Optimal Control in Light Traffic Markov Decision Processes

Ger Koole Olaf Passchier

ZOR - Mathematical Methods of Operations Research 45:63-79, 1997

Abstract

We consider Markov Decision Processes under light traffic conditions. We develop an algorithm to obtain asymptotically optimal policies for both the total discounted and the average reward criterion. This gives a general framework for several light traffic results in the literature. We illustrate the method by deriving the asymptotically optimal control of a simple ATM network.

1 Introduction

In this paper we consider Markov decision processes (MDPs) with as objective the computation of optimal decision rules. To do this in general a system of N equations in N unknowns has to be solved, where N is the number of states. In real problems this N is usually very big making this computation very complex. We focus on a class of MDPs with a special structure such that we can derive this solution in a recursive way. We develop an algorithm to compute optimal policies under light traffic, which means that there is a parameter of the system that tends to zero, for total discounted and average costs. There are two reasons for the use of such an algorithm: 1) With a computer the asymptotically optimal policy can be computed very fast (relatively to the regular optimal policy). 2) Structural properties of the asymptotically optimal policy can often be obtained easily.

Smith [11] considers a repairman model and expresses the average time the system is down as a function of some parameter. He computes an asymptotical optimal policy using a power series expansion of the stationary probabilities when this parameter is in a neighbourhood of zero, corresponding to cases of highly reliable and highly unreliable systems. This method is related to the power series algorithm (see [1]). Katehakis and Levine [4] are the first who use the optimality equation to obtain an asymptotically optimal policy. They look at a model where distinguishable servers have to be assigned to different types of jobs. The same method is used in Katehakis and Derman [3], Katehakis and Levine [5], Levine and Finkel [9] and Koole and Vrijenhoek [8]. These references only consider specific applications of the algorithm, the aim of this paper is to develop assumptions under which the method works, to derive an algorithm which can easily be applied and to illustrate the algorithm with an example. Note that all models studied in the papers mentioned above fit in our framework.

Model: The state space E and the action sets in all the states i, $A(i)$ are finite. We denote the traffic intensity with ρ. When action a is chosen in state i there is a transition to state j with probability $p(i,a,j)(\rho)$ and a direct cost $c(i,a)(\rho)$. We assume that there exists a $\rho^* > 0$ such that $p(i,a,\bullet)(\rho)$ is a probability distribution for all i,a and $\rho \in [0,\rho^*]$.

Because we have finite state and action sets it is known (Derman [2]) that for discounted as well as for average cost for every ρ there is at least one minimising policy that is stationary

and deterministic. Such a policy f is one which, whenever the system is in state $i \in E$, selects an action $f(i) \in A(i)$ as a deterministic function of i only. From now on we restrict ourselves to this class of policies, which we denote by F. Our main condition is:

Condition 1 *We can classify the states in levels* $0, 1, ..., M$ *such that*

i) The transitions are of the form

$$p(i, a, j)(\rho) = q(i, a, j)\rho^{(l(j)-l(i))^+}$$

where $l(i)$ *denotes the level of state* i *and* $q(i, a, j)$ *is independent of* ρ.

ii) We can order the states in each level in such a way that there are no transitions to higher ordered states in the same level.

iii) There are integers s_0 *and* s_1 *such that the costs are of the form*

$$c(i, a)(\rho) = \sum_{s=s_0}^{s_1} c^{(s)}(i, a)\rho^s$$

with $c^{(s)}(i, a)$ *independent of* ρ.

Remark: Similar assumptions can be found in [6]. There the parameter can be chosen completely arbitrary, and taken equal to 1 afterwards. Here however it is not just an artificial paramater but it is the case where ρ approaches 0 where are interested in. Thus taking the limit to zero of this parameter should mean something, for example high or low traffic or a very reliable or very unreliable system. Therefore we call ρ the traffic intensity.

For a policy $f \in F$ we define the direct cost vector c_f as $c_f(i) = c(i, f(i))$ and $c_f^{(s)}(i) = c^{(s)}(i, f(i))$. The transition matrix P_f has as entry in row i and column j the number $p_{ij}^f = p(i, f(i), j)$ for $i, j \in E$, and similarly $q_{ij}^f = q(i, f(i), j)$. Under condition 1 we can write $P_f = G_f^{(0)} + \rho G_f^{(1)} + \cdots + \rho^M G_f^{(M)}$ with $G_f^{(0)}$ a probability matrix and all rowsums of $G_f^{(s)}$ equal to zero for $0 < s \leq M$.

2 Discounted costs

Let us fix ρ for the moment. For a policy $f \in F$, initial state $i \in E$ and discount factor $\beta \in [0, 1)$ we denote the total discounted expected costs by $v_f^\beta(i)$. We know that the vector v_f^β solves the following system of equations:

$$v_f^\beta = c_f + \beta P_f v_f^\beta. \tag{1}$$

The discounted value of the process is denoted by v^β and defined as $v^\beta = \min_{f \in F} v_f^\beta$. We call a policy f^* optimal if $v_{f^*}^\beta = v^\beta$. This value vector solves the Discounted Optimality Equation, i.e.

$$v^\beta = \min_{f \in F}\{c_f + \beta P_f v^\beta\}. \tag{2}$$

As v_f^β and v^β depend on ρ, we denote them by $v_f^\beta(\rho)$ and $v^\beta(\rho)$ in the sequel.

Definition 2.1 *A policy* f^* *is called asymptotically optimal if there exists a* $\rho^* \in (0, 1)$ *such that for all* $\rho \in (0, \rho^*)$*:* $v^\beta(\rho) = v_{f^*}^\beta(\rho)$.

Theorem 1 *Fix β. Under condition 1 there exists a β-discounted asymptotically optimal policy f^*. We can compute f^* and its corresponding costs in a recursive way.*

Before we give the proof we formulate and prove some lemmas.

Lemma 2.1 *Fix a policy $f \in F$ and a discount factor β. We can express $v_f^\beta(\rho)$ as a power series expansion in ρ:*

$$v_f^\beta(\rho) = \sum_{t=s_0}^{\infty} \rho^t v_f^{\beta,(t)}.$$

Proof: From (1) it follows that

$$v_f^\beta(\rho) = \{I - \beta P_f(\rho)\}^{-1} c_f(\rho)$$

This matrix inversion can be done by Cramer's rule and hence the value can be written as the quotient of two polynomials in ρ with finite degree. It follows that the number of poles is also finite and that $\rho = 0$ cannot be a pole because $P_f(0) = G_f^{(0)}$ is a probability matrix. Hence, there exists a positive r_f such that $v_f^\beta(\rho)$ is analytic in a disc around $\rho = 0$ with radius r_f. Therefore we can express it as a powers series expansion at $\rho = 0$ that converges for $\rho < r_f$. If we substitute this expansion in (1) we obtain:

$$\sum_{t=0}^{\infty} \rho^t v_f^{\beta,(t)} = \sum_{t=s_0}^{s_1} \rho^t c_f^{(t)} + \beta \sum_{m=0}^{M} \rho^m G_f^{(m)} \sum_{t=0}^{\infty} \rho^t v_f^{\beta,(t)}$$

$$\sum_{t=0}^{\infty} \rho^t v_f^{\beta,(t)} = \sum_{t=s_0}^{s_1} \rho^t c_f^{(t)} + \beta \sum_{t=0}^{\infty} \rho^t \sum_{s=0}^{t} v_f^{\beta,(s)} G_f^{(t-s)}$$

Equating successive coefficients gives, if $s_0 > 0$:

$$\begin{cases} v_f^{\beta,(0)} &= v_f^{\beta,(0)} G_f^{(0)} \\ v_f^{\beta,(1)} &= v_f^{\beta,(0)} G_f^{(1)} + v_f^{\beta,(1)} G_f^{(0)} \\ & \vdots \\ v_f^{\beta,(s_0-1)} &= \sum_{s=0}^{s_0-1} v_f^{\beta,(s)} G_f^{(s_0-1-s)}, \end{cases}$$

and it is easy to see that $v_f^{\beta,(t)} = 0$ for $t = 0, \ldots, s_0 - 1$. □

Lemma 2.2 *For ρ small enough the power series expansion of the discounted value in ρ converges for every policy.*

Proof: The set F is finite and therefore if we take $\rho_1 = \min_{f \in F} r_f$ all the power series converge for $\rho \in (0, r)$. □

Lemma 2.3 *There exists ρ_2 and a strategy $f^* \in F$ such that f^* is optimal for $\rho \in (0, \rho_2)$.*

Proof: If there is no asymptotical optimal strategy f^*, there is a state x such that there exists a sequence of ρ that tends to zero and the corresponding sequence of optimal strategies has two limit points. So there exists a decreasing sequence $\{\rho_i\}_{i=1}^{\infty}$ and two policies f_1 and f_2 such that $v_{f_1}^\beta(x, \rho_{2i}) < v_{f_2}^\beta(x, \rho_{2i})$ and $v_{f_1}^\beta(x, \rho_{2i+1}) \geq v_{f_2}^\beta(x, \rho_{2i+1})$. From the continuity of v_f^β it follows that $v_{f_1}^\beta(x, \rho) - v_{f_2}^\beta(x, \rho)$ is equal to zero a infinite number of times. Fix

a state x, two policies f_1 and f_2 and discount factor β. By Cramer's rule we can express $v_f^\beta(x,\rho)$ as a quotient of two polynomials in ρ. So

$$v_{f_1}^\beta(x,\rho) - v_{f_2}^\beta(x,\rho) = \frac{\gamma(\beta,f_1,x,\rho)}{\delta(\beta,f_1,x,\rho)} - \frac{\gamma(\beta,f_2,x,\rho)}{\delta(\beta,f_2,x,\rho)}.$$

From the argument above we can conclude that the polynomial $\gamma(\beta,f_1,x,\rho)\delta(\beta,f_2,x,\rho) - \gamma(\beta,f_2,x,\rho)\delta(\beta,f_1,x,\rho)$ is equal to zero an infinite number of times, but not everywhere, which cannot be true. □

For the computation of these series it is convenient to introduce the following definitions.

Definition 2.2

$$\begin{aligned} L(i) &= \{j \in E : j < i\} \\ U(i) &= \{j \in E : j > i\} \\ \mathcal{L}(m) &= \{i \in E : l(i) = m\} \end{aligned}$$

Proof of Theorem 1: From the lemma's 2.2 and 2.3 it is clear that if we define $\rho^* = \min\{\rho_1,\rho_2\}$ the value can be expressed as a power series extension on $\rho \in (0,\rho^*)$. The only thing that still has to be proven is that we can do the computations in a recursive way.

If we put the transformed transitions and costs in (2) we obtain the following set of equations:

$$\begin{aligned} \sum_{t=s_0}^{\infty} \rho^t v_\beta^{(t)}(i) &= \min_{a\in A(i)}\Big[\sum_{s=s_0}^{s_1} \rho^s c^{(s)}(i,a) + \beta \sum_{t=s}^{\infty} \rho^t v_\beta^{(t)}(i) \\ &\quad + \beta \sum_{j\neq i} \rho^{(l(j)-l(i))^+} q(i,a,j)\Big\{\sum_{t=s}^{\infty} \rho^t v_\beta^{(t)}(j) - \sum_{t=s}^{\infty} \rho^t v_\beta^{(t)}(i)\Big\}\Big] \end{aligned}$$

If ρ is small enough then this minimising can also be done by first minimising the term with ρ^{s_0}, next minimising the term with ρ^{s_0+1} and so on. We derive the following minimisation problem, with $t > s_0$:

$$\begin{aligned} v_\beta^{(s_0)}(i) &= \min_{a\in A(i)}\Big[c^{(s_0)}(i,a) + \beta \sum_{j\in L(i)} q(i,a,j)\big\{v_\beta^{(s_0)}(j) - v_\beta^{(s_0)}(i)\big\}\Big] + \beta v_\beta^{(s_0)}(i) \\ v_\beta^{(t)}(i) &= \min_{a\in A_\beta^{(t-1)}(i)}\Big\{c^{(t)}(i,a) + \beta\Big[\sum_{j\in L(i)} q(i,a,j)\big\{v_\beta^{(t)}(j) - v_\beta^{(t)}(i)\big\} \\ &\quad + \sum_{m=1}^{M^*(i,t)} \sum_{j\in\mathcal{L}(l(i)+m)} q(i,a,j)\{v_\beta^{(t-m)}(j) - v_\beta^{(t-m)}(i)\} + v_\beta^{(t)}(i)\Big]\Big\} \end{aligned}$$

with

$$\begin{aligned} M^*(i,t) &= \min\{M - l(i), t - s_0\}, \\ A_\beta^{(s_0)}(i) &= \operatorname*{argmin}_{a\in A(i)}\Big[c^{(s_0)}(i,a) + \beta \sum_{j\in L(i)} q(i,a,j)\big\{v_\beta^{(s_0)}(j) - v_\beta^{(s_0)}(i)\big\}\Big] \\ A_\beta^{(t)}(i) &= \operatorname*{argmin}_{a\in A_\beta^{(t-1)}(i)}\Big\{c^{(t)}(i,a) + \beta\Big[\sum_{j\in L(i)} q(i,a,j)\big\{v_\beta^{(t)}(j) - v_\beta^{(t)}(i)\big\} \\ &\quad + \sum_{m=1}^{M^*(i,t)} \sum_{j\in\mathcal{L}(l(i)+m)} q(i,a,j)\{v_\beta^{(t-m)}(j) - v_\beta^{(t-m)}(i)\} + v_\beta^{(t)}(i)\Big]\Big\} \end{aligned}$$

This is equivalent to

$$v_\beta^{(s_0)}(i) = \min_{a\in A(i)}\Big[\frac{c^{(s_0)}(i,a)+\beta\sum_{j\in L(i)} q(i,a,j)v_\beta^{(s_0)}(j)}{1-\beta\{1-\sum_{j\in L(i)} q(i,a,j)\}}\Big] \tag{3}$$

$$\begin{aligned} v_\beta^{(t)}(i) = \min_{a\in A_\beta^{(t-1)}(i)} & \Big\{1-\beta\{1-\sum_{j\in L(i)} q(i,a,j)\}\Big\}^{-1}\cdot \\ & \Big\{c^{(t)}(i,a)+\beta\Big[\sum_{j\in L(i)} q(i,a,j)v_\beta^{(t)}(j) \\ & +\sum_{m=1}^{M^*(i,t)}\sum_{j\in\mathcal{L}(l(i)+m)} q(i,a,j)\{v_\beta^{(t-m)}(j)-v_\beta^{(t-m)}(i)\}\Big]\Big\} \end{aligned} \tag{4}$$

For the computation of $v_\beta^{(t)}(i)$ we only need to know $v_\beta^{(s_0)}$ up to $v_\beta^{(t-1)}$ and the components 0 to $i-1$ of $v_\beta^{(t)}$. So if we want to compute v^β we have to start with $v_\beta^{(s_0)}$, then compute $v_\beta^{(s_0+1)}$ and so on. Each vector has to be computed in the order according to its numbering. □

Algorithm for discounted costs:

for $i\in E$ **do**
  $A_\beta^{(s_0)}(i)=A(i)$
$s=s_0$
$M^*(i,s)=\min\{M-l(i),s-s_0\}$
repeat
  for $i\in E$ **do**
    for $a\in A_\beta^{(s)}(i)$ **do**

$$\begin{aligned} x_\beta^{(s)}(i,a) = & \Big\{1-\beta\{1-\sum_{j\in L(i)} q(i,a,j)\}\Big\}^{-1}\times \\ & \Big\{c^{(s)}(i,a)+\beta\Big[\sum_{j\in L(i)} q(i,a,j)v_\beta^{(s)}(j) \\ & +\sum_{m=1}^{M^*(i,s)}\sum_{j\in\mathcal{L}(l(i)+m)} q(i,a,j)\{v_\beta^{(s-m)}(j)-v_\beta^{(s-m)}(i)\}\Big]\Big\} \end{aligned}$$

    $v_\beta^{(s)}(i)=\min_{a\in A_\beta^{(s)}(i)} x_\beta^{(s)}(i,a)$
    $A_\beta^{(s+1)}(i)=\operatorname{argmin}_{a\in A_\beta^{(s)}(i)} x_\beta^{(s)}(i,a)$
  $s:=s+1;$
until $|A_\beta^{(s)}(i)|=1$ for $i\in E$

The remaining question is at which power of ρ we can stop. The answer to this cannot be given in general. If after a number of steps all the action spaces appearing in the minimisation contain one action we know that the corresponding policy is asymptotically optimal. However, we cannot bound the vectors $v_\beta^{(t)}$ so there is no bound on the error if we approximate v^β by $\sum_{s=s_0}^t \rho^s v_\beta^{(s_0)}$. The same holds if the action spaces contain more than one state. We can neither guarantee that all remaining policies are asymptotically optimal. In practice this problem does not seem to be a great disadvantage. In most applications the

action sets converge fast to one element. Compared to the usual ways of solving MDPs, a very accurate solution can be computed in a relatively short time.

3 Average costs

For the average costs our approach is similar to the discounted costs. First we introduce a condition under which the system will always be unichain and the expected average cost does not depend on the starting state.

Condition 2 *In every state $i \in E$ there is under every action $a \in A(i)$ a transition to a state in a lower level or to a state in the same level but with a lower order.*

We denote the average costs under a policy $f \in F$ by g_f and the bias vector by w_f. From now on we write the vector with each element equal to one as e. Under condition 2 the pair (g_f, w_f) satisfy the following system of equations.

$$g_f e + w_f = c_f + P_f w_f \tag{5}$$

The minimal average cost is defined by $g = \min_{f \in F} g_f$ and the minimal bias vector by $w = \min_{f \in F^*} w_f$ with $F^* = \{f : g_f = g\}$. They solve the optimality equation for average costs.

$$g = \min_f \{c_f + P_f w - w\} \tag{6}$$

Under condition 2 it is known that g is a unique solution and w is unique up to a constant. Similar as in the discounted case we can introduce the parameter ρ and write $g_f(\rho)$ and $w_f(\rho)$.

Lemma 3.1 *Under the conditions 1 and 2 there exists for every policy $f \in F$ power series expansions of the expected average costs and of the bias vector.*

$$g_f(\rho) = \sum_{t=s_0}^{\infty} \rho^t g_f^{(t)}$$

$$w_f(\rho) = \sum_{t=s_0}^{\infty} \rho^t w_f^{(t)}$$

Proof: We know that we are allowed to choose one component of $w_f(\rho)$. We choose $w_f(\rho, 0) = 0$. With this extra equation the system (5) has a unique solution. If we define the vector $h_f(\rho)$ and the matrix $S_f(\rho)$ as

$$[h_f(\rho)]_i = \begin{cases} g_f(\rho) & \text{if } i = 0 \\ w_f(\rho, i) & \text{if } i > 0, \end{cases}, \qquad [S_f(\rho)]_{ij} = \begin{cases} 1 & \text{if } j = 0, \\ -p_{ij}^f(\rho) & \text{if } j > 0, i \neq j, \\ 1 - p_{ij}^f(\rho) & \text{if } j > 0, i = j. \end{cases},$$

we can rewrite (5) with this extra equation as

$$S_f(\rho) h_f(\rho) = c_f(\rho).$$

It is known (??) that this matrix $S_f(\rho)$ has an inverse if the corresponding Markov chain has a unique stationary distribution which is true under condition 2 and so

$$h_f(\rho) = \{S_f(\rho)\}^{-1} c_f(\rho). \tag{7}$$

The proof of the the existence of the power series expressions of g_f and w_f is similar to lemma 2.1. □

Condition 2 is strong, it is possible to weaken it and obtain similar results. To simplify our theory and algorithm we do not study such weaker conditions in this paper. See [10].

The proof of the following two lemmas is identical to the proof of 2.2 and 2.3.

Lemma 3.2 *Under the conditions 1 and 2 the PSE of the average cost converges for ρ small enough for every policy $f \in F$.*

Lemma 3.3 *Under the conditions 1 and 2 there exist $\rho^* > 0$ and $f^* \in F$ such that f^* is average optimal for $\rho \in (0, \rho^*)$.*

So we can express $g(\rho)$ as power series for ρ small enough

$$g(\rho) = \sum_{t=s_0}^{\infty} \rho^t g^{(t)}$$

Algorithm for average costs:

for $i \in E$ **do**
$A^{(s_0)}(i) = A(i)$
$s = s_0$;
repeat
for $a \in A^{(s)}(0)$ **do**

$$h^{(s)}(a) = \{c^{(s)}(0a) + \sum_{m=1}^{\min\{M, s-s_0\}} \sum_{k \in \mathcal{L}(m)} p(0ak) w^{(s-m)}(k)\}$$

$$g^{(s)} = \min_{a \in A^{(s)}(0)} h^{(s)}(a)$$

$$A^{(s+1)}(0) = \operatorname*{argmin}_{a \in A^{(s)}(0)} h^{(s)}(a)$$

$w^{(s)}(0) = 0$
for $i = 1$ **to** N **do**
for $a \in A^{(s)}(i)$ **do**

$$x^{(s)}(ia) = \Big\{ \sum_{j \in L(i)} p(iaj) \Big\}^{-1} \cdot \Big\{ c^{(s)}(ia) - g^{(s)} + \sum_{j \in L(i)} p(iaj) w^{(s)}(j) + \sum_{m=1}^{M^*(i,s)} \sum_{j \in \mathcal{L}(m+l(i))} p(iaj) \{ w^{(s-m)}(j) - w^{(s-m)}(i) \} \Big\}$$

$$w^{(s)}(i) = \min_{a \in A^{(s)}(i)} x^{(s)}(ia)$$

$$A^{(s+1)}(i) = \operatorname*{argmin}_{a \in A^{(s)}(i)} x^{(s)}(ia)$$

s=s+1
until $|A^{(s)}(i)| = 1$ for $i \in E$.

Theorem 2 *Under the conditions 1 and 2 the PSE of $g(\rho)$ and the asymptotically optimal policy f^* can be computed recursively with the previous algorithm.*

Proof: To examine the expected average cost under a policy f we have to know the structure of the underlying Markov chain. From condition 2 it follows that the process is unichain and so there is precisely one closed and communicating set. In all states there

is a positive probability to visit state 0 in finite time so 0 has to be part of the closed communicating set. The component of equation (5) corresponding to state 0 is

$$g_f(\rho) = \sum_{s=s_0}^{s_1} \rho^s c_f^s(0) + \sum_{m=1}^{M} \rho^m \sum_{j \in \mathcal{L}(m)} q_{ij}^f \{w_f(j) - w_f(0)\}. \tag{8}$$

Since $w_f(\rho)$ is unique up to a constant we can choose $w_f(\rho, 0) = 0$. This choice does not influence the value of $g_f(\rho)$ and it follows that the first coefficient of $g_f(\rho)$ is equal to $c_f^{(s_0)}(0)$. For the other coefficient we have:

$$g_f^{(s)} = c_f^{(s)}(0) + \sum_{m=1}^{s-s_0} \sum_{j \in \mathcal{L}(m)} q_{ij}^f w_f^{(s-m)}(j) \tag{9}$$

These components does also depend on $w_f(\rho)$. To determine $w_f(\rho)$ we use the other equations of system (5). So

$$\begin{aligned} w_f^{(s)}(i) = \Bigg[& c_f^{(s)}(i) - g_f^{(s)} + \sum_{j \in L(i)} q_{ij}^f w^{(s)}(j) \\ & + \sum_{m=1}^{M^*(i,s)} \sum_{j \in \mathcal{L}(m+l(i))} q_{ij}^f \left\{ w^{(s-m)}(j) - w^{(s-m)}(i) \right\} \Bigg] \left[\sum_{j \in L(i)} q_{ij}^f \right]^{-1} \end{aligned} \tag{10}$$

The algorithm now follows in the same way as in the discounted case. □

It is clear that, if the algorithm stops it will produce the asymptotically optimal policy, and a number of terms of the power series expansion of the asymptotically minimal expected average reward. If this policy is unique the algorithm will stop in a finite number of steps, but we can not give any bound on the number of iterations. However, if the optimal policy is not unique, the algorithm will not stop. Depending on the considered problem one should bound the number of iterations in a way that if there are more policies possible after that number of iterations, the difference between the expected average reward under these policies is very small.

4 Example

Model With the next model we illustrate the method of the previous pages. We consider two service centers each with their own finite buffer, with the sizes N_1 and N_2. Both queues have arrivals from outside the system according to a Poisson process with parameters λ_1 and λ_2. All the customers that have been served at queue 1 are routed to queue 2,but the customers that finished service at queue 2 leave the system. The service times are exponentially distributed with parameters μ_1 and μ_2. In the first center we can stop the service. We want to compute a policy that minimises the probability that a customer finds a full queue and is lost. If we consider this model as part of an ATM network blocking probabilities are often very small and hence motivate the light traffic conditions This model is studied by Koole and Liu [7] with the average cost criterion. There the optimality of a monotone switching curve, increasing in the number of customers has been proven. After showing the optimal policy under light traffic we will relate both results.

Formulation The state is described by the number of customers in both queues. So the state space is

$$E = \{(i_1, i_2) : 0 \leq i_j \leq N_j, j = 1, 2\}.$$

The decision that has to be made in each state is whether or not to serve customers of type 1, so

$$A(i) = \begin{cases} \{0,1\} & \text{if } i_1 \neq 0; \\ \{0\} & \text{otherwise} \end{cases}$$

and the transition probabilities

$$p(i,a,j)(\rho) = \begin{cases} \rho\lambda_1 & \text{if } i_1 \neq N_1,\ j_1 = i_1 + 1 \text{ and } j_2 = i_2; \\ \rho\lambda_2 & \text{if } i_2 \neq N_2,\ j_1 = i_1 \text{ and } j_2 = i_2 + 1; \\ \mu_1 & \text{if } a = 1,\ j_1 = i_1 - 1 \text{ and } j_2 = i_2 + 1; \\ \mu_2 & \text{if } i_2 \neq 0,\ j_1 = i_1 \text{ and } j_2 = i_2; \\ 0 & \text{otherwise,} \end{cases}$$

for $j \neq i$ and

$$p(i,a,i)(\rho) = 1 - \sum_{j \neq i} p(i,a,j)(\rho).$$

The cost is the probability of blocking a customer, so

$$c(i,a)(\rho) = \rho\lambda_1 1_{\{i_1 = N_1\}} + \rho\lambda_2 1_{\{i_2 = N_2\}} + \mu_1 1_{\{i_2 = N_2, a = 1\}}$$

The level of a state i becomes the total number of customers and in a level the state with most type 1 customers has the highest ordering. We start with the discounted case and see that the discounted optimal policy will never serve if the number of free places in the buffer of the first queue is more than or equal to the number of free places in the buffer of the second queue. If the free places in the second buffer is one more than in the first buffer the optimal strategy does depend on the parameters. For $N_1 = 15$ and $N_2 = 10$ this policy can be found in table 1.

Theorem 3

$$A_\beta^{(0)}(i_1, N_2) = \{0\} \text{ if } i_1 = 0, \ldots, N_1,$$

$$v_\beta^{(0)}(i) = 0.$$

For $n = 1, 2, \ldots$

$$A_\beta^{(n)}(i_1,i_2) = \begin{cases} \{0\} & \text{if } N_1 - i_1 + 1 > N_2 - i_2, N_2 - i_2 \leq n; \\ \{1\} & \text{if } N_1 - i_1 \leq n-1, N_1 - i_1 < N_2 - i_2 - 1; \\ \{0\} \text{ or } \{1\} & \text{if } N_1 - i_1 \leq n-1, N_1 - i_1 = N_2 - i_2 - 1; \\ \{0,1\} & \text{otherwise} \end{cases}$$

$$v_\beta^{(n)}(i_1,i_2) = \begin{cases} \frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n} & \text{if } i_1 = N_1 - n + 1, i_2 \leq N_2 - n - 1; \\ \frac{\beta^{n-1}\lambda_2^n}{\{1-\beta(1-\mu_2)\}^n} & \text{if } i_1 \leq N_1 - n + 1, i_2 = N_2 - n + 1; \\ \frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n} + \frac{\beta^n\mu_1}{1-\beta(1-\mu_2)} \times & \\ \min\left\{ \frac{\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n}, \right. & \\ \left. \frac{\lambda_2^n}{\{1-\beta(1-\mu_1-\mu_2)\}\{1-\beta(1-\mu_2)\}^{n-1}} \right\} & \text{if } i_1 = N_1 - n + 1, i_2 = N_2 - n; \\ 0 & \text{if } i_1 \leq N_1 - n, i_2 \leq N_2 - n. \end{cases}$$

Proof: The proof is by induction. First see what happens in the first iteration. We compute $v_\beta^{(0)}$ by the following equation:

$$v_\beta^{(0)}(i) = \min_{a \in A(i)} \frac{\mu_1 1_{\{i_2 = N_2, a = 1\}} + \beta[1_{\{a=1\}}\mu_1 v_\beta^{(0)}(i_1 - 1, i_2 + 1) + \mu_2 v_\beta^{(0)}(i_1, i_2 - 1)]}{1 - \beta[1 - \mu_1 1_{\{a=1\}} - \mu_2]} \quad (11)$$

and from this it follows that

$$\begin{aligned} v_\beta^{(0)}(i) &= 0 \qquad\qquad \forall i \\ A_\beta^{(0)}(i) &= \begin{cases} \{0,1\} & \text{if } i_2 \neq N_2 \\ \{0\} & \text{if } i_2 = N_2 \end{cases} \end{aligned}$$

Next we show what happens in the second iteration. The equation that we use is:

$$\begin{aligned} v_\beta^{(1)}(i) &= \min_{a\in A^{(0)}(i)} \Big[\lambda_1 1_{\{i_1=N_1\}} + \lambda_2 1_{\{i_2=N_2\}} + \beta[1_{\{a=1\}}\mu_1 v_\beta^{(1)}(i_1-1,i_2+1) \\ &\quad +\mu_2 v_\beta^{(1)}(i_1,i_2-1)]\Big] \left[1-\beta[1-\mu_1 1_{\{a=1\}}-\mu_2]\right]^{-1} \end{aligned} \tag{12}$$

The cases mentioned in the lemma must all be considered separately, we show the case that $i_1 = N_1, i_2 \leq N_2 - 2$, the others are similarly.

$$\begin{aligned} v_\beta^{(1)}(N_1,0) &= \min_{a=0,1} \frac{\lambda_1 + \beta 1_{\{a=1\}}\mu_1 v_\beta^{(1)}(N_1-1,1)}{1-\beta[1-\mu_1 1_{\{a=1\}}]} \\ &= \min\left\{\frac{\lambda_1}{1-\beta}, \frac{\lambda_1}{1-\beta[1-\mu_1]}\right\} \\ &= \frac{\lambda_1}{1-\beta[1-\mu_1]}, \end{aligned}$$

$$\begin{aligned} v_\beta^{(1)}(N_1,i_2) &= \min_{a=0,1} \frac{\lambda_1 + \beta[1_{\{a=1\}}\mu_1 v_\beta^{(1)}(N_1-1,i_2+1) + \mu_2 v_\beta^{(1)}(N_1,i_2-1)]}{1-\beta[1-\mu_1 1_{\{a=1\}}-\mu_2]} \\ &= \min_{a=0,1} \frac{\lambda_1 + \beta\mu_2 \frac{\lambda_1}{1-\beta[1-\mu_1]}}{1-\beta[1-\mu_1 1_{\{a=1\}}-\mu_2]} \\ &= \frac{\lambda_1}{1-\beta[1-\mu_1]} \end{aligned}$$

and hence

$$A_\beta^{(1)}(N_1,i_2) = \{1\}.$$

Finally we consider iteration n+1.

$$\begin{aligned} v_\beta^{(n)}(i) &= \min_{a\in A^{(n-1)}(i)} \beta\Big[\mu_1 1_{\{a=1\}} v_\beta^{(n)}(i_1-1,i_2+1) + \mu_2 v_\beta^{(n)}(i_1,i_2-1) \\ &\quad +\lambda_1[v_\beta^{(n-1)}(i_1+1,i_2) - v_\beta^{(n-1)}(i)] + \lambda_2[v_\beta^{(n-1)}(i_1,i_2+1) - v_\beta^{(n-1)}(i)]\Big] \\ &\quad \left[1-\beta[1-\mu_1 1_{\{a=1\}}-\mu_2]\right]^{-1} \end{aligned}$$

Again, all cases have a similar approach and we show the case $i_1 = N_1 - n + 1, i_2 \leq N_2 - n - 1$.

$$\begin{aligned} v_\beta^{(n)}(N_1-n+1,0) &= \min_{a=0,1} \frac{\beta\lambda_1 v_\beta^{(n-1)}(N_1-n+2,0)}{1-\beta[1-\mu_1 1_{\{a=1\}}]} \\ &= \frac{\beta\lambda_1 \frac{\beta^{n-2}\lambda_1^{n-1}}{\{1-\beta(1-\mu_1)\}^{n-1}}}{1-\beta[1-\mu_1]} \\ &= \frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n} \end{aligned}$$

and

$$
\begin{aligned}
v_\beta^{(n)}(N_1-n+1,i-2) &= \min_{a=0,1}\beta\frac{\mu_2 v_\beta^{(n)}(N_1-n+1,i_2-1)+\lambda_1 v_\beta^{(n-1)}(N_1-n+2,i_2)}{1-\beta[1-\mu_1 1_{\{a=1\}}-\mu_2]}\\
&= \beta\frac{\mu_2\frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n}+\lambda_1\frac{\beta^{n-2}\lambda_1^{n-1}}{\{1-\beta(1-\mu_1)\}^{n-1}}}{1-\beta[1-\mu_1-\mu_2]}\\
&= \frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^{n-1}}\cdot\frac{\frac{\beta\mu_2}{1-\beta(1-\mu_1)}+1}{1-\beta[1-\mu_1-\mu_2]}\\
&= \frac{\beta^{n-1}\lambda_1^n}{\{1-\beta(1-\mu_1)\}^n}.
\end{aligned}
$$

□

It is important to note that if $N_1 > N_2$ in the states of the form $(i_1, 0)$ with $i_1 \leq N_1 - N_2$ there is no service at the first queue.

Next, we want to consider the average cost case. The problem is that condition 2 does not hold. We can deal with this problem by using the results of [7]. They showed that if in a state $(i_1, 0)$ there is no service at queue 1, the same holds in the states (i_1, i_2) with $i_2 > 0$. This means that there will always be at least i_1 customers in queue 1. With a coupling argument it can be showed that this can never lead to an optimal policy, so we can assume that in an optimal policy in all states of the form $(i_1, 0)$ there is service in the first service center. If we adjust the action spaces in this way condition 2 will hold and we can apply the algorithm. We state the results without proof, which would be similar to that for the discounted case.

Theorem 4 *The optimal policy under average cost f^* is of the form:*

$$
f^*(i) = \begin{cases} 0 & \text{if } N_1 - i_1 \geq N_2 - i_2, i_2 \neq 0; \\ 0 & \text{if } N_1 - i_1 + 1 = N_2 - i_2, i_2 \neq 0 \text{ and } \frac{\lambda_1^n}{\mu_1^n} < \frac{\lambda_2^n}{(\mu_1+\mu_2)\mu_2^{n-1}}; \\ 0 & \text{if } i_1 = 0 \\ 1 & \text{otherwise.} \end{cases}
$$

In the following tables we show the discounted and average optimal policies. A $*$ in the table means that the optimal action depends on the system parameters.

x_2:10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 *
8	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1
7	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1
6	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1
5	0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1 1
4	0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1 1 1
3	0 0 0 0 0 0 0 0 0 0 * 1 1 1 1 1 1
2	0 0 0 0 0 0 0 0 0 * 1 1 1 1 1 1 1
1	0 0 0 0 0 0 0 0 * 1 1 1 1 1 1 1 1
0	0 0 0 0 0 0 0 * 1 1 1 1 1 1 1 1 1
x_1:	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6

Table 1. Discounted optimal policy

x_2:10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 *
8	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1
7	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1
6	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1
5	0 0 0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1 1
4	0 0 0 0 0 0 0 0 0 0 0 0 * 1 1 1 1 1
3	0 0 0 0 0 0 0 0 0 0 * 1 1 1 1 1 1
2	0 0 0 0 0 0 0 0 0 * 1 1 1 1 1 1 1
1	0 0 0 0 0 0 0 0 * 1 1 1 1 1 1 1 1
0	0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
x_1:	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6

Table 2. Average optimal policy

The optimal policy is a special form of an increasing switching curve and is called the *balanced hole* policy. Simulations show that in general this policy behaves well if the load in the second center is high compared to the load in the first queue.

References

[1] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In L. Donatiello and R. Nelson, editors, *Performance evaluation of computer and communication systems*, Lecture notes in computer sciences 729, pages 53–80. Springer-Verlag, 1993.

[2] C. Derman. *Finite state Markov decision processes*. Academic Press, New York, 1970.

[3] M.N. Katehakis and C. Derman. On the maintenance of system composed of highly reiable components. *Management Science*, 35(5):551–560, 1989.

[4] M.N. Katehakis and A. Levine. A dynamic routing problem - numerical procedures for light traffic conditions. *Applied Mathematics and Computation*, 17:267–276, 1985.

[5] M.N. Katehakis and A. Levine. Allocation of distinguishable servers. *Computers and Operations Research*, 13:85–93, 1986.

[6] G. Koole. On the power series algorithm. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994. CWI Tract 105 & 106.

[7] G. Koole and Z. Liu. Nonconservative service for minimizing cell loss in ATM networks. In *Proceedings of the 1st Workshop on ATM Traffic Management*, 1995.

[8] G. Koole and M. Vrijenhoek. Scheduling a repairman in a finite source system. *ZOR - Mathematical Methods of Operations Research*, 44:333–344, 1996.

[9] A. Levine and D. Finkel. Load balancing in a multiserver queueing system. *Computers and Operations Research*, 17(1):17–24, 1990.

[10] O. Passchier. *The Theory of Markov Games and Queueing Control*. PhD thesis, Leiden University, 1996.

[11] D.R. Smith. Optimal repairman allocation - asymptotical results. *Management Science*, 24:665–674, 1978.