J. Appl. Prob. **42,** 15–26 (2005)
Printed in Israel

MULTI-ACTOR MARKOV DECISION PROCESSES

HYUN-SOO AHN,* *University of Michigan*
RHONDA RIGHTER,** *University of California, Berkeley*

Abstract

We give a very general reformulation of multi-actor Markov decision processes and show that there is a tendency for the actors to take the same action whenever possible. This considerably reduces the complexity of the problem, either facilitating numerical computation of the optimal policy or providing a basis for a heuristic.

Keywords: Markov decision process; multiarmed bandit; flexible server

2000 Mathematics Subject Classification: Primary 90C40
Secondary 90B22

1. Introduction

There have been many nice results establishing the optimality of index rules for classes of Markov decision processes with single actors. These include the traditional multiarmed bandit [3], and scheduling in networks of queues with a single server [4], [6], [7]. When there are multiple actors (players or servers), the problems become much more complicated, and simple index rules are generally no longer optimal. We give a very general reformulation of multi-actor Markov decision processes and give conditions under which there will be a tendency for the actors to take the same action, whenever possible, and for priority to be given to faster actors. This considerably reduces the complexity of the problem, either facilitating numerical computation of the optimal policy or providing a basis for a heuristic.

Our framework is very general. Since a simple index rule is no longer optimal, we can relax many of the assumptions required to obtain such a rule in previous work for single actors. We permit general, exogenous, random effects on the system, actors with different speeds, arbitrary constraints on which actors can take which actions, and all of these may be state dependent. Our model includes quite general queues with multiple servers, multiarmed bandits with multiple players, and data-flow models in which tokens (actors) can enable certain firings (state changes). We are also able to show that our structural results hold for stochastic optimality as long as such optimality is achievable. (By stochastic optimality we mean maximization of the net benefit in the stochastic sense, rather than just maximization of the mean net benefit.) We also give conditions under which the optimal policy can be implemented with distributed control. That is, each actor can choose its own action to maximize its own marginal return.

Many results in the literature follow from ours. Ahn *et al*. [1] considered a two-station queueing model with two flexible workers, Poisson arrivals, exponential service times, holding costs, and preemption permitted. Thus, there are two actors and two actions. They showed that,

Received 7 April 2004; revision received 22 July 2004.
* Postal address: Operations and Management Science, University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109-1234, USA. Email address: hsahn@umich.edu
** Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA. Email address: rrighter@ieor.berkeley.edu

in states where both workers can be assigned to either station, assigning both to the same station is always optimal, which follows from Corollary 3.1(i), below. They also showed that, when servers can collaborate, they always work at the same station, which follows from Corollary 3.3. Kaufman *et al*. [5] considered a two-station collaborative-service tandem queueing model in which workers may come (i.e. are hired) and go (i.e. quit or are fired), meaning that the rate of service changes over time. They showed that, if all servers are identical, it is optimal to allocate all available servers to one queue, and characterized the condition under which the allocation of servers to the chosen queue is optimal. These results can be shown to be a consequence of Corollary 3.3. Weiss and Pinedo [13] considered preemptive scheduling of jobs on parallel processors such that the processing time of a job on a processor is exponential, with a rate that is the product of the job and processor rates. They showed that processing the fastest job on the fastest processor, etc., minimizes the mean flowtime (the total time jobs spend in the system), while processing the slowest job on the fastest processor, etc., minimizes the mean makespan (the time taken to process all the jobs). This follows from our Corollary 3.2.

Our model also includes scheduling of project activities with arbitrary precedences, a finite number of resources, and technological constraints on which particular resources can be used for which activities – see Vairaktarakis [12] for a recent deterministic example.

2. Formulation

We consider a general Markov decision process on a countable state space $\mathcal{S}$, where state transitions occur after exponentially distributed times. To make things concrete, we relate the general framework to a multiplayer, multiarmed bandit. For the multiarmed bandit, the state might include the individual states of all arms present (let us call them 'arm states', to distinguish them from the overall state of the system), as well as environmental states and actor states. Let us fix an arbitrary state $s \in \mathcal{S}$. While the state is s, a cost at rate $g(s)$ is incurred. There are a finite number, N, of actors (players) and a countable set, $\mathcal{K}$, of possible actions (arms to play). For each actor i, there is a set $A_i(s) \subseteq \mathcal{K}$ of admissible actions, which permits us to model multiple skill sets. For each action $k \in \mathcal{K}$, there is a permissible number of actors, $M_k(s) \leq N$, that can be assigned to that action. For example, if the state space of the multiarmed bandit is such that it includes the number of arms in a particular arm state, then the number of players of those arms can be no more than the number of arms.

The rate at which the process changes state depends on the assignment of actors to actions in the following way. Actor i has a 'firing rate' $\mu_i(s)$, bounded by a finite μ_i, so that $\mu_i(s) \leq \mu_i$ for all $s \in \mathcal{S}$, and action $k \in \mathcal{K}$ has a firing rate $\nu_k(s) \leq \nu$ for some finite ν. If actor i is assigned to (admissible) action k, then the action will cause a state change with rate $\mu_i(s)\nu_k(s)$. Another interpretation of the firing rates is that $\mu_i(s)$ is the speed of actor i in state s, and $1/\nu_k(s)$ is the nominal mean time between transitions caused by action k in state s, i.e. the mean transition time when the actor has speed 1. It will be convenient to use the following equivalent interpretation. We will say that actor i fires in state s at rate $\mu_i(s)\nu$: if action k is chosen, this firing will cause a state transition with probability $\nu_k(s)/\nu$, while, otherwise, there is no transition. Later, we will generalize our model to permit more general firing rates.

If actor i is assigned to action k in state s, and this causes a state change, then a random reward $R_k(s)$ is earned (where $R_k(s) \leq R$ for some finite R, and may be negative) and the corresponding new state will be $S_k(s)$, chosen according to transition probabilities that depend only on s and k and not on the actions assigned to the other actors. Later, when we permit more general firing rates, the rate of the state change may depend on the actions of other actors. Our results also hold if the reward depends on the new state, or on both the original and the new state, $R_k(s, S_k(s))$, but we suppress this extra notation.

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

There may also be a state change to some state $\tilde{S}(s)$, when the current state is s, due to exogenous factors that are independent of the actions chosen. These occur at rate $\gamma(s)$, with $\gamma(s) \leq \gamma$ for all $s \in \mathcal{S}$, and with transition probabilities that depend only on the current state. Note that the state may include information about the actors, the actions, and environmental factors, as well as about the internal configuration of the system. For example, actor i may be unavailable in state s, in which case $A_i(s) = \varnothing$ or, equivalently, $\mu_i(s) = 0$. In the multiarmed bandit, arms may arrive or leave, arms that are not played may also change state, players may take a break, the speeds and skill sets of players may otherwise change, etc.

The traditional single-player multiarmed bandit is assumed to progress in discrete time but, with our uniformization, the continuous-time and discrete-time formulations are equivalent for one player. Also, in the traditional bandit problem, in showing the optimality of an index policy, exogenous state transitions are not permitted. Our results permit a much more general model but only provide a partial characterization of the optimal policy.

Our Markov decision process formulation includes very general queueing networks, with general routings (possibly with forks and joins), with servers of different speeds and availabilities that are trained to serve particular subsets of queues, with multiple types of customer, etc. The restriction on the number of actors that can perform an action can be used in a queueing network to ensure that servers cannot serve more customers than are present in a given queue.

We now summarize our notation (much of which will be introduced later). For actor i and state s,

- $A_i(s)$ are the admissible actions;
- $\mu_i(s)$ is the firing rate;
- $a_i(s)$ is the action chosen; and
- $I_i(s) = \mathbf{1}\{\text{actor } i \text{ fires}\}$,

where $\mathbf{1}\{\cdot\}$ is the indicator function of the event $\{\cdot\}$.

For action k and state s,

- $M_k(s)$ is the maximal number of actors that can be assigned to action k;
- $\nu_k(s)$ is the firing rate if action k is chosen;
- $R_k(s)$ is the reward if action k is chosen, fires, and causes the state to change;
- $S_k(s)$ is the new state if action k is chosen, fires, and causes the state to change;
- $J_k(s) = \mathbf{1}\{\text{action } k \text{ fires and causes the state to change} \mid k \text{ is chosen}\}$;
- $m_k^t(s) = J_k(s)[R_k(s) + V_t(S_k(s)) - V_t(s)]$ is the marginal value of choosing action k in state s at time t; and
- $c_k(s)$ is the number of available actors that can be assigned to action k.

For exogenous factors, for state s,

- $\gamma(s)$ is the rate of state change due to exogenous factors;
- $\tilde{S}(s)$ is the new state, given a state change due to an exogenous factor; and
- $\tilde{I}(s) = \mathbf{1}\{\text{the state changes due to an exogenous factor}\}$.

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

Other definitions for state s are as follows:

- $g(s)$ is the cost rate;
- $G(s)$ is the total cost between transitions;
- $A(s)$ is the set of admissible actions;
- $V_t^f(s)$ is the total net benefit under policy f for the next t decision epochs, starting in state s;
- $V_t(s)$ is the total net benefit under the optimal policy for the next t decision epochs, starting in state s;
- $H_t(s) = \tilde{I}(s)V_t(\tilde{S}(s)) + [1 - \tilde{I}(s)]V_t(s) - G(s)$;
- $N(s)$ is the number of available actors; and
- $K(s)$ is number of admissible actions.

3. Results

We use uniformization and assume, without loss of generality, that the total rate out of any state is $\sum_{i=1}^{N} \mu_i \nu + \gamma = 1$. Thus, we have dummy transitions from state s at rate $\beta(s) = 1 - \sum_{i=1}^{N} \mu_i(s)\nu + \gamma(s)$; these transitions cause the state to remain in state s. Note that we have already modeled a subset of the dummy transitions, at rate $\sum_{i=1}^{N} \mu_i(s)(\nu - \nu_k(s))$, with our reinterpretation of firing rates. We assume that there is a finite horizon, t, for the number of remaining decision epochs, where decision epochs occur at state transitions (including dummy transitions) and, thus, at rate 1. We define the current time, i.e. the time of the first decision epoch, to be time 0. Let $G(s)$ be the total (random) cost incurred between transitions when the state is s. Thus $G(s) = g(s)X$, where X is exponentially distributed with rate 1, i.e. $G(s)$ is exponentially distributed with rate $1/g(s)$.

For a fixed time t, let $a_i(s) \in \mathcal{K}$ be the action assigned to actor i in state s, and let $A(s)$ be the admissible set of actor–action combinations in state s. That is,

$$A(s) = \Bigg\{ (a_1(s), a_2(s), \ldots, a_N(s)) \colon a_i(s) \in A_i(s), i = 1, \ldots, N; \sum_{i=1}^{N} \mathbf{1}\{a_i(s) = k\} \le M_k(s) \text{ for all } k \in \mathcal{K} \Bigg\}.$$

3.1. Stochastically optimal policies

Let $V_t^f(s)$ be the total net benefit (i.e. rewards minus cost) starting in state s under some policy f for the next t decision epochs, assuming that the problem will stop at $t = 0$. Note that $V_t^f(s)$ is a random variable. For some problems, there may exist a stochastically optimal policy f^*, i.e. such that $V_t^{f^*}(s) \ge_{\text{st}} V_t^f(s)$ for all t and s. For example, consider the standard single-armed bandit problem with deteriorating bandits, in which arms that are pulled move to states with lower immediate returns. If the returns for arms in different states can be stochastically ordered, then it can easily be shown that the myopic policy of pulling the arm with the stochastically largest return is a stochastically optimal policy. Even if a stochastically optimal policy does not exist for a particular problem, there may be situations in which a certain class of policies can be shown to be stochastically better than others – for example, in preemptive scheduling

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

problems, this is often the case for the class of nonidling policies. We first develop our method assuming the existence of stochastically optimal policies, so we work with random variables instead of means, e.g. we use indicators of events rather than the probabilities of those events. This methodology easily extends to optimization of expected values; we discuss this further in Section 3.4.

Let $V_t(s) = V_t^{f^*}(s)$ be the total net benefit for the stochastically optimal policy. For fixed t, let $I_i(s)$ be an indicator for the event that actor i fires; then $\mathrm{P}(I_i(s) = 1) = \mu_i(s)\nu = 1 - \mathrm{P}(I_i(s) = 0)$ and at most one actor can fire at a time. Also, let $J_k(s)$ indicate whether such a firing causes a state transition (i.e. whether the action also fires); then $\mathrm{P}(J_k(s) = 1) = \nu_k(s)/\nu = 1 - \mathrm{P}(J_k(s) = 0)$ and $J_k(s)$ is independent of $I_i(s)$ for all i. Finally, let $\tilde{I}(s)$ be an indicator for a state transition due to an exogenous factor, in which case $\mathrm{P}(\tilde{I}(s) = 1) = \gamma(s) = 1 - \mathrm{P}(\tilde{I}(s) = 0)$ and $\mathrm{P}(\tilde{I}(s) = I_i(s) = 1) = 0$ for all i. Then we have $V_0(s) = 0$ and

$$\begin{aligned}
V_{t+1}(s) &= \max_{(k_1,\dots,k_N)\in A(s)} \Bigg\{ -G(s) + \sum_{i=1}^{N} I_i(s)\{J_{k_i}(s)[R_{k_i}(s) + V_t(S_{k_i}(s))] + (1 - J_{k_i}(s))V_t(s)\} \\
&\qquad + \tilde{I}(s)V_t(\tilde{S}(s)) + \Bigg[1 - \sum_{i=1}^{N} I_i(s) - \tilde{I}(s)\Bigg]V_t(s) \Bigg\} \\
&= \max_{(k_1,\dots,k_N)\in A(s)} \Bigg\{ -G(s) + \sum_{i=1}^{N} I_i(s)J_{k_i}(s)[R_{k_i}(s) + V_t(S_{k_i}(s)) - V_t(s)] \\
&\qquad + \tilde{I}(s)V_t(\tilde{S}(s)) + [1 - \tilde{I}(s)]V_t(s) \Bigg\} \\
&= \max_{(k_1,\dots,k_N)\in A(s)} \sum_{i=1}^{N} I_i(s)m_{k_i}^t(s) + H_t(s),
\end{aligned}$$

where

$$m_k^t(s) := J_k(s)[R_k(s) + V_t(S_k(s)) - V_t(s)]$$

is the marginal value of choosing action k in state s at time t, and

$$H_t(s) := \tilde{I}(s)V_t(\tilde{S}(s)) + [1 - \tilde{I}(s)]V_t(s) - G(s)$$

is independent of the actions chosen. Note that our assumption of the existence of a stochastically optimal policy implies that the random variables $m_k^t(s)$ can be ordered, in the stochastic sense, for all actions $k \in \mathcal{K}$. In general, the values of $m_k^t(s)$ will be difficult to obtain but, with the structural results given below, we can reduce the complexity of the problem significantly.

Also note that $m_k^t(s)$ does not depend on the actor, so we will prefer to assign as many actors as possible to action a rather than action to b when $m_a^t(s) \geq_{\mathrm{st}} m_b^t(s)$. Moreover, from the lemma below, we will also prefer to assign faster actors to action a rather than to action b. Theorem 3.1 then follows.

Lemma 3.1. *Suppose that X and Y are (not necessarily independent) random variables such that $X \geq_{\mathrm{st}} Y$, that I and J are Bernoulli random variables such that $I \geq_{\mathrm{st}} J$ and $\mathrm{P}(I = J = 1) = 0$, and that (X, Y) is independent of (I, J). Then $IX + JY \geq_{\mathrm{st}} JX + IY$.*

Proof. Let U be uniformly distributed on $[0, 1]$ and let $p = \mathrm{P}(I = 1)$ and $q = \mathrm{P}(J = 1) < p$. Also, let $J' = \mathbf{1}\{U \leq q\}$, $\Delta = \mathbf{1}\{q < U \leq p\}$, and $J'' = \mathbf{1}\{p < U \leq p + q\}$. Finally, let

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

$(X', Y') =_{\text{st}} (X, Y)$ be independent of (X, Y) and let both (X, Y) and (X', Y') be independent of U. Then

$$IX + JY =_{\text{st}} J'X + J''Y + \Delta X' \geq_{\text{st}} J'X + J''Y + \Delta Y' =_{\text{st}} JX + IY,$$

as required.

A consequence of Lemma 3.1 is that, if $\mu_i(s) \geq \mu_j(s)$ and $m_a^t(s) \geq_{\text{st}} m_b^t(s)$, then

$$I_i(s)m_a^t(s) + I_j(s)m_b^t(s) \geq_{\text{st}} I_i(s)m_b^t(s) + I_j(s)m_a^t(s).$$

Theorem 3.1. *Suppose that there is a stochastically optimal policy. Let the actors be arbitrarily ordered, let t be the number of remaining decision epochs, and let s be the initial state.*

(i) *Suppose that a, b, and $k_2, \ldots, k_N$ are such that $(a, k_2, \ldots, k_N) \in A(s)$, $(b, k_2, \ldots, k_N) \in A(s)$, and $m_a(s) \geq_{\text{st}} m_b(s)$. Let f and g be the policies that choose actions $(a, k_2, \ldots, k_N)$ and $(b, k_2, \ldots, k_N)$, respectively, at time 0, and which then follow the optimal policy. Then $V_t^f(s) \geq_{\text{st}} V_t^g(s)$ and g cannot be optimal.*

(ii) *Suppose that a, b, and $k_3, \ldots, k_N$ are such that $(a, b, k_3, \ldots, k_N) \in A(s)$, $(b, a, k_3, \ldots, k_N) \in A(s)$, $m_a(s) \geq_{\text{st}} m_b(s)$, and $\mu_1(s) \geq \mu_2(s)$. Let f and g be the policies that choose actions $(a, b, k_3, \ldots, k_N)$ and $(b, a, k_3, \ldots, k_N)$, respectively, at time 0, and which then follow the optimal policy. Then $V_t^f(s) \geq_{\text{st}} V_t^g(s)$ and g cannot be optimal.*

We say that an actor i is available in state s if $A_i(s) \neq \varnothing$ and $\mu_i(s) > 0$. Let $N(s)$ be the number of available actors in state s, and let $c_k(s)$ be the number of available actors for which action k is admissible in state s, i.e. $c_k(s) = \sum_{i=1}^{N} \mathbf{1}\{k \in A_i(s)\}$. Similarly, we call an action permissible in state s if $M_k(s) > 0$, and let $K(s) \leq \infty$ be the number of permissible actions in s.

Corollary 3.1. *Suppose that there is a stochastically optimal policy. For a given state s and time t, order the permissible actions in decreasing (stochastic) order of $m_k^t(s)$.*

(i) *If $c_k(s) \leq M_k(s)$ for all $k \leq K(s)$, then the optimal policy in state s is 'greedy', that is, it assigns each available actor to the lowest-indexed action it is permitted to.*

(ii) *Suppose that actors have the same set of admissible actions and that, for all $s \in S$, $M_1(s) \geq N(s)$. Then the optimal policy is to assign all actors to action 1 in all states (though the particular action that is referred to as action 1 will depend on the state). In this case, the actors effectively act as a single actor, or team, so any results for single-actor models will also be true for this multi-actor model.*

If there are few actors and many actions k with $M_k(s)$ large, and if the actors have similar admissible actions, then the state space can be considerably reduced using Theorem 3.1 and Corollary 3.1. For example, when there are two actors and they have the same sets of admissible actions, it will not be optimal, for all actions k and l with $M_k(s), M_l(s) \geq 2$, to assign one actor to action k and the other to action l. If $M_k(s) \geq 2$ for all k, then both actors should be assigned to the same action, and the number of possibilities for the optimal-action pair decreases from $K(s)^2$ to $K(s)$.

An application is to an extension of Klimov's model [6], [7] for a single server serving jobs in a queueing network; namely, to allow N fully flexible and failure-prone servers, where all service times, interarrival times, and failure and repair times are exponential and possibly

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

dependent on an exogenous state variable. If the state is such that all queues have either 0 or at least N jobs, then all servers should be assigned to the same queue. More generally, if there are M nonempty queues, if x_i is the number of jobs in the ith queue, and if, without loss of generality, we label the queues – starting with the nonempty ones – so that $x_1 \le x_2 \le \cdots \le x_M$, then the N servers will be assigned to at most $\hat{m}$ queues, where either $\hat{m} = \sum_{i=1}^{M} x_i$, if $\sum_{i=1}^{m} x_i \le N$, or $\hat{m}$ is the smallest m such that $\sum_{i=1}^{m} x_i \ge N$.

It is not hard to prove the following corollary to Theorem 3.1. Note that a special case of the ordering relation for admissible actions is that they are the same for all actors.

Corollary 3.2. *Suppose that there is a stochastically optimal policy. For state s, order the available actors in decreasing order of $\mu_i(s)$ and order the permissible actions in decreasing stochastic order of $m_k^t(s)$. If $A_1(s) \subseteq A_2(s) \subseteq \cdots \subseteq A_{N(s)}(s)$ then the optimal policy is to assign the ith actor to the best (lowest-indexed) remaining action after the first $i-1$ actors have been assigned.*

A consequence of this result is that, when the conditions of the theorem are satisfied, the optimal policy can be implemented in a distributed fashion, letting each actor choose the best action it can (in terms of $m_k(s)$), subject to the constraint that faster actors have higher priority in choosing.

3.2. Collaborative processing

Corollary 3.1 holds for the special case in which there is no constraint on the number of actors that can be assigned to an action, i.e. where $M_k(s) \ge N(s)$ for all $k \in \mathcal{K}$ and $s \in S$. Such a model is stochastically equivalent to a collaborative model, in which multiple actors can work together on the same action, and where firing rates and rewards for the actors working together are added. Indeed, similar reasoning gives us our next corollary.

Corollary 3.3. *Suppose that there is a stochastically optimal policy. For a given state s and time t, order the permissible actions in decreasing stochastic order of $m_k^t(s)$.*

(i) *When actors can collaborate, the optimal policy in state s is to assign each available actor to the lowest-indexed action permitted.*

(ii) *Suppose that a subset of actors have the same set of admissible actions and can collaborate. Then, it is optimal to assign all of the actors in this subset to the same action.*

Note that in case (ii) of Corollary 3.3, under the optimal policy the given subset of actors works as a single (fast) actor, so we can combine the actors into one. In particular, if all of the actors have the same set of admissible actions, the optimal policy is the one that is optimal for a single actor. This permits a slight generalization of the applicability of the Gittins index for optimizing the single-armed bandit problem. That is, for the standard single-armed bandit problem in continuous time, it is optimal always to devote all effort to the arm with the largest Gittins index, even when that effort can be divided among several arms. More generally, case (ii) of Corollary 3.3 provides insight into the optimality of so-called 'bang-bang' policies when processing rates can be chosen from an interval, say $[\mu_0, \mu_1]$, for some action k. With appropriate rescaling, we can think of there being one subset of the actors with total rate μ_0 and with their only admissible action being assignment to action k, and another subset with total rate $\mu_1 - \mu_0$ and with two admissible actions: assignment to k and idling. Then we know that, within each subset, it is optimal for all actors to take the same action, so the optimal service rate will be either μ_0 or μ_1.

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

Another model that is equivalent to the collaborative model in a scheduling context is the following. Consider a grid of parallel computers (actors) with different speeds, and arrivals of tasks of different types. It is permitted to send copies of the same task to more than one computer at a time, with the completion time of the task being the earliest completion time of any of its copies [8]. For exponential processing times, this copy option is equivalent to a collaboration option because the minimum of a set of exponential random variables is an exponential random variable with rate equal to the sum of the individual rates. Thus, it is optimal to process the same task on all of the computers, where the task chosen is the one that is optimal when there is only a single computer, e.g. it is chosen according to the '$c\mu$-rule' for appropriately defined c and μ.

Mandelbaum and Reiman [9] also considered a restricted form of collaboration, or resource pooling, in Jackson queueing networks. They compared steady-state sojourn times for a system with dedicated servers for each node in a network to one in which a single server with combined service rate can serve at all of the queues. However, in their model, when service is pooled there can be no preemption and jobs are processed on a first-come-first-served basis, and so the pooled system operates as an M/PH/1 queue. Under these constraints, the pooled model may have worse performance than the dedicated-server model. In the special case of tandem systems, Mandelbaum and Reiman showed that pooling is always better. This result follows from ours because, assuming that optimal policies are always followed, a system that permits collaboration and preemption, and in which all servers can perform all tasks (call this system 1, say) will perform better than a system that requires collaboration, does not permit preemption, and in which all servers can perform all tasks (the MR pooled system). System 1 will also perform better than a system that does not permit collaboration and in which each server can only perform one task (the MR dedicated-server system). From Corollary 3.3, the optimal policy in system 1 has all servers serving the same task at all times, i.e. acting as a single server. It is easy to show that the optimal policy for a single server when preemption is permitted is to always work on the task that is at the latest node in the tandem system, so preemption does not occur. Hence, for tandem queues, the performance of the MR pooled system is as good as that of system 1 and, hence, is better than that of the MR dedicated-server system. (See also [11], where the optimal, preemption-permitting, collaborative policy for tandem systems was shown to be the 'expedite policy', which, in fact, never preempts.)

3.3. Generalized firing rates

We now permit more generalized firing rates. Assigning a subset of actors, $\Gamma \subseteq \{1, 2, \ldots, N\}$, to action $k \in \mathcal{K}$ in state s will cause a state transition with rate $r_s(\Gamma)\nu_k(s)$, so the firing rate for the set of actors Γ is $r_s(\Gamma)$. In Section 3.1, we assumed that the firing rate was $r_s(\Gamma) = \sum_{i\in\Gamma} \mu_i(s)$. Now we let $I(\Gamma, s)$ be an indicator (analogous to $I_i(s)$) for whether the set of actors Γ fires, so $\mathrm{P}(I(\Gamma, s) = 1) = r_s(\Gamma)\nu = 1 - \mathrm{P}(I(\Gamma, s) = 0)$. Let $\mathcal{G}(s)$ be the set of admissible assignments of actors to actions in state s, i.e.

$$\mathcal{G}(s) = \{\Gamma_k, k \in \mathcal{K} : \Gamma_k \subseteq \{1, 2, \ldots, N\}; |\Gamma_k| \le M_k;\ j \in \Gamma_k \Rightarrow k \in A_j(s);\\ \Gamma_k \cap \Gamma_l = \varnothing \text{ for } l \in \mathcal{K}, l \ne k\}.$$

With our other definitions as before, we have

$$V_{t+1}(s) = \max_{\{\Gamma_k, k\in\mathcal{K}\}\in\mathcal{G}(s)} \sum_{k\in\mathcal{K}} I(\Gamma_k, s) m_k^t(s) + H_t(s). \tag{3.1}$$

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

We have the following variants of Corollary 3.2.

Corollary 3.4. *Suppose that* $r_s(\Gamma) = \rho_s(\sum_{j\in\Gamma}\mu_j(s))$, *where* $\rho_s(x)$ *is an increasing convex function of* x *for all* s, *and suppose that there is a stochastically optimal policy. For state* s *and time* t, *order the available actors in decreasing order of* $\mu_i(s)$ *and order the permissible actions in decreasing stochastic order of* $m_k^t(s)$. *If* $A_i(s) = A(s)$ *for all* i *and* s, *and* $M_k(s) = M(s)$ *for all* s *and all available* k, *then the optimal policy is a greedy policy that assigns the fastest* $\min\{M(s), N(s)\}$ *actors to action 1, the next fastest* $\min\{M(s), N(s) - M(s)\}$ *actors to action 2, etc., until all actors are assigned.*

Proof. Suppose that some policy f does not follow the greedy policy that we claim is optimal, and let $t+1$ be the smallest time to go at which this is true. Suppose that the state at time $t+1$ is s. Let $\{\Gamma_k, k\in\mathcal{K}\}$ be the assignments at time $t+1$ under policy f, let $\{\Gamma_k^*, k\in\mathcal{K}\}$ be the assignments under the greedy policy, and define $x_k = \sum_{j\in\Gamma_k}\mu_j(s)$. If $x_j > x_1$ for some $j > 1$, let $\Gamma_1' = \Gamma_j$, $\Gamma_j' = \Gamma_1$, and $\Gamma_l' = \Gamma_l$ for $l \neq 1, j$, and let f' be the policy that assigns actors at time $t+1$ according to $\{\Gamma_k', k\in\mathcal{K}\}$ and then agrees with f, following the greedy policy. Then

$$\begin{aligned} V_{t+1}^f(s) &= I_1 m_1^t(s) + I_j m_j^t(s) + \sum_{k\in\mathcal{K}, k\neq 1, j} I_k m_k^t(s) + H_t(s) \\ &\leq_{\text{st}} I_j m_1^t(s) + I_1 m_j^t(s) + \sum_{k\in\mathcal{K}, k\neq 1, j} I_k m_k^t(s) + H_t(s) \\ &= V_{t+1}^{f'}(s), \end{aligned}$$

where $I_k = 1$ with probability $\rho_s(x_k)\nu$ (and equals 0 otherwise) with $k\in\mathcal{K}$, and the inequality follows from Lemma 3.1 because $\rho_s(x_j) \geq \rho_s(x_1)$. We can repeat such interchanges until we have a policy f'' such that $x_1'' \geq x_j''$ for all $j > 1$. Now suppose that $|\Gamma_1''| < |\Gamma_1^*|$. Then we can assign another actor to action 1, and similarly show that the new policy has greater net benefit. Now, for a policy f''' such that $x_1''' \geq x_j'''$ for all $j > 1$ and $|\Gamma_1'''| = |\Gamma_1^*|$, if $\Gamma_1''' \neq \Gamma_1^*$ we can interchange actors as we did above and again improve the net benefit. Finally, for a policy f'''' with $\Gamma_1'''' = \Gamma_1^*$, we can think of action 1 as no longer being admissible for the remaining actors, and repeat the argument for action 2, etc. By induction on t, the result follows.

Now suppose that $r_s(\Gamma) = \rho_s(|\Gamma|)$, where $\rho_s(x)$ is increasing and concave in x for all s. That is, actors are indistinguishable in terms of their firing rates. Actors may, however, differ in terms of admissible actions, but we suppose that, for each s, the actors can be ordered so that $A_1(s) \subseteq A_2(s) \subseteq \cdots \subseteq A_{N(s)}(s)$. Then, the optimal policy can be determined by a greedy algorithm, as follows. For state s and time t, order the available actors in increasing order of their admissible sets and order the admissible actions in decreasing stochastic order of $m_k^t(s)$. Actor 1 should be assigned to the lowest-indexed action it can be. Suppose that the first j actors have been assigned, let $n(k)$ be the number of those actors assigned to action k, $k = 1, \ldots, N(s)$, and let $Y \subseteq \{1, 2, \ldots, N(s)\}$ be the set of eligible actions for $j+1$, i.e. $Y = \{k\colon n(k) < M(k) \text{ and } k \in A_{j+1}(s)\}$. Then actor $j+1$ should be assigned to action k if both $k\in Y$ and

$$m_k^t(s)[r_s(n(k)+1) - r_s(n(k))] \geq_{\text{st}} m_l^t(s)[r_s(n(l)+1) - r_s(n(l))]$$

for all $l\in Y$. The algorithm requires $O(N^2)$ computations.

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

Corollary 3.5. *Suppose that $r_s(\Gamma) = \rho_s(|\Gamma|)$, where $\rho_s(x)$ is increasing and concave in x for all s, that the actors can be ordered so that $A_1(s) \subseteq A_2(s) \subseteq \cdots \subseteq A_{N(s)}(s)$ for each s, and that there is a stochastically optimal policy. Then the optimal policy can be determined from the greedy algorithm described above.*

Proof. Suppose that some policy f does not follow the greedy algorithm, and let $t+1$ be the smallest time to go for which this is true. Suppose that the state at time $t+1$ is s. Let $\{\Gamma_k, k \in \mathcal{K}\}$ be the assignments at time $t+1$ under policy f, and let $\{\Gamma_k^*, k \in \mathcal{K}\}$ be the assignments under the greedy policy. Suppose first that, at time $t+1$, f assigns actor 1 to action $k > k_1$, where $k_1 = \min\{l : l \in A_1\}$, so that $1 \in \Gamma_k$. If f assigns some actor $j > 1$ to action k_1, we can interchange the actions for actors 1 and j without affecting the net benefit. If f does not assign an actor to action k_1, let f' assign actor 1 to action k_1 instead of action k, and let it otherwise agree with f. Let l be the number of actors assigned to action k under f', and let $I_i = 1$ with probability $\rho_s(i)\nu$ and $I_i = 0$ otherwise, $i = 0, 1, \ldots N$. Then

$$\begin{aligned} V_{t+1}^f(s) &= I_0 m_{k_1}^t(s) + I_{l+1} m_k^t(s) + \sum_{l \in \mathcal{K}, l \neq k, k_1} I_k m_k^t(s) + H_t(s) \\ &\leq_{\text{st}} I_1 m_{k_1}^t(s) + I_l m_k^t(s) + \sum_{l \in \mathcal{K}, l \neq k, k_1} I_k m_k^t(s) + H_t(s) \\ &= V_{t+1}^{f'}(s), \end{aligned}$$

where the inequality follows from Lemma 3.2, below. That is, the optimal policy must assign actor 1 according to the greedy algorithm. Now suppose, for the purposes of induction, that it is optimal to assign actors 1 through j according to the greedy algorithm. The problem of assigning the remaining actors is as if only these actors are available and only the actions in Y (defined in the algorithm) are admissible, so the argument for assigning actor $j+1$ is the same as the preceding argument for assigning actor 1. The result then follows by induction on t.

Lemma 3.2. *Suppose that X and Y are (not necessarily independent) random variables with $X \geq_{\text{st}} Y$, that I_1, I_2, I_3, and I_4 are Bernoulli random variables with $p_i = \mathrm{P}(I_i = 1)$, $p_1 \leq p_2 \leq p_3 \leq p_4$, $p_4 - p_3 \leq p_2 - p_1$, $p_1 + p_2 + p_3 + p_4 \leq 1$, and $\mathrm{P}(I_1 = I_4 = 1) = \mathrm{P}(I_2 = I_3) = 0$, and that (X, Y) is independent of (I_1, I_2, I_3, I_4). Then*

$$I_2 X + I_3 Y \geq_{\text{st}} I_1 X + I_4 Y.$$

Proof. Let $p_0 = 0$, let $q_i = p_i - p_{i-1}$, $i = 1, 2, 3, 4$, and let $q_5 = q_2 - q_4$. Additionally, let U be uniformly distributed on [0, 1], and let

$$\begin{aligned} J_1 &= \mathbf{1}\{2q_2 < U \leq 2q_2 + q_1\}, \\ J_1' &= \mathbf{1}\{2q_2 + q_1 + q_3 < U \leq 2q_2 + q_1 + q_3 + q_1\}, \\ J_3 &= \mathbf{1}\{2q_2 + q_1 < U \leq 2q_2 + q_1 + q_3\}, \\ J_4 &= \mathbf{1}\{U \leq q_4\}, \\ J_4' &= \mathbf{1}\{q_2 < U \leq q_2 + q_4\}, \\ J_5 &= \mathbf{1}\{q_4 < U \leq q_4 + q_5 = q_2\}, \\ J_5' &= \mathbf{1}\{q_2 + q_4 < U \leq q_2 + q_4 + q_5 = 2q_2\}. \end{aligned}$$

https://doi.org/10.1239/jap/1110381367 Published online by Cambridge University Press

Finally, let $(X', Y') =_{\text{st}} (X, Y)$ be independent of (X, Y) and let both (X, Y) and (X', Y') be independent of U. Then

$$\begin{aligned} I_2X + I_3Y &=_{\text{st}} J_1X + (J_4' + J_5')X' + (J_1' + J_3 + J_4 + J_5)Y \\ &\geq_{\text{st}} J_1X + (J_1' + J_3 + J_4 + J_5)Y + J_4'Y' \\ &=_{\text{st}} I_1X + I_4Y. \end{aligned}$$

Note that, for these generalized firing rates, the optimal policy can again be determined in a distributed fashion. That is, if actors are given priority based on their indices (actor 1 has highest priority, etc.), then each actor should choose an action that maximizes its marginal increase in the value function, given the choices of the higher priority actors.

3.4. Mean optimal policies

If there is not a stochastically optimal policy, all of our results still hold, except with the random variables replaced by their means (e.g. $\mathrm{E}\, V_t$ instead of V_t, ν instead of J, $\mathrm{E}\, m_k^t$, etc.) and our objective is to maximize the expected net benefit. Our results can also be extended to the infinite-horizon problem when the long-run average or discounted expected-net-benefit criterion is considered. Under appropriate conditions, one can show that there exists a solution of the optimality equation for the expected discounted benefit. If the solution exists, the expected benefit-to-go function starting from state s, $\mathrm{E}\, V_t(s)$, is replaced by the limit of the expected discounted benefit-to-go. Sufficient conditions would be, for example, a finite state space or an upper bound on rewards and costs. In the average-benefit case, the value function $\mathrm{E}\, V_t(\cdot)$ is replaced by the long-run average benefit, b^*, and the relative value function, $v(\cdot)$; the optimality equation can be rewritten by substituting $b^* + v(s)$ in place of $\mathrm{E}\, V_t(s)$. Under appropriate conditions (for example, the SEN assumptions of Sennott [10, p. 132]), the solution of the average-benefit optimality equation exists, the long-run average benefit is replaced by a scalar, b (independent of the initial state), and there exists a stationary, deterministic optimal policy. The proofs for the infinite-horizon problems immediately follow after substituting the corresponding value functions for $V_t(s)$.

4. Conclusions

We have given a formulation of multi-actor Markov decision processes that allows us to make general statements about optimal policies. The key assumptions are as follows.

1. Firing rates are multiplicative, so that some actors are uniformly faster or slower than others.

2. State transitions depend on the action chosen, and not on which actor chooses the action.

These assumptions imply a decomposition of the objective function, making it clear that actions should be chosen according to their marginal values, and that faster actors should be assigned to actions with higher marginal values. Such a decomposition will not hold without our key assumptions. For example, Andradóttir *et al.* [2] considered a preemptive tandem system with two stations and two servers that can collaborate on jobs, and in which service times are exponential, with rate μ_{ij} for server i serving a job at station j, such that $\mu_{11}\mu_{22} \geq \mu_{12}\mu_{21}$. In this case, it is better for each server to serve the station at which it is (relatively) more effective. That is, it is optimal to assign server 1 to the first station and server 2 to the second station whenever that assignment is nonidling; otherwise the servers should collaborate on tasks in the nonempty buffer.

Assumption 2 tends to rule out models that do not permit preemption. When preemption is not allowed, the state must describe which actors have been working on which actions, and so state transitions will depend on both the action and the actor.

Acknowledgement

We appreciate the efforts of the referee, which improved the presentation of the paper.

References

[1] Ahn, H.-S., Duenyas, I. and Lewis, M. E. (2002). The optimal control of a two-stage tandem queueing system with flexible servers. *Prob. Eng. Inf. Sci.* **16,** 453–469.
[2] Andradóttir, S., Ayhan, H. and Down, D. G. (2001). Server assignment policies for maximizing the steady-state throughput. *Manag. Sci.* **47,** 1421–1439.
[3] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *J. R. Statist. Soc. B* **14,** 148–177.
[4] Harrison, J. M. (1975). Dynamic scheduling of a multiclass queue: discount optimality. *Operat. Res.* **23,** 270–282.
[5] Kaufman, D., Ahn, H.-S. and Lewis, M. E. (2004). On the introduction of agile, temporary workers into a tandem queueing system. Work in progress.
[6] Klimov, G. P. (1974). Time-sharing service systems. I. *J. Theory Prob. Appl.* **19,** 532–551.
[7] Klimov, G. P. (1978). Time-sharing service systems. II. *J. Theory Prob. Appl.* **23,** 314–321.
[8] Koole, K. and Righter, R. (2004). Resource allocation in grid computing. Work in progress.
[9] Mandelbaum, A. and Reiman, M. I. (1998). On pooling in queueing networks. *Manag. Sci.* **44,** 971–981.
[10] Sennott, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems.* John Wiley, New York.
[11] Van Oyen, M. P., Gel, E. G. S. and Hopp, W. J. (2001). Performance opportunity for workforce agility in collaborative and noncollaborative work systems. *IIE Trans*. **33,** 761–777.
[12] Vairaktarakis, G. L. (2003). The value of resource flexibility in the resource-constrained job assignment problem. *Manag. Sci.* **49,** 718–732.
[13] Weiss, G. and Pinedo, M. (1980). Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J. Appl. Prob.* **17,** 187–202.