i|F_i) + K\sqrt{\text{Var}(\theta_i|F_i)}.$$

At each stage, the strategy chooses the arm for which $\text{UB}(F_i)$ is the largest. K is a parameter that regulates the amount of “learning” that our strategy does. A large K will tend to work better for large horizon discount sequences where we have time to exploit what we learn from pulling unfavorable arms that are less well known. If we start with the prior distribution $F_i = \text{beta}(a_i, b_i)$ for θ_i , this index after s_i successes and f_i failures with arm i will be:

$$\text{UB}(F_i) = \frac{a_i + s_i}{a_i + b_i + s_i + f_i} + K\sqrt{\frac{(a_i + s_i)(b_i + f_i)}{(a_i + b_i + s_i + f_i)^2(a_i + b_i + s_i + f_i + 1)}}$$

Lai (1987) presents an allocation rule φ^* that approximates his asymptotically optimal rule, and applies it to the Bernoulli case. Table 1 gives the worths normalized by N for the optimal strategy, Lai’s strategy and the upper bound strategy defined in this section. The worths for Lai’s φ^* strategy are taken from his paper. We computed the other worths exactly using backward induction with $K = 0.676$; when we face a tie, we choose the arm pulled the least and when used the same number of times we choose arm 2. (K was chosen arbitrarily to be a quartile of the normal distribution).

Even though all the strategies have similar total expected rewards, the new upper bound strategies appear to do better than Lai’s for these four combinations of independent beta

| | Parameters for the beta priors | | | |
|-------------------|--------------------------------|------------|------------|------------|
| STRATEGY | (1,1)(1,1) | (2,6)(2,6) | (4,4)(4,4) | (6,2)(6,2) |
| Lai's φ^* | 0.634 | 0.300 | 0.558 | 0.805 |
| UB | 0.6378 | 0.3029 | 0.5645 | 0.8064 |
| Optimal | 0.6399 | 0.3034 | 0.5650 | 0.8066 |

Table 1: $W(F_1, F_2, A; \tau)$ divided by N when τ is the Bayesian optimal strategy, the strategy φ^* proposed in Lai (1987) or the new upper bound strategy with $K = 0.676$. The discount is uniform with $N = 50$ and the priors are independent beta.

priors.

3.1 Influence of K ; Adaptive Upper Bound Strategies

Letting K vary yields a family of upper bound strategies that include the myopic strategy when $K = 0$. When K is large, the strategy always pulls the arm known the least as measured through the posterior standard deviation of θ . Figure 1 shows how the worth of these upper bound strategies changes with K for the uniform discount sequence with $N = 40$ and for four different combinations of independent beta priors. Each dot represents the exact worth of one strategy belonging to this upper bound family as computed through backward induction. The two horizontal lines are the worths obtained by the myopic and the optimal strategies. (The worth for the Lambda strategy overlaps the optimal one). We can see that for small K , the worth increases roughly with K though not strictly monotonically. It peaks close to the optimal value before $K = 2$ and for large K , the upper bound strategy behaves worse than myopic.

Notice that the improvement of the upper bound family over myopic is especially large for the $\text{beta}(1, 1) \times \text{beta}(6, 6)$ case. Under this prior, sampling starts with the assumption that the two arms have the same expectation on θ_i but that arm 2 is known better than arm 1. Since we choose arm 1 whenever $E(\theta_1|F_1) - E(\theta_2|F_2) > K(\sqrt{\text{Var}(\theta_2|F_2)} - \sqrt{\text{Var}(\theta_1|F_1)})$, the best K will relate the amount of immediate expected payoff we should be willing to give up to the difference in the standard deviation for the two θ_i 's. The best strategy in this family fails to be optimal in part because we keep K constant for all N stages. Lambda strategies improve on that by using, at each stage n , the best indexed strategy out of a richer family of indices that are functions of the $N - n$ first moments. To try to match the behavior of

$\Lambda(F_i, A)$ without actually computing it we could recompute the best $K = K_n^*$ at each stage n for the remaining horizon.

For this Bernoulli case, finding the best $K = K^*$ for a given problem does not make sense since it is easier to implement the optimal strategy. In Ginebra and Clayton (1994a) we propose an extension of this idea to the response surface bandit; there we find the best upper bound strategy for the given problem by estimating through simulation their worths as a function of K .

4. COMPARISON OF STRATEGIES

Jones (1976) and Jones and Kandeel (1985) compare myopic and play the winner strategies with the optimal strategy for two pairs of independent beta priors. Robinson (1983), using simulation, estimated the worth conditional on (θ_1, θ_2) for several strategies including the one that uses the Gittins index computed as if the discount sequence was geometric with $\alpha = 0.99, 0.995$ and 0.9999 ; Wang (1991) does another comparison, but with the exception of Berry (1978), we are unaware of anyone computing the exact Bayesian worth for suboptimal strategies using backward induction.

If we have an (F_1, F_2, A) two-armed Bernoulli bandit and we have had s_i successes and f_i failures on arm i ($i = 1, 2$), the backward induction equation for any strategy τ is:

$$W(\sigma^{s_1} \varphi^{f_1} F_1, \sigma^{s_2} \varphi^{f_2} F_2, A^{(n)}; \tau) = p_\tau W^{(1)}(\sigma^{s_1} \varphi^{f_1} F_1, \sigma^{s_2} \varphi^{f_2} F_2, A^{(n)}; \tau) \\ + (1 - p_\tau) W^{(2)}(\sigma^{s_1} \varphi^{f_1} F_1, \sigma^{s_2} \varphi^{f_2} F_2, A^{(n)}; \tau)$$

where $A^{(n)} = (\alpha_{n+1}, \alpha_{n+2}, \dots)$ and p_τ is 1 when τ selects arm 1 and 0 when it selects arm 2. $W^{(i)}$ is the worth obtained by pulling arm i first and then proceeding as dictated by τ ,

$$W^{(i)}(\sigma^{s_1} \varphi^{f_1} F_1, \sigma^{s_2} \varphi^{f_2} F_2, A^{(n)}; \tau) \\ = E(\theta | \sigma^{s_i} \varphi^{f_i} F_i) \{ \alpha_{n+1} + W(\sigma^{s_i+1} \varphi^{f_i} F_i, \sigma^{s_i} \varphi^{f_i} F_i, A^{(n+1)}; \tau) \} \\ + E(1 - \theta | \sigma^{s_i} \varphi^{f_i} F_i) \{ W(\sigma^{s_i} \varphi^{f_i+1} F_i, \sigma^{s_i} \varphi^{f_i} F_i, A^{(n+1)}; \tau) \}$$

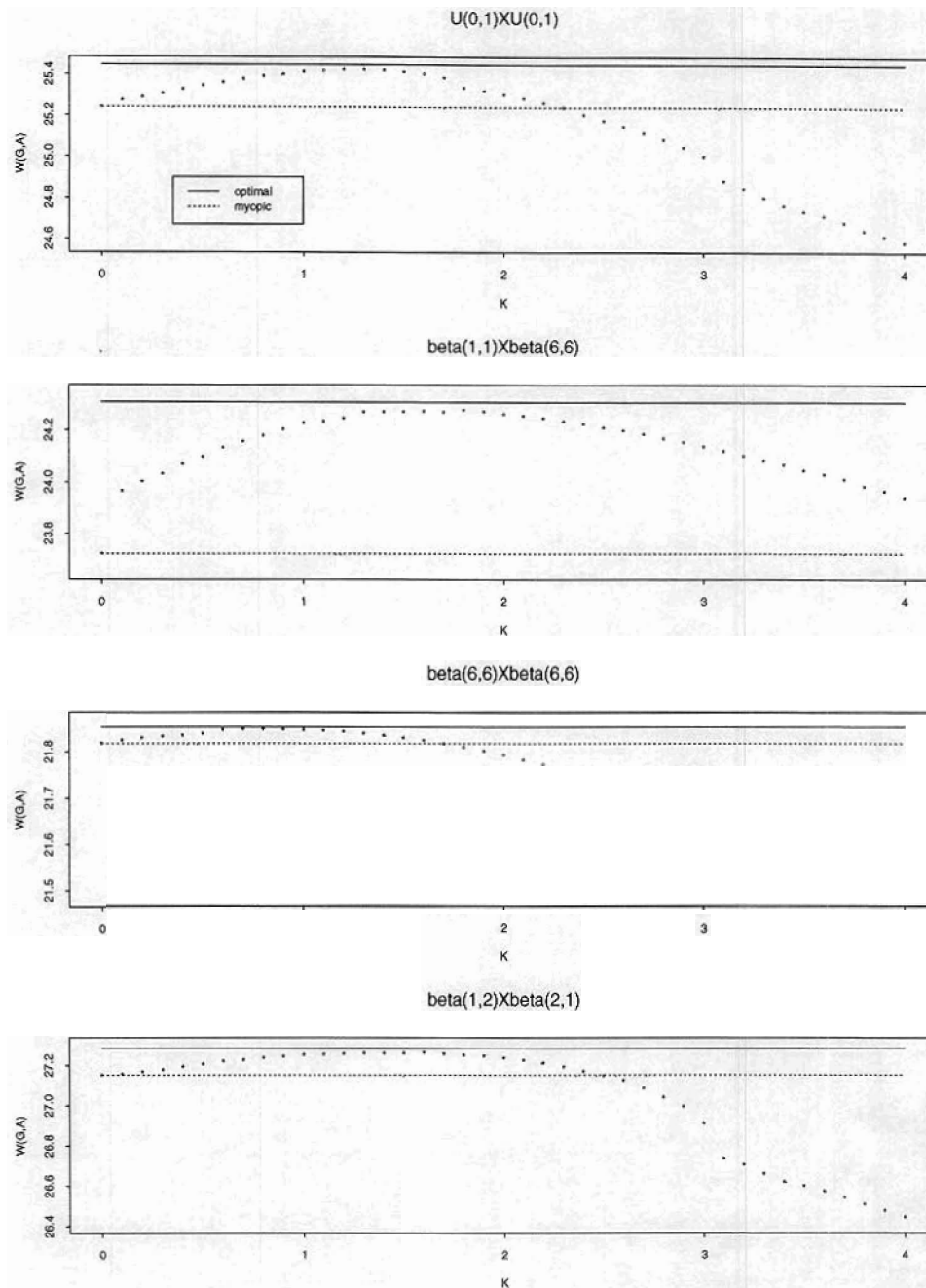


Figure 1: Exact worth for the upper bound strategies as a function of K ; when tied, they choose the arm pulled the least and when used the same number of times they randomize. The two horizontal lines correspond to the total expected reward for the optimal and the myopic strategies. The discounting is uniform with $N = 40$ and the priors are independent beta.

with $n = s_1 + s_2 + f_1 + f_2$ and \bar{i} meaning the arm that is not arm i . We computed the exact worth for the strategies in Sections 1, 2.1, 2.2, 2.3 and 4 using these equations. For the upper bound strategies we used $K = 0.676$. Whenever the strategies face a tie on both arms, we chose the arm used the least and if they have been used the same number of times, we used arm 2. In Table 2 we present the worths normalized by N for all the strategies listed plus the ones corresponding to the strategy described in Berry (1978) (in his case, he randomized ties). We use a subset of the combinations of beta priors used by Berry (1978) and we use uniform discounting with $N = 50$. Ginebra (1993) has the results for the complete set of priors used in Berry (1978) as well as for $N = 25$.

The most important fact that appears in Table 2 is that the myopic strategy yields worths that are quite close to the worths yielded by the optimal one. That closeness obscures the improvement we get using the upper bound and Lambda family of strategies instead of the myopic one. For the priors in the table, we observe that the Lambda and upper bound strategies always have larger worths than the myopic one and that the strategy from Berry (1978) is roughly equivalent to the myopic strategy. Also, the two indexed strategies presented in Section 2.2, based on the $\Lambda^{(1)}(F, A)$ and $\Lambda^{(2)}(F, A)$ indices, yield larger worths than the Lambda strategy; the one based on $\Lambda^{(1)}(F, A)$ has larger total expected payoffs than the one based on $\Lambda^{(2)}(F, A)$ in 10 cases, and smaller in 25 cases, with 1 tie. The worth of the Lambda strategy agrees with the optimal value at least up to the fourth significant digit while the strategies using $\Lambda^{(1)}(F, A)$ and $\Lambda^{(2)}(F, A)$ obtain worths that agree with the optimal value up to the fifth significant digit.

For these priors, Ginebra (1993) shows that the myopic strategy is optimal for N 's smaller than 6 to 8 while the upper bound strategies, Lambda strategies and their modifications sometimes are optimal for N as large as 25. Table 3 presents the worths divided by N for the same strategies and uniform discount sequences with $N < 50$ and beta(1,1) (uniform) independent priors. We included the m -step ahead strategies with $m = 3, 5$. In general, we can rank the strategies in terms of worth from best to worst as: $\Lambda^{(i)}(F, A)$ indexed strategies, Lambda, 5-step ahead, upper bound strategies with $K = 0.676$, 3-step ahead, the strategies that use the truncated $\Lambda(F, A')$ index from Section 2.3 and the myopic and Berry (1978) strategies. We have observed this same ranking whenever we have used pairs of beta priors,

with upper bound doing sometimes better and sometimes worse than three and five step ahead strategies. Only for very small N do the upper bound strategies behave sometimes worse than myopic, and $\Lambda^{(i)}(F, A)$ indexed strategies worse than Lambda strategies.

Overall we would say that when N is small and we have independent arms and beta priors, we can implement the optimal strategy. For intermediate and large N we can use any of these easy-to-use suboptimal strategies knowing that we are not losing much. This point is strengthened in Ginebra and Clayton (1994b), where it is shown that Bayesian optimal, Lambda and myopic strategies are also very close conditional on (θ_1, θ_2) .

| (a_1, b_1) | (a_2, b_2) | | | | | | | |
|--------------|--------------|----------|-----------|----------|----------|----------|----------|----------|
| | (6,2) | (2,1) | (1/2,1/2) | (1,1) | (2,2) | (6,6) | (1,2) | (2,6) |
| (6,2) | .805 | .791 | .774 | .763 | .756 | .750 | .749 | .748 |
| | .8052874 | .7887342 | .7593997 | .7567740 | .7547431 | .7528196 | .7502156 | .7500091 |
| | .8054064 | .7902691 | .7890280 | .7591964 | .7552628 | .7528407 | .7503527 | .7500110 |
| | .8063804 | .7961671 | .7789228 | .7646698 | .7572634 | .7529968 | .7507788 | .7500193 |
| | .8065852 | .7990069 | .7924898 | .7704806 | .7587264 | .7530484 | .7516007 | .7500258 |
| | .8066092 | .7990503 | .7924990 | .7706105 | .7587972 | .7530573 | .7516248 | .7500270 |
| | .8066083 | .7990504 | .7925033 | .7706123 | .7587977 | .7530568 | .7516327 | .7500270 |
| | .8066105 | .7990530 | .7925316 | .7706223 | .7588000 | .7530578 | .7516444 | .7500270 |
| (2,1) | .770 | .732 | .719 | .707 | .696 | .680 | .666 | .666 |
| | .7712827 | .7324578 | .7191628 | .7082695 | .6961745 | .6798328 | .6702370 | .6702370 |
| | .7717935 | .7368841 | .7203609 | .7086427 | .6985153 | .6804617 | .6702878 | .6702878 |
| | .7740095 | .7427235 | .7235377 | .7100510 | .6996022 | .6820156 | .6704767 | .6704767 |
| | .7749467 | .7486015 | .7262908 | .7110406 | .7003095 | .6836812 | .6705802 | .6705802 |
| | .7749996 | .7486429 | .7263434 | .7110900 | .7003445 | .6837311 | .6705921 | .6705921 |
| | .7750001 | .7486478 | .7263492 | .7110911 | .7003443 | .6837365 | .6705916 | .6705916 |
| | .7750012 | .7486687 | .7263606 | .7110921 | .7003453 | .6837484 | .6705923 | .6705923 |
| (1/2,1/2) | .670 | .649 | .632 | .615 | .575 | .535 | .535 | .535 |
| | .6737746 | .6593588 | .6461935 | .6344218 | .5781550 | .5429069 | .5429069 | .5429069 |
| | .6752072 | .6603792 | .6478735 | .6365903 | .5801008 | .5447893 | .5447893 | .5447893 |
| | .6774135 | .6622558 | .6422200 | .6390776 | .5817093 | .5466024 | .5466024 | .5466024 |
| | .6802168 | .6644922 | .6532809 | .6451395 | .5836602 | .5483707 | .5466443 | .5466443 |
| | .6802661 | .6645373 | .6533270 | .6451728 | .5837067 | .5483710 | .5466437 | .5466437 |
| | .6802695 | .6645401 | .6533289 | .6451766 | .5837102 | .5483712 | .5466437 | .5466437 |
| | .6802907 | .6645596 | .6533526 | .6451960 | .5837242 | .5483724 | .5466475 | .5466475 |
| (1,1) | .633 | .612 | .592 | .555 | .520 | .520 | .520 | .520 |
| | .6345883 | .6186463 | .5994570 | .5558664 | .5231916 | .5231916 | .5231916 | .5231916 |
| | .6354623 | .6194267 | .6023305 | .5564836 | .5242949 | .5242949 | .5242949 | .5242949 |
| | .6377594 | .6214397 | .6063087 | .5580583 | .5250167 | .5250167 | .5250167 | .5250167 |
| | .6398744 | .6232715 | .6106269 | .5596810 | .5255651 | .5255651 | .5255651 | .5255651 |
| | .6399212 | .6233327 | .6106936 | .5597329 | .5255992 | .5255992 | .5255992 | .5255992 |
| | .6399230 | .6233352 | .6106956 | .5597370 | .5255983 | .5255983 | .5255983 | .5255983 |
| | .6399342 | .6233423 | .6106997 | .5597448 | .5256004 | .5256004 | .5256004 | .5256004 |
| (2,2) | .595 | .571 | .535 | .508 | .508 | .508 | .508 | .508 |
| | .5956293 | .5737171 | .5343504 | .5099769 | .5099769 | .5099769 | .5099769 | .5099769 |
| | .5961095 | .5755855 | .5354518 | .5100921 | .5100921 | .5100921 | .5100921 | .5100921 |
| | .5978229 | .5784209 | .5382115 | .5105280 | .5105280 | .5105280 | .5105280 | .5105280 |
| | .5990290 | .5803591 | .5410411 | .5107736 | .5107736 | .5107736 | .5107736 | .5107736 |
| | .5990839 | .5804364 | .5411054 | .5107970 | .5107970 | .5107970 | .5107970 | .5107970 |
| | .5990853 | .5804366 | .5411064 | .5107964 | .5107964 | .5107964 | .5107964 | .5107964 |
| | .5990873 | .5804382 | .5411204 | .5107974 | .5107974 | .5107974 | .5107974 | .5107974 |
| (1,2) | .548 | .515 | .499 | .433 | .377 | .377 | .377 | .377 |
| | .5481328 | .5075315 | .5004853 | .4345095 | .3774645 | .3774645 | .3774645 | .3774645 |
| | .5482739 | .5105909 | .5005725 | .4351607 | .3795335 | .3795335 | .3795335 | .3795335 |
| | .5489731 | .5176908 | .5009109 | .4368658 | .3808392 | .3808392 | .3808392 | .3808392 |
| | .5491553 | .5258639 | .5011773 | .4385768 | .3818840 | .3818840 | .3818840 | .3818840 |
| | .5491818 | .5259728 | .5012069 | .4386384 | .3819390 | .3819390 | .3819390 | .3819390 |
| | .5491812 | .5259761 | .5012074 | .4386401 | .3819370 | .3819370 | .3819370 | .3819370 |
| | .5491830 | .5259856 | .5012077 | .4386475 | .3819406 | .3819406 | .3819406 | .3819406 |
| (2,6) | .302 | .302 | .302 | .302 | .302 | .302 | .302 | .302 |
| | .3018945 | .3018945 | .3018945 | .3018945 | .3018945 | .3018945 | .3018945 | .3018945 |
| | .3020908 | .3020908 | .3020908 | .3020908 | .3020908 | .3020908 | .3020908 | .3020908 |
| | .3029237 | .3029237 | .3029237 | .3029237 | .3029237 | .3029237 | .3029237 | .3029237 |
| | .3033579 | .3033579 | .3033579 | .3033579 | .3033579 | .3033579 | .3033579 | .3033579 |
| | .3034076 | .3034076 | .3034076 | .3034076 | .3034076 | .3034076 | .3034076 | .3034076 |
| | .3034063 | .3034063 | .3034063 | .3034063 | .3034063 | .3034063 | .3034063 | .3034063 |
| | .3034094 | .3034094 | .3034094 | .3034094 | .3034094 | .3034094 | .3034094 | .3034094 |

Table 2: $W(F_1, F_2, A; \tau)$ divided by N for eight strategies; when τ faces a tie, it chooses the arm pulled the least and when used the same number of times it chooses arm 2. We use 36 combinations of independent beta priors and uniform discounting with $N = 50$. The results for the strategy in Berry (1978) are taken from that paper.

| N | Myopic | $\Lambda(F, A')$ | Up.bd. | 3-Step | 5-Step | $\Lambda(F, A)$ | $\Lambda^{(1)}(F, A)$ | $\Lambda^{(2)}(F, A)$ | Optimal |
|----|---------|------------------|---------|---------|---------|-----------------|-----------------------|-----------------------|----------|
| 4 | .569444 | .569444 | .569444 | .569444 | .569444 | .569444 | .5694444 | .5694444 | .5694444 |
| 8 | .594940 | .594940 | .594668 | .594940 | .594940 | .594940 | .5949405 | .5949405 | .5949405 |
| 12 | .607229 | .607204 | .607557 | .607639 | .607643 | .607622 | .6076888 | .6076834 | .6077007 |
| 16 | .614725 | .614805 | .615445 | .615513 | .615669 | .615823 | .6158211 | .6158331 | .6158541 |
| 20 | .619775 | .619974 | .620887 | .620932 | .621226 | .621470 | .6215480 | .6215630 | .6215632 |
| 24 | .623440 | .623770 | .624933 | .624921 | .625335 | .625770 | .6258590 | .6258650 | .6258705 |
| 28 | .626244 | .626686 | .628076 | .627988 | .628515 | .629191 | .6292094 | .6292237 | .6292365 |
| 32 | .628458 | .628999 | .630598 | .631052 | .631917 | .631195 | .6319586 | .6319602 | .6319652 |
| 36 | .630260 | .630890 | .632674 | .632408 | .633132 | .634149 | .6342296 | .6342349 | .6342445 |
| 40 | .631756 | .632465 | .634416 | .634063 | .634871 | .636093 | .6361635 | .6361666 | .6361719 |
| 44 | .633020 | .633800 | .635896 | .635467 | .636346 | .637758 | .6378219 | .6378224 | .6378311 |
| 48 | .634102 | .634946 | .637182 | .636670 | .637615 | .639212 | .6392647 | .6392706 | .6392782 |

Table 3: $W(F_1, F_2, A; \tau)$ divided by N ; when τ faces a tie, it chooses the arm pulled the least and when used the same number of times it pulls arm 2. The priors are $U(0, 1)$ and the discounting is uniform with $N = 4, 8, 12, 16, 20, 24, 28, 32, 36, 44$ and 48 .

BIBLIOGRAPHY

- Bather, J.A. (1981) Randomized allocation of treatments in sequential experiments (with discussion). *Journal of the Royal Statistical Society B* 43:265-292.
- Bellman, R. (1956) A problem in the sequential design of experiments. *Sankhya A* 16: 221-229.
- Berry, D.A. (1972) A Bernoulli two-armed bandit. *The Annals of Mathematical Statistics* 43:871-897.
- Berry, D.A. (1978) Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association*, 73:339-345.
- Berry, D.A. and Fristedt, B. (1979) Bernoulli one-armed bandits-Arbitrary discount sequences. *Annals of Statistics*, 7: 1086-1105.
- Berry, D.A. and Fristedt, B. (1985) *Bandit Problems; Sequential Allocation of Experiments*. London; Chapman and Hall.
- Bradt, R.N., Johnson, S.M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics* 27:1060-1074.
- Feldman, D. (1962) Contributions to the 'two-armed bandit' problem. *The Annals of Mathematical Statistics* 33:847-856.

- Ginebra, J (1993) *Strategies for Response Surface Bandits and Bernoulli bandits*. Ph.D. thesis, University of Wisconsin, Madison, U.S.A.
- Ginebra, J and Clayton, M.K. (1994a) The response surface bandit. *Technical report 932*, Department of Statistics, University of Wisconsin, Madison.
- Ginebra, J and Clayton, M.K. (1994b) Small sample frequentist properties of Bernoulli two-armed bandit Bayesian strategies. *Technical report 933*, Department of Statistics, University of Wisconsin, Madison.
- Gittins, J.C. (1979) Bandit processes and dynamic allocation indices (with discussion). *Journal of the Royal Statistical Society B* 41:148-177.
- Gittins, J.C. (1989) *Multi-armed Bandit Allocation Indices*. New York: John Wiley and Sons.
- Gittins, J.C. and Jones, D.M. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (eds. Gani, J. et al) North-Holland: Amsterdam. pp. 241-266.
- Gittins, J.C. and Wang, Y.G. (1992) The learning component of dynamic allocation indices. *Annals of Statistics* 20:1625-1636.
- Jones, P.W. (1976) The two-armed bandit. *Biometrika* 62:523-524.
- Jones, P.W. and Kandeel, H.A. (1985) Numerical investigation of the two armed bandit. In *Mathematical Learning Models-Theory and Algorithms* (eds. U. Herkenrath, D. Kalin and W. Vogel). Springer-Verlag: Berlin pp. 101-107.
- Lai, T.L. (1987) Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics* 15:1091-1114.
- Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58:527-535.

Robinson, D. (1983) A comparison of sequential treatment allocation rules. *Biometrika* **70**:492-495.

Wang, Y.G. (1991) Gittins indices and constrained allocation in clinical trials. *Biometrika* **78**:101-111.

Wahrenberger, D.L., Antle, C.E. and Klimko, L.A. (1977) Bayesian rules for the two-armed bandit problem. *Biometrika* **64**:172-174.