# Choosing the Best Arm with Guaranteed Condence

Mohammad Javad Azizi
Sheldon M. Ross
Zhengu Zhang
Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089
azizim@usc.edu
smross@usc.edu
zhan892@usc.edu

The second author's work was supported by, or in part, by the National Science Foundation under contract/grant CMMI2132759.

Abstract

We consider the problem of nding, through adaptive sampling, which of n populations (arms) has the largest mean. Our objective is to determine a rule which identies the best arm with a xed mini-mum condence using as few observations as possible.We study such

problems when the population distributions are either Bernoulli or normal. We take a Bayesian approach that assumes that the unknown means are the values of independent random variables having a common specied distribution. We propose to use the classical vector at a time rule, which samples each remaining arm once in each round, eliminating arms whose cumulative sum falls k below that of another arm. We show how this rule can be implemented and analyzed in our Bayesian setting and how it can be improved by early elimination. We also propose and analyze a variant of the classical play the winner algorithm. Numerical results show that these rules perform quite well, even when considering cases where the set of means do not look like they come from the specied prior.

Keywords: best arm identication, vector at a time, Bayesian format

# 1 Introduction

Let F(x) be a family of distributions indexed by its mean : Suppose there are n populations, and that each new observation from population i is independent of all previous observations and is the value of a random variable with distribution $F_i$; where $_1;::::;_n$ are unknown. Furthermore, suppose that our objective is to decide which population has the largest mean. It is supposed that a decision is made at each stage as to which population to next take an observation from, with the decision made according to some rule which eventually calls for stopping and declaring which population has the largest mean. Our objective is to determine a rule that makes a decision in a relatively small expected time, subject to the condition that its probability of making a correct choice is at least : We will study such models both when the population distributions are Bernoulli and when they are normal with a xed variance. In the Benoulli case, we suppose that the result of an observation is either a success or a failure, and the objective is to nd the population having the largest success probability.

These type of models have many applications. Foremost is probably in clinical trials to determine which of several medical approaches (e.g., drugs, treatments, and procedures) yields the best results. Here a population would refer to a particular approach, with its use resulting in either a success (suitably dened) or not. Another application is to online advertising, where a decision maker is trying to decide which of n dierent advertisements to utilize. For instance, the advertisements might relate to army recruitment, and a success might refer to a subsequent clicking on the advertisement to obtain additional information. Another application is to choose among different methods for teaching a particular skill. Each day, a method can be used on a group of students, with the students being tested at the end of the day with each test resulting in a score which would be pass (1) or fail (0) in the Bernoulli case, and numerical in the normal case.

It should be noted that this problems has been studied for quite some time. However, the early work, such as in [3], [4], [9], [10], [11], [15], [16], was done under the assumption that the dierence between the largest and sec-ond largest population mean was at least some known positive value. More recent work, such as [1], [5], [6], [7], [8], [13], do not make this assump-tion. What primarily distinguishes our models from others considered in the literature is that we take a Bayesian approach that supposes the unknown means are the respective values of independent and identically distributed

(iid) random variables having a specied prior distribution F: Moreover, we present numerical evidence that the rules obtained when we assume that F is the uniform (0; 1) distribution tend to perform well for most sets of success probabilities.

In Section 2 we consider the Bernoulli case where we want to nd rules whose implementation results in a relatively small expected number of observations needed before a decision is made, subject to the condition that the rule results in the probability of a correct choice being at least some specied value . We reconsider the classical \vector at a time" (VT) rule, which is such that each alive population is sampled once in each round. At the end of a round, any population whose cumulative number of successes is k less than another is no longer alive. The process ends when only one population is alive. The appropriate value of k that results in the probability of a correct choice being at least was determined in [15] and [16] under the assumption that the dierence between the largest and second largest population mean is at least some specied positive value d > 0: We show, in this section, how this rule can be implemented and analyzed in our Bayesian setting. In particular, we present a lower bound as well as an accurate approximation for the probability that VT using critical value k makes the correct decision, as well as an approximation for the mean number of observations it takes. These bounds and approximations can be evaluated by a simple simulation, with each simulation run requiring either 2 or 3 random numbers. We also present some numerical evidence indicating that the version of this rule that results when we assume a uniform (0; 1) prior appears to outperform more recently proposed rules, even when the means do not come from the uniform prior.

In Section 3, we consider improving the VT rule by allowing for early elimination if a population is j behind another after j rounds. (That is, if one population has had 0 successes and another j successes in their rst j observations, then the former is no longer alive.) We show how to determine the probability that the population with largest mean is eliminated early, as well as the mean number of the non-best populations that are eliminated early.

In Section 4 we consider another classical rule: \play the winner" (PW). Our variant of the PW rule is such that in all but the nal round each alive population is continually sampled until it has a failure, where a population

is no longer alive if after a round its cumulative number of successes is at least k smaller than that of some other arm. This variant diers from the classical model of [15] which declared a population dead if at some point -not necessarily at the end of a round - it has k fewer successes than another population. We show how to analyze this rule in the Bayesian setting.

In Section 5 we consider the case of normal populations, where population distributions are all normal distributions with some xed variance $^2$; and where the standard normal is the prior distribution on the means. Among other things, by using that a normal conditioned to be positive has an increasing failure rate, we improve upon known bounds for the probability that a random walk of normal random variables reaches a before falling as low as b for given positive numbers a; b:

Section 6 gives the paper's conclusions.

# 2 The Vector at a Time Rule in the Bernoulli Case

Suppose there are n Bernoulli populations with respective means $p_1; : : : ; p_n$, and that at each stage we are allowed to make an observation from a population of our choice, stopping these observations at some point and making a decision as to which population has the largest mean. As noted earlier, we suppose that $p_1; : : : ; p_n$ are the respective values of iid random variables having a specied distribution F: Subject to the constraint that the policy used results in the probability of a correct choice being at least ; the objective is to choose a policy whose mean number of observations is relatively small.

Denition The vector at a time (VT) rule, introduced in [3], is dened as follows. For a given positive integer k, depending on the desired accuracy ; the policy is as follows.

Initially all populations are alive

A round consists of a single observation from each alive population.

At the end of a round, a population is no longer alive if its cumulative number of successes is k less than that of another population.

If only one population is alive, stop and declare it best. Otherwise perform a new round.

Let $p_{[1]} > p_{[2]} > \cdots > p_{[n]}$ be the ordered values of the unknown means $p_1, \ldots, p_n$; and let $C$ be the event that the correct choice is made. Under the assumption that there is a known positive value $d$ such that $p_{[1]} - p_{[2]} > d$; it was shown in [15] how to determine $k$ so that $P(C)$ : We now show how this can be done when $p_1, \ldots, p_n$ are the values of independent and identically distributed random variables having distribution $F$.

Notation: We use the notation $I\{A\}$ to be the indicator of the event $A$; equal to 1 if $A$ occurs and to 0 otherwise. Also, we use the notation $X =_{st} Y$ to indicate that $X$ and $Y$ have the same distribution.

Lemma 1: Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having distribution $F$, and let $U, U_1, \ldots, U_n$ be independent uniform $(0, 1)$ random variables. Then

$$\max_i X_i =_{st} F^{-1}(U^{1/n})$$

Proof:

$$\begin{aligned} \max_i X_i &=_{st} \max(F^{-1}(U_1), \ldots, F^{-1}(U_n)) \\ &= F^{-1}(\max_{i=1,\ldots,n} U_i) \\ &=_{st} F^{-1}(U^{1/n}) \quad \blacksquare \end{aligned}$$

We now show how to bound $P(C)$, the probability that the correct population is chosen. To begin, suppose that in each round we take an observation from each population, even those that may be dead. Let 0 be the best population, namely the one with largest mean, and randomly number the others as population $1, \ldots, n-1$: Imagine that the best population is playing a \gambler's ruin game" with each of the others, with the best one beating population $i$ if the dierence of the cumulative number of wins of the best to that of $i$ hits the value $k$ before $-k$: Let $B_i$ be the event that the best

population beats i; i = 1; : : : ; n 1; and note that the best population will be chosen if it wins all of its games. That is, if we let $B = B_1 B_2 B_{n-1}$; then $B \subset C$; giving that

$$P(C) \geq P(B)$$

Lemma 2 $P(B) \geq (P(B_1))^{n-1}$

Proof: Let $U_0; U_1; : : : ; U_{n-1}$ and $U_{i;j}; i = 0; 1; : : : ; n-1; j \geq 1$ all be independent uniform $(0; 1)$ random variables. Let $X_0 = F^{-1}(U_0^{1=n})$; $X_i = F^{-1}((1-U_i)U_0^{1=n}); i = 1; : : : ; n-1$: Using Lemma 1 along with the fact that conditional on the maximum, call it M; of n independent uniform $(0; 1)$ random variables, the $n-1$ of these uniforms whose values are less than M are distributed as independent uniform $(0; M)$ random variables, it follows that the joint distribution of $X_0; X_1; : : : ; X_{n-1}$ is exactly that of the mean of the best population, followed by the means of the other $n-1$ populations in a random order.

Let $I_{0;j} = If1-U_{0;j} < X_0 g; j \geq 1$; and $I_{i;j} = IfU_{i;j} < X_i g; i = 1; : : : ; n-1; j \geq 1$: Note that $I_{i;j}$ has the distribution of the $j^{th}$ observa-tion of population i; $i = 0; : : : ; n-1; j \geq 1$; and also that $I_{0;j}$ is increasing in $U_{0;j}$ whereas $I_{i;j}$ is decreasing in $U_{i;j}; i \geq 1$: Because $X_i$ is decreasing in $U_i$ for $i > 0$, it consequently follows that, conditional on $U_0$; the indicator variables $IfB_1 g; : : : ; IfB_{n-1} g$ are all increasing functions of the independent random variables $U_1; : : : ; U_{n-1}; U_{i;j}; i = 0; : : : ; n-1; j \geq 1$: Consequently, given $U_0$; the indicators $IfB_1 g; : : : IfB_{n-1} g$ are associated, implying that

$$P(BjU_0) \geq \prod_{i=1}^{n-1} P(B_i jU_0)$$

which, by symmetry gives

$$P(BjU_0) \geq (P(B_1 jU_0))^{n-1}:$$

Taking expectations gives

$$\begin{aligned} P(B) &\geq E[(P(B_1 jP_0))^{n-1}] \\ &\geq (E[P(B_1 jP_0)])^{n-1} = \\ &(P(B_1))^{n-1} \end{aligned}$$

where the last inequality follows from Jensen's inequality. ■

To obtain an upper bound on $P(B)$; let $B$ be the event that the population with largest mean wins its gambler's ruin game against the population with the second largest mean. Because $B \quad B$ we have

Lemma 3 $P(B) \quad P(B)$

Remark: It is possible for the best arm to be chosen even if it does not win all its games. Indeed, this will happen if the best arm loses to an arm that at an earlier time had become dead, However, it is intuitive that this event has a very small probability of occurrence. Consequently, $P(C) \quad P(B)$:

To compute $P(B_1)$ and $P(B)$ we will use simulation with a conditional expectation estimator. To begin, note that if populations with known probabilities $x$ and $y$ play a game that ends when one has $k$ more wins than the other, then the probability that the one with probability $x$ wins is the probability that a gambler, starting with fortune $k$, who wins each game with probability $p = \frac{x(1 \quad y)}{x(1 \quad y)+y(1 \quad x)}$ will reach a fortune of $2k$ before 0. But, with

$$r \quad \frac{1 \quad p}{p} = \frac{y(1 \quad x)}{x(1 \quad y)}$$

this gambler's ruin probability is

$$P(x \text{ wins}) = \frac{1 \quad r^k}{1 \quad r^{2k}} = \frac{1}{1 + r^k} \tag{1}$$

Also, using known results from the gambler's ruin problem along with Wald's equation (to account for the fact that not every round leads to a gain or a loss) it follows that the mean number of plays is

$$E[\text{number of plays}] = \frac{k(1 \quad r^k)}{(r^k + 1)(x \quad y)} \tag{2}$$

Proposition 1: Let $U$ and $V$ be independent uniform $(0; 1)$ random variables, and let

$$X = F^{\ 1}(U^{1=n}); \quad Y = F^{\ 1}(U^{1=n}V); \quad W = F^{\ 1}(U^{1=n}V^{1=(n \ 1)})$$

$$R = \frac{Y(1 - X)}{X(1 - Y)}; \quad S = \frac{W(1 - X)}{X(1 - W)}$$

then

$$P(B_1) = E\left[\frac{1}{1 + R^k}\right]; \quad P(B) = E\left[\frac{1}{1 + S^k}\right]$$

Proof: The result follows from equation (1) upon using that the joint distribution of $X; Y$ is that of the largest and a random one of the other means, and the joint distribution of $X; W$ is that of the largest and second largest means. ■

Letting $N$ be the number of observations, we can approximate $E[N]$ by approximating the mean number of plays of each of the non best populations by the mean number of plays in their game against the best population, and approximating the mean number of plays of the best population by the mean number of plays in its game against the second best one. Hence, using Equation (2) we see that

$$E[N] \quad A \quad (n \qquad 1)E\left[\frac{k(1 - R^k)}{(R^k + 1)(X - Y)}\right] + E\left[\frac{k(1 - S^k)}{(S^k + 1)(X - W)}\right] \qquad (3)$$

Using Proposition 1 and Equation (3) enables us to eciently estimate $P(B_1); P(B)$; and $A$ by a simulation. Indeed, a simulation of $1; 000; 000$ runs, yielded that when $n = 10$ and $F(x) = x; 0 \quad x \quad 1$; the value $k = 50$ resulted in the estimates $P(B_1)^{n \quad 1} = :9885$; $P(B) = :9889$; and that $A$ 5460:4, with an estimate of the standard deviation of the estimator of $A$ being 26:7: Thus, $:9885 < P(B) < :9889$; and $E[N]$ 5460:539: A much more time consuming simulation (the preceding takes a fraction of a second) consisting of 1,000,000 runs, each run generating a random variable distributed as $N$, yielded that $P(C)$ 0:9886; $E[N]$ 5466:318; with the standard deviation of the estimate of $E[N]$ being 17:34:

For another illustration of the utility of the approximations of $P(C)$ and of the mean number of observations, suppose $n = 5$ and VT with $k = 10$ is used. A simulation based on Proposition 1 and Equation (3) yielded that

$$0:9523404 < P(B) < 0:9561526; \quad A \quad 358:398320$$

with a standard deviation of the estimate of A being 0:8256477: A simulation with 500; 000 runs, with each run generating the value of N , yielded that

$$P(C) = 0:9540; E[N] = 358:3993; sd = 1:156$$

where sd refers to the standard deviation of the estimator of $E[N]$:

The following example compares the performance of VT with some recently proposed rules.

Table 1: Algorithm Comparison

| Case | TS1 | TS2 | CR | KL1 | KL2 |
|---|---|---|---|---|---|
| 1 | 3968 | 4052 | 4516 | 8437 | 9590 |
| 2 | 1370 | 1406 | 3078 | 2716 | 3334 |

Example 1. Comparison with Recent Literature One might hope that the vector at a time rule assuming a uniform $(0; 1)$ prior assumption performs well for any set of probabilities $p_1; : : : ; p_n$: The following compares its performance with some recent algorithms - with names such as track and stop, Cherno racing, and Kullback-Leibler racing (see [5]. [6], and [7]. These algorithms, some of which have some asymptotic optimality features as the desired accuracy goes to 1, solve optimization problems to determine which population to next observe (and, consequently, are much more dicult to implement than is the VT rule). Table 1 is taken from [8]. It gives the results of 5 of these algorithms for two cases: the rst case having $n = 4$ with probabilities: $(0:5; 0:45; 0:43; 0:4)$; and the second having $n = 5$ with probabilities $(0:3; 0:21; 0:20; 0:19; 0:18)$. The parameters of the algorithms are chosen to guarantee at least 90 percent accuracy. (TS1 and TS2 refer to two variants of the track and stop algorithm; CR refers to the Cherno racing algorithm, and KL1 and KL2 refer to two variants of the Kulback-Leibler racing algorithm.

Because our algorithm assumes knowledge of a prior distribution, in cases where there is no reason to assume that we know what the prior is, it seems reasonable to assume a uniform $(0; 1)$ prior and choose a larger accuracy than is actually desired. So suppose we do so and require an accuracy, un-der a uniform $(0; 1)$ prior, of 99 percent. When $n = 4$ the vector at a time algorithm assuming a uniform prior requires $k = 42$; and when $n = 5$ if requires $k = 47$: Simulation, using the probabilities in each case, yields that the average number of observations needed in case 1 is 2738 with the correct decision being made with probability 0:9999; whereas the average number of observations needed in case 2 is 2372 with the correct decision being made with probability 0:9999: If we were to be less conservative and only require 97 percent accuracy when assuming the uniform prior, then the required value of the vector at a time rule is $k = 15$ when $n = 4$ and $k = 16$ when $n = 5$:

Simulation yields that the average number of observations needed in case 1 is 905 with the correct decision being made with probability 0:9999; whereas the average number of observations needed in case 2 is 832 with the correct decision being made with probability 0:998: Thus the vector at a time algorithm signicantly outperforms the newer algorithms. (Though to be fair we should mention that, under the uniform (0; 1) prior, k = 5 is sucient when either n = 4 or n = 5 to obtain 90 percent accuracy. Using k = 5 in Case 1 yields that the average number of observations needed is only 166, but the accuracy is :601: Using k = 5 in Case 2 yields that the average number of observations needed is 213:2, with accuracy :81:)

Remark: The preceding example is very interesting in that it indicates that the vector at a time algorithm that assumes a uniform (0; 1) prior can signicantly outperform the newer algorithms even in cases where the probabilities are highly unlikely to have come from this prior. Thus, while we are not claiming that there are not cases where assuming a uniform prior will lead to a poor result (for instance, for any value of k chosen, if all the $p_i$ are approximately equal then the VT procedure will yield P(C) 1=n;) we do feel that it will typically perform quite well.

Remark on Variance Reduction:
In practice, we observe that the number of plays using VT may have a large variance. In the case where F is the uniform (0; 1) distribution, we can reduce the variance of a simulation estimator that on each run generates the value of N by using $Y = \frac{1}{P(1 \quad P)}$ as a control variable, where $P_1$ and $P_2$ are the means of the best and second best arms. That is, if let T denote the raw estimator, then the new estimator is

$$T + c(Y \quad E[Y])$$

where the variance is minimized when $c = \frac{Cov(T;Y)}{Var(Y)}$. To obtain the mean value of the control variable, we condition on $P_2$,

$$\begin{aligned} E[\frac{1}{P_1(1 \quad P_2)}] &= E[E[\frac{1}{P_1(1 \quad P_2)} | P_2]] \\ &= E[\frac{1}{1 \quad P_2} E[\frac{1}{P_1} | P_2]] \end{aligned}$$

$$= E[\frac{\log(\ ^2)}{(1 \quad P_2)^2}] \qquad \text{because } P_1 j P_2 \ \ \text{unif}(P_2; 1)$$

$$= n(n \quad 1)\int_0^1 \frac{x^{n\ 2}\log(x)}{(1 \quad x)}dx$$

$$n(n \quad 1)\frac{1}{r}\sum_{i=1}^{r} h(\frac{i \quad 0{:}5}{r})$$

where $r$ is a large integer, and $h(x) = \frac{x^{\ n}\ ^2\log(x)}{(1\ \ x)}$: The values of $Cov(T; Y)$ and $Var(Y)$ can be estimated from the simulation, and these can then be used to determine c. In our numerical examples, we observe that the variance is reduced by between 50 and 60 percent.

# 3 VT with Early Elimination

Suppose we use VT but with an early elimination possibility in that if an arm is $j$ behind after the first $j$ rounds (that is, if the arm had all failures in the first $j$ rounds while another arm had all successes) then that arm is eliminated. To see by how much that can reduce the accuracy of VT, let us compute $P(L)$; where $L$ is the event that the best arm is eliminated early. Let 0 be the best population, let $1; : : : ; n \quad 1$ be the other popu-lations in random order, and let $X_0; : : : ; X_{n\ 1}$ be their respective success probabilities. With $U; U_1; : : : ; U_{n\ 1}$ being iid uniform $(0; 1)$ random vari-ables, note that $(X_0; : : : ; X_{n\ 1}) =_{st} (F^{\ 1}(W); F^{\ 1}(W U_1); : : : ; F^{\ 1}(W U_{n\ 1}))$ where $W = U^{1=n}$: Consequently, with
$(X_0; : : : ; X_{n\ 1}) = (F^{\ 1}(W); F^{\ 1}(W U_1); : : : ; F^{\ 1}(W U_{n\ 1}))$; we have

$$P(L) = E[(1 \quad X_0)^j(1 \quad \prod_{i=0}^{Y}(1 \quad X_i^j)]$$

Let us now consider the expected number of non best populations that are eliminated early. Letting $I_k$ be the indicator or the event that population $k$ is eliminated early, we have for $k \ \ 1$

$$E[I_k] = E[(1 \quad X_k)^j(1 \quad (1 \quad X_0^j)\prod_{i=0;k}^{Y}(1 \quad X_i^j)]$$

Hence, with $N$ being the number of non best populations that are eliminated early, we have

$$E[N] = (n \quad 1)E[(1 \quad X_1)^j(1 \quad (1 \quad X_0^j)\prod_{i=2}^{Y^n}(1 \quad X_i^j)]$$

$P(L)$ and $E[N]$ are easily evaluated by a simulation.

The formulas for $P(L)$ and $E[N]$ considerably simplify when $F$ is the uniform $(0;1)$ distribution. Let $N_i; i = 0;:::;n \quad 1$; be the number of successes of population $i$ in the rst $j$ rounds. Because $F^{\quad 1}(x) = x$; we obtain when $i \quad 1$ that

$$P(N_i = jjW) = E[(U_iW)^jjW] = \frac{W^j}{j+1}$$

Using that $N_0;:::;N_{n\ 1}$ are conditionally independent given $W$, the preceding gives

$$P(LjW) = (1 \quad W)^j(1 \quad (1 \quad \frac{W^j}{j+1})^{n\ 1})$$

Taking expectations gives

$$P(L) = E[(1 \quad W)^j(1 \quad (1 \quad \frac{W^j}{j+1})^{n\ 1})]$$

where $W = U^{1;n}$: Let us now consider $E[N]$. Again using that $N_0;:::;N_{n\ 1}$ are conditionally independent given $W$, we have

$$P(N_1 = 0; \max_{i=1} N_i = jjW) = P(N_1 = 0jW)(1 \quad \prod_{i=1}^{Y} P(N_i = jjW)) \quad (4)$$

Now,

$$\begin{aligned} P(N_1 = 0jW) &= E[(1 \quad U_1W)^jjW] \\ &= \int_0^{Z\ 1} (1 \quad xW)^j\,dx \\ &= \frac{1 \quad (1 \quad W)^{j+1}}{(j+1)W} \end{aligned}$$

Also,

$$\begin{aligned} \prod_{i=1}^{Y} P(N_i = jjW) &= (1 \quad W^j)(E[1 \quad (U_1W)^jjW])^{n\ 2} \\ &= (1 \quad W^j)(1 \quad \frac{W^j}{j+1})^{n\ 2} \end{aligned}$$

Hence, Equation (4) yields

$$E[N] = (n \quad 1)E\left[\frac{1 \quad (1 \quad W)^{j+1}}{(j+1)W} \quad 1 \quad (1 \quad W^j)(1 \quad \frac{W^j}{j+1})^{n\ 2}\right]$$

where $W = U^{1=n}$:

Thus, when $F(x) = x$; both $P(L)$ and $E[N]$ are one dimensional integrals, easily evaluated by numerical methods.

Example 2. The following are the values of $P(L)$; the probability that the best population is eliminated early, and $E[N]$; the mean number of non best populations that are eliminated early, for a variety of values of $n$ and $j$ when $F$ is the Uniform $(0; 1)$ distribution.

| $n$ | $j$ | $P(L)$ | $E[N]$ |
|---|---|---|---|
| 5 | 2 | .02053 | 1.317 |
| | 3 | .00336 | 0.851 |
| | 4 | .00059 | 0.590 |
| | 5 | .00011 | 0.432 |
| 10 | 2 | .01278 | 3.234 |
| | 3 | .00201 | 2.310 |
| | 4 | .00033 | 1.731 |
| | 5 | .00006 | 1.343 |
| 20 | 2 | .00429 | 6.659 |
| | 3 | .00053 | 4.978 |
| | 4 | .00007 | 3.942 |
| | 5 | .00001 | 3.229 |

For the cases considered in the preceding table, for a xed $j$ the probability that the best population is eliminated early decreases in the number of populations $n$: Intuitively, the reason for this is that although it becomes much more likely that at least one of the non-best populations has $j$ successes in its rst $j$ trials as $n$ increases, because the success probability of the best population is distributed as the maximum of $n$ independent uniform $(0; 1)$ random variables - and thus stochastically increases in $n$ the larger $n$ is, the less likely it is that the rst $j$ observations of the best population will all be failures. ■

Let $P_k(C)$ be the probability of a correct choice when using $VT$ with critical value $k$; and suppose $P_{k\ 1}(C) < < P_k(C)$: The randomized rule that chooses $VT$ with critical value $k$ with probability $p = \frac{P(C_k) - }{P(C)_k P(C_k)_1}$ or $VT$ with critical value $k\ \ 1$ with probability $1\ \ p$ will yield the correct choice with probability : Another possibility is to use $VT$ along with early elimination parameter $j$; where $j$ is the smallest value $j$ for which using $VT$

with critical value k along with early elimination if behind by j after the rst j rounds results in a correct choice with probability at least : (Of course, we could use a policy that randomizes between VT with critical value k and early elimination at j 1 and VT with critical value k and early elimination at j :) The following is an example where randomizing among VT rules results in a smaller mean number of observations than does VT with early elimination.

Example 3. Suppose n = 5; = :95 and F(x) = x; 0 x 1: The following simulated results were based on 500; 000 runs. (The term sd refers to the standard deviation of the estimator of E[N].)

VT

k = 9 : P(C) = 0:948; E[N] = 313:64; sd = 2:11
k = 10 : P(C) = 0:954; E[N] = 358:40; sd = 1:156

VT early elimination when k = 10

j = 2 : P(C) = 0:9385; E[N] = 335:52; sd = 8:29
j = 3 : P(C) = 0:9523; E[N] = 348:27; sd = 2:70

Consequently, randomizing among VT with k = 9 and k = 10 to obtain P(C) = :95 results in the mean number of observations being (2=3)313:64 + (1=3)358:40 = 328:56; which is smaller than what can be obtained with VT with early elimination. (It is also better than the newer proposed rules. Of these, the Cherno bound algorithm performs best, giving a mean number of 423:4 with accuracy 0:953:)

# 4 Play the Winner rule

Another older rule that can also be utilized in the Bayesian setting is the play the winner (PW) rule, which in each but the last round continues to sample from each alive population until it has a failure. Our variation of PW, which is somewhat dierent than what has been previously considered, is as follows:

All populations are initially alive.

A round consists of subrounds. In a subround, each alive population is observed once, with the successful ones continuing to the next subround. If there is only one population that is successful in a subround,

then if that population currently has a cumulative number of successes that is at least k more than any other population the process stops and that population is declared the best; if not, that population moves to the next subround. If none of the populations in a subround are successful then the round ends.

At the end of a round, any population whose cumulative number of successes is less than that of another by at least k is no longer alive.

Remarks:

1. Note that the process ends after a subround which had exactly one successful arm, and that arm's cumulative number of successes is now at least k higher than all other populations and exactly k larger than at least one population.

2. If we had dened a round by saying that each alive population is observed until it had a failure, then when F is the uniform (0; 1) distribution, the expected number of plays until the rst population used has a failure is innite. On the other hand, dening rounds using subrounds results in the mean number of plays being nite. For instance, suppose the probabilities are the values of iid uniform (0; 1) random variables. Let $N_i$ denote the num-ber of plays in the rst round of the arm with $i^{th}$ largest success probability. Denoting this probability by $Y_i$, its density is

$$f_{Y_i}(p) = \frac{n!}{(i\ \ 1)!(n\ \ i)!} p^{n\ \ i}(1\ \ \ p)^{i\ \ 1} dp; \quad 0 < p < 1$$

it follows that $E[N_i]\ \ E[_1\ \frac{1}{Y_i}] < 1$ when $i > 1$: In addition,

$$N_1\ \ k + \max_{2 \text{ in}} N_i\ \ k + \sum_{i=2}^{n} N_i$$

giving that

$$E[N_1]\ \ k + \sum_{i=2}^{n} E[N_i] < 1$$

3. The PW rule as dened in [15] and [16] was such that the populations are initially randomly ordered. In each round, the alive populations were observed in that order, with each population being observed until it had a failure. If at any time one of the populations had k fewer successes than another population, then the former is no longer alive. The process ends

when only a single population is alive, and that population is declared best. Thus, for instance, in the original version if the rst population observed has k successes in a row then that population is declared best.

## 4.1 Analysis of P W

To begin, suppose there are only 2 arms and that their success probabilities are $p_1 > p_2$: Suppose we are going to choose an arm by using the procedure which in each round plays each arm until it has a failure, and then stopping at the end of a round if one of the arms has had a cumulative number of successes that is at least k more than the other, with that arm then being chosen. Let $q_i = 1 \quad p_i; i = 1; 2;$ and let $X_{i;r}; \ i = 1; 2; \ r \quad 1;$ be independent with $P(X_{i;r} = j) = q_i p_i^j; \ j \quad 0$: Interpret $X_{i;r}$ as the number of successes of arm i in round r; and let $Y_r = X_{1;r} \qquad X_{2;r}; \ r \quad 1$: Then,

$$E[e^{Y_r}] = \frac{q_1}{1 \quad p_1 e} \frac{q_2}{1 \quad p_2 e}$$

It is now easy to check that $E[e^{Y_r}] = 1$ if $e = p_2=p_1$. That is, $E[(p_2=p_1)^{Y_r}] = 1$: If we now let $S_m = \sum_{i=1}^{m} Y_i$; then $(p_2=p_1)^{S_m}; \ m \quad 1$ is a martingale with mean 1. Letting

$$N = \min(m : S_m \quad k \text{ or } S_m \quad k)$$

it follows by the martingale stopping theorem (see [11]) that

$$E[(p_2=p_1)^{S_N}] = 1$$

Let $p = P(S_N \quad k)$ be the probability that arm 1 is chosen. Then

$$\begin{aligned} 1 &= E[(p_2=p_1)^{S_N}] \\ &= E[(p_2=p_1)^{S_N} j S_N \quad k]p + E[(p_2=p_1)^{S_N} j S_N \quad k](1 \qquad p) \end{aligned}$$

Letting $X_i; \ i = 1; 2;$ have the distribution of $X_{i;r}$; it follows, by the lack of memory property of $X_i$, that

$$E[(p_2=p_1)^{S_N} j S_N \quad k] = (p_2=p_1)^k E[(p_2=p_1)^{X_1}] = (p_2=p_1)^k (q_1=q_2)$$

$$E[(p_2=p_1)^{S_N} j S_N \quad k] = (p_2=p_1)^{\ k} E[(p_2=p_1)^{\ X_2}] = (p_1=p_2)^k (q_2=q_1)$$

Substituting back yields that

$$p = \frac{1 - (p_1/p_2)^k(q_2/q_1)}{(p_2/p_1)^k(q_1/q_2) - (p_1/p_2)^k(q_2/q_1)} \tag{5}$$

Conditioning on which arm wins yields that

$$E[S_N] = (k + E[X_1])p + (-k - E[X_2])(1 - p)$$

Letting $m_i = E[X_i] = 1/q_i - 1 = p_i/q_i$; the preceding gives

$$E[S_N] = p(m_1 + m_2 + 2k) - (m_2 + k)$$

Wald's equation yields

$$E[N] = \frac{p(m_1 + m_2 + 2k) - m_2 - k}{m_1 - m_2} \tag{6}$$

Because $X_{1;r} + X_{2;r} + 2$ is the number of plays in round r; it follows that the total number of plays, call it T, is $\sum_{r=1}^{N}(X_{1;r} + X_{2;r} + 2)$: Applying Wald's equation and using (6) gives

$$E[T] = (p(m_1 + m_2 + 2k) - m_2 - k)\frac{m_1 + m_2 + 2}{m_1 - m_2} \tag{7}$$

Now, suppose that we utilize PW. Let $B(p_1; p_2)$ and $M(p_1; p_2)$ be, respectively, the probability that the arm with value $p_1$ is chosen and the mean number of plays before stopping. From equation (5) we have

$$B(p_1; p_2) = \frac{1 - (p_1/p_2)^k(q_2/q_1)}{(p_2/p_1)^k(q_1/q_2) - (p_1/p_2)^k(q_2/q_1)} \tag{8}$$

Because PW would stop play once the winning arm is ahead by k, whereas $E[T]$ is the mean number of plays when we continue on until a failure occurs, we obtain by conditioning on which arm wins that

$$M(p_1; p_2) = E[T] - \frac{B(p_1; p_2)}{q_1} - \frac{1 - B(p_1; p_2)}{q_2} \tag{9}$$

$$= (B(p_1; p_2)(m_1 + m_2 + 2k) - m_2 - k)\frac{m_1 + m_2 + 2}{m_1 - m_2} - \frac{B(p_1; p_2)}{q_1} - \frac{1 - B(p_1; p_2)}{q_2}$$

where $m_i = p_i/q_i$:

Now suppose there are n arms whose means are the values of independent random variables with distribution F, and let C be the event that the PW policy chooses the best arm. As in our analysis of VT, suppose that all arms participate in each round. Let arm 0 be the best arm, and randomly number the other arms as $1;:::;n\ \ 1$: Say that the best arm beats arm i if the end of round dierence between the cumulative number of successes of arm 0 and arm i is at least k before it is less than or equal to $\ \ k$: Letting $B_i; i = 1;:::;n\ \ 1$; be the event that arm 0 beats arm i, we have, by the same arguments as in Section 1, the following.

Lemma 4. With $B\ \ B_1B_2\ B_{n\ \ 1}$,

$$P(C)\ \ P(B)\ \ (P(B_1))^{n\ \ 1}$$

Also, $P(B)\ \ P(B)$; where B is the event that 0 beats the best of arms $1;:::;n\ \ 1$:
Our preceding analysis yields the following corollary.

Corollary 1. With U and V being independent uniform $(0;1)$ random variables, and

$$X = F^{\ \ 1}(U^{1=n}); \quad Y = F^{\ \ 1}(U^{1=n}V); \quad W = F^{\ \ 1}(U^{1=n}V^{1=(n\ \ 1)})$$

$$P(B_1) = E[B(X;Y)]$$
$$P(B) = E[B(X;W)]$$

Also, if we let M denote the mean number of plays, then

$$M\ \ A = (n\ \ \ \ 1)E[M(X;Y)] + E[M(X;W)]$$

## 4.2 PW with Early Elimination

Suppose we use PW with critical number k and add an early elimination on any population whose first j observations are all failures. Let $B_e$ be the event that the best population is eliminated early. Letting $f(p) = F'(p)$ be the density function of the prior distribution, the success probability of the best population has density function $f_b(p) = nF^{n-1}(p)f(p)$; $0 < p < 1$: Consequently, it follows that

$$P(B_e) = \int_0^1 (1-p)^j nF^{n-1}(p)f(p)dp$$

Let $N_{nb}$ be the number of nonbest populations that are eliminated early. To compute $E[N_{nb}]$; note that the probability a randomly chosen population is eliminated early is $\int_0^1 (1-p)^j f(p)dp$; giving that

$$n\int_0^1 (1-p)^j f(p)dp = E[\text{number eliminated early}] = E[N_{nb}] + P(B_e)$$

Hence,

$$E[N_{nb}] = n\int_0^1 (1-p)^j f(p)dp - \int_0^1 (1-p)^j nF^{n-1}(p)f(p)dp$$

When F is the uniform (0; 1) distribution

$$P(B_e) = \int_0^1 (1-p)^j np^{n-1}dp = \frac{n!j!}{(n+j)!}$$

and

$$E[N_{nb}] = \frac{n}{j+1} - \frac{n!j!}{(n+j)!}$$

For instance, if F is the uniform (0; 1) distribution, then when $n = 10; j = 5$, we have $P(B_e) = 0.000333$ and $E[N_{nb}] = 1.666$:

## 4.3 VT versus PW

Based on numerical experiments, VT and PW have roughly similar performances when $F(x) = x$:

Example 5. When $n = 5$; simulation yielded the following results for PW.

PW:

$k = 42 : P(C) = 0{:}9494;\ M = 319{:}78;\ sd = 1{:}64$

$k = 43 : P(C) = 0{:}9502;\ M = 327{:}80;\ sd = 1{:}65$

$k = 48 : P(C) = 0{:}9543;\ M = 375{:}4;\ sd = {:}899$

Thus, choosing PW with $k = 42$ with probability :25 and $k = 43$ with probability :75 results in $P(C) = {:}950$; and requires, on average, 325:795 observations, which is slightly less than the average of 328:56 which, as shown in Example 3, can be obtained by a randomization of VT rules to obtain $P(C) = {:}95$: On the other hand if we wanted $= {:}954$; then both VT with $k = 10$ and PW with $k = 48$ achieve that, with VT having a mean of 358:4 observations, compared to 375:4 for PW. (Because the average number of trials needed for PW with $k = 47$ is 367:05; randomizing between PW(47) and PW(48) still would not be as good as VT(10).)

# 5 The Normal Case

## 5.1 VT Rule in the Normal Case

Suppose that observations on population $i$ are the values of independent normal random variables with mean $_i$ and variance $^2; i = 1; : : : ; n$; where $^2$ is known and $_1; : : : ; _n$ are the unknown values of $n$ independent standard normal random variables. As before our objective is to determine, using relatively few observations, the population $i$ such that $i = \text{argmax}_i$ under the proviso that the probability of a correct decision is at least some specied value : The VT rule with parameter $c > 0$ is as follows:

Initially all populations are alive

A round consists of a single observation from each alive population.

At the end of a round, a population is no longer alive if its cumulative sum of observed values is more than $c$ less than that of another population. (That is, if $A$ is the set of alive populations after round $k$ 1, and $S_i(k)$ is the cumulative sum of the rst $k$ observations of population $i$, then $i \in A$ would no longer be alive after round $k$ if $S_i(k) < \max_{j \in A} S_j(k)$ c:)

If only one population is alive, stop and declare it best. Otherwise perform a new round.

Before showing how to determine the appropriate value of c we present some preliminaries concerning normal partial sums.

## 5.2 Some Preliminaries

Let $$ be the standard normal distribution function. Dene

$$R(a) = \frac{(a)}{1 \quad ^{(\ )}a} \tag{10}$$

Lemma 5. If W is a normal random variable with mean and variance 1, then

$$E[e^{\ 2W} jW > 0] = \quad R(\ )$$

$$E[W jW > 0] = \ + \ \frac{e^{\ 2=2}}{2\ ()R()}$$

$$E[e^{\ 2W} jW < 0] = \frac{e^{\ 2=2}}{2(1 \quad ())}$$

$$E[W jW < 0] =$$

Proof:

$$E[e^{\ 2W} jW > 0] = \frac{1}{2P(W > 0)} \int_0^{\ 1} e^{\ 2x}\, e^{(x\ )^2=2}\ dx$$

$$= \frac{1}{2()} \int_0^{\ 1} e^{\ (x+)^2=2}\ dx$$

$$= \frac{1 \quad ()\ ()}{}$$

Let $Z = W \quad .$

$$E[W jW > 0] = \ + E[Z jZ > \ ]$$

$$= \ + \ \frac{1}{2()} \int_{\ 1}^{\ 1} x e^{\ x^2=2}\, dx$$

$$= \ + \ \frac{e^{\ 2=2}}{2()}$$

Because $E[e^{2W}] = 1$; the third equality follows from the first upon using the identity

$$1 = E[e^{2W} | W > 0]() + E[e^{2W} | W < 0](1 \quad ())$$

Similarly, the fourth equality follows from the second since $= E[W | W > 0]() + E[W | W < 0](1 \quad ())$: ■

Lemma 6. Let $S_n = \sum_{i=1}^{n} W_i$; $n \quad 1$; where $W_i; i \quad 1$ are independent normal random variables with mean $> 0$ and variance 1: For given $b > 0$; let $N = \min\{n : \text{either } S_n < \quad b \text{ or } S_n > b\}$:
(a) $R(\quad)e^{2b} < E[e^{2S_N} | S_N > b] < e^{2b}$(b)
$e^{2b} < E[e^{2S_N} | S_N < \quad b] < e^{2b}R()$

Proof: The right hand inequality of (a) is immediate since $> 0$: To prove the left side of (a), note that conditional on $S_N > b$ and on the value $S_N \quad _1$; that $S_N$ is distributed as b plus the amount by which a normal with mean and variance 1 exceeds the positive amount $b \quad S_N \quad _1$ given that it does exceed that amount. But a normal conditioned to be positive is known to have strict increasing failure rate (see [2]) implying that $S_N | S_N > b; S_N \quad _1\}$ is stochastically smaller than $b + W_i | W_i > 0\}$. As this is true no matter what the value of $S_N \quad _1$; it follows that $S_N | S_N > b\}$ is stochastically smaller than $b + W_i | W_i > 0\}$, implying that $E[e^{2S_N} | S_N > b] > e^{2b}E[e^{2W_i} | W_i > 0]$: The result now follows from Lemma 5.
The left hand inequality of (b) is immediate. To prove the right hand inequality, note that the same argument as used in part (a) shows that $S_N | S_N < \quad b\} >_{st} \quad b + W_i | W_i < 0\}$; implying that $E[e^{2S_N} | S_N < \quad b] < e^{2b}E[e^{2W_i} | W_i < 0]$: Thus, the result follows from Lemma 5. ■

Proposition 2. Let $S_n = \sum_{i=1}^{n} W_i$; $n \quad 1$; where $W_i; i \quad 1$ are independent normal random variables with mean $> 0$ and variance 1: For given $b > 0$; let $N = \min\{n : \text{either } S_n < \quad b \text{ or } S_n > b\}$:

$$\frac{e^{2b} \quad 1}{e^{2b} \quad (\quad)e^{2b}} < P(S_N > b) < \frac{e^{2b} \quad R() \quad 1}{e^{2b}R() \quad e^{2b}} \qquad (11)$$

Proof: Let $p = P(S_N > b)$: Because $E[e^{2W_i}] = 1$; it follows that $\{e^{2S_n}; n \geq 1\}$ is a martingale with mean 1. Hence, by the martingale stopping theorem

$$\begin{aligned} 1 &= E[e^{2S_N}] \\ &= E[e^{2S_N} | S_N > b]p + E[e^{2S_N} | S_N \leq b](1 - p) \end{aligned}$$

Hence,

$$p = \frac{E[e^{2S_N} | S_N \leq b] - 1}{E[e^{2S_N} | S_N \leq b] - E[e^{2S_N} | S_N > b]} \tag{12}$$

Because $\frac{x-1}{x-y}$; $0 < y < 1 < x$; increases in both x and y, the inequalities (11) now follow from Lemma 6. ■

Although we don't directly use the following proposition, it is of independent interest.

Proposition 3. With N as previously defined.

$$E[N] \leq \frac{e^{2b}R(\theta) - 1}{e^{2b}R(\theta) - e^{2b\theta}}\left(\frac{2b}{\mu} + \frac{\theta}{\eta(\theta)}e^{2b\theta} + 1\right) - \frac{b}{\eta(\theta)} \tag{13}$$

$$E[N] \geq \frac{e^{2b\theta} - 1}{R(\theta)e^{2b\theta}}\left(\frac{2b}{\mu} - 1 + \frac{\theta}{2(1-\eta(\theta))}e^{2b\theta}\right) - \frac{b}{\eta(\theta)} + 1 - \frac{\theta}{2(1-\eta(\theta))} \tag{14}$$

Proof: Wald's equation gives

$$\begin{aligned} E[N] &= E[S_N | S_N > b]p + E[S_N | S_N \leq b](1 - p) \\ &\geq (b + E[W | W > 0])p - b(1 - p) \\ &= p\left(2b + \frac{\theta}{2\eta(\theta)}e^{2\theta\mu} + \mu\right) - b \end{aligned}$$

where the first inequality used, as shown in Lemma 6, that $S_N | \{S_N > b\}$ is stochastically smaller than $b + W | \{W > 0\}$. Inequality (13) now follows from Proposition 2. The lower bound follows from

$$E[N] = E[S_N | S_N > b]p + E[S_N | S_N \leq b](1 - p)$$

$$= \frac{bp + (b + E[W_1 | W_1 < 0])(1 - p)}{(2b + \sqrt{2\pi}\,\frac{e^{-\mu^2/2}}{\Phi(-\mu)})p + (b + \sqrt{2\pi}\,\frac{e^{-\mu^2/2}}{\Phi(-\mu)})(1-p)}$$

where the inequality used that $S_N | S_N < b$ is stochastically larger than $b + W_i | W_i < 0$. Inequality (12) now follows from Proposition 2. ■

Remark: One way to approximate $p = P(S_N > b)$ and $E[N]$ is to "neglect the excess" and assume $S_N | S_N > b \overset{st}{=} b$ and $S_N | S_N < b \overset{st}{=} b$. From (12) this gives that

$$p \approx \frac{e^{2b} - 1}{e^{2b} - e^{-2b}} \tag{15}$$

Also, $E[N] \approx bp + b(1-p)$, and so (15) gives that

$$E[N] \approx \frac{2b(e^{2b} - 1)}{(e^{2b} - e^{-2b})} + b \tag{16}$$

Example 6: If $b = 3$, $\mu = 1$, then (13), (14), and (16) yield that

$$2.9838 < E[N] < 4.2842, \qquad E[N] \approx 2.9852$$

Corollary 2: Let $S_n = \sum_{i=1}^{n} V_i$, $n \geq 1$, where $V_i, i \geq 1$ are independent normal random variables with mean $\mu > 0$ and variance $\sigma^2$. For given $c > 0$, let $N = \min\{n : \text{either } S_n < 0 \text{ or } S_n > c\}$.

$$\frac{e^{c\mu/\sigma^2} - 1}{e^{c\mu/\sigma^2} - R(\frac{\mu}{\sigma^2})e^{-c\mu/\sigma^2}} < P(S_N > c) < \frac{e^{c\mu/\sigma^2} - R(\frac{\mu}{\sigma^2}) - 1}{e^{c\mu/\sigma^2} - R(\frac{\mu}{\sigma^2}) - e^{-c\mu/\sigma^2}}$$

Moreover,

$$E[N] \approx \frac{2c}{c\mu}\left(\frac{(e^{c\mu/\sigma^2} - 1)}{e^{c\mu/\sigma^2} - e^{-c\mu/\sigma^2}}\right) - \frac{c}{\mu}$$

Proof: Let $W_i = \frac{V_i}{\sigma^2}$; note that $E[W_i] = \frac{\mu}{\sigma^2}$. Now, using $b = \frac{c\mu}{\sigma^2}$, apply Proposition 2 and Equation (16). ■

## 5.3 Analyzing the VT Rule in the Normal Case

We can obtain a lower bound and an eective approximation to $P(C)$; the probability that the correct choice is made, by a similar argument as in the Bernoulli case. Letting population 0 be the one with the largest mean, and randomly numbering the others as $1; \ldots; n-1$; we imagine a \gambler's ruin" game between populations 0 and $i$ in which the winner is the rst one whose cumulative sum is at least $c$ more than that of the other. With $B_i$ being the event that the best wins this game against population $i$, and $B$ the event that the best wins all these games, we can show exactly as before that

$$P(C) \geq P(B) \geq (P(B_1))^{n-1}; \qquad P(B) \geq P(B)$$

where $P(B)$ is the probability that the best beats the population with second largest mean. Given the values $_0; _1; \ldots; _{n-1}$, the dierence between the value of a population 0 observation and one from a dierent population is a normal random variable with variance $2^2$: Letting

$$L() = \frac{e^{c=2} - 1}{e^{c=2} R( - _2) - (- _2) e^{-c=2}}; \qquad U() = \frac{e^{c=2} - R(- _2) - 1}{e^{c=2} R(- _2) - e^{-c=2}} \tag{17}$$

be the lower and upper bounds on $P(S_N > c)$; Corollary 2 yields the following proposition.

Proposition 4: Let $U$ and $V$ be independent uniform $(0; 1)$ random variables, and let

$$X = {}^{-1}(U^{1=n}) - {}^{-1}(U^{1=n}V); \qquad Y = {}^{-1}(U^{1=n}) - {}^{-1}(U^{1=n}V^{1=(n-1)})$$

Then

$$P(C) \geq P(B) \geq (E[L(X)])^{n-1} \quad \text{and} \quad P(B) \geq E[U(Y)]$$

Letting $N$ be the number of observations, we can approximate $E[N]$ by approximating the mean number of plays of each of the non best populations by the mean number of plays in their game against the best population, and approximating the mean number of plays of the best population by the mean number of plays in its game against the second best one. Hence, using Equation (16) we see that

$$E[N] \quad A \quad (n \qquad 1)E[M(X)] + E[M(Y)] \tag{18}$$

where

$$M() = \frac{2c\,(e^{c=} \qquad - \quad 1)}{(\quad e^{c=}\ e\text{-} \qquad c=\ ^2)} \quad \underline{c}$$

The following table compares the performance of VT with the most quoted algorithms of the recent literature: LIL-UCB, TrackAndStop, and Cherno. The LIL-UCB algorithm (see [6]) uses upper condence bounds (UCB) based on the law of the iterated logarithm for the expected reward of the arms. At each stage it uses the arm with the largest upper bound, sim-ilar to the UCB algorithm of bandit problems. We use a heuristic variation of the LIL-UCB which has been shown to perform somewhat better than the original [6]. The TrackAndStop algorithm in [8] tracks lower bounds on the optimal proportions of the arm draws and uses a stopping rule based on Cherno's work on Generalized Likelihood Ratio statistic. The Cherno algorithm is similar to TrackAndStop, but rather than track the optimal proportions it instead chooses between the empirical best and second-best. In each of 10; 000 simulation runs we generate the n means by generating n standard normals and then simulate the results of the dierent algorithms.

Table 2: N is the average number of plays in 10; 000 simulation runs

| n | | Algorithm | c | N | p |
|---|---|---|---|---|---|
| 2 | 0.9 | VT | 5 | 15 | 0.9068 |
| | | TrackAndStop | - | 1351 | 0.9988 |
| | | LIL | - | 505 | 0.9989 |
| | | Cherno | - | 1476 | 1 |
| | 0.95 | VT | 11 | 35 | 0.9504 |
| | | TrackAndStop | - | 1545 | 1 |
| | | LIL | - | 504 | 0.999 |
| | | Cherno | - | 1546 | 1 |
| | 0.99 | VT | 40 | 489 | 0.9913 |
| | | TrackAndStop | - | 1759 | 1 |
| | | LIL | - | 503 | 0.9989 |
| | | Cherno 5 | - | 1760 | 1 |
| | 0.9 | VT | 7.8 | 62 | 0.910 |
| | | TrackAndStop | - | 1748 | 1 |
| | | LIL | - | 1657 | 0.9979 |
| | | Cherno | - | 1703 | 1 |
| | 0.95 | VT | 16.3 | 143 | 0.9537 |
| | | TrackAndStop | - | 1961 | 1 |
| | | lilUCB-H | - | 1654 | 0.9979 |
| | | Cherno | - | 1872 | 1 |
| | 0.99 | VT | 85 | 979 | 0.9909 |
| | | TrackAndStop | - | 2209 | 1 |
| | | lLIL | - | 1651 | 0.9979 |
| | | Cherno | - | 2151 | 1 |
| 10 | 0.9 | VT | 10.3 | 138 | 0.911 |
| | | TrackAndStop | - | 2071 | 1 |
| | | LIL | - | 2008 | 0.9989 |
| | | Cherno | - | 1965 | 1 |
| | 0.95 | VT | 20.6 | 304 | 0.9502 |
| | | TrackAndStop | - | 2173. | 1 |
| | | LIL | | 2005 | 0.9989 |
| | | -Cherno | - | 2055 | 1 |

# 6 Conclusions

We have presented a Bayesian model for nding the population having the largest mean when the populations under consideration are either all Bernoulli or all normal with a xed variance. In both cases we take a Bayesian approach that assumes the unknown means are the values of independent random variables having a specied distribution F: We consider two old rules that had previously been analyzed under the assumption that the largest mean diers from the second largest mean by at least some known positive number. The rst of these rules is the vector at a time rule (VT) which in each round takes a sample from each population, eliminates any population whose cumulative sum after a round is at least k less than that of another population, and continues until one one population is left. The second old rule, applicable in the Bernoulli case, is play the winner rule which in each but the last round continues to sample from each remaining population until it has a failure. For a given constant k we present easily computed bounds and approximations of the probability these rules yield the correct choice and the mean number of observations that are required. We also present numerical evidence showing in the Bernoulli case that the VT rule resulting when F is the uniform (0; 1) distribution has good results, when compared with more recent algorithms that make no assumptions about the set of means, even when the set of means does not look like it came from a uniform (0; 1) distribution. Although we recommend in any problem instance that one utilizes one's prior knowledge to determine the appropriate prior F , it is comforting to know that the method appears to work well even in cases when the actual means do not appear to have come from F:

# 7 Conict of Interest

On behalf of all authors, the corresponding author states that there is no conict of interest.

# References

[1] Audibert, J. Y., Bubeck, S., and R. Munos, Best Arm Identication in Multi-Armed Bandits, COLT 2010, The 23rd Conference on Learning

Theory, Haifa, Israel, 2010

[2] Barlow, R., and F. Proschan, Statistical Theory of Probability and Life Testing: Probability Models, Holt-Rinehart-Winston, 1975

[3] Bechhofer, R.E., Kiefer, J., and M. Sobel, Sequential Identication and Ranking Procedures, The University of Chicago Press, Chicago-London, 1968.

[4] Bechhofer, R.E. and R. V. Kulkarni, Closed adaptive sequential procedures for selecting the best of k > 2 Bernoulli populations, Statistical Decision Theory and Related Topics, III, S.S. Gupta and J.O. Berger, eds., Vol 1, Academic Press, New York, pp. 61-108, 1982.

[5] Even-Dar, E., Mannor, S., and Y. Mansour, Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. Journal of Machine Learning Research, 7, 1079-1105, 2006.

[6] Jamieson, K., Malloy, M., Nowak, R., and S. Bubeck, LiL UCB: An optimal exploration algorithm for multi-armed bandits. arxiv, 2014.

[7] Gabillon, V., Ghavamzadeh, M., and A. Lazaric. Best arm identication: A unied approach to xed budget and xed condence. in F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, 25, pages 3212?3220, 2012.

[8] Garivier, A. and E. Kaufmann, Optimal Best Arm Identication with Fixed Condence, JMLR Workshop and Conference Proceedings, vol 49: 1-30, 2016

[9] Hartman, M., An improvement on Paulson's sequential ranking procedure, Sequential Analysis, Volume 10, 363-372, 1991

[10] Hoel, D.G. and M. Mazumdar (1968), An extension of Paulson?s selection procedure. Ann. Math. Statist., 39, 2067-2074, 1968.

[11] Paulson, E., A sequential procedure for selecting the population with the largest mean from k normal populations. Ann. Math. Statist., 35, 174-180, 1964

[12] Ross, S. M., Stochastic Processes, ed. 2, Wiley, 1996

[13] Russo, D., Simple bayesian algorithms for best arm identication. CoRR, abs/1602.08448, 2016.

[14] Russo, D, Van Roy, B., Kazerouni, A., Osband, I., and Z. Wen, A Tutorial on Thompson Sampling, Stanford University Press, 2020

[15] Sobel, M., and G. Weiss, Play-the-winner rule and inverse sampling in selecting the better of two binomial populations. J. Amer. Statist. Assoc., Vol. 66, No. 335, pp. 545-551, 1971

[16] Sobel, M., and G. Weiss, Recent Results on Using the Play the Winner Sampling Rule with Binomial Selection Problems, Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Vol. 1, Univ. of Calif. Press, 717-736, 1972

[17] Thompson, W., On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 25(3/4): 285-294. 1933