FINITE DIMENSIONAL ALGORITHMS FOR THE HIDDEN MARKOV MODEL MULTI-ARMED BANDIT PROBLEM

Vikram Krishnamurthy and Josipa Mickova

Department of Electrical Engineering
University of Melbourne,
Parkville, Victoria 3052, Australia
email vikram@ee.mu.oz.au and josm@ee.mu.oz.au

ABSTRACT

The multi-arm bandit problem is widely used in scheduling of traffic in broadband networks, manufacturing systems and robotics. This paper presents a finite dimensional optimal solution to the multi-arm bandit problem for Hidden Markov Models. The key to solving any multi-arm bandit problem is to compute the Gittins index. In this paper a finite dimensional algorithm is presented which exactly computes the Gittins index. Suboptimal algorithms for computing the Gittins index are also presented and experimentally shown to perform almost as well as the optimal method. Finally an application of the algorithms to tracking multiple targets with a single intelligent sensor is presented.

Key words: Dynamic Programming, Gittins index, Multiarmed Bandit Problem, Hidden Markov Models

1. PROBLEM FORMULATION

The multi-armed bandit problem is a special case of a dynamic stochastic scheduling problem and has numerous applications in the scheduling of traffic in broadband networks, manufacturing systems and robotics. The standard multi-armed bandit problem involves fully observed Markov chains and is simply a Markov Decision Processes (MDP's) with a powerful structure. Several fast solutions have been recently proposed to solve this Markov chain multi-arm bandit problem (see [4] and [9]).

In this paper we give a solution of the hitherto unsolved multi-armed bandit problem for Hidden Markov Models (HMMs). The problem can be described as follows:

Problem: There are P parallel projects (or activities), $p = 1, \ldots, P$, of which only one can be worked on at any time instant. Let $s_k^{(p)}$ denote the state of project p at time k. We assume that each project p has a finite number of states $\mathcal{N}_p$, indexed by $i^{(p)} = 1, \ldots, \mathcal{N}_p$. If project p is worked on at time k, one receives an immediate reward of $\beta^k R(s_k^{(p)}, p)$ and the state $s_k^{(p)}$ evolves according to a $\mathcal{N}_p$ state Markov chain with transition probability matrix $A^{(p)} = (a_{ij}^{(p)})_{i,j \in \mathcal{N}_p}$. The states of all the other idle projects are unaffected, i.e., $s_{k+1}^{(p)} = s_k^{(p)}$, if the project p is idle at time k.

The total discounted reward over an infinite time horizon is given by:

$$J = \mathbf{E}[\sum_{k=0}^{\infty} \beta^k R(s_k^{(u_k)}, u_k)] \quad (1)$$

where, $0 < \beta < 1$ denotes the discount factor, $R(s_k^{(u_k)}, u_k)$ is the immediate reward of state s_k of project denoted by u_k, and u_k is the control at time k, which, in the multi-armed bandit case, corresponds to the project chosen to evolve at time k, so $u_k = 1, \ldots, P$. The sequence of controls, $\{u_k\}$, is called a policy, denoted by π. Define Π to the class of all possible policies π. Solving the multi-arm bandit problem involves determining the optimal policy $\pi \in \Pi$ which will maximize the total discounted reward (1).

The above problem formulation is the Markov chain multi-arm bandit problem and has been widely studied in the literature [9]. In this paper, motivated by the applications described below, we consider the HMM multi-arm bandit problem: We assume that the state of the project $s_k^{(p)}$ is not directly observed. Instead, noisy measurements (observations) $y_k^{(p)}$ of the active project state $s_k^{(p)}$ are available at time k. Assume that these observations $y_k^{(p)}$ belong to a finite set $\mathcal{M}_p$ indexed by $j^{(p)} = 1, \ldots, \mathcal{M}_p$. Let $M^{(p)} = (m_{ij}^{(p)})_{i \in \mathcal{N}_p, j \in \mathcal{M}_p}$ where $m_{ij}^{(p)} = P(y_k^{(p)} = j | s_k^{(p)} = i)$, denote the "symbol probabilities" of the HMM. *Our aim is to solve this HMM multi-arm bandit problem, i.e., determine the optimal policy $\pi \in \Pi$ which will yield the maximum possible reward in (1).* Let $J^* = \max_{\pi \in \Pi} J_\pi$ denote this optimal cost.

Application: Scheduling of a single intelligent sensor Numerous applications of the Markov chain multi-armed bandit problem in scheduling and manufacturing systems can be found in [2], [3] and [10]. Similar applications with noisy

This work was supported by an ARC large grant and CSSIP

observations can be formulated to fit the multi-arm bandit problem for HMMs. However, in this paper we present a novel application of the HMM multi-armed bandit problem in multi-target tracking. The problem is as follows:
There are P targets (e.g. aircraft). whose coordinates $s_k^{(p)}$, $p = 1, 2 \ldots, P$, evolve randomly according to P independent finite state Markov chains (see [] the use of Markov models in tracking). Assume that there is a single sensor which receives noisy measurements $y_k^{(p)}$ of **only one target** at any given time instant. Our aim is to answer the question: *Which single target should the sensor measure at each time instant?*
Since the sensor can measure only one target (say target p) at a given time, the coordinates of the other $P-1$ targets cannot be measured. The sensor simply assumes that the coordinates of these $P-1$ un-measured targets remain in the same state. To compensate for this risky assumption, the sensor places a penalty cost of $\sum_{q=1,\ q\neq p}^{P} \beta^k C(s_k^q, q)$ for these $P-1$ un-observed targets. This cost is typically depends on factors such as distance of the targets from the sensor when last measured. The aim is to minimize the cost incurred by the sensor for tracking the P targets over an infinite horizon.

The above problem is called the *tax problem* [1], [9]. It is similar to the HMM multi-arm bandit problem except that the non-evolving projects (targets) incur a cost. As shown in [1],[9] any tax problem can be converted to a multi-arm bandit problem. The above tax-problem is easily converted to the following HMM multi-arm bandit problem with reward function for project p equal to $R(s_k^{(p)}, p) = C(s_k^p, p) - \beta \mathbf{E}\left[C(s_{k+1}^{(p)}, p)\right]$.

As will be seen from the information state formulation below, the problem can be viewed as follows: Design an optimal scheduling policy to optimally choose at each time instant one target. Run a HMM filter to process these measurements to estimate the target's coordinates. (There is a computational cost associated with running the HMM filter for this target. However, since this cost is independent of the target it is hence irrelevant).

Information state formulation. As it stands, the above HMM multi-arm bandit problem is a partially observed infinite horizon stochastic control problem with a powerful structure. The structure considerably simplifies the solution, as will be shown later. But first, as is standard with such stochastic control problems – we convert the partially observed stochastic control problem to a fully observed stochastic control problem defined in terms of the *information state*.

The information state at time k, which we will denote by $x_k^{(p)}$, is merely the conditional filtered density of the Markov chain state $s_k^{(p)}$ given the past observations $Y_k^{(p)} = (y_0^{(p)}, \ldots, y_k^{(p)})$:

$$x_k^{(p)}(i) = P(s_k^{(p)} = i | Y_k^{(p)}) \tag{2}$$

The information state is a sufficient statistic to describe the current state of a HMM (see [5] and [1]) and thus converts the partially observed Markov chain to a fully observed one.

The information state update is computed straightforwardly by the HMM state filter (also known as the "forward algorithm"):

$$x_{k+1}^{(p)} = \frac{B^{(p)}(y_{k+1}) A^{(p)\prime} x_k^{(p)}}{\mathbf{1}' B^{(p)}(y_{k+1}) A^{(p)\prime} x_k^{(p)}} = T[x_k^{(p)}] \tag{3}$$

where $B^{(p)}(y_{k+1}) = (b_{ij}^{(p)})_{i \in \mathcal{N}_p, j \in \mathcal{M}_p}$ is a diagonal matrix formed by the y_{k+1}'th column of the observation matrix $M^{(p)}$, that is, $b_{ii}^{(p)} = m_{iy_{k+1}}^{(p)}, i \in \mathcal{N}_p, y_{k+1} \in \mathcal{M}_p$.

In terms of the information state, the total discounted reward (1) can be re-written as

$$J = \mathbf{E}[\sum_{k=0}^{\infty} \sum_{i=1}^{\mathcal{N}^{(u_k)}} R(i, u_k) x_k^{(u_k)}(i)] \tag{4}$$

where u_k and R are as in equation (1). The aim is to compute the optimal policy $\arg\max_{\pi \in \Pi} J_\pi$.

Notice that the information state is a continuous-state (infinite state) process. At each time it is a $\mathcal{N}_p$ dimensional vector, where each entry i of the vector is given by equation (2). Thus unlike the standard Markov chain multi-arm bandit problem – which can be optimally solved via finite dimensional dynamic programming, one might expect that the HMM multi-arm bandit problem requires infinite dimensional dynamic programming. The surprising result we will show is that there is a finite dimensional optimal algorithm.

Main Results: The general multi-arm bandit problem has a rich structure which makes possible a powerful solution methodology. It turns out that the optimal policies have an *index rule*; that is, for each project p, there is a a function $\gamma_p(s_k^{(p)})$ called the "Gittins index" such that the optimal policy at time k is to

$$\text{Work on project } q \text{ where } q = \arg\max_p \{\gamma_p(s_k^{(p)})\}$$

For a proof of this index rule for general multi-arm bandit problems please see [2], [10]. Thus computing the Gittins index is a key requirement for solving any multi-arm bandit problem. Thus far all algorithms for assigning the Gittins index have been for *finite state* MDPs.

The main contributions of this paper are:
1. We present a finite dimensional solution to compute the Gittins indices for the *infinite state* HMM multi-arm bandit problem. We will show that Gittins index can optimally be

assigned to all the states of an infinetely sized state space of a HMM, using dynamic programming.
2. The optimal algorithm, while finite dimensional, has a high computational cost. We will also present two suboptimal algorithms for calculating the Gittins index.
3. Finally we present numerical examples of the above algorithms for two applications: (i) The machine replacement problem – which is a universally used benchmark, (ii) The multi-target tracking problem.

The organization of this paper is as follows: section 2 presents the DP solution for calculating the Gittins indices of the POMDP information states; section 3 presents an algorithm on how to calculate the Gittins indices, optimally and suboptimally; and section 4 contains numerical examples which demonstrate the use of the Gittins index and compare the optimal and suboptimal methods of calculating it.

2. FINITE DIMENSIONAL SOLUTION FOR THE GITTINS INDEX

In this section we will formulate a finite dimensional Gittins index solution for the infinite state space of a HMM. The key ideas we will use are the following two results (we will drop the project index dependency since we will be working with only one bandit project):

Result 1 The return-to-state-x_0 problem. *In the return-to-state-x_0 problem [4], at any time, a maximization between the following two actions is carried out: continue project, that is, accumulate reward $\beta^k R(x_k)$ and evolve the project state as $x_{k+1} = T[x_k]$; or restart the project in state x_0, that is, accumulate reward $\beta^k R(x_0)$ and evolve the project state as $x_{k+1} = T[x_0]$. The value function of the return-to-state-x_0 problem is given by the finite horizon Bellman equation:*

$$\begin{aligned} V_k(x_k, x_0) &= \max[g' x_k + \beta \mathbf{E} V_{k+1}(T[x_k], x_0), \\ &\quad g' x_0 + \beta \mathbf{E} V_{k+1}(T[x_0], x_0)] \end{aligned} \tag{5}$$

where, $g = (R(1), \ldots, R(\mathcal{N})$, R is as in equation (1) without the project dependency, $0 < \beta < 1$ is a discount factor.

As k$\rightarrow\infty$, the value function from equation (5) evaluated at x_0 becomes the Gittins index of state x_0. To obtain the Gittins index for every state x_i, the return-to-state-x_i calculation has to be repeated for every x_i.

Result 2 Finite representation of the HMM value funtion. *At every time instant the finite horizon value function of a HMM is piecewise-linear, therefore it can be represented with a finite number of vectors (see [6], [7] and [8]).*

We can combine these two results, and state the following theorem:

Theorem 1 *The value function of a HMM return-to-state-x_0 problem (5) can be written in closed form, that is, it can be represented by a finite number of vectors in a new augmented state space defined as (x_k, x_0), in the following way:*

$$V_k(x_k, x_0) = \max_{i,j}[(c, d)_i'(x_k, x_0), (0, f)_j'(x_k, x_0)] \tag{6}$$

where the augmented space (x_k, x_0) is a concatenation of two information states, therefore is a vector of dimension $2\mathcal{N}$. Vectors $(c, d)_i$ and $(0, f)_j$ are also of this size.

PROOF. The proof is by induction. At time N,

$$V_N(x_N, x_0) = \max[(g, g)'(x_N, x_0), (0, g)'((x_N, x_0)] \tag{7}$$

where vector g is as in equation (5). Notice that equation (7) is of the required form stated in the theorem.

Assume that at time $k+1$, the value function also has the form of equation (6):

$$\begin{aligned} V_{k+1}(x_{k+1}, x_0) = \max_{i,j}[&(c, d)_i'(x_{k+1}, x_0), \\ &(0, f)_j'(x_{k+1}, x_0)] \end{aligned} \tag{8}$$

Substitute equation (8) into Bellman's equation (5),

$$\begin{aligned} V_k(x_k, x_0) = &\max_{i,j}[(g, 0)'(x_k, x_0) + \\ &\beta\, \mathbf{E}[\max_{i,j}[(c, d)_i'(x_{k+1}, x_0), (0, f)_j'(x_{k+1}, x_0)] | x_k = x_k], \\ &(g, 0)'(x_0, x_0) + \beta \mathbf{E}[\max_{i,j}[(c, d)_i'(x_{k+1}, x_0), \\ &(0, f)_j'(x_{k+1}, x_0) | x_k = x_0]]] \end{aligned}$$

After some algebraic manipulations this yields

$$\begin{aligned} &V_k(x_k, x_0) = \max_{i,j}[(g, 0)'(x_k, x_0) + \\ &\max_{i,j}[(c\beta \sum_m B(m) A' x_k, \beta d x_0)_j), (0, \beta f)_j(x_k, x_0)], \\ &(g, 0)'(x_0, x_0) + \max_{i,j}[(c\beta \sum_m B(m) A' x_0, \beta d x_0), \\ &(0, \beta f)_j(x_0, x_0)]] \end{aligned}$$

Taking out the inner maximizations:

$$\begin{aligned} &V_k(x_k, x_0) = \\ &\max_{i,j}[(g + \beta c \sum_m B(m) A', \beta d)_i'(x_k, x_0), \\ &(g, \beta f)_j'(x_k, x_0), (0, g + \beta c_i \sum_m B(m) A' + \beta d)_i'(x_k, x_0), \\ &(0, g + \beta f)_j'(x_k, x_0)] \end{aligned} \tag{9}$$

$$\tag{10}$$

Thus, the augmented state space has a closed form solution. □

3. ALGORITHMS

In this section we will state the optimal and suboptimal algorithms for computing the Gittins index.

Optimal. There are numerous standard algorithms [6], [7] and [8] that can be used to compute the finite set of vectors depicted in equation (10). They are all based on the on the result from theorem 1. After obtaining these vectors, the Gittins indices are given as follows: $\gamma_{opt}(x) = \max_i[\alpha_i'(x,x)]$, where the α_i's are vectors of the same size as the augmented state space defined in theorem 1. The Gittins index of state x is the return-to-state-x problem value function evaluated at state x.

Suboptimal. For large size problems, the above algorithm iteration can be time and memory consuming. For such cases further approximations of the index can be used. We state these next.

Definition 1 *The expected Gittins index of an information state x is defined as* $\gamma_{exp}(x) = \sum_i \gamma(i)x(i)$

where $\gamma(i)$ is the Gittins index of the i^{th} state of the process we are trying to observe, and $x(i)$ is given by equation (2).

Definition 2 *The MAP (maximum a posteriori) value of the Gittins index of an information state x is defined as* $\gamma_{MAP}(x) = \gamma(\max_{x_i}[i])$

The MAP Gittins index of the information state corresponds to the Gittins index of the state i which has the the largest $x(i)$ value in the information state x given by (2).

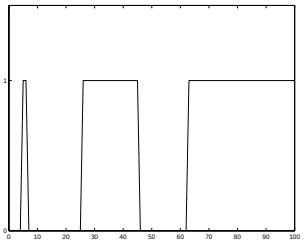

Figure 1: Sensor tracking both targets

4. NUMERICAL STUDIES

Machine replacement problem. The aim of this numerical example is to compare the optimality of the three different methods for calculating the Gittins index. Here, we have a 3-process multi-arm-bandit machine replacement problem. The Gittins indices for each of these processes were calculated using the optimal and the two suboptimal methods given in section 3.

The multi-armed bandit process was simulated once for each different algorithm from section 3, over a horizon length of 500. The optimal algorithm gave the largest total reward while the suboptimal algorithms both had a total reward which was 99% of the optimal result.

The two-state tracking problem. The aim of this numerical example is to demonstrate how the index is used to optimally choose the best project to evolve each time. Here, we set up two targets (two bandit processes) whose indices were computed via the optimal algorithm given in section 3. The targets were labeled 0 and 1, and their tracking was simulated for 100 time steps. Figure 1 demonstrates the time intervals when the sensor was tracking each of the targets. In this simulation both targets had comparable indices so the sensor switches between them.

5. REFERENCES

[1] D. Bertsekas, *Dynamic Programming and Optimal Control, Vol I*, Athena Scientific, 1995.

[2] J.C. Gittins, *Bandit Processes and Dynamic Allocation Indices*, J. R. Statist. Soc. B, Vol. 41, No. 2, pp. 148-177, 1979.

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

[4] M.N. Katehakis and A.F. Veinott, Jr., *The Multi-armed Bandit Problem: Decomposition and Computation*, Mathematics of Operations Research, Vol. 12, No. 2, pp. 262-268, 1987.

[5] P.R. Kumar, P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, 1985.

[6] W.S. Lovejoy, *A Survey of Logarithmic Methods for Partially Observed Markov Decision Processes*, Annals of Operations Research, Vol. 28, No. 1, pp. 47-66, 1991.

[7] G.E. Monahan, *A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algoritms*, Management Science, Vol. 28, No. 1, pp. 1-16, 1982.

[8] R.D. Smallwood and E.J. Sondik, *The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon*, Operations Research, Vol. 21, No. 5, pp. 1071-1088, 1973.

[9] P.P. Varaiya, J.C. Warland and C. Buyukkoc, *Extensions of the Multiarmed Bandit Problem: The Discounted Case*, IEE Trans. on Auto. Control, Vol. AC-30, No. 5, pp. 426-439, 1985.

[10] P. Whittle, *Multi-armed Bandits and the Gittins index*, J. R. Statist. Soc. B, Vol. 42, No. 2, pp. 143-149, 1980.