*J. Appl. Prob.* **42,** 754–765 (2005)
*Printed in Israel*
© *Applied Probability Trust* 2005

# TAX PROBLEMS IN THE UNDISCOUNTED CASE

P. WHITTLE,* *University of Cambridge*

## Abstract

The aim of this paper is to evaluate the performance of the optimal policy (the Gittins index policy) for open tax problems of the type considered by Klimov in the undiscounted limit. In this limit, the state-dependent part of the cost is linear in the state occupation numbers for the multi-armed bandit, but is quadratic for the tax problem. The discussion of the passage to the limit for the tax problem is believed to be largely new; the principal novelty is our evaluation of the matrix of the quadratic form. These results are confirmed by a dynamic programming analysis, which also suggests how the optimal policy should be modified when resources can be freely deployed only within workstations, rather than system-wide.

*Keywords:* Tax problem; multi-armed bandit; time sharing; undiscounted limit; dynamic programming

2000 Mathematics Subject Classification: Primary 60G40
Secondary 90B36; 68M20

## 1. Introduction

The title of this paper does not refer to the levy imposed by the state upon its grateful citizens, but rather to the problem of optimal scheduling in the case when all work stored in a system continues to incur cost until it is cleared. This is a variant of the classic multi-armed bandit problem, although it demonstrates a distinctly different structure. However, we begin with a discussion of the multi-armed bandit problem in order to set the scene and to prepare for the passage to the undiscounted version, which shows special simplifications in both cases.

In the literature since about 1980, the multi-armed bandit problem has been seen as one of choosing in which one of a number of 'projects' to engage, on the basis of the 'states' of the projects currently available. However, the tax problem is typically concerned with the scheduling of manufacturing processes, and so we shall speak rather of 'items' to be chosen for processing. For the same reason, we consider the open version of the problem, in which there is an inflow of new items to the system, balanced (if the system is not overloaded) by the discharge of completed items. In the multi-armed bandit case, the inflow is balanced by the effective discharge of items that have reached a state which makes further work on them unprofitable.

The papers by Klimov (1974), (1978) on time-sharing systems (essentially the tax problem) are now classic. Klimov considered the undiscounted version of the problem with a general distribution of processing times, proved optimality of an index policy, and found recursions for evaluation of the indices. However, this determination of the indices was far from explicit, and there was no evaluation of performance (i.e. average cost per unit time). Gittins' equally

Received 28 July 2004; revision received 2 December 2004.

* Postal address: Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. Email address: whittle@statslab.cam.ac.uk

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

classic solution of the multi-armed bandit problem (see Gittins (1979), (1989)) was given a dynamic programming treatment by Whittle (1980), (1981) which did evaluate performance, in that it gave explicit formulae for the value function (the minimal expected discounted future cost, conditional on current state occupation numbers). Varaiya *et al.* (1985) linked the tax and bandit problems and indicated how results could be extended to non-Markov models. Lai and Ying (1988) considered the determination of the Gittins indices for the tax problem in the undiscounted limit. However, neither of these last two sets of authors considered actual performance, once optimality had been established, and the topic seems to have attracted little attention since. In fact, it raises novel issues reflecting the very different characters of the bandit and tax problems.

In this paper, we address the evaluation of value functions, which in the undiscounted limit will mean the determination of minimal average cost and minimal differential cost. We assume Markov models and use a dynamic programming treatment, which leads to an analysis as compact and explicit as the subject allows, and suggests conjectures which can then be tested by other means. The Markov formulation given here forces a restriction to exponentially distributed processing times and scarcely distinguishes between a preemptive and nonpreemptive operation. These are limitations which could be removed by the assumption of passage through Markov substates during a given processing step. However, to follow that path now would obscure the line of the paper.

The treatment of the next two sections is formal, the aim being to see the argument and the further advances latent in it. The solutions thus suggested are confirmed by direct analysis in Sections 4 and 5.

The passage to the undiscounted case has a remarkably simplifying effect, and it is worthwhile to consider the implications of this move before beginning the analysis. Suppose that we have a ‘closed’ version of the multi-armed bandit problem, in that we consider the allocation of processing effort over a fixed number $N$ of statistically identical items. Suppose that all items (states) are communicating. Then the average reward is independent of the policy – it will be equal to the minimal value of the Gittins index no matter what policy we follow. (The policy will, however, affect the differential reward, i.e. the difference in expected total reward if we start from a given state rather than from some standardized initial condition.) If there are several distinct (non-communicating) classes of item then ultimately we will process only items for which this minimal (over states) index is maximal.

The analogous tax case would concern the processing of $N$ given items that must pass through various stages of manufacture before processing is complete, when the items effectively leave the system and are henceforth costless. In this case, the average cost is zero, but the total cost will certainly depend upon the order in which items are processed. This is because an item incurs a cost, which is in general state dependent, as long as it is present in the system.

If we now consider an open multi-armed bandit problem, in which new items are entering the system and items which are insufficiently rewarding can effectively be shelved, then average-optimality of policy does demand that we process only those items whose Gittins index is not less than some critical value $\overline{\nu}$, which in fact equals the average reward. The order in which we process items is immaterial as long as items whose index exceeds $\overline{\nu}$ are present. If there are no such items, then we must work on an item whose index equals $\overline{\nu}$, of which an infinite number will have accumulated.

In the tax case, we do not have the option of shelving an item: all items whose processing is incomplete incur a cost. The system must then satisfy a traffic stability condition, namely that it have the resources to complete items at a rate greater than that at which they arrive. Even under

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

the mild demand of average-optimality, the order of processing is critical and its optimization nontrivial, although much eased by the absence of discounting.

## 2. Generalities on the multi-armed bandit

We assume that items can take states $j = 1, 2, \dots, J$, plus an additional state 0, corresponding to completion and passage out of the system, in which no further cost is incurred or action required. Suppose that at a given time there are are $n_j$ items in state $j$ in the system ($j \neq 0$). We shall assume Markov rules which ensure that the $J$-vector $\boldsymbol{n} = \{n_j\}$ is indeed the Markov state of the system. Let $\boldsymbol{e}_j$ be a $J$-vector with component one in the $j$th place and zeros elsewhere. Then the transition in which an item of state $j$ changes to state $k$ corresponds to a transition $\boldsymbol{n} \to \boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_k$; we shall suppose that this has probability intensity $p\mu_{jk}$ if $p$ is the rate of effort in the processing of the item. The arrival of a new item of state $j$ corresponds to a transition $\boldsymbol{n} \to \boldsymbol{n} + \boldsymbol{e}_j$, which we shall suppose has transition intensity $\lambda_j$.

For notation simplicity, we assume that all items follow the same transition rules. This incurs no loss of generality: the occurrence of distinct classes of item will correspond to the presence of non-communicating classes of states under the prescribed transition rates.

First consider a system in which unit processing effort is available, the processing of an item in state $j$ yields rewards at rate $r_j$ per unit time, and the aim is to allocate processing in such a way as to maximize the expected discounted future reward over an infinite horizon. Let $F(\boldsymbol{n})$ be the value function, i.e. the maximal expected value of this discounted future reward, conditional on the current value $\boldsymbol{n}$ of the system state. Then $F$ will obey the dynamic programming equation

$$\max_j (r_j - \alpha F + \Lambda_j F + \Lambda_0 F) = 0, \tag{1}$$

where

$$\begin{aligned} \Lambda_j F &= \sum_k \mu_{jk}(F(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_k) - F(\boldsymbol{n})), \\ \Lambda_0 F &= \sum_k \lambda_k (F(\boldsymbol{n} + \boldsymbol{e}_k) - F(\boldsymbol{n})), \end{aligned} \tag{2}$$

and $\alpha$ is the continuous-time interest rate.

Let us write (1) as $\max_j \Omega_j F = 0$. Then there is interest in considering the more general equation

$$\max\big(\max_j \Omega_j F, M - F\big) = 0, \tag{3}$$

where $M$ is a prescribed positive constant. The implication of (3) is that, in addition to the option of choosing one of the items for processing, we have the option of drawing a lump reward $M$ and ceasing operations. This embedding of the problem was the key step in Gittins' argument. We shall speak of this stopping option as *resignation* and shall write the solution of (3) as $F(\boldsymbol{n}, M)$. Suppose that there is an upper bound $R$ on the value of the total discounted reward, and that the value of $M$ is such that resignation in a finite time is certain. It was shown in Whittle (1981) that the solution of (3) can then be written

$$F(\boldsymbol{n}, M) = R - \int_M^R \prod_j \left(\frac{\mathrm{d}\phi_j(m)}{\mathrm{d}m}\right)^{n_j} \mathrm{d}m. \tag{4}$$

Here, $\phi_j(M) = F(\boldsymbol{e}_j, M)$, which then obeys the equation

$$\max(\Omega_j F(\boldsymbol{e}_j, M), M - F(\boldsymbol{e}_j, M)) = 0. \tag{5}$$

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

Equations (4) and (5) determine the $\phi_j$ in principle, and so determine $F(\boldsymbol{n}, M)$. Furthermore, the value $M_j$ of $M$ for which the two options in (5) are equally attractive determines the *Gittins index*: if processing continues then it is optimal to process an item present which is of highest index.

Now consider the case of small $\alpha$, preparatory to considering the undiscounted case. We shall choose $M = \nu/\alpha$, so that the terminal reward is equivalent to a fixed income of $\nu$ in the indefinite future. Then $\phi_j(M) \geq M = \nu/\alpha$, the difference being extra rewards earned before resignation. We then conjecture that, for small $\alpha$,

$$\phi_j\left(\frac{\nu}{\alpha}\right) = \frac{\nu}{\alpha} + \psi_j(\nu) + o(1). \tag{6}$$

Here, $\psi_j(\nu)$ is nonnegative and is zero for $\nu \geq \nu_j$, where $\nu_j$ is the index value of retirement income corresponding to $M_j$ (in future, we shall use this version of the index). By inserting (6) into (4), we find that

$$\begin{aligned}F\left(\boldsymbol{n}, \frac{\nu}{\alpha}\right) &= R - \alpha^{-1}\int_\nu^{\alpha R}\prod_j\left(1 + \alpha\frac{\mathrm{d}\psi_j(\sigma)}{\mathrm{d}\sigma} + o(\alpha)\right)^{n_j}\mathrm{d}\sigma \\ &= R - \alpha^{-1}\int_\nu^{\alpha R}\prod_j\left(1 + \alpha\sum_j n_j\frac{\mathrm{d}\psi_j(\sigma)}{\mathrm{d}\sigma} + o(\alpha)\right)\mathrm{d}\sigma \\ &= \frac{\nu}{\alpha} + \sum_j \psi_j(\nu)n_j + o(1).\end{aligned} \tag{7}$$

This is the great simplification of the undiscounted case: the value function $F$ is linear in $\boldsymbol{n}$. Inserting (7) back into (5), and taking the limit as $\alpha \to 0$, we find the set of relations

$$\max\left(r_j - \nu + \sum_k \mu_{jk}(\psi_k - \psi_j) + \sum_k \lambda_k\psi_k, -\psi_j\right) = 0, \tag{8}$$

where the argument $\nu$ of the $\psi_j$ is understood. Relation (8) determines the $\psi_j(\nu)$, and the value of $\nu$ at which state $j$ is on the decision boundary (of resignation or continuation) is the Gittins index $\nu_j$. A direct proof of this assertion is immediate.

**Theorem 1.** *Assume, for notational convenience, that the states $j$ are numbered in order of decreasing index $\nu_j$, and suppose that, under operation of the index policy (which is known to be optimal), all items in states $j < h$ and some items in state $h$ are chosen for processing. Then, under optimal operation, the average reward is $\nu_h$ and the differential reward $\psi(\boldsymbol{n}) = F(\boldsymbol{n}) - F(\boldsymbol{0})$ has the form*

$$\psi(\boldsymbol{n}) = \sum_j \psi_j(\nu_h)n_j.$$

*Proof.* Suppose that, under the index rule, the average cost is $\gamma$ and the differential cost takes the linear form

$$\psi(\boldsymbol{n}) = \sum_j f_j n_j, \tag{9}$$

for some coefficients $f_j$. Since $n_j$ will in the course of time become infinite for $j \geq h$, we must have $f_j = 0$ in this range. The dynamic programming equation for operation of the index policy in the undiscounted case is

$$\gamma = r_j + \Lambda_j\psi + \Lambda_0\psi, \tag{10}$$

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

where $j$ (a function of $\boldsymbol{n}$) is the state of the item of maximal index which is present in the system. If we assume the linear form (9), then (10) becomes,

$$\gamma = r_j + \sum_k \mu_{jk}(f_k - f_j) + \sum_k \lambda_k f_k, \qquad j \le h,$$

with $f_j = 0$ for the remaining values of $j$. However, in this system we recognise (8) with the identifications $\gamma = \nu_h$ and $f_j = \psi_j(\nu_h)$.

The occurrence of an effective cutoff at state $h$ is something that also manifests itself in the discounted case. For the situation of (3), ultimate resignation must be certain if the solution (4) is to be valid. Choose the smallest value of $M$ for which this is so. This critical value will be the index $M_h$ of some state, which we shall again denote by $h$. Then all items of index greater than $M_h$ will be processed, no item of lesser index will be, and items of that index (of which an infinite reserve will have accumulated) will be used on those occasions (recurrent, by the definition of $h$) when no items of higher index are present. The value function $F(\boldsymbol{n}, M_h)$ can be separated into $M_h$ and $F - M_h$. The first term represents the future return from the system on those recurrent occasions when 'emptiness' (of the set of items with index greater than $M_h$) has been reached. The second represents the differential reward gained during passage from state $\boldsymbol{n}$ to the state of emptiness.

## 3. Generalities on the tax problem

For the tax case, the problem is one of minimizing cost. Denote the value function of the problem by $G(\boldsymbol{n})$, a minimal expected discounted future cost conditional on the current system state $\boldsymbol{n}$. If each item of state $j$ in the system incurs a cost of $c_j$ per unit time and costs are additive, then the total cost incurred per unit time is $\boldsymbol{c}^\top \boldsymbol{n} = \sum_j c_j n_j$. The dynamic programming equation analogous to (1) is then

$$\min_j(\boldsymbol{c}^\top \boldsymbol{n} - \alpha G + \Lambda_j G + \Lambda_0 G) = 0, \tag{11}$$

where $j$ is again restricted to values for which $n_j > 0$. This can be transformed into the form of (1) if we define $F(\boldsymbol{n}) = (\boldsymbol{c}^\top \boldsymbol{n}/\alpha) - G(\boldsymbol{n})$; then (11) becomes (1) with

$$r_j = \alpha^{-1}\biggl(\sum_k \mu_{jk}(c_j - c_k) - \sum_k \lambda_k c_k\biggr). \tag{12}$$

Effectively, the cost rate $c_j$ incurred during occupation of the state $j$ has been replaced by a *present* lump charge of $c_j/\alpha$ on entry to state $j$ and the same *present* reward upon leaving it. This amounts to a cost of $(c_j/\alpha)(1 - \mathrm{e}^{-\alpha\tau})$ incurred on arrival in state $j$, where $\tau$ is the length of sojourn in that state.

With this, we seem to have reduced the tax problem to the multi-armed bandit problem, but this is very far from the case; the occurrence of the factor $\alpha^{-1}$ in (12) has a profound effect. Nevertheless, the determination of the Gittins index is now immediate. We have to suppose a resignation reward (in the reward version of the problem associated with (1)) of $\nu/\alpha^2$ rather than $\nu/\alpha$, because of the factor $\alpha^{-1}$ in (12). For the quantity $\phi_j(\nu/\alpha^2)$ in the problem, we now assume the following form analogous to (6), where the $\theta_j$ are $\nu$-dependent coefficients:

$$\phi_j\left(\frac{\nu}{\alpha^2}\right) = \frac{\nu}{\alpha^2} + \frac{\psi_j(\nu)}{\alpha} + \theta_j(\nu) + o(1). \tag{13}$$

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

The analogue of the index-determining system (8) is then

$$\max\left(\sum_k \mu_{jk}(\Delta_k - \Delta_j) + \sum_k \lambda_k \Delta_k - \nu, -\psi_j\right) = 0, \tag{14}$$

where $\Delta_j(\nu) = \psi_j(\nu) - c_j$. Note that the matrix of the system (14) is nonsingular when $\nu = 0$ and, hence, that $\Delta_j(0) = 0$, or $\psi_j(0) = c_j$.

For what is probably the simplest example of interest, consider the case in which items proceed through two consecutive stages of processing before completion. Items arrive at rate $\lambda$ into state 2; when processed at unit effort they then move at rate $\mu_2$ to state 1. From that state they move at rate $\mu_1$ to the completion state 0 if processed at unit effort. The index-determining relations (14) then become

$$\max(\mu_2(\Delta_1 - \Delta_2) + \lambda\Delta_2 - \nu, -\psi_2) = 0,$$
$$\max(-\mu_1\Delta_1 + \lambda\Delta_2 - \nu, -\psi_1) = 0.$$

Suppose that $\nu$ is so small that the first option (of continuation) holds in both equations. We then find that

$$\psi_j = c_j - \kappa T_j, \qquad j = 1, 2, \tag{15}$$

where $\kappa = \nu - \lambda\Delta_2$ and $T_j$ is the expected time required for completion from state $j$, that is, $T_1 = \mu_1^{-1}$ and $T_2 = \mu_1^{-1} + \mu_2^{-1}$. As we increase $\nu$, a break point will occur when one of the $\psi$-values first becomes negative. We see from (15) that this will occur in state 2 or state 1 (meaning that items of state 1 or state 2 will have the higher index and, so, priority), according to whether the cost ratio $c_1/c_2$ is, respectively, greater than or less than the critical value $T_1/T_2 = \mu_2/(\mu_1 + \mu_2)$.

Since this is less than one, there is a prejudice in favour of processing first those items which are nearer completion.

Suppose for definiteness that it is $\psi_2$ which first becomes zero. We can solve for $\Delta_2$ and the critical value $\nu_2$ of $\nu$, and find that

$$\nu_2 = (1 - \lambda T_2)\frac{c_2}{T_2}. \tag{16}$$

This expression only makes sense if $\lambda T_2 < 1$, which is a stability condition, namely that the inflow rate of work must not exceed the capacity of the system. In this we see a difference from the case of the multi-armed bandit, in which we had the option of discarding unprofitable items. Now all items must be processed to completion, which sets a lower bound on system capacity.

For $\nu > \nu_2$ we have $\psi_2 = 0$ and, so, $\kappa = \nu - \lambda c_2$. We then find that the critical value of $\nu$ which makes $\psi_1$ zero is

$$\nu_1 = \mu_1 c_1 - \lambda c_2. \tag{17}$$

The presence of the factor $\alpha^{-1}$ in (12) indicates that this is a cost (or reward) which is maintained at a constant value in the indefinite future. The resignation reward which balances a future of such costs must then meet a constant stream of such permanent obligations, i.e. meet an obligation which increases linearly in time. Hence, the resignation reward takes the form $\nu/\alpha^2$. This potential embarrassment is in the end negated by the fact that, if the system processes all items (as it must do, and will do if it has adequate capacity), then idle periods will

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

recur and the resignment option will never be exercised. That is, we set $\nu = 0$ in the end, in that if $G(\boldsymbol{n}, M)$ is the tax analogue of $F(\boldsymbol{n}, M)$, we ultimately set $M = 0$.

We would now like to evaluate the performance of the index rule, which means evaluation of the value function $G$ and of the average cost rate which this implies in the undiscounted limit. We return again to (7), which now becomes

$$F\left(\boldsymbol{n}, \frac{\nu}{\alpha^2}\right) = R - \alpha^{-2} \int_{\nu}^{\alpha^2 R} \prod_j \left(1 + \alpha \frac{\mathrm{d}\psi_j(\sigma)}{\mathrm{d}\sigma} + \alpha^2 \frac{\mathrm{d}\theta_j(\sigma)}{\mathrm{d}\sigma} + o(\alpha^2)\right)^{n_j} \mathrm{d}\sigma.$$

Expansion of this expression in powers of $\alpha$ up to the zeroth-order term leads, if we recall that $\psi_j(0) = c_j$, to the expression

$$G(\boldsymbol{n}, 0) = \left(\frac{\boldsymbol{c}^\top \boldsymbol{n}}{\alpha}\right) - F(\boldsymbol{n}, 0) = -\sum_j \left(\theta_j(0) + \frac{Q_{jj}}{2}\right) n_j + \frac{1}{2} \boldsymbol{n}^\top \boldsymbol{Q} \boldsymbol{n} + o(1), \tag{18}$$

where

$$\boldsymbol{Q} = \int_0^\infty \frac{\mathrm{d}\boldsymbol{\psi}(\nu)}{\mathrm{d}\nu} \frac{\mathrm{d}\boldsymbol{\psi}(\nu)^\top}{\mathrm{d}\nu} \mathrm{d}\nu = \left[\int_0^\infty \frac{\mathrm{d}\psi_j(\nu)}{\mathrm{d}\nu} \frac{\mathrm{d}\psi_k(\nu)}{\mathrm{d}\nu} \mathrm{d}\nu\right]. \tag{19}$$

Here, $\boldsymbol{\psi}(\nu)$ is the column vector with elements $\psi_j(\nu)$. The argument makes it plausible that the essential term in the differential cost for the undiscounted tax problem is a *quadratic* form in $\boldsymbol{n}$, with matrix given by (19). This indicates the essential difference between the multi-armed bandit and tax problems. The value of $\theta_j(0)$ can be calculated from the determining equation for $\phi_j(\nu/\alpha^2)$, but we prefer to leave this until Section 5, when the calculation can be combined with the determination of the minimal average cost rate $\gamma$. This cost rate does not appear in the present calculation, because $G(\boldsymbol{n}, 0)$ evaluates costs up to first emptiness, rather than over an infinite horizon.

The only point in our calculations which cannot immediately be made rigorous is the supposition of a valid expansion (13). In the next two sections, we shall give direct derivations which support these results and give evaluations where they are still lacking. These confirm the evaluation (19) for the quadratic component of cost.

## 4. The undiscounted tax problem in deterministic form

We can imagine that effort is divided, so that an amount $p_j$ of effort is devoted to some unit of state $j$, with $\sum_j p_j \leq 1$ if, as we suppose, the total rate of effort available has been normalized to one. In this case, $\Lambda_j$ will be replaced in (11) by $\sum_j p_j \Lambda_j$ and minimization with respect to $j$ will be replaced by a minimization over the distribution $p$ (concentrated on those $j$ which are currently present). Now consider a deterministic version of the problem in which $\boldsymbol{n}$, instead of being a vector of integers following the Markov transitions thus specified, is a continuous variable obeying the equivalent deterministic equations

$$\dot{n}_j = \sum_k (p_k \mu_{kj} - p_j \mu_{jk}) + \lambda_j, \tag{20}$$

where a dot denotes a time derivative. We can regard this version as the limit case, as $\delta \downarrow 0$, of the stochastic model in which work comes in quanta of size $\delta$, the transition rates $\lambda$ and $\mu$ are changed to $\lambda/\delta$ and $\mu/\delta$, and $\boldsymbol{n}$ is now taken as $\delta$ times the numbers of quanta. The index

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

policy will remain optimal and the index values unchanged under this scaling, and the same will then be true of the limiting deterministic version.

Suppose that we write (20) in matrix form as $\dot{\boldsymbol{n}} = -\boldsymbol{L}\boldsymbol{p} + \boldsymbol{\lambda}$, so that the inflow $\boldsymbol{\lambda}$ could be exactly balanced if we chose $\boldsymbol{p} = \boldsymbol{L}^{-1}\boldsymbol{\lambda}$. The related inequality

$$\mathbf{1}\boldsymbol{L}^{-1}\boldsymbol{\lambda} < 1, \tag{21}$$

where $\mathbf{1}$ is a row $J$-vector of ones, is a classic necessary and sufficient condition for stability of the system. That is, the unit rate of effort available to the system is sufficient that $\boldsymbol{n}$ can be reduced to zero from any starting value.

The value function for this deterministic case is, in a sense, directly determinable. Consider, for concreteness, the two-state example of the previous section, and suppose that $n_1$ and $n_2$ are initially both positive. Under the assumptions of that example, we should initially allocate all attention to items of state 1, so that

$$\dot{n}_1 = -\mu_1 \quad \text{and} \quad \dot{n}_2 = \lambda. \tag{22}$$

Once $n_1$ has been reduced to zero, we hold it at that value and use the effort remaining to reduce $n_2$, so that

$$\dot{n}_1 = p_2\mu_2 - p_1\mu_1 = 0 \quad \text{and} \quad \dot{n}_2 = \lambda - p_2\mu_2 = \lambda - T_2^{-1}. \tag{23}$$

Here, we have given $p_1 = 1 - p_2$ the value required to hold $n_1$ at zero. There is then effort remaining to reduce $n_2$ if $\lambda T_2 < 1$, which is in fact just the stability condition (21) for this case. It follows from (22) and (23) that $\boldsymbol{n}$ is a piecewise linear function of $t$, and we readily verify the total cost (the integrated value over time of $c_1n_1 + c_2n_2$) to be

$$G(\boldsymbol{n}) = \tfrac{1}{2}\boldsymbol{n}^\top \boldsymbol{Q}\boldsymbol{n}, \tag{24}$$

where

$$Q_{11} = c_1T_1 + \frac{\lambda c_2 T_1^2}{1-\lambda T_2}, \qquad Q_{12} = \frac{c_2T_1}{1-\lambda T_2}, \qquad Q_{22} = \frac{c_2T_2}{1-\lambda T_2}.$$

These evaluations agree with (19), as can be verified from the evaluations of $\psi_1$ and $\psi_2$ (which are $\psi_1 = \psi_2 = 0$ for $\nu > \nu_1$, $\psi_1 = c_1 - \nu T_1$ and $\psi_2 = 0$ for $\nu_2 < \nu_1$, and

$$\psi_1 = \psi_2 = c_j - \frac{\nu T_j}{1-\lambda T_2}$$

for $\nu < \nu_2$), where the break values $\nu_j$ are given by (16) and (17). The average cost rate $\gamma$ is zero if the system obeys the stability condition, because $\boldsymbol{n}$ can be brought to zero and then held there.

It is not necessary to explicitly consider effort-sharing in the stochastic case, because (to take the above example) once $n_1$ has been reduced to zero, we can take time off to knock it back to zero every time a new item of state 1 appears. In this way, we achieve an effective effort-sharing over time. If we consider the $\delta$-scaled version of the process considered above, then this alternation of effort becomes ever more rapid as $\delta$ decreases, until in the deterministic limit we have an explicit effort-sharing.

We can now state and prove the result which has been suggested.

**Theorem 2.** *Suppose that the deterministic system specified by (20) satisfies the stability condition (21). Then the minimal cost (achieved by the index policy) of such stabilization is given by (24), with* $\boldsymbol{Q}$ *as given by (19).*

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

*Proof.* In the deterministic case, differentials will replace differences in the definitions (2) of the operators $\Lambda_j$, so that now

$$\Lambda_j = \sum_k \mu_{jk}(D_k - D_j), \qquad \Lambda_0 = \sum_k \lambda_k D_k,$$

where $D_j$ is the differential operator $\partial/\partial n_j$. With this understanding, the dynamic programming equation for the value function $G(\boldsymbol{n})$ is

$$\min_j \Omega_j G(\boldsymbol{n}) := \min_j (\boldsymbol{c}^\top \boldsymbol{n} + (\Lambda_j + \Lambda_0)G(\boldsymbol{n})) = 0, \tag{25}$$

where the choice of $j$ is again restricted to values for which $n_j > 0$. It is true that actual operation will demand effort-sharing, but (25) will single out the states which have immediate priority. We know that the index policy is optimal; assume again that the states have been numbered in order of nonincreasing index. Suppose that $j$ is indeed the state of highest index for which $n_j > 0$. We then wish to show that $\Omega_j(\boldsymbol{n}^\top \boldsymbol{Q}\boldsymbol{n}/2) = 0$. This amounts to the condition

$$c_i + \sum_k \mu_{jk}(Q_{ki} - Q_{ji}) + \sum_k \lambda_k Q_{ki} = 0, \qquad i \geq j. \tag{26}$$

The reason for the restriction on the set of $i$-values is that $n_i = 0$ for $i < j$, and so there is neither a need nor a basis for such a condition in this range. Let us denote $\psi_j'(\nu)$, the derivative of $\psi_j(\nu)$ with respect to $\nu$, by $f_j$ and $\sum_k \mu_{jk}(f_k - f_j) + \sum_k \lambda_k f_k$ by $L_j f$. Then (26) amounts to

$$c_i + \int_0^\infty (L_j f) f_i \, \mathrm{d}\nu = 0, \qquad i \geq j. \tag{27}$$

However, $L_j \Delta = \nu$, meaning that $L_j f = 1$ for $\nu < \nu_j$, and $\psi_i = 0$, meaning that $f_i = 0$ for $\nu > \nu_i$. The left-hand side of (27) thus equals

$$c_i + \int_0^{\nu_i} f_i \, \mathrm{d}\nu = c_i + \psi_i(\nu_i) - \psi_i(0) = c_i + 0 - c_i = 0.$$

Expression (24) thus satisfies the dynamic programming equation (25). This solution is unique, so (24) is indeed verified as the minimal cost.

There are then at least three ways of determining $\boldsymbol{Q}$: from (19), by following through the deterministic solution (as in the example we worked through), or by appeal to (26). Formula (19) certainly has the advantage of explicitness and elegance and, in fact, provided the most economical calculation in all the cases tested.

## 5. The undiscounted tax problem in stochastic form

We return to the Markov formulation considered in Sections 1–3. The dynamic programming equation for the average cost $\gamma$ and the differential cost $g(\boldsymbol{n}) = G(\boldsymbol{n}) - G(\boldsymbol{0})$ takes the form

$$\gamma = \min_j (\boldsymbol{c}^\top \boldsymbol{n} + \Lambda_j g(\boldsymbol{n}) + \Lambda_0 g(\boldsymbol{n})), \tag{28}$$

where the $\Lambda$ operators now revert to their original discrete-variable definitions (2). The differential cost $g(\boldsymbol{n})$ can be evaluated as follows:

$$g(\boldsymbol{n}) = \lim_{\alpha \downarrow 0}(G(\boldsymbol{n}, 0) - G(\boldsymbol{0}, 0)) = -\sum_j \left(\theta_j(0) + \frac{Q_{jj}}{2}\right) n_j + \frac{1}{2}\boldsymbol{n}^\top \boldsymbol{Q}\boldsymbol{n},$$

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

where $G(\boldsymbol{n}, 0)$ is defined according to (18). The coefficients $\theta_j(0)$ have not yet been determined; the following theorem gives us a direct way of doing so.

**Theorem 3.** *Suppose that the parameters $\lambda$ and $\mu$ satisfy the stability condition (21). Then (28) has the solution*

$$g(\boldsymbol{n}) = \boldsymbol{s}^\top \boldsymbol{n} + \frac{1}{2}\boldsymbol{n}^\top \boldsymbol{Q}\boldsymbol{n}, \qquad \gamma = \sum_k \lambda_k \left(\frac{Q_{kk}}{2} + s_k\right), \tag{29}$$

*where $\boldsymbol{Q}$ is again the matrix defined by (19) and $\boldsymbol{s}$ is the solution of the equation system*

$$\sum_k \mu_{jk}(s_k - s_j) + \frac{1}{2}\sum_k \mu_{jk}(Q_{jj} - 2Q_{jk} + Q_{kk}) = 0, \qquad 1 \le j \le J. \tag{30}$$

*Proof.* Again, the index rule is optimal. If we try solution (29) in (28), we find that it holds as far as the terms linear in $\boldsymbol{n}$ are concerned, by the calculations of Theorem 2. The second relation of (29) then follows by equating the constant terms in the case $\boldsymbol{n} = \boldsymbol{0}$. Relation (30) follows similarly from the cases in which $j$ is the state of maximum index represented in $\boldsymbol{n}$.

Equations (30) should be supplemented by $s_0 = 0$, so that if passage into the terminal state 0 is certain (as it must be for finiteness of costs), then the matrix $(\mu_{jk})$ is substochastic and the equation system nonsingular. The system has an immediate solution in those cases in which there is only one route from any given entry state to the terminal state. For the two-state example considered above, we find that

$$\gamma = \lambda(Q_{11} + Q_{22} - Q_{12}) = \frac{\lambda c_1}{\mu_1} + \frac{\lambda c_2(\lambda \mu_1^{-2} + \mu_2^{-1})}{1 - \lambda T_2}.$$

The referee, to whom the author is grateful, observed that relations (29) imply the identity

$$\gamma = \sum_k \lambda_k g(\boldsymbol{e}_k). \tag{31}$$

This is indeed so, and we can see the validity of (31) on general grounds. Since $g(\boldsymbol{0}) = 0$ under our normalization, we can interpret $g(\boldsymbol{n})$ as the differential cost of passage from state $\boldsymbol{n}$ to state $\boldsymbol{0}$. We then have the relationship

$$\gamma = \frac{0 + \sum_k (\lambda_k/\lambda)(g(\boldsymbol{e}_k) + \gamma\tau_k)}{1/\lambda + \sum_k (\lambda_k/\lambda)\tau_k}, \tag{32}$$

expressing $\gamma$ as the average cost over a recurrence cycle to emptiness. Here $\lambda = \sum_k \lambda_k$ and $\tau_k$ is the expected time needed to pass from $\boldsymbol{n} = \boldsymbol{e}_k$ to $\boldsymbol{n} = \boldsymbol{0}$, i.e. the expected length of the busy period if we start with a single item that is in state $k$. The system will stay empty for an expected time $1/\lambda$, incurring zero cost, and then move to state $\boldsymbol{n} = \boldsymbol{e}_k$ with probability $\lambda_k/\lambda$, so incurring a further expected cost of $g(\boldsymbol{e}_k) + \gamma\tau_k$ over an expected time of $\tau_k$, before returning to the empty state. Relation (32) immediately implies (31).

## 6. Final observations

The well-known fact that the multi-armed bandit problem and the tax problem differ fundamentally in the undiscounted limit has been confirmed very clearly. The multi-armed bandit

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

can shed the insufficiently profitable part of its load, has a differential reward linear in $\boldsymbol{n}$, and a maximal average reward which is quantized to take one of the index values $\nu_s$. The system of the tax problem must accept all load and so has a lower bound on capacity, has a differential cost quadratic in $\boldsymbol{n}$, and has a minimal average cost which could take any value.

The analysis of this paper could be generalized to allow nonexponential service times (by allowing passage through Markov substates during a given processing operation) and all the variants of preemptive and nonpreemptive service. However, what is most important is to see the essential structure; an extension that would be of greater practical interest would be to represent the effects of fixed processing capacity by setting constraints on the distribution of effort over the system. Our analysis in fact gives an indication of how the optimal scheduling should be modified in such a case.

Let us write the dynamic programming equation (28) in the form

$$\gamma = \min_j \Omega_j g(\boldsymbol{n}). \tag{33}$$

This is the equation appropriate to the case in which there is a single operator choosing a single action, with a free choice of actions (i.e. a free choice of which item to process). However, now consider the more realistic case in which processing is divided between workstations, indexed by $h$. Then each workstation can handle only items that are ready for its particular process, which for station $h$ corresponds to items (states) $j$ lying in a set $\mathcal{P}_h$, say. We shall suppose, for simplicity, that the workstations cannot substitute for each other, so that the sets $\mathcal{P}_h$ are disjoint and their union covers all processes.

The processing effort available must also be divided: we shall suppose that workstation $h$ disposes of a proportion $q_h$ of the total unit processing effort available. We further suppose that the workstations have been balanced, in that $q_h$ equals the proportion of effort spent on items in $\mathcal{P}_h$ under a free optimization. On the other hand, every workstation now has a decision to make, namely which of the available items in its work class to process next. Relation (33) must then be modified to

$$\gamma = \sum_h q_h \min_{j \in \mathcal{P}_h} \Omega_j g(\boldsymbol{n}). \tag{34}$$

We might expect that each workstation should concentrate its effort on one of the items of highest index in its buffer (queue). However, to take this view is to ignore the effect of the choice on the progress of items of higher index.

Consider the effect of the operator $\Omega_j$ on the differential cost $g(\boldsymbol{n})$ for the freely optimized policy. We find that

$$\Omega_j g(\boldsymbol{n}) = \sum_i \rho_{ij} n_i, \tag{35}$$

where $\rho_{ij}$ is the expression on the left-hand side of (26). We know this to be zero for $i \geq j$. By the argument used in proving this fact, we find that $\rho_{ij} = \int_{\nu_j}^{\nu_i} (L_j f - 1) f_i \, \mathrm{d}\nu$, $i < j$, where $L_j f$ and $f_i$ are the quantities defined immediately before (27). This expression for $\rho_{ij}$ is positive, since both factors in the integrand are negative in the interval indicated. We thus see that at workstation $h$, the item to be processed should be one whose state $j$ minimizes the linear criterion function

$$\Omega_j g(\boldsymbol{n}) = \sum_{i<j} \rho_{ij} n_i, \tag{36}$$

subject to there being such an item ($n_j > 0$, $j \in \mathcal{P}_h$) that is also ready for $h$-processing. Therefore, our concern is with the numbers of items present in the system that are of index

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press

higher than that now to be processed. Criterion (36) does reflect interference, in that it is concerned with actual numbers rather than just the simple presence or absence of an item of designated index. Its linear form makes its implementation relatively easy. The coefficient $\rho_{ij}$ is an asymmetric interference measure – a measure of the degree to which the choice of $j$ in $\mathcal{P}_h$ affects the progress of those items whose state $i$ gives them a higher priority.

Criterion (36) defines the policy that should be adopted if we have to operate under the workstation constraints for an instant before reverting to the optimal operation of the unconstrained case. It is plausible that this represents at least the direction in which the policy should develop if we have to operate under the workstation constraints indefinitely, but this is a matter yet to be investigated.

A paper that is relevant in the present context is that by Ansell *et al.* (2003). These authors considered what we might term a one-stage system, in which jobs of varying natures entering the system are routed to one of a number of workstations, at which their processing is completed. These workstations differ in their competence to process a given job, but there is a degree of mutual substitutability. The question is then one of choosing a workstation for a job rather than of a workstation choosing a job from its buffer. The authors sought a dynamic routeing rule that minimizes average cost, taking into account job priorities, the appropriateness of workstations for a given job, and also their current degree of congestion. The problems considered in this paper and that of Ansell *et al.* (2003) are respectively approached from opposite directions, in a sense, but there are parallels to be drawn between the results.

In this paper, we have determined the optimal policy in the ideal situation, when effort can be switched instantaneously to any part of the system. We then sought to adapt this policy 'downwards' to the case in which each workstation has a restricted role and effort can only be switched within workstations. In Ansell *et al.* (2003), the authors adapted 'upwards', in that they took the optimal state-independent policy (that which splits the job stream in fixed proportions) and then induced a dynamic response to the current congestion state by introducing a stage of policy improvement. This analysis led to the determination of an analogue of $g(\boldsymbol{n})$ and to an 'allocation' index similar to (35), in that it is linear in the queue sizes at the stations. The models and approaches of the two papers are too different for there to be complete agreement: Ansell *et al.*'s (2003) analogue of the matrix $(\rho_{ij})$ is symmetric and the allocation index contains a term independent of $\boldsymbol{n}$. Nevertheless, a similar pattern emerges from the two analyses.

## References

Ansell, P. S., Glazebrook, K. D. and Kirkbride, C. (2003). Generalised 'join the shortest queue' policies for the dynamic routing of jobs to multi-class queues. *J. Operat. Res. Soc.* **54,** 379–389.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *J. R. Statist. Soc. B* **41,** 148–177.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices.* John Wiley, Chichester.

Klimov, G. P. (1974). Time-sharing service systems. I. *Theory Prob. Appl.* **19,** 532–551.

Klimov, G. P. (1978). Time-sharing service systems. II. *Theory Prob. Appl.* **23,** 314–321.

Lai, T. L. and Ying, Z. (1988). Open bandit processes and optimal scheduling of neural networks. *Adv. Appl. Prob.* **20,** 447–472.

Varaiya, P., Walrand, J. C. and Buyukkoc, C. (1985). Extensions of the multi-armed bandit problem: the discounted case. *IEEE Trans. Automatic Control* **30,** 426–439.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. R. Statist. Soc. B* **42,** 143–149.

Whittle, P. (1981). Arm-acquiring bandits. *Ann. Prob.* **9,** 284–292.

https://doi.org/10.1239/jap/1127322025 Published online by Cambridge University Press