Infinite Horizon Analysis of a Hospital Admission Control Model

Jonathan E. Helm
Mark P. Van Oyen

Technical Report 10-01
May 28, 2010

University of Michigan
Industrial and Operations Engineering
Ann Arbor, Michigan 48109

Infinite Horizon Analysis of a Hospital Admission Control Model

Jonathan E. Helm and Mark P. Van Oyen
Department of Industrial and Operations Engineering
University of Michigan

May 31, 2010

1 Introduction

The purpose of this paper is to show that the infinite horizon average cost optimality equation that governs the dynamic control of hospital admissions presented in Helm et al. (2010) can be analyzed by instead analyzing the finite horizon discounted version of the problem. Please refer to that paper for the model formulation with $B \in \{1, 2, 3, \ldots\}$ beds.

The average cost optimality equation is presented below for the reader's reference:

$$Z^\pi = \limsup_{T\to\infty} E\frac{1}{T}\left[\int_0^T \big(h_1'(B - X_1^\pi(z))^+ + \tau'(X_1^\pi(z) - B)^+ + h_2' X_2^\pi(z)\big)dz + c' N^\pi(T)\right] . \tag{1}$$

A policy is then optimal if it achieves the optimal cost among all admissible policies Π as follows:

$$Z^* = \inf_{\pi\in\Pi} Z^\pi. \tag{2}$$

The rates that govern the stochastic processes $X_1^\pi(t)$ and $X_2^\pi(t)$ are the emergency patient arrival rate λ_e, the elective patient arrival rate λ_s, the call-in patient arrival rate λ_q and a discharge rate of μ, where there are B servers (beds) to serve patients.

The costs are the cancelation cost c, the cost per unit time of an *empty* bed h_1, the cost per unit time of holding a patient on the call-in queue h_2, and the cost of having too many people in the hospital τ (which is assessed for every patient in the hospital that exceeds the bed capacity B). The set of admissible policies, Π, is the set of policies defined by the following controls on admission to the hospital.

1. Elective admissions arriving at a rate of λ_s can be
   (a) rejected (canceled), or
   (b) admitted.
2. Call-in patients arriving at a rate of λ_q can either be

(a) admitted to the hospital, or

(b) placed on a call-in queue for later admission to the hospital.

3. Patients from the hospital are discharged at a rate of μ. When a discharge occurs, either

(a) the discharged patient can be backfilled by a patient from the call-in queue, so there is no change in hospital census, but the number in the call-in queue decreases by 1, or

(b) the bed can be left empty for future arrivals.

To show that the results from a finite-horizon discounted problem also hold for the optimality equation above, it is sufficient to show that the finite horizon discounted problem converges both in policy and in cost as the length of the horizon goes to infinity. The following theorem formalizes this concept, and it's proof is the subject of this paper.

Theorem 1.1. *For the average cost optimality equation the following hold.*

(i) There exists an average-cost optimal stationary policy.

(ii) The optimal average cost can be computed by: $Z^* = \inf_{\pi\in\Pi} Z^\pi = \lim_{\beta\to 1^-} \lim_{n\to\infty(1-\beta)} V_{n,\beta}(x)$, *where* $V_{n,\beta}(x)$ *is the n stage discounted value function.*

(iii) Let $\pi_{n,\beta}$, denote an optimal policy for the n-period (discounted cost) problem. Then any limit point π_β of the sequence $\{\pi_{\beta,n}\}_{n\geq 1}$ as $n\to\infty$ is optimal for the infinite-horizon discounted cost. Moreover, any limit point of the sequence $\{\pi_\beta\}_{\beta\in(0,1)}$ (as $\beta\to 1^-$) is average-cost optimal.

Sennott (1999) gives us sufficient conditions, called the SEN assumptions, for this theorem to hold. This paper presents in detail the arguments to show that

1. The SEN assumptions hold our MDP model, and

2. The SEN assumptions guarantee that Theorem 1.1 holds.

This paper is organized as follows. In Section 2 we show that the SEN assumptions hold for our model. In Section 3 we show how the SEN assumptions guarantee that Theorem 1.1 holds.

2 The SEN Assumptions

In this section we show that the SEN assumptions hold for our model. To do so, we draw upon a second set of assumptions, the CAV assumptions Sennott (1999), that are sufficient for the SEN assumptions to hold. We begin this section by defining important definitions that are needed for the CAV assumptions to hold in Section

2.1. After defining the necessary terminology taken from Sennott (1999), we present the CAV assumptions and prove that they hold for the hospital admission model, a controlled CTMC, defined by Equation 2, in Section 2.2.

2.1 Terminology

Definition 1. *$R(i, \mathcal{G})$ is the set of policies, θ, satisfying $P_\theta(X_n \in \mathcal{G} \quad \text{for some } n \geq 1 | X_0 = i) = 1$ and the expected time $m_{i\mathcal{G}}$ of a first passage from i to $\mathcal{G}$ is finite (Sennott p. 139).*

Definition 2. *$R^*(i, \mathcal{G})$ is the class of policies $\theta \in R(i, \mathcal{G})$ such that the expected cost $c_{i\mathcal{G}}(\theta)$ of a first passage from i to $\mathcal{G}$ is finite (Sennott p. 140).*

Definition 3. *$c_{i\mathcal{G}}(\theta)$ the expected cost of a first passage from i to $\mathcal{G}$ (Sennott p. 139).*

Definition 4. *$m_{i\mathcal{G}}$ the expected time of a first passage from i to $\mathcal{G}$ (Sennott p. 140).*

Definition 5. *Let d be a stationary policy. Then d is a* z standard policy *if the MC induced by d is z standard.*

Definition 6. *A MC is* z standard *if there exists a distinguished state z such that $m_{iz} < \infty$ and $c_{iz} < \infty$ for all $i \in S$.*

In these definitions the capital letter, $\mathcal{G}$, refers to a set of states whereas the lowercase z refers to a single in the final 2 definitions. While this is a small abuse of notation, this is the notation used by Sennott (1999) and the meaning should be clear that $m_{i\mathcal{G}} = m_{iz} \quad \forall z \in \mathcal{G}$.

2.2 Proof of the CAV Assumptions for a Hospital Admission Control Model

The CAV assumptions defined below provide a convenient set of conditions that guarantee that the SEN assumptions hold. The CAV assumptions are often used instead of directly applying the SEN assumptions because the CAV assumptions are easier to verify in many queueing systems. In this section we first present the CAV assumptions and then we prove that they hold for our hospital admission control model.

Definition 7. *CAV Assumptions*

(CAV1) *There exists a z standard policy d with positive recurrent class R_d.*

(CAV2) *Let $C(i, a)$ be the one stage cost associated with being in state i and taking action a. Given $U > 0$ the set $D_U = \{i | C(i, a) \leq U \quad \text{for some } a\}$ is finite.*

(CAV3) *Given $i \in S - R_d$, there exists a policy $\theta_i \in R^*(z, i)$.*

Before proving that the CAV assumptions hold for our model, we first define a policy, d, that can be used to show that CAV1 holds and a set of policies $\{d'_i\colon\ i \in S - R_d\}$ that can be used to show that CAV3 holds. After describing the properties of the Markov chain induced by these policies, we then define a stochastic process,

$x_m(n)$, based on the Markov chain induced by policy d and analyze the properties of this process that are used in the proof of CAV1. A similar process can be defined for policies d_i', but we omit the full exposition of CAV3 because the steps are very similar to those used to prove CAV1.

Consider the policy, d, that (1) cancels all arriving scheduled patients and (2) always admits call-in patients to the hospital when they arrive and (3) always admits a patient from the call-in queue, if there is at least one, when there is a system departure or when there are 0 full beds in the hospital. After ignoring the transient states in which $x_2 > 0$ the MDP reduces to a Birth Death process in x_1 with $\lambda = \lambda_e + \lambda_q$ in all states and $\mu_n = x_1\mu$ if $x_1 < B$ and $\mu_n = B\mu$ if $x_1 \geq B$. This policy induces the chain described in Figure 1.

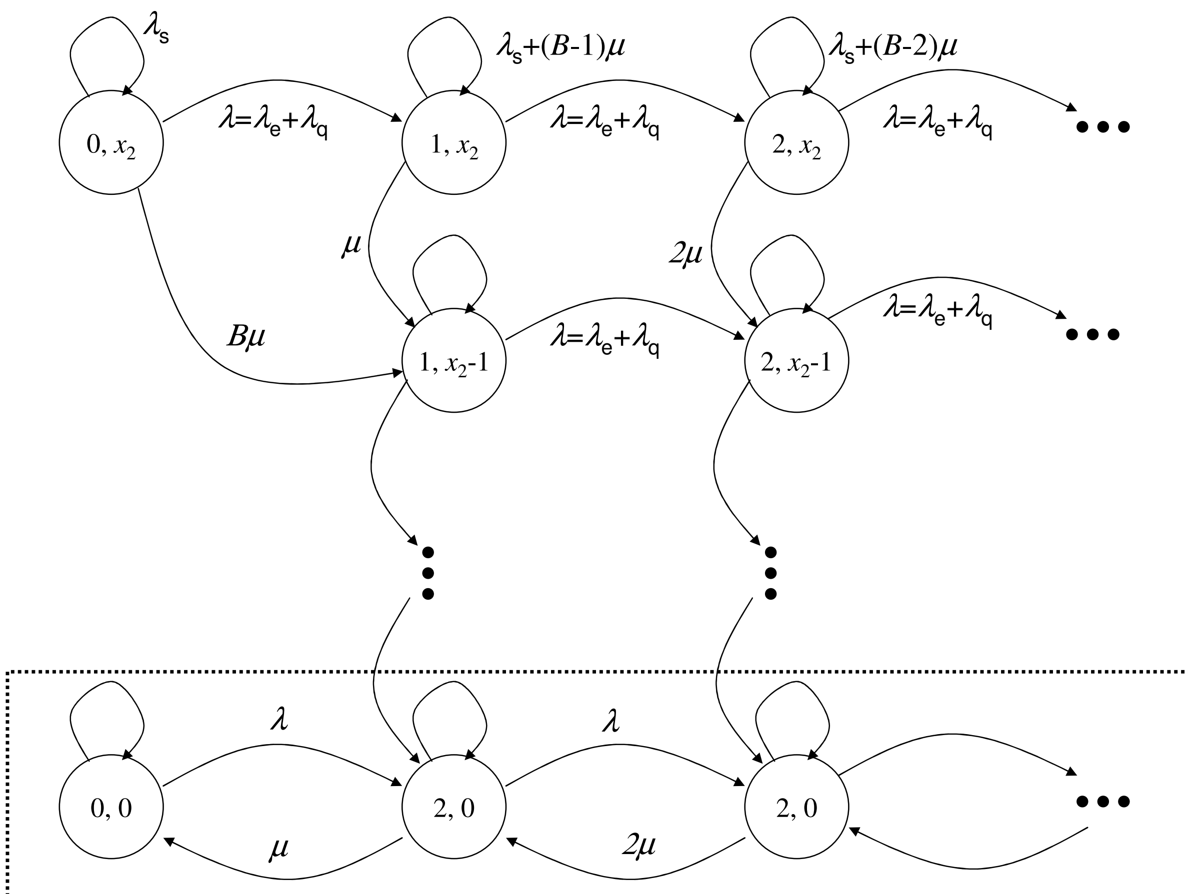

Figure 1: Markov chain induced by policy d. States in the box compose the only recurrence class, R_d.

The policy d_i' depends on the policy d and the z for which d induces a z-standard Markov chain. In our proof of CAV1, even though there are many states that could satisfy the role of z, we let $z = (0, 0)$ W.L.O.G. The policy d_i' is defined for states $i \in S - R_d$, for the policy d. Such states can be characterized by $\{(x, y)$ s.t. $y > 0, x \in \mathbb{R}\}$. For a given state $i = (x, y) \in S - R_d$ consider the policy d_i' that (1) when the number in the call-in queue is less than y, the controller places all arriving call-in patients on the call-in queue and does not call in any patients upon service completion (2) if the number in the call-in queue equals y the controller admits arriving call-in patients directly into the hospital but does not call in patients upon service completion and (3) if the number in the call-in queue is greater than y the controller admits arriving call-in patients directly into the hospital and calls in patients upon service completion. Regardless of state, the controller always cancels elective arrivals. This policy is depicted in Figure 2

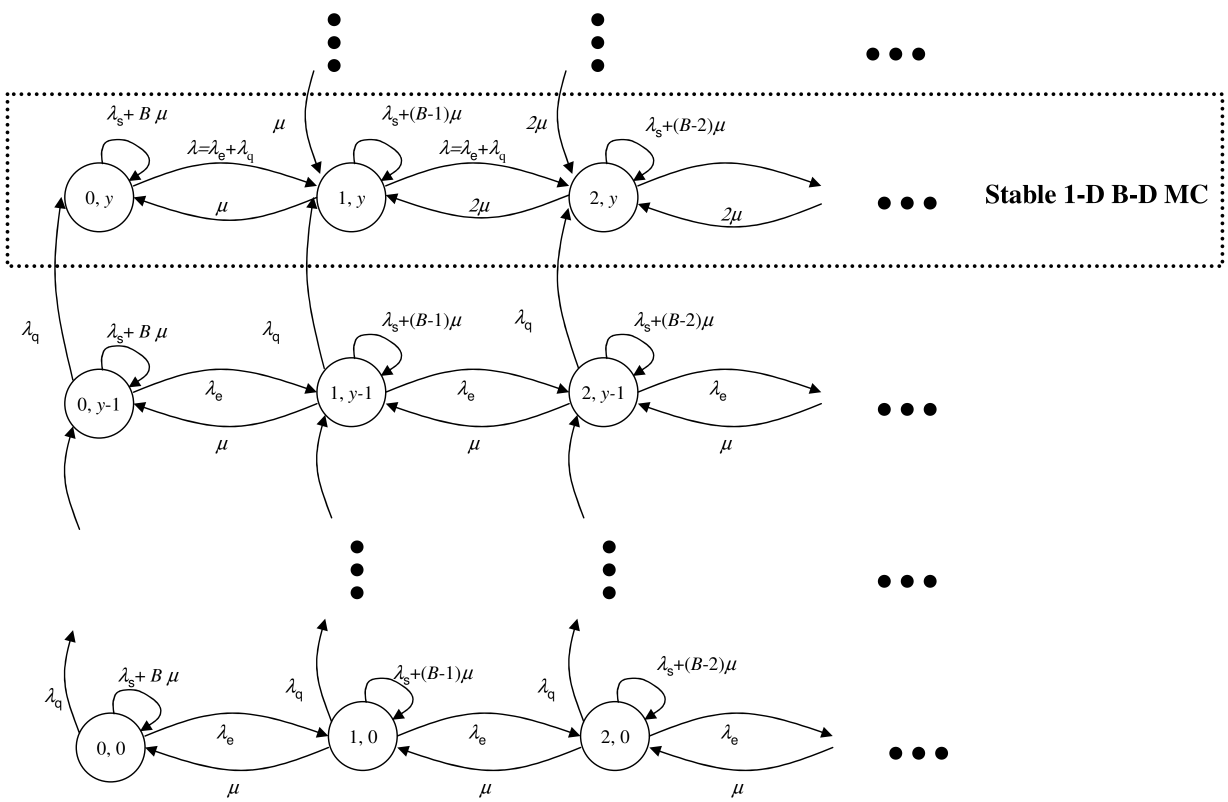

Figure 2: Markov chain induced by policy d'_y where $y \in S - R_d$. States in the box compose the only recurrence class, R_d.

The final step before proceeding to the proof of the CAV assumptions is to define a stochastic process $x_m(n)$ that will help us in analyzing the finiteness of the expected first passage time and expected cost of a first passage between various states and $z = (0, 0)$ for CAV1. An analogous process can be defined to aid the analysis of CAV3, but as mentioned we abbreviate this analysis because it follows very similar arguments to that of CAV1. In words $x_m(n)$ is a non-decreasing sequence (which can be seen from Figure 1) of the number of patients in the hospital after each downward transition (decrease) in call-in patients.

Definition 8. *$x_m(n)$ is the number of patients in the hospital when the number of call-in patients in the hospital transitions from $n + 1$ to n, given that there are initially m patients in the hospital.*

Note that in the definition, m remains fixed and n varies along the set $n = i_2, i_2 - 1, \ldots, 1, 0$ where i_2 is the number of call-in patients in the call-in queue when we initialize the process. This is a recursive process which requires an initial state. Suppose the initial state is $i = (i_1, i_2)$, then we initialize the stochastic process as $x_{i_1}(i_2) = i_1$ and we consider the entire process to be the set $\mathbf{x}(i_1, i_2) = \{x_{i_1}(i_2), x_{i_1}(i_2 - 1), \ldots, x_{i_1}(1), x_{i_1}(0)\}$. In order to prove CAV1, we will use the fact that the expectation of this stochastic process is finite. In other words, the expected number of patients in the hospital at each "row" (i.e., keeping a fixed length of the call-in queue) of the Markov chain induced by d is finite. This is formalized by the following lemma.

Lemma 2.1. *For a given state $i = (i_1, i_2) \in S - R_d$, where there are i_2 patients in the call-in queue and i_1 patients in the hospital, we have that $E[x_{i_1}(n)] < \infty \quad \forall n \leq i_2$.*

Proof. It suffices to consider only states where $i_1 \geq B$ (and therefore $x_{i_1}(n) \geq B \quad \forall n \leq i_2$). Some simple intuition is enough to see why this is sufficient. Consider any starting state $j = (j_1, j_2)$ s.t. $j_1 < B$. It is easy to see that $E[x_{j_1}(j_2)] \leq E[x_B(j_2)] \leq E[x_{i_1}(j_2)]$ for $i_1 \geq B$ when one considers the fact that the Markov chain only transitions to states with more patients in the hospital (see Figure 1) until it reaches the recurrent class R_d (which is the stopping condition for our process $x_m(n)$). So if $E[x_{i_1}(j_2)] < \infty$ for $i_1 \geq B$ so is $E[x_{j_1}(j_2)]$ for $j_2 < B$. Thus W.L.O.G. we proceed under the assumption that $i_1 \geq B$, where i_1 is the initial number of patients in the hospital.

Assuming $i_1 \geq B$ we let Z be the random variable for the number of patients admitted to the hospital in between two service completions. Since $i_1 \geq B$ and the chain only transitions to increasing levels of hospital census, the arrival and discharge parameters remain constant at λ and $B\mu$ respectively. Thus in all feasible states, the distribution on Z is the same for all call-in queue levels $n = i_2, i_2 - 1, \ldots, 1$ (see Figure 1). It is easy to show that Z is the modified Geometric($\frac{B\mu}{\lambda + B\mu}$), with $E[Z] = \frac{\lambda}{B\mu}$:

$$P(Z = k) = \Big(\frac{\lambda}{\lambda + B\mu}\Big)^k \cdot \Big(\frac{B\mu}{\lambda + B\mu}\Big), \quad k = 0, 1, 2, \ldots$$

Using Z we can define $x_1(n)$ using the recursive equation

$$x_{i_1}(n) = Z + x_{i_1}(n + 1). \tag{3}$$

To show that $E[x_{i_1}(n)] < \infty$ we rely on the recursive definition in Equation 3 and the base case $E[x_{i_1}(i_2)] = i_1$ to get

$$E[x_{i_1}(i_2 - 1)] = E[x_{i_1}(i_2) + Z] = E[x_{i_1}(i_2)] + E[Z] = i_1 + \frac{\lambda}{B\mu} < \infty.$$

Similarly $E[x_{i_1}(n)]$ can be show to be

$$\begin{aligned} E[x_{i_1}(n)] &= E[x_{i_1}(n + 1) + Z] = E[x_{i_1}(n + 2) + Z + Z] = \cdots = i_1 + (i_2 - n)E[Z] \\ &= i_1 + (i_2 - n)\frac{\lambda}{B\mu} < \infty. \end{aligned}$$

□

Now we are ready to prove the CAV assumptions.

Lemma 2.2. *CAV1 holds for our model.*

Proof. To prove CAV1, we need to show that policy d induces a positive recurrent class R_d and that d is z-standard.

We first show that (I) the Markov chain induced by d has a positive recurrent class R_d. Then we show that (II) for any state, i, $m_{iz} < \infty$, where $z = (0,0)$ and we want to show that the Markov chain induced by d is z-standard. Finally we show that (III) for any state, i, $c_{iz} < \infty$ where z is as in (II). These 3 conditions satisfy the requirements for d to be a z-standard policy with positive recurrent class R_d, and thus satisfy CAV1.

(I) We will show that the Markov chain associated with the one-dimensional Birth-Death process, composed of the states in the box in Figure 1, is ergodic in x_1, with x_2 fixed at 0. The recurrence class associated with this policy, d, is $R_d = \{(x_1, 0)|x_1 \geq 0\}$. All subsequent arguments refer to the one-dimensional birth death Markov chain highlighted by the box in Figure 1, whose states compose the recurrence class R_d. It is obvious that in this birth-death chain, all states communicate. Therefore the condition for the birth death process defined by policy d to be positive recurrent is $\exists N$ such that for all $n > N$ we have that $\lambda_n/\mu_n < \infty$ Kleinrock (1975). Therefore R_d is indeed positive recurrent, since $\lambda/(B\mu) < 1$.

(II) Now we need to show that for any state, i, the first passage time to $z \in R_d$ is finite; $m_{iz} < \infty$. We begin by showing this for states $i \in R_d$. For $i \in R_d$ the first passage time from $i \to z$ where $i = (x, 0), z = (y, 0), \quad x, y \in \mathbb{R}$, is $m_{iz} < \infty$ as shown in Jouini and Dallery (2008) since the MC induced by policy d is a stable birth-death process. Therefore, W.L.O.G., let us show the system is a z standard MC where $z = (0, 0)$.

Now we consider any state, i, not in recurrence class R_d, i.e. $i \in \{(i_1, i_2) \text{ s.t. } i_2 > 0\}$. Since i is a transient state, the amount of time spent in i is finite almost surely and the process enters the recurrence class in a finite amount of time with probability 1. In fact the expected time to enter the recurrence class is simply the expected time of i_2 service completions, since the state will decrease by $e_2 = (0, 1)$ after every service completion. In other words, after each service completion, the call-in queue will decrease by 1 and the number of patients in the hospital will remain the same (see Figure 1).

For $i_1 > 0$ this expected time is $\frac{i_2}{\mu}$, since it is simply the expected time to have i_2 service completions. Note that, when $i_1 > 0$, there will always be at least one patient in the hospital while the system is in a transient state. This can be seen in the diagram of the induced Markov chain, Figure 1. In this diagram, in all transient states the only system transitions are (1) self loops, (2) transition to a state with 1 more patient in the hospital, or (3) a transition to a state with one fewer patient on the call-in queue. Thus once $i_1 > 0$, we will have at least i_2 service completions before reaching the state with zero patients in the hospital in which no service completions occur.

If $i_1 = 0$, there is no one receiving service in the hospital, but a patient is called in at a rate of $B\mu$ and a patient arrives to the hospital with rate $\lambda = \lambda_e + \lambda_q$. Thus a patient enters the hospital with rate $\lambda + B\mu$. The expected time to reach 0 patients in the call-in queue given $i_1 = 0$ is then

$$\frac{1}{\lambda + B\mu} + \frac{\lambda}{\lambda + B\mu} \cdot \frac{i_2}{\mu} + \frac{B\mu}{\lambda + B\mu}\Big(\frac{i_2 - 1}{\mu}\Big) < \infty. \tag{4}$$

.

As noted previously, the key is that the number of patients in the hospital never decreases while the system resides in the transient states because every departure is backfilled by a patient from the call-in queue. The equation above can be understood as follows. First note that when there are zero patients in the hospital, the system transitions out of this state with a rate of $\lambda + B\mu$ according to the competing exponentials of an arrival (call-in λ_q or emergency λ_e) and a call-in ($B\mu$). Therefore the first term in the equation above represents the time for either an arrival to the hospital or a call in from the call-in queue to occur. The second term represents the expected time to empty the call-in queue if the event is an arrival to the hospital multiplied by the probability that the event triggering the initial transition was an arrival to the hospital. Note that once the arrival occurs, there process must now complete i_2 services successive, thereby admitting the i_2 patients on the call-in queue to the hospital and emptying the call-in queue. The third term represents the expected time to empty the call-in queue if the triggering event is a call in from the call-in queue multiplied by the probability that the triggering event is a call-in from the call-in queue. In this case, the triggering event itself causes the call-in queue to decrease by 1 and so to empty the call-in queue the process now needs only $i_2 - 1$ service completions to admit the remaining $i_2 - 1$ patients on the call-in queue to the hospital.

We have shown that the expected amount of time for the Markov chain to enter the recurrent class R_d is finite, but now we must consider *where* the process enters R_d to be sure that the remaining time to reach $z = (0, 0)$ is indeed finite in expectation. To do so we rely on the stochastic process $\mathbf{x}(i_1, i_2)$. In particular, the element from this process that we are interested in is $x_{i_1}(0)$, which is precisely the random variable that describes the state, $w(i_1) = \big(x_{i_1}(0), 0\big)$, at which the Markov chain initially enters R_d. To show that the time to reach $z = (0, 0)$ from $i \in S - R_d$ it remains to show that $E[m_{w(i_1)z}] < \infty$, where the expectation is taken over the random variable $x_{i_1}(0)$. To show this relationship, we first present a formula for m_{xz} in terms of x. Then we take the expectation of this formula, replacing x with $w(i_1) = \big(x_{i_1}(0), 0\big)$.

Recall that once the Markov chain enters R_d, it becomes a one dimensional birth-death process, which means we can rely on standard theory of stable birth-death processes. To this end it can be shown that $m_{xz} = \sum_{n=1}^{x} \overline{\theta}_n$ where

$$\overline{\theta}_n = \frac{1}{\lambda_{n-1}\pi_{n-1}} \sum_{m=n}^{\infty} \pi_m$$

is the expected cost to go from n patients in the hospital to $n-1$ patients in the hospital, from Jouini and Dallery (2008). In the above equation, π_n represents the equilibrium probability that there are n patients in the hospital. It is well known that the equilibrium distribution for this birth-death process can be written as

$$\pi_n = \begin{cases} \pi_0\left[\frac{\lambda^n}{\mu(n-1)!}\right] & \text{if } n < B \\ \pi_0\left[\frac{\lambda^{B-1}}{\mu(B-1)!}\cdot\left(\frac{\lambda}{B\mu}\right)^{n-B+1}\right] & \text{if } x_1 = 0, x_2 > 0 \end{cases}$$

For simplicity of analysis we again only treat those states $x = (x_1, 0)$ where $x_1 \geq B$ because again for states $j = (j_1, 0)$ where $j_1 < B$ it is obvious that $m_{jz} \leq m_{Bz} \leq m_{xz}$. It is, of course,intuitive that the expected time to reach 0 increases as the starting state moves farther from 0.

We begin our analysis of m_{xz} by defining

$$A = \sum_{n=1}^{B-1} \overline{\theta}_n < \infty.$$

Note that $A < \infty$ because it is a finite sum of finite terms (it was already shown that $m_{ab} < \infty \quad \forall a, b \in R_d$). This allows us to write

$$m_{xz} = A + \sum_{n=B}^{x} \overline{\theta}_n.$$

Let us introduce some more notation to simplify the calculations. Let $\rho = \frac{\lambda}{B\mu}$ and let $C = \pi_0\frac{\lambda^{B-1}}{\mu(B-1)!}$. Since we only need to consider $\overline{\theta}_n$ for $n \geq B$ we can obtain a simple formula that shows $\overline{\theta}_n$ is constant in n for $n \geq B$

$$\overline{\theta}_n = \frac{1}{\lambda}\cdot\frac{1}{C\cdot\rho^{n-B}}\cdot\sum_{m=n}^{\infty} C\cdot\rho^{m-B+1} = \frac{1}{\lambda}\cdot\frac{1}{\rho^{n-B}}\cdot\rho^{n-B+1}\sum_{m=0}^{\infty}\cdot\rho^m = \left(\frac{1}{\lambda}\right)\frac{\rho}{1-\rho} = \frac{\lambda}{(B\mu)^2}. \tag{5}$$

We are now ready to present a formula for m_{xn} in terms of x as

$$m_{xz} = A + \sum_{n=B}^{x}\overline{\theta}_n = A + \sum_{n=B}^{x}\frac{\lambda}{(B\mu)^2} = A + (x - B)\frac{\lambda}{(B\mu)^2}.$$

The last step in the analysis is to show that $E[m_{w(i_1)z}] < \infty$, where the expectation is taken over the random variable $x_{i_1}(0)$, which can be shown as

$$E[m_{w(i_1)z}] = E\Big[A + (x_{i_1}(0) - B)\frac{\lambda}{(B\mu)^2}\Big] = A - B\frac{\lambda}{(B\mu)^2} + \frac{\lambda}{(B\mu)^2}E[x_{i_1}(0)] < \infty$$

This equation follows because A, B, λ, and μ are finite constants and we $E[x_{i_1}(0)] < \infty$ by Lemma 2.1.

This completes the argument that $m_{iz} < \infty$ for $i \in S - R_d$ because we have that (1) $E[m_{iw(i_1)}] < \infty$ by 4 and (2) $E[m_{w(i_1)z}] < \infty$ by 6, which shows that $m_{iz} = E[m_{iw(i_1)}] + E[m_{w(i_1)z}] < \infty$.

(III) To see that c_{iz} is finite first, consider $i \in R_d$. Since the system is already in the recurrence class in which the size of the call-in queue is 0, we can analyze the cost of the first passage as the *cost of a first passage* in a stable birth death process. McNeil (1970) showed that the cost/reward of the first passage from state $i \to z$ for a general birth death process with parameters λ_n and μ_n and reward function $g(x)$ has the same distribution as the *first passage time* for what we will call our "cost modified birth death process" with parameters $\lambda_n^* = \frac{\lambda_n}{g(n)}$ and $\mu_n^* = \frac{\mu_n}{g(n)}$.

In the case of the birth death process induced by policy d, we have $\lambda_n^* = \frac{\lambda}{c+h_1(B-n)^+ + \tau(n-B)^+}$ and $\mu_n^* = \frac{(B \bigwedge n)\mu}{c+h_1(B-n)^+ + \tau(n-B)^+}$ since there are 0 patients on the call-in queue after entering the recurrence class R_d. Thus $c_{iz} = m_{iz}^*$ where m_{iz}^* is the expected first passage time for the cost modified birth death process with parameters λ_n^* and μ_n^*. Thus if $m_{iz}^* < \infty$ then $c_{iz} < \infty$. From Kleinrock (1975) and part (1) above, we know that $m_{iz}^* < \infty$ if $\exists m_0$ s.t. $m > m_0 \Rightarrow \frac{\lambda_n^*}{\mu_n^*} < 1$. Let $m_0 = B$. Then for $n > B$ we have

$$\frac{\lambda_n^*}{\mu_n^*} = \frac{\frac{\lambda}{c+\tau(n-B)}}{\frac{B\mu}{c+\tau(n-B)}} = \frac{\lambda}{B\mu} < 1. \tag{6}$$

Therefore the first moment of the first passage time between any 2 states in the modified birth death process is finite: $m_{iz}^* < \infty$. This in turn implies that $c_{iz} = m_{iz}^* < \infty$. Thus for any state $i \in R_d$, we now have $c_{iz} < \infty$.

Next consider $i \in S - R_d$. We already argued that for such a state $i = (i_1, i_2)$, the expected time to reach a state j in the recurrent class R_d is finite ($m_{ij} < \infty$). Furthermore, the costs in each state are bounded above by a linear function of x_1 (not necessarily a tight bound); i.e. $C'_{x_2}(x_1) = c + h_1 B + h_2 \cdot x_2 + \tau x_1$. The first 3 terms are constant since the maximum cost the system could incur in a unit of time from cancelation is c, and from empty beds is $h_1 B$. Finally since the policy d ensures that the number of patients on the call-in queue will never exceed the initial number of patients on the queue, x_2, the maximum call-in patient holding cost that can be incurred in a given period is $h_2 \cdot x_2$. Thus the costs are bounded above by a linear function of x_1.

In order to show that the expected cost to go from i_2 patients in the call-in queue (i.e. $i \in S - R_d$), to 0 patients in the call-in queue (a state $j \in R_d$), we break the transitions down into pieces. First we consider the expected cost to go from i_2 patients in the call-in queue to $i_2 - 1$ patients in the call-in queue. Then we consider the expected cost to go from $i_2 - 1$ to $i_2 - 2$ and continue iteratively until we reach the cost to go from 1 patient in the call-in queue to 0 patients in the call-in queue. The policy d guarantees that once a transition from i_2 to $i_2 - 1$ patients in the call-in queue has occurred, the chain will never go back up to i_2 again (as can be seen in Figure 1), which supports the use of such an approach.

To complete our analysis we first define the expected cost to go from a state with n patients in the call-in queue, to a state with $n-1$. Formally, we define $M_{x,n}$ as the expected cost for the number of patients in the call-in queue to decrease by 1 given the system is currently in state (x, n).

In order to show that $c_{iz} < \infty$ for $i \in S - R_d$ we first show that $M_{x,n} < \infty \quad \forall x, n > 0$. Then we show that c_{ij} for $j \in R_d$ can be calculated as a sum of $M_{(\cdot,\cdot)}$ terms and can thereby be guaranteed to be finite. To show that

c_{iz} is finite, however, we use the stochastic process, $x_{i_1}(i_2)$, that describes the number of patients in the hospital at specific epochs where a patient is called in off the call-in queue. Then we show that $E[x.(\cdot)] < \infty$ is sufficient for c_{iz} will be finite.

Before developing $M_{x,n}$ further, first consider the random variable Y_{x_1} that represents the number of arrival transitions that occur before a departure transition when there are x_1 patients in the hospital. If $x_1 \geq B$ it is clear that Y_{x_1} is simply a modified Geometric($\frac{B\mu}{\lambda+B\mu}$) with range $k = 0, 1, 2, \ldots$. If $x_1 < B$ then Y has a slightly more complicated distribution, which we express in terms of $P(Y_{x_1} \geq k)$:

$$P(Y_{x_1} \geq k) = \sum_{j=k}^{\infty} \Big(\frac{B \wedge (x_1 + j)\mu}{\lambda + B \wedge (x_1 + j)\mu}\Big) \prod_{m=0}^{j-1} \Big(\frac{\lambda}{\lambda + B \wedge (x_1 + j)\mu}\Big).$$

Rewriting this expression in a more usable form we get

$$\begin{aligned} P(Y_{x_1} \geq k) = & \sum_{j=k}^{B-x_1-1} \Big(\frac{(x_1 + j)\mu}{\lambda + (x_1 + j)\mu}\Big) \prod_{m=0}^{j-1} \Big(\frac{\lambda}{\lambda + (x_1 + j)\mu}\Big) + \\ & \sum_{j=\left(k-(B-x_1)^+\right)^+}^{\infty} \Big(\frac{B\mu}{\lambda + B\mu}\Big)\Big(\frac{\lambda}{\lambda + B\mu}\Big)^j \cdot \prod_{m=0}^{B-x_1-1} \Big(\frac{\lambda}{\lambda + (x_1 + m)\mu}\Big). \end{aligned}$$

Note that, for the sums and products, if the upper limit is lower limit the value taken to be 1 for products and 0 for sums by convention. Now that we have the distribution for Y, we can proceed to provide a bound on $M_{x,n}$ as follows

$$\begin{aligned} & M_{x,n} \leq \\ & \sum_{k=0}^{\infty} \Big[\Big(\frac{B \wedge (x + k)\mu}{\lambda + B \wedge (x + k)\mu}\Big) \cdot \prod_{j=0}^{k-1} \Big(\frac{\lambda}{\lambda + B \wedge (x + j)\mu}\Big)\Big] \cdot \Big[\sum_{m=0}^{k} \Big(\frac{1}{\lambda + B \wedge (x + j)\mu}\Big) C'_n(x + j)\Big] \\ & = \sum_{k=0}^{B-x-1} \Big(\frac{1}{\lambda + (x + k)\mu}\Big) C'_n(x + k) \cdot P(Y_x \geq k) + \sum_{k=(B-x)^+}^{\infty} \Big(\frac{1}{\lambda + B\mu}\Big) C'_n(x + k) \cdot P(Y_x \geq k). \end{aligned} \tag{7}$$

The first step follows because $C'_n(x)$ is an upper bound on the cost associated with being in state x. Note that the cost is multiplied by the expected length of time spent in the state, $\left(\frac{1}{\lambda+B\wedge(x+k)\mu}\right)$, and by the probability that there are $k-1$ arrivals before a service, $\prod_{j=0}^{k-1}\left(\frac{\lambda}{\lambda+B\wedge(x+j)\mu}\right)$. The second step follows from some algebra. Note that if $x \geq B$ then the first term on the second step will be zero. We want to show that $M_{x,n} < \infty$ for all finite n and x. To this end, note that the first term in the second step of the above equation is finite for all x so it is sufficient to instead focus solely on showing that $\sum_{k=(B-x)^+}^{\infty}\left(\frac{1}{\lambda+B\mu}\right)C'_n(x+k) \cdot P(Y_x \geq k) < \infty$. To simplify the following expression let $D_x = \left(\frac{\lambda}{\lambda+B\mu}\right)^{-(B-x)^+} \cdot \left(\frac{B\mu}{\lambda+B\mu}\right) \cdot \prod_{m=0}^{B-x-1}\left(\frac{\lambda}{\lambda+(x+m)\mu}\right)$. Now we can show that the quantity of interest is finite.

$$
\begin{aligned}
&\sum_{k=(B-x)^+}^{\infty} \Big(\frac{1}{\lambda+B\mu}\Big) C'_n(x+k)\cdot P(Y_x \ge k) \\
&= \sum_{k=(B-x)^+}^{\infty} \Big(\frac{1}{\lambda+B\mu}\Big) C'_n(x+k)\cdot \sum_{j=k-(B-x_1)^+}^{\infty} \Big(\frac{B\mu}{\lambda+B\mu}\Big)\Big(\frac{\lambda}{\lambda+B\mu}\Big)^j \cdot \prod_{m=0}^{B-x_1-1} \Big(\frac{\lambda}{\lambda+(x_1+m)\mu}\Big) \\
&= \sum_{k=(B-x)^+}^{\infty} \Big(\frac{1}{\lambda+B\mu}\Big) C'_n(x+k)\cdot \Big(\frac{\lambda}{\lambda+B\mu}\Big)^k D_x \sum_{j=0}^{\infty} \Big(\frac{\lambda}{\lambda+B\mu}\Big)^j \\
&= \Big(\frac{1}{\lambda+B\mu}\Big) D_x \Big(\frac{\lambda+B\mu}{B\mu}\Big) \sum_{k=(B-x)^+}^{\infty} C'_n(x+k)\cdot \Big(\frac{\lambda}{\lambda+B\mu}\Big)^k \\
&\le \frac{D_x}{B\mu} \sum_{k=0}^{\infty} \Big(c+h_1 B + h_2\cdot n + \tau(x+k)\Big)\cdot \Big(\frac{\lambda}{\lambda+B\mu}\Big)^k \\
&= \frac{D_x}{B\mu}\Big(\frac{\lambda+B\mu}{B\mu}\Big) C'_n(x) + \frac{\tau D_x}{B\mu}\sum_{k=0}^{\infty} k\cdot \Big(\frac{\lambda}{\lambda+B\mu}\Big)^k \\
&= \frac{D_x(\lambda+B\mu)}{(B\mu)^2} C'_n(x) + \frac{\tau D_x}{B\mu}\Big[\Big(\frac{\lambda}{B\mu}\Big)^2 + \frac{\lambda}{B\mu}\Big] < \infty. \qquad (8)
\end{aligned}
$$

The first step just applies the definition of $P(Y_x \ge k)$. The second step involves of change of variable in the second sum and taking constant multipliers outside the sum. The third step applies the formula for an infinite sum with $p < 1$ and takes multipliers that are not indexed by k outside the left most sum. The fourth step just applies the definition of $C'_n(x+k)$ and applies a loose upper bound by adding a non-negative number of additional positive terms to the sum. The fifth step separates the terms that are indexed by k (i.e. τk) and those that are not (i.e. $C'_n(x)$) and evaluating the infinite sum for those terms not indexed by k. The final step follows from the fact that if $p < 1$ then $\sum_{i=0}^{\infty} ip^i = \frac{p}{(1-p)^2}$. This shows that $M_{x,n} < \infty$ for all $x, n > 0$. The results is obviously finite for finite x and n.

The next piece of the proof is to use the elements of $\{M_{x,n} :\ 0 \le n \le i_2\ \}$ to build a bridge from i_2 patients in the call-in queue down to 0 patients in the call-in queue. To do so, we again rely on the process $x_{i_1}(i_2)$ of the number of patients in the hospital after each downward transition (decrease) in call-in patients. We show that $E[x_{i_1}(i_2)] < \infty$, guarantees that $c_{iz} < \infty$.

For our purpose, we let $x_1(i_2) = i_1$ for the starting state $i = (i_1, i_2) \in S - R_d$. The goal now is to use $x_1(n)$ and the function $M_{x,n}$ to show that the total cost of emptying the call-in queue is finite. From Equations 7 and 8 we can see that

$$M_{x,n} \le H_x + \frac{D_x(\lambda+B\mu)}{(B\mu)^2} C'_n(x) + \frac{\tau D_x}{B\mu}\Big[\Big(\frac{\lambda}{B\mu}\Big)^2 + \frac{\lambda}{B\mu}\Big]. \qquad (9)$$

Where $H_x = \sum_{k=0}^{B-x-1} \Big(\frac{1}{\lambda+(x+k)\mu}\Big) C'_n(x+k)\cdot P(Y_x \ge k) < \infty) \quad \forall x > 0$ is a finite constant. To further isolate the effect of x on the finiteness of $M_{x,n}$ we can again bound $M_{x,n}$ by replacing H_x with H defined as

$$\begin{aligned} H_x &= \sum_{k=0}^{B-x-1} \Big(\frac{1}{\lambda + (x+k)\mu}\Big) C'_n(x+k) \cdot P(Y_x \geq k) \\ &\leq \sum_{k=0}^{B} \Big(\frac{1}{\lambda + (x+k)\mu}\Big) C'_n(x+k) \cdot P(Y_x \geq k) = H. \end{aligned}$$

In the same manner we also replace D_x with D defined as

$$D_x = \Big(\frac{\lambda}{\lambda + B\mu}\Big)^{-(B-x)^+} \cdot \Big(\frac{B\mu}{\lambda + B\mu}\Big) \cdot \prod_{m=0}^{B-x-1} \Big(\frac{\lambda}{\lambda + (x+m)\mu}\Big) \leq \Big(\frac{\lambda + B\mu}{\lambda}\Big)^{B} = D.$$

If we now consider $M_{x,n}$ as a function of x note that all terms are constant except for $\frac{D(\lambda+B\mu)}{(B\mu)^2}\tau x$. If we let

$$H'_n = H + \frac{D(\lambda + B\mu)}{(B\mu)^2}(c + h_1 B + h_2 \cdot n) + \frac{D\tau}{B\mu}\Bigg[\Big(\frac{\lambda}{B\mu}\Big)^2 + \frac{\lambda}{B\mu}\Bigg].$$

represent all the terms that are constant in x we can obtain a bound on $M_{x,n}$ that is linear in x,

$$M_{x,n} \leq H'_n + \frac{D(\lambda + B\mu)}{(B\mu)^2}\tau x \tag{10}$$

To show that the cost to empty the call-in queue is finite it suffices to show $\sum_{n=1}^{i_2} M_{x_{i_1}(n),n} < \infty$. Recall that $x_m(n)$ is the stochastic process that represents the how many patients have entered the hospital by the time we have been able to admit patient $n+1$ from the call-in queue. In other words, $x_m(n)$ gives us the horizontal (see Figure 1) position in the Markov chain as we march down in the vertical dimension; i.e. emptying the call-in queue from i_2 patients on the queue to 0 patients on the call-in queue. The following relation shows that c_{iz} is indeed finite

$$E\Big[\sum_{n=1}^{i_2} M_{x_{i_1}(n),n}\Big] \leq E\Big[\sum_{n=1}^{i_2} H'_{i_2} + \frac{D(\lambda + B\mu)}{(B\mu)^2}\tau x_{i_1}(n)\Big] = i_2 \cdot H'_{i_2} + \frac{D(\lambda + B\mu)}{(B\mu)^2}\tau \sum_{n=1}^{i_2} E[x_{i_1}(n)] < \infty \tag{11}$$

The first step follows from Equation 10. The third step follows because for any finite i_1, i_2, $E[x_{i_1}(n)] < \infty \quad \forall n < i_2$ by Lemma 2.1. We have shown that the expected cost for the Markov chain to enter the recurrent class R_d is finite, but now we must consider *where* the process enters R_d to be sure that the remaining cost to reach $z = (0,0)$ is indeed finite in expectation. As in part (II) above we rely on the stochastic process $\mathbf{x}(i_1, i_2) = \{x_{i_1}(0), x_{i_1}(1), \ldots, x_{i_1}(i_2)\}$. In particular, the element from this process that we are interested in is $x_{i_1}(0)$, which is the random variable that describes the random state, $w(i_1) = \big(x_{i_1}(0), 0\big)$, at which the Markov chain initially enters R_d. To show that the expected cost to reach $z = (0,0)$ from $i \in S - R_d$ it remains to show that $E[c_{w(i_1)z}] < \infty$, where the expectation is taken over the random variable $x_{i_1}(0)$.

The following argument shows that $E[c_{w(i_1)z}] < \infty$. (1) Once we enter R_d, to analyze the cost to go from $w(i_1) \to z$ we can consider the cost modified birth-death chain (with $\lambda^*_n = \frac{\lambda_n}{g(n)}$ and $\mu^*_n = \frac{\mu_n}{g(n)}$) presented at the beginning of this part of the proof (part (III)). (2) From Equation 6 in part (II) we know that $E[m_{w(i_1)z}] < \infty$. From the same arguments used in part (II) it follows that

$$E[c_{w(i_1)z}] = E[m^*_{w(i_1)z}] < \infty. \tag{12}$$

where $m^*_{w(i_1)z}$ the expected time to reach state $z = (0,0)$ from state $w(i_1)$ in the cost modified birth death chain. This completes the argument that $c_{iz} < \infty$ for $i \in S - R_d$ because we have that (1) $E[c_{iw(i_1)}] < \infty$ by Equation 11 and (2) $E[c_{w(i_1)z}] < \infty$, by applying Equation 6 to the cost modified birth-death process which shows that $c_{iz} = E[c_{iw(i_1)}] + E[c_{w(i_1)z}] < \infty$.

We have now shown (I) the Markov chain induced by d has a positive recurrent class R_d. Then we show that (II) for any state, i, $m_{iz} < \infty$, where $z = (0,0)$ and we want to show that the Markov chain induced by d is z-standard. Finally we show that (III) for any state, i, $c_{iz} < \infty$. Therefore the MC induced by d is *z standard* and CAV1 holds. □

Lemma 2.3. *CAV2 holds for our model*

Proof. Take any $U > 0$. The set D_U contains at most those states (x,y) s.t. $h_1 y + (x-B)^+ \tau < U$ which is bounded above by states where $y < \frac{U}{h_1}$ and $x < \frac{U}{\tau} + B$ and bounded below by $(0,0)$. This set is clearly finite for non-zero parameters τ and h_1. □

Lemma 2.4. *CAV3 holds for our model.*

Proof. As in CAV 1 we need to show that d'_i satisfies CAV3. To show that this policy satisfies CAV3 we follow a similar approach to the proof of CAV1. We first show that (I) the Markov chain induced by d'_i reaches state i with probability 1. Then we show that (II) for $z = (0,0)$, $m_{zi} < \infty$. Finally we show that (III) for any state, $i \in S - R_d$, $c_{zi} < \infty$ under policy d'_i. These 3 conditions satisfy the requirements for CAV3.

(I) This policy guarantees with probability 1 that, starting at state z the process will reach state i for any $i \in S - R_d$. This is because the system will steadily build up to y patients on the call-in queue and then maintain exactly that number in the call-in queue from there on out as can be seen in Figure 2. Once reaching state (u,y) for an arbitrary u, the MC again becomes a stable birth death process in the first state variable, with the second state variable fixed at y. In a stable birth death process, given that you start in state (u,y), you will eventually visit state (w,y) with probability 1 and the mean first passage time between the two states will be finite as argued in the proof of CAV1.

(II) The expected time to reach a given state $i \in S - R_d$ from state $z = (0,0)$ can be defined as $\frac{i_2}{\lambda_q} + E[m_{w(i_2)z}]$ where $w(i_2)$ is the random variable for the state of the hospital when the process reaches a call-in queue length of i_2. Following the same arguments from the proof of Lemma 2.2 and defining a stochastic process $\mathbf{y}(i) = \{y(0), y(1), \ldots, y(i)\}$, which is analogous to the stochastic process $\mathbf{x}(i)$ used in Lemma 2.2 we can show the total expected time to go from $z \to i$ is $m_{zi} < \infty$ for all $i \in S$.

(III) The fact that $i \in S - R_d \quad c_{zi}(d'_i) < \infty$ follows using similar arguments to those used to prove CAV1 part (II) in Lemma 2.2. Therefore $\forall i \in S - R_d$ we have that $d'_i \in R^*(z,i)$.

□

This completes the proof of the CAV Assumptions. Now we proceed to show the main result, that Theorem 1.1 holds for our model of admission control in hospitals.

3 Proof of the Main Result

In Section 2 we showed that the CAV assumptions hold. The proof of Theorem 1.1 now follows by showing that the CAV assumptions imply the SEN assumptions and the SEN assumptions guarantee the results of the theorem.

Proof. of Theorem 1.1 First, from Corollary 7.5.9 and Theorem 7.5.6 (i) of Sennott (1999) the SEN assumptions hold if the CAV assumptions hold. Now the finite-horizon discounted problem converges to the infinite-horizon discounted problem by Proposition 4.3.1 of Sennott (1999). Finally, as a result of the SEN assumptions, the infinite-horizon discounted problem converges to the infinite-horizon average cost per unit time problem as the discount factor goes to 1 by Theorem 7.2.3 of Sennott (1999), which proves our Theorem 1.1 (i) and (ii). Part (iii) of Theorem 1.1 follows from Proposition 4.3.1 of Sennott (1999) for the infinite-horizon discounted problem and from Theorem 7.2.3 we get that the same holds for the infinite-horizon average cost per unit time problem. This concludes our main result by completing the proof of Theorem 1.1. □

References

Helm, J.E., S. AhmadBeygi, M.P. Van Oyen. 2010. Design and Analysis of Hospital Admission Control for Operational Effectiveness. *Production and Operations Management Society (Under Review)* .

Jouini, O., Y. Dallery. 2008. Moments of first passage times in general birth–death processes. *Mathematical Methods of Operations Research* **68**(1) 49–76.

Kleinrock, L. 1975. Queueing Systems, Vol. 1, 2. *Jon Wiley & Sons* .

McNeil, D.R. 1970. Integral functionals of birth and death processes and related limiting distributions. *The Annals of Mathematical Statistics* **41**(2) 480–485.

Sennott, L.I. 1999. *Stochastic dynamic programming and the control of queueing systems*. Wiley-Interscience.