Hypothesis Testing and Resource Allocation

Tom Cassey
cassey@cs.bris.ac.uk

December 12, 2008

1 Introduction

Technological developments, specifically those in areas such as communications, are resulting in information systems becoming increasingly interconnected and distributed. These developments are resulting in a transition from having myriad small scale, independent systems, towards having a smaller number of much larger systems, each of which is formed from the interconnection of many smaller systems.

As the complexity of these systems continues to grow it becomes increasingly difficult to obtain information, pertaining to the global state of the system, that classical state based control techniques depend on. Moreover, even with the availability of global state information, the non-linearity of the interactions between the various sub-systems makes it difficult to accurately predict the system's behaviour. In order to control these increasingly complex systems we believe a decentralised control methodology is required in which multiple controllers, each controlling a portion of the system and operating with incomplete knowledge of the system's state, work together to effect control over the entire system.

Such an approach will require a fusion of theories from a number of different fields, including statistical decision theory and game theory, in order to enable system controllers to effectively work together to achieve global system performance that approaches statistical optimality.

In this text we consider two areas of decision theory that we believe could form an integral part of a successful decentralised control solution. First, in Section 2, we introduce and discuss statistical hypothesis testing, starting with the Likelihood Ratio Test and subsequently showing how the methodology can be extended to form a sequential test, known as the Sequential Probability Ratio Test. Following this, in Section 3, we discuss a resource allocation problem, known as the bandit problem, in which a a single resource must be allocated between a set of consumers over a number of time intervals. Finally, we conclude in Section 4 by summarising the two areas we have considered.

2 Hypothesis Testing

A statistical hypothesis test is a method of selecting a single hypothesis from a set of one or more hypotheses, based on available data about the situation or process to which the hypotheses relate. In tests that consider only a a single hypothesis, the outcome of the test is either the acceptance or rejection of the hypothesis depending on the degree to which the hypothesis is supported by the data.

Numerous hypothesis tests have been developed, each of which has a different set of prior assumptions and characteristics. Choosing the 'best' hypothesis test to use is highly subjective and depends upon the specifics of the situation being hypothesised about.

In this this section we introduce and discuss two statistical hypothesis tests that can be used to make a decision between two simple hypotheses. The first hypothesis test considered is the Likelihood Ratio Test, which uses a set of observations, of fixed size, to select between two simple hypotheses. Following this, we introduce the Sequential Probability Ratio Test[5] and show how it extends the methodology of the Likelihood Ratio Test to create a hypothesis test that sequentially incorporates observations into the decision process, terminating once the uncertainty drops below a predetermined level. Finally, we conclude the section by summarising the two methods that have been discussed.

2.1 Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is a statistical test used to make a decision between a simple hypothesis H_0, often referred to as the null hypothesis, and a single alternative hypothesis H_1, based on a sequence of observations $X = \{x_1, x_2, \ldots, x_n\}$.

A decision between the two hypotheses is made through comparison of the likelihood function of each of the alternatives. The likelihood function, denoted $L(X \mid H_i)$, gives the probability of a given sequence of observations X occurring given that the hypothesis is true. Assuming that each observation x_t is independent of all others the likelihood function can expressed as follows.

$$L(X \mid H_i) = P(x_1 \mid H_i)P(x_2 \mid H_i) \ldots P(x_n \mid H_i) \tag{2.1}$$

Where $P(x_t \mid H_i)$ denotes the probability of x_t having been observed under hypothesis H_i.

Comparison of the two likelihood functions can be performed by taking the ratio $\Lambda(X)$ of the two functions as shown below in Equation 2.2.

$$\Lambda(X) = \frac{L(X \mid H_1)}{L(X \mid H_0)} = \frac{P(x_1 \mid H_1)P(x_2 \mid H_1) \ldots f(x_n \mid H_1)}{P(x_1 \mid H_0)P(x_2 \mid H_0) \ldots f(x_n \mid H_0)} \tag{2.2}$$

A decision rule $\delta(X)$ can be constructed that uses the ratio of the two likelihood functions and a decision threshold, denoted η, to determine which hypothesis to accept and which to reject,

as shown in Equation 2.3 below.

$$\delta(X) = \left\{ \begin{array}{ll} H_0 & \text{if } \Lambda(X) \leq \eta \\ H_1 & \text{if } \Lambda(X) > \eta \end{array} \right\} \tag{2.3}$$

Looking at Equation 2.3 it can be seen that the decision rule provides a mapping from each vector, X, in the observation space to a decision. This mapping divides the observation space into two mutually exclusive and exhaustive sets S_0 and S_1 each of which corresponds to the acceptance of a particular hypothesis. Since the sets are both mutually exclusive and exhaustive it must be true that $S_0 = S_1^c$ and conversely $S_1 = S_0^c$. A more formal definition of the sets and can be seen below in Equations 2.4 and 2.5.

$$S_1 = S_0^c = \left\{ X : \frac{L(X \mid H_1)}{L(X \mid H_0)} > \eta \right\} \tag{2.4}$$

$$S_0 = S_1^c = \left\{ X : \frac{L(X \mid H_1)}{L(X \mid H_0)} \leq \eta \right\} \tag{2.5}$$

The choice of decision threshold η affects the size of of the two sets. From Equation 2.3 it can be seen that a smaller value of η results in a larger number of vectors satisfying the inequality and thus a larger set S_1 and smaller set S_0. Conversely, a larger value of η leads to a larger set S_0 and a smaller set S_1. A threshold value of results in the hypothesis under which the observed sequence X is most likely to occur, being selected. Adjusting the threshold value biases the decision rule towards choosing the hypothesis H_i that corresponds to the set S_i which has grown relative to its size with threshold $\eta = 1$.

2.1.1 Neyman-Pearson Theorem

The Neyman-Pearson theorem shows that the LRT is the Uniformly Most Powerful (UMP) test, meaning that, for a given sample size n and significance level α the error rate for type 2 errors, β, is minimised. If S_1 is the set of observation vectors that result in the null hypothesis, H_0, being rejected, and S_A is the rejection set for an alternative hypothesis test and the probability that $X \in S$ given that X occurred under hypothesis H_i, denoted $C(X) = H_i$, is given by Equation 2.6.

$$P(X \in S \mid C(X) = H_i) = \sum_{X \in S} L(X \mid H_i) \tag{2.6}$$

If both tests have the same significance level α then $P(X \in S_1 \mid C(X) = H_0) = P(X \in S_A \mid C(X) = H_0)$. In addition, the observation space is problem specific and is therefore identical for both decision rules, thus $P(X \in S_1 \mid C(X) = H_i)$ and $P(X \in S_A \mid C(X) = H_i)$ can be broken down as shown in Equations 2.7 and 2.8.

$$P(X \in S_1 \mid C(X) = H_i) = P(X \in S_1 \cap S_A \mid C(X) = H_i) + P(X \in S_1 \cap S_A^c \mid C(X) = H_i) \tag{2.7}$$

$$P(X \in S_A \mid C(X) = H_i) = P(X \in S_A \cap S_1 \mid C(x) = H_i) + P(X \in S_A \cap S_1^c \mid C(X) = H_i) \tag{2.8}$$

Since the intersection operation is commutative, solving for equality of Equations 2.7 and 2.8 gives the following condition, shown below in Equation 2.9, that must be satisfied in order for both decision rules have the same significance level.

$$P(X \in S_1 \cap S_A^c \mid C(X) = H_0) = P(X \in S_A \cap S_1^c \mid C(X) = H_0) \tag{2.9}$$

By definition, the LRT decision rule is more powerful than the alternative if $P(X \in S_1 \mid C(X) = H_1) \geq P(X \in S_A \mid C(X) = H_1)$, substituting Equations 2.7 and 2.8 and simplifying, yields the following inequality.

$$P(X \in S_1 \cap S_A^c \mid C(X) = H_1) \geq P(X \in S_A \cap S_1^c \mid C(X) = H_1) \tag{2.10}$$

Satisfaction of the inequality in Equation 2.10 is both necessary and sufficient for the LRT to be more powerful than a given alternative decision rule. Therefore, demonstration that the condition holds for an arbitrary alternative, which has equal significance level, proves that the LRT is the uniformly most powerful decision rule for a given observation size N.

The remainder of this section will show, using the equations introduced above, that the inequality in Equation 2.10 holds for the LRT decision rule against an arbitrary alternative.

The condition for membership of set S_1 from Equation 2.4 means that for each $X \in S_1$ it must be the case that $L(X \mid H_1) \geq \eta L(X \mid H_0)$, therefore, the following must also be true, $\sum_{X \in S_1} L(X \mid H_1) \geq \eta \sum_{X \in S_1} L(X \mid H_0)$. From this and Equation 2.6, the following inequality can be derived about $P(X \in S \mid C(X) = H_i)$.

$$\begin{aligned} P(X \in S_1 \cap S_A^c \mid C(X) = H_1) &= \sum_{X \in S_A \cap S_1^c} L(X \mid H_1) \\ \sum_{X \in S_A \cap S_1^c} L(X \mid H_1) &\geq \eta \sum_{X \in S_A \cap S_1^c} L(X \mid H_0) \\ \eta \sum_{X \in S_A \cap S_1^c} L(X \mid H_0) &= \eta P(X \in S_1 \cap S_A^c \mid C(X) = H_0) \end{aligned} \tag{2.11}$$

However, since both rules have the same significance level, Equations 2.6 and 2.9 can be used to further expand the inequality in Equation 2.11 yielding the following result.

$$\begin{aligned} \eta P(X \in S_1 \cap S_A^c \mid C(X) = H_0) &= \eta P(X \in S_A \cap S_1^c \mid C(X) = H_0) \\ &= \eta \sum_{X \in S_A \cap S_1^c} L(X \mid H_0) \end{aligned} \tag{2.12}$$

Since the sets S_0 and S_1 are mutually exclusive and exhaustive $S_1^c = S_0$, thus membership of $S_A \cap S_1^c$ implies membership of the set S_0. Using the condition of membership of the set S_0 from Equation 2.5 it can be seen that for all $X \in S_0$ it must be true that $\eta L(X \mid H_0) \geq \eta L(X \mid H_1)$. Thus $\eta \sum_{S_0} L(X \mid H_0) \geq \sum_{S_1} \eta L(X \mid H_1)$ must also be true. Therefore, the inequality in Equation 2.12 can be further expanded to yield the following result.

$$\eta \sum_{X \in S_A \cap S_1^c} L(X \mid H_0) \geq \sum_{X \in S_A \cap S_1^c} L(X \mid H_1) = P(X \in S_A \cap S_1^c \mid C(X) = H_1) \tag{2.13}$$

Therefore, the condition $P(S_1 \cap S_A^c) \geq P(S_A \cap S_1^c, H_1)$ is satisfied and so $P(S_1, H_1) \geq P(S_A, H_1)$ must be true, thus, the LRT is uniformly most powerful for testing a simple hypothesis against a single alternative with fixed sample size n.

2.2 Sequential Probability Ratio Tests

The Sequential Probability Ratio Test (SPRT) builds upon the method outlined for the LRT by allowing the number of trials performed to vary depending on the level of uncertainty that remains. The benefit of this sequential approach is twofold. Firstly, the decision rule can be parameterised such that a maximum level of error can be specified for both α (rejection of a true null hypothesis) and β (acceptance of a false null hypothesis) errors. Secondly, it is expected that on average the number of observations required to make a decision is reduced.

At each time interval, m, the SPRT must decide whether to accept the null hypothesis, reject the null hypothesis or continue testing by making an additional observation and then reapplying the decision rule.

The SPRT decision, like the LRT decision, is based on the likelihood ratio of the two hypotheses. However, unlike the LRT which calculates the likelihood ratio once over the entire sequence of N observations, the SPRT calculates the likelihood ratio at each of the n intervals. Where N is a predetermined parameter of the LRT decision rule and n is the variable time index at which the SPRT terminates, which is dependent on the observations that are made.

The decision rule for the SPRT is shown below in Equation 2.14.

$$\delta_m(X_m) = \left\{ \begin{array}{ll} H_0 & \text{if } \Lambda_m(X) < B \\ H_1 & \text{if } \Lambda_m(X) > A \\ \delta_{m+1}(X_{m+1}) & \text{if } B \leq \Lambda_m(X) \leq A \end{array} \right\} \tag{2.14}$$

Where, $\Lambda_m(X_m)$ is the likelihood ratio at interval m, which is calculated over the set of observations, X_m, made at intervals $1..m$. Rather than using Equation 2.2 to calculate the likelihood ratio, as was done in the LRT, the calculation can be performed recursively, as shown in Equation 2.15.

$$\Lambda_m(X_m) = \Lambda_{m-1}(X_{m-1}) \frac{P(x_m \mid H_1)}{P(x_m \mid H_0)} \tag{2.15}$$

The initial value of the likelihood ratio, $\Lambda_0(X_0)$, is either given by the ratio of the prior probabilities of the two hypotheses, $\frac{\pi_1}{\pi_0}$, or if prior probabilities are unavailable or not applicable a value of $\Lambda_0(X_0) = 1$ should be used.

At each interval, m, at which the SPRT is applied the m dimensional observation space is divided into three exhaustive mutually exclusive sets, one for each for the possible decision outcomes. These sets are denoted S_{m0} and S_{m1} and S_{mc}, where S_{m0} is the set for which H_0 holds, S_{m1} is the set for which H_0 is rejected and S_{mc} is the set for which more observations

are required. A more formal definition of these sets is given below in Equations 2.16, 2.17, and 2.18.

$$S_{m1} = (S_{m0} \cup S_{mc})^c = \left\{ X : \frac{L(X \mid H_1)}{L(X \mid H_0)} \geq A \right\} \tag{2.16}$$

$$S_{m0} = (S_{m1} \cup S_{mc})^c = \left\{ X : \frac{L(X \mid H_1)}{L(X \mid H_0)} \leq B \right\} \tag{2.17}$$

$$S_{mc} = (S_{m0} \cup S_{m1})^c = \left\{ X : B < \frac{L(X \mid H_1)}{L(X \mid H_0)} < A \right\} \tag{2.18}$$

2.2.1 Decision Thresholds and Expected Error

As with the LRT decision rule, the threshold values that are used in the SPRT directly influence the expected error rates. The single decision threshold, η, that was used in the LRT decision rule allowed a maximum expected error rate to be specified for either α or β errors. The addition of a second decision threshold in the SPRT allows maximum error rates to be specified for both types of error. Thus, whereas the LRT rule allowed a single error type to be made arbitrarily small at the expense of the other error rate, the SPRT rule allows both error rates to be set arbitrarily small with the caveat that, as the error rates tend to zero, the number of observations required to make a decision tends to infinity.

Unlike the LRT, in which the error rates are dependent on both the threshold value η and the probability distribution over the observable variables, the maximum error rates in the SPRT are determined solely by the threshold values A and B.

Before relating the thresholds A and B with the error rates α and β we must first define the termination sets S_0 and S_1, where S_i contains all the observation vectors that result in the termination of the SPRT decision rule and the acceptance of hypothesis H_i. The formal definition of these sets is given below in Equations 2.19 and 2.20.

$$S_0 = \bigcup_{m=1}^{n} S_{m0} \tag{2.19}$$

$$S_1 = \bigcup_{m=1}^{n} S_{m1} \tag{2.20}$$

Where n is the eventual stopping time of the SPRT decision rule. Since, the sets S_{m0} and S_{m1} are mutually exclusive and the dimension i of an observation X excludes it from being a member of any of the sets S_{j0}, S_{j1}, and S_{jc} for which $i \neq j$, the supersets S_0 and S_1 are also mutually exclusive.

Assuming that X_n is a sequence of observations that will result in the SPRT terminating, the inequalities in Equations 2.21 and 2.22 must hold.

$$P(X_n \in S_1 \mid C(X_n) = H_1) \geq AP(X_n \in S_1 \mid C(X_n) = H_0) \tag{2.21}$$

$$P(X_n \in S_0 \mid C(X_n) = H_1) \leq BP(X_n \in S_0 \mid C(X_n) = H_0) \tag{2.22}$$

However, since $P(X_n \in S_1 \mid C(X_n) = H_0)$ is the probability of falsely rejecting the null hypothesis and $P(X_n \in S_0 \mid C(X_n) = H_1)$ which is denoted β, $P(X_n \in S_1 \mid C(X_n) = H_0) = \alpha$ and $P(X_n \in S_0 \mid C(X_n) = H_1) = \beta$. Furthermore, in order for the SPRT process to terminate, the observation vector X_n must be a member of either S_0 or S_1, thus Equation 2.23 must be true.

$$P(X_n \in S_0 \mid C(X_n) = H_i) + P(X_n \in S_1 \mid C(X_n) = H_i) = 1 \tag{2.23}$$

It follows from Equation 2.23 that $P(X_n \in S_0 \mid C(X_n) = H_0) = 1 - \alpha$ and $P(X_n \in S_1 \mid C(X_n) = H_1) = 1 - \beta$. Thus, by rearranging the inequalities in Equations 2.21 and 2.22 and substituting for $P(X_n \in S_0 \mid C(X_n) = H_0)$ and $P(X_n \in S_1 \mid C(X_n) = H_1)$ the thresholds A and B can be defined in terms of the error rates α and β, as shown below in Equation 2.24 and 2.25.

$$A \leq \frac{1-\beta}{\alpha} \tag{2.24}$$

$$B \geq \frac{\beta}{1-\alpha} \tag{2.25}$$

For simplicity, values of A and B are generally chosen such that $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$.

2.2.2 Illustration of the SPRT

The SPRT is generally easier to visualise when considering the logarithm of the likelihood function, $\log \Lambda_m(X_m)$, which can be calculated iteratively using the formula below.

$$\log \Lambda_m(X_m) = \log \Lambda_{m-1}(X_{m-1}) + \log P(x_m \mid H_1) - \log P(x_m \mid H_0) \tag{2.26}$$

When considering the logarithm of the likelihood function, the decision boundaries A and B must be mapped to their log space equivalents $\log A$ and $\log B$. From the decision rule in Equation 2.14 it can be seen that $A \geq B$ therefore, since the logarithm function is monotonic, it must also be true that $\log A \geq \log B$.

Since $P(x_m \mid H_i)$ is a probability with value in the range $[0, 1]$, the value of the logarithm $\log P(x_m \mid H_i)$ must lie in the range $[-\infty, 0]$, with $\log P(x_m \mid H_i) \to 0$ as $P(x_m \mid H_i) \to 1$. From this and the likelihood function in Equation 2.26 it can be seen that it is the improbability of a set of observations occurring under hypothesis H_i that causes the likelihood ratio to tend towards the boundary of the alternative hypothesis.

In order to illustrate the SPRT graphically, the SPRT has been implemented and applied to an artificial situation in which the aim is to select from two alternative probability distributions (hypotheses), the distribution that best describes the data observed from a given data source.

The data source used in the simulation has a Gaussian distribution with a mean of 4.0 and standard deviation of 1.0. The two hypotheses are both Gaussian distributions with standard

deviation of 1.0 but have differing means. Hypothesis H_0 is that the mean is 4.0 and hypothesis H_1 is that the mean is 3.95. The prior probabilities of the hypotheses are assumed to be equal, and thus inconsequential to the decision process. Finally we chose decision boundaries such that the expected error rate will be at most 0.05 for both types of error.

The graph in Figure 2.1 shows how the likelihood ratio changes as more data is integrated into the decision process. Looking at the graph it can be seen that as the decision process progresses the likelihood ratio drifts towards the lower decision boundary B. Using the SPRT decision rule in Equation 2.14, it can be seen that this corresponds to the acceptance of the null hypothesis, which, in the simulation described above, is the correct decision. Whilst the likelihood ratio does tend towards the correct decision boundary, the rate at which it does so varies from observation to observation, with some observations resulting in the ratio moving away from the boundary. This variability in the magnitude and direction of changes to the likelihood ratio makes the decision process similar to a random walk which with specified error tends towards the decision boundary corresponding to the hypothesis that most closely describes the data source.

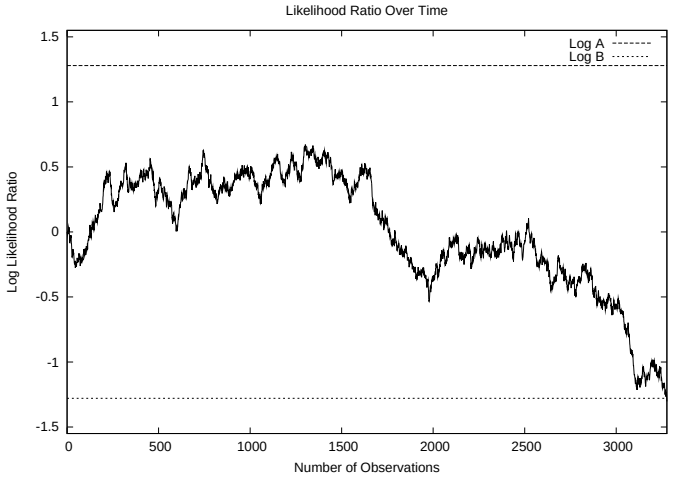


Figure 2.1: Change in Likelihood Ratio

2.3 Summary

In this section we have introduced and discussed two hypothesis testing techniques, namely, the Likelihood Ratio Test and the Sequential Probability Ratio Test. In the case of the Likelihood Ratio Test, we have demonstrated that for a given significance level α and observation size n, the error rate β is minimised.

Next, we showed how the SPRT builds upon the methodology of the LRT, resulting in a method that sequentially incorporates observations into the decision making process until the level of

uncertainty falls below a given threshold value. Thus, resulting in the SPRT using the smallest number of observations necessary, in order to ensure that the expected error rates are below predetermined levels α and β.

Finally, we illustrated how the sequential integration of observations into the decision process affects the level of uncertainty in the decision process. Furthermore, we considered how this changing uncertainty is similar to a random walk which, with specified error rate, drifts towards the decision boundary associated with the correct hypothesis.

3 Resource Allocation

The problem of optimally allocating a finite number of resources between a larger number of competing consumers is one that arises in a host of fields, including economics, medical sciences and computing.

In this section we concentrate on the situation in which a single resource is allocated between a set of consumers over a number of time intervals. A practical example of this situation is process scheduling in a single processor computing system. In this example the single processor is the resource that must be allocated between a number of processes, by selecting at each interval a process to be executed.

We begin by first introducing the analogy of a bandit process, which is commonly used as an abstraction of a resource allocation problem. Following this, we introduce the dynamic programming solution to the 'bandit problem' and discuss the limitations of this approach. Next, we introduce and discuss a solution to the bandit problem known as the Gittins index[4, 3]. Finally, we conclude the section with a summary of the bandit problem and the solutions to the problem that have been discussed.

3.1 Bandit Processes

A multi-armed bandit is a model of an abstract situation in which a resource must be dynamically allocated amongst a set of consumers for a number (possibly infinite) of time intervals. The bandit problem is so called because of the similarities to a gambler allocating their time (resource) between a number of slot machines (consumers), which are known as a one-armed bandits.

By choosing to play bandit i, at time t, the gambler either succeeds with probability θ_i and receives discounted reward a^t, where $0 < a < 1$, or fails and receives no reward. The distribution of successes and failures for each bandit is assumed to be unknown and independent of all other bandits.

The prior probability distribution for each bandit process is assumed to be a beta distribution $\pi_i(\theta) = Beta_i(\alpha_i(0), \beta_i(0), \theta)$. Where, $\alpha_i(0)$ and $\beta_i(0)$ are the hyper-parameters, or state of the bandit, which affect the distribution of possible values of θ. Initial values of $\alpha_i(0) = \beta_i(0) = 1$ result in an even distribution for all values of $0 \leq \theta \leq 1$.

Given that the likelihood function for the bandit process is a Bernoulli distribution and the prior probability function is a Beta distribution, the posterior probability, $P(\theta \mid x)$, is also a Beta distribution. Thus, at each interval t at which the gambler is playing, the prior probability distribution over θ_i for each bandit is a Beta distribution with hyper-parameters $\alpha_i(t) = \alpha_i(0) + l_i(t)$ and $\beta_i(0) = \beta_i(0) + m_i$, where $l_i(t)$ and $m_i(t)$ are the number of successes and failures that have been observed for bandit i up to time t.

$$\boldsymbol{E}(Beta(\alpha, \beta)) = \frac{\alpha}{\alpha + \beta} \tag{3.1}$$

The gambler's problem is to choose a sequence of bandits to play such that the total expected reward is maximised. Since the distribution of rewards is unknown, a trade-off must be made between gaining information through exploration of the possible actions and exploiting knowledge such that the action that is believed to be optimal is selected.

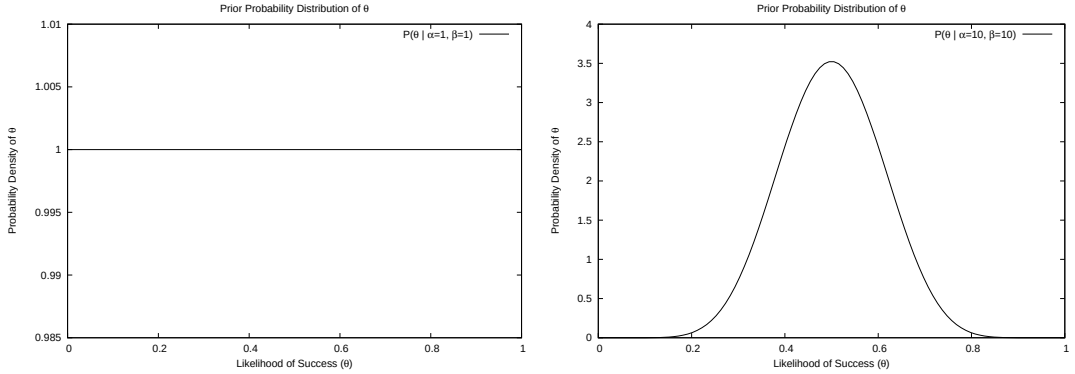
One possible solution to the decision problem is to use a decision rule that selects the bandit process at each interval that maximises the expected reward, as shown in Equation 3.2 below.

$$\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \underset{i}{\operatorname{argmax}} \frac{\alpha_i}{\alpha_i + \beta_i} \tag{3.2}$$

This decision rule selects the bandit that maximises the expected gain in each interval in which it is applied. Whilst it seems that maximising the expected gain in each interval would result in the maximum expected total reward it does not take into account the value that could be gained through exploration.

Consider the following situation in which there are two bandit processes, the first, bandit B_1, has parameters $\alpha = \beta = 1$ and the second, bandit B_2, has parameters $\alpha = \beta = 10$. From these parameters and Equation 3.1, it can be seen that the expected reward from the two bandits is identical. However, looking at the prior probability density functions, shown in Figures 1(a) and 1(b), it can be seen that there is much more uncertainty about success rate θ for bandit B_1. Therefore, whilst the immediate reward is expected to be identical for both bandit processes, the additional information gained from selecting bandit B_1 could impact the expected future gain. In this situation the optimality of selecting bandit B_1 is relatively obvious since there is no expected loss in the current interval from selecting B_1 and there is the potential to gain from making a more informed decision at subsequent intervals. However, in cases where the expected rewards are not identical it is not so obvious which bandit is likely to maximise the total reward over future intervals.

The following equation uses dynamic programming techniques to take into account the potential for the expected success probability to change as a result of information gain. Equation 3.3, shown below, provides a method for calculating the expected reward $R(\boldsymbol{\alpha}, \boldsymbol{\beta})$ given that the



(a) Beta Distribution with $\alpha = \beta = 1$

(b) Beta Distribution with $\alpha = \beta = 10$

Figure 3.1: Prior Probability Distributions over θ for two Bandits, B_1 and B_2

expected success probability changes as a result of information gain.

$$R(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \max \left\{ \frac{\alpha_i}{\alpha_i + \beta_i}[1 + aR(\text{inc}(\boldsymbol{\alpha}, i), \boldsymbol{\beta})] + \frac{\beta_i}{\alpha_i + \beta_i}[aR(\boldsymbol{\alpha}, \text{inc}(\boldsymbol{\beta}, i))] \right\} \tag{3.3}$$

Where, $\text{inc}(\boldsymbol{X}, i)$ is a function that increments the ith element of the vector $\boldsymbol{X}$.

The decision rule outlined below in Equation 3.4 utilises the reward function, $R(\boldsymbol{\alpha}, \boldsymbol{\beta})$, outlined above to determine the bandit that is expected to maximise the total reward received across the time horizon for which the decision process runs.

$$\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{argmax} \left\{ \frac{\alpha_i}{\alpha_i + \beta_i}[1 + aR(\text{inc}(\boldsymbol{\alpha}, i), \boldsymbol{\beta})] + \frac{\beta_i}{\alpha_i + \beta_i}[aR(\boldsymbol{\alpha}, \text{inc}(\boldsymbol{\beta}, i))] \right\} \tag{3.4}$$

Whilst this dynamic programming approach to solving the bandit problem provides a valid solution, calculating the reward, $R(\boldsymbol{\alpha}, \boldsymbol{\beta})$, over the process's time horizon becomes computationally infeasible as the number of bandits is increased.

3.2 Gittins Indices

The previous section demonstrated how dynamic programming techniques can be used to select a bandit to play at each time interval in order to maximise the expected total reward across all intervals. However, as shown in the previous section these techniques become computationally infeasible as the number of bandits in the system is increased.

In order to overcome the computational complexity of the dynamic programming solution outlined in the previous section, the approach taken by Gittins is to consider each bandit, B_i, as if it were in an allocation problem with a single alternative bandit, B_{iA} that has fixed probability of success λ_i.

The value of λ_i is calculated such that it is the critical probability of success, known as the Gittins Index, at which the reward given that bandit B_i is selected in the next interval is equal to selecting known bandit B_{iA} for each interval in the time horizon. That is, given the current Beta distribution for bandit B_i, at what point is the guaranteed reward from B_{iA} equal to the potential of B_i.

For a bandit process B_{iA} with fixed probability λ_i, the payout can be considered to be λ_i at each time interval. Thus the critical probability λ_i occurs when the following equality is satisfied.

$$\lambda_i \sum_{t=1}^{T} a^{t-1} = \frac{\alpha}{\alpha+\beta}\left[1 + aR(\alpha+1, \beta, T-1, a, \lambda)\right] + \frac{\beta}{\alpha+\beta} aR(\alpha, \beta+1, T-1, a, \lambda) \tag{3.5}$$

Where, $R(\alpha, \beta, T, a, \lambda)$ is the reward function calculated over the time interval $0..T$, with discount a and initial hyper-parameters α and β, which can be calculated using the algorithm below.

Algorithm 1: Recursive Implementation of Reward Function

Input: α, β, T, a, λ
Output: $R(\alpha, \beta, T, a, \lambda)$
if *N == 1* **then**
    reward = $\max\left(\lambda, \frac{\alpha}{\alpha+\beta}\right)$;
else
    reward = $\frac{\alpha}{\alpha+\beta}\left[1 + aR(\alpha+1, \beta, t-1, a, \lambda)\right] + \frac{\beta}{\alpha+\beta} aR(\alpha, \beta+1, t-1, a, \lambda)$;
    reward = $\max\left(reward, \lambda \sum_{i=1}^{t} a^{i-1}\right)$;
end
return reward;

From Equation 3.3, the value of λ_i for which the expect rewards are equal for selecting either bandit B_i or alternative B_{iA} can be found using the following equation.

$$\lambda_i = v(\alpha_i, \beta_i, T, a) = \frac{\frac{\alpha}{\alpha+\beta}\left[1 + aR(\alpha+1, \beta, T-1, a, \lambda)\right] + \frac{\beta}{\alpha+\beta} aR(\alpha, \beta+1, T-1, a, \lambda)}{\sum_{t=1}^{T} a^{t-1}} \tag{3.6}$$

However, since calculation of reward using the function $R(\alpha, \beta, t, a, \lambda)$ requires a maximisation step involving λ_i at each interval, the equation cannot be solved directly and must therefore be resolved using a 'brute force' algorithm, such as the one outlined below.

Once the indices have been calculated for each of the n bandits the optimal decision at each time interval can be made by selecting the bandit with the greatest value of λ. This decision rule is formalised below in Equation 3.7.

$$\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}, N) = \underset{i}{\operatorname{argmax}}\, v(\alpha_i, \beta_i, T) \tag{3.7}$$

The graph in Figure 3.2 shows how the Gittins index of a bandit with hyper-parameters $\alpha = \beta = 1$ varies as the time horizon of the decision problem increases for a number of discount

Algorithm 2: Function for Calculating Gittins Indices

Input: α, β, T, a, resolution
Output: $v(\alpha, \beta, T, a)$
terminate = false;
$\lambda = \frac{\alpha}{\alpha+\beta}$;
while *terminate == false* **do**
  reward = $\frac{\alpha}{\alpha+\beta}\left[1 + a * R(\alpha+1, \beta, T-1, a, \lambda)\right] + \frac{\beta}{\alpha+\beta} aR(\alpha, \beta+1, T-1, a, \lambda)$;
  if $\lambda \sum_{t=1}^{T} a^{t-1} > reward$ **then**
    terminate = true;
    $\lambda = \lambda - resolution$;
  else
    $\lambda = \lambda + resolution$;
  end
end
return λ;

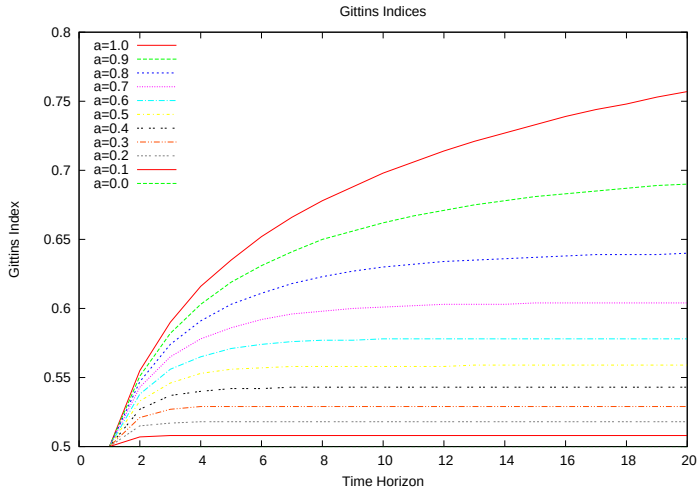
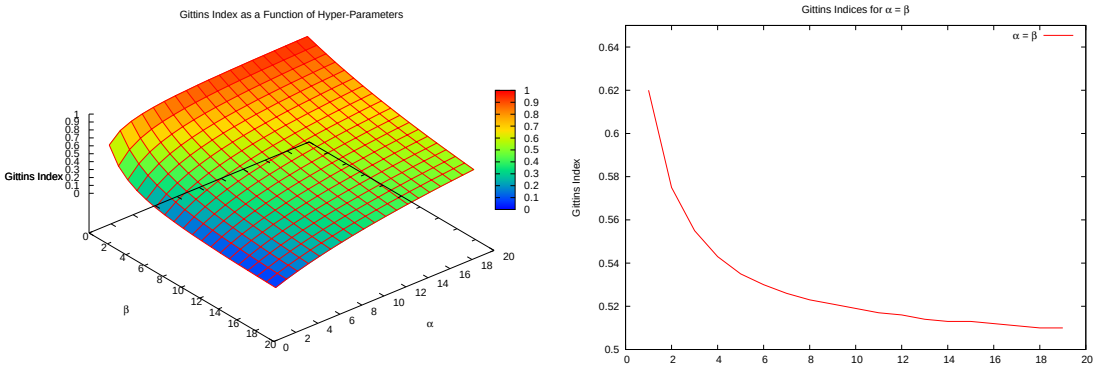


Figure 3.2: Gittins Indices

rates. The value of the discount rate that is used affects how much emphasis is placed on future rewards which are received. Larger discount values result in a more optimistic outlook on the potential of the bandit. This is due to the information gained allowing a more informed decision to be made at all subsequent intervals, and thus, the optimal bandit is more likely to be selected in the future.

The graphs in Figures 3(a) and 3(b) shows how the Gittins index varies as a function of the hyper-parameters α and β for a fixed discount rate of 0.75 and time horizon $T = 20$. From the cross-section of the surface taken when $\alpha = \beta$, shown in Figure 3(b), it can be seen that the

index becomes less optimistic about the potential of the Bandit B_i and approaches the expected value of $\frac{\alpha}{\alpha+\beta}$ as the number of samples $N = \alpha + \beta$ is increased.



(a) Gittins Indices as a function of α and β

(b) Cross Section of Figure 3(a) at $\alpha = \beta$

Figure 3.3: Gittins Indices

3.3 Summary

In this section, we have introduced the concept of a bandit process as an abstract model of a situation in which a single resource needs to be optimally allocated between a number of competing consumers, such that the reward recieved over the time horizon of the problem is maximised.

Next, we introduced a simple myopic decision rule, in Equation 3.2, that seeks only to maximise the reward at each interval in the decision process. Following this, we considered how dynamic programming techniques can be used to take into account the potential for information gain from each of the bandits and how it may affect future decisions and the reward received. In addition, we touched upon the computational complexity of the dynamic programming approach and how it becomes computationally intractable when considering even a small number of bandits.

Following this, we showed how the inherent complexity of the bandit problem can be overcome by considering each bandit individually, as if it were in a decision problem with a single alternative bandit with fixed probability of success λ. The value of λ, known as the Gittins index, is then calculated as the critical probability at which point it would be optimal to switch between bandits. The Gittins index incorporates the current expected success probability for the bandit and the level of optimism that there is for this value to improve as information is gained. Finally, we showed some examples of how the Gittins indices change as a function of the parameters of the model and the bandit for which it is calculated.

4 Summary

In this text we have provided an overview of two areas of statistical decision theory that we believe could form an integral part of a successful decentralised methodology for controlling complex distributed systems.

We began by discussing two hypothesis testing techniques, namely, the Likelihood Ratio Test and the Sequential Probability Ratio Test. In the case of the Likelihood Ratio Test, we have demonstrated that for a given significance level α and observation size n, the error rate β is minimised. Following this, for the Sequential Probability Ratio test we saw how for a given significance level α and type 2 error rate β the number of observations required to satisfy these limits is minimised.

Next, we considered a resource allocation problem in which a single resource must be shared amongst a number of competing consumers such that the reward received over the time horizon of the problem is maximised. Following the introduction of the problem we considered a myopic solution to the allocation problem that maximises the expected reward at each decision interval, based on current knowledge about the problem space. The pitfalls of this approach lead us to consider a dynamic programming approach, which whilst computationally complex does allows the expected reward across the time horizon of the problem to be maximised by taking into account the amount of uncertainty that remains about each of the alternative decisions. Finally, we considered how the complexities of the dynamic programming approach can be overcome to provide a solution that is both forward thinking and computationally feasible.

References

[1] Pranab Kumar Sen Bhaskar Kumar Ghosh. *Handbook of Sequential Analysis*. Marcel Dekker, 1991.

[2] Rafal Bogacz, Eric Brown, Jeffrey Moehlis, Phil Holmes, and Jonathan D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, 113(4):700–765, October 2006.

[3] J. C. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 1989.

[4] J. C. Gittins and D. M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[5] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[6] Richard Weber. On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.