A Bayesian Formulation of Search, Control and the Exploration/Exploitation Trade-off

Richard Rohwer and Huaiyu Zhu
Dept. of Computer Science and Applied Mathematics
Aston University, Birmingham, UK B4 7ET

August 15, 1995

Abstract

A new approach to optimisation is introduced based on a precise probabilistic statement of what is ideally required of an optimisation method. It is convenient to express the formalism in terms of the control of a stationary environment. This leads to an objective function for the controller which unifies the objectives of exploration and exploitation, thereby providing a quantitative principle for managing this trade-off. This is demonstrated using a variant of the multi-armed bandit problem. This approach opens new possibilities for optimisation algorithms, particularly by using neural network or other adaptive methods for the adaptive controller. It also opens possibilities for deepening understanding of existing methods. The realisation of these possibilities requires research into practical approximations of the exact formalism.

1 Introduction

Optimisation methods can be compared according to various criteria, such as the computation time they require, the accuracy of the solutions they produce (as a function of computation time), and the classes of functions on which they are effective. It is normal practice to invent a method and test it against these criteria using numerical simulations and/or theoretical results such as convergence proofs. This is useful, but it would be better still to be able to *derive* an optimisation method by first stating the criteria and then finding the method which best satisfies them. Even if it were not practical to search for the best method, which itself would entail solving an optimisation problem, a precise formula for evaluating optimisation methods provides the best starting point for approximations.

It is possible to formalise these ideas by regarding optimisation as the control of a stationary environment. The optimisation method is identified

with a "controller" π which selects arguments x to a function g. [1] The environment is fully specified by the function g, and simply responds to "control action" x with "response" $g(x)$ (and perhaps further information $h(x)$, such as some derivatives of g at x). The response of a non-stationary environment would depend on an internal state as well as the control action, which will not be considered here. A controller seeks to maximise some "reward" R, which is a function of the time-sequence of environmental responses. For an optimisation problem, (for definiteness, a maximisation problem) one sensible definition of the reward is the largest function value observed during the number of function evaluations T that one is willing to carry out.

Although they are not ordinarily defined this way, typically optimisation algorithms have an implicit controller, fixed once and for all, which prescribes how new function arguments are to be selected based on previous function evaluations. For example, in first- and second-order gradient descent methods [6, 17] the next argument is chosen as a function of the previous one or two function values and gradients, and perhaps other data structures which are incrementally updated, such as an inverse Hessian approximation. Simulated annealing [13] selects arguments from a prescribed Boltzmann distribution. Genetic algorithms use [9] an *ad hoc* formula to update a set of function evaluations. Regarding optimisation as a control problem leads naturally to the idea of upgrading these implicit controllers to *adaptive* controllers, possibly implemented by neural networks, in order to obtain an optimisation method best suited to a problem or class of problems.

An optimising controller itself requires an objective to optimise. Here a formal expression for such an objective is derived from first principles. This embodies a quantitative theory of the exploration/exploitation tradeoff, introducing an entirely new approach to this issue, as far as we are aware. This is illustrated with a simple example. We are using this as a starting point for workable approximations to obtain better optimisation methods, and to better understand existing ones.

An adaptive controller has two sub-tasks, system identification and, based on this, reward optimisation. In function maximisation, the identification step amounts to the creation of a (possibly quite crude) model of the function g, based on some of the "samples" $(x, g(x), h(x))$, normally the more recent ones. Optimisation is done on the basis of this model. For example, at time step t in gradient descent, g is modelled by its tangent plane in a neighbourhood of the latest argument sample x_t, and an optimal choice for x_{t+1} is made based on this model. These two tasks are partially conflicting. Whereas identification requires *exploration* to gather information about the environment, optimisation is best served by *exploitation* of

[1] We shall use the term "function" to mean either a deterministic function, ie. a function in the ordinary sense, or a stochastic function, ie. a random field. This more relaxed interpretation is advantageous not only because it has wider scope of applications, but also because some fundamental issues of optimisation are identical for both cases.

existing knowledge with the sole objective of increasing the reward. This exploration/exploitation tradeoff is a fundamental dilemma to which the control-based approach presented here supplies a quantitative solution, at least in principle.

Section 2 explains the main idea in detail while developing notation. Then a formula for "optimal optimisation" is developed in section 3 for a somewhat restricted case. The N-armed bandit problem is used in section 4 to show that this formula expresses the exploration/exploitation tradeoff. Discussion and conclusions follow in section 5.

2 Optimisation as a control problem

In a maximisation problem, the maximum of a fixed function g is sought. The word "fixed" is used cautiously because g is not known in a sense which makes the implied 'knowledge' of the solution $x^* = \arg\max_x g(x)$ particularly helpful. Effectively, g is only partially known because there is insufficient time to exhaustively evaluate it, even though the knowledge of how to do so is readily available. Only the partial knowledge is made available to the system identification task. Let this knowledge be called k_t at time t. In general, k_t will be a set of quantities computable from the past data $D_t \stackrel{\text{def}}{=} [x_1, y_1, ..., x_t, y_t]$, where $y_t = (r_t, h(x_t))$ with $r_t = g(x_t)$. For example, if past data is never discarded, as in Tabu search [7, 8], then k_t can be the data itself. The dimension of k_t would increase with t in this case. Another possibility is to choose a fixed-dimension form for k_t which can be updated using a function K of the form

$$k_{t+1} = K(x_{t+1}, y_{t+1}, k_t) \tag{1}$$

The search direction in the conjugate gradient algorithm constitutes knowledge of this type. Another example is a sample mean or sample variance of the data, together with the time t itself. Whether the relationship is of the form (1) or otherwise, the function relating k_t to D_t will be called $k_t = K(D_t)$, with the specific meaning of K being clear from the context.

In general, a controller π provides a distribution $P(x_{t+1}|k_t, \pi)$ from which the next argument is selected. This paper will focus on controllers which utilise all past data (or a sufficient statistic [5] of this data) to arrive at an optimal policy which is therefore deterministic [2, 10]. In this case $P(x_{t+1}|k_t, \pi)$ is a singular distribution. The more general case will be discussed briefly in section 5.

Bayesian probability theory provides the essentially unique logically consistent way to quantify uncertainty [4, 11] and and "reasonably good decision rules" are Bayesian decision rules [5, 3]. Therefore it is best to describe the partial knowledge of the environment g with a probability distribution $P(g|k_t, \pi)$, the probability that the function is g given that knowledge k_t

was acquired using controller π. Of course, this is not meant to suggest that g is produced by a random process, even though that is one way to interpret a probability distribution. Here $P(g|k_t, \pi)$ expresses only our ignorance of a definite deterministic or stochasitc[2] function g. (Whether g is known; ie., $P(g|\cdots)$ is singular, is independent of whether g is determinisitc; ie., $P(r_t|g, \cdots)$ is singular.) The distribution conditioned on no knowledge, $P(g)$, describes the class of functions to which the method is to be applied. Then with $P(g|\pi) = P(g)$ and $P(k_t|\pi) = \int_g P(k_t|g, \pi)P(g|\pi)$, Bayes' rule specifies $P(g|k_t, \pi)$ as

$$P(g|k_t, \pi) = P(k_t|g, \pi)P(g)/P(k_t|\pi). \tag{2}$$

If the controller π and the function g are both deterministic, then they completely determine the data D_t which in turn determines the knowledge k_t. Then formally, at least, there is a function $D_t(g, \pi)$, in terms of which $P(k_t|g, \pi) = \delta(k_t - K(D_t(g, \pi)))$ in terms of the Dirac delta distribution.

Other distributions conditioned on the knowledge can be defined in terms of $P(g|k_t, \pi)$, such as the probability that the next function evaluation will be r_{t+1} if the next argument supplied is x_{t+1}:

$$P(r_{t+1}|x_{t+1}, k_t, \pi) = \int_g P(r_{t+1}|x_{t+1}, g)P(g|k_t, \pi). \tag{3}$$

Here the fact that k_t and π contribute no more knowledge than g justifies using $P(r_{t+1}|x_{t+1}, g, k_t.\pi) = P(r_{t+1}|x_{t+1}, g)$, and the irrelevance of x_{t+1} to knowing g justifies $P(g|x_{t+1}, k_t, \pi) = P(g|k_t, \pi)$. The distribution $P(r_{t+1}|x_{t+1}, g)$ can have any form if g is stochastic. It is a singular distribution $P(r_{t+1}|x_{t+1}, g) = \delta(r_{t+1} - g(x_{t+1}))$ if g is deterministic.

A sensible reward in a maximisation problem is $R = \max_t r_t$ where $r_t = g(x_t)$. If knowledge of the best sample seen so far is retained, then a very similar reward is $R = r_T$, where T is the maximum number of time steps allowed, because the controller can simply re-select this remembered point at the final time step.

If the reward were known as a function $R(x_{t+1})$ of the argument to be selected at time $t + 1$, then the optimal policy would be simply to choose x_{t+1} to maximise R. The knowledge k_t suffices only to specify a distribution over rewards $P(R|x_{t+1}, k_t, \pi)$, but this can be used to define the expectation value $\langle R|x_{t+1}, k_t, \pi\rangle = \int_R RP(R|x_{t+1}, k_t, \pi)$. The function g can always be transformed to a utility function for which the expectation value expresses essentially arbitrary preferences about the distribution [19]. Let us restrict attention to the controller which always chooses the best sample x_{t+1} according to this expectation value, in which case the dependence on the controller

[2]Here we consider only stochastic functions which can be decomposed as a deterministic function added to a stationary independent random process. Technically, a subscript t should be appended to g to represent that the random process produces a different output at each time step, but this formality will be ignored.

π in every expression does not need to be explicitly noted. Let us further restrict attention to the reward $R = r_T$, so the control policy is to choose x_{t+1} to maximize $\langle r_T | x_{t+1}, k_t \rangle$. A formula for this expectation value is derived in the following section for the case of retaining all past data, $k_t = D_t$. A slightly more complicated formula can be obtained for the general case.

3 Expected reward given all past data

If the first t function evaluations D_t are known, an expression is required for $\langle r_T | x_{t+1}, D_t \rangle$ in terms of information available at time t, in order to make an optimal choice for the next argument to select, $x_{t+1} = \arg\max_{x_t} \langle r_T | x_{t+1}, D_t \rangle$. Such an expression can be obtained by working backwards from time $T-1$ to time t.

At time T, given knowledge of D_{T-1}, one would choose x_T to maximize

$$\langle r_T | x_T, D_{T-1} \rangle = \int_{r_T} r_T P(r_T | x_T, D_{T-1}). \tag{4}$$

This choice defines a function $\xi_T(D_{T-1})$.

Not all of the data D_{T-1} is known at time $t < T$. At time $T-1$, the best one can do is to choose x_{T-1} to maximize $\langle r_T | x_{T-1}, D_{T-2} \rangle$, which can be written

$$\langle r_T | x_{T-1}, D_{T-2} \rangle = \int_{r_{T-1}} \langle r_T | x_{T-1}, r_{T-1}, D_{T-2} \rangle \, P(r_{T-1} | x_{T-1}, D_{T-2}). \tag{5}$$

The expectation value in the integrand can be written as $\langle r_T | D_{T-1} \rangle$, or

$$\langle r_T | D_{T-1} \rangle = \int_{x_T} \langle r_T | x_T, D_{T-1} \rangle \, P(x_T | D_{T-1}). \tag{6}$$

Having established that the controller will select $x_T = \xi_T(D_{T-1})$, the distribution $P(x_T | D_{T-1})$ is seen to be a Dirac delta distribution $P(x_T | D_{T-1}) = \delta(x_T - \xi_T(D_{T-1}))$, so

$$\begin{aligned} &\langle r_T | x_{T-1}, D_{T-2} \rangle \\ &= \int_{r_{T-1}} \int_{x_T} \langle r_T | x_T, D_{T-1} \rangle \, \delta(x_T - \xi_T(D_{T-1})) P(r_{T-1} | x_{T-1}, D_{T-2}) \\ &= \int_{r_{T-1}} \langle r_T | x_T = \xi_T(D_{T-1}), D_{T-1} \rangle \, P(r_{T-1} | x_{T-1}, D_{T-2}). \end{aligned} \tag{7}$$

Continuing in this manner, the distributions over the arguments and rewards combine in a Markovian fashon to give

$$\begin{aligned} \langle r_T | x_t, D_{t-1} \rangle = \int_{r_t} \cdots \int_{r_{T-1}} &\langle r_T | x_T = \xi_T(D_{T-1}), D_{T-1} \rangle \\ &\prod_{\tau=t+1}^{T-1} P(r_\tau | x_\tau = \xi_\tau(D_{\tau-1}), D_{\tau-1}) \, P(r_t | x_t, D_{t-1}) \end{aligned} \tag{8}$$

for any $t < T$. Maximizing this expectation value with respect to x_t defines $\xi_t(D_{t-1})$, given that ξ_τ is already defined for $\tau > t$.

This expression gives the optimal sampling strategy for maximising the function. It involves an expected final reward $\langle r_T | x_T, D_{T-1} \rangle$ conditioned on data D_{T-1}, not all of which is available at time t. Different values for the unavailable data would have different implications for the expected reward, so an average is taken, weighted by the probabilities as known at time t. This average will be driven up if data is found which has higher r_t values than the current $\langle r_T \rangle$, but the probability of such data turning up may be low. This is the exploration/exploitation trade-off, and expression (8) gives it a quantatative form with each value of the integrand representing a different future scenario.

4 Illustration: An N-armed bandit

An "N-armed bandit" gives a simple illustration of the exploration/exploitation trade-off problem. The function g is a stochastic function of 1 N-valued variable x. The value $g(x)$ is given by a Gaussian distribution with mean μ_x and unit variance. This is a very simple example of a function which cannot be entirely determined by a finite amount of data. In this case this is because the function is stochastic, but similar conclusions can be expected if the source of the uncertainty is incomplete knowledge of a deterministic function.

There has been a large body of work on the N-armed bandit [21, 14, 12, 18, 1] with the objective of maximizing a possibly discounted sum of function values $\sum_t \gamma^t r_t$, with $0 < \gamma \leq 1$. However, the objective of interest here is quite different, to maximise r_T for some given final time T. Putting all the weight on the last time step can be accomplished by taking $\gamma \to \infty$, so it might be interesting to attempt to examine this case by conventional methods in order to make contact with the results below. We shall leave this aside, because the main point of the exercise is not to improve on bandit methodology but to illustrate that equation (8) does indeed quantify the exploration/exploitation tradeoff.

Let the prior distribution of μ_i be a Gaussian $N(a_{i0}, 1/n_{i0})$, where $N(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Let $a_{it} = \langle \mu_i | D_t \rangle$ and $n_{it} = n_{i0} + \#\{\tau \leq t : x_\tau = i\}$, with the notation $\#A$ meaning the number of elements in set A. These are sufficient statistics for μ_i. Then the posterior $P(\mu_i | D_t)$ is $N(a_{it}, 1/n_{it})$.

For the final step T,

$$\langle r_T | x_T = i, D_{T-1} \rangle = a_{i,T-1}, \tag{9}$$

so the optimal policy is

$$\xi_T(D_{T-1}) = \arg\max_i \{a_{i,T-1}\}, \tag{10}$$

with

$$\langle r_T | \xi_T(D_{T-1}), D_{T-1} \rangle = \max_i \{a_{i,T-1}\} . \tag{11}$$

Now consider step $T-1$. It holds that

$$\begin{aligned} &\langle r_T | x_{T-1}, D_{T-2} \rangle \\ &= \int_{r_{T-1}} \langle r_T | x_T = \xi_T(D_{T-1}), D_{T-1} \rangle \, P(r_{T-1} | x_{T-1}, D_{T-2}) \\ &= \int_{r_{T-1}} \max_i \{a_{i,T-1}\} \, P(r_{T-1} | x_{T-1}, D_{T-2}). \end{aligned} \tag{12}$$

In the rest of this section, except where explicitly noted, we shall only consider distributions conditional on D_{T-2} and $x_{T-1} = k$, so to simplify the notation, we shall make these two conditions implicit. With these conditions in mind, it can be shown from the updating rule for $a_{k,t}$ that

$$a_{j,T-1} \sim \begin{cases} N\left(a_{k,T-2}, \dfrac{1}{n_{k,T-2}(n_{k,T-2}+1)}\right), & j = k, \\ N\left(a_{k,T-2}, 0\right), & j \neq k. \end{cases} \tag{13}$$

Let $a^{(i)}_{T-2}$ denote the quantities $\{a_{i,T-2} : \forall i\}$, sorted in decreasing order, with $k^{(i)}_{T-2}$ denoting the original index of the ith sorted quantity. That is, $a^{(1)}_{T-2}$ is the maximum, which is identical with $a_{k^{(1)}_{T-2},T-2}$. [3] It then follows easily that

$$\max_i \{a_{i,T-1}\} = \begin{cases} \max\left\{a_{k,T-1}, a^{(1)}_{T-2}\right\}, & k \neq k^{(1)}_{T-2}, \\ \max\left\{a_{k,T-1}, a^{(2)}_{T-2}\right\}, & k = k^{(1)}_{T-2}. \end{cases} \tag{14}$$

It is then straightforward to derive from that (13) and (14) that

$$\langle r_T | x_{T-1} = k, D_{T-2} \rangle = \begin{cases} f\left(a^{(1)}_{T-2}, a_{k,T-2}, \dfrac{1}{n_{k,T-2}(n_{k,T-2}+1)}\right), & k \neq k^{(1)}_{T-2}, \\ f\left(a^{(2)}_{T-2}, a_{k,T-2}, \dfrac{1}{n_{k,T-2}(n_{k,T-2}+1)}\right), & k = k^{(1)}_{T-2}, \end{cases} \tag{15}$$

where $f(a, b, \sigma^2)$ is defined as $\langle \max\{x, b\} \rangle$ with $x \sim N(a, \sigma^2)$, and is given explicitly as

$$f(a, b, \sigma^2) = \frac{a+b}{2} + \frac{b-a}{2} \operatorname{erf}\left(\frac{b-a}{\sqrt{2}\sigma}\right) + \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(b-a)^2}{2\sigma^2}\right). \tag{16}$$

[3] We ignore the zero probability cases where $a^{(i)}_{T-2} = a^{(i+1)}_{T-2}$, which can be made to disappear by an infinitely small perturbation.

The optimal policy at step $T-1$ is

$$x_{T-1} = \xi_{T-1}(D_{T-2}) = \arg\max_k \langle r_T | x_{T-1} = k, D_{T-2} \rangle \,. \tag{17}$$

This favours those k such that $a_{k,T-2}$ is large or $n_{k,T-2}$ is small. Intuitively, this means that the optimal strategy is to choose the state which is most under-tested for its worth.

Note that by $f(a, b, \sigma^2) = f(b, a, \sigma^2)$ it can be easily verified that

$$\left\langle r_T | x_{T-1} = k^{(1)}_{T-2}, D_{T-2} \right\rangle = \left\langle r_T | x_{T-1} = k^{(2)}_{T-2}, D_{T-2} \right\rangle \tag{18}$$

whenever $n^{(1)}_{T-2} = n^{(2)}_{T-2}$. Therefore for the two-armed bandit problem, for $N = 2$, the only factor affecting $\xi_{T-1}(D_{T-2})$ is $n^{(1)}_{T-2} - n^{(2)}_{T-2}$; the optimal policy is simply to choose the less tested state with out any regard to the expected rewards of both states.

Figure 1 shows a contour plot of $\langle r_T | x_{T-1} = k, D_{T-2} \rangle$ as a function of $a_{k,T-2}$ and $n_{k,T-2}$ for particular values of $a^{(1)}_{T-2}$ and $a^{(2)}_{T-2}$. Other values give qualitatively similar plots. Given D_{T-2}, the N possible choices of $x_{T-1} = k$ will produce N points on this plot, and the one on the highest contour (toward the lower right) should be selected for the next evaluation. Points lying on the same contour are equally good choices. Therefore the contours show precisely how the need for exploration (low n) is balanced with the need for exploitation (high a).

5 Conclusions

An approach to optimisation has been introduced based on a precise formulation of what is required, ideally, of an optimisation method. It is convenient to express this in terms of the control of a stationary environment. This leads to an objective function for the controller which unifies the objectives of exploration and exploitation into a single objective, thereby providing a quantitative principle for managing this trade-off. We are not aware of any previous attempts of this nature, although there has been extensive research into objective functions for exploration [15], and many techniques for managing the exploration/exploitation tradeoff have been invented [20, 16]. One of our current research directions is to place some of these methods into the general context.

Here we have set out only the first steps of this approach, and demonstrated that it does indeed yield a quantitative expression of the exploration/exploitation trade-off in a simple case. It is also clear that in general it will not be practical to use the exact formalism; instead it must serve as a basis for approximations. Even though severe approximations may be necessary to make the problems of computing and optimising the controller's objective function less difficult than the original problem, we feel that this

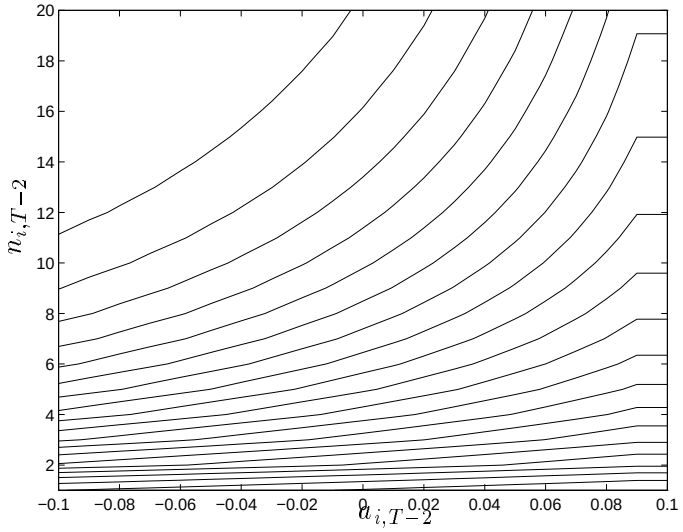

Figure 1: Contour plot of $\langle r_T | x_{T-1} = k, D_{T-2} \rangle$ as a function of $a_{k,T-2}$ and $n_{k,T-2}$ for $a^{(1)}_{T-2} = .1$ and $a^{(2)}_{T-2} = .09$.

basis for approximation is in a deep sense correct, and therefore a better starting point than any *ad hoc* method.

Only the case in which the controller has access to all past data, or a sufficient statistic of this data, has been investigated in any detail here. We are currently developing the more general case, which is of greater practical interest. In particular, it would be interesting to use a neural network for the controller, or an adaptive generalisation of the conjugate gradient method, or an adaptive genetic algorithm. All these methods employ a finite number of parameters, and must therefore lose track of information about past data. Without a sufficient statistic, the argument used in section 2 to conclude that the optimal controller is deterministic does not hold, so it is of interest to determine whether a stochastic method turns out to be optimal in some cases, exactly or as a good approximation. In particular, it would be of great interest to be able to evaluate simulated annealing [13] in this framework.

6 Acknowledgement

This work was partly supported by EPSRC grant GR/J17814.

References

[1] V. Anantharam, P. P. Varaiya, and J. C. Walrand. Asymptotically

efficient allocation rules for the multiarmed bandit problem with multiple playes. *IEEE Trans. Auto. Control*, AC-32(11):968–976,977–982, November 1987.

[2] R. E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[4] R. T. Cox. Probability, frequency and reasonable expectations. *Amer. J. Phys.*, 14:1–26, 1946.

[5] T. S. Fersuson. *Mathematical Statistics : A Decision Theoretic Approach.* Academic Press, New York, 1967.

[6] R. Fletcher. *Practical Optimization.* J. Wiley & Sons, Chichester, 1987.

[7] F. Glover. Tabu search — part i. *ORSA J. Comp.*, 1(3):190–206, 1989.

[8] F. Glover. Tabu search — part ii. *ORSA J. Comp.*, 2(1):4–32, 1990.

[9] J. H. Holland. *Adaptation in Natural and Artificial Systems.* Univ. of Michigan Press, Ann Arbor, 1975.

[10] R. A. Howard. *Dynamic Programming and Markov Processes.* MIT Press, New York, 1960.

[11] H. Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, 1961.

[12] R. Keener. Further contributions to the "two-armed bandit" problem. *Ann. Statist.*, 13(1):418–422, 1985.

[13] C. Kirkpatrick, D. Gelat, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[14] P. R. Kumar and T. I. Seidman. On the optimal solution of the one-armed bandit adaptive control problem. *IEEE Trans. Auto. Control*, AC-26(5):1176–1184, 1981.

[15] D. V. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005, 1956.

[16] D. J. C. MacKey. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604, 1992.

[17] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, UK, 1988.

[18] P. P. Varaiya, J. C. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Auto. Control*, AC-30(5):426–439, May 1985.

[19] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ., 1947.

[20] C. J. C. H. Watkins. *Learning with delayed rewards*. PhD thesis, Psychology Dept, Cambridge University, 1989.

[21] P. Whittle. Multi-armed bandits and the Gittins index. *J. Roy. Stat. Soc., B*, 42(2):143–149, 1980.