Exploitation vs. Exploration: Choosing a Supplier in an Environment of Incomplete Information ⋆

Rina Azoulay-Schwartz [a] Sarit Kraus [b,c,d]
Jonathan Wilkenfeld [e,c]

[a] *Department of Computer Science Bar-Ilan University, Ramat-Gan, 52900 Israel, Phone: +972-3-9324167; Fax:972-3-7360498; e-mail:schwart@cs.biu.ac.il*

[b] *Department of Computer Science Bar-Ilan University, Ramat-Gan, 52900 Israel, Phone: +972-3-5318762; Fax:972-3-7360498; e-mail: sarit@cs.biu.ac.il*

[c] *Institute for Advanced Computer Studies University of Maryland, College Park, MD 20742*

[d] **Correspondence author: Sarit Kraus**

[e] *Department of Government and Politics University of Maryland, College Park, MD 20742; Phone:+1-301-3147703; Fax: +1-301-314-9256; e-mail: jwilkenf@gvpt.umd.edu*

Abstract

An agent operating in the real world must often choose between maximizing its expected utility according to its current knowledge about the world, and trying to learn more about the world, since this may improve its future gains. This problem is known as the trade-off between exploitation and exploration.

In this research, we consider this problem in the context of electronic commerce. An agent intends to buy a particular product (goods or service). There are several potential suppliers of this product, but they differ in their quality, and in the price charged. The buyer cannot observe the average quality of each product, but he has some knowledge about the quality of previous goods purchased from the suppliers. On the one hand, the buyer is motivated to buy the goods from the supplier with the highest expected product quality, deducting the product price. However, when buying from a lesser known supplier, the buyer can learn about its quality and this can help him in the future, when he will purchase more products of this type.

We show the similarity of the suppliers problem to the k-armed bandit problem, and we suggest solving the suppliers problem by evaluating Gittins indices and choosing the supplier with the optimal index.

We demonstrate how Gittins indices are calculated in real world situations, where deals of different magnitudes may exist, and where product prices may vary. Finally, we consider the existence of suppliers with no history, and suggest how the original Gittins indices can be adapted in order to consider this extension.

Key words: Electronic commerce, Decision making, Incomplete information, Choosing a supplier.

1 Introduction

An agent in an electronic market often has to choose among several suppliers of a product or a service. The suppliers differ in the mean quality of the products they sell. The agent does not know with certainty the mean quality of each supplier, but he has some knowledge about qualities of previous items sold by each suppliers. We consider situations where the agent needs to repeatedly buy items, and he would like to buy high quality items for the lowest possible prices.

The problem of unknown quality of items appears also in traditional markets, but is much more pronounced in e-commerce since the buyer cannot view the product before purchasing it, and cannot form a personal impression of the supplier [10,13]. When automated agents represent the buyer, this problem is even more pronounced. Thus, there is considerable uncertainty about the quality of the goods, delivery time, and the reliability of the supplier.

The price of the goods cannot identify its quality, since unknown suppliers may exist, which may sell goods of high quality, but charge low prices since they are not known. Consider, for example, the market for books. The largest on-line supplier of books is Amazon [1], but there are other on-line suppliers that may sell the same books at a reduced price, and may also provide a better quality product, if quality is measured as delivery time, since a smaller company has fewer transactions, while the well-known supplier may be backlogged with orders, and delivery may sometimes be delayed. Similarly, empirical tests [4] show that price differences exist in the electronic market of airline ticket offerings.

In situations where the unknown supplier may sell superior goods at a lower price, the agents have to use an intelligent strategy in order to choose the most beneficial deals. Under these circumstances, the history of past transac-

⋆ This material is based upon work supported in part by NSF under Grant No. IIS-9820657 and IIS-0208608.

Email address: sarit@umiacs.umd.edu (Sarit Kraus).

tions with this supplier is crucial information in evaluating the quality of the products or services it sells.

On the one hand, the best option for the agent is to chose the supplier which maximizes his expected utility, i.e., with a low price and high average quality. However, there may be situations in which the agent will prefer to buy from an unknown supplier, in order to explore the quality of this supplier, and this may provide future benefits by buying from better suppliers.

The dilemma of the buyer is whether to choose the best known supplier or to try other suppliers in order to learn about their quality, in order to improve future gains. This dilemma is called in the literature the trade-off between exploitation and exploration. For this kind of problems, Gittins [5] suggests a method for calculating an index for each alternative, which considers the expected gains from choosing it, taking into account future gains from obtaining additional information. Gittins proves that the alternative with the highest calculated index is the optimal choice. However, there are several adaptations that are necessary in order to apply Gittins indices to real world situations such as those in the above examples. In this paper, we consider the problem of applying the Gittins technique to real problems of choosing among alternatives, and we demonstrate it on the problem of choosing an on-line supplier.

The paper is organized as follows. In Section 2 we discuss related work, and in Section 3, the formal model is presented. A theoretical background about Gittins indices is presented in Section 3.1 and how to solve the supplier problem using Gittins indices is explained in Section 3.2. Finally, in Section 4 we provide conclusions and suggestions for future extensions.

2 Related Work

The issue of product quality is considered in industrial organization literature [12]. Goods are classified into three different types: search goods are goods with a quality ascertained by consumers before a purchase, but the consumers differ in the way they value this quality; experience goods are goods with a quality that is learned after the goods have been bought; and credence goods, have a quality that can rarely be learned, even after consumption. In our model, we assume experience goods. Customers in our model differ according to their information about the firms, but we assume that two consumers will give the same evaluation to the goods if both have the same knowledge about their quality.

The quality of experience goods is learned only after an item is bought, and the question is how customers learn the quality of the goods, and what incentive firms have to supply high quality. If a one shot relationship is considered, then a problem of moral hazard exists, and the firms have an incentive to cut quality to the lowest possible level, because the market price cannot respond to the unobservable quality.

If repeated interactions are considered, as in our case, then there is an incentive to provide high quality goods. In this case, a firm can change the price in the short run, in order to signal its quality to the buyers. It may signal low prices, since it is able to lose money in order to earn it back in future interactions. However, a high quality producer may want to signal by a high price, since the high quality product is more costly to produce. Thus, a firm may signal high quality by either reducing prices in the short run, or by charging high prices.

To summarize, different strategies of the suppliers may be in equilibrium. In this paper, we do not consider a particular strategy of the supplier. Alternatively, we consider the real world situation, where different suppliers use different strategies, and the agent does not have full information about the product costs, so he cannot infer the quality of the goods from its price. We also do not discuss the problem of choosing quality or price, which is faced by the supplier, but we discuss what the best strategy to be taken by the consumers is, in a world of heterogeneous suppliers, with unknown price strategies and unknown qualities, which can be learned only by consuming their goods.

Empirical studies find differences in prices in on-line markets. Eric et al. [4] use data on airline ticket offering of online travel agents, and find that different travel agents offer tickets with different prices and characteristics when given the same customer request. Even after accounting for differences in ticket quality, ticket prices vary by as much as 18% across online travel agents. They explain this result by the existence of search costs of travel agents, and the costs involved in switching to a new agent (signing up, etc.). There is also uncertainty about which online travel agent will truly provide the best flight for the consumers preference. Their results also show that service differentiation is a key strategic component of online suppliers that offer access to heterogeneous goods. In this paper, we will consider this real world situation, and we will provide a strategy for the buyer in order to decide from which supplier to buy, given the search costs and switching costs, and given the uncertainty when trying an unknown supplier.

Since, as explained above, we consider experience goods, each time an agent needs the good it has to decide between choosing the best known supplier or trying other suppliers in order to learn about the quality of their products. When making this decision, the buyer can use techniques taken from Reinforcement learning in order to get better decisions. Reinforcement learning is learning how to behave through trial-and-error interactions with a dynamic environment. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them [11]. Kaelbling [7] divides the methods used for solving this trade-off into two main parts. First, there are ad-hoc heuristic techniques, that do not promise optimal solutions, but are computationally tractable, and do not need strong assumptions about the environment, as required by the formally justified techniques. Second, there are formally justified techniques, which may require stronger assumptions, but they are proven to be optimal (or near to optimal) given these assumptions. Among these techniques are: the dynamic programming approach [2], which is expensive in time and space; the Gittins allocation indices [5,14], which can be used in cases of immediate reward; and the automata learning approach [6], which is appropriate for cases of success and failure. In this paper, we adapt the formal approach for the exploration-exploitation trade-off. We prefer to consider this approach since it is formally proven to yield the highest expected utility. In particular, we suggest the use of the Gittins indices method for the problem of choosing the best supplier.

The dilemma of whether to choose the best known policy, or to try to learn other alternatives, appears in several domains, and reinforcement learning was successfully used to solve this trade-off in different situations. Schaerf et al. [9] study the problem of load balancing in a distributed system. In their model, the quality of a service provided by a resource at a given time deteriorates with the number of agents using it at that time. The rule by which agents select a resource for a new job is probabilistic, and clearly biased towards resources that have performed well in the past. The strength of the bias depends on n - the history length. Their experiments show how the interaction between different agent types affects the system's efficiency as well as the individual agent's efficiency.

Salganicoff and Ungar [8] use Gittins indices to select actions which optimally trade-off exploration and exploitation, in a continuous action space, where the result of an action is either a success or a failure. They combine Gittins indices with decision trees to develop a mapping from state and action to success or failure of that action. Salganicoff and Ungar illustrate their technique with a robot that learns to grasp objects.

In this paper, we use Gittins indices to optimally select a supplier. In our domain, the action space is finite (number of suppliers), while the action result is continuous. Moreover, we consider also some different issues, such as the dynamic arrival of suppliers, the stochastic arrival of purchase requests, and different sizes of deals.

The exploration vs exploitation problem is somewhat similar to the decision a planner agent has to make when planning under uncertainty [3]. The planner needs to decide whether to act upon the best plan it has so far, or to try to improve his plan. Dean and Wellman proposed a decision theory approach to this problem. Their method is appropriate for situations where different parameters can be observed in order to improve the information about the environment, and to improve the decision made. The case studied in this paper is different in the fact that the information revealed is, in any stage, about one of alternative, it can be revealed only by using this alternative, and the value of this information is different each time the alternative is used. Thus, it is not enough to reveal the information once, but it should be used at different occasions in order to reach satisfactory knowledge about this alternative.

3 The Formal Model

Consider a market of N suppliers who sell a given item. The mean quality of an item sold by supplier i is μ_i, its standard deviation is σ_i and its price is $price_i$. An agent needs to buy the goods often at discrete time periods. The probability that he will need to buy the item at a given time period t is $0 < p < 1$. Each time the agent buys the item, he must choose among the N suppliers. The agent does not know the mean quality and the standard deviation of the quality of the items sold by each supplier. He only knows the price $price_i$. In addition, for each supplier i, the agent has a history of n_i length of previous interactions. The average quality of the goods that were bought by the agent in the previous n_i interactions from supplier i is $\overline{x}_i$ and the standard deviation is $\hat{s}_i$. The utility of the agent from buying an item of quality x_i and paying $price_i$ is $x_i - price_i$. Thus, the expected utility of an agent from buying from supplier i is $\overline{x}_i - price_i$.

The agent has a discount factor of $0 < \delta < 1$ for each time delay. This means that the supplier is indifferent about obtaining a reward of u at the next time period or δu at the current time. This use of discounting can be viewed in at least two different lights: as a reflection of an agent's internal uncertainty about its predictions, or as an external factor, much like an interest rate, that allows an agent to assess the present value of income that will be earned in the future. In Section 3.2.1 we show the intuition for the existence of a discount factor.

We begin by assuming constant and equal prices of the different suppliers, i.e., the difference between the suppliers is only in their quality. The question is how will the agent choose a supplier of the item. That is, the agent has to choose $i \in \{1, ..., N\}$ in order to maximize his benefits. In the following, we will map the simple variation of the multi-suppliers problem to the classical multi armed bandit problem, and show a particular method to solve the problem using Gittins indices.

3.1 Restating the Formalization of Gittins Indices

Our problem can be viewed as a variation of the "multi-armed bandit" problem. The original problem deals with a slot machine with arms. Each arm, when pulled, pays some amount of money. In each time period, only one arm can be pulled. The amount paid from each arm i has an unknown distribution θ_i. The problem is to develop a strategy that yields the maximum payoff over time by choosing which arm to pull based on previous experience and payoffs. In our domain, there are N suppliers. At each time period, the agent has to choose one of them. The utility he obtained from each supplier has a predefined unknown distribution, which is based on the quality of the supplier. The agent has to choose a supplier at each decision node in order to maximize his future outcomes.

In the following, we give a formal definition of the multi-armed bandit problem, based on Gittins' definition [5]. We restate the definitions to match the case of multiple suppliers, with stochastic qualities. Then, we show how to translate the suppliers problem to Gittins' formulation.

In our domain, the agent, when buying a product from a supplier, obtains this product with a quality which can be any positive or negative number. We assume that the value obtained from choosing an action is drawn from a distribution which belongs to a known family of distributions. In our interpretation, as an agent chooses to buy from a particular supplier, the quality of the item sold is drawn according to a particular distribution, with some unknown parameters. Suppose also that the quality of the item of a supplier i is drawn from the normal distribution, but the mean and variance of this distribution are unknown. That is, in our domain D is the set of all normal distributions, and θ_i includes the mean and variance associated with supplier i.

Gittins [5] has developed a tractable method for the off-line calculation of a table that can be used to determine the optimal action without computing all possible combinations of the actions. The Gittins method rests on the realization that rather than comparing all possible actions against each other, one can compare each potential action against a reference arm with a known and constant reward λ. Thus, index values indicating the attractiveness of each action as a function of past successes and failures can be tabulated in a single two dimensional table. Gittins proved that selecting the proposed action with the highest index value - the action with an expected reward equal to a reference action with the highest equivalent reward per pooling - is optimal. The result assumes an exponentially discounted reward, and a particular distribution of rewards for each arm.

For a given discount factor a, one can tabulate the attractiveness of each history of an arm relative to the reference arm. The table is computed by iteratively back-solving the following recurrence relation:

Definition 3.1 Recurrence relation between an arm and the reference process

$$
\begin{aligned}
&R(\lambda, \overline{x}, \hat{s}, n) = \\
&\qquad max[\tfrac{\lambda}{1-a}, \\
&\qquad \overline{x} + a \int (R(\lambda, \chi(\overline{x}, x, n+1), \sigma(\hat{s}, \overline{x}, x, n+1), n+1) f(x|\overline{x}, \hat{s}, n) dx]
\end{aligned}
$$

where χ denotes the adjusted average of past trials, including the $n+1$ trial, with value x,

$$\chi(\overline{x}, x, n+1) = \frac{n * \overline{x}_1 + x}{n+1}$$

and σ denotes the adjusted value of the standard deviation, including the value x of the $n+1$ trial.

$$\sigma(\hat{s}, \overline{x}, x, n+1) = \sqrt{\frac{(n-1)\hat{s}^2}{n} + \frac{(x-\overline{x})^2}{n+1}}$$

The formula expresses the value of choosing between an uncertain arm with an average of $\overline{x}$ and a standard deviation of $\hat{s}$, and between an alternative arm with a constant value of λ. In order to explain the intuition for this formula, we start with the right argument of the maximum function. If the uncertain arm is chosen, then the expected value generated by the arm is $\overline{x}$, according to the current information known about this arm. Then, a new choice should be made between the uncertain arm and the alternative arm λ. The expected value of the next choice between the arms is discounted by a. This choice will be made according to the current information in this step, using the updated average of the uncertain arm values, the updated standard deviation of this arm, and the new number of trials ($n+1$ instead of n). Thus, the arguments of the recursive call of R are χ instead of $\overline{x}$, and σ instead of $\hat{s}$, and $n+1$ instead of n. The values of χ and σ are determined according to the value x generated by the arm, and this value is distributed according to $f(x|\overline{x}, \hat{s}, n)$.

The left argument of the maximum function is the cumulative payoff for always choosing the reference action, which achieves a reward of λ indefinitely, with a discount of a. The present value of this indefinite payoff is $\frac{\lambda}{1-a}$. The *index* of a given arm is a value of λ that causes both arguments of the maximum function to be equal, i.e., the λ value is equal to the value of choosing the uncertain arm in the present.

Gittins denotes by $v(\overline{x}, \hat{s}, n)$ the index value for an arm with a history of n length, an average of $\overline{x}$ and a standard deviation of $\hat{s}$. He proved in [5]:theorem 6.16 that

$$v(\overline{x}, \hat{s}, n) = \overline{x} + sv(0, 1, n)$$

This shows that given an arm with an average of $\overline{x}$ and a standard deviation of $\hat{s}$ of its rewards, there is an easy transformation that can be performed in order to obtain the index of the above arm, using the index value of an arm which generates rewards drawn from the standard normal distribution, with a mean of 0 and a standard deviation of 1. The transformation to perform is: multiplying the index value by the standard deviation of the arm rewards, and adding the average reward of the arm. It is clear that as the average reward increases, the index value increases as well. Moreover, as the risk involved with the arm increases, the index value increases. This shows that more information about this arm is required in order to reduce the risk involved in choosing it.

Gittins calculated the values of v for the standard normal distribution ($\hat{s} = 1$ and $\overline{x} = 0$), given different combinations of a and n.

3.2 Solving the Multi-supplier Problem

The original multiple armed problem is simplified with respect to many aspects. The problem of choosing between suppliers is much more complex with regard to several issues. First, buying is not performed in predefined steps, rather purchases occur at different times. Second, different suppliers may ask for different prices, and the agent has to consider this difference when deciding from which supplier to buy. Third, the same agent may buy goods of different sizes. Even if he buys the same kind of goods from the same electronic supplier, the size, or weight, of the goods may be changed. For example, the buyer may buy a package of disks from an online computer shop, and at another time he may buy a new computer. In this type of case, the agent will be much more careful when transacting a large deal, and he will prefer to buy from the supplier which is known to him from previous encounters to be the best, and not try to learn about lesser known suppliers.

In this section, we will show how to handle these issues in the multi-supplier problem. In general, we demonstrate how to transform the multi-supplier problem, taking into account its complex details, into a simplified multiple-armed bandit problem. Furthermore we explain how to use this transformation in order to decide from which supplier to buy using the Gittins indices.

3.2.1 Considering Stochastic Buying

As we mentioned above, in most cases, an agent cannot actually predict his future purchases. However, he may be able to determine the probability of p that he will need the product at a particular time period (i.e., day). In particular, he can determine how often he purchases a certain type of goods. For example, he can determine that he buys food every two days, he buys one music CD per week, and he buys a new piece of furniture or a new electric appliance twice a year. Using this information, when making a purchase, the agent can consider the expected time until he will again need to buy a product of the same type (i.e., a product which is sold by the same set of suppliers). Alternatively, the agent can calculate the probability that a certain type of product will be required at a particular time period (day, hour, etc.) For example, if the agent buys one music CD a week, then the probability for him to look for a disk on a particular day is $p = 1/7 = 14.285\%$.

The agent has a higher valuation for present rewards than for future rewards. This is expressed in his utility function using the discount factor $0 < \delta < 1$. Obtaining a utility of u at the next time period is worse than obtaining a utility of u at the current time. We assume that for each time period t, the agent is indifferent about obtaining u at time $t+1$, or obtaining δu at time t. We can interpret δ as an interest rate. If the interest rate per day is r, then we can state that $\delta = \frac{1}{1+r}$, and obtaining item δu today is valued the same as obtaining $\delta * u * (1+r) = u$ tomorrow.

Another interpretation of δ is to consider the uncertainty about the future. This includes the probability of price changing (in which the profitability of learning about suppliers' qualities may decrease), or the probability of leaving the market in the future. Suppose that at each time period, the probability of staying in the market is p. Consider a risk neutral agent, and time t. Suppose that he knows that if he will stay in the market at time $t+1$, he will receive u. In this case, he will be indifferent about receiving u at time $t+1$ (with a probability of p of staying in the market), or receiving pu at time t (with a probability of 1). Thus, if such a risk is involved, then $\delta = p$. In order to evaluate δ for the case of price changing, the agent should have beliefs about the behavior of prices.

In order to compare one supplier to another, we need to determine the index for each seller i. In other words, which value of λ, if given as a reward at each time period, gives the same utility as choosing supplier i in the present period, and choosing i or the alternative supplier associated with λ in the future.

However, we would like to adapt λ to the situation of stochastic buying. First, we must find the expected value of choosing the reference reward λ forever, given p, the probability of the agent buying at each time period, and δ, the discount factor per period. We can find the expected discount of the alternative supplier. Since in each time period from the present and forever, there is a probability of p for a buying event to occur, and in this case, a reward of λ is obtained. However, the reward is discounted by a factor of δ^t, for buying events accomplished at time t. The following lemma provides the expected value of choosing λ at each buying event. All the proofs of the lemmas and theorems are presented in the appendix.

Lemma 3.1 *Given p, the probability of the agent to buy at each time period, and δ, the present value of choosing an alternative arm with a quality of λ forever is*

$$\frac{p\lambda}{1-\delta}$$

Next, we would like to evaluate the expected discount factor of choosing between supplier i and reward λ at the next decision point.

Lemma 3.2 *Given p, the probability of the agent to buy at each time period, and δ, the discount factor, the expected discount factor a until the nearest event when the agent will buy a product, is*

$$E(a) = p\delta + p(1-p)\delta^2 + ...p(1-p)^{i-1}\delta^t + ... = \frac{p\delta}{1-\delta+p\delta}.$$

Note that the transformation done to the reward from the λ is different from the transformation done to the reward of choosing the uncertain supplier i. This means that, as opposed to the original Gittins model, where δ was used as a discount factor on both sides of the equation, here the loss over time behaves differently on both sides. This will require an adjustment from the original model to the supplier's domain. The adjustment is performed in the following lemma.

Lemma 3.3 *Consider a reward function $R(\lambda, \overline{x}, \hat{s}, n)$ with probability of p for a buying event, and a discount factor of δ. Denote by w the Gittins index, calculated by using $a = \frac{p\delta}{1-\delta+p\delta}$. Then, the actual index value which is equal to the reward function is*

$$\lambda = w\frac{(1-\delta+p\delta)}{p}$$

The above lemma shows the relationship between the original Gittins index, and the actual index, given an agent that has a fixed probability of p to buy the product at any particular time period. The value of λ is found by making both sides of definition 3.1 equal. λ is the index value after adapting it to the stochastic model, while w is the index value before any adaptation. Using the above theorem, we can calculate the Gittins value, in order to compare different suppliers.

δ and p are independent of the supplier. Thus, in order to compare two suppliers, i and j, we can show that $w_i > w_j$ if $\lambda_i > \lambda_j$. This is proven in the following lemma.

Lemma 3.4 *Given sellers i and j, index values w_i and w_j, and actual index values λ_i and λ_j are adapted to the case of stochastic agent, $w_i > w_j$ if $\lambda_i > \lambda_j$.*

Considering the above lemma, one can argue that in order to compare two suppliers, it is sufficient to compare their original Gittins values, and choose the supplier with the higher original Gittins value. However, the importance of the actual Gittins indices is most noticeable when using suppliers with an average quality different from 0 and with a standard deviation different from 1. In these cases, comparing the indices using w instead of λ is wrong, since during the adaptation, the influence of $\overline{x}$ and $\hat{s}$ may change the relation between the sellers. Thus, in such a case, the updated index value of each seller should be calculated according to the steps in Section 3.2.4, and then the seller with the highest updated index is the best choice.

3.2.2 Considering Differences in Prices

The next attribute of our model is the differences in the prices of different suppliers. We would like to determine how the price set by a supplier influences his index, and through this, its possible superiority to the other suppliers.

We assume that the price of each supplier i, $price_i$, is fixed, and is known to the agent. Alternatively, we can also assume that the difference between prices of different suppliers remains constant. For each supplier i, the agent knows the following parameters: $\overline{x}_i$ is the average quality of the goods provided by supplier i, n_i is the number of deals transacted with this supplier, $\hat{s}_i$ is the variance of the product's quality and $price_i$. The agent also knows p and δ, as defined above. We denote by $R(\lambda, \overline{x}_i, \hat{s}_i, n_i, price_i)$ the reward received by making the best choice among supplier i and the alternative constant utility λ forever. We denote by $v(\overline{x}_i, \hat{s}_i, n_i, price_i)$ the index of supplier i, which is the value of λ that is equivalent to supplier i. As we are able to calculate v for each supplier i, we will choose the supplier with the highest v, in order to attain the highest equivalent constant reward.

The following theorem shows how we can calculate R, the present value of a seller, and v, the value of its index, given the sellers' prices.

Lemma 3.5 *If the prices of the suppliers are constant, then the following holds*

$$R(\lambda, \overline{x}_i, \hat{s}_i, n_i, price_i) = R(\lambda, \overline{x}_i - price_i, \hat{s}_i, n_i)$$

and

$$v(\overline{x}_i, \hat{s}_i, n_i, price_i) = \overline{x}_i - price_i + \hat{s}_i \cdot v(0, 1, n_i).$$

and the supplier with the highest v is the best one to choose.

This lemma enables us to consider differences in prices as well as differences in average quality. For example, if there are two suppliers with the same standard deviation and length of history, then an agent will give these suppliers the same index value if and only if the difference between the average qualities of these suppliers, as known by the agent, is equal to the difference between their prices. This means that the agent will agree to pay more to obtain a higher quality. This result is derived from Section 3, where we state that the utility from buying a product with quality x_i for price p_i, is $p_i - x_i$. If the relation among prices and quality is different, this will influence the adaptation required for the index values of suppliers with different prices.

3.2.3 Considering Different Sizes of Purchases

In real world situations, the size of a deal is different at each time period. The size is also different among different agents. We assume that the quality of a product is determined per deal, and is not dependent on the deal size. However, the agent transacting a large deal will be more careful and will choose suppliers with a higher mean quality and less risk. For example, the same agent may first look into buying a small chest online, and later he may want to buy a living room set from an online catalog. In the online music domain, the agent may initially buy a CD for ten dollars, but at other time periods, he may seek to buy a rare collection of music for hundreds of dollars. In the Internet domain, a person may connect to the Internet for one minute in order to download his email, and at other times, he may want to connect for hours in order to surf.

Different strategies can be used when past deals of different sizes are considered, when evaluating the average quality of a supplier and his standard deviation: (a) A deal can be considered independently of its size. This means that deals with different sizes will be considered in the same way. (b) A deal can be considered with respect to its size. As the size of the deal increases its effect on the historical average will be higher.

For example, if there is a deal of size $s1$ with a quality of $q1$ and a deal of size $s2$ with a quality of $q2$, then the average quality of this seller, following (b), will be $(q1s1 + q2s2)/(s1 + s2)$. Similarly, the standard deviation will be $(\sqrt{((q1 - average)^2 * s1 + (q2 - average)^2 * s2)/(s1 + s2)}$. In this manner, the larger the deal, the greater its contribution to the average quality and standard deviation.

Alternatively, if (a) is in use, then the average quality is $(q1 + q2)/2$, and the standard deviation is $(\sqrt{((q1 - average)^2 + (q2 - average)^2)/2}$. Strategy (b) will be used in situations where the large deal is composed of small deals, so it should have a higher weight when it is considered. Strategy (a) will be used if the large deal is composed of one item: for example, while a trip for hundreds of people is considered as a one item (strategy (a)), a living room is composed of several different items (strategy (b)).

Another difference is due to the reward from the current deal. If the deal is large, then the present is much more important than in smaller deals, since low quality goods will cause much more harm. Consider, for example, a person buying a chest from an online catalog. He will care less about the quality of the single item than he will about the quality of a larger catalog purchase. This means that an agent has to consider the size of the current deal when deciding from which supplier to buy.

Future deals are considered as deals of size 1. This means that the deal size is normalized, and the average deal is 1. Thus, the expected size of each future deal is 1. The following theorem specifies the updated index of a supplier, given its mean quality and the standard deviation of its quality, and given the size of the current deal.

Theorem 3.1 *Given the size of the current deal, size, R and v are calculated as follows.*

$$R(\lambda, \overline{x}_i, \hat{s}_i, n_i, price_i, size) = R(\lambda, \overline{x}_i - price_i, \hat{s}_i, n_i) + (size - 1)(\overline{x}_i - price_i)$$

and

$$v(\overline{x}_i, \hat{s}_i, n_i, price_i, size_i) = v(\overline{x}_i, \hat{s}_i, n_i, price_i) + (1 - \delta) * (size - 1) * (\overline{x}_i - price_i)$$

and the supplier with the highest v is the best one to choose.

The above theorem shows how the size of a deal affects the decision concerning the given deal. As the size increases, the present becomes more important, and the expected benefits from the deal, which is $(\overline{x}_i - price_i)$, becomes more important when determining the index value of a given supplier. Thus, we add the expected benefits from the deal, multiplied by the additional size of this deal (the difference from 1)to the reward R. Similarly, we add the same formula, multiplied by $1-\delta$ to the index value, because of the transformation from reward to index. We can see that as the size of the current deal increases, the current expected benefits, $(\overline{x}_i - price_i)$, have greater importance.

3.2.4 Steps for Calculating the Gittins Index for a Given Seller

This section summarizes sections 3.2.1-3.2.3, and gives a list of steps required in order to compute the index for a supplier, given its parameters. The input for each supplier i are as follows. $\overline{x}_i$ is the average quality of products provided by supplier i. $\hat{s}_i$ is the standard deviation of the quality of these products. n_i is the products number, and $price_i$ is the price suggested by supplier i. Other inputs relevant to all the suppliers are: $size$, the size of the current deal; p, the probability of buying a product at each time period; and δ, the discount ratio per time period.

Using these arguments, the agent will evaluate for each supplier i the index value,

$$v(\overline{x}_i, \hat{s}_i, n_i price_i, size_i).$$

This index is the value of λ that is equivalent to choosing supplier i at the present time. The present value of obtaining λ, whenever a purchase occurs, is equivalent to choosing supplier i in the present. Given the indices of all the suppliers, the agent will choose the supplier who maximizes v. In the following, we summarize the calculation steps.

Step 1: Obtaining a

Given p and δ, as described in Section 3.2.1, the value of a is determined as follows:

$$a = \frac{p\delta}{1-\delta+p\delta}$$

This adjustment will enable us to use a as an expected discount rate when calculating the supplier reward and the supplier index.

Step 2: Finding Gittins Index

Using a, the value $w = v(0, 1, n_i)$ will be calculated for supplier i, according to Table A.1 in Appendix A. This value considers only the size n_i for each supplier i. In Step 4, we will also consider the other parameters related to each seller, such as the price and the average quality.

Step 3: Calculate Actual Index Using p and δ

Given p and δ, the value of λ is calculated as follows:

$$\lambda = w\frac{1-\delta+p\delta}{p}$$

This adjustment is made in order to adjust the indices to the stochastic model.

Step 4: Adding $\hat{s}$, $\overline{x}$ and price

In order to consider the standard variation and the average quality of the supplier, as well as its price, the value of v is calculated.

$$v = \overline{x}_i - price_i + \hat{s}_i \cdot \lambda.$$

The standard deviation of the quality is multiplied by the index of step 3, and the average benefits of the current product (the difference between quality and price) is added to the index.

Step 5: Taking the *size* of the deal into consideration

Finally, the value of $size$ is also considered. As the size of the deal increases, the size of the current deal becomes more important than future benefits from future learning.

As the size increases the agent's expected reward from the current transaction, $\overline{x}_i - price_i$, becomes more important relative to future possible rewards. Thus, as the size increases the expected benefits from learning decrease.

$$v(\overline{x}_i, \hat{s}_i, n_i, price_i, size) = v + (1-\delta) * (size-1) * (\overline{x}_i - price_i)$$

3.2.5 An Example of Choosing Among Sellers

In the following, we show a particular example of an agent who has to choose from three suppliers. Consider a consumer of online music who would like to buy CDs. His value from each CD is \$10. Suppose he can choose to buy a particular CD from three possible suppliers, but he has a preference for obtaining the CD as soon as possible. Suppose that the value of the discount factor is $\delta = 0.909$. Thus, if he obtains a CD after t days, his quality is $10 * 0.909^t$.

Consider the past information of the agent to be as follows:

He bought from supplier1 twice, and obtained the first purchase in 10 days, and the second purchase in one week. The agent bought 4 music CDs from supplier2, with the following delays: 1 day, 1 day, 10 days and 2 days, respectively. The agent bought 5 CDs from supplier3, with the following delays: 2 days, 2 days, 1 day, 5 days and 14 days, respectively. Now the agent once again would like to buy a CD, and he needs to decide which supplier to buy from.

Table 1 summarizes the details of the past transactions in this example, in-

Table 1
Example: details of 3 suppliers

supplier1's delay	supplier1's quality	supplier2's delay	supplier2's quality	supplier3's delay	supplier3's quality
10	3.85	1	9.09	2	8.26
7	5.12	1	9.09	2	8.26
		10	3.85	3	7.51
		2	8.26	5	6.20
				14	2.62

cluding the delay and the quality of each of the past transactions for each of the suppliers.

The agent wants to buy one particular CD. Thus, the size of its deal is 1. The prices of this CD are specified in the following table:

supplier	price
supplier1	5
supplier2	7
supplier3	6

Suppose also that the probability that the agent will buy a CD on each day is $p = 0.1$. Consider an agent facing this decision. If he considers the short run only, he will choose the seller that gives him the highest expected utility. The following table will evaluate for each supplier the average quality when buying a CD from it, using the data from Table 1. For future usage, we also add the standard deviation of the utility when using each supplier.

supplier	price	average quality	standard deviation of quality
supplier1	5	4.489	0.902
supplier2	7	7.573	2.511
supplier3	6	6.574	2.359

If the agent considers only the short run utility, then for each of the goods that he buys, he expects to receive the goods with the average quality, and to pay the required price. Since his utility from the goods is measured as quality minus price, then when he buys from supplier i, he expects to derive the average quality of supplier i minus the price of supplier i as a utility. Thus, according to the above table, if the agent considers only the short run utility, he will choose supplier3 since the agent's current utility from choosing supplier3 is the highest: the average quality, 6.574, minus the price 6, 6.574-6=0.574. This is higher than the short run utility from buying from supplier2, 7.573-7=0.573, and from buying from supplier1 4.489-5=-0.51. However, in the following we will describe the steps toward finding the optimal choice, and we will see that if the long run is taken into consideration, then the optimal choice is supplier2.

Step 1: Obtaining a

We start by calculating the value of a, the expected discount ratio from one purchase. The value depends on δ and p, and it is the same for the different suppliers.

$$a = \frac{p\delta}{1-\delta+p\delta} = \frac{0.1 * 0.909}{1-0.909+0.1 * 0.909} = 0.5$$

Step 2: Finding Gittins Index

We calculate the values of w, according to Table A.1 in Appendix A, given $a = 0.5$, and given n_i as in the table below.

	n	w
supplier1	2	0.726587142
supplier2	4	0.094863017
supplier3	5	0.070580465

Intuitively, we see that if the different suppliers would have the same average, standard deviation and prices, then supplier1 would be chosen, since if we consider only the effect of learning in the long run, then supplier1 has the highest value. The intuition behind this is clear, as there are less examples of a supplier, obtaining more information about a supplier becomes more important.

Step 3: Calculating the Actual Index Using p and δ

In this step, we perform the adaptation discussed in Section 3.2.1, and calculate the value of λ so that it will also consider the discount ratio of δ and the probability of p to make a purchase at a given time.

$$\lambda = w\frac{(1-\delta+p\delta)}{p} = w\frac{(1-0.909+0.1 * 0.909)}{0.1} = w * 1.819$$

	w	λ
supplier1	0.726587142	1.321662011
supplier2	0.094863017	0.172555828
supplier3	0.070580465	0.128385865

We can see that the relation between w of the different suppliers remains the same as the relation between the λ of the different suppliers. This will always happen, since the transformation will just multiply w of the different suppliers with a constant. However, this change may influence the relation between the indices of the suppliers, when adding other characteristics of the suppliers, such as quality average, standard deviation and prices.

Step 4: Adding $\hat{s}$, $\overline{x}$ and price

In this step, we consider the parameters related to each supplier: the average and standard deviation of past transactions, and its price.

$$v(\overline{x}_i, \hat{s}_i, n_i, price_i) = \overline{x}_i - price_i + \hat{s}_i \cdot \lambda.$$

The values obtained from this transformation are as follows.

	$\overline{x}$	price	$\hat{s}$	λ	$v(\overline{x}_i, \hat{s}_i, n_i, price_i)$
supplier1	4.489	5	0.902	1.321	0.682
supplier2	7.573	7	2.511	0.172	1.007
supplier3	6.574	6	2.359	0.128	0.877

We can see that $\overline{x}_i$ and $\hat{s}_i$ affect the relations between the indices of the suppliers. In step 3, supplier1 still has the best value, but in this step, supplier2 becomes the best. Although the values of $\overline{x} - price$ and $\hat{s}$ are higher for supplier3, in the decision for the short run, supplier3 should be chosen. Combining the short run criteria (i.e., highest difference between quality and price) with the long run criteria (i.e., shorter n) makes supplier2 the best choice.

The above example demonstrates a possible situation in which the short run decision is different from the long run decision, and the index value of the suppliers reflects all the aspects of choosing the supplier, yielding an optimal decision.

3.3 Considering New Sellers

Gittins indices can be used for suppliers with histories of a length of 2 or more. The values in table A.1 start from $n = 2$. The question that arises is how to choose among suppliers when there are new suppliers, i.e., suppliers with no history or with a history of a length of one. We should find out how to treat the new suppliers, since we do not have enough data on them for the calculation of optimal Gittins indices. In particular, the buyer can use its prior beliefs about the quality distribution of the different suppliers, in order to consider the new ones. However, the prior beliefs may be very inaccurate, and thus an index that is calculated using these prior beliefs will also be inaccurate. Therefore, we propose decision procedures where prior beliefs are not necessary for the calculation stage, since a buyer in the real world has usually no accurate beliefs to base a decision on.

The dilemma that a buyer with no accurate prior beliefs on new suppliers faces is as follows. On the one hand, the buyer would like to learn about the new supplier, in order to determine what its quality is, since there may be situations where the new supplier has the highest quality. However, if new suppliers appear frequently, then a buyer that always tests the new suppliers first may find itself spending his time on the new suppliers, with unknown expected utilities, instead of choosing the best known supplier, and on average this will reduce his utility.

Consider a set $\{s_1..s_n\}$ of potential suppliers, where a subset $\{c_1..c_m\} \subseteq \{s_1..s_n\}$ contains only new suppliers, with a history of a length 0 or 1.

Then, the following strategies can be used:

(1) Giving society parameters (called "society"): For each new supplier, we can complete the index parameters using the society parameters as temporary parameters, then compute the Gittins indices as usual. Then we can choose the supplier from the overall set of suppliers. Completing the index parameter of each new supplier will be done as follows. The average of the quality of this supplier, if it was never chosen before, will be the average of the quality of all other suppliers. The standard deviation for this supplier will be the average standard deviation, of all the suppliers. The size of n_i for the new seller i will be determined, for the index calculation, to be 2, since this is the smallest n value for which an index value exists. These data are used only for the index calculation. Whenever new information about this seller is collected, they are substituted for the society parameters.

    This strategy will often result in the choice of new suppliers. A new seller will have a relatively unattractive average and a standard deviation of quality, since we give it the average value of the suppliers. However, since the new seller has $n = 2$, this make its index value higher than the index value of an average supplier with a longer known history (since as n decreases, the corresponding index value increases).

    However, if the new supplier sets a high price, or has a history with a length of 1 and a very low quality, then it will not be chosen very early, since these parameters are in use when calculating the index value of this supplier (only the missing parameters are completed). In fact, this supplier will be tested only after testing other new suppliers who will possess better parameters. Thus, this strategy gives opportunities to new suppliers with relatively promising parameters.

(2) Randomizing according to probability ρ: The agent will decide to choose a new supplier with a probability of ρ. If there is more than one new supplier, then the agent has to decide which new supplier to choose. We

suggest that the agent will first try to find a supplier with a history of 1, and only if this does not exist, then he will buy from a supplier with no history. If there is more than one new supplier with a history of 1, then the agent will choose from them randomly. Similarly, if there are no suppliers with a history of 1, but there is more than one new supplier with a history of 0, then the agent will choose from them randomly. With a probability of $1-\rho$, a known supplier is chosen. In this case, the old supplier with the highest Gittins index will be chosen.

Using a randomized strategy for choosing a new seller is attractive, since it gives the new suppliers a chance to prove themselves, while assuring that a predefined proportion of the deals will be completed with the best known supplier.

Randomization can also be done using classical reinforcement learning methods, such as the Boltzmann distribution [7]. An agent that uses the Boltzmann distribution, will choose a new supplier with a high probability when the agent is new in the market, and this probability will be reduced over time. The motivation for using this distribution is that as time increases, the agent becomes more experienced, with better knowledge about the known sellers, so choosing a new supplier becomes less attractive to him.

(3) Choosing the new supplier first: (called "choose-new") According to this strategy, whenever there is a new supplier, it will be chosen first. If there is more than one new supplier, then suppliers with a history of a length of 1 will be chosen first (if there is more than one, then the agent will choose randomly from them), and suppliers with no history will be chosen only whenever there is no supplier with a history of a length of 1 (and if there is more than one, the decision among them is, again, made randomly).

This strategy will give the highest chance for a new supplier to be tested. It is beneficial if new suppliers are rare, and there are not enough old suppliers, since a new supplier, after two steps, becomes known, and it may have a better quality than the old suppliers. However, if new suppliers appear periodically, then this method is not beneficial, since the new suppliers will be chosen in most of the cases, even though they may not be the optimal choice.

(4) Choosing an old supplier whenever there is one: (denoted by strategy "0") According to this strategy, whenever there is one or more suppliers with a history of a length of 2 or larger, then one of the old ones will be chosen, according to the Gittins indices. Given that such a supplier exists, the new suppliers will never be tested (it is a special case of the randomizing strategy, but the probability of choosing a new supplier is 0). If there are no suppliers with a history of 2 or more, then one of the new suppliers will be tested. This strategy will never test the new suppliers, given that there is at least one known supplier. Thus, it is beneficial only in situations where there are enough known suppliers, and learning becomes non-beneficial.

new	1 known	2 known	3 known	4 known	5 known
1	first-new	society	society	0	0
2	first-new	society	society	0	0
3	0.9	society	0	0	0
4	first-new	society	0	0	0
5	0.9	society	0	0	0

Table 2
The strategy that yields the highest average utility given old and new suppliers. *First-new* indicates that strategy 3 is the best to use. *Society* indicates that strategy 1 is the best, 0.9 indicates that the best strategy is strategy 2 with probability 0.9 of choosing a new seller, and *0* indicates that the best strategy is not to choose a new supplier at all.

In order to test the performance of the different strategies in different domains, we developed a simulation tool. In our simulation, we defined several agents. Each agent behaves according to each of the strategies defined above. For the randomized strategy, we defined several agents, where each agent has a different probability for choosing a new seller.

Each of the agents needs to buy the same product several times. The agent can choose from several suppliers, where some of the suppliers are new. The quality of the goods produced by each supplier is derived from a normal distribution, but the details of this distribution are unknown to the agent.

We first ran two cycles of initial buying, where each agent (representing each type of strategy) buys the goods from each of the known suppliers, observes the quality of the goods, and saves this in his database. Then, we add the new suppliers, in which no information about them is currently known. Next, the simulation is run for 100 cycles. In each cycle, each agent is required to buy the product, and he chooses the supplier according to his strategy. After the purchase is made, the quality of the product is observed (it is derived from the supplier's distribution), and the agent saves this quality in his database in order to compute the future Gittins indices, and also in order to compute the actual utility when following his strategy. After the simulations are completed, the utility actually obtained by each agent is computed, using the data saved in his database, about the qualities actually obtained during the several purchases.

We ran this process several times, and in each run we computed the utility of each of the agents, (i.e., each of the strategies). Then, we chose the strategy used by the agent with the highest utility. For simplicity, we assume that all the deals performed are of size 1, and all the suppliers suggest the same price.

In our simulations we varied the number of known suppliers between 1 and 5 and varied the number of new suppliers between 1 and 5. We tested the following strategies: 1, 2 (with a probability of 0.1,0.2,..,0.9 of choosing a new supplier), 3 (which is the strategy of choosing a new supplier with a probability of 1), and 4 (which is the strategy of choosing a new supplier with a probability of 0).

Table 3.3 presents the simulation results of 90,000 runs for each combination of the number of known and new suppliers. For each set of 90,000 runs for specific number of known and new suppliers we tested different levels of δ (discount factor), and different values of p (probability to buy in a given time). The mean quality μ_i of supplier i was drawn randomly from the interval $[5, 20]$ in all the runs, and also the standard deviation, σ_i, of the quality of supplier i was drawn randomly from this interval.

In most of the combinations of parameters that were considered in the 90,000 runs, when there are 4 or 5 known suppliers, the strategy of ignoring the new suppliers, i.e., strategy 4, yields the average highest utility. In cases where there are fewer known suppliers, then in most of the cases, trying the new supplier is beneficial. The type of the strategy which is the best depends on the exact number of known suppliers and new suppliers in the simulation. For example, in the case of one known supplier and one or two new suppliers, the most beneficial strategy was strategy 3. However, when there were two known suppliers, the most beneficial strategy was strategy 1, and in the case of three known suppliers, learning the new suppliers qualities is still beneficial only if there are only few (one or two) new suppliers.

For a given set of parameters of the simulation, specific results can be presented. Table 3 presents the results of 10,000 runs, where $\delta = 0.99$,$p = 0.9$,$\sigma_{min} = 2.5$, $\sigma_{max} = 10$, $\mu_{min} = 5$ and $\mu_{max} = 20$. Also in this table, when there are 5 known suppliers, learning about the quality of a new one is, on average, not beneficial. As the number of known suppliers drops, learning the quality of a new one becomes more beneficial.

Our conclusions from the above simulation are as follows: First, we found that when there are 5 known suppliers or more, learning about new suppliers is not beneficial for the parameters we checked. (For a different set of parameters, for example, a larger p or δ, the number of known suppliers may decrease or increase). The explanation for this behavior lies in the fact that the quality of the new suppliers is derived from the same distribution of the quality of the known ones. Thus, as there are enough known suppliers, the probability for a new supplier to be better than all the known suppliers is very low, and in such a case, learning about a new supplier becomes non beneficial.

This remains true if there is more than one new supplier, since learning about

new	1 known	2 known	3 known	4 known	5 known
1	first-new	first-new	0.6	society	0
2	society	first-new	0.9	0.7	0
3	first-new	first-new	0.8	0	0
4	society	first-new	0.9	0	0
5	society	society	0.8	0	0

Table 3
The strategy that yields the highest average utility, given old and new suppliers, when $\delta = 0.99, p = 0.9, std = 5$. *First-new* indicates that strategy 3 is the best to use. *Society* indicates that strategy 1 is the best, a number $\rho = 0.1, ..., 0.9$ indicates that the best strategy is to choose a new supplier using strategy 2 with probability ρ of choosing a new seller, and 0 indicates that the best strategy is not to choose a new supplier at all: strategy 4

each new supplier is costly, and the probability to benefit from learning about new suppliers is low. However, as the number of known suppliers decreases, the probability of having a new supplier with a higher quality than the best known supplier increases, so the best strategy is to consider the new suppliers. As the number of old suppliers declines, the best strategy gives a higher probability of choosing a new supplier, since there is a higher probability that the new supplier will provide better goods than the old ones.

Another conclusion from the above results is that each of the strategies we discuss, strategy 1 - 4, may be in use for a different set of parameters. Strategy 1 is beneficial in situations where choosing a new supplier is recommended, as shown in Tables 3.3 and 3, since this strategy gives the new suppliers the value $n_i = 2$, so they will have a high index value, and will be chosen relatively early. However, there may be situations where always choosing the new supplier is more beneficial. In other situations never choosing a new supplier is the best, and yet in other situations choosing a new supplier according to a given probability, is more beneficial. Thus, we cannot dismiss any of these strategies.

Next, we evaluate how changing the parameters of the environment influences our results. We varied the δ, the discount factor over time, and p, the probability for the agent to need to buy at a given time period.

Intuitively, as the expected discount ratio decreases, learning about new suppliers is more beneficial, and choosing the new suppliers becomes more beneficial. This is demonstrated in Table 4. This table is based on the same data as Table 3.3, but analyzed in a different way. Each cell in the table is based on 10,000 runs. It is shown that as p or δ increase, i.e., the expected discount ratio from one buying period to the next buying period decreases, and learning

δ	p=0.1	p=0.5	p=0.9
0.9	0	0	0
0.95	0	society	society
0.99	0	society	0.9

Table 4
The strategy that yields the highest average utility given δ and p, for all the combinations of 1..4 old suppliers, and 1..4 new suppliers.

about the new suppliers becomes, on average, more beneficial. The reason lies in the fact that if the agent chooses a new seller, he may lose with respect to his current purchase, while the possible benefits may be obtained only in the future, if the new supplier is found to be better than the other suppliers. Thus, as the future becomes more important, the importance of future benefits from the suppliers becomes larger than the present possible losses, so the agent will be more willing to test new suppliers.

Finally, it appears that as the number of new suppliers increases, learning about them becomes less beneficial. The intuition behind this is that if there are more new suppliers, learning about them becomes more expensive, since more time is required in order to learn about all of them. So in such a case, it is enough to examine new suppliers with a lower probability, or to ignore them all.

4 Conclusion

In this research, we discuss the issue of solving problems which involve a trade-off between exploration and exploitation. We suggest using deterministic methods for solving these problems, in cases where the problem can be described formally, and some required properties about the parameters exist. To illustrate this idea, we introduce the problem of an agent deciding from which supplier to buy certain goods in the context of electronic commerce, where each supplier provides goods with different expected qualities, and the expected quality is unknown to the agent. The agent only has information about the quality of goods produced by the suppliers in the past. The agent has to decide from which supplier to buy, and whether to buy from the best known supplier, or to buy from an unknown supplier in order to learn about it.

We show that this problem is a special case of the "multi-armed bandit" problem [5], and suggest using Gittins allocation indices [5] for solving it. The *index* of a given supplier will indicate its value, and the supplier with the highest index will be chosen.

We first consider variations where the price of each supplier is constant, and the size of the goods sold is constant for all goods sold, and we proceed with additional details which appear in a real world variation of the suppliers problem. First, we consider the fact that the agent is likely to buy the goods in time periods which are not known in advance, and we suggest how Gittins indices for this case can be calculated. Second, we consider the case of suppliers with different prices, and suggest how the prices will be included in the indices calculation. Then we consider situations where the deals differ in their size. As the size of a deal increases, the value of the present deal becomes more important compared to the value of future deals. We suggest how to take into account the size of the deal when calculating the Gittins indices.

We also consider the decision problem when there are new suppliers, with an unknown history, or with a history of a single event. We suggest techniques for enabling the choice of new suppliers. Finally, we consider a situation where the item's price of each supplier varies over time. In this case, the superiority of a supplier decreases over time, and is terminated after the final time. We suggest how to solve the suppliers problem when considering changing prices.

References

[1] amazon. Amazon. http://www.amazon.com, 2001.

[2] D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, UK, 1985.

[3] T. L. Dean and M. P. Wellman. *Planning and Control*. Morgan Kaufmann Publishers Inc., 1991.

[4] C. Eric, K. Hann, and I. Hitt. The nature of competition in electronic markets: An empirical investigation of online travel agent offerings. WP, The Wharton School of the Univ. of Pennsylvania, 1998.

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

[6] E. R. Hilgard and G. H. Bower. *Theories of Learning*. Prentice-Hall, Englewood Cliffs, NJ, 1975.

[7] L. P. Kaelbling and A. W. Moore. Reinforcement learning: A survey. *JAIR*, 4:237–285, 1996.

[8] M. Salganicoff and L. H. Ungar. Active exploration and learning in real-valued spaces using multi-armed bandit allocation indices. In *Proceedings of the 12th International Conference on Machine Learning*, pages 480–487, San Francisco, CA, 1995.

[9] A. Schaerf, Y. Shoham, and M. Tennenholtz. Adaptive load balancing: A study in multi-agent learning. *JAIR*, 2:475–500, 1995.

[10] T. J. Strader and M. J. Shaw. Characteristics of electronic markets. *Decision Support Systems*, (21)3:185–198, 1997.

[11] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[12] J. Tirole. *The Theory of Industrial Organization.* MIT press, 1988.

[13] J. C. Westland. Transaction risk in electronic commerce. *Decision Support Systems*, (33)1:87–103, 2002.

[14] P. Whittle. Arm-acquiring bandits. *The Annals of Probability*, 9 (2):284–292, 1981.

A Table of Gittins Indices

Table A.1 demonstrates Gittins values, given n (the number of experiences for a supplier) and a (the expected discount factor from one time period to the next). This table is used in order to evaluate the index value in different real cases. The table was generated as follows: First, we used Table 1 from Gittins [5]: values of $n(1-a)^{1/2}v(0,n,1,a)$ for a normal reward process with a known variance. We eliminated $v(0,n,1,a)$ from this table. Then, we used Table 3, the ratio of indices for cases of unknown variance and known variance. The first 4 lines in that table were given as $\zeta\xi^{-1}-1$,
i.e., $index\_with\_unknown\_variance/index\_with\_known\_variance - 1$. We combined table 1 and 3 and eliminated $index\_with\_unknown\_variance$. The next lines of the table were $100(\zeta\xi^{-1}-1)$. Again, we eliminated the index of unknown variance cases. The index values are given in Table A.1

n	0.5	0.6	0.7	0.8	0.9	0.95	0.99
2	0.726587142	1.047405456	1.555451776	2.816295061	5.169212118	10.14091728	39.3343305
3	0.156203933	0.21476361	0.298041434	0.434252831	0.735711782	1.165610598	3.102001067
4	0.094863017	0.130008192	0.179135365	0.25663625	0.416059164	0.619336482	1.342794975
5	0.070580465	0.096733278	0.133229525	0.190465068	0.306080486	0.447756529	0.905235725
6	0.056787249	0.07791264	0.107422658	0.153685894	0.246655465	0.358998922	0.705420293
7	0.04778777	0.065636827	0.090613406	0.129833471	0.208662362	0.303516425	0.590103647
8	0.041349198	0.056851058	0.078577624	0.112777285	0.181653586	0.264506832	0.512330226
9	0.036489763	0.050213549	0.069484333	0.099880466	0.161278848	0.235252269	0.455568067
10	0.032676914	0.045000262	0.062338065	0.089742431	0.145267833	0.212338611	0.411874748

Table A.1
Gittins Indices.

B Proofs

B.1 *Proof of lemma 3.1*

In each time period from the present and forever, there is a Probability of p for a buying event to occur, and in this case, a reward of λ is obtained. The reward is discounted by a factor of δ^i, for each buying event completed at time i. Thus, the expected value of choosing λ forever is,

$$p\lambda + p\delta\lambda + p\delta^2\lambda + ... + p\delta^i\lambda + ...$$

This is an infinite geometric series, with a sum of

$$\frac{p\lambda}{1-\delta}$$

$\square$

B.2 *Proof of lemma 3.2*

The expected discount factor $E(a)$ can be calculated as follows: With a probability of p the agent will need to buy at time 1, in which case there is a discount factor of δ. With a probability of $p(1-p)$ the nearest buy will be at time 2, with a discount factor of δ^2. The progression becomes an infinite geometric series, with the sum $\frac{p\delta}{1-\delta+p\delta}$. $\square$

B.3 *Proof of lemma 3.3*

Using definition 3.1, and denoting the index value as w, we obtain,

$$\begin{aligned} &R(\lambda, \overline{x}, \hat{s}, n) = \\ &\qquad max[\tfrac{w}{1-a}, \\ &\qquad \overline{x} + a \int (R(w, \chi(\overline{x}, x, n+1), \sigma(\hat{s}, \overline{x}, x, n+1), n+1)dx] \end{aligned}$$

where χ and σ are independent of a. Consider now that instead of using a we use $\frac{p\delta}{1-\delta+p\delta}$, which is the expected discount until the next buying event, as proved in lemma 3.2. In fact, we will obtain a value w as an index for which

$$\frac{w}{1-a} = \overline{x} + a\int(R(\lambda, \chi(\overline{x}, x, n+1), \sigma(\hat{s}, \overline{x}, x, n+1), n+1)dx$$

However, we would like to find the equivalent λ such that obtaining λ at each time of a buying event will be equivalent to the right hand of the function. According to lemma 3.1, the expected value of choosing λ forever is

$$\frac{p\lambda}{1-\delta}$$

So we would like to find λ, such that

$$\frac{p\lambda}{1-\delta} = \frac{w}{1-a} = \frac{w}{1-\frac{p\delta}{1-\delta+p\delta}}$$

Solving the above equation, we obtain,

$$\lambda = w\frac{(1-\delta)(1-\delta+p\delta)}{p-p\delta}$$

□

B.4 Proof of lemma 3.4

Using lemma 3.3,

$$\lambda_i = w_i\frac{(1-\delta)(1-\delta+p\delta)}{p-p\delta},$$

and

$$\lambda_j = w_j\frac{(1-\delta)(1-\delta+p\delta)}{p-p\delta}.$$

Thus,

$$\lambda_i - \lambda_j = (w_i - w_j)\frac{(1-\delta)(1-\delta+p\delta)}{p-p\delta},$$

However, since $0 < p < 1$ and $0 < \delta < 1$, then $(1 - \delta) > 0$, and $\delta > p\delta$), and $(1 - \delta + p\delta) > 0$ too. Finally, $p > p\delta$, so $p - p\delta > 0$. Thus,

$$\frac{(1-\delta)(1-\delta+p\delta)}{p-p\delta} > 0$$

since both fraction sides are positive. Thus, $\lambda_i - \lambda_j > 0$ if $w_i - w_j > 0$. In other words, $w_i > w_j$ if $\lambda_i > \lambda_j$ $\square$

B.5 Proof of lemma 3.5

Since $price_i$ is constant, it is decreased from the utility $x_i(t)$ which the agent obtains from choosing seller i. This decrease is constant, thus, the expected value of the rewards from choosing seller i decreases with this constant. However, for the same reason, $price_i$ will not influence $\hat{s}_i$ on the whole, since standard deviation is not influenced by constant changes.

Gittins proved that

$$v(\overline{x}, \hat{s}, n) = \overline{x} + sv(0, 1, n)$$

In fact, this holds if the expected reward is $\overline{x}_i - price_i$. $\square$

B.6 Proof of theorem 3.1

Consider $\lambda = v(\overline{x}_i, \hat{s}_i, n_i, price_i, 1)$. Then, the buyer is indifferent to choosing seller i today, or the alternative process of λ forever. The utility of choosing λ forever, is $\frac{\lambda}{1-\delta}$. To this utility, we add the rewards due for the current size of deal, which is $(size - 1) * (\overline{x}_i - price_i)$.

So the total utility from the deal is $\frac{\lambda}{1-\delta} + (size - 1) * (\overline{x}_i - price_i)$.

Returning to the index form, we multiply the above term by $(1 - \delta)$, and obtain

$\lambda + (1 - \delta)(size - 1) * (\overline{x}_i - price_i)$. $\square$