Stochastic Systems

SCHEDULING RESIDENTS IN HOSPITALS: QUEUEING MODELS AND FLUID APPROXIMATIONS

By Ross Anderson and David Gamarnik

Massachusetts Institute of Technology

Major hospitals face a difficult challenge of designing shift schedules for their residents that satisfy demand, provide quality care, and are compliant with regulations restricting shift lengths. Motivated by empirical work conducted by the authors at the Brigham and Women's (B&W) Hospital in Boston, we analyze the impact of shift lengths on two key performance metrics. The first metric is admitting capacity—the largest patient arrival rate sustainable by a given shift schedule. The second metric is the number of reassigned patients—the number of patients admitted temporarily by one doctor and then permanently transferred to a resident.

We build a queueing model to compare two shift scheduling policies that are representative of the alternatives encountered in hospitals: one where residents work long shifts on alternating days, called Long Shifts (LS), and another where residents admit patients daily in short shifts, called Daily Admitting (DA). We determine the admitting capacity for our queueing model under each policy. Then we construct a fluid model—a large scale approximation of the underlying queueing model. We show that for each policy, the fluid model has a unique steady state solution. Finally, we establish an interchange of limits between the stochastic and fluid models in steady state. We use these results to compare the key performance metrics under the two policies.

Our analysis shows that the DA policy has a greater capacity to admit patients than the LS policy for all parameter choices. Furthermore, we numerically establish the existence of a threshold value, such that the number of reassigned patients is smaller for the DA policy than for the LS policy if and only if the arrival rate of patients is greater than the threshold value. Since most hospitals operate at near critical loads, our two findings lead to the conclusion that sched-

ules which rely on shorter more frequent shifts than those found in practice would increase admitting capacity and reduce the number of reassigned patients.

1. Introduction.

1.1. *Background and empirical results.* In this paper, we study the effect of medical resident schedules on hospitals' capacity to admit patients and the quality of care delivered. This work is motivated by an empirical study [1] conducted by the authors with Brigham and Women's hospital (B&W) in Boston. As a teaching hospital, the majority of patients at B&W have medical residents as their caring doctor. When our study began, it was a common practice for residents to have long work hours, with 80 hour work weeks and very long shifts, some extending to 36 continual hours spent in the hospital [35]. However, over the last twenty years, concerns over the dangers of resident fatigue, both for patients [15, 27, 33, 35], and residents [3, 4, 21, 35] have led to increasingly stringent regulations on resident shifts, especially for those who are in their first year of residency. Most recently, a new regulation that entered effect in August 2011 imposes a maximum shift length of 16 hours [23]. Proponents of the long shifts are concerned primarily with patient *continuity of care*. They argue that reducing shift lengths will result in more patient handoffs between caring practitioners, increasing the chance of miscommunication and accidents [8, 31]. A particularly undesirable type of handoff is a *reassignment*, when a patient is admitted temporarily by one doctor and then is transferred to a resident for a permanent care. Reassignments are dangerous as they greatly increase the risk of losing information that should be used in determining a course of treatment. Additionally, long shift advocates additionally argue that reducing residents' hours will force hospitals to increase staffing levels to compensate for lost capacity to admit patients [37, 38]. The impact of shift schedules on the number of reassignments and admitting capacity are two main questions we address in this paper.

We build a queueing model that compares the capacity and performance of different resident schedules to determine the quality of long shift alternatives. Using this model, we show that policies using shorter, more frequent shifts have a greater admitting capacity than policies using long less frequent shifts, when the total number of hours spent admitting patients is the same. Furthermore,

we show that scheduling policies based on frequent short shifts cause fewer reassignments when the patient load is high. Since most of the hospitals operate at high load, this observation is relevant to the majority of hospitals. Our theoretical results are substantiated with extensive empirical estimations which are reported in [1], where these two main findings are verified in a number of experiments based on real life data. In particular, we simulated the flow of patients in B&W hospital under a four day rotating cycle Long shift–Off–Short shift–Off (LOSO) resident schedule, and a similar six day variant. These were the schedules used in the hospital prior to the regulation of 2011. According to this schedule, at the beginning of the cycle a resident stays on a long (22-24 hour) shift, admits no patients the following day, then is on shift again for the morning of the third day (5-7 hours), and finally admits no patients on the fourth day. Residents were organized into a number of teams that was divisible by four with each team offset by a day so that the teams would provide uniform coverage. We have compared the performance of the LOSO schedule with an alternative schedule, called MMMO (Medium-Medium-Medium-Off). The schedule is also based on a four day rotating cycle where for the first three days residents are on shift for 8 to 10 hours and then have one day off. The shifts are set to provide more coverage during the peak of patient arrivals (see Figure 1 for the arrival rates throughout the day) and thus the schedule is expected to reduce the number of reassignments.

The results of our empirical studies are summarized on figure Figure 2, where we report the total number of reassignments at the current patient load of about 3250 patients a year, as well as for increased and decreased patient loads obtained by artificially changing the total number of patients while maintaining the hourly, daily, and monthly arrival rate patterns. The results are consistent with intuition one derives from queueing theory—we see a nonlinear degradation of the performance as the utilization of the system increases. However, the results indicate that the performance of LOSO degrades *faster* than the performance of MMMO, suggesting that the two policies have fundamentally different capacities despite using the same total number of hours. Also, we see that the MMMO schedule performs better at the current load of 3250 patients per year.

1.2. *Our results and discussion.* Motivated by these empirical findings, we build a stylized queueing model of the patient flow in a hospital. Patients arrive according to a non-homogenous Poisson process with rate λ_1 in one half of each day (say 10am till 10pm) and rate $\lambda_2 \leq \lambda_1$ in the

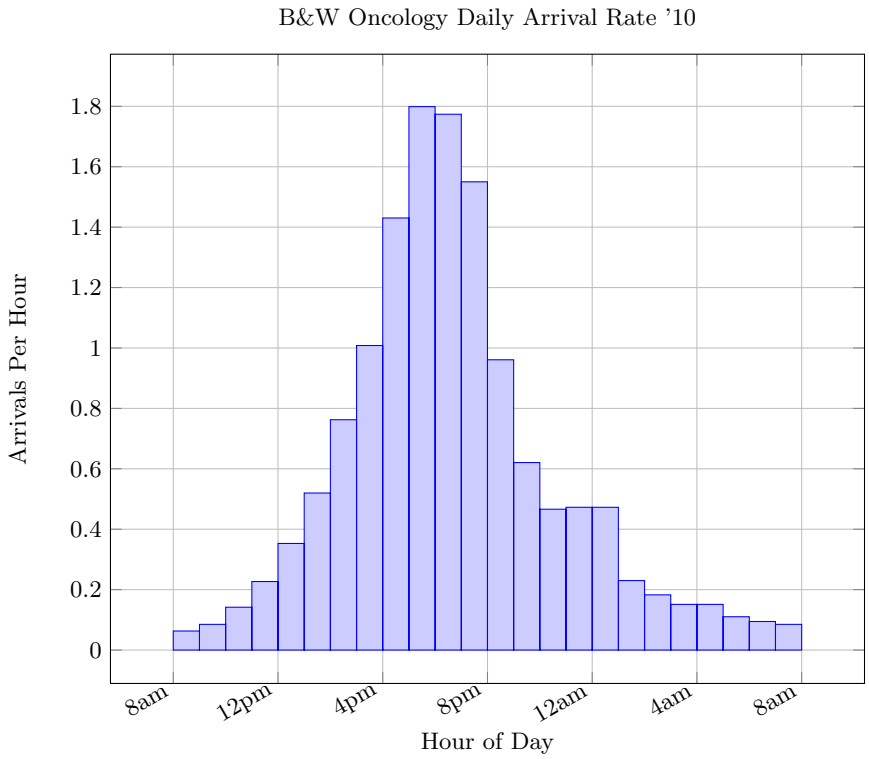

FIG 1. *Average patient arrival rate by hour of day, B&W Oncology dept., 2009-10.*

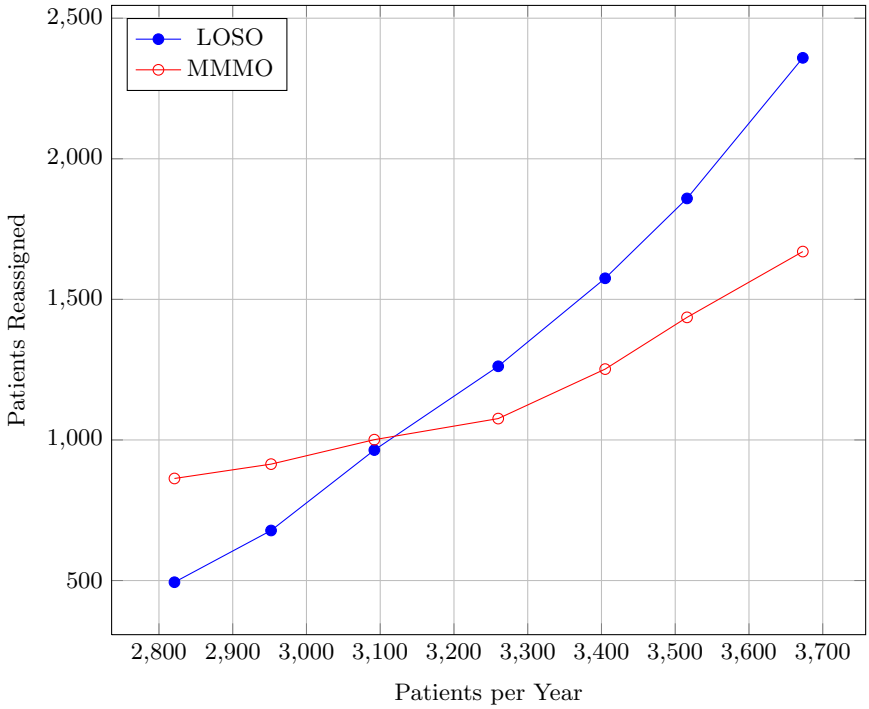

FIG 2. *Patient reassignments under increasing patient load, B&W Oncology dept., 2009-10. Actual patient load for 2009-10 was 3242.*

remaining part of the day. We consider two stylized policies to schedule the residents, which while significantly simplifying the actual LOSO and MMMO policies discussed above, capture the salient features of these policies. The formal description of the two policies is given in the next section. Now we just provide a high level description and discuss their main features.

The first policy we analyze in this paper is the *Long Shifts* (LS) policy, motivated by the LOSO policy above. According to this schedule two teams of residents with the same number of residents in each team work for the duration of a day every other day, taking a day off after each day on shift. Namely the first team works on days $2n+1, n \geq 0$ and the second team works on days $2n, n \geq 1$.

The second policy we analyze is called *Daily Admitting* (DA), and is motivated by the MMMO policy. According to the DA policy, two teams of residents each with the same number of residents as for the LS policy work every day during the high load (10am-10pm) half of the day, and a are off-shift for the other half of the day. Both policies organize the residents into two teams that are offset by a day in the rotating schedule, thus providing uniform coverage. The main distinction between the LS and DA policies is that DA is based on adopting *shorter more frequent* shifts.

The arriving patients are assigned to residents according to the following mechanism used both for the LS and DA schedules. Each resident has an upper bound (cap) on the number of patients he is allowed to have in care. Each arriving patient is assigned to a resident chosen uniformly at random from all residents on shift who have not reached their cap (in fact the analysis does not depend on how patients are assigned to available residents for these policies). If all the residents are capped at the arrival epoch, the patient joins the queue of unassigned patients and is cared for temporarily by a doctor from a back-up supply of care providers. We assume that we have an infinite supply of these providers, although using them is expensive in a sense to be described below. These patients are subsequently assigned to a caring resident on the first come first serve basis, as soon as one of them is available. Patients remain in the hospital for a random exponentially distributed amount of time, beginning from when they are assigned to a resident. We make this assumption because in practice, the newly arriving patients without an assigned resident are stabilized but the treatment plan is not determined until they are assigned to a resident.

We now summarize our results. First, we determine the throughput capacity of each policy. Specifically, for a given number of residents, we compute the maximum arrival rate at which patients

can arrive before the queueing system becomes unstable, i.e., the number of unassigned patients grows without bound. We show that DA has a greater throughput capacity than LS, independent of the parameters of our model, supporting our hypothesis from Figure 2 that MMMO has a greater capacity than LOSO.

Next, we compare the number of reassignments under LS and DA. In comparing policies, we are interested in the expected number of reassignments per day in steady state. As direct steady state analysis appears intractable, we instead resort to the method of fluid approximation of the underlying queueing model. We analyze the long term behavior of the fluid model and show that it converges to the unique steady state solution. The steady state fluid solution carries important information about the long-term performance of the underlying stochastic system. In particular, we prove an interchange of limits result, that the steady state number of patients being treated and the number of patients waiting to be reassigned converges to the steady state fluid solution under the appropriate rescaling. We obtain an implicit formula for the number of each type of reassignment per day in the fluid limit that can be solved numerically. Under minor technical assumptions, we also prove that the number of reassignments in the underlying stochastic model converges to this value in the fluid rescaling, thus justifying fluid approximation. These results provide important qualitative insights.

In particular, computing the number of reassignments under each policy from the fluid steady state solution, we find that the DA policy leads to *fewer* reassignments than the LS policy when the load is high, and leads to *more* reassignments than the LS policy when the load is low. These findings are consistent with the observed behavior of LOSO and MMMO shown in Figure 2. Since most hospitals tend to operate at high load, our results lead to the conclusion that the hospitals should consider implementing schedules with shorter more frequent shifts, as it will increase the capacity to admit patients and reduce the number of reassignments. In this sense the new regulation restricting further the length of shifts should not be perceived as an impediment to efficient handling of patients at hospitals.

1.3. *Related literature.* We now mention some other approaches considered in the field of operations research for capacity management in hospitals. A very general survey of capacity management in healthcare is given in [18]. Simulation studies of capacity have been done for medical resident

schedules [10, 25] and various other hospital resources [12, 24, 28, 32, 39]. However, as these simulations are incredibly sensitive to the details of each hospital's operations, the results do not generalize well to other hospitals. There are some recent papers which propose a queueing model of patient flow in a hospital, and then solve for important performance metrics, either analytically [40], asymptotically [9, 40, 41], numerically [34], or with heuristic methods [20]. Although not explicitly about healthcare, in [22, 29], fluid models for queues with time varying arrival rates alternating between overloaded and underloaded periods are considered. These models are more in the spirit of our work than the previously cited models from a technical perspective, as transient behavior and "end of day effects" (see [19]) play a prominent role.

1.4. *Organization and Notational conventions.* The remainder of the paper is organized as follows. The queueing model and its fluid limit are described Section 2 and the main results are stated there. In Section 3 we numerically solve for the steady state behavior of the fluid model and discuss the performance implications for our queuing model. Then we give some concluding remarks in Section 4. The proofs of the main results are then in the appendix. In Appendix A.1, we exactly characterize the stability of our queueing model under each policy using a simple linear Lyapunov function type argument. In Appendix A.2, we use quadratic Lyapunov functions to bound the expected steady state queue length. In Appendix A.3, we prove the existence of the fluid limits, applying the results in [30]. Then in Appendix A.4, we show that the fluid limit has a consistent periodic long run behavior under each policy, where the solution in each period is characterized by a simple system of differential equations. In Appendix A.5, we prove that the long run solution to the fluid model approximates the steady state queue lengths of the underlying queueing model. Justifying this requires an argument for an "interchange of limits." As in [16], we use our moment bound from Appendix A.2 to show tightness of the rescaled stationary distributions, and then we follow the technique of [13] and similarly [36] to prove the interchange of limits. In Appendix A.6, we use the result of Appendix A.5 to show that the long run number of daily reassignments converges in the fluid rescaling converges to a natural function of the fluid limit. Finally, we have two rather technical sections: Appendix A.7, where we show several elementary properties of the solution to a differential equation, and Appendix A.8, where we use another Lyapunov function argument to distinguish between the null recurrent and transient cases in our queueing model.

We conclude with a summary of the mathematical notation used in the paper. Throughout, $\mathbb{R}$ ($\mathbb{R}_+$) denotes the set of (nonnegative) reals, and likewise, $\mathbb{Z}$ ($\mathbb{Z}_+$) denotes the set of (nonnegative) integers. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ is the ℓ_p-norm. The ℓ_1 ball of radius r is denoted $B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 \mid \|\mathbf{x} - \mathbf{y}\|_1 < r\}$. For $x, y \in \mathbb{R}$, $x \wedge y = \min\{x, y\}$ and $(x)^+ = \max\{x, 0\}$. We define $f(t^-)$ as $\lim_{\tau \nearrow t} f(\tau)$ when the limit exists. We let $\text{Exp}(\mu)$, $\text{Pois}(\lambda)$, and $\text{Bin}(n, p)$, denote an exponential random variable with mean $1/\mu$, a Poisson random variable with mean λ, and a Binomial random variable with mean np and variance $np(1-p)$, respectively (these moments characterize the distributions). If the sequence of random vectors $\mathbf{X}^n$, $n = 1, 2, \ldots$, converges weakly (in distribution) to $\mathbf{X}$ as $n \to \infty$, we say $\mathbf{X}^n \Rightarrow \mathbf{X}$. For a stochastic process $X(t)$ in either discrete or continuous time, $\mathbb{E}_x[X(t)]$ denotes $\mathbb{E}[X(t) \mid X(0) = x]$. A sequence of continuous time vector valued stochastic processes $\mathbf{X}^n(t)$ on a common probability space Ω converges almost surely (a.s.) and uniformly on compact sets (u.o.c.) to a deterministic function $\mathbf{x}(t)$ if for every $t > 0$ and almost every $\omega \in \Omega$,

$$\lim_{n\to\infty} \sup_{0 \le s \le t} \{\|\mathbf{X}^n(s, \omega) - \mathbf{x}(s)\|_1\} = 0,$$

where $\|\cdot\|_1$ is the 1-norm for vectors. See [7] for more details. As in [11] sections 11.2-11.3, for functions $f \colon \mathbb{R}^n \to \mathbb{R}$, we let $\|f\|_L = \sup_{\mathbf{x},\mathbf{y} \in \mathbb{R}^n} |f(\mathbf{x}) - f(\mathbf{y})|/\|\mathbf{x} - \mathbf{y}\|_1$ denote Lipschitz semi-norm (when f is a Lipschitz function, the value of this norm is the smallest Lipschitz constant that f satisfies). We let $\|f\|_{\text{BL}} = \|f\|_L + \|f\|_\infty$. This quantity is a true norm.

2. Model, Assumptions and Main Results. We begin by introducing our model of residents admitting and treating a flow of incoming patients. The patients are assumed to arrive according to a non-homogeneous Poisson process. For each $k \in \mathbb{Z}_+$, the process has rate λ_1 over the time intervals, $[k, k+\frac{1}{2})$ and rate $\lambda_2 < \lambda_1$ over the time intervals $[k+\frac{1}{2}, k+1)$. Let let $\lambda = (\lambda_1+\lambda_2)/2$ denote the average arrival rate and

$$\lambda(t) = \begin{cases} \lambda_1 & t \in [k, k+\frac{1}{2}), \\ \lambda_2 & t \in [k+\frac{1}{2}, k+1). \end{cases}$$

The intervals $[k, k+1)$ represent, for example, 24 hour cycles, where $[k, k+\frac{1}{2})$ is the portion of the day, say from 10am to 10pm, in which the vast majority of patients arrive (see Figure 1).

The residents are combined into two teams, A and B, identical in size, which are eligible to admit patients (are on shift) according a schedule to be described below. Each team has capacity $c > 0$ bounding the maximum number of patients the team can have in care. Each arriving patient is assigned to one of the residents on a team, chosen uniformly at random, provided that at least one of the teams on shift has not reached its capacity c. Note, that this is equivalent to saying that each resident has in care capacity c/N, where N is the number of residents on each the team. If each on shift teams has reached its capacity, the patient joins a single queue and is cared for by one of the back-up doctors until one of the residents is available, at which point the waiting terminates, using the First-In-First-Out assignment policy. The availability occurs either when one of the assigned patients leaves the hospital freeing the capacity of one of the teams, or when one of the teams with load less than c begins a shift.

At any time, each team is in one of two states, *on shift* or *off shift*, as specified by a policy. Patients remain assigned to a team until they leave the hospital. The durations of hospital stays are assumed to be i.i.d. and exponentially distributed with rate μ. That the random length of treatment time each patient requires begins accumulating at the moment of assignment to a team, and continues to accumulate when team is off shift. How this corresponds to actual practices is explained in the introduction.

We consider two scheduling policies controlling when each team is on shift, *Long Shifts* (LS) motivated by LOSO, and *Daily Admitting* (DA), motivated by MMMO (see the introduction for descriptions of LOSO and MMMO). LS is a two day rotating schedule. Team A is on shift on odd days, i.e. $[2k+1, 2k+2)$ for all $k \in \mathbb{Z}_+$, and off shift otherwise. Similarly, team B is on shift for even days, i.e. $[2k, 2k+1)$ for all $k \in \mathbb{Z}_+$, and off otherwise. In DA, both teams A and B are on shift every day for the first half of each day, i.e. $[k, k+\frac{1}{2})$ for all $k \in \mathbb{Z}_+$, and off otherwise, effectively creating a single team with double the capacity.

To state our results, it will be convenient to introduce the following quantities describing the dynamics of our model. For $t \geq s \geq 0$, let $A(s,t)$ denote the number of patients that arrive in the time interval $[s,t]$ according to our non-homogeneous Poisson process. For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, let $Q_\theta(t)$ denote the number of patients in the queue not yet assigned to a team plus the number of patients assigned to the teams that are on shift at time t. For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, let $R_\theta(t)$ denote

the total number of patients currently assigned to teams which are off shift at time t. Further, for each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$ we introduce the random vector $\mathbf{S}_\theta(t) = (Q_\theta(t), R_\theta(t))$, which we take to be the state of our system. The processes $Q_\theta(t)$, $R_\theta(t)$, and $\mathbf{S}_\theta(t)$ are assumed to be right-continuous with left limits. For every $s < t$, we let $D_\theta^{\mathrm{on}}(s,t)$ denote the total number of patients which departed in the time interval $[s,t]$ from the teams which were on shift in this period. We define $D_\theta^{\mathrm{off}}(s,t)$ analogously.

Under the policy LS, for each $k \in \mathbb{Z}_+$ and each $t \in [0,1)$, Q_{LS} and R_{LS} satisfy

$$Q_{\mathrm{LS}}(k+t) = Q_{\mathrm{LS}}(k) + A(k,k+t) - D_{\mathrm{LS}}^{\mathrm{on}}(k,k+t), \tag{1}$$

$$R_{\mathrm{LS}}(k+t) = R_{\mathrm{LS}}(k) - D_{\mathrm{LS}}^{\mathrm{off}}(k,k+t). \tag{2}$$

On $[k,k+1)$, we see that in distribution $Q_{\mathrm{LS}}(t)$ behaves exactly as the total number of customers in system for an $M(t)/M/c$ queue with arrival rate $\lambda(t)$, service rate μ, and initial value $Q_{\mathrm{LS}}(k)$. Similarly, on $[k,k+1)$, $R_{\mathrm{LS}}(t)$ behaves as an $M/M/c$ system with no arrivals, service rate μ, and initial value $R_{\mathrm{LS}}(k)$. At each time $k \in \mathbb{Z}_+$ a transition occurs: the off shift team switches to on shift and vice versa, and patients that are waiting can get assigned to the team rotating on shift. In terms of Q_{LS} and R_{LS}, this can be described as follows:

$$\begin{aligned} Q_{\mathrm{LS}}(k) &= (Q_{\mathrm{LS}}(k^-) - c)^+ + R_{\mathrm{LS}}(k^-), \\ R_{\mathrm{LS}}(k) &= Q_{\mathrm{LS}}(k^-) \wedge c. \end{aligned}$$

We define operator $\Gamma : \mathbb{R}_+^2 \times \mathbb{R}_+ \to \mathbb{R}_+^2$ by

$$\Gamma(q,r;\kappa) \triangleq (q-\kappa)^+ + r, q \wedge \kappa), \tag{3}$$

and note that we can equivalently write

$$\mathbf{S}_{\mathrm{LS}}(k) = \Gamma(\mathbf{S}_{\mathrm{LS}}(k^-); c). \tag{4}$$

The equations (1), (2) and (4) along with a distribution of the initial states $\mathbf{S}_{\mathrm{LS}}(0)$ completely determine the distribution of $\mathbf{S}_{\mathrm{LS}}(t)$ for all $t \in \mathbb{R}_+$. It is immediate that on integer times, the process $\mathbf{S}_{\mathrm{LS}}(k)$ is a two dimensional Markov chain on the countable state space $\mathbb{Z}_+ \times \{0,1,\ldots,c\}$. Moreover, without loss of generality we can restrict the state space to the set,

$$\mathcal{S}_{\mathrm{LS}} \triangleq \{0,\ldots,c\}^2 \cup \{(q,c) \mid q \in \mathbb{Z}_+\}, \tag{5}$$

as this set is the image of $\Gamma(\cdot; c)$ when we take the domain to be $\mathbb{Z}_+ \times \{0, 1, \ldots, c\}$. Thus $\mathbf{S}_{\mathrm{LS}}(k) \in \mathcal{S}_{\mathrm{LS}}$ for all $k \geq 1$ with probability one.

We now describe similar relations for the policy DA. For every $k \in \mathbb{Z}_+$ and every $t \in [0, \frac{1}{2})$,

$$\begin{aligned} Q_{\mathrm{DA}}(k+t) &= Q_{\mathrm{DA}}(k) + A(k, k+t) - D^{\mathrm{on}}_{\mathrm{DA}}(k, k+t), \\ R_{\mathrm{DA}}(k+t) &= 0, \end{aligned} \tag{6}$$

where on $[k, k+\frac{1}{2})$, the process $Q_{\mathrm{DA}}(t)$ has the same distribution as an $M/M/2c$ queue with an arrival rate of λ_1, a service rate of μ and an initial value of $Q_{\mathrm{DA}}(k)$. At time $k+\frac{1}{2}$, both teams move off shift, resulting in the transition

$$\begin{aligned} Q_{\mathrm{DA}}(k+\tfrac{1}{2}) &= (Q_{\mathrm{DA}}((k+\tfrac{1}{2})^-) - 2c)^+, \\ R_{\mathrm{DA}}(k+\tfrac{1}{2}) &= Q_{\mathrm{DA}}((k+\tfrac{1}{2})^-) \wedge 2c. \end{aligned}$$

Equivalently, in terms of Γ,

$$\mathbf{S}_{\mathrm{DA}}(k+\tfrac{1}{2}) = \Gamma(\mathbf{S}_{\mathrm{DA}}((k+\tfrac{1}{2})^-); 2c). \tag{7}$$

On $[k+\frac{1}{2}, k+1)$,

$$\begin{aligned} Q_{\mathrm{DA}}(k+\tfrac{1}{2}+t) &= Q_{\mathrm{DA}}(k+\tfrac{1}{2}) + A(k+\tfrac{1}{2}, k+\tfrac{1}{2}+t), \\ R_{\mathrm{DA}}(k+\tfrac{1}{2}+t) &= R_{\mathrm{DA}}(k+\tfrac{1}{2}) - D^{\mathrm{off}}_{\mathrm{DA}}(k+\tfrac{1}{2}, k+\tfrac{1}{2}+t). \end{aligned} \tag{8}$$

Now on $[k+\frac{1}{2}, k+1)$, $Q_{\mathrm{DA}}(k+\frac{1}{2}+t)$ changes according to a Poisson process with arrival rate λ_2, and $R(t)$ is an $M/M/2c$ queue with no arrivals and service rate μ. At integer times, we have a second shift change, this time leading to

$$\begin{aligned} Q_{\mathrm{DA}}(k+1) &= Q_{\mathrm{DA}}((k+1)^-) + R_{\mathrm{DA}}((k+1)^-), \\ R_{\mathrm{DA}}(k+1) &= 0. \end{aligned}$$

Equivalently, in terms of Γ,

$$\mathbf{S}_{\mathrm{DA}}(k+1) = \Gamma(\mathbf{S}_{\mathrm{DA}}((k+1)^-); 0). \tag{9}$$

Equations (6), (7), (8) and (9) along with the distribution over the initial states $\mathbf{S}_{\mathrm{DA}}(0)$ determine the distribution of $\mathbf{S}_{\mathrm{DA}}(t)$ for all $t \in \mathbb{R}_+$. Again on integer times, the process $\mathbf{S}_{\mathrm{DA}}(k)$ is a Markov

chain on the countable state space. However, now the state space is the one-dimensional set

$$\mathcal{S}_{\mathrm{DA}} \triangleq \mathbb{Z}_+ \times \{0\}, \tag{10}$$

as $R_{\mathrm{DA}}(k) = 0$ for all $k \in \mathbb{Z}_+$.

We are now ready to discuss our main results. We define the stochastic process $\mathbf{S}_\theta(t)$ to be *stable* if the embedded discrete time process $\mathbf{S}_\theta(k)$ for $k \in \mathbb{Z}_+$ is positive recurrent, and *unstable* otherwise. Normally we define stability for this type of problem as positive Harris recurrence of the process $\mathbf{S}_\theta(t)$, $t \in \mathbb{R}_+$. However, it is easy to see that in our case these definitions are equivalent, and further that the former definition is much easier to work with.

Observe that under both policies, $R_\theta(k)$ is bounded hence $Q_\theta(k)$ is the only potential source of instability. Thus when $\mathbf{S}_\theta(t)$ is unstable, with probability one, the number of patients waiting to be assigned to a resident team will grow without bound. Note that in reality, when a hospital has a large number of patients waiting, it reroutes incoming patients to other hospitals to reduce congestion, so instability would actually correspond to the hospital frequently being forced to turn patients away—clearly a very undesirable situation.

We now discuss conditions under which $\mathbf{S}_\theta(t)$ is stable. Before formally stating our results, we provide some intuition. Let $L_c(t)$ and $L_{2c}(t)$ be the number of patients in system for an $M(t)/M/c$ queue and $M(t)/M/2c$ queue both driven by the arrival process $A(0,t)$, respectively. For LS, it is not difficult to see that we can couple $\mathbf{S}_{\mathrm{LS}}(t)$ with $L_c(t)$ and $L_{2c}(t)$ such that surely, for every t,

$$L_{2c}(t) \leq Q_{\mathrm{LS}}(t) + R_{\mathrm{LS}}(t) \leq L_c(t).$$

The inequalities hold as the process $\mathbf{S}_{\mathrm{LS}}(t)$ has capacity between c and $2c$ at all times t. Similarly, we can couple $\mathbf{S}_{\mathrm{DA}}(t)$ and $L_{2c}(t)$ such that

$$L_{2c}(t) \leq Q_{\mathrm{DA}}(t) + R_{\mathrm{DA}}(t).$$

Recall from basic queueing theory that the process $L_c(t)$ is positive recurrent iff $\lambda < c\mu$ and $L_{2c}(t)$ is positive recurrent iff $\lambda < 2c\mu$. In light of our coupling, we thus expect the maximum throughput (the largest $\lambda = (\lambda_1 + \lambda_2)/2$ such that $\mathbf{S}_\theta(t)$ is stable) of $\mathbf{S}_{\mathrm{LS}}(t)$ to lie between $c\mu$ and $2c\mu$, and likewise we expect the maximum throughput of $\mathbf{S}_{\mathrm{DA}}(t)$ to be at most $2c\mu$. This suggests that we need to determine to what extent each policy can utilize the $2c$ total capacity available to treat

patients, or conversely how much forced idling is caused under each policy by a team's inability to admit new patients when off shift. To this end, we let

$$\rho_{\mathrm{LS}} \triangleq \frac{\lambda}{c(1-e^{-\mu})+c\mu}, \tag{11}$$

$$\rho_{\mathrm{DA}} \triangleq \frac{\lambda}{2c(1-e^{-\mu/2})+c\mu}. \tag{12}$$

We intend to show that $\mathbf{S}_\theta(t)$ is positive recurrent iff $\rho_\theta < 1$ for each θ. These values of ρ_θ imply that

$$\lambda^*_{\mathrm{LS}} \triangleq \frac{\lambda}{\rho_{\mathrm{LS}}} = c(1-e^{-\mu})+c\mu, \qquad \lambda^*_{\mathrm{DA}} \triangleq \frac{\lambda}{\rho_{\mathrm{DA}}} = 2c(1-e^{-\mu/2})+c\mu,$$

give the maximum throughput of $\mathbf{S}_{\mathrm{LS}}(t)$ and $\mathbf{S}_{\mathrm{DA}}(t)$, respectively. Thus our main stability result is as follows.

THEOREM 1. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the process $\mathbf{S}_\theta(t)$ is positive recurrent when $\rho_\theta < 1$, null recurrent when $\rho_\theta = 1$, and transient when $\rho_\theta > 1$. Namely, the process $\mathbf{S}_\theta(t)$ is stable iff $\rho_\theta < 1$. Furthermore, $\rho_{\mathrm{DA}} < \rho_{\mathrm{LS}}$. In particular, the Daily Admitting policy has a greater maximum throughput.*

The intuition behind the result is that the queue will be stable as long as conditional on the queue being large, the expected number of arrivals per day is less than the expected number of departures per day. Independent of the initial queue length, we expect λ arrivals per day. For the sake of argument, assume the initial queue length were infinite, so the resident teams are only idle when they are off shift and have completed caring for their initial patients.

Under the policy LS, in a single day the team on shift has c patients in care at all times each recovering at rate μ, producing $\mathrm{Pois}(c\mu)$ departures. Thus the expected number of departures from the team on shift is $c\mu$. For the team off shift, as we assumed there were infinitely many patients initially in care, we begin with all c capacity utilized. Again patients depart at rate μ, but now when they leave they are not replaced. The probability a patient will depart is $\mathbb{P}(\mathrm{Exp}(\mu) \le 1) = 1 - e^{-\mu}$. As whether or not each patient departs is independent, we have $\mathrm{Bin}(c, 1-e^{-\mu})$ departures, so the expected number of departures from the team off shift is $c(1-e^{-\mu})$. Thus the expected change for the number of patients in the system is given by

$$-\gamma_{\mathrm{LS}} \triangleq \lambda - c\mu - c(1-e^{-\mu}).$$

Recalling that we expect the system to be stable when $\gamma_{\text{LS}} > 0$, we see from (11) that this is equivalent to $\rho_{\text{LS}} < 1$. Performing a similar computation for DA, we see that in a single day there are $\text{Pois}(c\mu)$ on shift departures and $\text{Bin}(2c, 1 - e^{-\mu/2})$ off shift departures, giving an expected change in the number of patients in system of

$$-\gamma_{\text{DA}} \stackrel{\Delta}{=} \lambda - c\mu - 2c(1 - e^{-\mu/2}).$$

Thus the queue should be stable if $\gamma_{\text{DA}} > 0$, or equivalently from (12), when $\rho_{\text{DA}} < 1$. To show $\rho_{\text{DA}} < \rho_{\text{LS}}$, it suffices to show that $2 - 2e^{-\mu/2} > 1 - e^{-\mu}$, which follows since

$$1 - 2e^{-\mu/2} + e^{-\mu} = (1 - e^{-\mu/2})^2 > 0. \tag{13}$$

As $c\mu + 2c(1 - e^{-\mu/2}) < 2c\mu$, we see that DA still results in fewer expected departures than an $M/M/2c$ queue. However, if we are willing to consider schedules with more shift changes per day, we can achieve an expected number of departures arbitrarily close to our "upper bound" of $2c\mu$ by generalizing the policy DA. Given $k > 0$ integer and even, consider the schedule where both teams are on shift for $[i/k, (i + 1)/k)$ for all i even (the case of $i = 2$ is simply the policy DA). This divides the day into k equally sized pieces, where for $k/2$ such pieces both teams are on shift, and for the remaining $k/2$ periods both teams are off shift. We see immediately that independent of k, each team still spends half of each day on shift. In this half day on shift, our two teams' $2c$ capacity will again have $\text{Pois}(c\mu)$ departures. Now in each of our off shift periods, the probability of a patient leaving is $\mathbb{P}(\text{Exp}(\mu) \leq 1/k) = 1 - e^{-\mu/k}$, so we have $\text{Bin}(2c, 1 - e^{-\mu/k})$ off shift departures in each of our $k/2$ off shifts, or $\text{Bin}(kc, 1 - e^{-\mu/k})$ off shift departures per day. Thus the expected off shift departures per day is $kc(1 - e^{-\mu/k})$. Letting $k \to \infty$, we see through Taylor expansion that our off shift departures tend to $c\mu$, giving $2c\mu$ total departures as with the $M/M/2c$ queue. While in practice, we cannot have arbitrarily short shifts, we do see a general trend that shorter shifts increase capacity.

The stability property however is not the only relevant performance measure. An important quantity to look at is the number of *patient reassignments* (i.e. the number of arriving patients forced to wait due to the non-availability of residents, as discussed in the introduction). For each policy $\theta \in \{\text{LS}, \text{DA}\}$, we can easily verify that $\mathbf{S}_\theta(k)$ is irreducible and aperiodic on $\mathcal{S}_\theta$. Thus under the condition $\rho_\theta < 1$, there exists a unique steady state distribution for $\mathbf{S}_\theta(t)$, and we denote this

random vector by $\mathbf{S}_\theta(\infty)$. Analyzing $\mathbf{S}_\theta(\infty)$ directly appears to be intractable. Instead, we resort to the method of fluid approximation, which we now define.

Given the parameters of our queueing model λ_1, λ_2, μ and c, we consider a sequence of approximate models $n = 1, 2, \ldots$, where we change the parameters so that in the nth model, $\lambda_1^n = \lambda_1 n$, $\lambda_2^n = \lambda_2 n$, $\mu^n = \mu$, and $c^n = cn$. In words, the rate of patient recovery is fixed, but the patient arrival rates and patient capacity scale up linearly. For each policy $\theta \in \{\text{LS}, \text{DA}\}$, we let $Q_\theta^n(t), R_\theta^n(t)$ and $\mathbf{S}_\theta^n(t)$ be the corresponding processes. We let $\mathcal{S}_\theta^n$ be the set $\mathcal{S}_\theta$ as defined in (5) and (10) for the processes $\mathbf{S}_\theta^n(t)$. The process associated with fluid rescaling is defined as $\mathbf{S}_\theta^n(t)/n$. We immediately note by (11) and (12) that ρ_θ does not change with n, so the stability criteria for each $\mathbf{S}_\theta^n(t)$ is the same. Thus for $\rho_\theta < 1$ the sequence $\mathbf{S}_\theta^n(\infty)/n$ is well defined. Our next main result is that as $n \to \infty$, the sequences $\mathbf{S}_\theta^n(t)/n$ and $\mathbf{S}_\theta^n(\infty)/n$ converge meaningfully to some deterministic process $\mathbf{s}_\theta(t) = (q_\theta(t), r_\theta(t))$, and its unique fixed point $\lim_{k\to\infty} \mathbf{s}_\theta(k)$, respectively. We now provide details.

For LS, we define the process $\mathbf{s}_{\text{LS}}(t) = (q_{\text{LS}}(t), r_{\text{LS}}(t))$ on $\mathbb{R}_+ \times [0, c]$ inductively on intervals $[k, k+1)$. For each interval, consider the system of ordinary differential equations (ODEs)

$$\dot{q}_{\text{LS}}(t) = \lambda(t) - \mu(q_{\text{LS}}(t) \wedge c), \tag{14}$$

$$\dot{r}_{\text{LS}}(t) = -\mu r_{\text{LS}}(t). \tag{15}$$

At integer times $k \geq 1$, the process jumps as did $\mathbf{S}_{\text{LS}}(k)$. Specifically, we let

$$\mathbf{s}_{\text{LS}}(k) = \Gamma(\mathbf{s}_{\text{LS}}(k^-); c). \tag{16}$$

In analogy with $\mathcal{S}_{\text{LS}}$, we will show that at integer times $k \geq 1$ this process is actually restricted to

$$\mathcal{T}_{\text{LS}} \stackrel{\Delta}{=} [0, c]^2 \cup \mathbb{R}_+ \times \{c\}. \tag{17}$$

We now give a similar construction for $\mathbf{s}_{\text{DA}}(t) = (q_{\text{DA}}(t), r_{\text{DA}}(t))$ on $\mathbb{R}_+ \times [0, 2c]$. Again for each interval $[k, k+\frac{1}{2})$, we let $r_{\text{DA}}(t) = 0$ and define $q_{\text{DA}}(t)$ by

$$\dot{q}_{\text{DA}}(t) = \lambda_1 - \mu(q_{\text{DA}}(t) \wedge 2c). \tag{18}$$

At times $k + \frac{1}{2}$, $k \in Z_+$, we let

$$\mathbf{s}_{\text{DA}}(k + \tfrac{1}{2}) = \Gamma(\mathbf{s}_{\text{DA}}((k + \tfrac{1}{2})^-); 2c).$$

For each interval $[k+\frac{1}{2}, k+1)$, $q_{\mathrm{DA}}(t)$ and $r_{\mathrm{DA}}(t)$ are defined by the following ODEs:

$$\dot{q}_{\mathrm{DA}}(t) = \lambda_2,$$
$$\dot{r}_{\mathrm{DA}}(t) = -\mu r_{\mathrm{DA}}(t).$$

Again at integer times $k \geq 1$, we define

$$\mathbf{s}_{\mathrm{LS}}(k) = \Gamma(\mathbf{s}_{\mathrm{DA}}(k^-); 0).$$

We let

$$\mathcal{T}_{\mathrm{DA}} \stackrel{\Delta}{=} \mathbb{R}_+ \times \{0\}. \tag{19}$$

We will show that this is the set of possible values $\mathbf{s}_{\mathrm{DA}}(k)$ can take for integer $k \geq 1$.

PROPOSITION 1. *For every $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, and every $\mathbf{s}_\theta(0) \in \mathcal{T}_\theta$, $\mathbf{s}_\theta(t)$ exists and is uniquely defined for all $t \in \mathbb{R}_+$. Further, for all integer $k \geq 1$, $\mathbf{s}_\theta(k) \in \mathcal{T}_\theta$.*

The result is shown in Appendix A.4. We now formally relate $\mathbf{S}_\theta(t)$ to $\mathbf{s}_\theta(t)$.

THEOREM 2. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, if $\mathbf{S}^n_\theta(0)/n \to \mathbf{s}_\theta(0)$ a.s., then*

$$\lim_{n\to\infty} \frac{\mathbf{S}^n_\theta(t)}{n} = \mathbf{s}_\theta(t),$$

a.s. and u.o.c.

While this theorem allows us to approximate $\mathbf{S}_\theta(t)$ by the simpler process $\mathbf{s}_\theta(t)$, we have not established any relationship between $\mathbf{S}_\theta(\infty)$ and $\mathbf{s}_\theta(k)$ as $k \to \infty$. We do this next, but first we need some definitions.

Suppose we are given a discrete time dynamical system on a state space $\mathcal{X} \subset \mathbb{R}^n$ defined by $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$, i.e. $\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k)$ for all k. A point $\mathbf{x}^*$ is defined to be *attractive* if for all $\mathbf{x}_0 \in \mathcal{X}$,

$$\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}^*.$$

Note that there can be at most one attractive point. We now state our next result relating $\mathbf{s}_\theta(\infty)$ and $\mathbf{S}_\theta(\infty)$.

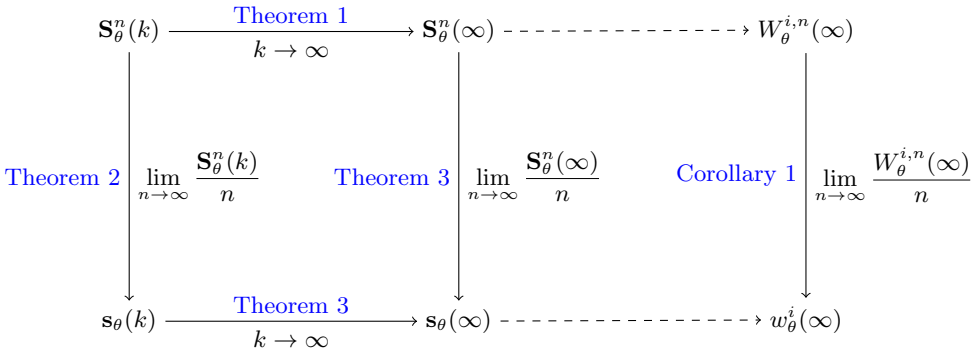

FIG 3. *A diagram explaining how each of our theorems relate our stochastic process and the fluid limit, for finite times, at steady state, and then finally for the steady state number of reassignments.*

THEOREM 3. *For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the sequence $\mathbf{s}_\theta(k)$ has a unique attractive point $\mathbf{s}_\theta(\infty) \in \mathcal{T}_\theta$ iff $\rho_\theta < 1$. Moreover, when $\rho_\theta < 1$, the following convergence in probability takes place:*

$$\lim_{n\to\infty} \frac{\mathbf{S}_\theta^n(\infty)}{n} = \mathbf{s}_\theta(\infty).$$

Notice that condition for the existence of an attractive point for $\mathbf{s}_\theta(t)$ is exactly the same as the stability condition for $\mathbf{S}_\theta(t)$. In the second claim of Theorem 3, we are essentially justifying an interchange of limits, as informally we are "equating" $\lim_{n\to\infty}\lim_{k\to\infty} \mathbf{S}_\theta^n(k)/n$ with $\lim_{k\to\infty}\lim_{n\to\infty} \mathbf{S}_\theta^n(k)/n$, as shown in the left half of Figure 3.

We now use this result to approximate the steady state number of reassignments in the queueing model. For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$ and each $k \in \mathbb{Z}_+$, let $W_\theta^1(k)$ (resp. $W_\theta^2(k)$) be the number of arriving patients during $[k, k+\frac{1}{2})$ (resp. $[k+\frac{1}{2}, k+1)$) that are forced to wait a nonzero amount of time before assignment to a resident, i.e. the number of reassignments. Similarly, when $\rho_\theta < 1$, we define $W_\theta^1(\infty)$ (resp. $W_\theta^2(\infty)$) to be the steady state number patients forced to wait during $[0, \frac{1}{2})$ (resp. $[\frac{1}{2}, 1)$). Next we define variables for the fluid approximations of these quantities. Let $b_\theta(t)$ be

$$b_{\mathrm{LS}}(t) = \mathbb{I}_{\{q_{\mathrm{LS}}(t) \geq c\}}, \tag{20}$$

$$b_{\mathrm{DA}}(t) = \begin{cases} \mathbb{I}_{\{q_{\mathrm{DA}}(t) \geq 2c\}} & t \in [k, k+\frac{1}{2}), \\ 1 & t \in [k+\frac{1}{2}, k+1), \end{cases} \tag{21}$$

i.e. $b_\theta(t)$ is the indicator that the on shift teams are saturated. We define

$$w_\theta^1(k) = \int_k^{k+\frac{1}{2}} b_\theta(t)\lambda_1 dt, \tag{22}$$

$$w_\theta^2(k) = \int_{k+\frac{1}{2}}^{k+1} b_\theta(t)\lambda_2 dt. \tag{23}$$

When $\rho_\theta < 1$, we let $w_\theta^1(\infty) = w_\theta^1(0)$ and $w_\theta^2(\infty) = w_\theta^2(0)$ assuming the fluid system begins in steady state, i.e. $\mathbf{s}_\theta(0) = \mathbf{s}_\theta(\infty)$. We next argue that $w_\theta^1(\infty)$ and $w_\theta^2(\infty)$ asymptotically describe the steady state number reassignments. Let $W_\theta^{1,n}(k)$ and $W_\theta^{2,n}(k)$ be the number of reassignments for $\mathbf{S}_\theta^n(t)$ from our fluid approximation. Then

COROLLARY 1. *For policy LS, assuming $\lambda_1, \lambda_2 \neq c\mu$, and for policy DA, assuming $\lambda_1 \neq 2c\mu$, the following convergence in probability takes place:*

$$\lim_{n\to\infty} \frac{W_\theta^{1,n}(\infty)}{n} = w_\theta^1(\infty),$$
$$\lim_{n\to\infty} \frac{W_\theta^{2,n}(\infty)}{n} = w_\theta^2(\infty).$$

The case when $\lambda_j = c\mu$ for either $j = 1, 2$ presents some annoying technical difficulties. As realistically we will never have exact equality, we do not pursue this issue further. This sequence of results justifies approximating $W_\theta^1(\infty)$ and $W_\theta^2(\infty)$ by $w_\theta^1(\infty)$, $w_\theta^2(\infty)$, respectively. The result of Corollary 1 are summarized in the right half of Figure 3.

3. Numerical Results. In this section, we numerically solve for the steady state solution of the fluid model of each policy. We then compare the cost of the reassignments in a single day starting at steady state under each policy as we vary the average arrival rate. We relate our numerical observations to our empirical observations from Figure 2.

Throughout this section, we use the following parameters in our model: $\mu = 1/2$, $c = 40$, $\lambda_1 = 9\lambda/5$, and $\lambda_2 = \lambda/5$. Our choice of μ and c imply that $\lambda_{\text{LS}}^* \approx 35.7388$ and $\lambda_{\text{DA}}^* \approx 37.6959$. The value of c and the ratio of λ_1 to λ_2 were chosen to be representative of a department from a large hospital such as B&W. The value of μ must be chosen more carefully. In light of Remark 1, we set μ to control the relative sizes of the average length of stay and length of time between shifts. At B&W under the policy LOSO, there is a long shift every four days and the average patient length

of stay is four days. Thus in our model we set the average length of stay $(1/\mu)$ to be two days as the policy LS has a long shift every two days.

In Figure 4, we fix λ at 34 and observe the steady state behavior of our two policies in the fluid limit over the course of a day. Notice that $\lambda < \lambda^*_{\text{LS}} < \lambda^*_{\text{DA}}$, so under both policies the fluid model is stable, but heavily loaded. We see that for both policies, under these particular parameters, the number of patients being treated by the teams on shift plus the number of patients waiting, $q_\theta(t)$, increases over the first half of day. For LS, the capacity of 40 for the teams on shift (as indicated by the dotted black line) is exceeded, and resulting in some reassignments. For DA however, as both teams are working during the first half of the day, we stay below the capacity of 80 patients and have no reassignments. In the second half of the day, under LS we see that the backlog of patients subsides and we return below 40 patients by the end of the day. For DA, as both teams are off shift during the second half of the day, we see a jump at time 1/2 between q_{DA} and r_{DA} and then small backlog of arrivals accumulate in the second half of the day.

In Figure 5, we show the number of reassignments for each policy in $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ as we vary λ. The dotted vertical lines indicate λ^*_{LS} and λ^*_{DA}, the largest patient arrival rates such that LS and DA are stable. We see that our observation from Figure 4, that LS had many reassignments in $[0, \frac{1}{2})$ while DA had no reassignments in this period, is typical when the system is heavily loaded (for λ near λ^*_{LS}). We also see in Figure 5 that when λ is low, both polices cause no reassignments in the first half of the day, and only DA causes reassignments in the second half the day. As λ increases towards λ^*_{LS}, we see LS begin to reassign nearly all patients, while DA continues to only reassign patients arriving in the second half of the day. Finally, for very large λ, we eventually see DA reassigning some patients during the first half of the day. While for these particular parameter settings, we only see DA reassignments in the first half of the day for λ so large that LS is unstable, this does not hold for all parameter settings. Interestingly, we see that under DA for λ near λ^*_{DA}, the number of reassignments does not approach λ, while it does for LS. This is occurring as under these parameters, we have more patients leaving than arriving in the second half of the day, creating some spare capacity during the start of the first half of the following day.

Comparing Figure 5 with our empirical observations from Figure 2 we see that the relationship between LS and DA is qualitatively similar to the relationship between the B&W policies LOSO

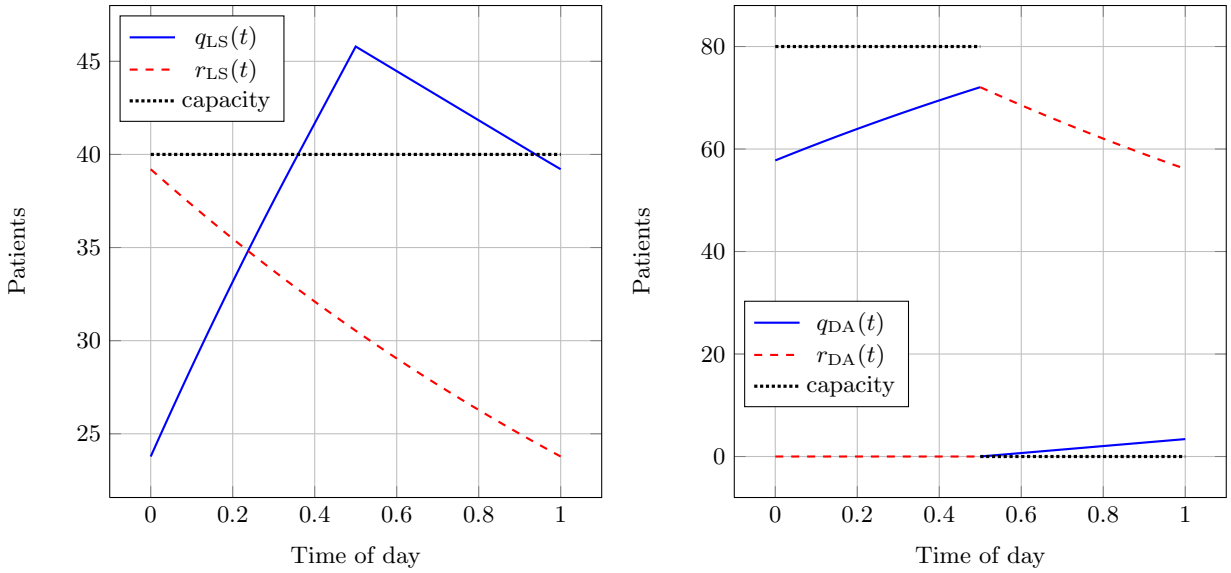

FIG 4. *Steady state number of patients in system in the fluid limit under each policy. Parameter values:* $\lambda = 34$, $\lambda_1 = 9\lambda/5$, $\lambda_2 = \lambda/5$, $\mu = 1/2$, $c = 40$.

and MMMO. Most importantly, we have preserved the property that shorter more frequent shifts are the better policy when the patient load is heavy.

4. Conclusion. We have developed a queueing model to determine the effect of long shifts in medical resident schedules on the hospital's capacity to admit patients and the quality of care delivered. Our model was motivated by the empirical work [1] on scheduling medical residents for B&W hospital. In this paper, we compared the stylized schedules *Long Shifts* (LS), where residents worked 24 hour shifts on alternating days, and *Daily Admitting* (DA), where residents worked every day but only during peak arrival hours. We used Lyapunov function techniques to characterize the stability of our queueing model under each policy. We found that DA has a greater capacity to admit patients than LS for all parameter choices. To analyze the long-run performance of our queueing model, we first considered the associated fluid model, which is a deterministic system with periodic dynamics. We showed that under each policy, when the queueing model is stable, the fluid model had a unique periodic steady state solution. We showed that our queueing model under the fluid rescaling converges to the fluid model on finite time intervals. Then we used an interchange of limits argument to show that the steady state queue lengths under the fluid rescaling converge to the unique steady state solution of the fluid model. We use these results to approximate the

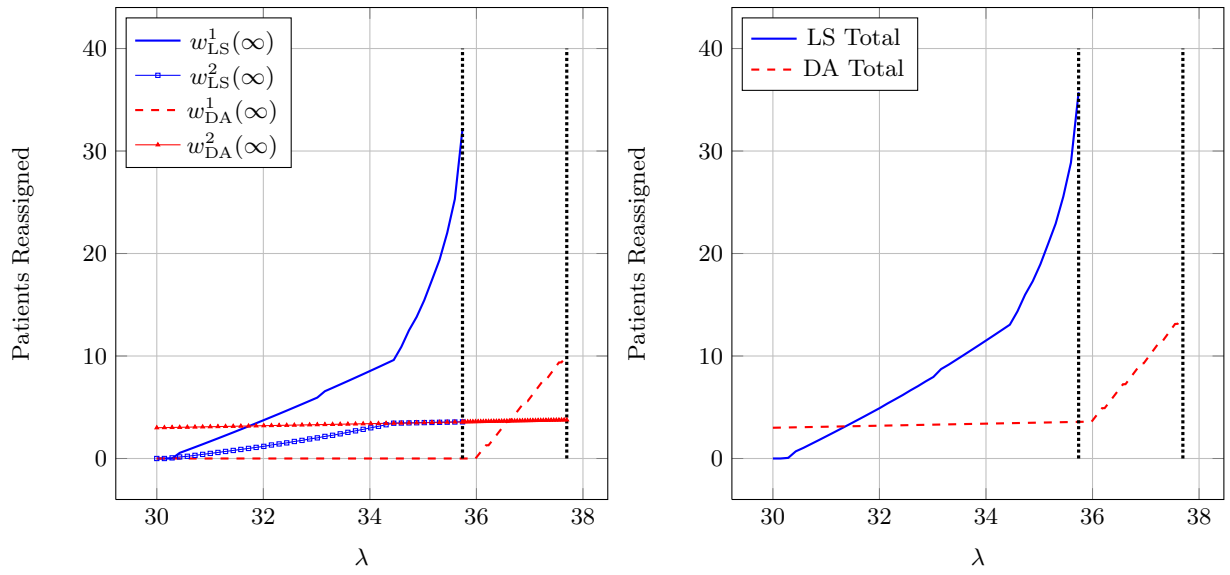

FIG 5. *Steady state number of patient reassignments in the fluid limit for different λ under each policy. Parameter values:* $\lambda_1 = 9\lambda/5$, $\lambda_2 = \lambda/5$, $\mu = 1/2$, $c = 40$.

steady state number of reassignments in our queueing model by the steady state behavior of the fluid model. Numerically solving for the steady state of the fluid model under various parameter choices, we found evidence suggesting the existence of a threshold value on the arrival rate such that DA causes fewer reassignments than LS iff the arrival rate exceeds the threshold value. These results substantiate the main empirical findings in [1]. The issue of resident schedules is currently quite pertinent, as new regulations restrict residents to a maximum shift length of 16 hours [23]. Our work contributes to understanding the implication of the new regulation. As hospitals tend to operate in heavily loaded regimes, we find that schedules relying on shorter more frequent shifts could increase capacity and reduce reassignments.

Acknowledgments. The authors thank Brigham and Women's Hospital for providing the financial support for this work. The authors also wish to thank David Goldberg for his initial work on the B&W hospital project and technical discussions.

REFERENCES

[1] R. Anderson and D. Gamarnik. Impact of resident shift lengths on the delivery of care. Forthcoming, 2012.

[2] S. Asmussen. *Applied probability and queues*. Springer Verlag, 2003.

[3] N.T. Ayas, L.K. Barger, B.E. Cade, D.M. Hashimoto, B. Rosner, J.W. Cronin, F.E. Speizer, and C.A. Czeisler. Extended work duration and the risk of self-reported percutaneous injuries in interns. *JAMA: the journal of the American Medical Association*, 296(9):1055, 2006.

[4] L.K. Barger, B.E. Cade, N.T. Ayas, J.W. Cronin, B. Rosner, F.E. Speizer, C.A. Czeisler, et al. Extended work shifts and the risk of motor vehicle crashes among interns. *New England Journal of Medicine*, 352(2):125, 2005.

[5] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

[6] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Verlag, 2006.

[7] H. Chen and D.D. Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*. Springer Verlag, 2001.

[8] R.I. Cook, M. Render, and D.D. Woods. Gaps in the continuity of care and progress on patient safety. *British Medical Journal*, 320(7237):791, 2000.

[9] F. de Véricourt and O.B. Jennings. Nurse-to-patient ratios in hospital staffing: a queuing perspective. *ESMT Working Paper No. 08-005*, 2008.

[10] R.S. Dittus, R.W. Klein, D.J. DeBrota, M.A. Dame, and J.F. Fitzgerald. Medical resident work schedules: design and evaluation by simulation modeling. *Management Science*, 42(6):891–906, 1996.

[11] R.M. Dudley. *Real analysis and probability*. Cambridge University Press, 2002.

[12] E. El-Darzi, C. Vasilakis, T. Chaussalet, and PH Millard. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149, 1998.

[13] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence*. Wiley New York, 1986.

[14] G. Fayolle, V.A. Malyshev, and M.V. Menshikov. *Topics in the constructive theory of countable Markov chains*. Cambridge University Press, 1995.

[15] D.M. Gaba and S.K. Howard. Fatigue among clinicians and the safety of patients, 2002.

[16] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, 16(1):56–90, 2006.

[17] P.W. Glynn and A. Zeevi. Bounding stationary expectations of Markov processes. In *Markov processes and related topics: A Festschrift for Thomas G. Kurtz. Selected papers of the conference, Madison, WI, USA, July*, pages 10–13, 2006.

[18] L. Green. Capacity planning and management in hospitals. *Operations Research and Health Care*, pages 15–41, 2005.

[19] L.V. Green, P.J. Kolesar, and W. Whitt. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management*, 16(1):13–39, 2007.

[20] L.V. Green, J. Soares, J.F. Giglio, and R.A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[21] M.M. Hutter, K.C. Kellogg, C.M. Ferguson, W.M. Abbott, and A.L. Warshaw. The impact of the 80-hour resident workweek on surgical residents and attending surgeons. *Annals of surgery*, 243(6):864, 2006.

[22] R. Ibrahim and W. Whitt. Wait-Time Predictors for Customer Service Systems With Time-Varying Demand and Capacity. *Oper. Res., forthcoming. Columbia University, NY, NY*, 2010.

[23] J.K. Iglehart. The ACGME's Final Duty-Hour Standards Special PGY-1 Limits and Strategic Napping. *New England Journal of Medicine*, 363(17):1589–1591, 2010.

[24] S.H. Jacobson, S.N. Hall, and J.R. Swisher. Discrete-event simulation of health care systems. *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 211–252, 2006.

[25] R.W. Klein, R.S. Dittus, D.J. DeBrota, and M.A. Dame. Using discrete event simulation to evaluate housestaff work schedules. In *Proceedings of the 22nd conference on Winter simulation*, pages 738–742. IEEE Press, 1990.

[26] J. Lamperti. Criteria for the recurrence or transience of stochastic process, part I. *J. Math. Anal. and Appls.*, 1, 1960.

[27] C.P. Landrigan, J.M. Rothschild, J.W. Cronin, R. Kaushal, E. Burdick, J.T. Katz, C.M. Lilly, P.H. Stone, S.W. Lockley, D.W. Bates, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine*, 351(18):1838, 2004.

[28] E. Litvak and Inc Joint Commission Resources. *Managing Patient Flow in Hospitals: Strategies and Solutions*. Joint Commission Resources, 2010.

[29] Y. Liu and W. Whitt. A fluid model for a large-scale service system experiencing periods of overloading. 2010.

[30] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1):149–201, 1998.

[31] L.A. Petersen, T.A. Brennan, A.C. O'Neil, E.F. Cook, and T.H. Lee. Does housestaff discontinuity of care increase the risk for preventable adverse events? *Annals of internal medicine*, 121(11):866, 1994.

[32] M.D. Rossetti, G.F. Trzcinski, and S.A. Syverud. Emergency department simulation and determination of optimal attending physician staffing schedules. In *wsc*, pages 1532–1540. IEEE, 1999.

[33] J.S. Samkoff and CH Jacques. A review of studies concerning effects of sleep deprivation and fatigue on residents' performance. *Academic Medicine*, 66(11):687, 1991.

[34] T. Schoenmeyr, P.F. Dunn, D. Gamarnik, R. Levi, D.L. Berger, B.J. Daily, W.C. Levine, and W.S. Sandberg. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, 110(6):1293, 2009.

[35] R. Steinbrook. The debate over residents' work hours. *New England Journal of Medicine*, 347(16):1296, 2002.

[36] J.N. Tsitsiklis and K. Xu. On the Power of (even a little) Centralization in Distributed Processing. *ACM Sigmetrics, San Jose*, 2011.

[37] D.F. Weinstein. Duty hours for resident physicians–tough choices for teaching hospitals. *New England Journal of Medicine*, 347(16):1275, 2002.

[38] E.E. Whang, M.M. Mello, S.W. Ashley, and M.J. Zinner. Implementing resident work hour limitations: lessons from the New York State experience. *Annals of surgery*, 237(4):449, 2003.

[39] K.P. White Jr. A survey of data resources for simulating patient flows in healthcare delivery systems. In *Proceedings of the 37th conference on Winter simulation*, pages 926–935. Winter Simulation Conference, 2005.

[40] G. Yom-Tov. Queues in Hospitals: Semi-Open Queueing Networks in the QED Regime. *IE PhD thesis proposal*, 2008.

[41] G. Yom-Tov and A. Mandelbaum. Time-varying QED queues with reentrant customers in support of healthcare staffing. *The Technion, Haifa, Israel*, 2010.

APPENDIX A: PROOFS

A.1. Stability conditions for two schedules. Proof of Theorem 1. In this section, we prove Theorem 1, characterizing the stability of $\mathbf{S}_\theta(t)$. We show these results using the method of Lyapunov functions, which we describe next. Let $\{\mathbf{X}_k\}$ be a discrete time irreducible Markov chain on a countable state space $\mathcal{X} \subset \mathbb{Z}^d$. First we give a condition for the positive recurrence of $\{\mathbf{X}_k\}$ due to Foster (see [2]).

PROPOSITION 2. *If there exists a function $V\colon \mathcal{X} \to \mathbb{R}$, $\gamma > 0$, and a finite set $B \subset \mathcal{X}$ such that for all $\mathbf{x} \in B$,*

$$\mathbb{E}_{\mathbf{x}}[V(\mathbf{X}_1) - V(\mathbf{X}_0)] < \infty, \tag{24}$$

and for all $\mathbf{x} \in \mathcal{X} \setminus B$,

$$\mathbb{E}_{\mathbf{x}}[V(\mathbf{X}_1) - V(\mathbf{X}_0)] \leq -\gamma,$$

then $\{\mathbf{X}_k\}$ is positive recurrent.

A function V satisfying these properties is usually called a *Lyapunov function*. Lyapunov functions can also be used to prove $\{\mathbf{X}_k\}$ is not positive recurrent when the drift is nonnegative. The following is a special case of Proposition 5.4 from [2].

PROPOSITION 3. *Suppose there exists a Lyapunov function $V\colon \mathcal{X} \to \mathbb{R}$, a finite set $B \subset \mathcal{X}$ and*

a state $\mathbf{y} \in \mathcal{X} \setminus B$ *satisfying*

$$\sup_{\mathbf{x}\in B} V(\mathbf{x}) < V(\mathbf{y}), \tag{25}$$

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{x}}\left[(V(\mathbf{X}_1) - V(\mathbf{X}_0))^2\right] < \infty, \tag{26}$$

$$\inf_{x\in\mathcal{X}\setminus B} \mathbb{E}_{\mathbf{x}}[V(\mathbf{X}_1) - V(\mathbf{X}_0)] \geq 0. \tag{27}$$

Then $\{\mathbf{X}_k\}$ *is either null recurrent or transient.*

Returning to our model, we now let the Lyapunov function $V\colon \mathcal{S}_\theta \to \mathbb{R}_+$ be defined by $V(q,r) = q + r$ for the Markov chain $\{\mathbf{S}_\theta(k)\}$. We first analyze the drift under the policy LS, namely $\mathbb{E}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))]$. Let $A \triangleq A(0,1)$, $D_{\mathrm{LS}}^{\mathrm{on}} \triangleq D_{\mathrm{LS}}^{\mathrm{on}}(0,1)$, and $D_{\mathrm{LS}}^{\mathrm{off}} \triangleq D_{\mathrm{LS}}^{\mathrm{off}}(0,1)$ denote the number of arrivals and departures in a single day under LS.

LEMMA 1. *We have*

$$-\gamma_{\mathrm{LS}} = \lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))] = \inf_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))]. \tag{28}$$

Additionally, there exists a constant $C_{\mathrm{LS}} > 0$ *depending only on* μ *such that*

$$\sup_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2] \leq \lambda^2 + \lambda + C_{\mathrm{LS}}(c^2 + c). \tag{29}$$

PROOF. First, observe that the value of the Lyapunov function does change at time 1:

$$V(\mathbf{S}_{\mathrm{LS}}(1)) = V(\Gamma(\mathbf{S}_{\mathrm{LS}}(1^-); c)) = V(\mathbf{S}_{\mathrm{LS}}(1^-)), \tag{30}$$

as applying Γ does not change the number of patients in system. Thus for $\ell = 1, 2$,

$$\begin{aligned} \mathbb{E}_{(q,r)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^\ell\right] &= \mathbb{E}_{(q,r)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1^-)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^\ell\right] \\ &= \mathbb{E}_{(q,r)}\left[(A - D_{\mathrm{LS}}^{\mathrm{on}} - D_{\mathrm{LS}}^{\mathrm{off}})^\ell\right]. \end{aligned} \tag{31}$$

Let $\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \stackrel{d}{=} \mathrm{Pois}(c\mu)$ and $\tilde{D}_{\mathrm{LS}}^{\mathrm{off}} \stackrel{d}{=} \mathrm{Bin}(c, 1 - e^{-\mu})$ such that $\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}$, $\tilde{D}_{\mathrm{LS}}^{\mathrm{off}}$, and A are independent. As $D_{\mathrm{LS}}^{\mathrm{on}}(0,t)$ for $0 \leq t < 1$ has the distribution of the departure process for an $M(t)/M/c$ queue, we can couple $D_{\mathrm{LS}}^{\mathrm{on}}$ with $\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}$ such that regardless of $\mathbf{S}_{\mathrm{LS}}(0)$,

$$D_{\mathrm{LS}}^{\mathrm{on}} \leq \tilde{D}_{\mathrm{LS}}^{\mathrm{on}}. \tag{32}$$

For the off shift departures, as the patient length of stay is exponential, each of the $r = R(0) \leq c$ patients in care will depart in the interval $[0, 1)$ with probability $1 - e^{-\mu}$ independently of other patients. Thus $D_{\mathrm{LS}}^{\mathrm{off}} \stackrel{d}{=} \mathrm{Bin}(r, 1 - e^{-\mu})$, so trivially it can be coupled with $\tilde{D}_{\mathrm{LS}}^{\mathrm{off}}$ such that

$$D_{\mathrm{LS}}^{\mathrm{off}} \leq \tilde{D}_{\mathrm{LS}}^{\mathrm{off}}, \tag{33}$$

with equality when $r = c$. From (32) and (33) we obtain that for any initial condition $\mathbf{S}(0) = (q, r) \in \mathcal{S}_{\mathrm{LS}}$

$$A - D_{\mathrm{LS}}^{\mathrm{on}} - D_{\mathrm{LS}}^{\mathrm{off}} \geq A - \tilde{D}_{\mathrm{LS}}^{\mathrm{on}} - \tilde{D}_{\mathrm{LS}}^{\mathrm{off}}.$$

Taking expectations and then the infimum over all $(q, r) \in \mathcal{S}_{\mathrm{LS}}$, we obtain that

$$\inf_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}(0))] \geq \lambda - c\mu - c(1 - e^{-\mu}) = -\gamma_{\mathrm{LS}}.$$

Observe that for any realization where $Q(0) = q$ and $D_{\mathrm{LS}}^{\mathrm{on}} \leq \tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q - c$, we also have $Q_{\mathrm{LS}}(t) \geq c$ for all $t \in [0, 1^-)$, and thus $D_{\mathrm{LS}}^{\mathrm{on}} = \tilde{D}_{\mathrm{LS}}^{\mathrm{on}}$. As a result,

$$\mathbb{E}_{(q,c)}[D_{\mathrm{LS}}^{\mathrm{on}}] \geq \mathbb{E}_{(q,c)}\left[D_{\mathrm{LS}}^{\mathrm{on}} \mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q-c\}}\right] = \mathbb{E}\left[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q-c\}}\right]. \tag{34}$$

Since almost surely

$$\lim_{q\to\infty} \tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q-c\}} = \tilde{D}_{\mathrm{LS}}^{\mathrm{on}},$$

by monotonicity of expectation and then the Monotone Convergence Theorem, we obtain that

$$\liminf_{q\to\infty} \mathbb{E}_{(q,c)}[D_{\mathrm{LS}}^{\mathrm{on}}] \geq \lim_{q\to\infty} \mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q-c\}}] = \mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}] = c\mu.$$

Combining this inequality with (32), we obtain $\lim_{q\to\infty} \mathbb{E}_{(q,c)}[D_{\mathrm{LS}}^{\mathrm{on}}] = c\mu$ and thus

$$\lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))] = \lambda - c(1 - e^{-\mu}) - c\mu = -\gamma_{\mathrm{LS}}.$$

Lastly, to show (29), using independence, (32), and (33),

$$\begin{aligned}\sup_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[(A - D_{\mathrm{LS}}^{\mathrm{on}} - D_{\mathrm{LS}}^{\mathrm{off}})^2] &\leq \mathbb{E}[A^2] + \mathbb{E}[(\tilde{D}_{\mathrm{LS}}^{\mathrm{on}})^2] + \mathbb{E}[(\tilde{D}_{\mathrm{LS}}^{\mathrm{off}})^2] + 2\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}]\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{off}}] \\ &= \lambda^2 + \lambda + c^2\mu^2 + c\mu + c(1 - e^{-\mu})e^{-\mu} + (c(1 - e^{-\mu}))^2 + 2c^2\mu(1 - e^{-\mu}).\end{aligned}$$

□

We now give an analogous result to the previous lemma for DA. As the proof is nearly the same, some details have been omitted. As before, let $D^{\text{on}}_{\text{DA}} \triangleq D^{\text{on}}_{\text{DA}}(0,\frac{1}{2})$ and $D^{\text{off}}_{\text{DA}} \triangleq D^{\text{off}}_{\text{DA}}(\frac{1}{2},1)$, give the number of departures under policy DA in a single day (note that there is no one on shift during $[\frac{1}{2},1)$ and no one off shift during $[0,\frac{1}{2})$ under DA).

LEMMA 2. *We have*

$$-\gamma_{\text{DA}} = \lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\text{DA}}(1)) - V(\mathbf{S}_{\text{DA}}(0))] = \inf_{(q,r)\in\mathcal{S}_{\text{DA}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\text{DA}}(1)) - V(\mathbf{S}_{\text{DA}}(0))]. \tag{35}$$

Additionally, there exists a constant $C_{\text{DA}} > 0$ depending only on μ such that

$$\sup_{(q,r)\in\mathcal{S}_{\text{DA}}} \mathbb{E}_{(q,r)}[(V(\mathbf{S}_{\text{DA}}(1)) - V(\mathbf{S}_{\text{DA}}(0)))^2] \le \lambda^2 + \lambda + C_{\text{DA}}(c^2+c). \tag{36}$$

PROOF. Again, for any initial state $\mathbf{S}_{\text{DA}}(0) = (q,0)$, $V(\mathbf{S}_{\text{DA}}(1)) = V(\mathbf{S}_{\text{DA}}(1^-))$ and $V(\mathbf{S}_{\text{DA}}(\frac{1}{2})) = V(\mathbf{S}_{\text{DA}}(\frac{1}{2}^-))$, as applying Γ at times $\frac{1}{2}$ and 1 does not change the number of patients in system. Thus for $\ell = 1, 2$,

$$\mathbb{E}_{(q,0)}[(V(\mathbf{S}_{\text{DA}}(1)) - V(\mathbf{S}_{\text{DA}}(0)))^\ell] = \mathbb{E}_{(q,0)}\left[(A - D^{\text{on}}_{\text{DA}} - D^{\text{off}}_{\text{DA}})^\ell\right]. \tag{37}$$

Let $\tilde{D}^{\text{on}}_{\text{DA}} \stackrel{d}{=} \text{Pois}(c\mu)$ and $\tilde{D}^{\text{off}}_{\text{DA}} \stackrel{d}{=} \text{Bin}(2c, 1-e^{-\mu/2})$ such that $\tilde{D}^{\text{on}}_{\text{DA}}$, $\tilde{D}^{\text{off}}_{\text{DA}}$, and A are independent. As $D^{\text{on}}_{\text{DA}}(0,t)$ for $0 \le t < \frac{1}{2}$ has the distribution of the departure process for an $M(t)/M/2c$ queue, we can couple $D^{\text{on}}_{\text{DA}}(0,\frac{1}{2}^-)$ with $\tilde{D}^{\text{on}}_{\text{DA}}$ such that regardless of $\mathbf{S}_{\text{DA}}(0)$,

$$D^{\text{on}}_{\text{DA}} \le \tilde{D}^{\text{on}}_{\text{DA}}. \tag{38}$$

For the off shift departures, as the patient length of stay is exponential and thus memoryless, each of the $R_{\text{DA}}(\frac{1}{2}) \le 2c$ patients in care will depart in the interval $[\frac{1}{2},1)$ with probability $1-e^{-\mu/2}$ independently of other patients. Thus $D^{\text{off}}_{\text{DA}} \stackrel{d}{=} \text{Bin}(R_{\text{DA}}(\frac{1}{2}), 1-e^{-\mu/2})$, so it can be coupled with $\tilde{D}^{\text{off}}_{\text{DA}}$ such that

$$D^{\text{off}}_{\text{DA}} \le \tilde{D}^{\text{off}}_{\text{DA}}, \tag{39}$$

with equality when $R_{\text{DA}}(\frac{1}{2}) = 2c$. From (38) and (39) we obtain that for any $(q,0) \in \mathcal{S}_{\text{DA}}$,

$$A - D^{\text{on}}_{\text{DA}} - D^{\text{off}}_{\text{DA}} \ge A - \tilde{D}^{\text{on}}_{\text{DA}} - \tilde{D}^{\text{off}}_{\text{DA}},$$

and thus by taking expectations

$$\inf_{(q,r)\in\mathcal{S}_{\mathrm{DA}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))] \geq \lambda - c\mu - 2c(1-e^{-\mu/2}) = -\gamma_{\mathrm{DA}}.$$

As before, to complete showing the three term equality in (35), it suffices to show $\lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))] = -\gamma_{\mathrm{DA}}$. Given $Q_{\mathrm{DA}}(0) = q$, for any realization such that $D^{\mathrm{on}}_{\mathrm{DA}} \leq \tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \leq q-2c$, we have $Q_{\mathrm{DA}}(t) \geq 2c$ for all $t \in [0, \frac{1}{2}^-)$, and thus $D^{\mathrm{on}}_{\mathrm{DA}} = \tilde{D}^{\mathrm{on}}_{\mathrm{DA}}$. Further, $Q_{\mathrm{DA}}(\frac{1}{2}^-) \geq 2c$ ensures $R_{\mathrm{DA}}(\frac{1}{2}) = 2c$ and thus $D^{\mathrm{off}}_{\mathrm{DA}} = \tilde{D}^{\mathrm{off}}_{\mathrm{DA}}$. As a result,

$$\mathbb{E}[D^{\mathrm{on}}_{\mathrm{DA}}] \geq \mathbb{E}\left[D^{\mathrm{on}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \leq q-2c\}}\right] = \mathbb{E}\left[\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \leq q-2c\}}\right],$$
$$\mathbb{E}[D^{\mathrm{off}}_{\mathrm{DA}}] \geq \mathbb{E}\left[D^{\mathrm{off}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \leq q-2c\}}\right] = \mathbb{E}\left[\tilde{D}^{\mathrm{off}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \leq q-2c\}}\right].$$

We can apply the Monotone Convergence Theorem as before but now on both $\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}} \leq q-2c\}}$ and $\tilde{D}^{\mathrm{off}}_{\mathrm{DA}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}} \leq q-2c\}}$ to obtain the desired limit. The rest of the proof is exactly as in the previous lemma. □

Proof of Theorem 1.. Suppose $\rho_\theta < 1$, i.e. $\gamma_\theta > 0$. From (28) of Lemma 1, we obtain that

$$-\gamma_{\mathrm{LS}} = \lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))].$$

Similarly, from (35) of Lemma 2, we

$$-\gamma_{\mathrm{DA}} = \lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))].$$

Recall by (5) that for every $(q,r) \in \mathcal{S}_{\mathrm{LS}}$, when $q \geq c$, we must have $r = c$, and by (10) for every $(q,r) \in \mathcal{S}_{\mathrm{DA}}$ we have $r = 0$. Thus the sets

$$B_\theta = \left\{ (q,r) \in \mathcal{S}_\theta \;\middle|\; \mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] > -\gamma_\theta/2 \right\},$$

are finite. Observe that for both LS and DA, (24) is satisfied by (29) and (36), respectively. Applying Proposition 2, taking $B = B_\theta$ and $\gamma = \gamma_\theta/2$, we conclude that $\{\mathbf{S}_\theta(k)\}$ is positive recurrent.

Now suppose instead that $\rho_\theta \geq 1$, i.e. $\gamma_\theta \leq 0$. In the setting of Proposition 3, for both θ we take $B_\theta = \{(0,0)\}$ and observe that (25) is trivially satisfied by taking $y = (q,r)$ for any nonzero $(q,r) \in \mathcal{S}_\theta$. The condition in (26) is satisfied by (29) for LS and (36) for DA. Finally, (27) is satisfied as by (28) and (35), we have

$$\inf_{(q,r)\in\mathcal{S}_\theta} \mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] = -\gamma_\theta \geq 0.$$

Thus from Proposition 3 we conclude that $\{\mathbf{S}_\theta(k)\}$ is either null recurrent or transient.

It remains to show that $\{\mathbf{S}_\theta(k)\}$ is null recurrent when $\rho_\theta = 1$ and transient when $\rho_\theta > 1$. To do so, we use another Lyapunov function argument, Theorem 3.2 from [26], (see also [14] section 3.6). However, as the statement of this theorem is rather technical, we defer this part of the proof to Appendix A.8.

Finally, that $\rho_{\mathrm{DA}} < \rho_{\mathrm{LS}}$ follows from (13). □

REMARK 1. Note that as $\mu \to 0$, $(1 - e^{-\mu/2})^2 \to 0$, so by (13) we see that $\rho_{\mathrm{LS}} - \rho_{\mathrm{DA}} \to 0$ as well. Intuitively, if patients take many days to recover, the amount of forced idle time due to not being able to admit patients while off shift will be negligible. Conversely, when μ is large,

$$\rho_{\mathrm{LS}} \approx \frac{\lambda}{c\mu + c}, \qquad \rho_{\mathrm{DA}} \approx \frac{\lambda}{c\mu + 2c}.$$

In this regime, nearly all patients recover in each off shift. When c is also large, we see DA has a larger stability region, due to the off shifts for each team being shorter. While this regime isn't particularly relevant in the hospital setting, where $\mu \approx 1/4$, it exposes another interesting and relevant parameter, namely the length of the time between shifts relative to the recovery rate μ.

A.2. Uniform bounds for stationary performance measures. In this section, we consider the sequence of systems under the fluid rescaling $\mathbf{S}^n_\theta(t)/n$ with the assumption that $\rho_\theta < 1$, and give bounds on the expected stationary queue lengths that are independent of n. In the notation of Proposition 2 and Proposition 3, again suppose $\{\mathbf{X}_k\}$ is a discrete time irreducible Markov chain on a countable state space $\mathcal{X} \subset \mathbb{Z}^d$. Further, suppose that $\{\mathbf{X}_k\}$ is positive recurrent, and let $\mathbf{X}_\infty$ denote the unique steady state distribution. We now give a bound on the first moment of $f(\mathbf{X}_\infty)$ for any function f. The bound is similar to results from [16, 17].

PROPOSITION 4. *Suppose there exists $\alpha, \beta, \gamma > 0$, a bounded set $B \subset \mathcal{X}$ and a Lyapunov function $U\colon \mathcal{X} \to \mathbb{R}_+$ where $f\colon \mathcal{X} \to \mathbb{R}$ is such that for $\mathbf{x} \in \mathcal{X} \setminus B$,*

$$\mathbb{E}_{\mathbf{x}}[U(\mathbf{X}_1) - U(\mathbf{X}_0)] \leq -\gamma f(\mathbf{x}), \tag{40}$$

and for $\mathbf{x} \in B$,

$$f(\mathbf{x}) \leq \alpha, \tag{41}$$

$$\mathbb{E}_{\mathbf{x}}[U(\mathbf{X}_1) - U(\mathbf{X}_0)] \leq \beta. \tag{42}$$

Then

$$\mathbb{E}[f(\mathbf{X}_\infty)] \leq \alpha + \frac{\beta}{\gamma}.$$

PROOF. For every $z > 0$ let $U_z \colon \mathcal{X} \to \mathbb{R}_+$ and $f_z \colon X \to \mathbb{R}_+$ be given by

$$U_z(\mathbf{x}) \triangleq \min\{U(\mathbf{x}), z\}, \qquad f_z(\mathbf{x}) \triangleq f(\mathbf{x})\mathbb{I}_{\{U(\mathbf{x})<z\}}.$$

Trivially for all sufficiently large z, we have $f_z(\mathbf{x}) = f(\mathbf{x})$ and $U_z(\mathbf{x}) = U(\mathbf{x})$ for all $\mathbf{x} \in B$, so (41) and (42) are satisfied by f_z and U_z for all large z. We claim that (40) is satisfied as well. Suppose that $\mathbf{x}$ is such that $U(\mathbf{x}) \geq z$. Then

$$\mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)] = \mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1)] - z \leq 0 = f_z(\mathbf{x}) = -\gamma f_z(\mathbf{x}).$$

Alternatively, when $\mathbf{x}$ is such that $U(\mathbf{x}) < z$, using that $U_z(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x}$,

$$\mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)] = \mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U(\mathbf{X}_0)] \leq \mathbb{E}_{\mathbf{x}}[U(\mathbf{X}_1) - U(\mathbf{X}_0)] \leq -\gamma f(\mathbf{x}) = -\gamma f_z(\mathbf{x}).$$

Thus for all z and all $\mathbf{x} \in \mathcal{X} \setminus B$, $\mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)] \leq -\gamma f_z(\mathbf{x})$ in analogy with (40). As for all z, U_z is bounded by construction, $\mathbb{E}[U_z(\mathbf{X}_\infty)]$ is finite. By stationarity and the finiteness of $\mathbb{E}[U_z(\mathbf{X}_\infty)]$,

$$\begin{aligned}
0 &= \mathbb{E}\left[\mathbb{E}_{\mathbf{X}_\infty}\left[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)\right]\right] \\
&= \sum_{\mathbf{x}\in B} \mathbb{P}(\mathbf{X}_\infty = \mathbf{x})\mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)] + \sum_{\mathbf{x}\notin B} \mathbb{P}(\mathbf{X}_\infty = \mathbf{x})\mathbb{E}_{\mathbf{x}}[U_z(\mathbf{X}_1) - U_z(\mathbf{X}_0)] \\
&\leq \beta\mathbb{P}(\mathbf{X}_\infty \in B) - \gamma \sum_{\mathbf{x}\notin B} \mathbb{P}(\mathbf{X}_\infty = \mathbf{x}) f_z(\mathbf{x}) \\
&\leq \beta - \gamma\mathbb{E}[f_z(\mathbf{X}_\infty)\mathbb{I}_{\{\mathbf{X}_\infty \notin B\}}],
\end{aligned}$$

or rearranging terms,

$$\mathbb{E}[f_z(\mathbf{X}_\infty)\mathbb{I}_{\{\mathbf{X}_\infty \notin B\}}] \leq \frac{\beta}{\gamma}.$$

Thus

$$\begin{aligned}\mathbb{E}[f_z(\mathbf{X}_\infty)] &= \mathbb{E}[f_z(\mathbf{X}_\infty)\mathbb{I}_{\{\mathbf{X}_\infty \in B\}}] + \mathbb{E}[f_z(\mathbf{X}_\infty)\mathbb{I}_{\{\mathbf{X}_\infty \notin B\}}] \\ &\leq \alpha \mathbb{P}(\mathbf{X}_\infty \in B) + \frac{\beta}{\gamma} \\ &\leq \alpha + \frac{\beta}{\gamma}.\end{aligned}$$

Now by Fatou's lemma, as $f_z(\mathbf{X}_\infty) \to f(\mathbf{X}_\infty)$ almost surely as $z \to \infty$,

$$\mathbb{E}[f(\mathbf{X}_\infty)] = \mathbb{E}\left[\lim_{z\to\infty} f_z(\mathbf{X}_\infty)\right] \leq \liminf_{z\to\infty} \mathbb{E}[f_z(\mathbf{X}_\infty)] \leq \alpha + \frac{\beta}{\gamma},$$

giving the result. □

We need a property of sample paths of the $M(t)/M/m$ queue. For every initial queue length $q \in \mathbb{Z}_+$, we create a separate queue length process with the same arrival and service rates on a common probability space Ω. Let $f\colon \mathbb{Z}_+ \times \mathbb{R}_+ \times \Omega \to \mathbb{Z}_+$ map an initial queue length q, a time t, and a realization $\omega \in \Omega$ to the number of patients in the $M(t)/M/m$ system length at time t. The queues are coupled such that they share a single common Poisson process determining arrival times, and a single independent common Poisson process determining potential departure times (which only result in departures when there are patients in care). The relationship between the arrival process, the potential departure process, and the actual departures is the same as the relationship between $A(0,t)$, $\tilde{D}_{\text{on}}(0,t)$ and $D_{\text{on}}(0,t)$ from Appendix A.1.

LEMMA 3. *For every realization $\omega \in \Omega$, every time $t \in \mathbb{R}_+$, and all initial queue lengths $q, r \in \mathbb{Z}_+$ such that $q \geq r$, $f(\cdot, t, \omega)$ satisfies*

$$0 \leq f(q,t,\omega) - f(r,t,\omega) \leq q - r.$$

Namely, $f(\cdot, t, \omega)$ is monotone increasing and 1-Lipschitz continuous with respect to the ℓ_1 norm in the initial queue length.

PROOF. Fix ω, and consider $q, r \in \mathbb{Z}_+$, $q > r$. Let $\tau_0 = 0$ and for $i = 1, 2, \ldots$, let τ_i be the time of the ith event from our processes driving arrivals and departures. It suffices to prove that

$$0 \leq |f(q,\tau_i,\omega) - f(r,\tau_i,\omega)| \leq q - r, \tag{43}$$

for all τ_i, as the queue length can only change at the times of these events. Trivially the claim holds at τ_0. Suppose the claim holds until τ_i. At time τ_{i+1}:

1. Suppose the event was an arrival. Then $f(q, \tau_{i+1}, \omega) = 1 + f(q, \tau_i, \omega)$ and $f(r, \tau_{i+1}, \omega) = 1 + f(r, \tau_i, \omega)$, so

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - f(r, \tau_i, \omega),$$

   thus (43) holds.

2. Suppose the event was a potential departure. By our inductive hypothesis, we must be in one of the two cases below:

   (a) $f(q, \tau_i, \omega) = f(r, \tau_i, \omega)$. Then by our coupling, the system under initial condition q and under initial condition r must both either have an actual departure or have no departure at τ_{i+1}. As the change in queue lengths will be the same, by the same reasoning as when we have an arrival, we continue to satisfy (43).

   (b) $f(q, \tau_i, \omega) > f(r, \tau_i, \omega)$. Now either both systems have a departure, or only the system with initial queue length q has a departure (as it has more active servers), which with the inductive hypothesis implies

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) \leq f(q, \tau_i, \omega) - f(r, \tau_i, \omega) \leq q - r.$$

   When both systems experience an actual departure, $f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - f(r, \tau_i, \omega) \geq 0$ where the inequality is by the inductive hypothesis. When only the system under initial condition q has a departure, we still have

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - 1 - f(r, \tau_i, \omega) \geq 0,$$

   where the inequality holds by initial assumption for this case. Thus (43) holds.

□

Next we establish a few properties of Γ, defined in (3).

LEMMA 4. *For all $\kappa \geq 0$ and all $(q,r),(q',r') \in \mathbb{R}^2_+$, the operator $\Gamma(\cdot;\kappa)$ satisfies*

$$n\Gamma\left(\frac{(q,r)}{n};\kappa\right) = \Gamma(q,r;n\kappa), \tag{44}$$

$$\|\Gamma(q,r;\kappa) - \Gamma(q',r';\kappa)\|_1 \leq \|(q,r)-(q',r')\|_1. \tag{45}$$

Namely, with respect to the ℓ_1 norm, $\Gamma(\cdot;\kappa)$ is 1-Lipschitz continuous. Further, each component of $\Gamma(q,r;\kappa)$ increases monotonically in both q and r.

PROOF. The first property, follows immediately by definition of Γ, as

$$n\Gamma\left(\frac{(q,r)}{n};\kappa\right) = n\left(\frac{q}{n}\wedge\kappa, \left(\frac{q}{n}-\kappa\right)^+ - \frac{r}{n}\right) = \left(q\wedge n\kappa, (q-n\kappa)^+ - r\right) = \Gamma(q,r;n\kappa).$$

To show the second part, we find that

$$\begin{aligned}\|\Gamma(q,r;\kappa) - \Gamma(q',r';\kappa)\|_1 &= |(q-\kappa)^+ + r - (q'-\kappa)^+ - r'| + |(q\wedge\kappa) - (q'\wedge\kappa)| \\ &\leq |(q-\kappa)^+ - (q'-\kappa)^+| + |r-r'| + |(q\wedge\kappa) - (q'\wedge\kappa)|.\end{aligned}$$

Without loss of generality, suppose $q \geq q'$. Now by considering the exhaustive cases $q' \geq \kappa$, $q > \kappa > q'$, and $\kappa \geq q$, the Lipschitz continuity follows trivially. The monotonicity property is an immediate consequence of (3). □

COROLLARY 2. *For $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, fix any $\mathbf{s},\bar{\mathbf{s}} \in \mathcal{S}_\theta$. Assume that the processes $\mathbf{S}_\theta(t)$ has $\mathbf{S}_\theta(0) = \mathbf{s}$ and let $\tilde{\mathbf{S}}_\theta(t)$ be a version of $\mathbf{S}_\theta(t)$ with instead $\tilde{\mathbf{S}}_\theta(0) = \tilde{\mathbf{s}}$. There is a coupling between these processes such that for all $t \geq 0$,*

$$\|\mathbf{S}_\theta(t) - \tilde{\mathbf{S}}_\theta(t)\|_1 \leq \|\mathbf{s}-\tilde{\mathbf{s}}\|_1.$$

Moreover, $\mathbf{s} \geq \tilde{\mathbf{s}}$ componentwise implies $\mathbf{S}_\theta(t) \geq \tilde{\mathbf{S}}_\theta(t)$ componentwise for all t.

PROOF. It suffices to show the claim for all $t \in (0,1]$, as then the result follows by induction. For $\theta = \mathrm{LS}$, for all $t \in (0,1)$, $Q_{\mathrm{LS}}(t)$ and $\tilde{Q}_{\mathrm{LS}}(t)$ are $M(t)/M/c$ queues with the same arrival and service rates. Thus by Lemma 3, we obtain that $Q_{\mathrm{LS}}(t)$ is monotone increasing and 1-Lipschitz in q. Likewise, $R_{\mathrm{LS}}(t)$ and $\tilde{R}_{\mathrm{LS}}(t)$ are $M(t)/M/c$ queues with arrival rate zero and the same service rate, so again by Lemma 3, we obtain that $R_{\mathrm{LS}}(t)$ is monotone increasing and 1-Lipschitz in r. As

$Q_{\mathrm{LS}}(t)$ does not depend on r and $R_{\mathrm{LS}}(t)$ does not depend on q, the claim holds for $t \in (0,1)$. For $t = 1$,

$$\begin{aligned} \left\| \mathbf{S}_{\mathrm{LS}}(1) - \tilde{\mathbf{S}}_{\mathrm{LS}}(1) \right\|_1 &= \left\| \Gamma(\mathbf{S}_{\mathrm{LS}}(1^-); c) - \Gamma(\tilde{\mathbf{S}}_{\mathrm{LS}}(1^-); c) \right\|_1 \\ &\leq \left\| \mathbf{S}_{\mathrm{LS}}(1^-) - \tilde{\mathbf{S}}_{\mathrm{LS}}(1^-) \right\|_1 \qquad (46) \\ &\leq \left\| \mathbf{s} - \tilde{\mathbf{s}} \right\|_1, \qquad (47) \end{aligned}$$

where (46) follows from Lemma 4 and (47) follows from our analysis of the case $t \in (0,1)$. Monotonicity follows as each component of $\mathbf{S}_\theta(1)$ is the composition of monotone increasing functions and thus monotone increasing in every input, again by Lemma 4 and our analysis of the case $t \in (0,1)$. For $\theta = \mathrm{DA}$, the proof is very similar. □

We also need a simple uniform bound on a sequence of Poisson random variables.

LEMMA 5. *If $X_n \stackrel{d}{=} \mathrm{Pois}(\gamma n)$, then for all $k > 2\gamma(e-1)$ and all $n = 1, 2, \ldots,$*

$$\mathbb{P}(X_n \geq kn) \leq e^{-kn/2}.$$

PROOF. We have

$$\mathbb{P}(X_n \geq kn) \leq \exp(-kn)\mathbb{E}[\exp(X_n)] = \exp(-kn)\exp(\gamma n(e-1)) \leq \exp(-kn/2).$$

□

We now analyze our system in the fluid scaling. As before let $A^n \stackrel{\Delta}{=} A^n(0,1)$, $D^{\mathrm{on},n}_{\mathrm{LS}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{LS}}(0,1^-)$, $D^{\mathrm{on},n}_{\mathrm{DA}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{DA}}(0,\frac{1}{2}^-)$, $D^{\mathrm{off},n}_{\mathrm{LS}} \stackrel{\Delta}{=} D^{\mathrm{off},n}_{\mathrm{LS}}(0,1^-)$, and $D^{\mathrm{off},n}_{\mathrm{DA}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{DA}}(\frac{1}{2},1^-)$. Using Lemma 5, we show that when $\rho_\theta < 1$, the convergence as q goes to infinity in Lemma 1 is uniform in n.

LEMMA 6. *We have*

$$\lim_{k\to\infty} \sup_{n>0} \mathbb{E}_{(nk,nc)}\left[\frac{V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))}{n}\right] = -\gamma_{\mathrm{LS}}.$$

PROOF. From (31), we have that for all k and n,

$$\begin{aligned} \mathbb{E}_{(nk,nc)}\left[\frac{V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))}{n}\right] &= \mathbb{E}_{(nk,nc)}\left[\frac{A^n - D^{\mathrm{on},n}_{\mathrm{LS}} - D^{\mathrm{off},n}}{n}\right] \\ &= \lambda - c(1-e^{-\mu}) - \mathbb{E}_{(nk,nc)}\left[\frac{D^{\mathrm{on},n}_{\mathrm{LS}}}{n}\right]. \end{aligned}$$

where the second equality follows as $A^n \stackrel{d}{=} \text{Pois}(n\lambda)$ and $D_{\text{LS}}^{\text{off},n} \stackrel{d}{=} \text{Bin}(nc, 1-e^{-\mu})$ (see discussion (33)). As in (32) we have $\tilde{D}_{\text{LS}}^{\text{on},n} \stackrel{d}{=} \text{Pois}(nc\mu)$ coupled with $D_{\text{LS}}^{\text{on},n}$ such that $\tilde{D}_{\text{LS}}^{\text{on},n} \geq D_{\text{LS}}^{\text{on},n}$. Thus $\mathbb{E}[D_{\text{LS}}^{\text{on},n}] \leq nc\mu$ for all n and all k, so it suffices to show that

$$\liminf_{k\to\infty} \sup_{n>0} \mathbb{E}_{(nk,nc)} \left[\frac{D_{\text{LS}}^{\text{on},n}}{n} \right] \geq c\mu.$$

As in (34), we find that

$$\liminf_{k\to\infty} \sup_{n>0} \frac{1}{n} \mathbb{E}_{(nk,nc)}[D_{\text{LS}}^{\text{on},n}] \geq \lim_{k\to\infty} \sup_{n>0} \frac{1}{n} \mathbb{E}[\tilde{D}_{\text{LS}}^{\text{on},n} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on},n} \leq (k-c)n\}}].$$

Now

$$\begin{aligned}
\lim_{k\to\infty} \sup_{n>0} \left| \frac{1}{n} \mathbb{E}[\tilde{D}_{\text{LS}}^{\text{on},n} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on},n} \leq (k-c)n\}}] - c\mu \right| &= \lim_{k\to\infty} \sup_{n>0} \left| \frac{1}{n} \mathbb{E}[\tilde{D}_{\text{LS}}^{\text{on},n} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on},n} \leq (k-c)n\}} - \tilde{D}_{\text{LS}}^{\text{on},n}] \right| \\
&= \lim_{k\to\infty} \sup_{n>0} \frac{1}{n} \mathbb{E}[\tilde{D}_{\text{LS}}^{\text{on},n} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on},n} \geq (k-c)n\}}] \\
&\leq \lim_{k\to\infty} \sup_{n>0} \frac{1}{n} \sqrt{\mathbb{E}[(\tilde{D}_{\text{LS}}^{\text{on},n})^2] \mathbb{E}[\mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on},n} \geq (k-c)n\}}]} \\
&\leq \lim_{k\to\infty} \sup_{n>0} \frac{1}{n} \sqrt{((nc\mu)^2 + nc\mu) \exp(-kn/2) \exp(cn)} \qquad (48) \\
&= \lim_{k\to\infty} \sqrt{((c\mu)^2 + c\mu) \exp(-k/2) \exp(c)} \qquad (49) \\
&= 0.
\end{aligned}$$

Here (48) follows from Lemma 5, and (49) follows as when k, is large, the supremum is attained by taking $n = 1$. □

LEMMA 7. *We have*

$$\lim_{k\to\infty} \sup_{n>0} \mathbb{E}_{(nk,0)} \left[\frac{V(\mathbf{S}_{\text{DA}}^n(1)) - V(\mathbf{S}_{\text{DA}}^n(0))}{n} \right] = -\gamma_{\text{DA}}.$$

The proof is very similar to previous lemma and omitted. We can give the uniform moment bounds for $\mathbf{S}_\theta^n(\infty)$.

LEMMA 8. *For each policy $\theta \in \{\text{LS}, \text{DA}\}$, when $\rho_\theta < 1$, there exists a constant M_θ depending on λ, c and μ such that for every $n > 0$, $\mathbb{E}[V(\mathbf{S}_\theta^n(\infty))] \leq M_\theta n$.*

PROOF. As $\rho_\theta < 1$, we have $\gamma_\theta > 0$. By Lemma 6, there exists $\bar{k}_{\mathrm{LS}}$ such that for all $k > \bar{k}_{\mathrm{LS}}$ and all n,

$$\frac{1}{n}\mathbb{E}_{(nk,nc)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] \le -\frac{\gamma_{\mathrm{LS}}}{2}.$$

For any $\bar{q} > q$,

$$0 \le \mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] \le \bar{q} - q,$$

by the monotonicity and 1-Lipschitz of Corollary 2. Thus

$$\begin{aligned}
&\mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] \\
&\quad = \mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] + q - \bar{q} \\
&\quad \le 0.
\end{aligned}$$

As a result, we obtain that for every n, for all $q > \bar{k}_{\mathrm{LS}} n$,

$$\frac{1}{n}\mathbb{E}_{(q,nc)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] \le -\frac{\gamma_{\mathrm{LS}}}{2}. \tag{50}$$

We define $\bar{k}_{\mathrm{DA}}$ analogously using Lemma 7 and γ_{DA}. Let

$$b_\theta \triangleq \max\left\{\bar{k}_\theta, \frac{1}{\gamma_\theta}\left(\lambda^2 + \lambda + C_\theta(c^2 + c)\right)\right\},$$

where C_θ is as defined by (29) for LS and (36) for DA. Thus by Lemma 1 and Lemma 2, for all $\mathbf{s} \in \mathcal{S}^n_\theta$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}}[(V(\mathbf{S}^n_\theta(1)) - V(\mathbf{S}^n_\theta(0)))^2] &\le (\lambda^n)^2 + \lambda^n + C_\theta((c^n)^2 + c^n) \\
&= n^2\lambda^2 + n\lambda + C_\theta(n^2c^2 + nc) \\
&\le b_\theta\gamma_\theta n^2. \qquad (51)
\end{aligned}$$

We let

$$\begin{aligned}
B^n_\theta &\triangleq \{(q,r) \in \mathcal{S}_{\mathrm{LS}} \mid q + r \le 2nb_\theta\}, \qquad (52)\\
U(q,r) &\triangleq (q+r)^2, \\
\alpha^n_\theta &\triangleq 2nb_\theta, \\
\beta^n_\theta &\triangleq n^2 b_\theta(4\lambda + \gamma_\theta), \\
\gamma^n_\theta &\triangleq n\gamma_\theta/2.
\end{aligned}$$

and apply Proposition 4 for each n, using U as our Lyapunov function and $f(q,r) = q + r$. We observe that (41) holds trivially. For $(q,r) \in \mathcal{S}_\theta^n$ we have

$$\begin{aligned}
\mathbb{E}_{(q,r)}[U(\mathbf{S}_\theta^n(1)) - U(\mathbf{S}_\theta^n(0))] &= \mathbb{E}_{(q,r)}[(q + r + A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n})^2] - (q+r)^2 \\
&= 2(q+r)\mathbb{E}_{(q,r)}[A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n}] \\
&\quad + \mathbb{E}_{(q,r)}\left[(A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n})^2\right] \\
&= 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] \\
&\quad + \mathbb{E}_{(q,r)}[(V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0)))^2] \\
&\le 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta, \qquad (53)
\end{aligned}$$

where (53) is a consequence of (51). Now, for $(q,r) \in B_\theta^n$, using (52) with (53), we see that

$$\begin{aligned}
2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta &\le 4nb_\theta \mathbb{E}_{(q,r)}[A^n(0,1^-)] + n^2\gamma_\theta b_\theta \\
&\le n^2 b_\theta(4\lambda + \gamma_\theta) \\
&= \beta_\theta^n,
\end{aligned}$$

showing that (42) holds. Finally, for $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$,

$$\begin{aligned}
\mathbb{E}_{(q,r)}[U(\mathbf{S}_\theta^n(1)) - U(\mathbf{S}_\theta^n(0))] &\le 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta \qquad (54) \\
&\le -\gamma_\theta n(q+r) + n^2\gamma_\theta b_\theta \qquad (55) \\
&\le -\frac{\gamma_\theta}{2} n(q+r) - \frac{\gamma_\theta}{2} n(2nb_\theta) + n^2\gamma_\theta b_\theta \qquad (56) \\
&= -\frac{\gamma_\theta n}{2} f(q,r),
\end{aligned}$$

where (54) follows from (53), (55) follows as $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$ so we can apply (50), and (56) follows as $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$ ensures $q + r > 2nb_\theta$. This shows that (40) is satisfied for each n. Thus for every n we can apply Proposition 4 to obtain that

$$\mathbb{E}[f(\mathbf{S}^n(\infty))] \le \alpha^n + \frac{\beta^n}{\gamma^n} = n\left(2b_\theta\left(4\lambda/\gamma_\theta + 1\right)\right),$$

showing the result. □

A.3. Fluid model approximations. Proof of Theorem 2. In this section, we establish Theorem 2. First, we introduce the following additional notation to be used throughout the section.

Let u_θ^i be the time of the ith shift change under policy θ, i.e. for $i = 0, 1, 2, \ldots$, $u_{\text{LS}}^i \triangleq i$ and $u_{\text{DA}}^i \triangleq i/2$. Let c_θ^i be the number of residents on shift during $[u_i, u_{i+1})$, i.e. $c_{\text{LS}}^i \triangleq c$ for $i = 0, 1, 2, \ldots$, $c_{\text{DA}}^i \triangleq 2c$ for even i, and $c_{\text{DA}}^i \triangleq 0$ for odd i.

To prove the result, we will invoke a theorem from [30] that shows the convergence of multidimensional Markovian queueing processes to its fluid limit under the so called "uniform acceleration." Consider a Markov process $\mathbf{X}(t)$ on state space $\mathbb{Z}_+^m$ with transition rates that depend both on the current state and the time. The process $\mathbf{X}(t)$ is driven by a finite set of independent rate one exogenous Poison process $E_i(t)$, $i = 1, \ldots, k$. The events from these processes trigger a "jump" $\mathbf{v}_i \in \mathbb{Z}^m$ in $\mathbf{X}(t)$. For each process i, there is a rate function $\alpha_i(\mathbf{x}, t) \colon \mathbb{R}_+^m \times \mathbb{R}_+ \to \mathbb{R}_+$ that depends both on the state $\mathbf{x}$ and the time t. Assume that for each i and $\bar{t} \in \mathbb{R}_+$, $\alpha_i(\cdot, \bar{t})$ is γ_i-Lipschitz in $\mathbf{x}$ where γ_i does not depend on $\bar{t}$. We define $\mathbf{X}(t)$ by

$$\mathbf{X}(t) \triangleq \mathbf{X}(0) + \sum_{i=1}^{k} \mathbf{v}_i E_i \left(\int_0^t \alpha_i(\mathbf{X}(\tau), \tau) d\tau \right),$$

In Theorem 9.2 from [30], it is shown that this procedure uniquely defines the process $\mathbf{X}(t)$. Next, we consider a deterministic process $\mathbf{x}(t)$ on $\mathbb{R}_+^m$ defined by

$$\mathbf{x}(t) \triangleq \mathbf{x}(0) + \sum_{i=1}^{k} \mathbf{v}_i \int_0^t \alpha_i(\mathbf{x}(\tau), \tau) d\tau.$$

Again the existence and uniqueness of such an $\mathbf{x}(t)$ is shown in Theorem 11.4 from [30]. To approximate $\mathbf{X}(t)$ by $\mathbf{x}(t)$, we consider a sequence of processes $\mathbf{X}^n(t)$, $n = 1, 2, \ldots$, defined by

$$\mathbf{X}^n(t) \triangleq \mathbf{X}^n(0) + \sum_{i=1}^{k} \mathbf{v}_i E_i \left(n \int_0^t \alpha_i \left(\frac{\mathbf{X}^n(\tau)}{n}, \tau \right) d\tau \right),$$

i.e. $\mathbf{X}^1(t)$ is the original process. A special case of their result is as follows.

PROPOSITION 5 ([30], Theorem 2.2). *If $\mathbf{X}^n(0)/n \to \mathbf{x}(0)$ a.s., then*

$$\lim_{n \to \infty} \frac{\mathbf{X}^n(t)}{n} = \mathbf{x}(t),$$

a.s. and u.o.c.

We now return to our model. On the intervals $[u_\theta^i, u_\theta^{i+1})$ between shift changes, our processes $\mathbf{S}_\theta(t)$ and $\mathbf{s}_\theta(t)$ are of the form of $\mathbf{X}(t)$ and $\mathbf{x}(t)$ from the theorem. Specifically, we can take $\mathbf{v}_1 = (1, 0)$

and $\alpha_1((q,r),t) = \lambda(u_\theta^i + t)$ so that the arrival process $A(u_\theta^i, u_\theta^i + t) = E_1(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. Similarly, we take $\mathbf{v}_2 = (-1, 0)$ and

$$\alpha_2((q,r),t) \triangleq (c_\theta^i \wedge q)\mu = \begin{cases} (c \wedge q)\mu & \theta = \text{LS}, \\ (2c \wedge q)\mu & \theta = \text{DA}, \ i \text{ even}, \\ 0 & \theta = \text{DA}, \ i \text{ odd}, \end{cases}$$

then $D^{\text{on}}(u_\theta^i, t) = E_2(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. Finally, we take $\mathbf{v}_3 = (0, -1)$ and $\alpha_3((q,r),t) \triangleq r\mu$ so that $D^{\text{off}}(u_\theta^i, t) = E_3(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. We satisfy the Lipschitz condition on $\alpha_i(\cdot, t)$ as α_1 does not depend on the state and both α_2 and α_3 are μ-Lipschitz in (q,r) independent of t.

Thus the proposition immediately yields that if $\mathbf{S}_\theta^n(u_\theta^i)/n \to \mathbf{s}_\theta(u_\theta^i)$ a.s., then $\mathbf{S}_\theta^n(t)/n \to \mathbf{s}_\theta(t)$ u.o.c. From this point, the primary difficulty in proving Theorem 2 is showing that $\mathbf{S}_\theta(t)$ jumping at each shift change does not ruin the convergence. We can now prove the main result of the section.

Proof of Theorem 2.. For each policy $\theta \in \{\text{LS}, \text{DA}\}$, we will show by induction on i that $\mathbf{S}_\theta^n(t)/n \to \mathbf{s}_\theta(t)$ a.s. and uniformly on $[0, u_\theta^i]$. The case of $i = 0$ holds by the assumption of the theorem.

Suppose the claim holds for i. We notice that

$$\sup_{0 \le \tau \le u_\theta^{i+1}} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1 = \max \left\{ \sup_{0 \le \tau \le u_\theta^i} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1, \sup_{u_\theta^i \le \tau < u_\theta^{i+1}} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1, \left\| \frac{\mathbf{S}_\theta^n(u_\theta^{i+1})}{n} - \mathbf{s}_\theta(u_\theta^{i+1}) \right\|_1 \right\}.$$

By the definitions of u_θ^i and c_θ^i, it follows immediately that when applying Γ for the shift change at time u_{i+1}, we use $\Gamma(\cdot; c_\theta^i)$. Recalling that $\mathbf{S}_\theta^n$ and $\mathbf{s}_\theta$ are a.s. RCLL, and applying (44) and then (45) from Lemma 4,

$$\begin{aligned} \left\| \frac{\mathbf{S}_\theta^n(u_\theta^{i+1})}{n} - \mathbf{s}_\theta(u_\theta^{i+1}) \right\|_1 &= \left\| \frac{\Gamma(\mathbf{S}_\theta^n((u_\theta^{i+1})^-); nc_\theta^i)}{n} - \Gamma(\mathbf{s}_\theta((u_\theta^{i+1})^-); c_\theta^i) \right\|_1 \\ &= \left\| \Gamma\left(\frac{\mathbf{S}_\theta^n((u_\theta^{i+1})^-)}{n}; c_\theta^i \right) - \Gamma\left(\mathbf{s}_\theta((u_\theta^{i+1})^-); c_\theta^i \right) \right\|_1 \\ &\le \left\| \frac{\mathbf{S}_\theta^n((u_\theta^{i+1})^-)}{n} - \mathbf{s}_\theta((u_\theta^{i+1})^-) \right\|_1. \end{aligned}$$

For $\tau \in [u_\theta^i, u_\theta^{i+1}]$, we let $\bar{\mathbf{S}}_\theta^n(t)$ and $\bar{\mathbf{s}}_\theta(t)$ be the continuous extension of $\mathbf{S}_\theta^n(t)$ and $\mathbf{s}_\theta(t)$, respectively, from $[u_\theta^i, u_\theta^{i+1})$ to $[u_\theta^i, u_\theta^{i+1}]$, i.e.,

$$\bar{\mathbf{S}}_\theta^n(\tau) \triangleq \begin{cases} \mathbf{S}_\theta^n(\tau) & \tau < u_\theta^{i+1}, \\ \mathbf{S}_\theta^n((u_\theta^{i+1})^-) & \tau = u_\theta^{i+1}, \end{cases} \qquad \bar{\mathbf{s}}_\theta(\tau) \triangleq \begin{cases} \mathbf{s}_\theta(\tau) & \tau < u_\theta^{i+1}, \\ \mathbf{s}_\theta((u_\theta^{i+1})^-) & \tau = u_\theta^{i+1}. \end{cases}$$

Thus we obtain that

$$\sup_{0 \le \tau \le u_\theta^{i+1}} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1 = \max \left\{ \sup_{0 \le \tau \le u_\theta^i} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1, \sup_{u_\theta^i \le \tau \le u_\theta^{i+1}} \left\| \frac{\bar{\mathbf{S}}_\theta^n(\tau)}{n} - \bar{\mathbf{s}}_\theta(\tau) \right\|_1 \right\}.$$

By induction, the first term in the above maximum goes to zero a.s. and in particular $\mathbf{S}_\theta^n(u_\theta^i)/n \to \mathbf{s}_\theta(u_\theta^i)$ a.s. By Proposition 5, the second term converges to zero a.s. as well, as previously discussed.

□

REMARK 2. The result from [30] is actually stronger than what we suggested. It implies that a.s. and u.o.c., as $n \to \infty$, $\|\mathbf{S}_\theta^n(t)/n - \mathbf{s}_\theta(t)\|_1 \le O(\log n)$. This can be generalized to our case inductively in the same manner, but we do not pursue this further.

A.4. Long run behavior of the fluid model. In this section, for each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, we show that the fluid limit at integer times $\{\mathbf{s}_\theta(k)\}$, which is a deterministic discrete time dynamical system, has a simple long run behavior. To show this, we need to recall definitions from Section 2. For a set $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X} \neq \emptyset$, a function $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$, and an initial condition $\mathbf{x}_0 \in \mathcal{X}$, let $\mathbf{x}_k$, $k \in \mathbb{Z}_+$ be defined by $\mathbf{f}(\mathbf{x}_k) = \mathbf{x}_{k+1}$. Recall from Section 2 that a point $\mathbf{x}^* \in \mathcal{X}$ is *attractive* if for every $\mathbf{x}_0 \in \mathcal{X}$, $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}^*$. As previously mentioned, such an $\mathbf{x}^*$ must be unique. Further, when $\mathbf{f}$ is continuous on $\mathcal{X}$, it immediately follows that $\mathbf{f}(\mathbf{x}^*) = \mathbf{x}^*$, i.e. $\mathbf{x}^*$ is a *fixed point* of $\mathbf{f}$. First, we give a known (e.g. [6] page 183) criterion for identifying attractive points.

PROPOSITION 6. *Suppose $\mathcal{X}$ is nonempty and compact, $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$ is continuous on $\mathcal{X}$, and for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_p < \|\mathbf{x} - \mathbf{y}\|_p,$$

for some $p \ge 1$. Then there exists a unique attractive point $\mathbf{x}^* \in \mathcal{X}$.

We now give a sufficient condition to ensure that in finite time $\{\mathbf{x}_k\}$ will reach a bounded set, such as the compact set in the previous theorem.

PROPOSITION 7. *If there is a function $V\colon \mathcal{X} \to \mathbb{R}_+$, $\gamma > 0$ and $B \subset \mathcal{X}$ such that for all $\mathbf{x} \in \mathcal{X} \setminus B$,*

$$V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \leq -\gamma,$$

then for all $\mathbf{x}_0 \in \mathcal{X} \setminus B$, there exists $m \leq \lceil V(\mathbf{x}_0)/\gamma \rceil$ such that $\mathbf{x}_m \in B$.

PROOF. Let $n = \lceil V(\mathbf{x}_0)/\gamma \rceil + 1$ and assume for contradiction that $\mathbf{x}_0, \ldots, \mathbf{x}_n$ are all not in B. Then

$$V(\mathbf{x}_n) = V(\mathbf{x}_0) + \sum_{k=1}^{n} V(\mathbf{x}_k) - V(\mathbf{x}_{k-1}) \leq V(\mathbf{x}_0) - n\gamma < 0,$$

contradicting the non-negativity of V. □

Finally, we give a criteria for instability.

PROPOSITION 8. *Suppose $V\colon \mathcal{X} \to \mathbb{R}_+$ is a continuous function such that $\sup\{V(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\} = \infty$, and for all $\mathbf{x} \in \mathcal{X}$,*

$$V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \geq 0.$$

Then an attractive point does not exist.

PROOF. Assume for contradiction there were an attractive point $\mathbf{x}^*$. Let $\mathbf{x}_0 \in \mathcal{X}$ be such that $V(\mathbf{x}_0) > V(\mathbf{x}^*)$. Such a point exists as have assumed that the supremum of V is infinite. However, as $\mathbf{x}^*$ is attractive and V is continuous,

$$V(\mathbf{x}^*) = \lim_{n\to\infty} V(\mathbf{x}_n) = V(\mathbf{x}_0) + \lim_{n\to\infty} \sum_{k=1}^{n} V(\mathbf{x}_k) - V(\mathbf{x}_{k-1}) \geq V(\mathbf{x}_0),$$

contradicting $V(\mathbf{x}_0) > V(\mathbf{x}^*)$. □

We now consider the differential equation that controls the evolution of the state for the fluid model.

LEMMA 9. *Given parameters* $(\gamma, m, \mu, x(0)) \in \mathbb{R}^4_+$, *the differential equation*

$$\dot{x}(t) = \gamma - (x(t) \wedge m)\mu, \tag{57}$$

has a unique solution. The solution $x(t)$ *is monotone in* t *and satisfies* $x(t) \geq 0$ *for all* $t \geq 0$. *Further, if* $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ *is defined by* $g(x(0)) = x(\frac{1}{2})$, *then* g *is strictly increasing and 1-Lipschitz. Finally, if we let*

$$\tilde{x} \triangleq \begin{cases} m & \gamma \geq m\mu, \\ m - (\gamma - m\mu)/2 & \gamma < m\mu, \end{cases}$$

then $x(0) \geq \tilde{x}$ *implies that:*

(a) $x(t) \geq m$ *for* $t \in [0, \frac{1}{2})$,

(b) $x(\frac{1}{2}) = x(0) + (\gamma - m\mu)/2$,

(c) For all $y > x(0)$, $g(y) - g(x(0)) = y - x(0)$.

On the other hand, if $x(0) < \tilde{x}$, *then each of (a), (b) and (c) above are violated. In particular,*

(a') There exists $s \in [0, \frac{1}{2})$ *such that* $x(s) < m$,

(b') $x(\frac{1}{2}) > x(0) + (\gamma - m\mu)/2$,

(c') For all $y > x(0)$, $g(y) - g(x(0)) < y - x(0)$.

The proof is rather lengthy but no difficult, so it is deferred to Appendix A.7. We now use this to analyze the fluid limits of LS and DA. Let $g^1_{\mathrm{LS}}(q)$ and $g^2_{\mathrm{LS}}(q)$ be the function g from Lemma 9 when the parameters (λ_1, c, μ, q) and (λ_2, c, μ, q) are used, respectively. Let $\mathbf{f}_{\mathrm{LS}}\colon \mathcal{T}_{\mathrm{LS}} \to \mathcal{T}_{\mathrm{LS}}$ be given by

$$\mathbf{f}_{\mathrm{LS}}(q, r) \triangleq \Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), re^{-\mu}; c). \tag{58}$$

Similarly, let $g_{\mathrm{DA}}(q)$ be the function from Lemma 9 using parameters $(\lambda_1, 2c, \mu, q)$, and let

$$\mathbf{h}^1_{\mathrm{DA}}(q, r) \triangleq \Gamma(g_{\mathrm{DA}}(q), 0; 2c), \tag{59}$$

$$\mathbf{h}^2_{\mathrm{DA}}(q, r) \triangleq \Gamma\left(q + \frac{\lambda_2}{2}, re^{-\mu/2}; 0\right), \tag{60}$$

$$\mathbf{f}_{\mathrm{DA}}(\mathbf{s}) \triangleq \mathbf{h}^2(\mathbf{h}^1(\mathbf{s})). \tag{61}$$

We now prove Proposition 1.

PROOF OF PROPOSITION 1.. To prove existence and uniqueness of $\mathbf{s}_\theta(t)$, it suffices to show that for every $k \in \mathbb{Z}_+$, $\mathbf{s}_\theta(t)$ exists and is uniquely defined for all $0 \le t \le k$. For $k = 0$, we are given $\mathbf{s}_\theta(0)$ in the statement of the proposition. Suppose the claim holds for k. We now consider cases on θ.

LS – By induction $\mathbf{s}_{\mathrm{LS}}(k)$ is uniquely defined. As $q_{\mathrm{LS}}(t)$ solves the differential equation on $[k, k+\frac{1}{2})$ as used to define $g^1_{\mathrm{LS}}(q)$ with initial condition $q_{\mathrm{LS}}(k)$, and likewise solves the differential equation on $[k+\frac{1}{2}, 1)$ used to define $g^2_{\mathrm{LS}}(q)$ with initial condition $q_{\mathrm{LS}}(k+\frac{1}{2})$, we obtain by Lemma 9 that $q_{\mathrm{LS}}(t)$ is uniquely determined on $[k, k+1)$. As $\dot{r}_{\mathrm{LS}}(t) = -\mu r_{\mathrm{LS}}(t)$ and by induction $r_{\mathrm{LS}}(k)$ is uniquely defined, we obtain that for $t \in [k, k+1)$, $r(t) = r(k)e^{-\mu(t-k)}$, uniquely defining $\mathbf{s}_{\mathrm{LS}}(t)$ on that interval as well. Thus we immediately obtain that for every $\mathbf{s}_{\mathrm{LS}}(k) \in \mathcal{T}_{\mathrm{LS}}$,

$$\mathbf{s}_{\mathrm{LS}}(k+1) = \mathbf{f}_{\mathrm{LS}}(\mathbf{s}_{\mathrm{LS}}(k)),$$

showing the hypothesis.

DA – The argument is similar. Briefly, for all $\mathbf{s}_{\mathrm{DA}}(k) \in \mathcal{T}_{\mathrm{DA}}$,

$$\mathbf{s}_{\mathrm{DA}}(k+\tfrac{1}{2}) = \mathbf{h}^1(\mathbf{s}_{\mathrm{DA}}(k)),$$
$$\mathbf{s}_{\mathrm{DA}}(k+1) = \mathbf{h}^2(\mathbf{s}_{\mathrm{DA}}(k+\tfrac{1}{2})),$$
$$\mathbf{s}_{\mathrm{DA}}(k+1) = \mathbf{f}_{\mathrm{DA}}(\mathbf{s}_{\mathrm{DA}}(k)),$$

and at intermediate times in $t \in (k, k+\frac{1}{2})$ and $t \in (k+\frac{1}{2}, k+1)$, $\mathbf{s}_{\mathrm{DA}}(t)$ is the unique solution to a linear ODE either with constant coefficients or of the type from Lemma 9.

Finally, we must show that $\mathbf{s}_\theta(k) \in \mathcal{T}_\theta$, $k \ge 1$. For LS, we must show that $r_{\mathrm{LS}}(k) < c$ implies $q_{\mathrm{LS}}(k) \le c$. By definition

$$\mathbf{s}_{\mathrm{LS}}(k) = \Gamma(\mathbf{s}_{\mathrm{LS}}(k^-); c) = \begin{cases} (q_{\mathrm{LS}}(k^-) - c + r_{\mathrm{LS}}(k^-), c) & q_{\mathrm{LS}}(k^-) \ge c, \\ (r_{\mathrm{LS}}(k^-), q_{\mathrm{LS}}(k^-)) & q_{\mathrm{LS}}(k^-) < c. \end{cases}$$

Suppose $r_{\mathrm{LS}}(k) < c$. Note that $r_{\mathrm{LS}}(k) < c$ only in the second case. As $q_{\mathrm{LS}}(k) = r_{\mathrm{LS}}(k^-) \le r_{\mathrm{LS}}(k-1) \le c$, we obtain that $q_{\mathrm{LS}}(k) \le c$, showing the claim. For DA, must simply show that $r_{\mathrm{DA}}(k) = 0$, which holds as

$$r_{\mathrm{DA}}(k) = \Gamma(\mathbf{s}_{\mathrm{DA}}(k^-); 0) = (q(k^-) + r(k^-), 0).$$

□

Next we give some simple structural properties of the functions g^i_{LS} and $\mathbf{f}_\theta$ that will be needed in the analysis of the long run behavior of the fluid limits.

LEMMA 10. *There exists a unique $\tilde{q}_{\text{LS}}$ such that when $q_{\text{LS}}(0) \geq \tilde{q}_{\text{LS}}$,*

(a) $q_{\text{LS}}(t) \geq c$ *for* $t \in [0,1)$,

(b) $q_{\text{LS}}(1^-) = q_{\text{LS}}(0) + \lambda - c\mu$,

(c) For $\bar{q} > q_{\text{LS}}(0)$, $g^2_{\text{LS}}(g^1_{\text{LS}}(\bar{q})) - g^2_{\text{LS}}(g^1_{\text{LS}}(q_{\text{LS}}(0))) = \bar{q} - q_{\text{LS}}(0)$.

and when $q_{\text{LS}}(0) < \tilde{q}_{\text{LS}}$, (a), (b) and (c) are violated. In particular,

(a') *There exists* $s \in [0,1)$ *such that* $q_{\text{LS}}(s) < c$,

(b') $q_{\text{LS}}(1^-) > q_{\text{LS}}(0) + \lambda - c\mu$,

(c') *For* $\bar{q} > q_{\text{LS}}(0)$, $g^2_{\text{LS}}(g^1_{\text{LS}}(\bar{q})) - g^2_{\text{LS}}(g^1_{\text{LS}}(q_{\text{LS}}(0))) < \bar{q} - q_{\text{LS}}(0)$.

PROOF. For $i = 1, 2$, let $\tilde{q}^i_{\text{LS}}$ be $\tilde{x}$ from Lemma 9 when used to create g^i_{LS}. It is easy to see that properties (a), (b) and (c) will hold when for all $i = 1, 2$, properties (a), (b) and (c) from the application Lemma 9 to create g^i_{LS} hold, i.e. we have both $q_{\text{LS}}(0) \geq \tilde{q}^1_{\text{LS}}$ and $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$.

Similarly, it is easy to see that properties (a'), (b') and (c') will hold if there exists $i \in \{1, 2\}$ such that that properties (a'), (b') and (c') from the application of Lemma 9 to create g^i_{LS} hold, i.e. if either $q_{\text{LS}}(0) < \tilde{q}^1_{\text{LS}}$ or $q_{\text{LS}}(\frac{1}{2}) < \tilde{q}^2_{\text{LS}}$.

As $q_{\text{LS}}(\frac{1}{2}) = g^1_{\text{LS}}(q_{\text{LS}}(0))$ and by Lemma 4 g^1_{LS} is strictly increasing, there is a threshold q^* such that $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$ iff $q_{\text{LS}}(0) \geq q^*$. Thus by taking $\tilde{q}_{\text{LS}} = \max\{q^*, q^1_{\text{LS}}\}$, we will have both $q_{\text{LS}}(0) \geq \tilde{q}^1_{\text{LS}}$ and $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$ iff $q_{\text{LS}}(0) \geq \tilde{q}_{\text{LS}}$. This gives the result. □

LEMMA 11. *For $\theta \in \{\text{LS}, \text{DA}\}$, $\mathbf{f}_\theta$ is 1-Lipschitz with respect to the ℓ_1 norm and both outputs of the function $\mathbf{f}_\theta$ are monotonically increasing in both inputs.*

PROOF. For LS, the functions g^1_{LS} and g^2_{LS} are monotonically increasing and 1-Lipschitz by Lemma 9. Similarly $re^{-\mu}$ as a function of r is increasing and 1-Lipschitz, and by Lemma 4, $\Gamma(\cdot; c)$ is 1-Lipschitz and each component is monotonically increasing in every input. Thus $\mathbf{f}_{\text{LS}}(q, r)$ is

a composition of monotone increasing 1-Lipschitz functions and thus monotone increasing and 1-Lipschitz.

The argument is similar for DA. The functions g_{DA}, $q + \lambda_2/2$ as a function of q, $re^{-\mu/2}$ as a function of r, $\Gamma(\cdot; 2c)$ and $\Gamma(\cdot; 0)$ are all 1-Lipschitz and monotonically increasing in every input (by Lemma 9 for g_{DA} and by Lemma 4 for $\Gamma(\cdot, 2c)$ and $\Gamma(\cdot, 0)$). Therefore $\mathbf{f}_{\mathrm{DA}}(q, r)$ is a composition of 1-Lipschitz monotone increasing functions and thus 1-Lipschitz and monotone increasing. □

We can now analyze the long run behavior of the fluid limits. First we will show that when $\rho_\theta < 1$, $\mathbf{f}_\theta$ restricted to some $T_\theta \subset \mathcal{T}_\theta$ has an attractive fixed point using contractive mapping (Proposition 6). Then we will use a Lyapunov function argument to show that the attractive point over T_θ is in fact attractive over all of $\mathcal{T}_\theta$ (Proposition 7). Finally, we will use another Lyapunov function argument to show that $\mathbf{f}_\theta$ has no attractive points when $\rho_\theta \geq 1$ (Proposition 8).

PROPOSITION 9. *The process $\{\mathbf{s}_\theta(k)\}$ has a unique attractive fixed point iff $\rho_\theta < 1$.*

PROOF. Assume $\rho_{\mathrm{LS}} < 1$. Recall $\tilde{q}_{\mathrm{LS}}$ from Lemma 10, and let $\tilde{q}_{\mathrm{DA}}$ be $\tilde{x}$ from Lemma 9 as applied to create g_{DA}. Let

$$T_\theta \triangleq \mathcal{T}_\theta \setminus \{(q, r) \mid q > \tilde{q}_\theta\}. \tag{62}$$

We now check the assumptions of Proposition 6 are satisfied by $\mathbf{f}_\theta$ restricted to T_θ. We can immediately verify by definition that $\mathbf{f}_\theta$ is the composition of continuous functions and thus continuous (recall that g^1_{LS}, g^2_{LS}, and g_{DA} are continuous by Lemma 9 and $\Gamma(\cdot, \kappa)$ is continuous for all κ by Lemma 4). Noting that $\tilde{q}_{\mathrm{LS}} \geq c$ by part (a) of Lemma 10, we see that $T_{\mathrm{LS}} = [0, c] \times [0, c] \cup \{(q, c) \mid 0 \leq q \leq \tilde{q}_{\mathrm{LS}}\}$ and thus it is a nonempty compact set. Likewise $T_{\mathrm{DA}} = \{(q, 0) \mid 0 \leq q \leq \tilde{q}_{\mathrm{DA}}\}$ where by Lemma 9 we see that $\tilde{q}_{\mathrm{DA}} \geq 2c$, thus T_{DA} is a nonempty compact set as well. We still need to check that $\mathbf{f}_\theta \colon T_\theta \to T_\theta$ and that $\mathbf{f}_\theta$ is contractive on T_θ.

To show $\mathbf{f}_{\mathrm{LS}} \colon T_{\mathrm{LS}} \to T_{\mathrm{LS}}$, we have that for any $(q, r) \in T_{\mathrm{LS}}$,

$$\mathbf{f}_{\mathrm{LS}}(q, r) \leq \mathbf{f}_{\mathrm{LS}}(\tilde{q}_{\mathrm{LS}}, c) \tag{63}$$

$$= \Gamma\left(\tilde{q}_{\mathrm{LS}} + \lambda - c\mu, ce^{-\mu}; c\right) \tag{64}$$

$$= \left(\tilde{q}_{\mathrm{LS}} + \lambda - c\mu - c + ce^{-\mu}, c\right) \tag{65}$$

$$\leq (\tilde{q}_{\mathrm{LS}}, c), \tag{66}$$

where the inequalities are componentwise. Here (63) holds by Lemma 11. We obtain (64) by Lemma 10 part (b). Then (65) holds as we have $\tilde{q}_{\mathrm{LS}}+\lambda-c\mu = q_{\mathrm{LS}}(1^-) \geq c$ by part (a) of Lemma 10. Finally (66) holds as $\rho_{\mathrm{LS}} < 1$ implies $\lambda - c\mu - c + ce^{-\mu} = -\gamma_{\mathrm{LS}} < 0$. Thus we obtain that $\mathbf{f}_{\mathrm{LS}} \colon T_{\mathrm{LS}} \to T_{\mathrm{LS}}$.

To show $\mathbf{f}_{\mathrm{DA}} \colon T_{\mathrm{DA}} \to T_{\mathrm{DA}}$, we first compute that

$$\mathbf{h}_1(\tilde{q}_{\mathrm{DA}}, 0) = \Gamma\left(\tilde{q}_{\mathrm{DA}} + \frac{\lambda_1}{2} - c\mu, 0; 2c\right) \tag{67}$$

$$= \left(\tilde{q}_{\mathrm{DA}} + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right). \tag{68}$$

Here (67) follows from part (b) of Lemma 9 on g_{DA}, and (68) follows from part (a) of the lemma. Thus for all $(q, 0) \in T_{\mathrm{DA}}$,

$$\begin{aligned} \mathbf{f}_{\mathrm{DA}}(q,0) &\leq \mathbf{f}_{\mathrm{DA}}(\tilde{q}_{\mathrm{DA}}, 0) \\ &= \mathbf{h}_2\left(\tilde{q} + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right) \\ &= \Gamma\left(\tilde{q} + \frac{\lambda_1}{2} - c\mu - 2c + \frac{\lambda_2}{2}, 2ce^{-\mu/2}; 0\right) \\ &= \left(\tilde{q} + \lambda - c\mu - 2c + 2ce^{-\mu/2}, 0\right) \end{aligned} \tag{69}$$

$$\leq (\tilde{q}_{\mathrm{DA}}, 0). \tag{70}$$

where (69) holds by the monotonicity of $\mathbf{f}_{\mathrm{DA}}$ and (70) holds as $\rho_{\mathrm{DA}} < 1$. Again the inequalities are componentwise. Thus we obtain that $\mathbf{f}_{\mathrm{DA}} \colon T_{\mathrm{DA}} \to T_{\mathrm{DA}}$.

We show that $\mathbf{f}_{\mathrm{LS}}$ is contractive on T_{LS} with respect to the $\|\cdot\|_1$ norm. Consider $(q, r), (q', r') \in T_{\mathrm{LS}}$, such that $(q, r) \neq (q', r')$. If $q \neq q'$, then by part (c') of Lemma 10

$$|g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) - g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q'))| < |q - q'|. \tag{71}$$

Similarly, when $r \neq r'$, then

$$|re^{-\mu} - r'e^{-\mu}| < |r - r'|. \tag{72}$$

Thus we obtain that

$$\begin{aligned} \|\mathbf{f}_{\mathrm{LS}}(q,r) - \mathbf{f}_{\mathrm{LS}}(q',r')\|_1 &= \|\Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), re^{-\mu}; c) - \Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q')), r'e^{-\mu}; c)\|_1 \\ &\leq |g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) - g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q'))| + |re^{-\mu} - r'e^{-\mu}| \end{aligned} \tag{73}$$

$$\begin{aligned} &< |q - q'| + |r - r'| \\ &= \|(q,r) - (q',r')\|_1. \end{aligned} \tag{74}$$

Here (73) holds by Lemma 4 and (74) follows from (71) if $q \neq q'$ and from (72) if $r \neq r'$.

Showing that $\mathbf{f}_{\mathrm{DA}}$ is contractive on T_{DA} with respect to the $\|\cdot\|_1$ norm is very similar to the LS case. Briefly, we observe that when $q < \tilde{q}_{\mathrm{DA}}$, that $\mathbf{h}^1_{\mathrm{DA}}$ is strictly contractive by (c') of Lemma 9. It is easy to see that $\mathbf{h}^2_{\mathrm{DA}}$ is non-expansive for all $(q, r) \in \mathbb{R}^2_+$. Thus $\mathbf{f}_{\mathrm{DA}}$ on T_{DA} is the composition of a contractive function and a non-expansive function and hence contractive.

Thus the assumptions of Proposition 6 are satisfied by $\mathbf{f}_\theta$ on T_θ when $\rho_\theta < 1$. This implies that once $\mathbf{s}_\theta(k)$ enters T_θ it will converge to a unique fixed point. For the case $\rho_\theta < 1$, it remains to show that for $\mathbf{s}_\theta(0) \notin T_\theta$, we reach T_θ in finite time.

To this end, we apply Proposition 7 using the Lyapunov function $V(q, r) \stackrel{\Delta}{=} q + r$, the set of exceptions as T_θ, and γ to be γ_θ (we have $\gamma_\theta > 0$ as $\rho_\theta < 1$). We now show that the drift condition is satisfied. For LS, $(q, r) \notin T_{\mathrm{LS}}$ implies $q \geq \tilde{q}_{\mathrm{LS}} \geq c$, thus we must have $r = c$. Thus,

$$
\begin{aligned}
V(\mathbf{f}_{\mathrm{LS}}(q, c)) - V(q, c) &= V(\Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), ce^{-\mu}; c)) - (q + c) \\
&= g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) + ce^{-\mu} - (q + c) \qquad (75) \\
&= q + \lambda - c\mu + ce^{-\mu} - (q + c) \qquad (76) \\
&= -\gamma_{\mathrm{LS}}.
\end{aligned}
$$

Here (75) follows from the same argument justifying (30), and (76) follows from part (b) of Lemma 10. For DA, $(q, 0) \notin T_{\mathrm{DA}}$ implies $q \geq \tilde{q}_{\mathrm{DA}}$. We obtain

$$
\mathbf{h}_1(q, 0) = \left(q + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right),
$$

just as we justified (67) and (68). Thus for such q,

$$
\begin{aligned}
V(\mathbf{f}_{\mathrm{DA}}(q, 0)) - V(q, 0) &= V\left(\mathbf{h}_2\left(q + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right)\right) - q \\
&= V(\Gamma(q + \lambda - c\mu - 2c, 2ce^{-\mu/2}; 0)) - q \\
&= -\gamma_{\mathrm{DA}}.
\end{aligned}
$$

Thus the assumptions of Proposition 7 are satisfied, establishing the claim in the case when $\rho_\theta < 1$.

Finally, we show that $\{\mathbf{s}_\theta(k)\}$ has no attractive point when $\rho_\theta \geq 1$ by applying Proposition 8. We again take $V(q, r) = q + r$, and immediately verify that it is continuous and unbounded on $\mathcal{T}_\theta$.

For LS, we compute that for any $\mathbf{s}_{\mathrm{LS}}(0)$,

$$V(\mathbf{s}_{\mathrm{LS}}(1)) - V(\mathbf{s}_{\mathrm{LS}}(0)) = q_{\mathrm{LS}}(1^-) - q_{\mathrm{LS}}(0) + r_{\mathrm{LS}}(1^-) - r_{\mathrm{LS}}(0) \tag{77}$$

$$= \int_0^1 \lambda(t) - \mu(q_{\mathrm{LS}}(t) \wedge c)dt - r_{\mathrm{DA}}(0)(1 - e^{-\mu}) \tag{78}$$

$$\geq \lambda - c\mu - c(1 - e^{-\mu}) \tag{79}$$

$$= -\gamma_{\mathrm{LS}}$$

$$\geq 0, \tag{80}$$

where (77) follows similarly to (30), (78) follows from the definition of $\dot{q}_{\mathrm{LS}}(t)$, (79) follows as $r_{\mathrm{LS}}(0) \leq c$ and $q_{\mathrm{LS}}(t) \wedge c \leq c$, and finally (80) follows as $\rho_{\mathrm{LS}} \leq 1$. For DA, we compute that for any $\mathbf{s}_{\mathrm{DA}}(0)$ that

$$V(\mathbf{s}_{\mathrm{DA}}(1)) - V(\mathbf{s}_{\mathrm{DA}}(0)) = q_{\mathrm{DA}}(1^-) - q_{\mathrm{DA}}(0) + r_{\mathrm{DA}}(1^-) \tag{81}$$

$$= \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) + r_{\mathrm{DA}}(\tfrac{1}{2})e^{-\mu/2}$$

$$= \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) + r_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) - r_{\mathrm{DA}}(\tfrac{1}{2})(1 - e^{-\mu/2})$$

$$\geq \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) + r_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) - 2c(1 - e^{-\mu/2}) \tag{82}$$

$$= \frac{\lambda_2}{2} + q_{\mathrm{DA}}\left(\tfrac{1}{2}^-\right) - q_{\mathrm{DA}}(0) - 2c(1 - e^{-\mu/2}) \tag{83}$$

$$= \frac{\lambda_2}{2} + \int_0^{\frac{1}{2}} \lambda_1 - \mu(q_{\mathrm{DA}}(t) \wedge 2c)dt - 2c(1 - e^{-\mu/2}) \tag{84}$$

$$\geq \lambda - \int_0^{\frac{1}{2}} 2c\mu\, dt - 2c(1 - e^{-\mu/2})$$

$$= -\gamma_{\mathrm{DA}}$$

$$\geq 0, \tag{85}$$

where (81) follows similarly to (30), (82) follows as $r_{\mathrm{DA}}(\frac{1}{2}) \leq 2c$, (83) follows as by Γ, $q_{\mathrm{DA}}(\frac{1}{2}^-) + r_{\mathrm{DA}}(\frac{1}{2}^{-1}) = q_{\mathrm{DA}}(\frac{1}{2})$, (84) follows from the definition of $\dot{q}_{\mathrm{DA}}(t)$, and finally (85) holds as $\rho_{\mathrm{DA}} \geq 1$. Thus we see by Proposition 8 that $\rho_\theta \geq 1$ implies that $\{\mathbf{s}_\theta(k)\}$ has no attractive point, completing the proof of Proposition 9. □

Finally, we give a result providing a uniform bound on the distance moved towards the fixed point in each iteration of the fluid model. This result will be useful in the proof of Theorem 3. Let

$V_\theta \colon \mathcal{T}_\theta \to \mathbb{R}_+$ be given by

$$V_\theta(\mathbf{s}) \triangleq \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1. \tag{86}$$

COROLLARY 3. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, when $\rho_\theta < 1$, for every $z > 0$, there exists $\gamma > 0$ such that*

$$\inf_{\mathbf{s} \in \mathcal{T}_\theta \setminus B_z(\mathbf{s}_\theta(\infty))} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq -\gamma.$$

PROOF. Recall from in the proof of Proposition 9 that for each θ, we defined sets T_θ such that that $\mathbf{f}_\theta$ is contractive on T_θ and for all $\mathbf{s} \in \mathcal{T}_\theta \setminus T_\theta$,

$$V(\mathbf{f}_\theta(\mathbf{s})) - V(\mathbf{s}) = -\gamma_\theta < 0.$$

Suppose $\mathbf{s} = (q, r) \notin T_\theta$. Trivially $\mathbf{s} \geq \mathbf{s}_\theta(\infty)$ componentwise as $\mathbf{s}_\theta(\infty) \in T_\theta$. By Lemma 11, as $\mathbf{s} \geq \mathbf{s}_\theta(\infty)$ componentwise, we have $\mathbf{f}_\theta(\mathbf{s}) \geq \mathbf{f}_\theta(\mathbf{s}_\theta(\infty)) = \mathbf{s}_\theta(\infty)$ componentwise as well. Thus for $\mathbf{s} \notin T_\theta$, letting $(q', r') = \mathbf{f}_\theta(\mathbf{s})$,

$$\begin{aligned} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{s}_\theta(\infty)\|_1 - \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1 \\ &= q' - q_\theta(\infty) + r' - r_\theta(\infty) - (q - q_\theta(\infty) + r - r_\theta(\infty)) \\ &= V(\mathbf{f}_\theta(\mathbf{s})) - V(\mathbf{s}) \\ &\leq -\gamma_\theta. \end{aligned}$$

For all $\mathbf{s} \in T_\theta$, $\mathbf{s} \neq \mathbf{s}_\theta(\infty)$, as $\mathbf{f}_\theta$ is contractive on T_θ, we have

$$\begin{aligned} V_\theta(\mathbf{f}_\theta(\mathbf{s})) &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{s}_\theta(\infty)\|_1 \\ &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{f}_\theta(\mathbf{s}_\theta(\infty))\|_1 \\ &< \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1 \\ &= V_\theta(\mathbf{s}). \end{aligned}$$

Thus for $\mathbf{s} \in T_\theta$, $V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq 0$, holding with equality only when $\mathbf{s} = \mathbf{s}_\theta(\infty)$. Fix z from the statement of the Lemma. As $V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s})$ is a composition of continuous functions and thus continuous and as T_θ is compact (as shown in the proof of Proposition 9), we have that given our z there exists $\varepsilon > 0$ such that

$$\inf_{\mathbf{s} \in T_\theta \setminus B_z(\mathbf{s}_\theta(\infty)} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq -\varepsilon.$$

Thus by taking $\gamma = \min\{\varepsilon, \gamma_\theta\}$, we obtain the result. □

A.5. Interchange of Limits. Proof of Theorem 3. In this section, we prove Theorem 3, showing that the rescaled steady state distributions $\mathbf{S}_\theta^n(\infty)/n$ converge in probability to the fixed point $\mathbf{s}_\theta(\infty)$ of the fluid limit at integer times.

Recall that a set of random vectors $\{\mathbf{X}_n\}$ is defined to be *tight* if for every ε there exists k such that for every n, $\mathbb{P}(\|\mathbf{X}_n\|_1 > k) \leq \varepsilon$. As a direct consequence of Lemma 8 and Markov's inequality, we obtain:

COROLLARY 4. *For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, when $\rho_\theta < 1$, the set of random vectors $\{\mathbf{S}_\theta^n(\infty)/n\}$ is tight.*

By Prokhorov's theorem, this implies that $\{\mathbf{X}_n\}$ is relatively compact. That is, for every subsequence $\mathbf{X}_{n_i}$ there exists a random vector $\mathbf{X}$ and a subsubsequence $\mathbf{X}_{n_{i_j}}$ such that $\mathbf{X}_{n_{i_j}} \Rightarrow \mathbf{X}$ (see [5]). Thus for every subsequence n_i there is a subsubsequence n_{i_j} and a random vector $\bar{\mathbf{S}}_\theta$ such that as $j \to \infty$

$$\frac{\mathbf{S}_\theta^{n_{i_j}}(\infty)}{n_{i_j}} \Rightarrow \bar{\mathbf{S}}_\theta.$$

Thus to show Theorem 3, it is sufficient to show that for every sequence n_i, the resulting $\bar{\mathbf{S}}_\theta$ equals $\mathbf{s}_\theta(\infty)$ with probability one, as convergence in distribution to a constant implies convergence in probability.

PROOF OF THEOREM 3.. First, we claim that for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$,

$$\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta) \stackrel{d}{=} \bar{\mathbf{S}}_\theta, \tag{87}$$

where $\mathbf{f}_{\mathrm{LS}}$ and $\mathbf{f}_{\mathrm{DA}}$ are defined by (58) and (61), respectively. By Proposition 11.3.2 from [11], we can equivalently check that the Lévy-Prokhorov distance between these variables is zero, i.e. that for all $g\colon \mathcal{T}_\theta \to \mathbb{R}$ such that $\|g\|_{\mathrm{BL}} \leq 1$, we have

$$\mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g(\bar{\mathbf{S}}_\theta)] = 0.$$

Here, we use the three term estimate as devised by [13], Chapter 4, Theorem 9.10, in a similar continuous time interchange of limits argument. See [36] for similar but less terse argument. Assume

for every n that $\mathbf{S}_\theta^n(0) \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)$. Now for every n, we have

$$\begin{aligned}\left|\mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g(\bar{\mathbf{S}}_\theta)]\right| \leq & \left|\mathbb{E}\left[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right)\right]\right| \\ & + \left|\mathbb{E}\left[g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right) - g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right]\right| \\ & + \left|\mathbb{E}\left[g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right) - g\left(\bar{\mathbf{S}}_\theta\right)\right]\right|.\end{aligned}$$

As $\mathbf{f}_\theta$ and g are continuous and g is bounded, $g \circ \mathbf{f}_\theta$ is a bounded continuous function. For the first term, $\mathbf{S}_\theta^n(0)/n \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)/n \Rightarrow \bar{\mathbf{S}}_\theta$, so we can apply the Continuous Mapping Theorem and then the Bounded Convergence Theorem to obtain that $\mathbb{E}[g\left(\mathbf{f}_\theta\left(\mathbf{S}_\theta^n(0)/n\right)\right)] \to \mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta))]$ as $n \to \infty$ along n_{i_j}. By stationarity $\mathbf{S}_\theta^n(1)/n \stackrel{d}{=} \mathbf{S}_\theta^n(0)/n \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)/n$, implying that the third term converges to zero along n_{i_j} by a similar argument.

Finally we bound the second term. Let $h_\theta^n \colon \mathcal{T}_\theta \to \mathbb{R}$ and $h_\theta \colon \mathcal{T}_\theta \to \mathbb{R}$ be given by

$$\begin{aligned}h_\theta^n(\mathbf{s}) &\stackrel{\Delta}{=} \mathbb{E}_{\lfloor n\mathbf{s} \rfloor}\left[g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right], \\ h_\theta(\mathbf{s}) &\stackrel{\Delta}{=} g(\mathbf{f}_\theta(\mathbf{s})),\end{aligned}$$

so that

$$\left|\mathbb{E}\left[g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right) - g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right]\right| = \left|\mathbb{E}\left[h_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right) - h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right]\right|. \tag{88}$$

We need some properties of h_θ^n and h to make an estimate. First, we claim that for all $\mathbf{s}$, $h_\theta^n(\mathbf{s}) \to h_\theta(\mathbf{s})$ as $n \to \infty$. As a consequence of Theorem 2, we have that for every $\mathbf{s} \in \mathcal{T}_\theta$, if $\mathbf{S}_\theta^n(0) = \lfloor n\mathbf{s} \rfloor$ so that $\mathbf{S}_\theta^n(0)/n \to \mathbf{s}$ a.s., then $\mathbf{S}_\theta^n(1)/n \to \mathbf{f}_\theta(\mathbf{s})$ a.s. as well. By the continuity of g, it follows from that $g(\mathbf{S}_\theta^n(1)/n) \to g(\mathbf{f}_\theta(\mathbf{s}))$ a.s. as well. Noting that $g(\mathbf{S}_\theta^n(1)/n)$ is bounded, we can apply the Bounded Convergence Theorem to obtain that for $\mathbf{s} \in \mathcal{T}_\theta$,

$$\lim_{n\to\infty} h_\theta^n(\mathbf{s}) = h_\theta(\mathbf{s}). \tag{89}$$

We can now bound (88) with a coupling argument. By the Skorokhod Representation Theorem, let Ω be a common probability space for $\{\mathbf{S}_\theta^n(0)\}$ and $\bar{\mathbf{S}}_\theta$ such that for $\omega \in \Omega$, $\mathbf{S}_\theta^n(0, \omega) \to \bar{\mathbf{S}}_\theta(\omega)$ a.s.

Now we have

$$\begin{aligned}\left|h_\theta\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)-h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right| \leq \quad & \left|h_\theta\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)-h_\theta\left(\bar{\mathbf{S}}_\theta(\omega)\right)\right| \\ & + \left|h_\theta\left(\bar{\mathbf{S}}_\theta(\omega)\right)-h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right)\right| \\ & + \left|h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right)-h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right|.\end{aligned}$$

We claim each of these terms converges to zero a.s. The first term converges to zero as h_θ is a continuous function and $\mathbf{S}_\theta^n(0,\omega)/n \to \bar{\mathbf{S}}_\theta(\omega)$ a.s. The second term converges to zero by (89). Let $\tilde{\mathbf{S}}_\theta^n(t)$ be another version of the process $\mathbf{S}_\theta^n(t)$ with the initial condition $\tilde{\mathbf{S}}_\theta^n(0) = \lfloor n\bar{\mathbf{S}}_\theta \rfloor$ that is coupled to $\mathbf{S}_\theta^n(t)$ as in Corollary 2. Then

$$\begin{aligned}\left|h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right)-h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right| &= \left|\mathbb{E}\left[g\left(\frac{\tilde{\mathbf{S}}_\theta^n(1)}{n}\right)-g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\;\middle|\;\tilde{\mathbf{S}}_\theta^n(0)=\lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor,\,\mathbf{S}_\theta^n(0)=\mathbf{S}_\theta^n(0,\omega)\right]\right| \\ &\leq \frac{\|g\|_{\mathrm{BL}}}{n}\mathbb{E}\left[\left\|\tilde{\mathbf{S}}_\theta^n(1)-\mathbf{S}_\theta^n(1)\right\|_1\;\middle|\;\tilde{\mathbf{S}}_\theta^n(0)=\lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor,\,\mathbf{S}_\theta^n(0)=\mathbf{S}_\theta^n(0,\omega)\right] \\ &\leq \frac{\|g\|_{\mathrm{BL}}}{n}\left\|\lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor-\mathbf{S}_\theta^n(0,\omega)\right\|_1 \qquad (90) \\ &\leq \|g\|_{\mathrm{BL}}\left(\left\|\bar{\mathbf{S}}_\theta(\omega)-\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right\|_1+\frac{1}{n}\right),\end{aligned}$$

showing that the third term converges to zero a.s. as well. Here (90) follows from Corollary 2. Finally, as h_θ^n and h_θ are bounded by one, the Bounded Convergence Theorem implies that the right hand side of (88) converges to zero. Thus we have shown (87).

Now we show that $\bar{\mathbf{S}}_\theta = \mathbf{s}_\theta(\infty)$ a.s. We assume the conclusion is false to show a contradiction. By assumption, there exists some $\tilde{\mathbf{s}}_\theta \neq \mathbf{s}_\theta(\infty)$ and $\varepsilon > 0$ such that $\mathbf{s}_\theta(\infty) \notin B_\varepsilon(\tilde{\mathbf{s}}_\theta)$ and

$$\mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) > 0.$$

Let N be such that

$$\mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N) < \mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)). \qquad (91)$$

Let $z = \|\tilde{\mathbf{s}} - \mathbf{s}_\theta(\infty)\|_1 - \varepsilon$ and let $Z = B_z(\mathbf{s}_\theta(\infty))$ be the largest ball around $\mathbf{s}_\theta(\infty)$ disjoint from $B_\varepsilon(\tilde{\mathbf{s}}_\theta)$. We now use V_θ from (86), and let

$$-d \triangleq \inf_{\mathbf{s}\notin Z} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}).$$

Note that $d > 0$ by Corollary 3. Let $n \in \mathbb{Z}_+$ be such that

$$nd > \sup_{\|\mathbf{s}\|_1 < N} V_\theta(\mathbf{s}),$$

(the supremum is bounded as $\{\mathbf{s} \in \mathcal{T}_\theta | \|\mathbf{s}\|_1 < N\}$ is compact and V_θ is continuous). Let $\mathbf{f}_\theta^{(m)}$ be the function $\mathbf{f}_\theta$ composed with itself m times. We now claim that for all $\mathbf{s} \in \mathcal{T}_\theta$,

$$\mathbf{f}^{(n)}(\mathbf{s}) \notin Z \text{ implies that } \|\mathbf{s}\|_1 > N. \tag{92}$$

We show the contrapositive using Proposition 7. We take our bounded set of exceptions as Z, use the Lyapunov function V_θ, and drift $-d$. The drift condition is satisfied as $d > 0$. Thus $\|\mathbf{s}\|_1 < N$ implies that there exists m with $0 \le m \le n$ such that $\mathbf{f}_\theta^{(m)}(\mathbf{s}) \in Z$. Notice that Corollary 3 implies that $\mathbf{f}_\theta(Z) \subset Z$. Thus $\|\mathbf{s}\|_1 < N$ in fact implies that $\mathbf{f}_\theta^{(n)}(\mathbf{s}) \in Z$, showing (92).

Thus we have the inequalities

$$\mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N) < \mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) \tag{93}$$

$$= \mathbb{P}(\mathbf{f}_\theta^{(n)}(\bar{\mathbf{S}}_\theta) \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) \tag{94}$$

$$\le \mathbb{P}(\mathbf{f}_\theta^{(n)}(\bar{\mathbf{S}}_\theta) \notin Z) \tag{95}$$

$$\le \mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N), \tag{96}$$

where (93) follows from (91), (94) follows from (87), (95) follows as Z and $B_\varepsilon(\tilde{\mathbf{s}}_\theta)$ are disjoint, and (96) follows from (92). Thus we have obtained a contradiction, which shows that $\bar{\mathbf{S}}_\theta$ equals $\mathbf{s}_\theta(\infty)$ with probability one. This completes the proof. □

A.6. Convergence of reassignments in the fluid limit. Before proving Corollary 1, we need a simple monotonicity result for the fluid limit $q_\theta(t)$.

LEMMA 12. *Under the assumptions of Corollary 1, for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, $q_\theta(t)$ is monotone on $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$. Further, for any $t^* \in [0,1]$ such that $q_{\mathrm{LS}}(t^*) = c$, (resp. $t^* \in [0, \frac{1}{2})$ such that $q_{\mathrm{DA}}(t^*) = 2c$), $q_\theta(t)$ is strictly monotone in a neighborhood of t^*. Consequently, there exists ε_0 such that for every $\varepsilon < \varepsilon_0$ there exists δ such that $|q_\theta(t) - q_\theta(t^*)| < \delta$ implies $|t - t^*| < \varepsilon$.*

PROOF. The monotonicity on $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ follows as on each such interval, $q_\theta(t)$ is the solution to the differential equation of the type from Lemma 9. For LS, for any t^* such that $q_{\mathrm{LS}}(t^*) = c$,

we have $\dot{q}_{\text{LS}}(t^*) = \lambda(t^*) - c\mu$. As we have assumed that $\lambda_1, \lambda_2 \neq c\mu$, we have that $\dot{q}_{\text{LS}}(t^*) \neq 0$. Noting that $q_{\text{LS}}(t)$ is continuously differentiable, it follows that $q_{\text{LS}}(t)$ is strictly monotone. A similar argument applies for DA. □

Finally, we show the convergence of the number reassignments (the number of patients forced to wait per day), completing the commutative diagram in Figure 3.

Proof of Corollary 1.. We first show that $W_{\text{LS}}^{1,n}(\infty)/n \to w_{\text{LS}}^1(\infty)$ in probability. Noting that the limit is a constant, it sufficient to show convergence in distribution. By Proposition 11.3.3 from [11], $W_{\text{LS}}^{1,n}(\infty)/n \Rightarrow w_{\text{LS}}^1(\infty)$ iff for all $g\colon \mathbb{R}_+ \to \mathbb{R}$ such that $\|g\|_{\text{BL}} \leq 1$, we have

$$\lim_{n\to\infty} \mathbb{E}\left[g\left(\frac{W_{\text{LS}}^{1,n}(\infty)}{n}\right)\right] = g(w_{\text{LS}}^1(\infty)).$$

For each n we take each $\mathbf{S}_{\text{LS}}^n(0) \overset{d}{=} \mathbf{S}_{\text{LS}}^n(\infty)$. Using Theorem 3 and the Skorokhod Representation Theorem, we put the $\mathbf{S}_{\text{LS}}^n(0)$ on a common probability space such that $\mathbf{S}_{\text{LS}}^n(0)/n \to \mathbf{s}_{\text{LS}}(\infty)$ a.s. We use this process to generate the $W_{\text{LS}}^{1,n}(\infty)$ and $w_{\text{LS}}^1(\infty)$ all on a common probability space.

It follows from Lemma 12 that there can be at most one time $t^* \in [0, \frac{1}{2})$ such that $q_{\text{LS}}(t^*) = c$, and that $q_{\text{LS}}(t)$ must be strictly monotone at t^*. Let ε_0 be from the lemma and fix some $\varepsilon \in (0, \varepsilon_0)$. Let δ be from Lemma 12 such that $|q_{\text{LS}}(t) - q_{\text{LS}}(t^*)| < \delta$ implies $|t - t^*| < \varepsilon$. Recall from Theorem 2 that when $\mathbf{S}_{\text{LS}}^n(0)/n \to \mathbf{s}_{\text{LS}}(0)$ a.s., then $\sup_{0\leq t\leq \frac{1}{2}} \|\mathbf{S}_{\text{LS}}^n(t)/n - \mathbf{s}_{\text{LS}}(t)\|_1 \to 0$ a.s. As $|Q_{\text{LS}}^n(t)/n - q_{\text{LS}}(t)| \leq \|\mathbf{S}_{\text{LS}}^n(t)/n - \mathbf{s}_{\text{LS}}(t)\|_1$, we also have $\sup_{0\leq t\leq \frac{1}{2}} |Q_{\text{LS}}^n(t)/n - q_{\text{LS}}(t)| \to 0$ a.s. As a result, we also have convergence in probability. In particular, for our δ,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{2}} \left|\frac{Q^n(t)}{n} - q(t)\right| > \delta\right) = 0.$$

Let E_δ^n be the event

$$E_\delta^n \triangleq \left\{\sup_{0\leq t\leq \frac{1}{2}} \left|\frac{Q^n(t)}{n} - q(t)\right| > \frac{\delta}{2}\right\},$$

i.e. $\lim_{n\to\infty}\mathbb{P}(E_\delta^n)=0$ for all δ. Let $\bar{E}_\delta^n$ denote the complement of this event. We have that

$$\begin{aligned}\left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right)\right]-g(w_{\mathrm{LS}}^1(\infty))\right| \leq & \left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right)-g(w_{\mathrm{LS}}^1(\infty))\;\middle|\;E_\delta^n\right]\right|\mathbb{P}(E_\delta^n)\\ &+\left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right)-g(w_{\mathrm{LS}}^1(\infty))\;\middle|\;\bar{E}_\delta^n\right]\right|\mathbb{P}(\bar{E}_\delta^n)\\ &\leq 2\|g\|_{\mathrm{BL}}\mathbb{P}(E_\delta^n)+\|g\|_{\mathrm{BL}}\mathbb{E}\left[\left|\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}-w_{\mathrm{LS}}^1(\infty)\right|\;\middle|\;\bar{E}_\delta^n\right], \qquad (97)\end{aligned}$$

where in (97), we are using both that g is bounded by $\|g\|_{\mathrm{BL}}$ and has Lipschitz constant at most $\|g\|_{\mathrm{BL}}$. Letting $n\to\infty$, we see the first term go to zero. For the second term, we consider two cases:

1. Suppose that $\inf_{t\in[0,\frac{1}{2})}|q_{\mathrm{LS}}(t)-c|=\gamma>0$. We can assume without loss of generality that $\delta<\gamma$, as we can always take δ smaller without interfering in the convergence of our first term, and doing so will only increase the proposed infimum. As $\delta<\gamma$, we ensure that conditional on $\bar{E}_\delta^n$, for every time $t\in[0,\frac{1}{2})$ that

$$\begin{aligned}\left|\frac{Q_{\mathrm{LS}}^n(t)}{n}-c\right| &\geq \left||q_{\mathrm{LS}}(t)-c|-\left|\frac{Q_{\mathrm{LS}}^n(t)}{n}-q_{\mathrm{LS}}(t)\right|\right| \qquad (98)\\ &= |q_{\mathrm{LS}}(t)-c|-\left|\frac{Q_{\mathrm{LS}}^n(t)}{n}-q_{\mathrm{LS}}(t)\right| \qquad (99)\\ &\geq \gamma/2,\end{aligned}$$

   where (98) is the reverse triangle inequality and (99) follows by our choice of δ. We have two further cases:

   (a) If $q_{\mathrm{LS}}(t)<c$ and thus $Q_{\mathrm{LS}}(t)/n<c$ for all t, then both $W_{\mathrm{LS}}^{1,n}(\infty)$ and $w_{\mathrm{LS}}^1(\infty)$ will be zero, so we will have that (97) converges to zero as $n\to\infty$.

   (b) Similarly, if $q_{\mathrm{LS}}(t)>c$ and thus $Q_{\mathrm{LS}}(t)/n>c$ for all t, then $W_{\mathrm{LS}}^{1,n}(\infty)=A^n(0,\frac{1}{2})$ and $w_{\mathrm{LS}}^1(\infty)=\lambda_1/2$. Thus

$$\begin{aligned}\mathbb{E}\left[\left|\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}-w_{\mathrm{LS}}^1(\infty)\right|\;\middle|\;\bar{E}_\delta^n\right] &= \frac{1}{\mathbb{P}\left(\bar{E}_\delta^n\right)}\mathbb{E}\left[\left|\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}-w_{\mathrm{LS}}^1(\infty)\right|\mathbb{I}_{\bar{E}_\delta^n}\right]\\ &= \frac{1}{\mathbb{P}\left(\bar{E}_\delta^n\right)}\mathbb{E}\left[\left|\frac{A^n(0,\frac{1}{2})}{n}-\frac{\lambda_1}{2}\right|\mathbb{I}_{\bar{E}_\delta^n}\right]\\ &\leq \frac{1}{\mathbb{P}\left(\bar{E}_\delta^n\right)}\mathbb{E}\left[\left|\frac{A^n(0,\frac{1}{2})}{n}-\frac{\lambda_1}{2}\right|\right].\end{aligned}$$

   The above converges to zero almost surely since $\mathbb{E}[A^n(0,\frac{1}{2})/n]\to\lambda_1/2$ as $n\to\infty$. Thus (97) converges to zero as $n\to\infty$.

2. Suppose instead that $q_{\mathrm{LS}}(t)$ crosses c. Suppose $\lambda_1 > c\mu$, so by Lemma 12 $q_{\mathrm{LS}}(t)$ is monotonically increasing. Let t^* be the time such that $q_{\mathrm{LS}}(t^*) = c$. Then

$$w^1_{\mathrm{LS}}(\infty) = \int_{t^*}^{\frac{1}{2}} \lambda_1 dt = (\tfrac{1}{2} - t^*)\lambda_1.$$

We claim that conditional on $\bar{E}^n_\delta$,

$$|t - t^*| \geq \varepsilon \text{ implies that } |Q^n_{\mathrm{LS}}(t)/n - c| \geq \delta/2. \tag{100}$$

We will show the contrapositive. We have that when $|Q^n_{\mathrm{LS}}(t)/n - c| < \delta/2$, then

$$|q_{\mathrm{LS}}(t) - c| \leq \left| q_{\mathrm{LS}}(t) - \frac{Q^n_{\mathrm{LS}}(t)}{n} \right| + \left| \frac{Q^n_{\mathrm{LS}}(t)}{n} - c \right| < \delta.$$

Now by Lemma 12, $|q_{\mathrm{LS}}(t) - c| < \delta$ implies $|t - t^*| < \varepsilon$, showing the claim.

Next, we claim that then conditional on $\bar{E}^n_\delta$, for all $t > t^* + \varepsilon$, $Q^n_{\mathrm{LS}}(t)/n > c$. Assume not for contradiction. Then

$$\begin{aligned} 0 &\leq c - \frac{Q^n_{\mathrm{LS}}(t)}{n} - \frac{\delta}{2} && (101) \\ &\leq c - q_{\mathrm{LS}}(t) + \left| \frac{Q^n_{\mathrm{LS}}(t)}{n} - q_{\mathrm{LS}}(t) \right| - \frac{\delta}{2} \\ &\leq c - q_{\mathrm{LS}}(t) && (102) \\ &< 0, && (103) \end{aligned}$$

giving a contradiction. Here (101) holds by (100) in conjunction with $Q^n_{\mathrm{LS}}(t)/n < c$, (102) holds as we are assuming $\bar{E}^n_\delta$, and finally (103) holds as $\lambda_1 > c\mu$ and Lemma 12 implies that $q_{\mathrm{LS}}(t)$ is increasing in t, $q_{\mathrm{LS}}(t^*) = c$, and $t > t^*$.

By an analogous argument, we can show that for all $t < t^* - \varepsilon$, $Q^n_{\mathrm{LS}}(t)/n < c$. As the number of reassignments $W^{1,n}_{\mathrm{LS}}(\infty)$ is the number of arrivals such that $Q^n_{\mathrm{LS}}(t) \geq cn$ at the time t of arrival, we thus have that all arrivals $A^n(t^* + \varepsilon, \frac{1}{2})$, will be reassignments, none of the arrivals $A^n(0, t^* - \varepsilon)$ will be reassignments, and the remaining arrivals are to be determined. This implies

$$A^n(t^* + \varepsilon, \tfrac{1}{2})\mathbb{I}_{\bar{E}^n_\delta} \leq W^{1,n}_{\mathrm{LS}}(\infty)\mathbb{I}_{\bar{E}^n_\delta} \leq A^n(t^* - \varepsilon, \tfrac{1}{2})\mathbb{I}_{\bar{E}^n_\delta}.$$

Thus

$$\begin{aligned}\mathbb{E}\left[\left|\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}-w_{\mathrm{LS}}^{1}(\infty)\right|\,\middle|\,\bar{E}_{\delta}^{n}\right] &\leq \mathbb{E}\left[\left|\frac{A_{\mathrm{LS}}^{1,n}(t^*+\varepsilon)}{n}-w_{\mathrm{LS}}^{1}(\infty)\right|+\left|\frac{A_{\mathrm{LS}}^{1,n}(t^*-\varepsilon)}{n}-w_{\mathrm{LS}}^{1}(\infty)\right|\,\middle|\,\bar{E}_{\delta}^{n}\right]\\ &\leq \mathbb{E}\left[\left|\frac{A_{\mathrm{LS}}^{1,n}(t^*+\varepsilon)}{n}-w_{\mathrm{LS}}^{1}(\infty)\right|+\left|\frac{A_{\mathrm{LS}}^{1,n}(t^*-\varepsilon)}{n}-w_{\mathrm{LS}}^{1}(\infty)\right|\right]\Big/\mathbb{P}(\bar{E}_{\delta}^{n})\\ &\leq 2\varepsilon\lambda_1/\mathbb{P}(\bar{E}_{\delta}^{n}).\end{aligned}$$

Letting $n\to\infty$, the above converges to $2\varepsilon\lambda_1$. As ε was arbitrary, the above term and thus (97) must converge to zero as $n\to\infty$.

It is not hard to see that if instead $\lambda_1 < c\mu$, then as $q_{\mathrm{LS}}(t)$ will be decreasing, we will obtain a similar bound of the form

$$A^n(0,t^*-\varepsilon)\mathbb{I}_{\bar{E}_{\delta}^{n}} \leq W_{\mathrm{LS}}^{1,n}(\infty)\mathbb{I}_{\bar{E}_{\delta}^{n}} \leq A^n(0,t^*+\varepsilon)\mathbb{I}_{\bar{E}_{\delta}^{n}},$$

After taking expectations, we could again show the convergence of (97) to zero. Thus we conclude that $W_{\mathrm{LS}}^{1,n}(\infty)/n \Rightarrow w_{\mathrm{LS}}^{1}(\infty)$.

Showing that $W_{\mathrm{LS}}^{2,n}(\infty) \Rightarrow w_{\mathrm{LS}}^{2}(\infty)$, $W_{\mathrm{DA}}^{1,n}(\infty)/n \Rightarrow w_{\mathrm{DA}}^{1}(\infty)$, and $W_{\mathrm{DA}}^{2,n}(\infty)/n \Rightarrow w_{\mathrm{DA}}^{2}(\infty)$ is very similar and the details are omitted. □

A.7. Proof of Lemma 9. Here we give a series of lemmas about the differential equation

$$\dot{x}(t) = \gamma - \mu(x(t)\wedge m),$$

with $x(0)\in\mathbb{R}_+$, $\gamma,\mu,m>0$, that will ultimately allow us to prove Lemma 9. For convenience, we let $g(x,t)$ equal $x(t)$ when $x(0)=x$ (making the function $g(x)$ defined in Lemma 9 equal to $g(x,\frac{1}{2})$).

LEMMA 13. *The differential equation given by $\dot{x}(t)$ with initial condition $x(0)\in\mathbb{R}_+$ has a unique solution $x(t)$ with $x(t)\geq 0$ for all $t\in[0,\frac{1}{2}]$. Further, for every $x\in\mathbb{R}_+$, $g(x,t)$ is either strictly increasing in t for all t, strictly decreasing in t for all t, or equal to $g(x,0)$ for all t.*

PROOF. The differential equations

$$\dot{y}(t) = \gamma - m\mu,$$

$$\dot{z}(t) = \gamma - z(t)\mu,$$

with an initial condition $y(s) \in \mathbb{R}_+$ and $z(s) \in \mathbb{R}_+$ both have unique solutions for all $t > 0$ given by

$$y(t) = y(s) + (t-s)(\gamma - m\mu),$$
$$z(t) = \frac{\gamma}{\mu} + \exp(-\mu(t-s))\left(z(s) - \frac{\gamma}{\mu}\right),$$

respectively. We claim that these two differential equations in combination determine the path of $x(t)$. Given our formula for $\dot{x}(t)$, we observe that $x(t) \geq m$ and $x(t) = y(t)$ implies $\dot{x}(t) = \dot{y}(t)$. Suppose that for some s we have $x(s) \geq m$ and let $y(s) = x(s)$. Then for all $t \geq s$ such that $y(t) \geq m$, we will have $x(t) = y(t)$. Analogously, we observe that $x(t) < m$ and $x(t) = z(t)$ implies that $\dot{x}(t) = \dot{z}(t)$. Suppose that for some s we have $x(s) < m$, and let $z(s) = x(s)$. Then for all $t \geq s$ such that $z(t) \leq m$, we will have $x(t) = z(t)$. Thus we can show that $x(t)$ has a unique solution on $[0, \frac{1}{2})$ for all initial conditions by showing that there is a clean exchange at the boundary $\{t \mid x(t) = m\}$. In particular, it suffices to show that we cross the boundary at most one time, which follows from the monotonicity claim in the second part of the Lemma.

First however, we need to analyze the long run behavior of $z(t)$. We can immediately see from $\dot{z}(t)$ that if $z(0) = \gamma/\mu$, then $\dot{z}(t) = 0$ so $z(t) = \gamma/\mu$ for all t. Similarly, when $z(t) < \gamma/\mu$, z will be strictly increasing at t, and when $z(t) > \gamma/\mu$, z will be strictly decreasing at t. Further, from the solution for $z(t)$, we see that if at any time s, $z(s) < \gamma/\mu$ then for all times $t > s$, we will still have $z(t) < \gamma/\mu$. Namely, z will approach γ/μ but never reach it. Likewise, when $z(s) > \gamma/\mu$, we will have $z(t) > \gamma/\mu$ for all $t > s$.

We can now finish the Lemma by considering three cases:

1. Suppose $\gamma > m\mu$, or equivalently $\gamma/\mu > m$. If $x(0) < m$, then as the attractive point of $z(t)$ is greater than m, we will have $x(t)$ strictly increasing until either time $\frac{1}{2}$ or s such that $z(s) = m$, if $s < \frac{1}{2}$. There is nothing left the prove in the first case, so we consider the second. Once $x(t) \geq m$, as $\gamma \geq m\mu$, we have $\dot{x}(t) = \dot{y}(t) = \gamma - m\mu > 0$, so $x(t)$ will increase strictly and never again fall before m. Thus in all cases, $x(t)$ is strictly increasing for all t.
2. Suppose $\gamma < m\mu$. Then if $x(t) \geq m$, we will have $\dot{x}(t) = \dot{y}(t) = \gamma - m\mu < 0$, so $x(t)$ will be strictly decreasing. Again there are two possibilities, either there is a time $s < \frac{1}{2}$ such that $x(s) = m$, or $x(s)$ will not reach m before time $\frac{1}{2}$. As $x(t)$ is uniquely defined in the second case, we need only consider the first case further. Once we reach m, the dynamics of $x(t)$ will

be that of $z(t)$. Recall that $z(t)$ will strictly decrease towards the fixed point γ/μ for all $t > s$ when $z(s) < \gamma/\mu$. Thus for all $x > \gamma/\mu$, $g(x,t)$ is strictly decreasing in t. When $x(0) = \gamma/\mu$, then by our previous analysis of $z(t)$ we have that $g(x(0),t) = \gamma/\mu$ for all t. Finally, when $x(0) < \gamma/\mu$, we know that $x(t)$ will be strictly increasing for all t towards γ/μ. Thus in all cases on $x(0)$ the criteria of the Lemma are met.

3. Suppose $\gamma = m\mu$. Then for all $x \geq m$, $\dot{x}(t) = \gamma - m\mu = 0$, so $g(x,t) = g(x,0)$. For all $x < m$, by our analysis of $z(t)$, we know that $x(t)$ will be strictly increasing towards $\gamma/\mu = m$ but never reach it.

Thus we can conclude that $x(t)$ has a unique solution for all $t \in [0, \frac{1}{2})$. □

LEMMA 14. *When $x > y$, we have $g(x,t) > g(y,t)$ for all t.*

PROOF. We will make a coupling argument. By Lemma 13, we know $g(x,t)$ is either strictly increasing in t, strictly decreasing in t, or constant. Assume for contradiction that there is a time s such that $g(y,s) \geq g(x,s)$. We now consider cases:

1. Suppose that $g(x,t)$ is strictly increasing in t. As $g(y,t)$ is continuous in t, and at time s, $g(y,s) \geq g(x,s) > g(x,0)$, by the Intermediate Value Theorem there must be some time r with $0 < r \leq s$ such that $g(y,r) = g(x,0)$. But as $\dot{x}(t)$ is not a function of t, only $x(t)$, we thus obtain that $g(y,s) = g(x,s-r) < g(x,s)$, giving a contradiction.
2. Suppose that $g(x,t)$ is constant. As $g(y,t)$ is continuous in t, and at time s, we have $g(y,s) \geq g(x,s) = g(x,0)$, by the Intermediate Value Theorem there is a time $r \leq s$ such that $g(y,r) = g(x,0)$. But then $g(y,t)$ is constant at r, contradicting that $g(y,t)$ must either be strictly increasing, strictly decreasing, or constant for all t.
3. Suppose that $g(x,t)$ is strictly decreasing and $g(y,t)$ is either strictly decreasing or constant. Then by taking $\bar{g}(x,t) \stackrel{\Delta}{=} -g(y,t)$ and $\bar{g}(y,t) \stackrel{\Delta}{=} -g(x,t)$, we can apply cases one and two to $\bar{g}(x,t)$ to show the claim.
4. Finally, suppose that $g(x,t)$ is strictly decreasing and $g(y,t)$ is strictly increasing. Under our assumption that $g(y,s) \geq g(x,s)$, again by the Intermediate Value Theorem, there must be some time $0 < r \leq s$ such that $g(x,r) = g(y,r)$. But as $\dot{x}(t)$ depends only $x(t)$, not t, we would then have that for all $t \geq r$, $g(x,t) = g(y,t)$. This creates a contradiction, as we have

assumed that $g(x,t)$ is strictly decreasing in t and $g(y,t)$ is strictly increasing in t.

□

LEMMA 15. *For $x \geq \tilde{x}$ as defined in Lemma 9, $g(x,t) \geq m$ for $0 \leq t \leq \frac{1}{2}$ and $g(x,\frac{1}{2}) = x + (\gamma - c\mu)/2$. For $x < \tilde{x}$, there exist $0 \leq s < t \leq \frac{1}{2}$ such that for $\tau \in (s,t)$, $g(x,\tau) < m$, and $g(x,\frac{1}{2}) < x + (\gamma - c\mu)/2$.*

PROOF. We compute for $0 \leq t \leq \frac{1}{2}$ that

$$\begin{aligned} g(\tilde{x},t) &= \tilde{x} + \int_0^t \gamma - \mu(m \wedge g(\tilde{x},\tau))d\tau \\ &\geq \tilde{x} + \int_0^t \gamma - m\mu d\tau \\ &= \tilde{x} + t(\gamma - m\mu). \end{aligned} \tag{104}$$

We first consider $g(\tilde{x},t)$ in cases:

1. Suppose that $\gamma \geq m\mu$. Then $\tilde{x} = m$, and as $\gamma - m\mu \geq 0$, we obtain from (104) that $g(\tilde{x},t) \geq m$ for all $t \geq 0$.
2. Suppose that $\gamma < m\mu$. Then $\tilde{x} = m - (\gamma - m\mu)/2$, so by (104) for all $t \geq 0$,

$$g(\tilde{x},t) \geq m + (\gamma - m\mu)\left(t - \frac{1}{2}\right) \geq m.$$

We thus conclude that $g(\tilde{x},t) \geq m$ for all $0 \leq t \leq \frac{1}{2}$. As a result, we can now make the exact computation that for all $x \geq \tilde{x}$,

$$\begin{aligned} g(x,\tfrac{1}{2}) &= x + \int_0^{\frac{1}{2}} \gamma - \mu(m \wedge g(x,t))dt \\ &= x + \int_0^{\frac{1}{2}} \gamma - \mu m dt \\ &= x + \frac{1}{2}(\gamma - m\mu). \end{aligned} \tag{105}$$

where (105) holds as $g(x,t) \geq g(\tilde{x},t) \geq m$ by Lemma 14 and then the above analysis, making $m \wedge g(x,t) = m$ for all t.

We now consider $x < \tilde{x}$, and find s and t such that for all $\tau \in (s,t)$, $g(x,\tau) < m$, as in the statement of the lemma. We consider cases:

1. Suppose $\gamma > m\mu$ and thus $\tilde{x} = m$. Then $g(x, 0) = x < \tilde{x} = m$ so obviously we can take $s = 0$ and t small to show the claim.
2. Suppose instead that $\gamma < m\mu$ and thus $\tilde{x} = m - (\gamma - m\mu)/2$. For $x < m$, again the claim obviously holds as then $g(x, t) < m$ for all $t \le \frac{1}{2}$. For x such that $m \le x < \tilde{x}$, observe that as $\tilde{x} = m - (\gamma - m\mu)/2$,

$$0 \le t^* \triangleq \frac{x - m}{m\mu - \gamma} < \frac{\tilde{x} - m}{m\mu - \gamma} = \frac{1}{2},$$

and thus

$$g(x, t^*) = x + (\gamma - m\mu)\frac{x - m}{m\mu - \gamma} = m.$$

As $\dot{x}(t) = \gamma - \mu(m \wedge x(t)) \le \gamma - m\mu < 0$ by our assumptions, we can take $s = t^*$ and $t = \frac{1}{2}$.

Finally, using s and t from the statement of the lemma, we show that $g(x(0), \frac{1}{2}) > x(0) + (\gamma - c\mu)/2$ when $x(0) < \tilde{x}$. We compute that

$$\begin{aligned}
g(x, \tfrac{1}{2}) &= x + \int_0^{\frac{1}{2}} \gamma - \mu(g(x, \tau) \wedge m) d\tau \\
&\ge x + \int_0^s \gamma - \mu m d\tau + \int_s^t \gamma - \mu g(x, \tau) d\tau + \int_t^{\frac{1}{2}} \gamma - \mu m d\tau \\
&= x + (\gamma - m\mu)(\tfrac{1}{2} - (t - s)) + \int_s^t \gamma - \mu g(x, \tau) d\tau \\
&> x + (\gamma - m\mu)/2.
\end{aligned}$$

The final inequality is strict as $g(x, \tau) < m$ for all $\tau \in (s, t)$. □

LEMMA 16. *For all $x > y \ge \tilde{x}$, where $\tilde{x}$ is defined in Lemma 9,*

$$g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) = x - y,$$

and for all $x > y$, $y < \tilde{x}$,

$$0 < g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) < x - y.$$

PROOF. For $x > y \ge \tilde{x}$, by Lemma 15, we have that

$$g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) = x + (\gamma - m\mu)/2 - y - (\gamma - m\mu)/2 = x - y.$$

For $x > y$, $y < \tilde{x}$, let s and t be from Lemma 15 such that for all $\tau \in (s,t)$, $g(y,\tau) < m$. Then

$$g(x,\tfrac{1}{2}) - g(y,\tfrac{1}{2}) = x - y + \mu \int_0^{\frac{1}{2}} (g(y,\tau) \wedge m) - (g(x,\tau) \wedge m) d\tau.$$

As $x > y$, by Lemma 14 we have $g(x,\tau) > g(y,\tau)$ for all τ, and thus $g(x,\tau) \wedge m \geq g(y,\tau) \wedge m$ for all τ, making the integrand nonpositive. Thus

$$\begin{aligned} g(x,\tfrac{1}{2}) - g(y,\tfrac{1}{2}) &= x - y + \mu \int_s^t (g(y,\tau) \wedge m) - (g(x,\tau) \wedge m) d\tau \\ &= x - y + \mu \int_s^t g(y,\tau) - (g(x,\tau) \wedge m) d\tau \\ &\leq x - y + \mu \int_s^t g(y,\tau) - g(x,\tau) d\tau \\ &< x - y. \end{aligned}$$

That $g(x,\frac{1}{2}) - g(y,\frac{1}{2}) > 0$ follows immediately from Lemma 14. □

Proof of Lemma 9.. The Lemma follows immediately from Lemma 13, Lemma 14, Lemma 15, and Lemma 16. □

A.8. Null Recurrence and Transience. We give a sufficient condition to distinguish between the null recurrent and transient cases from [26], Theorem 3.2, (see also Section 3.6 from [14]). We do not present the theorem in full generality. As in Appendix A.1, suppose we have a discrete time irreducible Markov chain $\{\mathbf{X}_k\}$ taking values in $\mathcal{X} \subset \mathbb{Z}^d$.

Proposition 10. *Given a finite set $B \subset \mathcal{X}$ and Lyapunov function $V \colon \mathcal{X} \to \mathbb{R}_+$, assume that*

$$\mathbb{P}\left(\limsup_{k\to\infty} V(\mathbf{X}_k) = \infty\right) = 1, \tag{106}$$

$$\inf_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{x}}\left[(V(\mathbf{X}_1) - V(\mathbf{X}_0))^2\right] > 0, \tag{107}$$

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{x}}\left[(V(\mathbf{X}_1) - V(\mathbf{X}_0))^4\right] < \infty. \tag{108}$$

If for all $\mathbf{x} \in \mathcal{X} \setminus B$,

$$\mathbb{E}_{\mathbf{x}}\left[V(\mathbf{X}_1) - V(\mathbf{X}_0)\right] \leq \frac{\mathbb{E}_{\mathbf{x}}\left[(V(\mathbf{X}_1) - V(\mathbf{X}_0))^2\right]}{2V(\mathbf{x})}, \tag{109}$$

then $\{X_k\}$ is recurrent. Alternatively, if there exists $\varepsilon > 0$ such that for all $x \in \mathcal{X} \setminus B$,

$$\mathbb{E}_{\mathbf{x}}\left[V(\mathbf{X}_1) - V(\mathbf{X}_0)\right] \geq (1+\varepsilon)\frac{\mathbb{E}_{\mathbf{x}}\left[(V(\mathbf{X}_1) - V(\mathbf{X}_0))^2\right]}{2V(\mathbf{x})}, \tag{110}$$

then $\{X_n\}$ is transient.

REMARK 3. As noted in [26], a sufficient condition for (106) is that for every $z \geq 0$,

$$\inf_{\mathbf{x} \in \mathcal{X}} \mathbb{P}(V(\mathbf{X}_1) \geq z \mid \mathbf{X}_0 = x) > 0. \tag{111}$$

We now return to our model. As before, we will use the Lyapunov function $V(q,r) \stackrel{\Delta}{=} q + r$.

LEMMA 17. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the following limit exists, is finite, and is non-zero:*

$$F_\theta \stackrel{\Delta}{=} \lim_{\substack{q \to \infty \\ (q,r) \in \mathcal{S}_\theta}} \mathbb{E}_{(q,r)} \left[\left(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0)) \right)^2 \right]. \tag{112}$$

Further,

$$\sup_{\mathbf{s} \in \mathcal{S}_\theta} \mathbb{E}_{\mathbf{s}} \left[\left(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0)) \right)^4 \right] < \infty. \tag{113}$$

Finally, when $\rho_\theta = 1$, as $q \to \infty$, we have for r such that $(q,r) \in \mathcal{S}_\theta$ that

$$\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] = O(\exp(-q/2)). \tag{114}$$

PROOF. First consider $\theta = \mathrm{LS}$. Recall from (31) that for any $\ell \geq 0$,

$$\mathbb{E}_{(q,c)} \left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)) \right)^\ell \right] = \mathbb{E}_{(q,c)} \left[\left(A + D^{\mathrm{on}}_{\mathrm{LS}} - D^{\mathrm{off}}_{\mathrm{LS}} \right)^\ell \right].$$

Further, recall that we coupled $D^{\mathrm{on}}_{\mathrm{LS}}$ with $\tilde{D}^{\mathrm{on}}_{\mathrm{LS}}$ that had distribution $\mathrm{Pois}(c\mu)$ such that $D^{\mathrm{on}}_{\mathrm{LS}} < \tilde{D}^{\mathrm{on}}_{\mathrm{LS}}$, and $D^{\mathrm{off}}_{\mathrm{LS}}$ with $\tilde{D}^{\mathrm{off}}_{\mathrm{LS}}$ that had distribution $\mathrm{Bin}(c, 1-e^{-\mu})$ such that $D^{\mathrm{off}}_{\mathrm{LS}} \leq \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}$. Using these couplings, we can show (113) just as we showed (29).

Recall that under our coupling, we have $D^{\mathrm{off}}_{\mathrm{LS}} = \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}$ for initial (q,r) such that $r = c$. Likewise, we have that $D^{\mathrm{on}}_{\mathrm{LS}}$ was equal to $\tilde{D}^{\mathrm{on}}_{\mathrm{LS}}$ under an initial condition (q,r) for those realizations where $\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} < q - c$. This led to (34), namely that for $q > c$,

$$\mathbb{E}_{(q,c)} \left[D^{\mathrm{on}}_{\mathrm{LS}} \right] \geq \mathbb{E} \left[\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} < q-c\}} \right].$$

We now can show (114). Assuming $\rho_\theta = 1$ and thus $\gamma_\theta = 0$, we have

$$\begin{aligned}
\mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))] &= \mathbb{E}\left[A - D^{\mathrm{off}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}}\right] + \mathbb{E}_{(q,c)}\left[\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - D^{\mathrm{on}}_{\mathrm{LS}}\right] \\
&\leq -\gamma_\theta + \mathbb{E}[\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \geq q-c\}}] \\
&\leq \sqrt{\mathbb{E}\left[\left(\tilde{D}^{\mathrm{on}}_{\mathrm{LS}}\right)^2\right] \mathbb{E}\left[\mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \geq q-c\}}\right]} \\
&\leq \sqrt{((c\mu)^2 + c\mu)\mathbb{E}\left[\exp\left(\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - q + c\right)\right]} \\
&\leq \exp(-q/2)\sqrt{((c\mu)^2 + c\mu)\exp(c)\exp(c\mu(e-1))} \\
&= O\left(\exp(-q/2)\right),
\end{aligned}$$

as $q \to \infty$. Here we use that for any random variable X, $\mathbb{I}_{\{X \geq t\}} \leq \exp(X - t)$.

It remains to show (112). We will show that

$$F_{\mathrm{LS}} = \mathbb{E}[(A - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{off}}_{\mathrm{LS}})^2].$$

For any $q > 0$, observe that

$$\begin{aligned}
\mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^2\right] &= \mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right] \\
&\quad + \mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right].
\end{aligned}$$

For the second term, we have that

(115)
$$\lim_{q\to\infty} \mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right] = \lim_{q\to\infty} \mathbb{E}\left[\left(A - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right]$$

(116)
$$\begin{aligned}
&= \mathbb{E}\left[\left(A - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}\right)^2\right] \\
&= F_{\mathrm{LS}},
\end{aligned}$$

where (115) follows by our coupling and (116) follows from the Monotone Convergence Theorem.

For the first term, we have

$$\begin{aligned}
&\lim_{q\to\infty} \mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right] \\
&\qquad \le \lim_{q\to\infty} \sqrt{\mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^4\right] \mathbb{E}_{(q,c)}\left[\mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right]} \\
&\qquad \le \lim_{q\to\infty} \sqrt{\mathbb{P}\left(\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - c > q\right)} \sqrt{\sup_{\mathbf{s}\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{\mathbf{s}}\left[\left(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))\right)^4\right]} \\
&\qquad = 0,
\end{aligned}$$

where in the final equality we use (113). This shows (112) and thus the Lemma in the case of LS. The case of DA is similar. □

We now complete the proof of Theorem 1 by showing that for each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, $\{\mathbf{S}_\theta(k)\}$ is null recurrent when $\rho_\theta = 1$ and transient when $\rho_\theta > 1$.

Proof of Theorem 1.. We have already established in section Appendix A.1 that when $\rho_\theta \ge 1$, $\{\mathbf{S}_\theta(k)\}$ is either null recurrent or transient, and when $\rho_\theta < 1$, $\{\mathbf{S}_\theta(k)\}$ is positive recurrent. Thus it suffices to show that $\{\mathbf{S}_\theta(k)\}$ is recurrent when $\rho_\theta = 1$ and transient otherwise. We proceed using Proposition 10. We must check that the assumptions of the proposition are satisfied by $\{\mathbf{S}_\theta(k)\}$ for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$. By (113) from Lemma 17, (108) is satisfied. It is obvious that (107) is satisfied. Finally, to check (106), we verify the sufficient condition (111). Recalling Corollary 2, it is immediate that for all $z > 0$,

$$\inf_{\mathbf{s}\in\mathcal{S}_\theta} \mathbb{P}(V(\mathbf{S}_\theta(1)) \ge z \mid \mathbf{S}_\theta(0) = \mathbf{s}) = \mathbb{P}(V(\mathbf{S}_\theta(1)) \ge z \mid \mathbf{S}_\theta(0) = (0,0)) > 0.$$

Now suppose $\rho_\theta = 1$. We will show that (109) holds. For a constant $b_\theta > 0$, we will take B_θ of the form $\{(q,r) \in \mathcal{S}_\theta \mid q < b_\theta\}$. By (114) from Lemma 17, as $q \to \infty$, we have

$$\mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right] = O(\exp(-q/2)),$$

and by (112) from Lemma 17, as $q \to \infty$,

$$\frac{\mathbb{E}_{(q,r)}\left[\left(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right)^2\right]}{2V(q,r)} = \Theta\left(\frac{1}{q}\right).$$

Thus by taking b_θ sufficiently large and using that $\exp(-q/r) = o(1/q)$, we have (109) for all $\mathbf{s} \in \mathcal{S}_\theta \setminus B_\theta$, showing that $\{\mathbf{S}_\theta(k)\}$ is null recurrent.

Alternatively, suppose that $\rho_\theta > 1$. We now must show that (110) holds. We use b_θ to define B_θ in the same way. By Lemma 1 and Lemma 2 we have

$$\lim_{\substack{q\to\infty \\ (q,r)\in\mathcal{S}_\theta}} \mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right] = -\gamma_\theta > 0,$$

and again by (112) we have

$$\frac{\mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right]}{2V(\mathbf{S}_\theta(0))} = \Theta\left(\frac{1}{q}\right).$$

Thus by taking b_θ sufficiently large and $\varepsilon = 1$, we have (110), showing $\{\mathbf{S}_\theta(k)\}$ is transient, completing the proof. □

Operations Research Center
Massachusetts Institute of Technology, E40-148
Cambridge, Massachusetts, 02139
E-mail: rma350@mit.edu
URL: http://rma350.scripts.mit.edu/home/

Sloan School of Management
Massachusetts Institute of Technology, E62
Cambridge, Massachusetts, 02139
E-mail: gamarnik@mit.edu
URL: http://www.mit.edu/ gamarnik/home.html