ZERO-SUM MARKOV GAMES AND WORST-CASE OPTIMAL CONTROL OF QUEUEING SYSTEMS

Eitan ALTMAN
INRIA
2004 Route des Lucioles
BP93, 06902 Sophia-Antipolis Cedex
France

Arie HORDIJK
Dept. of Mathematics and Computer Science
Leiden University
P.O.Box 9512, 2300RA Leiden
The Netherlands

Submitted: February 1994, Revised: July 1995

Abstract

Zero sum stochastic games model situations where two persons, called players, control some dynamic system, and both have opposite objectives. One player wishes typically to minimize a cost which has to be paid to the other player. Such a game may also be used to model problems with a single controller who has only partial information on the system: the dynamic of the system may depend on some parameter that is unknown to the controller, and may vary in time in an unpredictable way. A worst-case criterion may be considered, where the unknown parameter is assumed to be chosen by "nature" (called player 1), and the objective of the controller (player 2) is then to design a policy that guarantees the best performance under worst-case behaviour of nature. The purpose of this paper is to present a survey of stochastic games in queues, where both tools and applications are considered. The first part is devoted to the tools. We present some existing tools for solving finite horizon and infinite horizon discounted Markov games with unbounded cost, and develop new ones that are typically applicable in queueing problems. We then present some new tools and theory of expected average cost stochastic games with unbounded cost. In the second part of the paper we present a survey on existing results on worst-case control of queues, and illustrate the structural properties of best policies of the controller, worst-case policies of nature, and of the value function. Using the theory developed in the first part of the paper, we extend some of the above results, which were known to hold for finite horizon costs or for the discounted cost, to the expected average cost.

Keywords: Zero-sum stochastic games, discounted and expected average cost, Worst case control of queueing networks, value iteration, structural properties of optimal policies and value function.

1 Introduction

The purpose of this paper is to present the topic of worst-case control of queues, both from the point of view of the tools it requires, i.e. stochastic games with unbounded cost, and from the point of view of the applications. The zero-sum stochastic game model that we use is quite a pessimistic one. It assumes that an imaginary player, called nature, is playing against us, and tries to control some dynamic unknown parameters in the worst possible way. This view point is supported by a rich experimental history, that is best summarized by the well known Murphy's laws, that state, among others, that "anything that can go wrong, will go wrong". If one accepts this view point, one is naturally faced with the problem of guaranteeing the best performance under the worst possible behaviour of nature. An optimistic person, on the other hand, could assume that nature is on his side. He would then be faced with the problem of guaranteeing the best performance under the best performance of nature. This gives rise to a standard control view point (with several agents) that have the same objectives. Finally, if one has some knowledge about nature's objective, one could use tools from non-zero sum non-cooperative games.

The theory on optimal control of queues is well established, see e.g. the survey on flow control models in [36], the tutorial review by Stidham [35] on control of admission, routing, and service in queues and networks of queues, and [37] by Walrand. Both tools and applications have been extensively studied for control problems both with a single agent and with several agents (see e.g. [16]). The theory on worst-case control as well as other types of dynamic games arising in queueing models (e.g. when several users compete for some resource) are quite recent and partial, due in part to the fact that structural results are simpler to obtain in infinite queues, for which the holding costs, or costs related to waiting times, are typically unbounded. Tools for solving such problems were not available until recently, especially for the expected average cost (which is the cost the most used in computers and telecommunications applications). Indeed, whereas for finite horizon and infinite horizon discounted cost, the theory of stochastic games with unbounded cost goes back to the seventies, see [39] (and [5, 31, 33] for more recent papers) the theory for the expected average cost has not yet been published, and appears in the recent works [5, 10, 33].

An interesting situation in worst-case control of queueing networks occurs when not every pair of stationary policies gives rise to a recurrent Markov chain. Typically, the controller is the vulnerable player, in the sense that if he plays "badly" then nature can cause the Markov chain to be transient (or null recurrent). Since costs are unbounded, this typically results in an infinite expected average cost. On the other hand, the controller can usually enforce stability, i.e., he has some set of stationary policies for which the resulting Markov chain will be stable under any policy of nature. We call this situation a weak stability condition. Important questions are, first, can the controller restrict to stationary policies? Are there worst-case policies (of nature) which are stationary? Another typical question is whether the game has a value, or, equivalently, would the cost be the same if nature chooses his policy first and then the controller chooses his, knowing the choice of nature, or if the reverse happens. In the situation described in the beginning of the paragraph it is not intuitively clear that a value exists. Indeed, it is clear that if the controller chooses first his policy then he is obliged to choose one that stabilizes the system (i.e., for which the resulting Markov chain is recurrent) under any stationary policy of nature. Now, if nature plays first and the controller responds, one could expect that a best response of the controller need not be restricted to those stabilizing policies, and then perhaps he could do better.

In this paper we present general conditions for the value of the game as well as optimal stationary policies for both players to exist, thus extending the conditions in [5, 10, 33]. (Indeed, the conditions in [5, 10] are restrictive as they require the strong stability condition, i.e., under any pair of stationary policies of both players, the state process is a recurrent Markov chain; the conditions in [33] are restrictive as they require a nonnegative immediate cost, only a finite number of actions at any state, and other conditions). The conditions that we present here are weaker, and thus more general than those in [5, 10, 33]). On the other hand, we present conditions more restrictive than in [5, 10, 33] that have the advantage that they are easily verifiable in queueing models. Some of them are in the spirit of Borkar's conditions [9] and Weber and Stidham's [38] for control problems.

We then present a survey of recent structural results in stochastic zero-sum games in queueing network. We present both flow control [1, 2], service control [5], routing control [3, 6, 8] and a service assignment model [6]. Typical monotonicity and other properties of optimal policies are described, that are obtained by value iteration techniques, after showing that the value function has some nice properties (monotonicity, convexity, supermodularity, Schur convexity etc...) The type of monotonicity results are richer than in standard control problems, since optimal policies usually require randomization, and the way that monotonicity translates to randomized policies is not straight-forward. For the different models above, we apply our new theorems from the first part of the paper, to show that the existing results that hold for the discounted cost or for the finite horizon case carry over to the expected average cost. We have not yet been able to solve all the above problems under weak stability conditions, and some of them remain open.

Finally, we would like to mention that some results on structure of policies exist also for non-zero sum stochastic games in queueing models, see [4, 7, 8, 15, 17, 18, 24, 26]. It seems however that there is still much to do in that area, especially in developing tools, since the standard techniques of value iteration do not hold for non zero-sum games.

2 Zero-sum stochastic game: the model

We begin by a brief description of stochastic games. We present the model, and then present existing tools and develop new ones in order to handle problems with unbounded cost.

We consider a countable state space $\mathbf{X}$; at each state $x \in \mathbf{X}$ the compact sets of actions $\mathbf{A}(x), \mathbf{B}(x)$ are available to player 1 and 2, respectively. Let $\mathbf{A} = \cup_x \mathbf{A}(x)$, $\mathbf{B} = \cup_x \mathbf{B}(x)$. Let $K := (x, a, b : x \in \mathbf{X}, a \in \mathbf{A}(x), b \in \mathbf{B}(x))$. $P = \{P_{xaby}\}$ are the transition probabilities, where P_{xaby} is the the probability to go from state x to state y given that actions a and b are chosen by the players. For all x, y, P_{xaby} is continuous in a and b. $c : K \to \mathbb{R}$ is an immediate cost, assumed to be continuous in a and b. Let R and U denote the set of policies for players 1 and 2. A policy $r \in R$ is a sequence $r = (r_1, r_2, ...)$, where r_t is a probability over $\mathbf{A}(x_t)$, x_t being the state at time t, conditioned on the history of all actions of both players and of the history of all states until time $t-1$, as well as the state at time t. Define similarly the policies for the other player. The actions of the two players at time t are chosen independently according to r_t and u_t. Let $\mathcal{P}_z^{\rho,\eta}$ and $E_z^{\rho,\eta}$ denote the (unique) probability measure induced by an initial state z and policies ρ, η, and the corresponding expectation. Let $\{X_t, A_t, B_t\}$ be the resulting stochastic process describing the states and actions.

We denote by $R(M)$, $U(M)$ the set of Markov policies for players 1 and 2, (where u_t and r_t are functions only of the state at time t and of the time t), and by $R(S)$ and $U(S)$ the stationary (randomized) policies (where u_t and r_t are functions only of the state at time t that are the same for all t). When a pair of Markov policies $\rho \in R(S), \eta \in U(S)$ is used by the players, then the state process becomes a Markov chain (non homogeneous in time). Its transition probabilities at step t are denoted by $P(\rho_t, \eta_t) = \{P_{xy}(\rho_t, \eta_t)\}$. In particular, if ρ and η are stationary then the Markov chain is time homogeneous and the transition probabilities are denoted by $P(\rho, \eta)$. We denote then the Cezaro limit by

$$P^\infty(\rho,\eta) := \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^t P^s(\rho,\eta).$$

By $P(a,\eta)$ we shall understand the transition probabilities corresponding to player 1 using the same action a in all states, and by $P(\alpha,\eta)$ we denote the transition probabilities corresponding to player 1 using the same randomizing rule α in all states. A pair of stationary policies (r,u) is said to be unichain, if the corresponding Markov chain does not consist of more than one closed class of states.

We consider the following cost criteria.
The finite horizon discounted cost: $V^{t,\xi}_{\rho,\eta}(z) = E^{\rho,\eta}_z \sum_{s=1}^t \xi^{s-1} c(X_s,A_s,B_s)$ where $\xi \le 1$ is the discount factor;
the infinite horizon discounted cost: $V^{\xi}_{\rho,\eta}(z) = E^{\rho,\eta}_z \sum_{s=1}^\infty \xi^{s-1} c(X_s,A_s,B_s)$ (in this case we assume $\xi < 1$),
the infinite horizon expected average cost: $g_{\rho,\eta}(z) = \overline{\lim}_{t\to\infty} t^{-1} E^{\rho,\eta}_z \sum_{s=1}^t c(X_s,A_s,B_s)$.

Define the following lower values of the stochastic game:

$$\underline{V}^{t,\xi}(z) = \sup_\rho \inf_\eta V^{t,\xi}_{\rho,\eta}(z), \qquad \underline{V}^{\xi}(z) = \sup_\rho \inf_\eta V^{\xi}_{\rho,\eta}(z), \qquad \underline{g}(z) = \sup_\rho \inf_\eta g_{\rho,\eta}(z), \tag{1}$$

and the upper values

$$\overline{V}^{t,\xi}(z) = \inf_\eta \sup_\rho V^{t,\xi}_{\rho,\eta}(z), \qquad \overline{V}^{\xi}(z) = \inf_\eta \sup_\rho V^{\xi}_{\rho,\eta}(z), \qquad \overline{g}(z) = \inf_\eta \sup_\rho g_{\rho,\eta}(z). \tag{2}$$

For the finite horizon case, the objective of the controller (player 2) is to minimize $\sup_\rho V^{t,\xi}_{\rho,\eta}(z)$. If a minimizing policy η^* exists, it is called an optimal policy for player 2. In that case, $\sup_\rho V^{t,\xi}_{\rho,\eta^*}(z) = \overline{V}^{t,\xi}(z)$. If moreover, $\sup_\rho V^{t,\xi}_{\rho,\eta^*}(z) = \underline{V}^{t,\xi}(z)$ then it is called strongly optimal. Similarly, ρ^* is an optimal policy for nature (player 1) if $\inf_\eta V^{t,\xi}_{\rho^*,\eta}(z) = \underline{V}^{t,\xi}(z)$. It is called strongly optimal if $\inf_\eta V^{t,\xi}_{\rho^*,\eta}(z) = \overline{V}^{t,\xi}(z)$. If $\underline{V}^{t,\xi} = \overline{V}^{t,\xi}$ then we denote $V^{t,\xi} = \underline{V}^{t,\xi} = \overline{V}^{t,\xi}$ and call it the value of the finite horizon problem. In that case, any optimal policy is strongly optimal. We define similarly optimality of policies and the values for the other cost criteria.

3 Notation

For any set S, we shall use $\mathcal{M}(S)$ to denote the set of probability measures over S. For any decision rules $\alpha : \mathbf{X} \to \mathcal{M}(\mathbf{A}(x))$ $\beta : \mathbf{X} \to \mathcal{M}(\mathbf{B}(x))$ and any function $d : \mathbf{A}(x) \times \mathbf{B}(x) \to \mathbb{R}$, we shall define, with some abuse of notation,

$$d(x,\alpha,\beta) \quad = \quad \int_{\mathbf{A}} \int_{\mathbf{B}} d(x,a,b)\alpha(da)\beta(db)$$

(whenever the integrals are well defined). Denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$. and $\mathbb{R}^+ = [0,\infty]$.

We shall often use $c(\alpha,\beta)$ to denote the column vector whose entries are $c(x,\alpha,\beta)$, etc... $P(\rho,\eta)V^{t,\xi}$ then stands for the vector whose x entry is $\sum_{y\in\mathbf{X}} P_{x\rho(x)\eta(x)y} V^{t,\xi}(y)$.

Denote the unit vector $e : \mathbf{X} \to R$ such that $e(x) = 1$ for all $x \in \mathbf{X}$.

For any $x \in \mathbf{X}$ and function $d : \mathbf{A}(x) \times \mathbf{B}(x) \to \mathbb{R}$, if

$$\sup_{\alpha \in \mathcal{M}(\mathbf{A}(x))} \inf_{\beta \in \mathcal{M}(\mathbf{B}(x))} d(\alpha, \beta) = \inf_{\beta \in \mathcal{M}(\mathbf{B}(x))} \sup_{\alpha \in \mathcal{M}(\mathbf{A}(x))} d(\alpha, \beta)$$

then we define $\mathrm{val}[d] = \sup_{\alpha \in \mathcal{M}(\mathbf{A}(x))} \inf_{\beta \in \mathcal{M}(\mathbf{B}(x))} d(\alpha, \beta)$. α^* and β^* are said to be optimal policies for the dummy game d if

$$\sup_{\alpha \in \mathcal{M}(\mathbf{A}(x))} d(\alpha, \beta^*) \leq \mathrm{val}[d] \leq \inf_{\beta \in \mathcal{M}(\mathbf{B}(x))} d(\alpha^*, \beta).$$

An important class of stochastic games that will be considered in queueing applications consists of the contracting stochastic games for which a value exists and optimal policies are known to exist for both players; both the value and the optimal policies can be obtained from the optimality equation [5, 39].

Let $\mu : \mathbf{X} \to [1, \infty)$ be some positive function. Following Dekker and Hordijk [12] and Spieksma [34], define the μ-norm of vectors $\pi \in \mathbb{R}^{\mathbf{X}}$ and of matrices $Q \in \mathbb{R}^{\mathbf{X} \times \mathbf{X}}$ as

$$\begin{aligned} \|\pi\|_\mu &= \sup_{x \in \mathbf{X}} \mu_x^{-1} |\pi_x| \\ \|Q\|_\mu &= \sup_{x \in \mathbf{X}} \mu_x^{-1} \sum_{y \in \mathbf{X}} |Q_{xy}| \mu_y \end{aligned} \tag{3}$$

For a subset $M \subset \mathbf{X}$, let ${}_M P$ be the taboo matrix corresponding to P, i.e.

$${}_M P_{xy} = \begin{cases} P_{xy}, & y \notin M \\ 0, & y \in M \end{cases} \tag{4}$$

A stochastic game is said to be μ-uniform geometrically recurrent (μ-UGR), if a finite set M and a $\zeta < 1$ exist, such that for any $r \in R(S), u \in U(S)$,

$$\|{}_M P_{ru}\|_\mu \leq \zeta. \tag{5}$$

We introduce the following assumptions related to contracting stochastic games. For some μ as defined above,
Assumption 1: (i) the immediate cost is μ-bounded, i.e.

$$\sup_{x \in \mathbf{X}} \sup_{a \in \mathbf{A}(x)} \sup_{b \in \mathbf{B}(x)} \frac{|c(x, a, b)|}{\mu(x)} < \infty.$$

(ii) The transition probabilities are μ-continuous, i.e. for all $x \in \mathbf{X}$ and any sequences $a(n) \to a$, $b(n) \to b$,

$$\lim_{n \to \infty} \sum_{y \in \mathbf{X}} |P_{xa(n)b(n)y} - P_{xaby}| \mu_y \to 0.$$

Assumption 2(ξ): There is a finite $M \subset \mathbf{X}$ and a constant $\zeta < 1$, such that for any $x \in \mathbf{X}$, $a \in \mathbf{A}$, $b \in \mathbf{B}$,

$$\xi \sum_{y \in \mathbf{X}} {}_M P_{xaby} \mu_y < \zeta \mu_x.$$

This condition is equivalent to the μ-UGR for $\xi = 1$ [12].

4 Shapley's optimality equations, optimal policies, value iteration

We now present the basic tools for obtaining the values in stochastic games and optimal policies for both players. We summarize known results for the finite horizon and the infinite horizon discounted cost, and present new results on expected average cost.

4.1 The finite horizon case

For any $\xi \in [0,1]$ we define recursively $v^{\xi,t} : \mathbf{X} \to \mathbb{R}$ as

$$\begin{aligned} v^{0,\xi}(x) &= 0 \\ v^{t+1,\xi}(x) &= \mathrm{val}_{a,b}\left[c(x,a,b) + \xi \sum_y P_{xaby} v^{t,\xi}(y)\right], \qquad t = 0,1,... \end{aligned} \tag{6}$$

We call the argument of the "val" the $(t+1)$-dummy game. The value $v^{t,\xi}$ is well defined for all $t = 0,1,...$ and both players have optimal policies [5, 28] for these dummy games (under the conditions of the Lemma below). These randomized decision rules will be called "conservative policies".

Lemma 4.1 *Consider the problem with finite horizon T. Assume that c is bounded, or that Assumptions 1 and $2(\xi)$ hold. Then*
(i) the stochastic game has a value $V^{T,\xi}$ and $V^{T,\xi} = v^{T,\xi}$.
(ii) Consider the randomized decision rules r_t and u_t obtained as the conservative (optimal) policies for the t dummy games, $t = 1,...,T$. Then the Markov policies $r = (r_T,...,r_1)$, $u = (u_T,...,u_1)$ are strongly optimal for players 1 and 2.

Proof. For the bounded cost see [28], and for the other case, see [5]. ∎

We note that in all queueing applications that we consider, it is sufficient to consider the above Lemma for bounded cost, as the applications satisfy the following nearest neighbor property: from any given state, only finitely many states are accessible within one transition. Therefore, any finite horizon problem is equivalent to a stochastic game with a finite number of states if we fix the initial state at time 1, and hence the immediate cost can be assumed to be bounded.

4.2 The infinite horizon discounted cost

We summarize results on contracting stochastic games as obtained in [5], which generalize the approach of [39].

Lemma 4.2 *[5] Assume that Assumptions 1 and $2(\xi)$ hold, with $\xi < 1$. Then*
(i) There exists a μ-bounded solution of the optimality equation,

$$v^{\xi}(x) = \mathrm{val}_{a,b}\left[c(x,a,b) + \xi \sum_y P_{xaby} v^{\xi}(y)\right].$$

This solution is unique in the class of μ-bounded functions and it satisfies $v^\xi = \lim_{t\to\infty} v^{\xi,t}$ (defined in (6)). Morever, the stochastic game has a value $V^\xi = v^\xi$.
(ii) Let α, β be any decision rules that are optimal in the dummy game $[c(a,b) + \xi P_{ab} v^\xi]$. Then the stationary policies r, u such that $r_t = \alpha, u_t = \beta$ (for all t) are strongly optimal for both players.

The above contracting framework will be again suitable for all the queueing examples, even under nonstable conditions (where the size of some queues tend to infinity). This follows from the following property. If the current total number of customers in all queues is y, then after one transition they can are bounded by $y + L$, where L is some integer. The state in all queueing examples that we consider is either a vector x of the number of customers in all queues, or sometimes a vector of the form $\underline{x} = (x_1, x_2)$, where x_1 belongs to some finite set, and x_2 is the vector of queue lengths. As we shall see, this implies the contracting framework for immediate cost functions that are polynomially bounded, and for some that grow exponentially fast. We shall use below the notation $|x_2| = \sum_{k=1}^d x_2(k)$, for $x_2 \in N^d$.

Lemma 4.3 *Let d be some integer and assume that $\mathbf{X} = \mathbf{X}_1 \times \mathbf{X}_2$ where $\mathbf{X}_1$ is a finite set and $\mathbf{X}_2 = \mathbb{N}^{\mathrm{d}}$. Assume that*
(i) there is a constant L such that from any state $x = (x_1, x_2) \in \mathbf{X}$, only states $y = (y_1, y_2)$ satisfying $|y_2| \le |x_2| + L$ may be accessible.
(ii) Let $\gamma > 1$ be some constant with $\zeta := \gamma^L \xi < 1$, and assume that the immediate cost satisfies

$$\sup_{x\in\mathbf{X}} \sup_{a,b} \frac{c(x,a,b)}{\gamma^{|x_2|}} < \infty. \tag{7}$$

Then Assumptions 1 and $2(\xi)$ are satisfied with $\mu(x_1, x_2) = \gamma^{|x_2|}$.

Proof. Assumption 1(i) holds by the definition of μ. Assumption 1(ii) follows since the summation there is over a finite set, and hence the limit and summation may be interchanged. For Assumption 2, we choose $M = \emptyset$, and obtain indeed for all a, b

$$\xi \sum_{y\in\mathbf{X}} {}_M P_{xaby} \mu_y = \xi \sum_{y\in\mathbf{X}} {}_M P_{xaby} \gamma^{|y_2|} \le \xi \gamma^{|x_2|+L} < \zeta \mu_x.$$

■

4.3 The expected average cost: general conditions and characterization of optimal policies

We begin by presenting general conditions for existence and characterizations of optimal stationary policies for the expected average cost. We then present in the following subsection verifiable sufficient conditions.

The following set of conditions **A** will serve in establishing the existence of optimal stationary policies for both players, obtained as limits of ξ-discounted optimal policies.

Assumption A1: For all ξ sufficiently close to one, there exist stationary policies $\rho^\xi \in R(S), \eta^\xi \in U(S)$ and a function $v^\xi : \mathbf{X} \to \overline{\mathbb{R}}$ such that for all a, b:

$$c(a, \eta^\xi) + \xi P(a, \eta^\xi) v^\xi \le v^\xi \le c(\rho^\xi, b) + \xi P(\rho^\xi, b) v^\xi. \tag{8}$$

■

Define $\hat{v}^\xi = \inf_x v^\xi(x)$ and denote $w^\xi = v^\xi - \hat{v}^\xi e$ (provided that $\hat{v}^\xi \neq \infty$). Let $\xi(k)$ be a sequence along which the following limits exist:

$$\begin{aligned}
\lim_{k\to\infty} \xi(k) &= 1, \\
\lim_{k\to\infty} \rho^{\xi(k)} &=: \rho^*, \\
\lim_{k\to\infty} \eta^{\xi(k)} &=: \eta^*, \\
\lim_{k\to\infty} (1-\xi(k))\hat{v}^{\xi(k)} &=: g^*, \\
\lim_{k\to\infty} w^{\xi(k)} &=: w.
\end{aligned}$$

Assumption A2: $w^{\xi(k)} \neq \infty$ and $\hat{v}^{\xi(k)} \neq \infty$ (componentwise) for all k large enough, and componentwise,

$$\underline{\lim}_{k\to\infty} P(\rho^{\xi(k)}, b) w^{\xi(k)} \leq P(\rho^*, b) w \quad \text{for all} \quad b$$

Note that since w is nonnegative, by Fatou's Lemma, the dual of A2 holds (componentwise):

$$\overline{\lim}_{k\to\infty} P(a, \eta^{\xi(k)}) w^{\xi(k)} \geq P(a, \eta^*) w \quad \text{for all} \quad a$$

Next we present condition **B1**, which is weaker than A1-A2 and which, together with condition B2, will imply the existence of optimal stationary policies for both players, obtained as solutions of some optimality equation.

Assumption B1: There exist stationary policies $\rho^* \in R, \eta^* \in U$, and $g^* \in \mathbb{R}^+$, $w : \mathbf{X} \to \mathbb{R}^+$, such that

$$c(a, \eta^*) - g^* e + P(a, \eta^*) w \leq w \leq c(\rho^*, b) - g^* e + P(\rho^*, b) w. \tag{9}$$

Assumption B2:

$$g_{\rho^*, \eta} \geq g^* e$$

for all $\eta \in U(S)$.

Lemma 4.4 *A1 and A2 imply B1.*

Proof. Follows immediately by applying (8) to the subsequence $\xi(k)$ and subtracting $\hat{v}^{\xi(k)} e$ from both sides, and then going to the limit. ■

Proposition 4.1 *Assume that conditions B hold and $w < \infty$ (componentwise). Then (ρ^*, η^*) is average optimal, i.e.*

$$g_{\rho, \eta^*} \leq g^* e = g_{\rho^*, \eta^*} \leq g_{\rho^*, \eta} \qquad \forall \rho \in R, \eta \in U. \tag{10}$$

Proof. Due to an extension of a Theorem of Derman-Strauch to the denumerable state space ([19] Theorem 13.2), it is sufficient to show that (10) holds for Markovian policies only, i.e. $\forall \rho \in R(M), \eta \in U(M)$. Fix some $\rho = (\rho_1, ...) \in R(M)$. From the optimality equation (9) it follows that

$$c(\rho_t, \eta^*) - g^* e + P(\rho_t, \eta^*) w \leq w.$$

Iterating this we get

$$\sum_{t=1}^{T} P(\rho_1, \eta^*) ... P(\rho_{t-1}, \eta^*) c(\rho_t, \eta^*) - T g^* e + P(\rho_1, \eta^*) w \leq w,$$

where $P(\rho_1, \eta^*) ... P(\rho_{t-1}, \eta^*) = I$ for $t = 0$. Since $w \geq 0$ we conclude that

$$\overline{\lim_{T \to \infty}} \frac{1}{T} \sum_{t=1}^{T} P(\rho_1, \eta^*) ... P(\rho_{t-1}, \eta^*) c(\rho_t, \eta^*) \leq g^* e + \lim_{T \to \infty} \frac{w}{T}.$$

Hence $g_{\rho, \eta^*} \leq g^* e$. ∎

Proposition 4.2 *Assume A1, A2 and B2 hold. Then any limit policy-pair of $\xi(k)$-discounted optimal policies $(\rho^{\xi(k)}, \eta^{\xi(k)})$ (from Assumption A1) is optimal if $w < \infty$.*

Proof. Follows by combining Lemma 4.4 and Proposition 4.1. ∎

4.4 The average cost: verifiable sufficient conditions for conditions A and B

We first have the following sufficient conditions for A1:

Lemma 4.5 *[5] Assume that Assumptions 1 and $2(\xi)$ from Section 3 hold for all $\xi \in (0,1)$. Then Assumption A1 holds, and any stationary ξ-discounted optimal policies ρ^ξ and η^ξ satisfy the optimality equation (8), with finite v^ξ.*

Recall that Assumptions 1 and $2(\xi)$ are easily verifiable in queueing models, due to Lemma 4.3.

For Assumption A2 we have the following obvious sufficient condition:

Lemma 4.6 *If there are only finitely many actions for player 1 then A2 holds.*

Next we present an important sufficient conditions for B2. With the definitions as above A2, consider
Assumption D1: Assume A1; assume that there exists a class stationary policies for player 2, denoted by $\tilde{U}(S)$, such that for all $\eta \in \tilde{U}(S)$, the policy pair (ρ^*, η) is unichain (i.e. the Markov chain for any such policy pair does not have two or more closed sets), where ρ^* is as defined above A2; moreover, for all $\overline{\eta} \in U(M)$, there is a $\eta \in \tilde{U}(S)$ with

$$g_{\rho^*, \overline{\eta}} \geq g_{\rho^*, \eta}$$

Assumption D2: There is some (partial) order on the state space; for all $\eta \in \tilde{U}(S)$, there exist sequences $\{\rho^k\}, \{\eta^k\}$ of stationary policies with $\lim_{k\to\infty} \rho^k = \rho^*$ and $\lim_{k\to\infty} \eta^k = \eta$, such that for all $x \in \mathbf{X}$, $T \in \mathbb{N}$, $[P^T]_{x\rho^k\eta^k\bullet} \le [P^T]_{x\rho^*\eta\bullet}$ in the stochastic order corresponding to the partial order on the state (see definition of stochastic order in Ross [32]). We shall use the notation $P^\infty(\rho^k, \eta^k) \le_{st} P^\infty(\rho^*, \eta)$.
Assumption D3: The immediate cost is separable, i.e. $c(x, a, b) = c_1(a, b) + h(x)$; h is a monotone nondecreasing function in x and for any constant $q \in \mathbb{R}$, the set $\{y \in \mathbf{X} : h(y) < q\}$ is finite (this is known as a moment condition). Moreover, c_1 is bounded.
Assumption D4: $g_{\rho^k,\eta^k} \ge g^* e, \qquad \forall k \in \mathbb{N}$.

Lemma 4.7 *D (i.e. D1-D4) implies B2.*

Proof. For checking B2 we may restrict to $\eta \in \tilde{U}(S)$ (as defined in D1). Assume that for some x, $[P^\infty(\rho^*, \eta)h](x) = \infty$. It follows from the fact that c_1 is bounded together with Fatou's Lemma that $g_{\rho^*,\eta}(x) = \infty$, so that $g_{\rho^*,\eta} \ge g^* e$. Hence it suffices to consider some x for which $[P^\infty(\rho^*, \eta)h](x) < \infty$. This, together with the moment condition in D3 implies that $\sum_{y\in\mathbf{X}} P^\infty_{xy}(\rho^*, \eta) = 1$ (see e.g. [9]). Next we show that this implies that the family $\{P^\infty_{x\bullet}(\rho^k, \eta^k)\}_{k=1}^\infty$ is a tight family. Fix some $\epsilon > 0$, and let $q(\epsilon)$ be such that for all T,

$$\sum_{y\le q(\epsilon)} [P^T]_{xy}(\rho^*, \eta) \ge 1 - \epsilon.$$

Then by D2 it follows that

$$\sum_{y\le q(\epsilon)} [P^T]_{xy}(\rho^k, \eta^k) \ge \sum_{y\le q(\epsilon)} [P^T]_{xy}(\rho^*, \eta) \ge 1 - \epsilon$$

for any k, T. Hence $\{P^\infty_{x\bullet}(\rho^k, \eta^k)\}_{k=1}^\infty$ is tight. Suppose $\hat{P}^\infty$ is a limit measure of this family, obtained along some subsequence $k(n)$. Clearly it is a probability measure. Since

$$\sum_{y\in\mathbf{X}} P^\infty_{xy}(\rho^{k(n)}, \eta^{k(n)}) P^\infty_{yz}(\rho^{k(n)}, \eta^{k(n)}) = P^\infty_{xz}(\rho^{k(n)}, \eta^{k(n)}),$$

it follows from the bounded convergence theorem that

$$\hat{P}^\infty P(\rho^*, \eta) = \hat{P}^\infty.$$

Hence $\hat{P}^\infty$ is a matrix whose rows are the invariant measures of $P(\rho^*, \eta)$. Since $P(\rho^*, \eta)$ is unichain, it has at most one invariant measure (thus all rows are equal) and $P^\infty(\rho^*, \eta) = \hat{P}^\infty$. Since the subsequence $k(n)$ is arbitrarily chosen, we have

$$\lim_{k\to\infty} P^\infty(\rho^k, \eta^k) = P^\infty(\rho^*, \eta).$$

Since h is nondecreasing and $P^\infty(\rho^k, \eta^k) \le_{st} P^\infty(\rho^*, \eta)$ it follows that

$$\overline{\lim_{k\to\infty}} P^\infty(\rho^k, \eta^k)h \le P^\infty(\rho^*, \eta)h.$$

From Fatou's Lemma we have

$$\underline{\lim}_{k\to\infty} P^\infty(\rho^k,\eta^k)h \geq P^\infty(\rho^*,\eta)h.$$

Hence,

$$\lim_{k\to\infty} P^\infty(\rho^k,\eta^k)h = P^\infty(\rho^*,\eta)h.$$

Since c_1 is bounded, we also have

$$\lim_{k\to\infty} P^\infty(\rho^k,\eta^k)c_1 = P^\infty(\rho^*,\eta)c_1.$$

From $P^\infty(\rho^*,\eta)h < \infty$, it follows that

$$\lim_{t\to\infty} t^{-1}\sum_{s=1}^{t} P^s(\rho^*,\eta)h = P^\infty(\rho^*,\eta)h < \infty$$

and similarly with $P^\infty(\rho^k,\eta^k)$ (see [20]). Combining the last three equations, we have finally

$$g_{\rho^*,\eta} = \lim_{k\to\infty} g_{\rho^k,\eta^k} \geq g^* e.$$

■

Next we present sufficient conditions for Assumption D, and other sufficient assumptions for $B2$. It turns out that the approach by Weber and Stidham [38] is very useful. We collect the contribution needed later on in a Lemma which we call the Weber-Stidham Lemma; it is a direct adaptation from [38]. Define

$$X_\xi := \{x \in \mathbf{X} | v^\xi(x) = \inf_{y\in\mathbf{X}} v^\xi(y)\}.$$

We use again the notation introduced above Assumption A2.

Lemma 4.8 *Assume A1, and the existence of a stationary policy $\overline{\eta}$ for player 2 such that for all $0 < \xi < 1$, $g_{\rho^\xi,\overline{\eta}} < \infty$, and assume that the immediate cost function satisfies the following moment condition: for any finite q, the set $\{x \in \mathbf{X} : \inf_{a\in\mathbf{A}(x)} \inf_{b\in\mathbf{B}(x)} c(x,a,b) < q\}$ is finite. Then*
(i) X_ξ is a nonempty finite set, and for any $x \in X_\xi$,

$$\min_b c(\rho^\xi, b) \leq (1-\xi)V^\xi(x) \leq g_{\rho^\xi,\overline{\eta}}.$$

(ii) Moreover, suppose that for some infinite sequence $\xi(k) \to 1$, we have $\rho^{\xi(k)} = \rho^*$, and $g_{\rho^*,\overline{\eta}} < \infty$. Then for any $x \in X_\xi$,

$$(1-\xi)V^\xi(x) \leq g_{\rho^*,\eta}, \qquad \forall \eta \in U,$$

and B2 holds. If for some stationary η it is possible under (ρ^*,η) to go from any state x to any state y with finite expected cost then $w < \infty$ (componentwise), and (ρ^*,η^*) are expected average optimal policies.

Next we present a sufficient condition for assumption A2. It will also imply that $w^{\xi(k)}$ is finite in an appropriate norm, which we need later for developing further sufficient conditions for conditions D.

Assumption Q: Assumption 1 (from Section 3) holds; assume that there exists a stationary policy $\overline{\eta} \in U(S)$ for player 2 such that for some finite set $M \subset \mathbf{X}$ and constant $\zeta < 1$,

$$\sum_{y \in \mathbf{X}} {}_M P_{xy}(a, \overline{\eta}) \mu_y \leq \zeta \mu_x, \qquad \forall x \in \mathbf{X}, a \in \mathbf{A}(x).$$

We will refer to this condition as the μ-uniform geometric recurrence condition for player 1 with respect to policy $\overline{\eta}$ of player 2. Note that it is weaker than Assumptions 1 and 2(1) (from Section 3). In short, we call this assumption "weak stability condition".

Lemma 4.9 *Assume that Q holds. Then* $\sup_k \left\| w^{\xi(k)} \right\|_\mu < \infty$*, which implies A2.*

Proof. We use a standard argument that goes back to S. Ross in 1968 (see Section 12 of [19] and references therein); it is the observation that for $x_\xi \in X_\xi$, $w^\xi(x) = v^\xi(x) - v^\xi(x_\xi)$ is smaller than or equal to the total ξ-discounted cost incurred from state x at time 1 until the stochastic process reaches state x_ξ, when player 1 and player 2 use ρ^ξ and $\overline{\eta}$, respectively. The assertion now follows from Key Theorem 1 in [13]. ∎

We present next other sufficient conditions for Assumption $D4$. We use again the notation introduced above Assumption A2.

Lemma 4.10 *Assume that Assumptions 1 (from Section 3) and Assumption B1 hold. Assume for some* $\mu : \mathbf{X} \to [1, \infty)$*,* $\|w\|_\mu < \infty$*, and assume that for some stationary policy* η *there is a finite set* $M \subset \mathbf{X}$ *and a constant* $\zeta < 1$ *such that*

$$_M P(\rho^*, \eta)\mu \leq \zeta \mu. \tag{11}$$

Then $g_{\rho^*, \eta} \geq g^* e$*.*

Proof. With an argument as in the proof of Lemma 2.1 in [5], one shows that (11) implies

$$\bar{q} := sup_T \left\| P^T(\rho^*, \eta) \right\|_\mu < \infty$$

(see e.g. [23]). Iterating $w \leq c(\rho^*, \eta) - g^* e + P(\rho^*, \eta) w$, T times and dividing by T yields

$$\frac{1}{T} \sum_{t=0}^{T-1} P^t(\rho^*, \eta) c(\rho^*, \eta) \geq g^* e + \frac{w}{T} - \frac{P^T(\rho^*, \eta) w}{T}.$$

Also,

$$\lim_{T \to \infty} \frac{P^T(\rho^*, \eta) w}{T} \leq \lim_{T \to \infty} \frac{1}{T} \left\| P^T(\rho^*, \eta) \right\|_\mu \|w\|_\mu \to 0.$$

The assertion then follows by combining the last two inequalities. ∎

Proposition 4.3 *Assume that Assumptions 1 and 2(1) (from Section 3). Then Assumptions A and B hold.*

Proof. Assumptions A and B1 follow from [5]. B2 follows from Lemma 4.10. ∎

Remark 4.1 *Recently Sennott [33] obtained a set of conditions that imply Conditions B, if the immediate costs are bounded below. Although the conditions are obtained under the assumption of finite number of actions per state, they can be generalized in the setting of our paper to compact sets of actions and continuous immediate cost.*

4.5 Conclusions

In most queueing applications, even in unstable situations, Assumptions A1 hold by Lemma 4.5, whereas Assumption A2 follows from Lemma 4.6 or 4.9. Then condition B1 holds by Lemma 4.4, and it remains to check B2 in order to obtain Propositions 4.1 and 4.2. In many applications we shall have $\rho^{\xi(n)} = \rho^*$ for all n large enough. In that case B2 follows from natural assumptions on the immediate cost by the Weber-Stidham Lemma (Lemma 4.8). An alternative approach to establish B2 in some models is by verifying conditions Q, D. D1-D3 are quite often satisfied. We then show that for any stationary policy $\eta \in U(S)$, there exists a sequence of policies η^k such that (11) holds for any (ρ^*, η^k). This will imply D4, and hence, by Lemma 4.7, B2. Finally, when the above approaches do not work one can obtain conditions A1-A2 directly from Proposition 4.3.

5 A survey of worst-case control of queues

We now present a survey of worst-case control of queues, solved via stochastic games. Most of the properties on the structure of optimal policies below were obtained for the finite horizon case only. In this paper we extend them to the infinite horizon discounted and average cost, using the theory from the previous section.

5.1 Flow and service control

This first example, studied by Altman in [1] and [2], illustrates monotonicity properties in problems with a one dimensional state space. Since, in general, optimal policies in stochastic games require randomization, the monotonicity properties of policies in games are more refined and involved than those arising when only one controller exists (where no randomization is required). We shall show that a strong type of monotonicity arises, and that only a finite number of randomizations are needed.

Considered is a discrete-time single-server queue with a buffer of size $L \leq \infty$. We assume that at most one customer may join the system in a time slot. This possible arrival is assumed to occur at the beginning of the time slot. The state corresponds to the number of customers in the queue at the beginning of a time slot.

Let a_{min} and a_{max} be two real numbers satisfying $0 < a_{min} \leq a_{max} < 1$. At the end of the slot, if the queue is non-empty and if the action of the server is a, then a service of a customer is successfully completed with probability $a \in \mathbf{A}$ where $\mathbf{A}$ is a finite subset of $[a_{min}, a_{max}]$. If the service fails the customer remains in the queue, and if it succeeds then the customer leaves the system.

Let b_{min}, b_{max} be two real numbers satisfying $0 \leq b_{min} \leq b_{max} < 1$. At the beginning of each time slot, if the state is x then the flow controller chooses an action b from a finite set $\mathbf{B}(x) \subset [b_{min}, b_{max}]$ In this case, the probability of having one arrival during this time slot is equal to b. If the buffer is finite ($L < \infty$) we assume that $0 \in \mathbf{B}(x)$ for all x; moreover, when the buffer is full, no arrivals are possible ($\mathbf{B}(L) = \{0\}$). In all states other than L we assume that the available actions are the same, and we denote them by $\mathbf{B}(x) = \mathbf{B}$.

We assume that a customer that enters an empty system may leave the system (with probability a, when action a is used) at the end of this same time slot.

The transition law P is:

$$P_{xaby} := \begin{cases} \bar{b}a, & \text{if } L \geq x \geq 1,\ y = x - 1; \\ ba + \bar{b}\bar{a}, & \text{if } L \geq x \geq 1,\ y = x; \\ b\bar{a}, & \text{if } L > x \geq 0,\ y = x + 1; \\ 1 - b\bar{a}, & \text{if } y = x = 0; \end{cases}$$

(for any number $\chi \in [0, 1]$, $\bar{\chi} := 1 - \chi$).

We define an immediate payoff

$$c(x, a, b) := h(x) + \theta(a) + \zeta(b) \tag{12}$$

for all $x \in \mathbf{X}$, $a \in \mathbf{A}$ and $b \in \mathbf{B}$. We assume that $h(x)$ is a real-valued *increasing convex* function on $\mathbf{X}$ which is polynomially bounded, θ is a real function on $\mathbf{A}$ and ζ is a real function on $\mathbf{B}$. It is natural to assume that θ is increasing in a and $\theta \geq 0$ whereas ζ is decreasing in b and $\zeta \leq 0$. h can be interpreted as a function which gives the holding cost rates, ζ as a reward function related to the acceptance of incoming customer, and θ as a cost function per quality of service.

We thus consider a flow control problem, where the real controller is the one that chooses actions b, and it is playing in an unknown environment of service conditions, so that nature represents an imaginary service controller. We could however consider the opposite situation, where the real controller is the server, playing in an environment of unkown flow parameter b, which would then be represented as nature.

We now describe the type of monotonicity of optimal policies that will occur in the above problems. Let $u : \mathbf{X} \to \mathcal{M}(\mathbf{B})$. Denote $b_x^{sup}(u) :=$ the greatest b in the support of u_x, i.e. the greatest $b \in \mathbf{B}$ that is chosen by u with positive probability in state x. Denote $b_x^{inf}(u) :=$ the smallest b in the support of u_x. We say that a decision rule u_t at time t is **strongly monotone decreasing** if for any $x \in \mathbf{X}$ and y with $y < x$, $b_y^{inf}(u_t) \geq b_x^{sup}(u_t)$. The analogous definitions hold naturally for player 1 (nature). As a direct consequence from the definition of strongly monotone policies we have

Lemma 5.1 *If r is strongly monotone then it randomizes in at most $|\mathbf{A}| - 1$ states. If u is strongly monotone then it randomizes in at most $|\mathbf{B}| - 1$ states.*

The main result is

Theorem 5.1 *If the holding cost h is convex nondecreasing, and either $h(1) > h(0)$ or $h(2) - h(1) > h(1) - h(0)$, then*
(i) there exist strongly optimal Markov policies for both players for the finite horizon problem, which are strongly monotone decreasing for each t.
(ii) there exist strongly optimal stationary policies for both players for the discounted infinite horizon problem and for the average cost problem, which are strongly monotone decreasing.

Proof. The proof for the finite horizon and for the discounted cost can be found in [1, 2]. The main step consists of showing that $V^{t,\xi}$ and V^{ξ} exist and are convex nondecreasing. This is done by standard value iteration arguments. This is then shown in [2] to imply the strong monotonicity for both players when the holding cost is polynomially bounded. By Lemma 4.2 and Lemma 4.3, this can be generalized to some cases of exponential holding costs.

Finally, we consider the average cost problem. The case of a finite buffer is proved in [2].

The main difficulty for the infinite buffer is that under some pairs of policies the cost may be infinite; restricting to stationary policies we may have situations where the resulting state process is a transient or a null recurrent Markov chain. To handle this, we first consider the degenerate case $a_{min} \leq b_{min}$. Then nature can use in all states a_{min}, which is trivially a monotone policy. Whatever policy is used by the flow controller, X_t will tend in probability to infinity, so that the expected average cost is infinite. (This follows since X_t is stochastically minimized (for all t) if the flow controller uses b_{min} in all states. In that case X_t is a null recurrent (or transient) Markov chain. Note also that the holding cost tends to infinity as the number of customers grows to infinity.) Hence the above policy of nature is strongly optimal, and so is any policy for the flow controller.

It thus remains to consider the case $a_{min} > b_{min}$. We shall use Lemma 4.7, as suggested in the end of Subsection 4.5. This, together with the Lemmas 4.3, 4.5, 4.6 and 4.9 imply propositions 4.1 and 4.2, from which the result follows.

We first verify that Q holds with $\overline{\eta}(x) = b_{min}$ for all x, and $\mu(x) = z^x$ for some $z > 1$ (see [5, 23]). Hence from Lemma 4.9, $\|w\|_\mu < \infty$. Next we have to show that conditions D hold. We first note that the immediate cost satisfies the condition D3. In D1 we choose for $\tilde{U}(S)$ the class of all stationary policies. It follows from the condition $a_{min} > b_{min}$ that for all pairs of stationary policies (r, u), the Markov chain is unichain (possibly transient). D1 then follows from theorem 6.2 in [11]. We now define the policies ρ^k, η^k in conditions D2, D4. Let $\rho^k = \rho^*$ for all k. Choose some $\eta \in U(S)$ and define

$$\eta^k(x) = \begin{cases} \eta(x) & \text{if } x < k \\ \\ \delta_{b_{min}}(b) & \text{if } x \geq k \end{cases}$$

In other words, in all states greater than or equal to k, the policy η^k chooses action b_{min} w.p.1, and in other states it behaves like η. It follows that the assumptions of Lemma 4.10 hold for the policy η^k (note that M may depend on k), where again the μ function can be chosen to have the form $\mu(x) = z^x$ for some $z > 1$ (see [23]). By Lemma 4.10, this implies that D4 holds. It remains to check D2. We will show that $P^t(\rho^k, \eta^k) \leq_{st} P^t(\rho^{k+1}, \eta^{k+1})$, for all integers t, which clearly implies D2. We use a coupling argument for that purpose. We shall consider a stochastic process X corresponding to policy η^k, and a stochastic process $\overline{X}$ corresponding to policy η^{k+1} starting from the same given initial state. The evolution equations for the systems are

$$X_{t+1} = (X_t + \beta_t - \alpha_t)^+, \qquad \overline{X}_{t+1} = (\overline{X}_t + \overline{\beta}_t - \overline{\alpha}_t)^+,$$

where $\beta_t = 1$ if an arrival occurs, and $\alpha_t = 1$ if a (potential) service occurs in system X at time t ($\overline{\alpha}_t$ and $\overline{\beta}_t$ are similarly defined). We assume in fact that $X_1 \leq \overline{X}_1$, which is certainly satisfied if both processes have the same initial state, and we show recursively by coupling that $X_t \leq \overline{X}_t$ for all t. Assume that $X_t \leq \overline{X}_t$. Clearly, if $X_t + 2 \leq \overline{X}_t$ then $X_{t+1} \leq \overline{X}_{t+1}$. It remains to consider $X_t = \overline{X}_t$, and $X_t + 1 = \overline{X}_t$.

It is possible now to couple the systems so that the following holds. If at time t $X_t = \overline{X}_t$, then $\alpha_t = \overline{\alpha}_t$. Moreover, if $\eta^k(X_t) = \eta^{k+1}(\overline{X}_t)$, then $\beta_t = \overline{\beta}_t$; now if $\eta^k(X_t) \neq \eta^{k+1}(\overline{X}_t)$ (which means that $X_t = k$) then $\overline{\beta}_t = 0$ implies $\beta_t = 0$ (since X uses the lowest input probability b_{min} at state k). It remains to check $X_t + 1 = \overline{X}_t$. Since ρ^* is a limit of strongly monotone policies (that are optimal for the discounted cost), it turns out that it is itself strongly monotone decreasing, and hence we may couple the systems in a way such that if $\overline{\alpha}_t = 1$ then $\alpha_t = 1$ (since the probability of a successful service decreases with the state). Consequently, $X_{t+1} \leq \overline{X}_{t+1}$ for all t. This implies finally condition D2, which establishes conditions D. ∎

For the case of an infinite buffer analogous results as obtained in Theorem 5.1 hold for the dual problem of service control, where nature is the flow controller. The decreasing monotonicity of the policies of both players is replaced by increasing monotonicity. The case with a finite buffer seems however more involved, and remains open. The proof of the results for the infinite buffer are basically the same, where condition D is obtained by a similar coupling argument as above.

A continuous time version of the above service control problem was considered in [5]. The service controller chooses a rate of service, instead of a probability of service success, and the flow controller chooses the rate of arrival. Optimal monotone policies were shown to exist for both players, that do not require randomization. The fact that no randomization is required is due to the fact that the stochastic game has an AR-AT structure (i.e. Additive Reward, Additive Transition Games, see [30]). This is typical to continuous time stochastic games, as illustrated als in the following sections.

5.2 Routing games: introduction

We now present some results on routing problems. In all problems, there are several queues with exponential service and finite or infinite capacity. The decisions of the router are immediately after an arrival to the system occurs, and specify to which queue the arriving customer is routed. Typical statements on the structure of optimal policies are: (1) join the shortest queue (SQP) [8], in a symmetrical setting, and, in the nonsymmetrical case, a characterization of the type: do not join a queue if there is another shorter queue whose server is faster (SQFSP) [6]. (2) monotonicity of the policies [3]. We consider three scenarios where the unknown parameters of the systems are modelled by different kinds of controls of nature. All models are continuous time Markov games. Restricting to state dependent policies, we may employ a standard uniformization procedure [27] and formulate the problems as (equivalent) discrete time stochastic games. We present below three different routing models where the router plays against some unknown dynamically varying parameter. We then present a dual problem: where the server plays against an unknown routing scheme.

5.3 A routing game against an extra service capacity, optimality of SQP

We assume a Poisson arrival process with rate λ. There are N parallel queues, each queue $1 \leq i \leq N$ having a buffer of length $L_i \leq \infty$. Denote $L = (L_1, \ldots, L_N)$. At each arrival the router takes a decision $b \in \{1, ..., N\}$ to which queue the arrival should be routed to. If all queues are full the customer is lost. The server in each queue is assumed to be exponential. The service rate of each server is known to be at least μ (for some $\mu \geq 0$). Moreover, it is known that the sum of the service rates is $\sum_{i=1}^{N} \mu_i = N\mu + \Delta$, with Δ a positive constant. Subject to these constraints, the μ_i's may change in time in a way unpredictable by the router. One may assume that each queue has a basic service rate μ, and extra service rates a_i, $1 \leq i \leq N$, with $\sum_{i=1}^{N} a_i = \Delta$ are allocated to the queues.

Let the state of the system be the number of customers in the different queues (including in service) just prior to a transition. The cost is assumed to be a function of the state only, i.e. $c(x, a, b) = h(x)$.

Using uniformization, the system is observed at time instants which form a Poisson process with rate $\vartheta = \lambda + N\mu + \Delta$. The probability that a transition corresponds to an arrival, a successful uncontrolled service in queue j, or a successful controlled service in queue j are λ/ϑ, μ/ϑ and a_j/ϑ respectively. Let $\mathcal{A}_j y$ and $\mathcal{D}_j y$ be the state obtained by an arrival to queue j and by a departure (possibly dummy) from queue j, when the state was previously y. The transition probabilities are

$$P_{xaby} = \begin{cases} \lambda/\vartheta & \text{if } \ y = \mathcal{A}_b x \leq L \ \text{ or } \ x = y = L\,, \\ (\mu + a_i)/\vartheta & \text{if } \ y = \mathcal{D}_i x\,. \end{cases}$$

We introduce the following properties of a function $f : \mathbf{X} \to \mathbb{R}$.
C1 (monotonicity) $f(z) \leq f(y)$ for $z \leq y$,
C2 (symmetry) $f(y) = f(\pi y)$ for any permutation πy of y such that $\pi y \in \mathbf{X}$.
C3 $f(\mathcal{A}_k y) \leq f(\mathcal{A}_l y)$ for y such that $y_k \leq y_l$, $\mathcal{A}_k \mathcal{A}_l y \leq L$.

Theorem 5.2 *Assume that the holding cost h satisfies* **C1,C2,C3**. *Consider a routing policy u^o that always joins a (non-full) shortest queue. Consider a server's policy r^o that always serves a shortest queue. Then (r^o, u^o) are optimal for the three follosin cases for finite and infinite capacity queues (i) finite time horizon, (ii) infinite time horizon and discounted cost, provided that the immediate cost satisfies (7), (iii) expected average cost, provided that the immediate cost are polynomially bounded.*

Proof. [8] established the case of finite capacity queues. The results for the infinite horizon discounted cost are obtained by using Lemmas 4.1, 4.2, 4.3. For the expected average cost we use the Weber-Stidham Lemma, in particular Lemma 4.8 (ii), since the policy $\rho^* = r^o$ of nature is optimal for all discount factors. Note that we also need to check that w is finite, which follows from Lemma 4.8 (ii), or simpler, from Lemma 4.9. ∎

5.4 MDAP and routing: a game against worst-case arrival rate, optimality of SQFSP

In all examples until now, there was an interplay between an input (flow or routing) and an output (service) controller. The next model is of a game between two aspects of the input process. The real controller is the router, whereas the nature controls the varying rate of arrivals from previous nodes. This setting is typical for worst-case control in the last node of a network, and it allows the arrival process to depend not only on some independent random environment, as is the case in MAP (Markov Arrival Process)) and MMPP (Markov Modulated Poisson Process), but also on the whole state of the system, including in particular the state of the last node. We use the formalism of MDAP (Markov Decision Arrival Process) introduced in [21, 22] for the MDP context, and generalized in [6] to the stochastic game context. The MDAP generalizes the MAP by allowing the transition rates and arrival probabilities to be controlled dynamically, i.e. the transition rates and arrival probabilities depend on the actions that are sequentially chosen by a controller. In this way, the transition rates of the MDAP are dependent on the state of the system where the customers arrive, through the actions chosen by the controller. A typical example of the use of an MDAP in the routing context is the following. Customers arrive according to a Poisson process at m parallel $M|M|1$ queues. The customers have to be assigned to one of the queues by a dynamic policy. After being served the customers arrive at a second station, where we have again m parallel queues to choose from. The question is how to assign the customers at the second center (for example in the case that all service parameters are equal), assuming that the the objective of the router in the first center is unknown to the controller in the second one. In general, the optimal action in the first center will not only depend on the state of the first center, but also on the state of the second one. Thus, the arrival process at the second center depends, through the actions in the first center, on the state of the second center. Therefore we cannot use the standard results on the optimality of shortest queue routing for independent arrivals.

We thus consider the following generic stochastic game. The state space is given by a product of two spaces: the state space of the MDAP $\mathbf{X}_1$, assumed to be finite, and the state space of the queueing system $\mathbf{L} = \prod_{j=1}^{N} \mathbf{L}_j$ where $\mathbf{L}_j = \{0, 1, ..., L_j\}$, and $L_j > 0$ is the capacity of queue j (which may again be either finite or infinite). Let $L = (L_1, ..., L_N)$ be the vector of queue capacities. A typical element of the state space is denoted by $\underline{x} = (x, i)$, with $x \in \mathbf{X}_1$ and $i = (i_1, \dots, i_m) \in \mathbf{L}$ the number of customers in the m queues *including* the ones in service.

The probability of a successful service in queue j is μ_j. Without loss of generality we assume that $\mu_1 \geq ... \geq \mu_m$.

The finite space of actions of the MDAP (player 1) is $\mathbf{A}$ (different actions may be available in different states). λ_{xay} is the probability that the MDAP moves from x to y if action a was chosen by player 1, and q_{xay} is the probability that a customer arrives if the arrival process moves from x to y under action a. We assume without loss of generality that for any (x, i), $\sum_y \lambda_{xay} + \sum_{j=1}^{m} \mu_j = 1$. The actions for the router are as in Subsection 5.3. The transition probabilities are thus

$$P_{(x,i),a,b,(y,k)} = \begin{cases} \lambda_{xay} q_{xay} & \text{if } k = \mathcal{A}_b(i), \\ \lambda_{xay}(1 - q_{xay}) + \sum_{j=1}^{m} \mu_j 1\{i_j = 0, y = x\} & \text{if } k = i, \\ \mu_j & \text{if } y = x, k = \mathcal{D}_j i, i_j > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(We assume that the rates are already normalized so that for any (x, i) we have $\sum_{y,k} P_{(x,i),a,b,(y,k)} =$

1.) If all queues are full then

$$P_{(x,i),a,b,(y,k)} = \begin{cases} \lambda_{xay} & \text{if } k = i, \\ \mu_j & \text{if } y = x, k = \mathcal{D}_j i, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that player 2 takes an action immediately after an arrival occurs, (hence after a transition in the MDAP occurs) already knowing the new state of the MDAP. A precise description of the decision process (transition probabilities, and the state space) is given in [25] Section 5.1. We use below notations from Subsection 5.3.

We assume here an immediate cost which is only a holding cost, i.e. $c(\underline{x}, a, b) = h(x, i)$ for $\underline{x} = (x, i)$, and we assume that it satisfies the following properties:

$$h(x, \mathcal{A}_{j_1} i) \le h(x, \mathcal{A}_{j_2} i) \text{ if } i_{j_1} \le i_{j_2},\ j_1 \le j_2,\ \mathcal{A}_{j_1}\mathcal{A}_{j_2} i \le L \tag{13}$$

$$h(x, i) \le h(x, \mathcal{A}_j i) \text{ if } \mathcal{A}_j i \le L \tag{14}$$

$$h(x,i) \le h(x, i^*) \text{ if } \begin{cases} \max(i_{j_1}, i_{j_2}) \le \max(L_{j_1}, L_{j_2}), \\ i_{j_1} > i_{j_2},\ j_1 \le j_2 \end{cases} \text{ and } i^*_j = \begin{cases} i_j \text{ if } j \ne j_1, j_2, \\ i_{j_2} \text{ if } j = j_1, \\ i_{j_1} \text{ if } j = j_2. \end{cases} \tag{15}$$

In particular, we may choose $h(x, i) = \sum_{j=1}^N i_j$.

For the finite horizon problem, Altman and Koole [6] established the existence of a strongly optimal policy for the router of the SQFSP type. The proof is based on using iteratively the following dynamic programming equation to show that $V^{t,\xi}$ possesses properties (13), (14) and (15), from which the required structure is obtained. If the system is not full then the dynamic programming equation is

$$\begin{aligned} V^{t+1,\xi}(x,i) &= h(x,i) + \xi \max_a \Big\{ \sum_y \lambda_{xay} \Big(q_{xay} \min_b \{ V^{t,\xi}(y, \mathcal{A}_b i) \} + (1 - q_{xay}) V^{t,\xi}(y, i) \Big) \Big\} \\ &\quad + \xi \sum_{j=1}^m \mu_j V^{t,\xi}(x, \mathcal{D}_j i). \end{aligned}$$

When the system is full then the dynamic programming equation is

$$V^{t+1,\xi}(x,i) = h(x,i) + \xi \max_a \Big\{ \sum_y \lambda_{xay} V^{t,\xi}(y,i) \Big\} + \xi \sum_{r=1}^m \mu_r V^{t,\xi}(x, \mathcal{D}_r i).$$

We note that the decisions of the two players are taken in different time instants (this is known as a game with complete information, see [30]). The router takes a decision right after an arrival occurs, whereas the the MDAP controller takes actions after a departure, or after a customer has joined one of the queues. This implies that no randomization is required by either player (as the game is then of perfect information, see [14]). In fact, it can be seen directly from the dynamic programming equations that the minimizer has dominant actions that do not depend on nature's choice, and hence can restrict to pure actions. Thus nature does not benefit either by randomizing.

We call the above problem (P0). We consider next a slight variation ([6] Section 3) of (P0). The dynamics are unchanged, but the information structure is a little different. Player 2 takes an

action immediately after an arrival occurs; however, due to information delay, it does not have the knowledge of the new state of the MDAP. As a result, we may consider this action to have been taken already prior to the arrival (since no new information is obtained by player 2 in the arrival epoch). We shall thus assume that the decision instants for the players are the same; each time a transition occurs (departure, or a transition in the MDAP), both players take a decision. The decision of player 2 should be interpreted however as the action to be taken when there will be a future arrival. We further consider two versions of that game, depending on whether or not the information on the action of player 1 is delayed too.
(P1) When a customer arrives then player 2 already has the information on the last action of player 1. Hence, at each decision epoch, player 1 takes a decision first and only then player 2 takes a decision, knowing the decision of player 1.
(P2) When a customer arrives then player 2 does not yet have the information on the last action of player 1. Hence, at each decision epoch, the players take their actions independently.

To summarize, the information available to each player at a given decision epoch consists of all previous states and actions of both players, as well as the current state of the system. Moreover, in (P1), at any time t, player 2 has the information on the decision of player 1 at time t. Problem (P1) is known as a stochastic game with complete information. It is known that for these games there exist optimal policies which do not require randomizations (for both players), whereas in (P2) randomized policies are usually needed to obtain optimality. Since the action of player 2 is interpreted as the decision to be taken when a future arrival occurs, the knowledge of the current state indeed grasps the fact that information is delayed, and thus when that arrival will occur and the MDAP will change its state, the new state will not be available to player 2.

Since the amount of information that player 2 possesses in (P1) when making a decision is less than in (P2), and that is less than the information he has in (P0) (of the previous section), the value v^n will satisfy

$$V^{t,\xi}_{(P0)} \leq V^{t,\xi}_{(P1)} \leq V^{t,\xi}_{(P2)}.$$

For the finite horizon problem, Altman and Koole [6] established the existence of a strongly optimal policy for the router of the SQFSP type for problems (P1) and (P2). The proof is again based on using iteratively the dynamic programming equation to show that $V^{t,\xi}$ satisfies properties (13), (14) and (15), from which the required structure is obtained. The dynamic programming equation for the two problems are as follows. If the system is not full then

$$\begin{aligned}
(P1):\ V^{t+1,\xi}(x,i) &= h(x,i) + \xi \max_a \min_b \Big\{ \sum_y \lambda_{xay} \Big(q_{xay} V^{t,\xi}(y, \mathcal{A}_b i) + (1-q_{xay}) V^{t,\xi}(y,i) \Big) \Big\} \\
&\quad + \xi \sum_{j=1}^m \mu_j V^{t,\xi}(x, \mathcal{D}_j i), \\
(P2):\ V^{t+1,\xi}(x,i) &= h(x,i) + \xi \mathrm{val} \Big\{ \sum_y \lambda_{xay} \Big(q_{xay} V^{t,\xi}(y, \mathcal{A}_b i) + (1-q_{xay}) V^{t,\xi}(y,i) \Big) \Big\} \\
&\quad + \xi \sum_{j=1}^m \mu_j V^{t,\xi}(x, \mathcal{D}_j i).
\end{aligned}$$

When the system is full then we have for both problems:

$$V^{t+1,\xi}(x,i) = h(x,i) + \xi \max_a \Big\{ \sum_y \lambda_{xay} V^{t,\xi}(y,i) \Big\} + \xi \sum_{r=1}^m \mu_r V^{t,\xi}(x, \mathcal{D}_r i).$$

For the infinite horizon discounted cost optimal SQFSP policies exist for all three problems, (P0), (P1), and (P2), provided that the immediate cost satisfies (7). This follows from Lemma 4.2, 4.3. Let λ_{sup} be the supremum of the expected average rate under all policies of the MDAP that are stationary, and depend only on $\mathbf{X}_1$. Then $\lambda_{sup} < \sum_{i=1}^{M} \mu_i$ seems to be a sufficient condition for the SQFSP structure of optimal policies to carry over to the expected average cost.

5.5 Monotonicity results in routing games: unknown service rates

Consider two infinite queues. Customers arrive to the system according to a Poisson process with rate λ. The service duration of a customer in queue i is exponentially distributed with a parameter $\mu(i)$ that lies in the interval $\mathbf{A}_i = [\underline{\mu}(i), \overline{\mu}(i)]$. This parameter, may change in time in a way unknown to the router and is assumed to be controlled by nature. The state space is $\mathbf{X} = \mathbb{N}^2$ where $\mathbb{N}$ are the natural numbers. We have $\mathbf{A} = \mathbf{A}(x) = \mathbf{A}_1 \times \mathbf{A}_2$ for all $x \in \mathbf{X}$. The transition probabilities are

$$P_{xaby} = \begin{cases} \lambda, & \text{if } b = i, y = \mathcal{A}_i x, i = 1,2; \\ \mu(i), & \text{if } y \neq x, y = \mathcal{D}_i x, i = 1,2; \\ 1 - (\lambda + \sum_{i=1}^{2} \mu(i) 1\{x_i > 0\}), & \text{if } y = x \end{cases}$$

where $a = (\mu(1), \mu(2))$. We assume again that the rates are normalized so that $\lambda + \overline{\mu}(1) + \overline{\mu}(2) \leq 1$. The immediate cost $c(x, a, b)$ is separable, and has the form:

$$c(x,a,b) = h(x) + \sum_{i=1}^{2} \theta_i(\mu(i)) + \sum_{i=1}^{2} \zeta_i 1\{b = i\}.$$

It is composed of a holding cost h, a cost θ_i that depends on the quality of the service at queue i, and an admission cost ζ_i if a customer is to be admitted to queue i. We assume that θ_i (and thus c) are continuous in the actions.

In this example we have again the property of AR-AT (Additive Reward, Additive Transition Games, see [30]) which implies that pure optimal policies exist for both players.

Define for all $x \in \mathbf{X}$ $V^{0,\xi}(0) = 0$, and denote

$$R_i^t(\mu, x) = \mu(i)[V^{t,\xi}(\mathcal{D}_i x) - V^{t,\xi}(x)] + \theta_i(\mu(i)) + \overline{\mu}(i) V^{t,\xi}(x), \quad \mu_i \in \mathbf{A}_i$$

$$S^t(i,x) = V^{t,\xi}(\mathcal{A}_i x) + \zeta_i, \quad i = 1,2.$$

for $t = 0, 1, 2, ..., T$. The dynamic programming equation for the finite horizon cost has the form:

$$\begin{aligned} V^{t+1,\xi}(x) &= h(x) + \xi \sum_{i=1,2} \max_{\mu(i) \in \mathbf{A}_i} R_i^t(\mu(i), x) + \xi \lambda \min_{b=1,2} S^t(b,x) \qquad (16) \\ &\quad + \xi(1 - \lambda - \overline{\mu}(1) - \overline{\mu}(2)) V^{t,\xi}(x), \end{aligned}$$

for $t = 0, ..., T-1$.

We say that a decision rule is of the monotone switching curve type (see [16]) if it has the following monotonicity property. It is described by a curve in $\mathbf{X}$ with a monotone slope, that separates $\mathbf{X}$ into two connected regions, $\mathbf{X}_1$ and $\mathbf{X}_2$ and there exists two actions, $j = 1, 2$, such that in region $\mathbf{X}_j$ it is optimal to use action j, $j = 1, 2$.

We describe below the sufficient conditions from [3] which guarantees that the router has an optimal nondecreasing switching curve policy, i.e., if it is optimal to route a customer to queue 1 for a given length of the queues $x = (x_1, x_2)$, then it is also optimal to route the customer to queue 1 when the length of the queues is $y = (y_1, y_2)$ provided that $y_1 \leq x_1$ and $y_2 \geq x_2$. (This implies indeed that the curve that separates $\mathbf{X}_1$ and $\mathbf{X}_2$ is nondecreasing). A similar monotonicity property holds also for routing to queue 2.

Consider the following property of a function $z : \mathbf{X} \to \mathbb{R}$:
$\mathbf{\Pi_1}$: $\quad z(\mathcal{A}_i^2 x) - z(\mathcal{A}_i \mathcal{A}_j x) \geq z(\mathcal{A}_i x) - z(\mathcal{A}_j x), \qquad i, j = 1, 2,\ i \neq j.$
If $V^{t,\xi}$ satisfies $\mathbf{\Pi_1}$, then there exists an action for the minimizer in (16) which has the above monotone switching curve structure. Next consider the property of a function $z : \mathbf{X} \to \mathbb{R}$:
$\mathbf{\Pi_2}$: $\quad z(\mathcal{A}_i \mathcal{A}_j x) - z(\mathcal{A}_j x) \geq z(\mathcal{A}_i x) - z(x), \qquad i, j = 1, 2.$
If $V^{t,\xi}$ satisfies $\mathbf{\Pi_2}$, then there exist actions for both maximizers in (16) which are monotone nonincreasing in x. Moreover, if θ_i is convex for some $i = 1, 2$, then server i has an action achieving the maximum in (16) which is an element of $\{\underline{\mu}(i), \overline{\mu}(i)\}$ (bang-bang policies); combining these we have that server i has a nonincreasing monotone switching curve structure (this means in particular that if at some state x server i uses $\underline{\mu}(i)$, then it also uses $\underline{\mu}(i)$ for all states $y \geq x$, componentwise).

In [3], Altman shows using standard value iteration, that indeed $V^{t,\xi}$ satisfies $\mathbf{\Pi_1}$ and $\mathbf{\Pi_2}$ (using arguments similar to [16]), provided that $h(x)$ does, and that it is monotone nondecreasing. This is shown in [3] to imply that there exist optimal policies with the above structure for the finite time horizon case. These results carry over to the infinite time horizon discounted cost provided that the immediate cost satisfies (7), as follows from Lemma 4.2, 4.3. The optimal policies for player 1 and 2 (having the above structure) can be chosen as maximizer and minimizer (respectively) of the dynamic programming equation

$$\begin{aligned} V^\xi(x) \quad &= \quad h(x) + \xi \sum_{i=1,2} \max_{\mu(i) \in \mathbf{A}_i} R_i(\mu(i), x) + \xi \lambda \min_{b=1,2} S(b, x) \\ &\quad + \xi(1 - \lambda - \overline{\mu}(1) - \overline{\mu}(2)) V^\xi(x), \end{aligned} \tag{17}$$

where R_i and S are defined as above Eq. (16), with V^ξ replacing $V^{t,\xi}$. (The value is the unique function, bounded in the appropriate norm introduced in Lemma 4.3, that solves the dynamic programming equation).

Next we discuss the infinite time horizon expected average cost. Assume that h is polynomially bounded. The structural results carry over the expected average case under the following strong stability assumption

$$\lambda < \min\{\underline{\mu}(1), \underline{\mu}(2)\}, \tag{18}$$

by Proposition 4.3, by choosing an exponential μ norm (see [23]). If $\lambda \geq \underline{\mu}(1) + \underline{\mu}(2)$, then nature has a strongly monotone policy of using $\underline{\mu}(1)$ and $\underline{\mu}(2)$ in all states, for which the expected average cost can easily be shown to be infinite under any policy of the controller. It remains thus to consider the following weak stability condition

$$\lambda < \underline{\mu}(1) + \underline{\mu}(2).$$

In order to prove that the monotone structure still holds, we need the following Lemma.

Lemma 5.2 *Assume that h satisfies $\mathbf{\Pi_1}$ and $\mathbf{\Pi_2}$, and that $h(x) \to \infty$ as $|x| = x_1 + x_2$ goes to* *infinity. Then for any $x \in \mathbf{X}$, and $k = 1, 2$,*

$$\lim_{\xi \to 1} \lim_{x_1 + x_2 \to \infty} [V^\xi(\mathcal{A}_k x) - V^\xi(x)] = \infty.$$

Proof. The proof uses an idea from [5] for a worst-case service control problem in a single queue. Fix some discount factor $\xi < 1$. For any function real function f on $\mathbf{X}$, define $\Delta_k f(x) = f(\mathcal{A}_k x) - f(x)$. We assume for simplicity that

$$\inf_{x \in \mathbf{X}, k=1,2} \Delta_k h(x) \geq \overline{c} \tag{19}$$

for some constant $\overline{c} > 0$. Note that a sufficient condition for (19) is that $h(0,1) > h(0,0)$ and $h(1,0) > h(0,0)$. Fix an arbitrary $x \in \mathbf{X}$. Denote by $\tilde{\mu}(i)$ some action of nature which achieves the max in (17) for that x, and denote by $\tilde{b}$ some action of the router which achieves the min (of $S(b,x)$) there. From (17), we have for $k = 1, 2$,

$$\begin{aligned}
\Delta_k V^\xi(x) &\geq \overline{c} + \xi \sum_{i=1,2} \Delta_k R_i(\tilde{\mu}(i), x) + \xi \lambda \Delta_k S(\tilde{b}, x) \\
&= \overline{c} + \xi \sum_{i=1,2} \tilde{\mu}(i) \left[\Delta_k V^\xi(\mathcal{D}_i x) - \Delta_k V^\xi(x)\right] + \xi(1-\lambda)\Delta_k V^\xi(x) + \xi\lambda\Delta_k V^\xi(\mathcal{A}_{\tilde{b}} x) \\
&= \overline{c} + \xi\lambda \left[\Delta_k V^\xi(\mathcal{A}_{\tilde{b}} x) - \Delta_k V^\xi(x)\right] + \xi \left(1 - \sum_{i=1,2} \tilde{\mu}(i)\right) \Delta_k V^\xi(x) \\
&\quad + \xi \sum_{i=1,2} \tilde{\mu}(i) \left[\Delta_k V^\xi(\mathcal{D}_i x)\right] \\
&\geq \overline{c} + \xi \left(1 - \sum_{i=1,2} \tilde{\mu}(i)\right) \Delta_k V^\xi(x) + \xi \sum_{i=1,2} \tilde{\mu}(i) \left[\Delta_k V^\xi(\mathcal{D}_i x)\right]
\end{aligned} \tag{20}$$

where we used property $\mathbf{\Pi_2}$. Assume that $x_1 = 0$. Then $V^\xi(x) = V^\xi(\mathcal{D}_1 x)$, and

$$\begin{aligned}
(20) &= \overline{c} + \xi\,(1 - \tilde{\mu}(2))\,[\Delta_k V^\xi(x) - \Delta_k V^\xi(\mathcal{D}_2 x)] + \xi \Delta_k V^\xi(\mathcal{D}_2 x) \\
&\geq \overline{c} + \xi \Delta_k V^\xi(\mathcal{D}_2 x)
\end{aligned} \tag{21}$$

where we used property $\mathbf{\Pi_2}$. Similarly, if $x_2 = 0$ then

$$(20) \geq \overline{c} + \xi \Delta_k V^\xi(\mathcal{D}_1 x). \tag{22}$$

Assume $x_1 > 0, x_2 > 0$. Then

$$\begin{aligned}
(20) &\geq \overline{c} + \xi \left(1 - \sum_{i=1,2} \tilde{\mu}(i)\right) \Delta_k V^\xi(x) + \xi \sum_{i=1,2} \tilde{\mu}(i) \min_{j=1,2} \left[\Delta_k V^\xi(\mathcal{D}_j x)\right] \\
&= \overline{c} + \xi \left(1 - \sum_{i=1,2} \tilde{\mu}(i)\right) \left\{\Delta_k V^\xi(x) - \min_{j=1,2} \left[\Delta_k V^\xi(\mathcal{D}_j x)\right]\right\} \\
&\quad + \xi \min_{j=1,2} \left[\Delta_k V^\xi(\mathcal{D}_j x)\right] \\
&\geq \overline{c} + \xi \min_{j=1,2} \left[\Delta_k V^\xi(\mathcal{D}_j x)\right]
\end{aligned} \tag{23}$$

where we used property $\mathbf{\Pi_2}$. Define

$$Z_k(n) = \min_{\{x:x_1+x_2=n\}} \Delta_k V^\xi(x).$$

By combining (21)-(23) we conclude that

$$Z_k(n) \geq \overline{c} + \xi Z_k(n-1)$$

for all $n > 0$ and $k = 1, 2$. This implies that

$$Z_k(n) \geq \overline{c}\left(1 + \xi + ... + \xi^{n-1}\right)$$

The Lemma now follows by the definition of Z_k. ∎

From the above Lemma, from the dynamic programming equation (17), and from the fact that conservative policies in (17) are ξ-optimal, we obtain:

Corollary 5.1 *There exists some integer N_0, such that for all ξ close enough to 1, there is a pure stationary optimal policy ρ^ξ for player 1, that uses $(\underline{\mu}(1), \underline{\mu}(2))$ in all states x with $x_1 + x_2 > N_0$.*

As a result of the above corollary, there is only a finite number of distinct (monotone) policies ρ^ξ, for all ξ close enough to 1. One may thus choose some sequence $\xi(k)$ converging to 1 so that $\rho^{\xi(k)}$ are the same, which we denote by ρ^*. It now follows from the Weber-Stidham Lemma, in particular Lemma 4.8 (ii), that B2 holds. The policy $\overline{\eta}$ can be chosen as one that with probability z_1 routes an arrival to queue 1, and with probability $z_2 = 1 - z_1$ to queue 2; where z_1 is chosen such that $\lambda_i z_i < \underline{\mu}_i, i = 1, 2$. The finiteness of $g_{\rho^*,\overline{\eta}}$ can be established with the use of Lemma 4.9 (using the fact that h is assumed to be polynomially bounded). Hence the structural properties established for the infinite horizon discounted cost carry over to the infinite horizon average cost.

5.6 The dual game: service control versus unknown routing

We use the same model as in Subsection 5.5, except that the roles are changed: the "good guy" is the server control, playing against an unknown routing scheme, represented by "nature". The only change in the dynamic programming is that the min and max are interchanged. One can show that similar structure as before still holds.

The router (now nature) has again an "optimal" monotone nondecreasing switching curve policy with the following structure. If it routes a customer to queue 1 for a given length of the queues $x = (x_1, x_2)$, then it also routes an arriving customer to queue 1 when the length of the queues is $y = (y_1, y_2)$ provided that $y_1 \geq x_1$ and $y_2 \leq x_2$. This implies that a similar monotonicity property holds also for routing to queue 2. The servers have optimal monotone policies, where higher service rates are used in higher states.

By slight a modification of the proof in [3], the above structure is again obtained for the finite time horizon cost provided that the holding cost h is nondecreasing and that it satisfies $\mathbf{\Pi_1}$ and $\mathbf{\Pi_2}$. Again, by assuming (7), Lemma 4.2 and 4.3 imply that the structure carries over to the infinite horizon discounted cost.

Assume next that h is polynomially bounded. Using Proposition 4.3, it can be shown similarly to [3], that if the strong stability condition (18) holds then the monotonicity properties also hold for the expected average cost. We assume (without too much loss of generality) that $h(x) \to \infty$ as $x_1 + x_2 \to \infty$, and we assume that $\lambda \geq \min(\overline{\mu}(1), \overline{\mu}(2))$. Let $i^* \in argmin\overline{\mu}(i)$, then nature has the obvious monotone optimal policy of routing always to queue i^*, which results, for any server's policy, in an infinite average cost (by arguments similar to those in Subsection 5.1). Hence this is a degenerate case for which we have trivially monotone optimal policies for both players. It remains thus to consider the case $\lambda < \min(\overline{\mu}(1), \overline{\mu}(2))$, which is analyzed in [29].

5.7 Server assignment model: a game against the arrival process

We describe in this subsection the model of [6] Section 4. We consider N infinite capacity queues served by a single server. Customers in queue j have an exponential service time distribution with parameter μ_j. Let $\mu = \max_i \mu_i$. A decision $b \in \mathbf{B} = \{1, ..., N\}$ of Player 2 has the interpretation that the server will be assigned to queue b. The game is obtained by considering the general dependent arrival process described in Subsection 5.4, described as an MDAP. The states of the MDAP evolve according to the controlled transition rates λ_{xay}, and q^j_{xay} is the probability of an arrival to queue j when a transition from state x to y occurs under an action a. We consider a cost $c((x,i),a,b) = h(i) = \sum_{j=1}^N c_j i_j$ where i_j are the number of customers in queue i, and c_j are nonnegative constants. The dynamic programming equation has the form:

$$\begin{aligned} V^{t+1,\xi}(x,i) &= h(i) + \xi \max_a \Big\{ \sum_y \lambda_{xay} \Big(\sum_{j=1}^m q^j_{xay} V^{t,\xi}(y, \mathcal{A}_j i) + (1 - \sum_{j=1}^m q^j_{xay}) V^{t,\xi}(y,i) \Big) \Big\} \\ &\quad + \xi \min_b \Big\{ \mu_b V^{t,\xi}(x, \mathcal{D}_b i) + (\mu - \mu_b) V^{t,\xi}(x,i) \Big\}. \end{aligned}$$

For independent arrivals, the μc-rule is known to be optimal for the server (serving the nonempty queue j if and only if it has the largest product $\mu_j c_j$). Reorder the queues such that $\mu_1 c_1 \geq \cdots \geq \mu_m c_m$. For arrivals according to an MDAP the extra condition $\mu_1 \leq \cdots \mu_m$ was needed in [22]. In the present setting, where maximizing actions are chosen in the MDAP, we have to assume $\mu_1 \geq \cdots \geq \mu_m$ instead. Under that condition, Altman and Koole show in [6] Section 4 that the μc rule is strongly optimal for the server for the finite time horizon case. This result is extended in [6] for the case of several servers, provided that $\mu_1 = \cdots = \mu_m$.

Since the immediate cost is linear, it satisfies (7), and so Lemma 4.2 and 4.3 imply that this structure carries over to the infinite time horizon discounted cost. We conjecture that this structure also holds in the case of the average cost.

Acknowledgement The authors wish to thank the referees and associate editor for their helpful comments and suggestions. This research has been supported through visitor's grants from INRIA, NWO (Netherlands Organization for Scientific Research) and the Thomas Stieltjes Institute for Mathematics. The authors are greatful for the hospitality at INRIA and the Mathematical Institute of Leiden University during their visits.

References

[1] E. Altman, "Flow control using the theory of zero-sum Markov games", *IEEE Trans. Automatic Control*, pp. 814-818, 1994.

[2] E. Altman, "Monotonicity of optimal policies in a zero sum game: a flow control model", *Advances of dynamic games and applications*, pp. 269-286, 1994.

[3] E. Altman, "A Markov game approach for optimal routing into a queueing network", INRIA report No. 2178, submitted, 1994.

[4] E. Altman, "Non zero-sum stochastic games in admission, service and routing control in queueing systems", submitted to *QUESTA*.

[5] E. Altman, A. Hordijk and F. M. Spieksma, "Contraction conditions for average and α-discounted optimality in countable state Markov games with unbounded rewards", submitted to *MOR*, 1994.

[6] E. Altman and G. Koole, "Stochastic Scheduling Games with Markov Decision Arrival Processes", *Journal Computers and Mathematics with Appl.*, 3rd special issue on Differential Games, pp. 141-148, 1993.

[7] E. Altman and N. Shimkin, "Individually Optimal Dynamic Routing in a Processor Sharing System: Stochastic Game Analysis", EE Pub No. 849, August 1992. Submitted to *Operations Research*.

[8] E. Altman and N. Shimkin, "Worst-case and Nash routing policies in parallel queues with uncertain service allocations", IMA Preprint Series No. 1120, Institute for Mathematics and Applications, University of Minnesota, Minneapolis, USA, 1993, submitted to *Operations Research*.

[9] V. Borkar, "Control of Markov chains with long-run average cost criterion", *Proc., Stochastic Differential Systems*, Fleming and Lions, (eds.) pp. 57-77, Springer Verlag 1986.

[10] V. Borkar and M. K. Ghosh, "Denumerable state stochastic games with limiting average payoff", *JOTA*, pp. 539-560, 1993.

[11] R. Cavazos-Cadena, "Recent results on conditions for the existence of average optimal stationary policies", *Annals of Operations Research* **28**, special issue on "Markov Decision Processes", Eds. O. Hernandez-Lerma and J. B. Lasserre, 1991.

[12] R. Dekker and A. Hordijk, "Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards", *Mathematics of Operations Research*, **13**, pp. 395-421, 1988.

[13] R. Dekker, A. Hordijk and F. M. Spieksma, "On the relation between recurrence and ergodicity properties in denumerable Markov decision chains", *Math. Operat. Res* **19**, pp. 539-559, 1994.

[14] D. Gillette, "Stochastic games with zero stop probabilities", *Contribution to the Theory of Games*, III, M. Dresher, A. W. Tucker, P. Wolfe, eds., Princeton University Press, Princeton, 1957, pp. 179-187.

[15] A. Glazer and R. Hassin, "Stable priority purchasing in queues", *Operations Research Letters*, **6**, pp. 285-288, 1986.

[16] B. Hajek, "Optimal control of two interacting service stations", *IEEE Trans. Automatic Control*, **29**. No. 6, pp. 491-499, 1984.

[17] R. Hassin and M. Haviv, "Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying", *Commun. Statist. - Stochastic Models*, **10**, pp. 415-435, 1994.

[18] M. Haviv, "Stable strategies for processor sharing systems", *European J. of Operations Research* **52**, pp. 103-106, 1991.

[19] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Second Edition, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1977.

[20] A. Hordijk and P. J. Holewijn, "On the convergence of moments in stationary Markov chains", *Stoch. Proc. Appl.*, **3**, pp. 55-64, 1975.

[21] A. Hordijk and G. Koole, "On the assignment of customers to parallel queues", *Probability in the Engineering and Informational Sciences* **6**, pp. 495-511, 1992.

[22] A. Hordijk and G. Koole, "On the optimality of LEPT and μc rules for parallel processors and dependent arrival processes", *Advances in Applied Probability*, **25**, pp. 979-997, 1993.

[23] A. Hordijk and F. M. Spieksma, "On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network", *Advances in Applied Probability*, **24**, pp. 343-376, 1992.

[24] M. T. Hsiao and A. A. Lazar, "A game theoretic approach to decentralized flow control of Markovian queueing Networks", *Performance '87*, Courtois & Latouche (eds.), pp. 55-73, 1988.

[25] G. Koole, *Stochastic scheduling and dynamic programming*, Ph.D. thesis, Leiden University, 1992. (On request available from the author)

[26] Y. A. Korilis and A. Lazar, "On the Existence of Equilibria in Noncooperative Optimal Flow Control", 1994. To appear in the *Journal of the ACM.*

[27] S. A. Lippman, "Applying a new device in the optimization of exponential queueing systems", *Opns. Res.* **23**, pp. 687–710, 1975.

[28] A. S. Nowak, "On zero-sum stochastic games with general state space I". Prob and Math Statistics, Vol. IV, Fasc. 1, pp. 13-32, 1984.

[29] O. Passchier, "Optimal service control against worst case admission policies", preprint, 1995.

[30] T. E. S. Raghavan and J. A. Filar, "Algorithms for Stochastic Games - A survey", *ZOR*, **35**, pp. 437-472, 1991.

[31] U. Rieder, "Non-Cooperative Dynamic Games with General Utility Functions", *Stochastic Games and related topics*, T.E.S. Raghavan et al (eds), pp. 161-174, Kluwer Academic Publishers, 1991.

[32] S. M. Ross, *Stochastic Processes*, John Wiley, New York, 1983.

[33] L. I. Sennott, "Zero-sum stochastic games with unbounded costs: discounted and average cost cases", *ZOR*, **40**, pp. 145-162, 1994.

[34] F. M. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, 1990, Leiden University (available on request from the author).

[35] S. Stidham, "Optimal control of admission, routing, and service in queues and networks of queues: a tutorial review", *Proceedings ARO Workshop: Analytic and Computational Issues in Logistics R and D*, George Washington University, pp. 330-377, 1984.

[36] S. Stidham, "Optimal Control of Admission to a Queueing System," *IEEE Trans. Aut. Contr.*, **30**, pp. 705-713, 1985.

[37] J. Walrand, *An Introduction to Queueing Networks*, Prentice Hall, Cliffs, NJ, 1988.

[38] R. R. Weber and S. Stidham, "Optimal control of service rates in networks of queues", *Advances in Applied Probability*, **19**, pp. 202-218, 1987.

[39] J. Wessels, "Markov Games with unbounded rewards", *Dynamische Optimierung*, M. Schäl (editor) Bonner Mathematische Schriften, Nr. 98, Bonn, 1977.