

Proofs of auxiliary lemmas for the manuscript titled:  
**Centralized vs. Decentralized Ambulance Diversion:  
A Network Perspective**

This document has three parts. In the first part (§T.1) we provide the proofs of the key results that were stated in the manuscript but not proved in its appendix. Specifically, we prove here Theorems 1, 3, 6 and 7. Also, at the end of §T.1 we prove the existence of optimal solutions to the social planner’s problem as defined in (5).

In the second part of this report (§T.2) we prove the convergence of (sequence of) steady-state distributions of the the truncated chains to that of the original chain as the truncation parameter increases. This claim was used to justify the methodology for the numerical experiments in the manuscript.

Finally, §T.3 includes the proofs of auxiliary results that where stated and used within the appendix of the manuscript but whose proofs where relegated to this document. These are Lemma A.1, Propositions A.1-A.3 and Corollary A.2.

## **T.1 Proofs of remaining theorems**

### **T.1.1 Proof of Theorem 1:**

The intuition for stability is simple. At times  $t$  with  $Q_i^K(t) \geq K_i \vee N_i$ , ED  $i$  experiences an arrival rate that is smaller or equal to  $\lambda_i$  and the depletion rate of the queue is  $\mu N_i$ . By condition (1),  $\lambda_i < \mu N_i$  and this guarantees that the queue does not “explode”. The argument below formalizes this intuition.

First, it is easily verified that the Markov chain  $X^K(\cdot)$  is irreducible. A simple coupling argument shows that, initializing  $X^k(0) = (0, 0)$ , the process  $X^K(\cdot)$  can be constructed on the same sample space with two  $M/M/1$  queues  $\mathcal{M}_1(\cdot)$  and  $\mathcal{M}_2(\cdot)$  so that  $\mathcal{M}_i(\cdot)$  has arrival rate  $\lambda_i$ , and service rate  $\mu N_i$ . Initializing  $\mathcal{M}_1(0) = \mathcal{M}_2(0) = 0$ ,  $X_i^K(t) \leq K_i \vee N_i + \mathcal{M}_i(t)$  for all  $t \geq 0$  almost surely and for  $i = 1, 2$ .

Consequently, for all  $t \geq 0$ ,

$$\mathbb{E}[X_i^K(t)] \leq K_i \vee N_i + \mathbb{E}[\mathcal{M}_i(t)], \quad i = 1, 2. \quad (\text{T1})$$

Condition (1) guarantees for the  $M/M/1$  queue (by its monotonicity; see e.g. Chapter 9 of Ross [1996]) that  $\mathbb{E}[\mathcal{M}_i(t)]$  is increasing in  $t$  and converges to the steady-state mean  $\mathbb{E}[\mathcal{M}_i(\infty)]$ . Plugging this into (T1) we have that

$$\sup_{t \geq 0} \mathbb{E}[X_i^K(t)] < \infty. \quad (\text{T2})$$

Since the chain is irreducible all states are either recurrent or transient together. Transience implies that  $\mathbb{E}[X_1^K(t) + X_2^K(t)] \rightarrow \infty$  as  $t \rightarrow \infty$  so that, from (T2), the chain must be recurrent. To show that it is, in fact, positive recurrent assume, to reach a contradiction, that it is null recurrent. In that case,  $P_{xy}(t) := \mathbb{P}\{X^K(t) = y | X^K(0) = x\} \rightarrow 0$ , for all  $x, y \in \mathbb{Z}_+^2$  (see, e.g., Corollary 4.7 in Asmussen [2003]). In turn, given  $M > 0$ , there exists  $t_0(M)$  such that  $\mathbb{P}\{X_1^K(t) + X_2^K(t) > M\} \geq 1/2$  for all time  $t \geq t_0(M)$  and, in particular,  $\mathbb{E}[X_1^K(t) + X_2^K(t)] \geq M/2$  for all such  $t$ . Consequently,  $\sup_{t \geq 0} \mathbb{E}[X_1^K(t) + X_2^K(t)] = \infty$

contradicting (T2). We thus conclude that  $X^K(\cdot)$  is a positive recurrent CTMC.

We next show that given a threshold pair  $(K_1, K_2) \neq 0$ ,  $X^k(\cdot)$  is reversible if and only if  $\lambda_1^a = \lambda_2^a = 0$ . For one direction note that if  $\lambda_1^a = \lambda_2^a = 0$  the system consists of two independent  $M/M/N$  queues, each of which is reversible (see e.g. Proposition 5.6.1 in Ross [1996]), so that  $X^K(\cdot)$  is trivially reversible.

To prove the other direction it suffices to find a number  $L \in \mathbb{Z}_+$  and a sequence of states  $\{x^m \in \mathbb{Z}_+^2, m = 0, \dots, L\}$  that satisfy  $\prod_{m=1}^L q_{x^{m-1}, x^m} q_{x^L, x^0} \neq q_{x^0, x^L} \prod_{m=1}^L q_{x^m, x^{m-1}}$ , where  $q_{x^i, x^j}$  is the transition rate from  $x^i$  to  $x^j$ ; see, e.g., Theorem 6.33 in Kulkarni [1995].

To that end, fix a threshold pair  $K = (K_1, K_2) \neq 0$ . Assuming  $K_2 > 0$  consider the four states  $x^0 = (K_1 + 1, K_2 - 1)$ ,  $x^1 = (K_1 + 2, K_2 - 1)$ ,  $x^2 = (K_1 + 2, K_2)$  and  $x^3 = (K_1 + 1, K_2)$ . Then,

$$A := \prod_{m=1}^3 q_{x^{m-1}, x^m} q_{x^3, x^0} = \lambda_1^w (\lambda_2 + \lambda_1^a) \mu((K_1 + 2) \wedge N_1) \mu(K_2 \wedge N_2),$$

while

$$B := q_{x^0, x^3} \prod_{m=1}^3 q_{x^m, x^{m-1}} = (\lambda_2 + \lambda_1^a) \lambda_1 \mu(K_2 \wedge N_2) \mu((K_1 + 2) \wedge N_1).$$

Note that  $A = B$  if and only if  $\lambda_1 = \lambda_1^w$  (i.e.,  $\lambda_1^a = 0$ ). If  $K_2 = 0$  but  $K_1 > 0$ , we can similarly choose the states to be  $x^0 = (K_1 - 1, K_2 + 1)$ ,  $x^1 = (K_1 - 1, K_2 + 2)$ ,  $x^2 = (K_1, K_2 + 2)$  and  $x^3 = (K_1, K_2 + 1)$  the equality now holds only if  $\lambda_2^a = 0$ . Thus, the chain is reversible if and only if  $\lambda_1^a = \lambda_2^a = 0$ . ■

### T.1.2 Proof of Theorem 3

First, we claim that it suffices to restrict attention to finite and strictly positive threshold pairs. Indeed, infinite thresholds are ruled out as potential equilibria by Corollary A.1. Thresholds pairs  $K = (K_1, K_2) \neq 0$  for which  $K_i = 0$  for one  $i = 1, 2$  are ruled out as potential equilibria by Theorem 2.

To prove the theorem we need to show that, given a finite and strictly positive threshold pair  $K$ , there exists a choice of  $\lambda_1^a$  and  $\lambda_2^a$  such that  $K$  cannot be an equilibrium. We will prove this by showing that for all sufficiently small (but strictly positive)  $\lambda_1^a$ , ED 2's best response to  $K_2$  is to use  $K_1 = 0$  regardless of  $K_2$ . An identical argument can be applied to  $\lambda_2^a$  and ED 1.

The proof of the theorem builds on two sub-results: (a) a continuity result that establishes that, as  $\lambda_1^a$  approaches 0 (and, in turn,  $\lambda_1^w$  approaches  $\lambda_1$ ), the performance metrics of the network approach those of a *limit* network in which  $\lambda_1^a = 0$  and, (b) a comparison result for this *limit* network showing that  $K_1 = 0$  is ED 1's best response. In passing, we note that part of the complexity of the argument, especially in the comparison result (item (b) above), arises because of strict inequalities.

We fix the parameters  $\lambda_1, \lambda_2^a, \lambda_2^w, N_1, N_2$  and  $\mu$  and set  $\zeta := \lambda_1^a$ . We will let  $\zeta$  go to 0. To make explicit the dependence on  $\zeta$ , we add  $\zeta$  to the superscript. Accordingly, for example,  $Q_i^{\zeta, K}$  is the steady-state queue length at ED  $i$  when the threshold pair is  $K$  and  $\lambda_1^a = \zeta$ .

We next state two auxiliary results. The first corresponds to items (a) above.

**Lemma T.2** *For any finite threshold pair  $K$*

$$X^{\zeta, K} \Rightarrow X^{0, K}, \text{ and } \mathbb{E}[Q_i^{\zeta, K}] \rightarrow \mathbb{E}[Q_i^{0, K}], \quad i = 1, 2, \text{ as } \zeta \rightarrow 0.$$

The comparison results that we state next corresponds to item (b) in the outline above. Here, we consider the network with  $\lambda_1^a = \zeta = 0$  (since  $\zeta = 0$  is fixed we omit it from the notation). We define a Discrete Time Markov Chain (DTMC) by “sampling”  $X^K(\cdot)$  at arrival epochs to ED 1, where arrivals are generated by:

- A walk-in arrival to ED 1 (there are no exogenous ambulance arrivals because  $\lambda_1^a = 0$ ).
- An ambulance diverted from ED 2.

More formally, letting  $A_1^w(t)$  be the number of walk-in arrivals to ED 1 by time  $t$  and letting  $A_2^a(t)$  be the number of ambulance arrivals from catchment area 2 by time  $t$ , the number of arrivals to ED 1 by time  $t$  are given by:

$$\bar{A}_1(t) = A_1^w(t) + \int_0^t \mathbb{1}\{X_1(s) < K_1, X_2(s) \geq K_2\} dA_2^a(s).$$

Let  $\tau^n = \inf\{t \geq 0 : \bar{A}_1(t) = n\}$  so that  $\tau^n$  is the time of the  $n^{\text{th}}$  such arrival. Finally, define

$$\begin{aligned} \bar{X}^K(0) &= (X_1^K(0), X_2^K(0)), \\ \bar{X}^K(n) &= (X_1^K(\tau^n-), X_2^K(\tau^n-)), \quad n \geq 1. \end{aligned}$$

The process  $\{\bar{X}^K(n), n \in \mathbb{Z}_+\}$  is a DTMC that is easily verified to be irreducible and aperiodic. Furthermore, as the CTMC  $X^K(\cdot)$  has a unique steady-state distribution  $\pi^K$ , so does  $\bar{X}^K(\cdot)$  and its (unique) steady-state distribution  $\bar{\pi}^K$  can be constructed explicitly from  $\pi^K$ . Let  $\bar{X}^K$  be a random variable with the distribution  $\bar{\pi}^K$ . The following comparison result shows that (with  $\lambda_1^a = 0$ ) the number of patients found by arriving patients to ED 1 is greater under  $(K_1, K_2)$  than it is under  $(0, K_2)$ .

**Lemma T.3** *Assume that  $\bar{X}_1^{(K_1, K_2)}(0) \geq_{st} \bar{X}_1^{(0, K_2)}(0)$ . Then,*

$$\bar{X}_1^{(K_1, K_2)}(n) \geq_{st} \bar{X}_1^{(0, K_2)}(n), \quad \text{for all } n \in \mathbb{Z}_+,$$

and, consequently,

$$\bar{X}_1^{(K_1, K_2)} \geq_{st} \bar{X}_1^{(0, K_2)}. \tag{T3}$$

Furthermore,

$$\mathbb{P}\left\{\bar{X}_1^{(K_1, K_2)} > j\right\} > \mathbb{P}\left\{\bar{X}_1^{(0, K_2)} > j\right\}, \tag{T4}$$

for all  $j \geq N_1$  and, consequently,

$$\mathbb{E}\left[(\bar{X}_1^{(K_1, K_2)} - N_1)^+\right] > \mathbb{E}\left[(\bar{X}_1^{(0, K_2)} - N_2)^+\right]. \tag{T5}$$

Note that stochastic ordering does not imply strict inequalities. Hence, (T4) is a strict addition over (T3). The proof of Lemmas T.2 and T.3 appear at the end of this subsection and we use them now to complete the proof of Theorem 3.

**Proof of Theorem 3:** Because service times are exponential the expected waiting time of patients arriving to ED 1 under a threshold pair  $(K_1, K_2)$  is given for the network with  $\lambda_1^a = 0$ , by

$$\mathbb{E}[W_1(K_1, K_2)] = \frac{\mathbb{E}[(\bar{X}^{(K_1, K_2)} - N_1)^+]}{N_1\mu}, \quad (\text{T6})$$

and by Lemma T.3, we have that

$$\mathbb{E}[W_1(K_1, K_2)] = \frac{\mathbb{E}[(\bar{X}^{(K_1, K_2)} - N_1)^+]}{N_1\mu} > \frac{\mathbb{E}[(\bar{X}^{(0, K_2)} - N_1)^+]}{N_1\mu} = \mathbb{E}[W_1(0, K_2)].$$

The waiting time can also be computed from the CTMC using Little's law to get

$$\frac{\mathbb{E}[Q_1^{(K_1, K_2)}]}{\lambda_1(K_1, K_2)} = \mathbb{E}[W_1(K_1, K_2)] > \mathbb{E}[W_1(0, K_2)] = \frac{\mathbb{E}[Q_1^{(0, K_2)}]}{\lambda_1(0, K_2)}. \quad (\text{T7})$$

Lemma T.2 (for the CTMC) guarantees that the inequality in (T7) is preserved for all  $\lambda_1^a$  sufficiently small. For all such  $\lambda_1^a$ ,  $K_1 = 0$  is ED 1's best response to  $K_2$  so that  $(K_1, K_2)$  can not be a Nash equilibrium. ■

**Proof of Lemma T.2:** We fix a threshold pair  $K$  throughout and we omit it from the notation.

First, almost identically to the proof of Lemma T.8 it can be argued via a coupling argument that the family  $(X^\zeta; \zeta \geq 0)$  is uniformly integrable and, in turn, tight. We omit the detailed argument and refer the reader to the proof of Lemma T.8.

Tightness allows us to fix a sequence  $\{\zeta_l, l \in \mathbb{Z}_+\}$  such that  $\zeta_l \rightarrow 0$  and  $X^{\zeta_l} \Rightarrow Y$  for some limit random variable  $Y$  (that may depend on the chosen subsequence). Let  $\tilde{\pi}$  be the distribution of  $Y$ . Applying Theorem 1 and Lemma 1 of Whitt [1980] it then follows that  $\tilde{\pi}$  is a stationary distribution for the CTMC  $X^0(\cdot)$ . The sufficient conditions of Theorem 1 and Lemma 1 in Whitt [1980] apply to general Semi-Markov Processes and are easy to verify in our setting. By our Theorem 1  $X^0(\cdot)$  has a unique stationary distribution which is also its steady-state distribution. In turn,  $\tilde{\pi}$  must be the unique stationary distribution of  $X^0$ .

Since the same argument can be repeated for every convergent subsequence we conclude, using Prohorov's theorem, that  $X^\zeta \Rightarrow X^0$  as  $\zeta \rightarrow 0$ . Finally, the uniform integrability implies the convergence of the expectations. ■

**Proof of Lemma T.3:** This lemma is concerned with the comparison of a single coordinate of the two-dimensional Markov chains  $\bar{X}^{(K_1, K_2)}(\cdot)$  and  $\bar{X}^{(0, K_2)}(\cdot)$ . Our result is similar in spirit to Theorem 5.2.11 in Müller and Stoyan [2002] which deals with such partial ordering of Markov chains. Nevertheless, we provide a self-contained argument that also allows us to get a strict inequality as in the statement of the lemma.

Key in the proof is establishing the two properties of the chains  $\bar{X}^{(K_1, K_2)}(\cdot)$  and  $\bar{X}^{(0, K_2)}(\cdot)$ .

1. *Stochastic monotonicity of  $\bar{X}^{(0, K_2)}(\cdot)$  in its first coordinate:* for any  $i, j, n \in \mathbb{Z}_+$  and  $l \leq i$ ,

$$\mathbb{P}\{\bar{X}_1^{(0, K_2)}(n) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \geq \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = l\}. \quad (\text{T8})$$

Note that with the threshold pair  $(0, K_2)$  and with  $\lambda_1^a = 0$  ED 1 does not divert nor does it accept any diversions from ED 2. Hence, the process  $\bar{X}_1^{(0, K_2)}(\cdot)$  has the law of an  $M/M/N$  queue with arrival rate  $\lambda_1$ , service rate  $\mu$  and  $N_1$  servers. In turn, the monotonicity in equation T8 follows from known stochastic monotonicity of the  $M/M/N$  queue; see e.g. Chapter 9 of Ross [1996].

2. *Comparison of the two chains with respect to the first-coordinate transition probabilities:* for all  $i, j, k, l$

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j, \bar{X}_2^{(K_1, K_2)}(1) = l | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \\ \geq \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\}. \end{aligned} \quad (\text{T9})$$

The inequality in (T9) is strict when  $j = i$  and  $i < K_1$ . That is, for  $i < K_1$

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > i, \bar{X}_2^{(K_1, K_2)}(1) = l | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \\ > \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > i | \bar{X}_1^{(0, K_2)}(0) = i\}. \end{aligned} \quad (\text{T10})$$

Note that, by the definition of the underlying DTMC (sampling on arrivals) we have (with probability 1) that  $\bar{X}_1^{(K_1, K_2)}(n+1) \leq \bar{X}_1^{(K_1, K_2)}(n) + 1$  for all  $n \in \mathbb{Z}_+$  so that (T10) becomes equivalent to

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) = i+1, \bar{X}_2^{(K_1, K_2)}(1) = l | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \\ > \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > i | \bar{X}_1^{(0, K_2)}(0) = i\}, \end{aligned} \quad (\text{T11})$$

for  $i < K_1$ . Also, summing over  $l \in \mathbb{Z}_+$  in (T9) we have that

$$\mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \geq \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\}. \quad (\text{T12})$$

Before proving these two properties we apply them to the proof of the lemma. Using (T8) and (T9), we have

for  $j \geq 0$  that

$$\begin{aligned}
& \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j\} \\
&= \sum_{(i, k) \in \mathbb{Z}_+^2} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \\
&\geq \sum_{(i, k) \in \mathbb{Z}_+^2} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \\
&= \sum_{i \in \mathbb{Z}_+} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \sum_{k \in \mathbb{Z}_+} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_1^{(K_1, K_2)}(0) = k\} \\
&= \sum_{i \in \mathbb{Z}_+} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \\
&\geq \sum_{i \in \mathbb{Z}_+} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(0) = i\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \\
&= \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j\}
\end{aligned}$$

Here, the first inequality follows from (T9). Recall that for two random variables  $R_1$  and  $R_2$ , we have that  $R_1 \leq_{st} R_2$  if and only if  $\mathbb{E}[f(R_1)] \leq \mathbb{E}[f(R_2)]$  for every non-decreasing function  $f$  for which the expectations are defined so that the second inequality above follows from the assumption that  $\bar{X}_1^{(K_1, K_2)}(0) \geq_{st} \bar{X}_1^{(0, K_2)}(0)$  and from the fact (implied by (T8)) that the function  $f_j^n(\cdot)$  defined by

$$f_j^n(i) := \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n) > j | \bar{X}_1^{(0, K_2)}(0) = i\} \quad (\text{T13})$$

is a non-decreasing function of  $i$  for every  $n$  (and, in particular, for  $n = 1$ ).

Thus, we have proved that  $\bar{X}_1^{(K_1, K_2)}(1) \geq_{st} \bar{X}_1^{(0, K_2)}(1)$  whenever  $\bar{X}_1^{(K_1, K_2)}(0) \geq_{st} \bar{X}_1^{(0, K_2)}(0)$ . Proceeding by induction we conclude that

$$\bar{X}_1^{(K_1, K_2)}(n) \geq_{st} \bar{X}_1^{(0, K_2)}(n), \text{ for all } n \in \mathbb{Z}_+. \quad (\text{T14})$$

To prove the steady-state ordering in (T3) we fix an arbitrary initial state  $(i, k)$  and set  $\bar{X}^{(K_1, K_2)}(0) = \bar{X}^{(K_1, K_2)}(0) = (i, k)$ . This trivially satisfies the condition  $\bar{X}_1^{(K_1, K_2)}(0) \geq_{st} \bar{X}^{(0, K_2)}(0)$  so that (T14) applies. Using the convergence to steady-state as  $n \rightarrow \infty$  we conclude that

$$\bar{X}^{(K_1, K_2)} \stackrel{d}{=} \lim_{n \rightarrow \infty} \bar{X}^{(K_1, K_2)}(n) \geq \lim_{n \rightarrow \infty} \bar{X}^{(0, K_2)}(n) \stackrel{d}{=} \bar{X}^{(K_1, K_2)},$$

which establishes (T3).

We turn to prove (T4). For this, we initialize both  $\bar{X}_1^{(0, K_2)}(\cdot)$  and  $\bar{X}_1^{(K_1, K_2)}(\cdot)$  at time  $n = 0$  with their steady-state distributions. In turn, by (T3) we have that  $\bar{X}_1^{(K_1, K_2)}(\cdot) \geq_{st} \bar{X}_1^{(0, K_2)}(\cdot)$ . The resulting processes are stationary so that,

$$\bar{X}^{(K_1, K_2)}(n) \stackrel{d}{=} \bar{X}^{(K_1, K_2)}(0) \text{ and } \bar{X}^{(0, K_2)}(n) \stackrel{d}{=} \bar{X}^{(0, K_2)}(0), \text{ for all } n \in \mathbb{Z}_+, . \quad (\text{T15})$$

Given  $j \geq N_1$  we will identify a fixed time index  $n_0$  for which

$$\mathbb{P}\{\bar{X}^{(K_1, K_2)}(n_0) > j\} > \mathbb{P}\{\bar{X}^{(0, K_2)}(n_0) > j\}, \quad (\text{T16})$$

so that (T4) will then follows from (T15).

To prove (T16) fix  $i < K_1 \wedge N_1$  and  $j \geq N_1$  and set  $n_0 = j - i + 1$ . Note, that the only path to get from  $i$  to (above)  $j$  in  $n_0$  steps is to have  $n_0$  consecutive arrivals with no service completions at ED 1. Using the Markov property we then have that

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} &= \\ \sum_{\{l^n, n \leq n_0\}} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n) = i + n, \bar{X}_2^{(K_1, K_2)}(n) = l^n; n = 1, \dots, n_0 | \bar{X}^{(K_1, K_2)}(0) = i, \bar{X}_2^{(K_1, K_2)}(0) = k\} &= \\ \sum_{\{l^n, n \leq n_0\}} \prod_{n=1}^{n_0} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n) = i + n, \bar{X}^{(K_1, K_2)}(n) = l^n | \bar{X}^{(K_1, K_2)}(n-1) = i + n - 1, \bar{X}_2^{(K_1, K_2)}(n-1) = l^{n-1}\}, \end{aligned}$$

where the sum is over sequences  $\{l^n, n \leq n_0\}$ . For each of the elements in the right-hand-side we apply (T11) to obtain

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n) = i + n | \bar{X}_1^{(K_1, K_2)}(n-1) = i + n - 1, \bar{X}_2^{(K_1, K_2)}(n-1) = l^{n-1}\} \\ > \mathbb{P}\{\bar{X}_1^{(0, K_2)}(m) = i + 1 | \bar{X}^{(0, K_2)}(m-1) = i, \}. \end{aligned} \quad (\text{T17})$$

We are now going to use this to conclude the proof of the strict inequality in (T4). Note that

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n_0) > j\} &= \sum_l \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = l\} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = l\} \\ &= \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i\} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = i\} \\ &\quad + \sum_{l \neq i} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = l\} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = l\} \\ &> \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = i\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = i\} \\ &\quad + \sum_{l \neq i, K_1} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = l\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n_0) > j | \bar{X}_1^{(0, K_2)}(0) = l\} \\ &= \sum_l \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(0) = l\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n_0) > j | \bar{X}_1^{(0, K_2)}(0) = l\} \\ &\geq \sum_l \mathbb{P}\{\bar{X}_1^{(0, K_2)}(0) = l\} \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n_0) > j | \bar{X}_1^{(K_1, K_2)}(0) = l\} \\ &= \mathbb{P}\{\bar{X}_1^{(0, K_2)}(n_0) > j\}, \end{aligned}$$

where the last step follows from the stochastic ordering in (T3) and from the monotonicity of the function  $f_j^{n_0}(\cdot)$ ; see equation (T13).

This concludes the proof of (T16) and in turn that of (T4). Equation (T5) now follows directly from (T3) and (T4) using the tail formula for expectation  $\mathbb{E}\left[(\bar{X}_1^{(K_1, K_2)} - N_1)^+\right] = \sum_{j \geq N_1} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)} > j\}$

and similarly for  $\bar{X}_1^{(0,K_2)}$ .

To complete the proof of the lemma it then only remains to prove (T9). For that purpose, we introduce a few definitions. Let  $(D_l(t), t \geq 0)$  be a pure death process with death rate  $\mu d$  when in state  $d$  and with initial condition  $D_l(0) = l$ . Let  $D_l(s, t) = D_l(t) - D_l(s)$  that is  $D_l(s, t)$  is a random variable that has the distribution of the change in this death process on the interval  $(s, t]$ . Between arrivals to ED 1 the number of patients there evolves like the above pure death process. Formally,

$$X_1^{(K_1, K_2)}(n+1) = X_1^{(K_1, K_2)}(n) + 1 - D_{X_1^{(K_1, K_2)}(n)+1}(\tau^n, \tau^{n+1}).$$

Note that  $\tau^{n+1} - \tau^n$  is independent of  $X^{(K_1, K_2)}(n)$ . Also, for  $K = (0, K_2)$  the only arrivals are those from catchment area 1 (there are no diversion from ED 2) and also recall that  $\lambda_1^a = 0$  by assumption so that there are not diversion from catchment area 1 to ED 2. Hence,  $\tau^{n+1} - \tau^n$  is exponentially distributed with rate  $\lambda_1$ . Letting  $Z^a$  be an exponential random variable with rate  $\lambda_1$ , we hence have that

$$\mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)}(0) = i\} = \mathbb{P}\{D_{i+1}(0, Z^a) \leq j - i\},$$

for all  $j \leq i$ . Under the threshold pair  $K = (K_1, K_2)$  there can be arrivals also through diversions from ED 2 so that  $\tau^{n+1} - \tau^n \leq_{st} Z^a$ . In turn,

$$\mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j | \bar{X}_1^{(K_1, K_2)} = i, \bar{X}_2^{(K_1, K_2)} = k\} \geq \mathbb{P}\{D_{i+1}(0, Z^a) \leq j - i\} = \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)} = i\}, \quad (\text{T18})$$

where the inequality follows here because with  $(K_1, K_2)$  an arrival might happen earlier than  $Z^a$  (a diversion from ED 2).

To show that the inequality in (T18) is strict for  $i < K_1$  and  $j = i$ , define

$$\tilde{\tau} = \inf\{t \geq 0 : A_2(t) - A_2(t-) = 1, X_2^{(K_1, K_2)}(t-) \geq K_2\}.$$

(note that  $\tilde{\tau}$  is defined using the CTMC  $X^{(K_1, K_2)}$  is the original CTMC rather than the DTMC). In words,  $\tau$  is the first time that an arrival from catchment area 2 (a jump of the process  $A_2(t) = A_2^w(t) + A_2^a(t)$ ) arrives after time 0 to find ED 2 above its threshold. If  $\bar{X}_1^{(K_1, K_2)}(0) = i < K_1$  ED 1 is not on diversion until  $\tau^1$  (the first arrival after time 0) so that this first arrival can be a diverted ambulance from ED 2 at which case  $\tau^1 = \tilde{\tau}$ . Hence, we have that

$$\begin{aligned} \mathbb{P}\{\bar{X}_1^{(K_1, K_2)}(1) > j | \bar{X}_1^{(K_1, K_2)} = i, \bar{X}_2^{(K_1, K_2)} = k\} &= \mathbb{P}\{D_{i+1}(0, \tilde{\tau} \wedge Z^a) \leq j - i\} \\ &> \mathbb{P}\{D_{i+1}(0, Z^a) \leq j - i\} = \mathbb{P}\{\bar{X}_1^{(0, K_2)}(1) > j | \bar{X}_1^{(0, K_2)} = i\}. \end{aligned}$$

The strict inequality follows from the fact that  $\tilde{\tau}$  and  $Z^a$  are independent random variables each with a support over  $\mathbb{R}_+$  so that  $\mathbb{P}\{\tilde{\tau} < Z^a\} > 0$  (in words, there is a strictly positive probability of an overflow from ED 2 before an exogenous arrival to ED 1). This concludes the proof of (T9) and, in turn, the proof of the lemma.  $\blacksquare$

### T.1.3 Proof of Theorem 6

We start with preliminary bounds on the steady-state queues in the  $(N_1, N_2)$ -threshold system.

**Lemma T.4** For all  $\hat{B} \geq 0$ , and  $i = 1, 2$ ,

$$\mathbb{E}[Q_i^N] \leq \hat{B} + \mathbb{P}\{Q_i^N \geq \hat{B}\} \frac{\lambda_i}{\mu N_i - \lambda_i}.$$

The proof of Lemma T.4 appears at the end of this subsection and we proceed with the proof of Theorem 6. Fixing  $\hat{B} > 0$  we have the following sequence of inequalities for  $i = 1, 2$ :

$$\begin{aligned} \mathbb{P}\{Q_i^N \geq \hat{B}\} &\leq \mathbb{P}\{Q_\Sigma^N \geq \hat{B}\} \leq \mathbb{P}\{Q_\Sigma^N \geq \hat{B}, Q_\Sigma^P = 0\} + \mathbb{P}\{Q_\Sigma^P > 0\} \\ &\leq \mathbb{P}\{Q_\Sigma^P > 0\} + \mathbb{P}\{|Q_\Sigma^N - Q_\Sigma^P| \geq \hat{B}\}, \end{aligned} \quad (\text{T19})$$

For the last inequality we used the inclusion of events  $\{Q_\Sigma^N \geq \hat{B}, Q_\Sigma^P = 0\} \subseteq \{|Q_\Sigma^N - Q_\Sigma^P| \geq \hat{B}\}$ . We now bound each of the elements on the last line of (T19). First, recall that  $Q_\Sigma^P = (X_\Sigma^P - N_\Sigma)^+$  so that

$$\mathbb{P}\{Q_\Sigma^P > 0\} \leq \mathbb{P}\{(X_\Sigma^P - N_\Sigma)^+ \geq 0\} = \mathbb{P}\{W^P > 0\}. \quad (\text{T20})$$

Second, by Markov's inequality

$$\mathbb{P}\{|Q_\Sigma^N - Q_\Sigma^P| \geq \hat{B}\} \leq \frac{\mathbb{E}[|Q_\Sigma^N - Q_\Sigma^P|]}{\hat{B}} = \frac{\mathbb{E}[Q_\Sigma^N] - \mathbb{E}[Q_\Sigma^P]}{\hat{B}},$$

where the last equality follows from the fact that  $Q_\Sigma^N \geq_{st} Q_\Sigma^P$ ; see (A11). Using the first inequality in (A17) we conclude that

$$\mathbb{P}\{|Q_\Sigma^N - Q_\Sigma^P| \geq \hat{B}\} \leq \frac{2}{\hat{B}} \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \frac{\epsilon}{\hat{B}} \mathbb{E}[Q_\Sigma^P]. \quad (\text{T21})$$

Combining (T20) and (T21) in (T19) we get

$$\mathbb{P}\{Q_i^N \geq \hat{B}\} \leq \mathbb{P}\{W^P > 0\} + \frac{2}{\hat{B}} \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \frac{\epsilon}{\hat{B}} \mathbb{E}[Q_\Sigma^P]. \quad (\text{T22})$$

Since  $W^P$  has the distribution of the waiting time in an  $M/M/N$  we have that  $p(\lambda_\Sigma, \mu, N_\Sigma) = \mathbb{P}\{W^P > 0\}$  where  $p(\cdot, \cdot, \cdot)$  is as defined prior to the statement of Theorem 6. Similarly, since  $W_i(0, 0)$  has the distribution of the waiting time in an  $M/M/N$  queue with parameters  $\lambda_i, \mu$  and  $N_i$ ,  $\mathbb{P}\{W_i(0, 0) > 0\} = p(\lambda_i, \mu, N_i)$ . By the assumptions of the theorem  $p(\lambda_\Sigma, \mu, N_\Sigma) \leq \min\{p(\lambda_1, \mu, N_1), p(\lambda_2, \mu, N_2)\}$  so that  $\mathbb{P}\{W^P > 0\} \leq \mathbb{P}\{W_i(0, 0) > 0\}$ . Using this together with (T22) and Lemma T.4 we have that

$$\begin{aligned} \mathbb{E}[Q_i^N] &\leq \hat{B} + \mathbb{P}\{W_i(0, 0) > 0\} \frac{\lambda_i}{\mu N_i - \lambda_i} + \left( \frac{2}{\hat{B}} \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \frac{\epsilon}{\hat{B}} \mathbb{E}[Q_\Sigma^P] \right) \frac{\lambda_i}{N\mu - \lambda_i} \\ &= \hat{B} + \mathbb{E}[Q_i(0, 0)] + \left( \frac{2}{\hat{B}} \mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \frac{\epsilon}{\hat{B}} \mathbb{E}[Q_\Sigma^P] \right) \frac{\lambda_i}{N\mu - \lambda_i}. \end{aligned}$$

Here, the last equality follows from the fact that,  $Q_i^0$  has the steady-state distribution of the queue length in an  $M/M/N$  with the parameters  $\lambda_i, \mu$  and  $N_i$  so that  $\mathbb{E}[Q_i^0] = \mathbb{P}\{W_i(0, 0) > 0\} \lambda_i (\mu N_i - \lambda_i)$ . Applying Little's law we conclude that:

$$\mathbb{E}[W_i(N_1, N_2)] = \frac{\mathbb{E}[Q_i^N]}{\bar{\lambda}_i} \leq \frac{\lambda_i}{\bar{\lambda}_i} \mathbb{E}[W_i(0, 0)] + \frac{\lambda_i}{\bar{\lambda}_i} \left( \frac{\hat{B}}{\lambda_i} + \frac{1}{\hat{B}} (2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \mathbb{E}[Q_\Sigma^P]) \frac{1}{N_i \mu - \lambda_i} \right), \quad (\text{T23})$$

where  $\bar{\lambda}_i := \bar{\lambda}_i(N_1, N_2)$  is the effective arrival rate to ED  $i$  under the threshold pair  $(N_1, N_2)$ ; see equation (2). As a function of  $\hat{B} > 0$  the right hand side of (T23) is convex and optimizing over  $\hat{B}$  we get

$$\hat{B}^* = \sqrt{\frac{\lambda_i 2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \mathbb{E}[Q_\Sigma^P]}{(\mu N_i - \lambda_i)}}. \quad (\text{T24})$$

In turn,

$$\mathbb{E}[W_i(N_1, N_2)] \leq \frac{\lambda_i}{\bar{\lambda}_i} \mathbb{E}[W_i(0, 0)] + 2 \frac{\lambda_i}{\bar{\lambda}_i} \sqrt{\frac{2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2) + \epsilon \mathbb{E}[Q_\Sigma^P]}{\lambda_i (N_i \mu - \lambda_i)}}. \quad (\text{T25})$$

By basic properties of the  $M/M/N$  queue  $\mathbb{E}[Q_\Sigma^P] \leq \lambda / (\mu N - \lambda)$  and

$$\frac{1}{\mu N_i - \lambda_i} = \mathbb{E}[W_i(0, 0) | W_i(0, 0) > 0] = \frac{\mathbb{E}[W_i(0, 0)]}{\mathbb{P}\{W_i(0, 0) > 0\}},$$

so that, in (T25), we can replace  $1/(\mu N_i - \lambda_i)$  with  $(\mu N_i - \lambda_i) (\mathbb{E}[W_i(0, 0)] / \mathbb{P}\{W_i(0, 0) > 0\})^2$  to get

$$\mathbb{E}[W_i(N_1, N_2)] \leq \frac{\lambda_i}{\bar{\lambda}_i} \mathbb{E}[W_i(0, 0)] + 2 \frac{\lambda_i}{\bar{\lambda}_i} \frac{\mathbb{E}[W_i(0, 0)]}{\mathbb{P}\{W_i(0, 0) > 0\}} \sqrt{(\mu N_i - \lambda_i) \left( \frac{2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_i} + \frac{\epsilon \lambda}{\lambda_i (\mu N - \lambda)} \right)}.$$

Define

$$\delta_i := \sqrt{(\mu N_i - \lambda_i) \left( \frac{2\mu T(\epsilon) C(\tilde{\rho}_1, \tilde{\rho}_2)}{\lambda_i} + \frac{\epsilon \lambda}{\lambda_i (\mu N - \lambda)} \right)}.$$

Then,

$$\mathbb{E}[W_i(N_1, N_2)] \leq \frac{\lambda_i}{\bar{\lambda}_i} \mathbb{E}[W_i(0, 0)] \left( 1 + \frac{\delta_i}{\mathbb{P}\{W_i(0, 0) > 0\}} \right).$$

The proof is now concluded by applying Theorem 7 to bound  $\lambda_i / \bar{\lambda}_i$ . ■

**Proof of Lemma T.4:** Fix  $\hat{B} > 0$  as in statement of the lemma. Since the threshold pair is fixed to  $(N_1, N_2)$  we omit it from the notation. Note that whenever  $X_1(t) \geq N_1$  there are no diversions into ED 1. Let  $\pi_1(i) = \sum_j \pi(i, j)$ . Multiplying and dividing by the same elements we write

$$\pi_1(N_1 + \hat{B} + l) = \pi_1(N_1 + \hat{B}) \prod_{m=1}^l \frac{\pi_1(N_1 + \hat{B} + m)}{\pi_1(N_1 + \hat{B} + m - 1)}$$

and, in turn,

$$\mathbb{P}\left\{X_1 = N_1 + \hat{B} + l | X_1 \geq N_1 + \hat{B}\right\} = \frac{\pi_1(N_1 + \hat{B} + l)}{\sum_{k=0}^{\infty} \pi_1(N_1 + \hat{B} + k)} = \frac{\prod_{m=1}^l \frac{\pi_1(N_1 + \hat{B} + m)}{\pi_1(N_1 + \hat{B} + m - 1)}}{1 + \sum_{k=1}^{\infty} \prod_{m=1}^{i=k} \frac{\pi_1(N_1 + \hat{B} + m)}{\pi_1(N_1 + \hat{B} + m - 1)}}. \quad (\text{T26})$$

Consider now the transition rates between the set of states  $\mathcal{A}_m = \{(N + \hat{B} + m, j), j \in \mathbb{Z}_+\}$  and the set of states  $\mathcal{A}_{m+1} = \{(N + \hat{B} + m + 1, j), j \in \mathbb{Z}_+\}$ . Then, by a standard result

$$\sum_{x \in \mathcal{A}_m, y \in \mathcal{A}_{m+1}} \pi(x) q_{xy} = \sum_{y \in \mathcal{A}_{m+1}, x \in \mathcal{A}_m} \pi(y) q_{yx},$$

where  $q_{xy}$  is the transition rate from state  $x$  to state  $y$ ; see e.g. Exercise 5.34 in Ross [1996]. Note that the only transitions from  $\mathcal{A}_m$  to  $\mathcal{A}_{m+1}$  be transitions of the form  $(i, j) \rightarrow (i + 1, j)$  and transition in the reverse direction can be only of the form  $(i + 1, j) \rightarrow (i, j)$ . Also, note that transitions of the latter form happen at rate  $N_1\mu$  and transition of the former happen only from state in which  $j \geq N_2$  (otherwise the arrival is diverted). Hence, we have that  $\sum_{x \in \mathcal{A}_m, y \in \mathcal{A}_{m+1}} \pi(x) q_{xy} \leq \lambda_1 \sum_j \pi(i, j)$  and  $\sum_{y \in \mathcal{A}_m, x \in \mathcal{A}_{m+1}} \pi(y) q_{yx} = \sum_j N_1\mu \pi(i + 1, j)$ . Since we can repeat this argument for any  $m$ , we have that for all  $m \geq 1$ ,

$$\frac{\pi_1(N_1 + \hat{B} + m)}{\pi_1(N_1 + \hat{B} + m - 1)} \leq \frac{\lambda_1}{N_1\mu}. \quad (\text{T27})$$

We will now use a simple calculus result. Let  $\{a_n\}_{n \geq 1}, \{\epsilon_n\}_{n \geq 1}$  be sequences of non-negative numbers. Let  $\bar{a}_n = a_n(1 + \epsilon_n)$ ,  $b_m := \prod_{n=1}^m a_n$  and  $\bar{b}_m = \prod_{n=1}^m \bar{a}_n$ . Then,

$$\frac{\sum_{l=1}^{\infty} l b_l}{\sum_{l=1}^{\infty} b_l} \leq \frac{\sum_{l=0}^{\infty} l \bar{b}_l}{\sum_{l=0}^{\infty} \bar{b}_l}. \quad (\text{T28})$$

We omit the proof of this simple claim. For our setting choose  $a_0 = b_0 = \bar{a}_0 = \bar{b}_0 = 1$ ,  $a_m = \frac{\pi_1(N_1 + \hat{B} + m)}{\pi_1(N_1 + \hat{B} + m - 1)}$  and  $\bar{a}_m = \lambda_1 / (N_1\mu)$  and let  $b_m$  and  $\bar{b}_m$  be defined from these as above. Then, using (T26) and we have that

$$\begin{aligned} \mathbb{E}[(X_1 - (N_1 + \hat{B})) | X_1 \geq N_1 + \hat{B}] &= \sum_{l=0}^{\infty} l \mathbb{P}\left\{X_1 = N_1 + \hat{B} + l | X_1 \geq N_1 + \hat{B}\right\} \\ &= \frac{\sum_{l=0}^{\infty} l b_l}{\sum_{l=0}^{\infty} b_l} \leq \frac{\sum_{l=0}^{\infty} l \bar{b}_l}{\sum_{l=0}^{\infty} \bar{b}_l} \leq \frac{\sum_{l=0}^{\infty} l \bar{b}_l}{\sum_{l=0}^{\infty} \bar{b}_l} \\ &= \left(1 - \frac{\lambda_1}{N_1\mu}\right) \sum_{l=0}^{\infty} l \left(\frac{\lambda_1}{N_1\mu}\right)^l = \frac{\lambda_1}{N_1\mu - \lambda_1}. \end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}[Q_1] &= \mathbb{E}[(X_1 - N_1)^+] \\
&= \mathbb{E}[(X_1 - N_1)^+ | X_1 < N_1 + \hat{B}] \mathbb{P}\{X_1 < N_1 + \hat{B}\} \\
&\quad + \mathbb{P}\{X_1 \geq N_1 + \hat{B}\} \left( \hat{B} + \mathbb{E}[X_1 - (N_1 + \hat{B}) | X_1 \geq N_1 + \hat{B}] \right) \\
&\leq \hat{B} \mathbb{P}\{X_1 < N_1 + \hat{B}\} + \hat{B} \mathbb{P}\{X_1 \geq N_1 + \hat{B}\} + \mathbb{P}\{X_1 \geq N_1 + \hat{B}\} \frac{\lambda_1}{N_1 \mu - \lambda_1},
\end{aligned}$$

which concludes the proof. ■

### T.1.4 Proof of Theorem 7

The threshold pair is fixed to  $(N_1, N_2)$  and we omit it from the notation. We focus on ED 1 and the proof is identical for ED 2. The logic of the proof is as follows: since ED 1 does not divert ambulances on time periods on which its patient count is below  $N_1$ , all arrivals (ambulance or walk-ins) from catchment area 1 are routed to ED 1 rather than diverted so that the arrival rate is  $\lambda_1$  during these periods. In terms of service completions, when the patient count is smaller than

$$n_1 = \left\lfloor \frac{\lambda_1}{\mu} - \sqrt{\frac{\lambda_1}{\mu}} \right\rfloor, \quad (\text{T29})$$

the rate of service completions is at most  $\lambda_1 - \sqrt{\mu\lambda_1}$ . Consequently, the excursions below the level  $\lambda_1 - \sqrt{\mu\lambda_1}$  can be bounded, through a coupling argument, by those of an appropriately defined  $M/M/1$  queue with arrival rate  $\lambda_1 - \sqrt{\mu\lambda_1}$  and service rate  $\lambda_1$ .

We will use such a coupling argument to show that

$$\mathbb{E} \left[ \left( Z_1 - \frac{\lambda_1}{\mu} \right)^- \right] \leq \sqrt{\frac{\lambda_1}{\mu}} + 1 + \frac{(\lambda_1 - \sqrt{\mu\lambda_1})/\lambda_1}{1 - (\lambda_1 - \sqrt{\mu\lambda_1})/\lambda_1} \leq \sqrt{\frac{\lambda_1}{\mu}} + 1 + \frac{\lambda_1 - \sqrt{\mu\lambda_1}}{\sqrt{\mu\lambda_1}} \leq 2\sqrt{\frac{\lambda_1}{\mu}}, \quad (\text{T30})$$

so that (since for  $x, y \in \mathbb{R}$ ,  $x - y = (x - y)^+ - (x - y)^-$ ) we have  $\mathbb{E}[Z_1] \geq \lambda_1/\mu - \mathbb{E}[(Z_1 - \lambda_1/\mu)^-]$ , and, in turn,  $\mu\mathbb{E}[Z_1] \geq \lambda_1 - 2\sqrt{\lambda_1\mu}$ . In steady-state, the effective arrival rate is equal to the departure rate so that

$$\bar{\lambda}_i(N_1, N_2) = \mu\mathbb{E}[Z_1] \geq \lambda_1 - 2\sqrt{\lambda_1\mu},$$

which will conclude the proof the theorem.

The rest of the proof is the formalization of the coupling argument that leads to (T30). For that purpose, we use a different (and simpler) construction of the sample paths than the one we used in §A.2.1. The difference is in terms of the generation of service completions. Here, instead of having a single Poisson process  $S(\cdot)$  from which we construct the service completions, we have two such processes:  $S_1(\cdot)$  for ED 1 and  $S_2(\cdot)$  for ED 2 with respective rates  $\mu N_1$  and  $\mu N_2$ . We have two sequences of i.i.d. Uniform  $[0, 1]$  random variables  $\{U_n^1, n \in \mathbb{Z}_+\}$  and  $\{U_n^2, n \in \mathbb{Z}_+\}$ . When  $S_i(\cdot)$  jumps for the  $n^{\text{th}}$  time (let this jump time be denoted by  $\tau_{s,i}^n$ ) this jump triggers and actual service completion in ED  $i$  if  $U_n^i \in [0, Z_1(\tau_{s,i}^n -)/N_1]$  and

it does not trigger such a completion otherwise. As before we use  $\tau_{i,a}^n$  to denote the time of the  $n^{\text{th}}$  jump of the process  $A_i^a(\cdot)$ . We then write:

$$\begin{aligned} X_1(t) &= X_1(0) + A_1^w(t) \\ &+ \sum_{n=1}^{A_1^a(t)} \left( \mathbb{1}\{X_1(\tau_{1,a}^n -) < N_1\} + \mathbb{1}\{X_1(\tau_{1,a}^n -) \geq N_1, X_2(\tau_{1,a}^n -) \geq N_2\} \right) \\ &+ \sum_{n=1}^{A_2^a(t)} \mathbb{1}\{X_2(\tau_{2,a}^n -) \geq N_2, X_1(\tau_{2,a}^n -) < K_1\} - \sum_{n=1}^{S_1(t)} \mathbb{1}\{U_n^1 \in [0, Z_1(\tau_{s,1}^n -)/N_1]\}. \end{aligned} \quad (\text{T31})$$

We write a similar equation for ED 2. These generate the correct probability law for the CTMC  $X(\cdot)$ .

We proceed to prove (T30). Note that, by (1),  $n_1 < N_1 = K_1$ . Let  $A_1(\cdot) = A_1^w(\cdot) + A_1^a(\cdot)$  (i.e.,  $A_1(t)$  captures the total number of patients arriving from catchment area 1 by time  $t$ ). Let  $(s, t]$  be an interval such that  $X_1(u) \leq n_1 < K_1$  for all  $u \in (s, t]$ . Then, by (T31), for all such  $u$ ,

$$\begin{aligned} X_1(u) - X_1(s) &= A_1(t) - A_1(s) - \sum_{n=1}^{S_1(t)} \mathbb{1}\{U_n^1 \in [0, Z_1(\tau_{s,1}^n -)/N_1]\} \\ &\geq A_1(t) - A_1(s) - \sum_{n=1}^{S_1(t)} \mathbb{1}\{U_n^1 \in [0, n_1/N_1]\}. \end{aligned}$$

Define a process  $M_1(\cdot)$  as follows:

$$M_1(t) = M_1(0) + \sum_{n=1}^{S_1(t)} \mathbb{1}\{U_n^1 \in [0, n_1/N_1]\} - \int_s^t \mathbb{1}\{M_1(u) > 0\} dA_1(u).$$

Then,  $M_1(t) - M_1(s) = -(X_1(t) - X_1(s))$  on intervals  $(s, t]$  s.t.  $X_1(u) < n_1$  for all  $u \in (s, t]$  so that almost surely

$$(X_1(t) - n_1)^- \leq M_1(t), \text{ for all } t \geq 0.$$

Since  $n_1 < N_1$  and since each ED is work conserving (there can be no positive queue while there are available beds) we have that  $(X_1(t) - n_1)^- = (Z_1(t) - n_1)^-$ . In turn,

$$(Z_1(t) - n_1)^- = (X_1(t) - n_1)^- \leq M_1(t), \text{ for all } t \geq 0. \quad (\text{T32})$$

Moreover,  $M_1(\cdot)$  has the law of an  $M/M/1$  queue with arrival rate  $\mu n_1 \leq \lambda_1 - \sqrt{\mu \lambda_1}$  and service rate  $\lambda_1$ . It has the utilization  $\tilde{\rho} = \mu n_1 / \lambda_1 < 1$ . In turn,  $M_1(t) \Rightarrow M_1$  as  $t \rightarrow \infty$  where  $M_1$  has the steady-state distribution of this  $M/M/1$  queue. In turn, taking  $t \rightarrow \infty$  on both sides of (T32) and applying expectations we have that

$$\mathbb{E}[(Z_1 - n_1)^-] \leq \mathbb{E}[M_1] \leq \frac{\tilde{\rho}}{1 - \tilde{\rho}} \leq \frac{(\lambda_1 - \sqrt{\mu \lambda_1})/\lambda_1}{1 - (\lambda_1 - \sqrt{\mu \lambda_1})/\lambda_1},$$

which by the definition of  $n_1$  establishes (T30) and concludes the proof of the theorem. ■

### T.1.5 Existence of optimal solutions

For the following recall that  $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+ \cup \{\infty\}$ .

**Lemma T.5** *There exists a threshold pair  $K^* := (K_1^*, K_2^*) \in \bar{\mathbb{Z}}_+^2$  such that*

$$\mathbb{E}[W(K_1^*, K_2^*)] = \min_{K_1, K_2} \mathbb{E}[W(K_1, K_2)].$$

**Proof:** We first extend the definition of  $\mathbb{E}[W(\cdot, \cdot)]$  to allow for instability. To that end, given  $K = (K_1, K_2)$  define

$$\mathbb{E}[W(K_1, K_2)] := \limsup_{T \rightarrow \infty} \frac{1}{\lambda_\Sigma T} \int_0^T (Q_1^K(s) + Q_2^K(s)) ds, \quad (\text{T33})$$

where, as before  $\lambda_\Sigma = \lambda_1 + \lambda_2$  and where the lim sup maybe infinite. When the CTMC  $X^K(\cdot)$  is stable this definition is, by Little's law and the ergodic theorems, consistent with  $\mathbb{E}[W(K_1, K_2)] = \mathbb{E}[Q_\Sigma^K]/\lambda_\Sigma$ . For purposes of optimality it suffices to consider threshold pairs in the set  $\mathcal{A} := \{K \in \bar{\mathbb{Z}}_+^2 : \mathbb{E}[W(K_1, K_2)] < \infty\}$ . Note that the set  $\mathcal{A}$  may contain also threshold pairs  $K$  with  $K_1 \vee K_2 = \infty$ .

With these definitions  $\inf_{K \in \bar{\mathbb{Z}}_+^2} \mathbb{E}[W(K_1, K_2)] = \inf_{K \in \mathcal{A}} \mathbb{E}[W(K_1, K_2)]$ . By definition,

$$0 \leq \inf_{K \in \mathcal{A}} \mathbb{E}[W(K_1, K_2)] \leq \mathbb{E}[W(0, 0)] = \frac{\mathbb{E}[Q_\Sigma^0]}{\lambda_\Sigma} = \frac{\mathbb{E}[Q_1^0] + \mathbb{E}[Q_2^0]}{\lambda_\Sigma} < \infty, \quad i = 1, 2 \quad (\text{T34})$$

where the last inequality follows from the fact that  $Q_i^0$  is the steady-state queue length in an  $M/M/N$  queue with parameters  $\lambda_i, \mu$  and  $N_i$  and satisfying  $N_i > \lambda_i/\mu$  by (1); see Remark 2 in the manuscript.

To show that a minimizer exists for  $\inf_{K \in \mathcal{A}} \mathbb{E}[W(K_1, K_2)]$  we define, for  $B \in \mathbb{Z}_+$ ,

$$\mathcal{A}^B = \{K \in \mathbb{Z}_+^2 : \max\{K_1, K_2\} \leq B\}.$$

The set  $\mathcal{A}^B$  is compact and the function  $\mathbb{E}[W(K_1, K_2)]$  is trivially continuous on this set. In turn,

$$\mathcal{A}^{*,B} := \operatorname{argmin}_{K \in \mathcal{A}^B} \mathbb{E}[W(K_1, K_2)] \neq \emptyset,$$

and we can construct a sequence of (locally optimal) threshold pairs  $\{K^B; B \in \mathbb{Z}_+\}$  by choosing one element of  $\mathcal{A}^{*,B}$  for each  $B$ . By definition, the sequence  $\{\mathbb{E}[W(K_1^B, K_2^B)]; B \in \mathbb{Z}_+\}$  is non-increasing sequence in  $B$  and has a limit (which may be infinite). Also, for each finite  $B$ ,  $\mathbb{E}[Q_\Sigma^{K^B}] = \lambda_\Sigma \mathbb{E}[W(K_1^B, K_2^B)]$  by Little's law so that the sequence  $\{\mathbb{E}[Q_\Sigma^{K^B}], B \in \mathbb{Z}_+\}$  also has a (possibly infinite) limit.

There are two cases to consider. If the sequence  $\{K^B; B \in \mathbb{Z}_+\}$  can be chosen so that

$$\sup_{B \in \mathbb{Z}_+} \max\{K_1^B, K_2^B\} \leq D < \infty$$

for some constant  $D$ , then, an optimal solution to (5) trivially exists and it is an element of  $\left\{ \operatorname{argmin}_{K \in \mathcal{A}^D} \mathbb{E}[W(K_1, K_2)] \right\}$ . Consider the case in which any such sequence of locally optimal thresholds has

$$\sup_{B \in \mathbb{Z}_+} \max\{K_1^B, K_2^B\} = \infty. \quad (\text{T35})$$

We will show that in this case there exists a minimizer to (5) that has  $\bar{K}_i = \infty$  for some  $i = 1, 2$  or both. We will now make use of the following auxiliary lemma whose proof appears at the end of this subsection.

**Lemma T.6** *Let  $\{K^l, l \in \mathbb{Z}_+\}$  be a sequence of finite threshold pairs such that*

$$\sup_{l \in \mathbb{Z}_+} \mathbb{E}[W(K_1^l, K_2^l)] \leq \mathbb{E}[W(0, 0)],$$

*and such that  $K^l \rightarrow \bar{K} \in \bar{\mathbb{Z}}_+^2$  as  $l \rightarrow \infty$ . Then,  $X^{K^l} \Rightarrow Y$  as  $l \rightarrow \infty$ , as  $l \rightarrow \infty$ , where the random variable  $Y$  has the steady-state distribution,  $\pi^{\bar{K}}$  of the CTMC  $X^{\bar{K}}(\cdot)$ .*

An implicit by-product of this result is that a steady-state distribution exists for the CTMC  $X^{\bar{K}}(\cdot)$ . We apply it now to conclude the proof of Lemma T.5. Note that if (T35) holds, then there exists a further subsequence  $\{B_l\}_{l \geq 1}$  that converges to a limit (that is possibly infinite). Let  $\bar{K} = (\bar{K}_1, \bar{K}_2)$  be this limit. Recall that  $Q_\Sigma^K = (X_1^K - N_1)^+ + (X_2^K - N_2)^+$ . From Lemma T.6 we have that as  $l \rightarrow \infty$ ,  $Q_\Sigma^{K^{B_l}} \Rightarrow Q_\Sigma^{\bar{K}}$ , and, using Fatou's lemma that

$$\mathbb{E}[Q_\Sigma^{\bar{K}}] \leq \liminf_{l \rightarrow \infty} \mathbb{E}[Q_\Sigma^{K^{B_l}}] \leq \mathbb{E}[Q_\Sigma^0] < \infty.$$

Applying Little's law we have that  $\mathbb{E}[W(\bar{K}_1, \bar{K}_2)] \leq \lim_{l \rightarrow \infty} \mathbb{E}[W(K_1^{B_l}, K_2^{B_l})] < \infty$ , so that  $(\bar{K}_1, \bar{K}_2)$  is optimal for (5). This concludes the proof.  $\blacksquare$

**Proof of Lemma T.6:** First, note that if  $\max\{\bar{K}_1, \bar{K}_2\} < \infty$  then, since  $\bar{K}_i$  obtains only integer values,  $K^l \rightarrow \bar{K}$  implies that  $(K_1^l, K_2^l) = (\bar{K}_1, \bar{K}_2)$  for all large enough  $l$  so that  $X^{K^l} = X^{\bar{K}}$  for all  $l$  large enough and it trivially follows that  $X^{K^l} \Rightarrow X^{\bar{K}}$ .

We only need consider then the case  $\bar{K}_1 \vee \bar{K}_2 = \infty$ . From here, the argument is similar in spirit to the proof of Lemma T.7.

$Q^{K^l}$  be the generator matrix for the CTMC  $X^{K^l}(t)$ . Theorem 1 and the assumed finiteness of  $K^l$  guarantee the existence of a steady-state distribution  $\pi^{K^l}$  that is the unique non-negative solution to  $\pi^{K^l} Q^{K^l} = 0$  and  $\sum_{(i,j) \in \mathbb{Z}_+^2} \pi^{K^l}(i, j) = 1$ . Similarly, let  $Q^{\bar{K}}$  be the generator matrix of the CTMC  $X^{\bar{K}}(\cdot)$ .

We will establish that (a) the sequence  $\pi^{K^l}$  is tight (b) every limit point,  $\pi$ , satisfies  $\pi Q^{\bar{K}} = 0$  and  $\sum_{(i,j) \in \mathbb{Z}_+^2} \pi(i, j) = 1$  so that  $\pi$  is equal to  $\pi^{\bar{K}}$ , the steady-state distribution of the chain  $X^{\bar{K}}(t)$ . The tightness of the sequence  $Q_\Sigma^{K^l}$  follows immediately from the assumption that  $\mathbb{E}[Q_\Sigma^{K^l}] \leq \mathbb{E}[Q_\Sigma^0]$  for all  $l \in \mathbb{Z}_+$ . Since  $X_1^{K^l} + X_2^{K^l} \leq N_1 + N_2 + Q_\Sigma^{K^l}$  the tightness of the sequence  $X^{K^l}$  follows from that of the sequence  $Q_\Sigma^{K^l}$ .

We consider only the case that  $\bar{K}_1 < \infty$  and  $\bar{K}_2 = \infty$  (at the end of the proof we indicate how to modify the argument for the other cases). In this case, for all  $l$  large enough  $(K_1^l, K_2^l) = (\bar{K}_1, K_2^l)$ . For all large

enough  $l$  the entries of the generator  $\mathcal{Q}^{K^l}$  are equal to those of  $\mathcal{Q}$  for all states  $(i, j) \in \mathbb{Z}_+^2$  with  $j < K_1^l$ . Since the entries of the generator are bounded by  $\lambda + N\mu$  we get

$$\sup_{(i,j)} |[\pi^l(\mathcal{Q}^{K^l} - \mathcal{Q}^{\bar{K}})]_{(i,j)}| \leq 4(\lambda + N\mu) \sum_{(i,j): j \geq K_2^l/2} \pi_{i,j}^l \rightarrow 0 \text{ as } l \rightarrow \infty, \quad (\text{T36})$$

where  $[\pi^l(\mathcal{Q}^{K^l} - \mathcal{Q}^{\bar{K}})]_{(i,j)}$  is the  $(i, j)$  entry of  $\pi^l(\mathcal{Q}^{K^l} - \mathcal{Q}^{\bar{K}})$  and the convergence follows from the tightness of the sequence  $X^{K^l}$ . Since  $\pi^l \mathcal{Q}^l = 0$  for all  $l \in \mathbb{Z}_+$  we conclude from (T36) that  $\pi^l \mathcal{Q}^{\bar{K}} \rightarrow 0$ . Using the tightness, pick now a convergent subsequence  $\{\pi^{l_k}, k \in \mathbb{Z}_+\}$  and let  $\tilde{\pi}$  be the limit point of this subsequence. In particular,  $\tilde{\pi}$  satisfies  $\sum_{(i,j) \in \mathbb{Z}_+^2} \pi(i, j) = 1$ . By the above it also satisfies that  $\pi^{l_k} \mathcal{Q}^{\bar{K}} \rightarrow 0$ . Using the bounded entries of the  $\mathcal{Q}^{\bar{K}}$  we conclude that  $\tilde{\pi}$  is a probability distribution that solves  $\tilde{\pi} \mathcal{Q}^{\bar{K}} = 0$  with  $\mathcal{Q}$  being the generator of the CTMC  $X(\cdot)$ .

Finally, since the CTMC  $X(\cdot)$  has bounded transition rates it is, in particular, non explosive and we may conclude that the CTMC  $X(\cdot)$  is ergodic and  $\tilde{\pi}$  is its steady-state distribution; see e.g. Theorem II.4.3 in Asmussen [2003]. Since the same argument can be applied to any convergent subsequence we have, by Prohorov's theorem, that  $X^{K^l} \Rightarrow X^{\bar{K}}$  as required.

The argument above requires only minor modifications in the case  $\bar{K}_1 = \infty$  and  $\bar{K}_2 < \infty$  or in the case that  $\bar{K}_1 = \bar{K}_2 = \infty$ . For the former, the right hand side in (T36) is replaced with  $4(\lambda + N\mu) \sum_{(i,j): i \geq K_1^l} \pi_{i,j}^l$  and in the latter it is replaced by  $4(\lambda + N\mu) \sum_{(i,j): i \geq K_1^l, j \geq K_2^l} \pi_{i,j}^l$ . ■

## T.2 Convergence of truncated chains

In §6 of the manuscript, we truncate the state-space to approximate the steady-state distribution of  $X^K(\cdot)$ . The validity of such approximation is justified by a limit argument that shows that as the truncation parameter  $B$  grows, the steady-state distribution of the truncated chain approaches, in an appropriate sense, that of the original chain. In this section we formally state and prove this result.

Throughout, we add the superscript  $B$  to make explicit the dependence on the truncation parameter. Specifically,  $X^{K,B}(\cdot) = (X_1^{K,B}(\cdot), X_2^{K,B}(\cdot))$  is the CTMC corresponding to the truncated chain. When the superscript  $B$  is removed the notation refers to the original non-truncated chain. We note that a unique steady-state distribution exists for  $X^{K,B}(\cdot)$  by its irreducibility and its bounded state space. We remove the time argument when referring to steady-state quantities so that  $X^{K,B}$  has the corresponding steady-state distribution  $\pi^{K,B}$ . As before, we define  $\mathcal{A}^B := \{K \in \mathbb{Z}_+^2 : K_1 \vee K_2 \leq B\}$ .

**Lemma T.7** *Fix a finite threshold pair  $(K_1, K_2)$ . Then, as  $B \rightarrow \infty$ ,*

$$X^{K,B} \Rightarrow X^K, \text{ and } \mathbb{E}[Q_i^{K,B}] \rightarrow \mathbb{E}[Q_i^K], \quad i = 1, 2.$$

*In turn, for any  $B \in \mathbb{Z}_+$ , as  $M \rightarrow \infty$ ,*

$$\sup_{K \in \mathcal{A}^B} \left| \mathbb{E}[Q_i^{K,B}] - \mathbb{E}[Q_i^K] \right| \rightarrow 0, \quad i = 1, 2. \quad (\text{T37})$$

**Remark T.1** The literature on Markov chains provides numerous sufficient conditions for the convergence of the (sequence of) steady-state distributions to that of the infinite-state chain; see e.g. Wolf [1980] and the references therein. These sufficient conditions correspond to tightness of the sequence of steady-state distributions,  $\pi^{K,B}$ , and are often rather involved because of their generality. The structure of our Markov chain allows us to have a self-contained argument and prove this tightness and its implications from basic principles rather than applying any of the sufficient conditions provided by the literature. ■

A key step then is the following tightness result whose proof appears at the end of this subsection.

**Lemma T.8** *Fix a finite threshold pair  $K$ . Then, the sequence  $\{X^{K,B}, B \in \mathbb{Z}_+\}$  is uniformly integrable and, in particular, tight.*

**Proof of Lemma T.7:** We fix a finite pair  $K \in \mathbb{Z}_+^2$  and omit it from the notation. Since the sequence  $X^{B_l}$  is tight we can choose a subsequence  $\{B_l, l \in \mathbb{Z}_+\}$  such that  $X^{B_l}$  converges to a limit random variable  $Y$ . Denote the distribution of  $Y$  by  $\tilde{\pi}$ . Note that, since  $\tilde{\pi}$  is a distribution it satisfies  $\sum_{(i,j) \in \mathbb{Z}_+^2} \tilde{\pi}(i,j) = 1$ . We claim that  $\tilde{\pi}$  must then be the steady-state distribution of the (non-truncated) chain  $X(t)$ .

To that end, let  $Q^B$  be the generator of the Markov chain  $X^B(\cdot)$  expanded to all of  $\mathbb{Z}_+^2$  by defining  $Q_{i,j}^B = 0$  for all  $(i,j)$  with  $\max\{i,j\} > B$ . Let  $Q$  be the generator matrix of the original (non-truncated) chain. Similarly, we expand the matrix  $\pi^B$  to all of  $\mathbb{Z}_+$  by setting  $\pi^B(i,j) = 0$  for all  $(i,j)$  with  $\max\{i,j\} > B$ . Since transitions happen only to neighboring states and transition rates are bounded by  $\lambda + N\mu$ , we have that

$$\sup_{(i,j)} |[\pi^B(Q^B - Q)]_{(i,j)}| \leq 4(\lambda + N\mu) \sum_{(i,j): i \geq B/2 \text{ or } j \geq B/2} \pi^B(i,j),$$

where  $[\pi^B(Q^B - Q)]_{(i,j)}$  is the  $(i,j)$  entry of the matrix  $\pi^B(Q^B - Q)$ . The tightness of  $X^{K,B}$  guarantees that the right-hand side converges to 0 as  $B \rightarrow \infty$  and, in turn, that

$$\sup_{(i,j)} |[\pi^{B_l}(Q^{B_l} - Q)]_{(i,j)}| \rightarrow 0.$$

By definition,  $\pi^{B_l} Q^{B_l} = 0$  for all  $l \in \mathbb{Z}_+$ , so that  $\pi^{B_l} Q \rightarrow 0$  as  $l \rightarrow \infty$ .

Since  $\pi^{B_l} \Rightarrow \tilde{\pi}$  and  $Q$  has entries that are bounded by  $\lambda + N\mu$ , we conclude that  $\tilde{\pi}$  is a distribution that satisfies  $\tilde{\pi} Q = 0$  and  $\sum_{(i,j) \in \mathbb{Z}_+^2} \tilde{\pi}(i,j) = 1$ . Hence  $\tilde{\pi}$  is a stationary distribution for the stable chain  $X(\cdot)$ . By Theorem 1 there is a unique stationary distribution for  $X(\cdot)$ , hence,  $\tilde{\pi}$  must be this unique distribution. The same argument can be repeated for every convergent subsequence. Using Prohorov's theorem (see page 59 of Billingsley [1999]) we conclude that the whole sequence converges. Namely, that  $X^{K,B} \Rightarrow Y$  as  $B \rightarrow \infty$  where  $Y$  is distributed according to the steady-state distribution of the (un)truncated chain  $X^K(\cdot)$ .

This concludes the proof of weak convergence. Convergence of expectations then follows from the uniform integrability in Lemma T.8. Finally, (T37) follows from the finiteness of the set  $\mathcal{A}^B$ . ■

**Proof of Lemma T.8:** The sample path construction that we use here is the one we used in the proof of Theorem 6 rather than the one introduced in §A.2.1. The only difference is that here we work with truncated chains so we augment the construction by specifying that when there is a jump of  $A_1^w(\cdot)$  or  $A_1^a(\cdot)$  at time  $t$  in which  $X_1^B(t) = M$  and  $X_2^B(t) \geq N_2$  (or the symmetric case with ED 2), then  $X^B(t)$  remains unchanged.

This construction generate the correct probability law for  $X^B(t)$ .

We now define two  $M/M/1$  queues on the same probability space. Let  $\mathcal{M}_i(\cdot)$  be the process that goes up by 1 whenever there is a jump of any of the processes  $A_i^w(t)$  or  $A_i^a(t)$  and goes down by 1 whenever  $\mathcal{M}_i(t) > 0$  and there is a jump of  $S_i(\cdot)$ . With this construction  $\mathcal{M}_i(t)$  has the probability law of an  $M/M/1$  queue with arrival rate  $\lambda_i$  and service rate  $\mu N_i$ .

From here it is a matter of a simple coupling argument that is omitted to show that, initializing  $X_1^B(0) = X_2^B(0) = 0$  and initializing  $\mathcal{M}_i(t)$  with its stationary distribution (that of an  $M/M/1$  queue as above)

$$Q_i^B(t) \leq (K_i - N_i)^+ + \mathcal{M}_i(t), \text{ for all } t \geq 0, i = 1, 2$$

Taking  $t \rightarrow \infty$  we have that  $Q_i^B \leq (K_i - N_i)^+ + M_i$  where  $M_i$  has the unique steady-state distribution of  $(\mathcal{M}_i(t), t \geq 0)$  and, in turn, has a finite expectation by condition (1). Since this bound is independent of the truncation parameters  $B$ , the sequence  $Q^B = (Q_1^B, Q_2^B)$  is uniformly integrable and, in particular, tight and so is  $X^B$  because  $X_i^B \leq N_i + Q_i^B$ . ■

### T.3 Proofs of auxiliary lemmas and propositions

**Proof of Lemma A.1:** To prove (A6) consider the following three cases:

1. Let  $t$  be such that  $X_1(t) \geq N_1$  and  $X_2(t) < N_2$ . There are two further cases to consider
  - (a) Assume first that  $X_\Sigma(t) \geq N$  so that the total number of patients in the network is greater than  $N$ . In this region all the idleness is in ED 2, i.e,  $I_2(t) = I_1(t) + I_2(t)$ . Moreover, we must have that  $Q_1(t) \geq I_2(t)$  (otherwise, we would not have that  $X_\Sigma(t) \geq N$ ). Hence, in the case in which  $X_\Sigma(t) \geq N$ ,

$$Z_\Sigma(t) = N - I_2(t) = X_\Sigma(t) \wedge N_\Sigma - Q_1(t) \wedge I_2(t).$$

- (b) If, on the other hand,  $X_\Sigma(t) < N$  then we must have that  $I_2(t) \geq Q_1(t)$ . Here, we then have that  $Z_\Sigma(t) = X_\Sigma - Q_1(t) = X_\Sigma(t) \wedge N_\Sigma - Q_1(t) \wedge I_2(t)$ .

2. Let  $t$  be such that  $X_1(t) < N_1$  and  $X_2(t) \geq N_2$ . We then have (by similar arguments) that

$$Z_\Sigma(t) = X_\Sigma(t) \wedge N_\Sigma - Q_2(t) \wedge I_1(t).$$

3. Let  $t$  be such that  $X_1(t) \leq N_1$  and  $X_2(t) \leq N_2$  or such that both  $X_1(t) \geq N_1$  and  $X_2(t) \geq N_2$ . Here we have that  $Z_\Sigma(t) = X_\Sigma(t) \wedge N_\Sigma$  and also that  $Q_1(t) \wedge I_2(t) = Q_2(t) \wedge I_1(t) = 0$ . ■

**Proof of Proposition A.1:** We start with (A9). We will prove that almost surely,  $X_\Sigma(t) \geq X_\Sigma^P(t)$  and  $Z_\Sigma(t) \leq Z_\Sigma^P(t)$  for all  $t \geq 0$ . Since  $Q_\Sigma(t) = X_\Sigma(t) - Z_\Sigma(t)$  this will establish (A9).

We use an induction on the jumps of  $A(t)$  and  $S(t)$  defined in §A.2.1. To that end, put  $\tau_0 = 0$  and define inductively, for  $k \geq 1$ ,

$$\tau_k = \inf\{t \geq 0 : \tau_{k-1} : A(t) - A(t-) > 0 \text{ or } S(t) - S(t-) > 0\}.$$

By assumption  $X_\Sigma^P(0) = X_\Sigma(0)$  and  $Z_\Sigma^P(0) = X_\Sigma(0) \wedge N_\Sigma \geq Z_\Sigma(0)$  where the last inequality follows from the properties of the threshold system; see equation (A6). Hence, the ordering holds trivially for all  $t < \tau_1$ . Fix now  $k > 1$ . Assume that the ordering is valid for all  $t < \tau_{k-1}$  and consider the time  $\tau_k$ . There are two cases to consider:

1. The event at time  $\tau_k$  is a jump of  $A(\cdot)$ : In this case the arrival will be an arrival to both systems and both  $X_\Sigma(t)$  and  $X_\Sigma^P(t)$  will increase by one, thus preserving the ordering. As  $Z_\Sigma^P(t) \geq Z_\Sigma(t)$  for all  $t < \tau_{k-1}$  we have the following at  $\tau_k$ : If  $Z_\Sigma^P(\tau_k-) < N$  then  $Z_\Sigma^P$  will go up by one with this arrival while  $Z_\Sigma$  will not necessarily go up by one (for example if  $Z_1 = N_1$  and the arrival is a walk-in to ED 1). If, on the other hand,  $Z_\Sigma^P(\tau_k-) = N$  but  $Z_\Sigma(\tau_k-) < N$  then, necessarily,  $Z_\Sigma^P(\tau_k-) > Z_\Sigma(\tau_k-)$  so that the comparison is preserved even if  $Z_\Sigma(\cdot)$  jumps up by one.
2. The event at time  $\tau_k$  is a jump of  $S(\cdot)$ : If  $Z_\Sigma^P(\tau_k-) = Z_\Sigma(\tau_k-)$ , since we use the same sequence  $\{U_n\}_{n \geq 1}$  for both systems, this jump will be an actual service completion in both systems or in none thus preserving the ordering. If  $Z_\Sigma^P(\tau_k-) > Z_\Sigma(\tau_k-)$ , this jump might trigger a completion only in the pooled system in which case we will still have  $Z_\Sigma^P(\tau_k) \geq Z_\Sigma(\tau_k)$ . Note that  $X_\Sigma^P(\cdot)$  might decrease at  $\tau_k$  while  $X_\Sigma(\cdot)$  does not. This, however, preserves the inequality  $X_\Sigma^P(t) \leq X_\Sigma(t)$ .

Since the probability of a simultaneous jump of  $S(\cdot)$  and  $A(\cdot)$  is 0 there are no additional cases to consider.

This concludes the proof of (A9) and we turn to prove (A10). Using (A8) and (A5) we have:

$$X_\Sigma(t) - X_\Sigma^P(t) = \sum_{n=1}^{S(t)} \mathbb{1} \left\{ U_n \in \left[ 0, \frac{X_\Sigma^P(\tau_s^n-) \wedge N_\Sigma}{N_\Sigma} \right] \right\} - \mathbb{1} \left\{ U_n \in \left[ 0, \frac{X_\Sigma(\tau_s^n-) \wedge N_\Sigma - C(\tau_s^n-)}{N_\Sigma} \right] \right\}.$$

Using (A9) we have, however, that

$$\mathbb{1} \left\{ U_n \in \left[ 0, \frac{X_\Sigma(\tau_s^n-) \wedge N_\Sigma - C(\tau_s^n-)}{N_\Sigma} \right] \right\} \geq \mathbb{1} \left\{ U_n \in \left[ 0, \frac{X_\Sigma^P(\tau_s^n-) \wedge N_\Sigma - C(\tau_s^n-)}{N_\Sigma} \right] \right\},$$

for all  $n \in \mathbb{Z}_+$  so that

$$\begin{aligned} 0 &\leq X_\Sigma(t) - X_\Sigma^P(t) \\ &\leq \sum_{n=1}^{S(t)} \mathbb{1} \{ U_n \in [0, X_\Sigma^P(\tau_s^n-) \wedge N_\Sigma N_\Sigma] \} - \mathbb{1} \{ U_n \in [0, X_\Sigma^P(\tau_s^n-) \wedge N_\Sigma - C(\tau_s^n-) N_\Sigma] \}. \end{aligned} \quad (\text{T38})$$

Taking expectations we get

$$\mathbb{E} \left[ X_\Sigma(t) - X_\Sigma^P(t) \right] \leq \mathbb{E} \left[ \mu N \int_0^t C(s) / N_\Sigma ds \right] = \mathbb{E} \left[ \mu \int_0^t C(s) ds \right],$$

which concludes the proof of the proposition. ■

**Proof of Proposition A.2:** By the stability of the CTMC  $X(\cdot)$ , we have that for any deterministic initial condition  $X(0) = x \in \mathbb{Z}_+^2$ ,  $C(t) \Rightarrow C$  where  $C(t)$  is as defined in (A7) and  $C$  has the corresponding steady-state expression, i.e.,

$$C = (Q_1 \wedge I_2) \mathbb{1}\{X \in \mathcal{K}_1\} + (Q_2 \wedge I_1) \mathbb{1}\{X \in \mathcal{K}_2\}.$$

To show that  $\mathbb{E}[C] \leq C(\tilde{\rho}_1, \tilde{\rho}_2)$ , it then suffices to fix the initial condition to  $X(0) = (0, 0)$  and show that the weak limit  $C$  of  $C(t)$  then satisfies the desired bound. Initializing  $X(\cdot)$  with its steady-state distribution as in the statement of the proposition we then have that  $C(t) \stackrel{d}{=} C$  for all  $t \geq 0$ , and, in turn, that  $\mathbb{E}[C(t)] \leq C(\tilde{\rho}_1, \tilde{\rho}_2)$  for all  $t \geq 0$ .

To complete the details in the above outline, let

$$C_1(t) := (Q_1(t) \wedge I_2(t)) \mathbb{1}\{X(t) \in \mathcal{K}_1\} \text{ and } C_2(t) := (Q_2(t) \wedge I_1(t)) \mathbb{1}\{X(t) \in \mathcal{K}_2\},$$

so that  $C(t) = C_1(t) + C_2(t)$ . We define two processes  $M_1(t)$  and  $M_2(t)$  such that almost surely  $C_1(t) \leq M_1(t)$  and  $C_2(t) \leq M_2(t)$  for all  $t \geq 0$ , and such that  $M_i(t) \Rightarrow M_i$  as  $t \rightarrow \infty$  for  $i = 1, 2$  where

$$\mathbb{E}[M_1] + \mathbb{E}[M_2] \leq \frac{1}{1 - \tilde{\rho}_1} + \frac{1}{1 - \tilde{\rho}_2}.$$

To that end, let  $M_1(\cdot)$  be the process that evolves as follows:

1.  $M_1(t) = M_1(t-) + 1$  if

- (a) there is a jump of  $A_1^w(\cdot)$ , i.e.  $A_1^w(t) - A_1^w(t-) = 1$ , or if
- (b) there is a jump of  $S(\cdot)$ ,  $S(t) - S(t-) = 1$  and  $Y^n \notin [0, N_1/N_\Sigma]$  where  $n$  is such that  $t = \tau_s^n$ . (recall the construction of the sample paths in §A.2.1).

2.  $M_1(t) = M_1(t-) - 1$  if  $M_1(t-) > 0$  and

- (a) there is a jump of either  $A_2^w(t)$ ,  $A_2^a(t)$  or  $A_1^a(t)$ , i.e.  $A_2^w(t) + A_2^a(t) + A_1^a(t) - A_2^w(t-) - A_2^a(t-) - A_1^a(t-) = 1$ , or
- (b) there is a jump of  $S(\cdot)$ ,  $S(t) - S(t-) = 1$  and  $Y^n \in [0, N_1/N_\Sigma]$  where  $n$  is such that  $t = \tau_s^n$ .

In terms of the probability law,  $M_1(\cdot)$  evolves as an  $M/M/1$  queue with arrival rate  $\lambda_1^w + \mu N_2$  and service rate  $\mu N_1 + \lambda_1^a + \lambda_2$ . Since by assumption  $\frac{\lambda_1^w + \mu N_2}{\mu N_1 + \lambda_2 + \lambda_1^a} =: \tilde{\rho}_1 < 1$  we have that

$$M_1(t) \Rightarrow M_1, \text{ as } t \rightarrow \infty \tag{T39}$$

where  $M_1$  has the steady-state distribution of the queue length in an  $M/M/1$  queue with the above parameters. In particular,  $\mathbb{E}[M_1] \leq 1/(1 - \tilde{\rho}_1)$ .

We let  $M_2(\cdot)$  be the process that evolves as follows:

1.  $M_2(t) - M_2(t-) = 1$  if

(a) there is a jump of  $A_2^w(\cdot)$ , i.e.,  $A_2^w(t) - A_2^w(t-) = 1$ , or if

(b) there is a jump of  $S(\cdot)$ ,  $S(t) - S(t-) = 1$  and  $Y^n \in [0, N_1/N_\Sigma]$  where  $n$  is such that  $t = \tau_s^n$ .

2.  $M_2(t) - M_2(t-) = -1$  if  $M_2(t) > 0$  and

(a) there is a jump of either  $A_1^w(t)$ ,  $A_1^a(t)$  or  $A_2^a(t)$ , i.e.,  $A_1^w(t) + A_1^a(t) + A_2^a(t) - A_1^w(t-) - A_1^a(t-) - A_2^a(t-) = 1$ , or

(b) there is a jump of  $S(\cdot)$ ,  $S(t) - S(t-) = 1$  and  $Y^n \notin [0, N_1/N_\Sigma]$  where  $n$  is such that  $t = \tau_s^n$ .

Note that  $M_2(t)$  has the probability law of an  $M/M/1$  queue with arrival rate  $\lambda_2^w + \mu N_1$  and service rate  $\lambda_2^a + \lambda_1 + \mu N_2$ . Since, by assumption,  $\frac{\lambda_2^w + \mu N_1}{\mu N_2 + \lambda_1 + \lambda_2^a} =: \tilde{\rho}_2 < 1$ ,

$$M_2(t) \Rightarrow M_2, \text{ as } t \rightarrow \infty, \quad (\text{T40})$$

where  $M_2$  has the steady-state distribution of the queue length in an  $M/M/1$  queue with the above parameters. In particular,  $\mathbb{E}[M_2] \leq 1/(1 - \tilde{\rho}_2)$ .

Initializing at  $X(0) = (0, 0)$  and  $M_1(0) = M_2(0) = 0$  we claim that, almost surely,

$$C_1(t) \leq M_1(t), \text{ and } C_2(t) \leq M_2(t) \text{ for all } t \geq 0. \quad (\text{T41})$$

This is proved by a simple coupling argument whose proof appears, for completeness, at the end of this section. Combining (T41), (T39) and (T40) we have that  $C(t) \Rightarrow C$  where

$$\mathbb{E}[C] \leq \mathbb{E}[M_1 + M_2] \leq \frac{1}{1 - \tilde{\rho}_1} + \frac{1}{1 - \tilde{\rho}_2}.$$

Equation (A13) then follows from (A6) by which  $Q_\Sigma(t) = (X_\Sigma(t) - N_\Sigma)^+ + C(t)$ . ■

**Proof of Proposition A.3:** The proof is based on a comparison result between the  $(N_1, N_2)$ -threshold system and a network with randomized routing.

We start by introducing a network with randomized routing. As before, walk-ins are routed to their own ED. The ambulances are routed randomly as follows: fixing non-negative numbers  $p_{ij}$ ,  $i, j \in \{1, 2\}$  such  $p_{i1} + p_{i2} = 1$ ,  $i = 1, 2$  we thin the Poisson process  $A_1^a(\cdot)$  and  $A_2^a(\cdot)$  according to these probabilities so that  $A_i^a(\cdot)$  is split into two independent Poisson process with rate  $\lambda_1^a p_{i1}$  and  $\lambda_1^a p_{i2}$ ; see e.g. Chapter 5 of Kulkarni [1995]. In other words, an arrival of an ambulance from catchment area  $i$  is routed to ED  $j$  with probability  $p_{i,j}$  independently of all previous arrivals.

Let  $X_i^R(t)$ ,  $i = 1, 2$  be the total number of patients in ED  $i$  under this randomization scheme. Note that, starting empty,  $X_i^R(t)$  has the law of an  $M/M/N$  queue with service rate  $\mu$ ,  $N_i$  servers and arrival rate  $\lambda_i^w + \sum_{j=1}^2 \lambda_j^a p_{ji}$ . We claim that with *any* fixed thinning probabilities  $p_{ij}$ , initializing  $X^R(0) = X(0) = (0, 0)$  we have that

$$X_\Sigma(t) \leq_{st} X_1^R(t) + X_2^R(t), \quad t \geq 0, \quad (\text{T42})$$

where,  $X_\Sigma(t)$  is the total number of patients in the  $(N_1, N_2)$ -threshold system at time  $t$ . The stochastic dominance in equation (T42) is established via a coupling argument that is omitted as it is similar to the argument used in the proof of Proposition A.1. The essence of the coupling is that the threshold system routes arriving ambulance to an empty bed whenever there is such in the system. The randomized system, in contrast, can route an ambulance to an occupied ED even if there are available beds in the other one.

The required bounds are then obtained by choosing the probabilities  $p_{ij}$  so that

$$\frac{\lambda_1^w + \sum_{j=1}^2 \lambda_j^a p_{j1}}{\mu N_1} = \frac{\lambda_2^w + \sum_{j=1}^2 \lambda_j^a p_{j2}}{\mu N_2} = \frac{\lambda_1 + \lambda_2}{\mu(N_1 + N_2)} = \frac{\lambda}{\mu N_\Sigma} = \rho. \quad (\text{T43})$$

Such choice of the parameters is possible because of (7). With this choice, the steady-state distribution of  $X_i^R(t)$  is that of an  $M/M/N$  queue with service rate  $\mu$ ,  $N_i$  servers and arrival rate  $\mu\rho N_i$  so that the result follows by taking  $t \rightarrow \infty$  in (T42). ■

**Proof of Corollary A.2:** By Proposition A.3  $X_{\Sigma,\pi}^P(0) \leq_{st} \bar{\mathcal{M}}_\Sigma$  where  $\bar{\mathcal{M}}_\Sigma$  has the distribution  $\bar{\nu}_\Sigma$ . By the stochastic monotonicity of the  $M/M/N$  queue it is initial condition (see e.g. Chapter 9 of Ross [1996]), we then have that  $X_{\Sigma,\pi}^P(t) \leq_{st} X_{\Sigma,\bar{\nu}_\Sigma}^P(t)$  for all  $t \geq 0$ . Since  $(x - N_\Sigma)^+$  is a non-decreasing function of  $x$ , we also have that  $Q_{\Sigma,\pi}^P(t) = (X_{\Sigma,\pi}^P(t) - N_\Sigma)^+ \leq_{st} (X_{\Sigma,\bar{\nu}_\Sigma}^P(t) - N_\Sigma)^+ = Q_{\Sigma,\bar{\nu}_\Sigma}^P(t)$  for all  $t \geq 0$  so that the corollary now follows from the definition of  $T(\epsilon)$ . ■

**Proof of equation (T41):** Here we prove equation (T41) that was used in the proof of Proposition A.2. We perform a coupling argument to show that  $C_1(t) := Q_1(t) \wedge I_2(t) \mathbb{1}\{X(t) \in \mathcal{K}_1\} \leq M_1(t)$  for all  $t \geq 0$ . To that end, note that: (i)  $Q_1(t) > 0$  implies that  $Z_1(t) = N_1$  and, since  $Z_\Sigma(t) \leq N$ , also that  $Z_1(t)/Z_\Sigma(t) \geq N_1/N_\Sigma$ , and (ii) when  $I_2(t) > 0$  and  $X(t) \in \mathcal{K}_1$ , then  $Z_1(t) = N_1$  so that  $Z_1(t)/Z_\Sigma(t) = N_1/(N_1 + Z_2(t)) \geq N_1/N_\Sigma$ .

We use induction on the time of the event, whether it is a jump of  $A_i^w(\cdot)$ ,  $A_i^a(\cdot)$  or  $S(\cdot)$ . Note that at time 0,  $C_1(0) = 0$  since we initialized the system empty so that (T41) holds at time 0. To establish that (T41) is preserved for all  $t$  we now show that  $M_1(t)$  jumps up whenever  $C_1(t)$  does and that whenever  $M_1(t)$  decreases so does  $C_1(t)$  if  $C_1(t) > 0$ . Assume that the induction assumption holds up to the  $(n-1)^{th}$  jump and consider the  $n^{th}$  jump. Then, assume that  $X(t) \in \mathcal{K}_1$  (otherwise  $C_1(t) = 0$ ).

1. If the event is a jump of  $A_1^w(\cdot)$ , i.e.  $A_1^w(t) - A_1^w(t-) = 1$ , then  $C_1(t) - C_1(t-) = 1$  and also, by definition,  $M_1(t) - M_1(t-) = 1$  so that the ordering is preserved.
2. If the event is a jump of  $S(\cdot)$ , then  $C_1(t) - C_1(t-) = 1$  only if  $Y^n \notin [0, Z_1(t)/Z_\Sigma(t)]$  (with  $n$  such that  $\tau_s^n = t$ ). On the other hand,  $M_1(t) - M_1(t-) = 1$  if  $Y^n \notin [0, N_1/N_\Sigma]$ . As we noted  $Z_1(t)/Z_\Sigma(t) \geq N_1/N_\Sigma$  so  $M_1(t-) - M_1(t) = 1$  whenever  $C_1(t) - C_1(t-) = 1$  and the ordering is preserved.

On the other hand, with a jump of  $S(\cdot)$ ,  $M_1(t) - M_1(t-) = -1$  if  $Y^n \in [0, N_1/N_\Sigma]$  whereas  $C_1(t) = C_1(t-) - 1$  if  $Y^n \in [0, Z_1(t)/Z_\Sigma(t)]$ . By our previous observations  $Z_1(t)/Z_\Sigma(t) \geq N_1/N_\Sigma$  so that  $C_1(t) - C_1(t-) = -1$  if  $M_1(t) - M_1(t-) = 1$  and the ordering is preserved.

3. Assume that the jump is of  $A_2^w(\cdot)$ ,  $A_2^a(\cdot)$  or  $A_1^a(\cdot)$ . Then we have both  $C_1(t) - C_1(t-) = -1$  and  $M_1(t) - M_1(t-) = -1$  (provided they are both positive) and the ordering is preserved.

Note that it suffices to consider individual jumps because the probability of simultaneous jumps is 0. The argument to show that  $M_2(t) \leq C_2(t)$  for all  $t \geq 0$  is very similar and is omitted. ■

## Appendix references

- S. Asmussen. *Applied probability and queues*. Springer Verlag, 2003.
- P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 1999.
- V. Kulkarni. *Modeling and analysis of stochastic systems*. Chapman & Hall/CRC, 1995.
- A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. John Wiley & Sons Inc, 2002.
- S. Ross. *Stochastic processes*. Wiley New York, 1996.
- W. Whitt. Continuity of generalized semi-markov processes. *Mathematics of Operations Research*, 5(4):494–501, 1980.
- D. Wolf. Approximation of the invariant probability measure of an infinite stochastic matrix. *Advances in Applied Probability*, 12(3):710–726, 1980.