S.G. Henderson and B.L. Nelson (Eds.), *Handbook in OR & MS, Vol. 13*

DOI: 10.1016/S0927-0507(06)13002-9

Chapter 2
Mathematics for Simulation

Shane G. Henderson

School of Operations Research and Industrial Engineering, Cornell University, USA
E-mail: sgh9@cornell.edu

Abstract

This chapter surveys certain mathematical results and techniques that are pervasive in the analysis of stochastic simulation. The concepts are introduced through the study of a simple model of ambulance operation to ensure clarity, concreteness and cohesion.

1 Introduction

Stochastic simulation (henceforth just "simulation") is a tremendously broad subject that draws from diverse mathematical fields including, but certainly not limited to, applied probability, statistics, number theory, and mathematical programming. One of the goals of this handbook is to survey stochastic simulation, communicating the key concepts, techniques and results that serve as a foundation for the field. This chapter contributes to that goal by surveying a collection of mathematical techniques and results that pervade simulation analysis.

Given the breadth of the field of simulation, it is necessary to focus the discussion somewhat. This chapter describes a set of mathematical tools and techniques that can be used to explore the estimation of performance measures in both the terminating and steady-state simulation context. The focus throughout is on *large-sample properties* of estimators, i.e., those properties of estimators that persist when simulation runlengths become large. Any practical simulation study works with a finite simulation runlength and so may not reach the regime where large-sample properties emerge. It is therefore of great practical importance to consider the *small-sample properties* of estimators as well. We do not do so here, since to do so would enlarge the scope beyond what is manageable in a single chapter.

A refined statement of the goal of this chapter is then to survey a subset of mathematical techniques and results that are useful in understanding the large-

sample behavior of estimators of performance measures. It would be very easy to provide a smorgasbord of such results, but such a chapter would read like a dictionary. Therefore, many of the results are applied to a simple model of ambulance operation that serves to unify the discussion and hopefully make it more interesting.

This chapter is an outgrowth of Henderson (2000) and Henderson (2001), in which there were 2 main topics. First, in the terminating simulation context, performance measures were rigorously defined through the strong law of large numbers for i.i.d. random variables. The performance of estimators of these performance measures was studied via the central limit theorem. Variants of these results were used to study performance measures that, instead of being expectations of random variables, were *functions* of expectations of random variables. Second, in the steady-state context, performance measures were rigorously defined and analyzed by appealing to asymptotic results for general state-space Markov chains. Lyapunov conditions were used to provide sufficient conditions under which the asymptotic results hold.

All of the performance measures described in Henderson (2000) take the form of an expectation of a random variable, or a differentiable function of a finite number of expectations. Such performance measures are particularly useful when the goal is to compare many different stochastic systems, as they provide a concrete basis for the comparison. If instead the goal is to enhance one's *understanding* of a single stochastic system, then it is often more useful to analyze the *distribution* of certain random variables, perhaps through density estimation techniques. This was the focus of Henderson (2001). This chapter combines elements of both of those papers, with a leaning toward density estimation.

In Section 2 we review some approaches to performance-measure estimation in a particularly transparent context, namely that of estimating the density of the completion time in a stochastic activity network. The analysis in this section requires the use of the strong law of large numbers (SLLN) and central limit theorem (CLT). We also review the continuous mapping theorem and converging together lemma.

Section 3 sets the stage for the remainder of the chapter by introducing a simple model of ambulance operation. In Section 4 we specialize this model to the terminating simulation context. Even the *definition* of certain performance measures leads to the use of some interesting techniques and results.

In Section 5 we modify the ambulance model slightly to obtain a steady-state simulation. To rigorously define performance measures for this model, it is necessary to define an appropriate stochastic process with which to work. A great deal is known about the class of Markov processes evolving on general (not necessarily countable) state spaces, and so a general state space Markov chain is defined. To ensure that long-run averages exist, it is necessary to show that this chain is, in a certain sense, positive recurrent.

A very practical approach to establishing that a Markov chain is positive recurrent is to use Lyapunov functions, and this approach is the central math-

ematical tool illustrated in Section 5. We use Lyapunov theory to show that certain Markov chains are positive recurrent, that our performance measures are well defined, and that certain estimators are consistent and satisfy central limit theorems. An important consideration in the steady-state context is that of initialization bias. We also use Lyapunov theory to characterize this bias. Sometimes a natural choice of Lyapunov function does not work, at least at first sight. Section 5 concludes with a discussion of one approach to dealing with such problems that is especially applicable to queueing examples.

The underlying theme of Section 5 is that Lyapunov functions provide an enormously powerful and easily applied (at least relative to many other methods) approach to establishing results that underlie steady-state simulation methodology. It is fair to say that Lyapunov functions have not been broadly applied in simulation analysis. It is far more common, for example, to see results based on mixing hypotheses for stationary processes. But such hypotheses are difficult to verify in practice, which helps to explain the attention devoted to Lyapunov techniques in this chapter.

Throughout the chapter results are rigorously quoted and references given for the proofs. To simplify the exposition it is often the case that results are quoted using stronger hypotheses than are strictly necessary, but tighter hypotheses can be found in the references provided.

Notation is occasionally reused from section to section, but is consistently applied within each section.

2 Static simulation: Activity networks

Our running example in this section will be that of a stochastic activity network (SAN), as introduced in Chapter 1. A SAN is a directed graph that represents some set of activities that, taken together, can represent some project/undertaking. Each arc in the graph represents a task that needs to be completed, and the (random) length of an arc represents the time required to complete the associated task. Nodes are used to indicate the precedence relationships between tasks. The time required to complete the project is indicated by the longest path between the designated "source" and "sink" nodes.

Example 1. The stochastic activity network in Figure 1 is adapted from Avramidis and Wilson (1996). Nodes a and i are the source and sink nodes respectively. The arcs are labeled for easy identification. For simplicity we assume that the task durations associated with each arc are independent of one another.

Let Y be the (random) network completion time, i.e., the longest path from node a to node i. We are interested in computing both $\mathrm{E}Y$ (assuming it is finite) and the distribution of Y. We first consider $\mathrm{E}Y$.

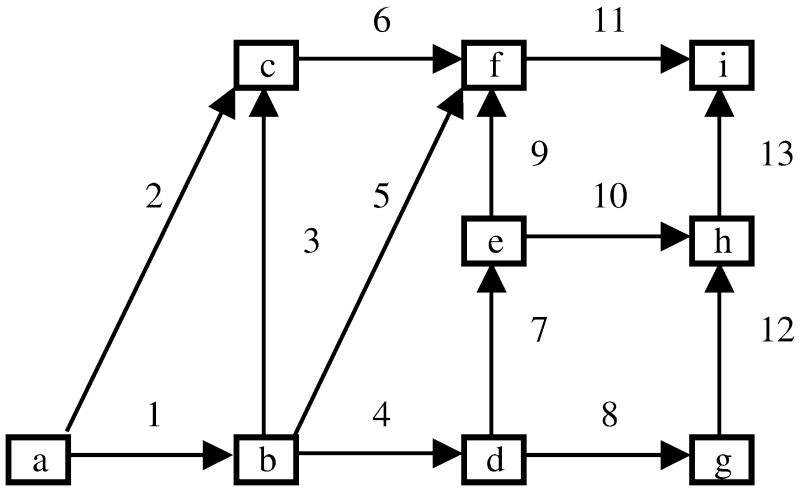

Fig. 1. A stochastic activity network.

As noted in Chapter 1, it is easy to estimate $\mathrm{E}Y$. One simply generates i.i.d. replicates $Y_1, Y_2, \ldots, Y_n$ of Y, and then forms the sample average

$$\overline{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

The strong law of large numbers ensures that the estimator $\overline{Y}_n$ is a strongly consistent estimator of $\mathrm{E}Y$. (An estimator is *strongly consistent* if it converges almost surely to the appropriate value. It is *consistent* if it converges in probability to the appropriate value. In simulation we don't usually concern ourselves with the distinction between these two concepts since it is impossible to distinguish between the types of convergence based on a finite runlength. However, the difference is important in establishing the validity of sequential stopping (Glynn and Whitt, 1992), and there may be other contexts where the difference plays a role.)

Theorem 1 (SLLN). *If $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $\mathrm{E}|X_1| < \infty$, then*

$$\frac{\sum_{i=1}^{n} X_i}{n} \to \mathrm{E}X_1 \quad a.s.$$

as $n \to \infty$.

For a proof, see Billingsley (1986, p. 290).

To apply this result we need to ensure that $EY_1 < \infty$. Let T_i be a random variable giving the completion time for task i. Let $\mathcal{A}$ be the set of all arcs. Then

$$Y \leqslant \sum_{i \in \mathcal{A}} T_i, \tag{1}$$

so that a simple sufficient condition for the strong law to hold is that all task durations have finite mean.

Under this condition we know that the estimator $\overline{Y}_n$ converges almost surely to EY as $n \to \infty$. But how accurate is it for a finite value of n? The central limit theorem (CLT) provides an answer to this question.

Theorem 2 (CLT). *If $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $EX_1^2 < \infty$, then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - EX_1\right) \Rightarrow \sigma N(0, 1)$$

as $n \to \infty$, where $\sigma^2 = \operatorname{var} X_1$, "$\Rightarrow$" denotes convergence in distribution and $N(0, 1)$ denotes a normal random variable with mean 0 and variance 1.

For a proof, see Billingsley (1986, p. 367).

To apply the CLT we need $EY_1^2 < \infty$. From (1) this follows if $ET_i^2 < \infty$ for all i. If $EY_1^2 < \infty$, then the CLT basically establishes that the error in the estimator $\overline{Y}_n$ is asymptotically normally distributed with mean 0 and variance s^2/n, where $s^2 = \operatorname{var} Y_1$, and this is the basis for obtaining confidence intervals for EY. In particular, an approximate 95% confidence interval for EY_1 is given by

$$\overline{Y}_n \pm 1.96\frac{s}{\sqrt{n}}. \tag{2}$$

Of course, s^2 must invariably be estimated. The usual estimator is the sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2.$$

The confidence interval that is reported is the same as (2) with s replaced by its sample counterpart s_n. But is the modified confidence interval then valid?

If $EY_1^2 < \infty$, then the SLLN implies that $s_n^2 \to s^2$ as $n \to \infty$ a.s. Hence, from Billingsley (1986, Exercise 29.4) we have that

$$\begin{pmatrix} n^{1/2}(\overline{Y}_n - EY) \\ s_n^2 \end{pmatrix} \Longrightarrow \begin{pmatrix} s\,N(0, 1) \\ s^2 \end{pmatrix}. \tag{3}$$

The joint convergence in (3) is a direct result of the fact that s^2 is a deterministic constant. In general, marginal convergence does not imply joint convergence.

The natural tool to apply at this point is the continuous mapping theorem. For a function $h : \mathbb{R}^d \to \mathbb{R}$, let D_h denote its set of discontinuities (in $\mathbb{R}^d$).

Theorem 3 (Continuous Mapping Theorem). *Let $(X_n\colon n \geqslant 1)$ be a sequence of $\mathbb{R}^d$-valued random variables with $X_n \Rightarrow X$ as $n \to \infty$ and let $h : \mathbb{R}^d \to \mathbb{R}$. If $\mathrm{P}(X \in D_h) = 0$, then $h(X_n) \Rightarrow h(X)$ as $n \to \infty$.*

For a proof, see Billingsley (1986, p. 391).

Define $h(x, y) = x/y^{1/2}$, and then apply the continuous mapping theorem to (3), to obtain

$$\frac{n^{1/2}(\overline{Y}_n - \mathrm{E}Y_1)}{s_n} \Rightarrow \mathrm{N}(0, 1) \tag{4}$$

as $n \to \infty$ when $s^2 > 0$, and so the confidence interval procedure outlined above is indeed valid.

The argument leading to (4) is commonly known as a *converging together argument*. It is based on the *converging together lemma*, sometimes known as *Slutsky's theorem*, which is a direct corollary of the continuous mapping theorem. We have essentially supplied a proof of this result in the argument above.

Corollary 4 (Converging Together Lemma). *If $X_n \Rightarrow X$, $U_n \Rightarrow u$ and $V_n \Rightarrow v$ as $n \to \infty$, where $X \in \mathbb{R}^d$, and u and v are constants, then*

$$U_n X_n + V_n \quad \Longrightarrow \quad uX + v$$

as $n \to \infty$.

Let us now see how these same tools can be used to estimate another performance measure. Recall that we are interested in the distribution of Y beyond its mean. One way to represent the distribution is through a density function. So when does Y have a density?

Before answering this question, we first need to understand what we mean when we say that a random variable has a density. The standard definition is based on the Radon–Nikodym theorem; see Billingsley (1986, pp. 434, 443).

Definition 1. We say that a real-valued random variable X has a density if $\mathrm{P}(X \in A) = 0$ for all Lebesgue-measurable sets A with Lebesgue measure 0. This is equivalent to saying that there exists a nonnegative function f (a density function) with the property that for all $x \in \mathbb{R}$,

$$F(x) \equiv \mathrm{P}(X \leqslant x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$

The first part of this definition may not be as familiar as the second part. Heuristically speaking, X has a density if the probability that X takes on values in "insignificant" sets is 0. The second part of the definition is perhaps more familiar. We will use the 2 definitions interchangeably in what follows. The proof of the following result demonstrates the use of the definitions.

Proposition 5. *Consider a SAN with a finite number of arcs/tasks, where the individual task durations are independent. Suppose that every path from the source to the sink contains an arc for which the corresponding task duration has a density. Then the time to complete the project has a density.*

Proof. Let P be a path from the source to the sink. The length L of the path P is the sum of the times required to traverse each arc in the path. At least one of these times has a density, and since the task durations are independent, it follows that L has a density. Let $m < \infty$ denote the number of such paths from the source to the sink, and let $L_1, \ldots, L_m$ denote the times required to traverse each path. Then $Y = \max\{L_1, \ldots, L_m\}$.

Now, the maximum of 2 random variables, X_1 and X_2 say, that have densities, also has a density. To see why, let A denote an arbitrary (measurable) subset of the real line, and let $Z = \max\{X_1, X_2\}$. Then

$$\begin{aligned} \mathrm{P}(Z \in A) &\leqslant \mathrm{P}\big(\{X_1 \in A\} \cup \{X_2 \in A\}\big) \\ &\leqslant \mathrm{P}(X_1 \in A) + \mathrm{P}(X_2 \in A), \end{aligned}$$

where the first inequality follows since $Z \in A$ implies that at least one of X_1 and X_2 must be in A, and the second is Boole's inequality. Now, we know that X_1 and X_2 have densities, so if the Lebesgue measure of A is 0, then

$$\mathrm{P}(X_1 \in A) = \mathrm{P}(X_2 \in A) = 0.$$

Hence $\mathrm{P}(Z \in A) = 0$ and so, by Definition 1, Z has a density.

We can now apply this result inductively to $Y = \max\{L_1, \ldots, L_m\}$ to conclude that Y has a density. □

Applying this result to Example 1, we see that the network completion time Y will have a density if T_{11} and T_{13}, the completion times for tasks 11 and 13, have densities. But then, how can we estimate this density?

In general, density estimation is difficult. However, the special structure in this problem allows us to use a simple device. The look-ahead density estimators developed by Henderson and Glynn (2001) are easily analyzed, and have excellent statistical properties. The mathematics of look-ahead density estimation are intimately related to those of gradient estimation via conditional Monte Carlo. See Chapter 19 for more on conditional Monte Carlo and gradient estimation, and Fu and Hu (1997) for a comprehensive account.

Let L_f and L_h be the lengths of the longest paths from the source node to nodes f and h respectively. Recall that T_{11}, T_{13} denote the (random) task

durations for tasks 11 and 13. Let F_{11}, F_{13} be the corresponding distribution functions which we assume to be continuously differentiable, and let f_{11}, f_{13} be the corresponding derivatives (and therefore densities). Then

$$\begin{aligned}
&\mathrm{P}(Y \leqslant t) \\
&\quad = \mathrm{E}\,\mathrm{P}(Y \leqslant t | L_f, L_h) \\
&\quad = \mathrm{E}\,\mathrm{P}(T_{11} \leqslant t - L_f, T_{13} \leqslant t - L_h | L_f, L_h) \\
&\quad = \mathrm{E}\big[\mathrm{P}(T_{11} \leqslant t - L_f | L_f, L_h)\,\mathrm{P}(T_{13} \leqslant t - L_h | L_f, L_h)\big] \qquad (5) \\
&\quad = \mathrm{E}\big[F_{11}(t - L_f)F_{13}(t - L_h)\big] \\
&\quad = \mathrm{E}\int_{-\infty}^{t} \frac{\mathrm{d}}{\mathrm{d}x}\big\{F_{11}(x - L_f)F_{13}(x - L_h)\big\}\,\mathrm{d}x \qquad (6) \\
&\quad = \mathrm{E}\int_{-\infty}^{t} \big\{F_{11}(x - L_f)f_{13}(x - L_h) + f_{11}(x - L_f)F_{13}(x - L_h)\big\}\,\mathrm{d}x \\
&\quad = \int_{-\infty}^{t} \mathrm{E}\big\{F_{11}(x - L_f)f_{13}(x - L_h) + f_{11}(x - L_f)F_{13}(x - L_h)\big\}\,\mathrm{d}x. \qquad (7)
\end{aligned}$$

Equality (5) follows since all of the task durations are independent, (6) is just the fundamental theorem of calculus and (7) follows since the integrand is nonnegative.

Thus, we can conclude that Y has a density f say, where

$$f(x) = \mathrm{E}\big[F_{11}(x - L_f)f_{13}(x - L_h) + f_{11}(x - L_f)F_{13}(x - L_h)\big]. \qquad (8)$$

(See Avramidis and Wilson, 1996, Section 4.1, for a related discussion.)

The expression (8) has an intuitive interpretation. The first term in (8) is related to the probability that the longest path from the source to the sink through node f has length at most x and at the same time, the longest path from the source to the sink through node h is exactly of length x. The second term can be interpreted similarly.

The expression (8) immediately suggests a density estimator for f. We generate i.i.d. replicates $L_f(i)$, $L_h(i)$ for $i = 1, \ldots, n$, and estimate $f(x)$ by

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \Gamma(i; x),$$

where

$$\begin{aligned}
\Gamma(i; x) = {} & F_{11}\big(x - L_f(i)\big)f_{13}\big(x - L_h(i)\big) \\
& + f_{11}\big(x - L_f(i)\big)F_{13}\big(x - L_h(i)\big).
\end{aligned}$$

Applying the SLLN, we see that $f_n(x) \to f(x)$ a.s., for all x for which (8) is finite. The set of values x for which this does not hold has Lebesgue measure 0,

and so the convergence of f_n to f occurs for almost all x. This is all that can be expected, because densities are only defined up to a set of Lebesgue measure 0.

We can apply the CLT to report confidence intervals for $f(x)$ exactly as in the case of estimating EY if $E[I^2(1; x)] < \infty$.

To this point we have established pointwise consistency of the estimator f_n of f. One typically estimates the density to get a sense of what the *entire* density looks like, and not just the density at a point. To this end, we might attempt to establish *global* measures of convergence such as uniform (in x) convergence. To do so here would carry us too far afield. See Henderson and Glynn (2001) for such results for look-ahead density estimation in the Markov chain context.

The look-ahead density estimator described above has very appealing and easily-derived asymptotic properties. These attractive properties are a result of carefully investigating the model to identify exploitable properties. One might ask if this somewhat specialized density estimation technique can be complemented by a more general-purpose approach that does not require as much tailoring to specific applications. The field of *nonparametric functional estimation* encompasses several such approaches; see, e.g., Prakasa Rao (1983). Wand and Jones (1995) is a very readable introduction to the field of *kernel density estimation*, which is also discussed in Chapter 8. We will not go into this area in any detail because the mathematical techniques used to analyze kernel density estimators are beyond the scope of this chapter.

This section has introduced several mathematical techniques that are repeatedly used in simulation analysis. The SLLN and CLT need no introduction. The continuous mapping theorem and converging together lemma are not as well known, but are also ubiquitous in simulation analysis. Conditional Monte Carlo can also be viewed as a variance reduction technique; see Chapter 10.

3 A model of ambulance operations

We now describe a very simple model that will serve as a vehicle for the concepts to follow. The purpose of the example is simplicity, and certainly not realism, although with a few straightforward extensions, the model could be considered to be quite practical.

Suppose that a single ambulance serves calls in a square region. By translating and rescaling units, we may assume that the square is centered at the origin, with lower left-hand corner at $(-1/2, -1/2)$ and upper right-hand corner at $(1/2, 1/2)$. The combined hospital/ambulance base is located at the origin.

Calls arrive (in time) according to a homogeneous Poisson process with rate λ calls per hour. The location of a call is independent of the arrival process, and uniformly distributed over the square. To serve a call, the ambulance travels at unit speed in a Manhattan fashion (i.e., at any given time, movement is restricted to lie only in the x direction or the y direction) from its present location to the location of the call. For definiteness we assume that travel in the y direction is completed before travel in the x direction. A random amount of

time, independent of all else, is then spent at the scene treating the patient and successive scene times are i.i.d. For definiteness we assume that scene times are gamma distributed (see Law and Kelton, 2000, p. 301, for details on this distribution). After the scene time is complete, and independent of all else, with probability p the ambulance is required to transport and admit the patient to the hospital. Hospital admission occurs instantaneously once the ambulance reaches the hospital. If the patient does not require transport to the hospital then the ambulance is immediately freed for other work. It then returns to the hospital/base. If a call requires service before the free ambulance reaches the base, then the ambulance responds to the call from its current location.

4 Finite-horizon performance

In this section, we assume that the ambulance only receives calls from (say) 7 a.m. until 11 p.m. each day. At 11 p.m., the ambulance completes the call that it is currently serving (if any) and returns to base. We will further assume that if the ambulance is engaged with a call when another call is received, then some outside agency, such as another emergency service, handles the other call. Finally, we assume that the random variables associated with each day are independent of those for all other days.

We will be primarily concerned with two performance measures.

- α The long-run fraction of calls attended by the ambulance.
- r The conditional density of the response time to a call given that the ambulance attends the call.

The utilization, or fraction of time that the ambulance is busy, is also of interest but the performance measures α and r are sufficient for our purposes.

We first consider α, the long-run fraction of calls attended by the ambulance. Let N_i denote the total number of calls received on day i, and for $j = 1, \ldots, N_i$, let A_{ij} be 1 if the ambulance is available when the jth call arrives on day i and 0 otherwise. Then the number of calls A_i attended by the ambulance on day i is $\sum_{j=1}^{N_i} A_{ij}$. After n days, the fraction of calls attended by the ambulance is given by

$$\frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} N_i}. \tag{9}$$

Dividing both the numerator and denominator of (9) by n, and applying the SLLN separately to both the numerator and denominator, we see that

$$\frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} N_i} \to \alpha = \frac{\mathrm{E}A_1}{\mathrm{E}N_1} \quad \text{a.s.}$$

as $n \to \infty$. But $\mathrm{E}N_1 = 16\lambda$, so we can estimate α by

$$\alpha_n = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{16\lambda}.$$

The SLLN establishes that α_n is a consistent estimator of α, and the CLT allows us to construct confidence intervals for α based on α_n.

Notice that α is defined as $\mathrm{E}A_1/\mathrm{E}N_1$ and not as $\mathrm{E}(A_1/N_1)$. The latter quantity is not really defined, since $\mathrm{P}(N_1 = 0) > 0$. Even if we were to define A_1/N_1 to be, say, 1 on the event $N_1 = 0$, the latter quantity is not our desired performance measure. Observe that A_i/N_i ($\mathrm{E}(A_1/N_1)$) gives the actual (expected) fraction of calls on day i that are attended by the ambulance. This expected fraction weights days equally, irrespective of the number of calls received. In contrast, the quantity $\mathrm{E}A_1/\mathrm{E}N_1$ weights days by the number of calls that are received on the day, and should be preferred.

We now develop a look-ahead density estimator for r, the conditional density of the response time given that the ambulance responds to a call. But first let us understand exactly what the density r represents. Let R_{ij} be the response time for the jth call on day i when the ambulance responds to the call ($A_{ij} = 1$) and let $R_{ij} = -1$ if the ambulance does not respond to the jth call ($A_{ij} = 0$). For $t > 0$ define

$$\begin{aligned} R(t) &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \sum_{j=1}^{N_i} A_{ij} I(R_{ij} \leqslant t)}{\sum_{i=1}^{n} A_i} \\ &= \frac{\mathrm{E} \sum_{j=1}^{N_1} A_{1j} I(R_{1j} \leqslant t)}{\mathrm{E}A_1} \end{aligned} \tag{10}$$

to be the long-run fraction of calls answered by the ambulance with response time less than or equal to t. The maximum distance the ambulance can drive is 2 units from one corner of the square to the opposite corner, so all response times are bounded by 2, and consequently $R(t)$ is 1 for all $t \geqslant 2$. We define $r(\cdot)$ to be the derivative of $R(\cdot)$ on $(0, 2)$. Of course, we have not yet established that $R(\cdot)$ is differentiable. In the process of establishing this fact we will also arrive at a look-ahead estimator for r. In essence we are again applying conditional Monte Carlo which, in this setting, is called filtering (Glasserman, 1993).

Consider the numerator $R(t)\,\mathrm{E}A_1$ of (10). We can write

$$\begin{aligned} R(t)\,\mathrm{E}A_1 &= \mathrm{E} \sum_{j=1}^{\infty} I(j \leqslant N_1) A_{1j} I(R_{1j} \leqslant t) \\ &= \sum_{j=1}^{\infty} \mathrm{E}\big[I(j \leqslant N_1) A_{1j} I(R_{1j} \leqslant t)\big], \end{aligned} \tag{11}$$

where (11) follows since the summands are nonnegative. For $i \geqslant 1$ and $j = 1, \ldots, N_i$, let B_{ij} (C_{ij}) denote the vector location of the ambulance (new call) at the time at which the jth call on day i is received. Let $d(b, c)$ denote the time required for the ambulance to travel from location b to location c. Also, let

$$g(t, b) = \mathrm{P}\big(d(b, C) \leqslant t\big) \tag{12}$$

be the probability that the travel time for the ambulance from location b to the random call location C is less than or equal to t. Observe that on the event $A_{ij} = 1$ (the ambulance answers the jth call on day i),

$$\mathrm{P}(R_{ij} \leqslant t | A_{ij}, B_{ij}) = g(t, B_{ij}).$$

So we see that

$$\begin{aligned}
&\mathrm{E}\big[I(j \leqslant N_1) A_{1j} I(R_{1j} \leqslant t)\big] \\
&\quad = \mathrm{E}\big\{\mathrm{E}\big[I(j \leqslant N_1) A_{1j} I(R_{1j} \leqslant t) | A_{1j}, B_{1j}\big]\big\} \\
&\quad = \mathrm{E}\big\{I(j \leqslant N_1) A_{1j} g(t, B_{1j})\big\}.
\end{aligned} \tag{13}$$

The final step (13) requires some care, but we omit the details.

Combining (13) with (11) we find that

$$R(t) = \frac{\mathrm{E} \sum_{j=1}^{N_1} A_{1j} g(t, B_{1j})}{\mathrm{E} A_1}. \tag{14}$$

We now wish to differentiate both sides of (14). For each fixed b, $g(\cdot, b)$ is continuously differentiable in t with bounded derivative. To see why, notice that $g(t, b)$ is the area of the intersection of the unit square with a diamond (recall that the ambulance travels in Manhattan fashion) centered at b with "radius" t; see Figure 2. For each fixed b this is a piecewise quadratic with continuous derivative at the breakpoints. So define

$$f(t, b) = \frac{\partial g(t, b)}{\partial t}. \tag{15}$$

For each fixed b, $f(t, b)$ is piecewise linear and continuous in t, as discussed in the caption of Figure 2.

We now see that

$$\begin{aligned}
r(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathrm{E} \sum_{j=1}^{N_1} A_{1j} g(t, B_{1j})}{\mathrm{E} A_1} \\
&= \frac{\mathrm{E} \frac{\mathrm{d}}{\mathrm{d}t} \sum_{j=1}^{N_1} A_{1j} g(t, B_{1j})}{\mathrm{E} A_1} \\
&= \frac{\mathrm{E} \sum_{j=1}^{N_1} A_{1j} \frac{\mathrm{d}}{\mathrm{d}t} g(t, B_{1j})}{\mathrm{E} A_1}
\end{aligned} \tag{16}$$

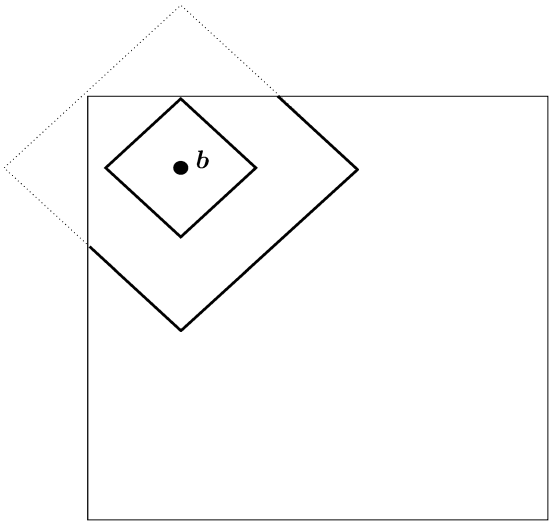

Fig. 2. Each diamond represents the set of points that are a fixed L_1 distance from the point b. The value $f(t, b)$ is proportional, for each fixed t, to the length of the (solid) line segments of the appropriate diamond that fall within the unit square. The value $g(t, b)$ gives the area of the intersection of the unit square with the appropriate diamond, and can also be viewed as an integral of f. Since f is continuous and piecewise linear, it follows that g is continuously differentiable and piecewise quadratic.

$$= \frac{\mathrm{E}\sum_{j=1}^{N_1} A_{1j} f(t, B_{1j})}{\mathrm{E}A_1}$$
$$= \frac{\mathrm{E}Y_1(t)}{\mathrm{E}A_1}, \tag{17}$$

where

$$Y_i(t) \triangleq \sum_{j=1}^{N_i} A_{ij} f(t, B_{ij}).$$

Of course, we need to justify the interchange of expectation and derivative in (16). We use the following result, which is stated in Glasserman (1991, p. 15) and proved in Dieudonné (1960, Section 8.5).

Theorem 6 (Generalized Mean-Value Theorem). *Let h be a continuous real-valued function on the closed interval $[a, b]$ which is differentiable everywhere except possibly on a set D of at most countably many points. Then for all x and $x + \delta$ in $[a, b]$,*

$$\left|\frac{h(x+\delta) - h(x)}{\delta}\right| \leqslant \sup_{y \in [a,b] \setminus D} \left|h'(y)\right|.$$

To justify (16) notice that $g(t, B_{1j})$, when viewed as a function of t, is continuously differentiable with derivative $f(t, B_{1j})$. Now, $0 \leqslant f(\cdot, \cdot) \leqslant 2$. Hence the generalized mean value theorem together with the dominated convergence theorem allow us to conclude that

$$\begin{aligned}
&\mathrm{E}\,\frac{\mathrm{d}}{\mathrm{d}t}\sum_{j=1}^{N_1} A_{1j}g(t, B_{1j}) \\
&\quad= \mathrm{E}\lim_{\delta\to 0}\frac{\sum_{j=1}^{N_1} A_{1j}g(t+\delta, B_{1j}) - \sum_{j=1}^{N_1} A_{1j}g(t, B_{1j})}{\delta} \\
&\quad= \lim_{\delta\to 0}\mathrm{E}\frac{\sum_{j=1}^{N_1} A_{1j}g(t+\delta, B_{1j}) - \sum_{j=1}^{N_1} A_{1j}g(t, B_{1j})}{\delta} \\
&\quad= \frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{E}\sum_{j=1}^{N_1} A_{1j}g(t, B_{1j}),
\end{aligned}$$

and the justification is complete.

Hence, the expression (17) rigorously defines $r(t)$. We now also have a means for estimating $r(t)$ using the estimator

$$r_n(t) = \frac{\sum_{i=1}^{n} Y_i(t)}{\sum_{i=1}^{n} A_i}.$$

So how can we assess the accuracy of the estimator $r_n(t)$? Certainly, the standard central limit theorem cannot be applied, because $r_n(t)$ is a *ratio* of sample means of i.i.d. observations. We first consider a strongly related question, and then return to the problem at hand.

Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with finite mean $\mu = \mathrm{E}X_1$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ denote the sample mean. If the real-valued function h is continuous at μ, it follows that $h(\overline{X}_n) \to h(\mu)$ a.s. as $n \to \infty$. So how does the error $h(\overline{X}_n) - h(\mu)$ behave, for large n? For large n, $\overline{X}_n$ will be very close to μ, and so the asymptotic behavior of the error should depend only on the local behavior of h near μ. Indeed, if h is appropriately differentiable, then Taylor's theorem implies that

$$h(\overline{X}_n) - h(\mu) \approx h'(\mu)(\overline{X}_n - \mu),$$

and so if X_1 has finite variance, then

$$\begin{aligned}
n^{1/2}\big(h(\overline{X}_n) - h(\mu)\big) &\approx h'(\mu)n^{1/2}(\overline{X}_n - \mu) \\
&\Rightarrow \eta\mathrm{N}(0, 1)
\end{aligned}$$

as $n \to \infty$, where $\eta^2 = h'(\mu)^2 \operatorname{var} X_1$.

This argument can be made rigorous and generalized to higher dimensions to obtain the following result, sometimes referred to as the delta method.

Theorem 7. *Suppose that* $(X_n\colon\ n \geqslant 1)$ *is an i.i.d. sequence of* $\mathbb{R}^d$*-valued random variables with* $\mathrm{E}\|X_1\|_2^2 < \infty$*. Let* $\mu = \mathrm{E}X_1$ *denote the mean vector and* $\Lambda = \operatorname{cov} X_1$ *denote the covariance matrix. Let* $\overline{X}_n$ *denote the sample mean of* $X_1, \ldots, X_n$*. If* $h : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable in a neighborhood of* μ *with nonzero gradient* $g = \nabla h(\mu)$ *at* μ*, then*

$$n^{1/2}\big(h(\overline{X}_n) - h(\mu)\big) \Rightarrow \sigma \mathrm{N}(0, 1)$$

as $n \to \infty$*, where* $\sigma^2 = g^\top \Lambda g$*.*

For a proof, see Serfling (1980, p. 122).

To apply this result in our context, let

$$X_i = \big(Y_i(t),\, A_i\big),$$

and define $h(y, a) = y/a$. Theorem 7 then implies that

$$n^{1/2}\big(r_n(t) - r(t)\big) \Rightarrow \sigma(t)\mathrm{N}(0, 1),$$

where

$$\sigma^2(t) = \frac{\mathrm{E}(Y_1(t) - r(t)A_1)^2}{(\mathrm{E}A_1)^2}.$$

Using the SLLN, one can show that $\sigma^2(t)$ can be consistently estimated by

$$s_n^2(t) = \frac{n^{-1}\sum_{i=1}^n (Y_i(t) - r_n(t)A_i)^2}{(n^{-1}\sum_{i=1}^n A_i)^2},$$

and the same continuous mapping argument discussed earlier establishes that

$$r_n(t) \pm \frac{1.96 s_n(t)}{\sqrt{n}}$$

is an approximate 95% confidence interval for $r(t)$.

The estimator $r_n(t)$, being a ratio estimator, is biased. Taylor's theorem can be used to examine this bias. Reverting to our one-dimensional digression for the moment, Taylor's theorem implies that

$$h(\overline{X}_n) - h(\mu) \approx h'(\mu)(\overline{X}_n - \mu) + \frac{1}{2}h''(\mu)(\overline{X}_n - \mu)^2.$$

Taking expectations, we find that

$$\mathrm{E}h(\overline{X}_n) - h(\mu) \approx \frac{1}{2}h''(\mu)\frac{\operatorname{var} X_1}{n},$$

i.e., we have an explicit expression for the asymptotic bias. This argument can be made rigorous and generalized to higher dimensions.

Theorem 8. *Suppose that* $(X_n\colon n \geqslant 1)$ *is an i.i.d. sequence of* $\mathbb{R}^d$*-valued random variables with* $\mathrm{E}\|X_1\|_2^4 < \infty$*. Let* $\mu = \mathrm{E}X_1$ *denote the mean and* $\Lambda = \operatorname{cov} X_1$ *denote the covariance matrix. Let* $\overline{X}_n$ *denote the sample mean of* $X_1, \ldots, X_n$*. If* $h : \mathbb{R}^d \to \mathbb{R}$ *is such that* $h(\overline{X}_n)$ *is bounded for all n with probability* 1*, and twice continuously differentiable in a neighborhood of* μ*, then*

$$n\big(\mathrm{E}h(\overline{X}_n) - h(\mu)\big) \to \frac{1}{2}\sum_{i,j=1}^{d} \nabla^2 h(\mu)_{ij}\Lambda_{ij}$$

as $n \to \infty$*.*

The proof is a slight modification of Glynn and Heidelberger (1990, Theorem 7).

One of the hypotheses of Theorem 8 is that $h(\overline{X}_n)$ is bounded for all n a.s. This regularity condition is used in the proof to show that a certain sequence of random variables is uniformly integrable, so that one can pass expectations through to the limit. We did not need this condition in Theorem 7 because that result is about convergence of *distributions* and not about convergence of *expectations*. The boundedness condition can be replaced by less-stringent conditions like moment conditions.

We would like to apply Theorem 8 to the estimator $r_n(t)$. To that end, define $h(y, a) = y/a$. The only condition that is not obviously satisfied is the one that requires that $h(\overline{X}_n)$ be bounded for all n with probability 1. But the function f is bounded by 2, and so $h(\overline{X}_n(t)) = r_n(t)$ is also bounded by 2. We have therefore established that the bias in the estimator $r_n(t)$ is of the order n^{-1}.

It is reasonable to ask whether this bias is sufficient to noticeably affect the performance of the confidence intervals produced earlier for a given runlength n. Recall that the widths of the confidence intervals are of the order $n^{-1/2}$. Thus, the bias decreases at a (much) faster asymptotic rate than the width of the confidence intervals, and so when runlengths are sufficiently large it is reasonable to neglect bias.

In this section we used the delta method, which is very well known and sees a great deal of use. It has been applied in, for example, the regenerative method of steady-state simulation analysis, e.g., Shedler (1987), quantifying the impact of input uncertainty on simulation outputs, e.g., Cheng and Holland (1998) and in analyzing transformation-based simulation metamodels (Irizarry et al., 2003). We also used conditional Monte Carlo to enable the taking of derivatives.

5 Steady-state simulation

We now turn to useful mathematical techniques and results for steady-state simulation analysis. To this end we modify the assumptions of the previous

section on the dynamics of the ambulance model. In addition to the assumptions given in Section 3, we assume that the ambulance operates 24 hours a day, 7 days a week. Furthermore, calls that arrive while the ambulance is busy are queued and answered in first-in first-out order. (The assumption that calls queue is not needed to create an interesting steady-state model. We introduce it to allow a nontrivial discussion about stability.) Once the current call is complete, either at the hospital if the patient required hospitalization or at the scene of the call if not, the ambulance then responds to the next call if any. (Recall that a call is completed either at the scene, with probability $1 - p$, or when the ambulance drops the patient off at the hospital, with probability p.) If the patient does not require hospitalization and no calls are queued, then the ambulance returns to the hospital/ambulance base, but can respond to newly-arriving calls before it reaches the hospital/ambulance base.

We are still interested in ambulance utilization and the distribution of response times. But the ambulance now handles all incoming calls, and so the fraction of calls answered by the ambulance is no longer relevant. Our performance measures are the following.

- ρ The long-run utilization of the ambulance, i.e., the percentage of time that the ambulance is occupied with a call. The ambulance is not considered to be occupied when returning to the hospital/base without a patient.
- r The long-run density of the response time to a call.

Notice that ρ is a deterministic constant, while r is a density function.

In the previous section we attempted to rigorously define the suggested performance measures, and also to derive asymptotic results that lay at the heart of confidence interval methodology for estimating them. We will proceed in a similar fashion in this section. Both performance measures involve the term "long-run". In order that such long-run measures exist, it is first necessary that the ambulance model be stable, in the sense that calls do not "pile up" indefinitely. In order to be able to make statements about the stability, or lack thereof, of the model it is first necessary to define an appropriate stochastic process from which our performance measures can be derived. Statements about the stability of the model really relate to the stability of the stochastic process.

There are typically a host of stochastic processes that may be defined from the elements of a simulation. The choice of stochastic process depends partly on the performance measures. Given that our measures are related to response time, it is natural to consider a stochastic process that yields information on response times. Furthermore, for mathematical convenience, it is often helpful to ensure that one's stochastic process is Markov.

For $n \geqslant 1$, let T_n denote the time at which the nth call is received, and define $T_0 = 0$. For $n \geqslant 1$, let W_n be the *residual workload* of the ambulance at time T_n+, i.e., just after the nth call is received. By residual workload at some time t, we mean the amount of time required for the ambulance to complete

any current call, along with calls that might also be queued at time t. We assume that the ambulance is idle at the hospital at time $T_0 = 0$, so that $W_0 = 0$.

Unfortunately, $(W_n\colon n \geqslant 0)$ is not a Markov process, because the response time for a future call, and hence the workload, depends on the location of the ambulance when the ambulance clears the previous workload. (Here, the ambulance may clear the previous workload either at the location of a call if hospitalization is unnecessary, or at the hospital/base if hospitalization is necessary.) So if we also keep track of the location $\beta_n = (\beta_n(1), \beta_n(2))$ of the ambulance at the instant at which the workload W_n is first cleared, then the resulting process $Z = (Z_n\colon n \geqslant 0)$ is Markov, where $Z_n = (W_n, \beta_n)$.

The process Z is a general state space Markov chain, and evolves on the state space

$$S = [0, \infty) \times \left[-\tfrac{1}{2}, \tfrac{1}{2}\right]^2.$$

The first step in ensuring that our long-run performance measures are defined is to establish that Z exhibits some form of positive recurrence. One way to achieve this is to verify that the chain Z satisfies the following condition, which will be explained shortly.

To avoid confusion between general results and those for our particular model, we will state general results in terms of a Markov chain $X = (X_n\colon n \geqslant 0)$ evolving on a state space $\mathcal{S}$.

The First Lyapunov Condition (FLC). There exists a nonempty $B \subseteq \mathcal{S}$, positive scalars $a < 1$, b and δ, an integer $m \geqslant 1$, a probability distribution φ on $\mathcal{S}$, and a function $V : \mathcal{S} \to [1, \infty)$ such that

(1) $\mathrm{P}(X_m \in \cdot | X_0 = z) \geqslant \delta\varphi(\cdot)$ for all $z \in B$, and
(2) $\mathrm{E}(V(X_1)|X_0 = z) \leqslant aV(z) + bI(z \in B)$ for all $z \in \mathcal{S}$.

The FLC (sometimes called a Foster–Lyapunov condition) is a stronger Lyapunov condition than we really require, but it simplifies the presentation. The function V is called a Lyapunov (think of energy) function. The second requirement basically states that when the chain X lies outside of the set B, the energy in the system tends to decrease, and when the chain lies inside B, the energy in the system cannot become too big on the next step. This condition implies that the set B gets hit infinitely often. Of course, if one takes $B = \mathcal{S}$, the entire state space, then this requirement is trivially satisfied. The first condition is needed to ensure that the set B is not too "big".

In any case, the point is that if a chain X satisfies the FLC, then X is appropriately positive recurrent, and in particular has a unique stationary probability distribution. In fact, the FLC is essentially equivalent to a strong form of ergodicity. It is therefore reasonable (but nonstandard) to *define* a chain as being appropriately ergodic if the FLC holds.

Definition 2. We say that a discrete time Markov chain X is V-uniformly ergodic if it satisfies the Lyapunov condition and is aperiodic.

The aperiodicity condition is not strictly necessary for much of what follows, but we impose it to be consistent with the term "ergodic". For more on V-uniform ergodicity and its relationship to the FLC see Meyn and Tweedie (1993, Chapter 16).

Does our chain Z satisfy the FLC? The answer is yes, and it is instructive to go through a proof. However, on a first reading one may skip the following development up to the statement of Proposition 9 without loss of continuity.

For many systems, the function V may be taken to be $e^{\gamma v}$, where v is some measure of the work in the system. In fact, as we now show, one may take $V(w, b) = e^{\gamma w}$ for some yet to be determined constant $\gamma > 0$.

Consider what happens on a single transition of the chain Z starting from the point $Z_n = (w, b)$, where $n \geqslant 0$. The workload decreases at unit rate, at least until it hits 0, until the arrival of the next call over an interval of length $\tau_{n+1} = T_{n+1} - T_n$. At time T_{n+1} a new call arrives at location C_{n+1} and adds some work to the workload. In particular, there will be some travel time η_{n+1} to the scene of the call, some time U_{n+1} spent at the scene, and then potentially some travel time ξ_{n+1} to transport the patient to the hospital. If the patient requires transport to the hospital then $\beta_{n+1} = (0, 0)$, which is the location of the hospital. If not, then $\beta_{n+1} = C_{n+1}$, which is the location of the call, and $\xi_{n+1} = 0$. If hospitalization is not necessary and no calls are queued when the ambulance completes service at the scene, then the ambulance returns to the hospital/base, but this travel time is not counted as workload because the ambulance is free to respond to a new call, albeit not necessarily from the hospital/base.

So for $n \geqslant 0$, the new workload W_{n+1} is given by $W_{n+1} = [W_n - \tau_{n+1}]^+ + Q_{n+1}$, where $[x]^+ = \max\{x, 0\}$, and $Q_n = \eta_n + U_n + \xi_n$. Recall that we assume that the scene times $(U_n\colon\ n \geqslant 1)$ are i.i.d. gamma-distributed random variables, and are independent of all other quantities. We assume that the call location sequence $(C_n\colon\ n \geqslant 1)$ is i.i.d. and independent of all other quantities.

Equipped with this Lindley-type recursion for the workload, we can now attempt to identify conditions under which the Lyapunov condition will hold. We use the fact that $Q_1 \leqslant 3 + U_1$ because $\eta_1 \leqslant 2$ and $\xi_1 \leqslant 1$ (recall that the ambulance travels distances as measured by the Manhattan metric). If $z = (w, b)$, then $\mathrm{E}[V(Z_1)|Z_0 = z]$ is given by

$$
\begin{aligned}
\mathrm{E}e^{\gamma([w-\tau_1]^+ + Q_1)} &\leqslant \mathrm{E}\left[e^{\gamma[w-\tau_1]^+} e^{\gamma(3+U_1)}\right] \\
&= \mathrm{E}e^{\gamma[w-\tau_1]^+}\, \mathrm{E}e^{\gamma(3+U_1)} \\
&\leqslant \left[\mathrm{E}e^{\gamma(w-\tau_1)} + \mathrm{P}(w - \tau_1 < 0)\right] \mathrm{E}e^{\gamma(3+U_1)} \\
&= e^{\gamma w}\left[\mathrm{E}e^{-\gamma\tau_1} + e^{-(\lambda+\gamma)w}\right] \mathrm{E}e^{\gamma(3+U_1)} \qquad (18) \\
&= e^{\gamma w}\left[1 + \frac{\lambda+\gamma}{\lambda} e^{-(\lambda+\gamma)w}\right] \mathrm{E}e^{\gamma(3+U_1-\tau_1)} \qquad (19) \\
&= V(z)\left[1 + \frac{\lambda+\gamma}{\lambda} e^{-(\lambda+\gamma)w}\right] \phi(\gamma), \qquad (20)
\end{aligned}
$$

where ϕ is the moment generating function of $3 + U_1 - \tau_1$. Equation (18) uses the fact that $\mathrm{P}(\tau_1 > w) = \mathrm{e}^{-\lambda w}$, while (19) follows since $\mathrm{E}\mathrm{e}^{-\gamma\tau_1} = \lambda/(\lambda+\gamma)$ (when $\gamma > -\lambda$).

Since $\mathrm{E}\mathrm{e}^{tU_1}$ is finite in a neighborhood of 0, i.e., U_1 has a moment generating function defined near 0, we have that $\phi(0) = 1$, and

$$\phi'(0) = \mathrm{E}(U_1 + 3 - \tau_1).$$

So if $\mathrm{E}U_1 + 3 < \mathrm{E}\tau_1$, then $\phi'(0) < 0$, and so $\phi(t) < 1$ for $t > 0$ in some neighborhood of 0. So fix $\gamma > 0$ so that $\phi(\gamma) < 1$.

Now, there is some $K > 0$ such that if $w > K$, then

$$\left[1 + \frac{\lambda+\gamma}{\lambda}\mathrm{e}^{-(\lambda+\gamma)w}\right]\phi(\gamma) < 1. \tag{21}$$

Furthermore, for $w \leqslant K$, we have that

$$\mathrm{E}\big[V(Z_1)|Z_0 = z\big] \leqslant \mathrm{E}\mathrm{e}^{\gamma(K+3+U_1)} < \infty. \tag{22}$$

Thus, if we take $B = [0, K] \times [-\frac{1}{2}, \frac{1}{2}]^2$, then it follows from (20)–(22) that the second requirement in the FLC is met.

It remains to check the first requirement. Suppose that $Z_n = (w, b) \in B$ so that the current workload $w \leqslant K$. If the time τ_{n+1} till the next call is large enough, then irrespective of whether the nth patient requires transport to the hospital or not, the ambulance will have reached the hospital and be available to respond to a new call by the time the $(n+1)$st call arrives. So if $\tau_{n+1} > K+1$, the $(n+1)$st call will be served immediately by the ambulance from the base. In fact, the chain regenerates at such times. Let

$$\delta = \mathrm{P}(\tau_1 > K+1) = \mathrm{e}^{-\lambda(K+1)}$$

and φ denote the distribution of $Z_1 = (W_1, B_1)$ assuming that just before time T_1 the ambulance is free and located at the hospital. Then we have that for all $z \in B$,

$$\mathrm{P}(z, \cdot) \geqslant \delta\varphi(\cdot),$$

and the first requirement in the FLC is satisfied.

We have established that Z satisfies the FLC. It is straightforward to show that Z is aperiodic, and so we arrive at the following result.

Proposition 9. *If* $\mathrm{E}U_1 + 3 < \mathrm{E}\tau_1$, *then the chain* Z *is* V*-uniformly ergodic*, *where* $V(w, b) = \mathrm{e}^{\gamma w}$ *for some* $\gamma > 0$.

The stability condition

$$\mathrm{E}U_1 + 3 < \mathrm{E}\tau_1$$

has an appealing interpretation. The left-hand side of the inequality gives an upper bound on the expected amount of work (travel time to the scene + time

at the scene + travel time from the scene to the hospital) brought in by an arriving call. We require this to be smaller than the expected amount of time that the ambulance has available between calls to deal with the work. This condition can be weakened by being more careful about defining how much work each call brings to the system, but this is not something that we will pursue further.

The main point is that Proposition 9 gives *easily-verifiable* conditions under which the system is stable. While it may have appeared somewhat difficult to verify the Lyapunov condition, the argument used is actually quite straightforward once one picks an appropriate function V. The difficult part (in general), and the part where one's insight into the underlying process plays a key role, is in choosing an appropriate V. Thankfully, we will see that the payoff from verifying the Lyapunov condition is certainly worth the effort. Based on this result, we can now define our performance measures rigorously, and also construct estimators that are consistent and satisfy central limit theorems.

As in Section 4, the rigorous definition of our performance measures is based on the strong law of large numbers. For simplicity, we state this theorem under stronger hypotheses than are really necessary. Let E_ν denote expectation for the path space of a Markov chain with initial distribution ν.

Theorem 10 (MCSLLN). *Let X be a V-uniformly ergodic Markov chain on state space $\mathcal{S}$ with stationary probability distribution π. Let $h : \mathcal{S} \to \mathbb{R}$ be a real-valued function on $\mathcal{S}$. If $\pi|h| = \mathrm{E}_\pi|h(X_0)| = \int_{\mathcal{S}} |h(x)|\pi(\mathrm{d}x) < \infty$, then*

$$\frac{1}{n}\sum_{i=0}^{n-1} h(X_i) \to \pi h \quad a.s.$$

as $n \to \infty$.

For a proof, see Meyn and Tweedie (1993, Theorem 17.0.1).

Assuming V-uniform ergodicity we see that if h is bounded, then the MCSLLN holds. This will be sufficient for our purposes in this section. However, sometimes one is also interested in unbounded h. So long as $|h(z)| \leqslant cV(z)$ for some $c > 0$ and all z, then $\pi|h| < \infty$; see Meyn and Tweedie (1993, Theorem 14.3.7).

We turn now to the performance measure ρ, the long-run utilization of the ambulance. The actual utilization of the ambulance over the time interval $[0, T_n)$, i.e., up until the time of the nth arrival is

$$\frac{n^{-1}\sum_{i=0}^{n-1} \min\{W_i, \tau_{i+1}\}}{n^{-1}\sum_{i=1}^{n} \tau_i}. \tag{23}$$

Now, the SLLN for i.i.d. random variables implies that the denominator converges to λ^{-1}. We would like to apply the MCSLLN to the numerator, but it is not yet in an appropriate form. There are several ways to proceed here. One

way is to enlarge the state space of the chain, and we will demonstrate this method shortly in analyzing the density r. Another approach is to apply filtering; see Glasserman (1993), which can be viewed as a variant of conditional Monte Carlo. We have that

$$\begin{aligned}\mathrm{E}\min\{w, \tau_1\} &= w\,\mathrm{P}(\tau_1 > w) + \mathrm{E}\tau_1 I(\tau_1 \leqslant w)\\ &= \lambda^{-1}\bigl(1 - \mathrm{e}^{-\lambda w}\bigr),\end{aligned}$$

and so we replace (23) by

$$\rho_n = \frac{1}{n}\sum_{i=0}^{n-1}\bigl(1 - \mathrm{e}^{-\lambda W_i}\bigr). \tag{24}$$

Notice that ρ_n is in exactly the form that we need to apply the MCSLLN, with $h(w, b) = 1 - \mathrm{e}^{-\lambda w}$ which is bounded, and so we find that

$$\rho_n \to \rho \quad \text{a.s.}$$

as $n \to \infty$. This then is a rigorous definition of ρ, and also a proof that the estimator ρ_n is strongly consistent.

Let us turn now to r, the steady-state density of the response time to a call. For $n \geqslant 0$, the response time L_{n+1} to the call arriving at time T_{n+1} is the sum of the workload $[W_n - \tau_{n+1}]^+$ just before the arrival of the call and the time η_{n+1} for the ambulance to travel to the location of the new call. The travel time to the new call is given by the distance $d(B_{n+1}, C_{n+1})$ between the call location C_{n+1} and the location B_{n+1} of the ambulance at the time when the ambulance responds to the $(n + 1)$st call. Now, $B_{n+1} = \beta_n$ if the $(n + 1)$st call arrives before the previous workload is cleared, i.e., $W_n \geqslant \tau_{n+1}$. If $W_n < \tau_{n+1}$ then the new call arrives after the ambulance completes the previous workload, so the ambulance may be on its way to the hospital, or at the hospital, when the $(n+1)$st call arrives. In any case, the location B_{n+1} is a deterministic function of $Z_n = (W_n, \beta_n)$ and τ_{n+1}. So the response time L_{n+1} depends not only on Z_n, but also on τ_{n+1} and C_{n+1}.

This dependence of the response time on additional quantities beyond those in the state space of our chain causes some difficulties in our analysis. We could again apply filtering, but let us consider an alternative approach. We expand the state space of the Markov chain so that it is "sufficiently rich" to supply all of the needed information.

Define $\widetilde{Z} = (\widetilde{Z}_n\colon\ n \geqslant 0)$ where, for $n \geqslant 0$, $\widetilde{Z}_n = (W_n, \beta_n, \tau_{n+1}, C_{n+1})$. Using techniques that are very similar to those used for the chain Z, we can show that $\widetilde{Z}$ is a $\widetilde{V}$-uniformly ergodic chain on the state space

$$\widetilde{S} = [0, \infty) \times \left[-\tfrac{1}{2}, \tfrac{1}{2}\right]^2 \times [0, \infty) \times \left[-\tfrac{1}{2}, \tfrac{1}{2}\right]^2,$$

where

$$\widetilde{V}(w, b, t, c) = \mathrm{e}^{\gamma[w-t]^+}$$

for some $\gamma > 0$.

To define the density r we first define the corresponding distribution function R, and then differentiate.

Recall that for $n \geqslant 0$, the response time is given by

$$L_{n+1} = [W_n - \tau_{n+1}]^+ + d(B_{n+1}, C_{n+1}).$$

Consider the *empirical* response time distribution function based on the first n response times

$$\frac{1}{n}\sum_{i=1}^{n} I(L_i \leqslant \cdot).$$

Notice that $I(L_i \leqslant t)$ is a deterministic and bounded function of $\widetilde{Z}_{i-1}$, so we can apply the MCSLLN to assert that

$$\frac{1}{n}\sum_{i=1}^{n} I(L_i \leqslant t) \to R(t)$$

as $n \to \infty$ a.s., for any fixed t, where

$$R(t) = \mathrm{E}_{\tilde{\pi}} I(L_1 \leqslant t).$$

Here $\tilde{\pi}$ refers to the stationary distribution of $\widetilde{Z}$.

It is not yet clear how to obtain an expression for the density r, since the indicator functions that we used to define R are not differentiable. We need to perform some sort of smoothing. We again use conditional Monte Carlo. Notice that

$$\begin{aligned} R(t) &= \mathrm{E}_{\tilde{\pi}} I(L_1 \leqslant t) \\ &= \mathrm{E}_{\tilde{\pi}} \mathrm{P}_{\tilde{\pi}}\big([W_0 - \tau_1]^+ + d(B_1, C_1) \leqslant t | W_0, \beta_0, \tau_1\big) \\ &= \mathrm{E}_{\tilde{\pi}} \mathrm{P}_{\tilde{\pi}}\big(d(B_1, C_1) \leqslant t - [W_0 - \tau_1]^+ | W_0, \beta_0, \tau_1\big) \\ &= \mathrm{E}_{\tilde{\pi}} g\big(t - [W_0 - \tau_1]^+, B_1\big), \end{aligned}$$

where the function $g(\cdot, \cdot)$ was introduced in (12). (Here we extend the definition so that $g(t, b) = 0$ for $t < 0$.) Notice that B_1 can be determined from W_0, β_0 and τ_1. So we can estimate $R(t)$ using

$$R_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} g\big(t - [W_i - \tau_{i+1}]^+, B_{i+1}\big),$$

and again the MCSLLN shows that $R_n(t) \to R(t)$ as $n \to \infty$ a.s., for each fixed t.

We can now define r, via

$$\begin{aligned} r(t) &= \frac{\mathrm{d}}{\mathrm{d}t} R(t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{E}_{\tilde{\pi}} g\big(t - [W_0 - \tau_1]^+, B_1\big) \\ &= \mathrm{E}_{\tilde{\pi}} \frac{\mathrm{d}}{\mathrm{d}t} g\big(t - [W_0 - \tau_1]^+, B_1\big) \qquad (25) \\ &= \mathrm{E}_{\tilde{\pi}} f\big(t - [W_0 - \tau_1]^+, B_1\big), \qquad (26) \end{aligned}$$

where $f(\cdot, \cdot)$ was defined in (15). Of course, we need to justify the interchange of derivative and expectation in (25). This is virtually identical to the justification given for the interchange (16), and so we omit the details.

Equation (26) defines r, and immediately suggests an estimator for $r(t)$ given by

$$r_n(t) = R_n'(t) = \frac{1}{n} \sum_{i=0}^{n-1} f\big(t - [W_i - \tau_{i+1}]^+, B_{i+1}\big).$$

The MCSLLN shows that $r_n(t) \to r(t)$ as $n \to \infty$ a.s. for each fixed t. Hence, we have rigorously defined the density r, and established that it can be consistently estimated by r_n.

We now turn to the error in the estimators. As before, error can be assessed through confidence intervals derived from a central limit theorem. In great generality, the error $n^{-1} \sum_{i=0}^{n-1} h(X_i) - \pi h$ is approximately normally distributed with mean 0 and variance σ^2/n, exactly as in the i.i.d. case. The difference here is that we are averaging dependent random variables rather than independent ones, and this difference is exhibited through the variance constant which now includes covariance terms in addition to the variance (under π) of $h(X_0)$.

For simplicity we state the Markov chain central limit theorem under stronger conditions than are strictly necessary.

Theorem 11 (MCCLT). *Suppose that the chain X is V-uniformly ergodic. Then, for any function $h : \mathcal{S} \to \mathbb{R}$ with $h^2(z) \leqslant cV(z)$ for some $c > 0$ and all z,*

$$\sqrt{n}\left(\frac{1}{n} \sum_{i=0}^{n-1} h(X_i) - \pi h\right) \Rightarrow \sigma \mathrm{N}(0, 1),$$

where π is the stationary probability distribution of X, and

$$\sigma^2 = \mathrm{var}_{\pi}\big[h(X_0)\big] + 2 \sum_{k=1}^{\infty} \mathrm{cov}_{\pi}\big[h(X_0), h(X_k)\big]. \qquad (27)$$

For a proof, see Meyn and Tweedie (1993, Theorem 17.0.1).

We immediately obtain the following result.

Proposition 12. *Under the conditions of Proposition 9,*

$$\sqrt{n}(\rho_n - \rho) \Rightarrow \sigma_\rho \mathrm{N}(0,1)$$

as $n \to \infty$, *for an appropriately defined constant* σ_ρ^2. *In addition, for all* $t > 0$,

$$\sqrt{n}\big(r_n(t) - r(t)\big) \Rightarrow \sigma(t)\mathrm{N}(0,1)$$

as $n \to \infty$, *for an appropriately defined constant* $\sigma^2(t)$.

Thus, just as in the terminating simulation case, the error in the estimators ρ_n and $r_n(t)$ is approximately normally distributed with mean 0 and variance on the order of n^{-1}.

Proposition 12 serves as a foundation for constructing confidence intervals for our performance measures. One approach is to estimate the variance constants directly using the regenerative method (see Chapter 16), which is certainly easily applied to our example. But the method of batch means is, at least currently, more widely applicable and so we instead consider this approach. See Chapter 15 for a more extensive discussion of the method of batch means.

Suppose that we have a sample path $X_0, X_1, \ldots, X_{n-1}$. Divide this sample path into l batches each of size m, where for convenience we assume that $n = ml$, so that the kth batch consists of observations $X_{(k-1)m}, \ldots, X_{km-1}$. (We use "$l$" for the number of batches instead of the traditional "b" since we have already used "b" to describe ambulance locations.) Now, for $k = 1, \ldots, l$, let M_k be the sample mean over the kth batch, i.e.,

$$M_k = \frac{1}{m} \sum_{i=(k-1)m}^{km-1} h(X_i),$$

and let $\overline{M}_l$ denote the sample mean of the l batch means $M_1, \ldots, M_l$. Finally, let

$$s_l^2 = \frac{1}{l-1} \sum_{k=1}^{l} (M_k - \overline{M}_l)^2$$

denote the sample variance of the M_k's. The method of batch means provides a confidence interval for πh of the form $\overline{M}_l \pm t s_l/\sqrt{l}$, for some constant t, and relies on the assumption that for large n, $(\overline{M}_l - \pi h)/(s_l/\sqrt{l}\,)$ is approximately t-distributed, with $l-1$ degrees of freedom.

The MCCLT above suggests that as $n \to \infty$ with l, the number of batches, held fixed, all of the batch means are asymptotically normally distributed with mean πh and variance $l\sigma^2/n$. If each of the batch means are also asymptotically independent, then a standard result (e.g., Rice, 1988, p. 173) shows that the above confidence interval methodology is valid.

A sufficient condition that supplies both the asymptotic normality and asymptotic independence is that the chain X satisfy a functional central limit theorem; see Schruben (1983) and Glynn and Iglehart (1990), from which much of the following discussion is adapted.

Definition 3. Let X be a Markov chain on state space $\mathcal{S}$, and let $h:\mathcal{S} \to \mathbb{R}$. For $0 \leqslant t \leqslant 1$, let

$$\overline{X}_n(t) = n^{-1} \sum_{k=0}^{\lfloor (n-1)t \rfloor} h(X_k)$$

and set

$$\zeta_n(t) = n^{1/2}\big(\overline{X}_n(t) - \kappa t\big)$$

for some constant κ. We say that X satisfies a functional central limit theorem (FCLT) if there exists a $\xi > 0$ such that $\zeta_n \Rightarrow \xi B$ as $n \to \infty$, where B denotes a standard Brownian motion.

Observe that if X satisfies an FCLT, then the jth batch mean M_j can be expressed as

$$\begin{aligned} M_j &= l\left[\overline{X}_n\left(\frac{j}{l}\right) - \overline{X}_n\left(\frac{j-1}{l}\right)\right] \\ &= \kappa + n^{-1/2} l\left[\zeta_n\left(\frac{j}{l}\right) - \zeta_n\left(\frac{j-1}{l}\right)\right]. \end{aligned}$$

Since the increments of Brownian motion are independent and normally distributed, the FCLT then implies that the M_j's are asymptotically independent and normally distributed with mean κ and variance $l\xi^2/n$. Thus, under an FCLT assumption, the batch means confidence interval methodology outlined above is asymptotically valid as $n \to \infty$ with l fixed.

So when can we be sure that X satisfies an FCLT? One sufficient condition is the following result.

Theorem 13. *Suppose that X is V-uniformly ergodic and $h^2(z) \leqslant cV(z)$ for all z and some $c > 0$. If the constant σ^2 defined in (27) above is positive, then X satisfies a functional central limit theorem with $\kappa = \pi h$ and $\xi^2 = \sigma^2$.*

For a proof, see Meyn and Tweedie (1993, Theorems 17.4.4, 17.5.3).

Notice that we have already established that the conditions of Theorem 13 hold for our estimators. Thus, we immediately arrive at the conclusion that the method of batch means yields asymptotically valid confidence intervals when used in conjunction with our estimators. In fact, an FCLT is sufficient to ensure that any standardized time series method (batch means with a fixed number of batches is one such method) is asymptotically valid (Schruben, 1983).

As in the terminating simulation case, the performance of confidence interval procedures may be negatively impacted by bias. The bias depends on the initial distribution, μ say, of the chain. The bias in the estimator ρ_n is $\mathrm{E}_\mu \rho_n - \rho$, with a similar expression for the bias in $r_n(t)$ for each $t > 0$.

We give the appropriate calculations for ρ, as those for r are similar. Let $h(w, b) = 1 - \mathrm{e}^{-\lambda w}$. Using a standard technique (e.g., Glynn, 1995), we see that the bias in ρ_n under initial distribution μ is

$$\begin{aligned}
&\mathrm{E}_\mu \frac{1}{n} \sum_{i=0}^{n-1} \bigl[h(Z_i) - \pi h\bigr] \\
&\quad = \frac{1}{n} \sum_{i=0}^{\infty} \bigl[\mathrm{E}_\mu h(Z_i) - \pi h\bigr] - \frac{1}{n} \sum_{i=n}^{\infty} \bigl[\mathrm{E}_\mu h(Z_i) - \pi h\bigr] \\
&\quad = \frac{\nu}{n} + \mathrm{o}\bigl(n^{-1}\bigr),
\end{aligned}$$

where

$$\nu = \sum_{i=0}^{\infty} \bigl[\mathrm{E}_\mu h(Z_i) - \pi h\bigr]$$

provided that

$$\sum_{i=0}^{\infty} \bigl|\mathrm{E}_\mu h(Z_i) - \pi h\bigr| < \infty. \tag{28}$$

So the bias in the estimator ρ_n will be of the order n^{-1} if (28) holds. This result holds in great generality.

Theorem 14. *Suppose that X is V-uniformly ergodic. Let π be the stationary probability distribution of X. If $|h(z)| \leqslant cV(z)$ for all z and some $c < \infty$, and $\mu V < \infty$, then*

$$\sum_{i=0}^{\infty} \bigl|\mathrm{E}_\mu h(X_i) - \pi h\bigr| < \infty$$

and

$$\mathrm{E}_\mu \frac{1}{n} \sum_{i=0}^{n-1} h(X_i) - \pi h = \frac{\nu}{n} + \mathrm{O}\bigl(q^n\bigr)$$

as $n \to \infty$, where $q < 1$ and

$$\nu = \sum_{i=0}^{\infty} \bigl[\mathrm{E}_\mu h(X_i) - \pi h\bigr].$$

This result is a straightforward consequence of Meyn and Tweedie (1993, Theorem 16.0.1).

We can conclude from Theorem 14 that if the initial conditions are chosen appropriately (e.g., if $\widetilde{Z}_0$ is chosen to be deterministic), then the bias of our estimators is of the order n^{-1}.

Recall that the batch means $M_1, \ldots, M_l$ are asymptotically normally distributed with variance $l\chi^2/n$. Their standard deviation is therefore of the order $n^{-1/2}$, and so the width of the batch means confidence interval is also of the order $n^{-1/2}$. The bias in the estimators is of the order n^{-1}, and so it follows that bias will not play a role for large runlengths.

5.1 Multiple ambulances

Sometimes a good choice of Lyapunov function immediately presents itself. In other cases the choice is not so clear, and the process of finding a good function becomes more of an art than a science. Here we consider the case where multiple ambulances operate in the unit square from potentially different bases. We will again look for a good choice of Lyapunov function. Some natural choices do not work, at least at first sight. However, one of those choices *does* work if we use an extension of the FLC.

Suppose now that we have ℓ identical ambulances where ambulance i operates out of a base located at the point $d_i \in [-1/2, 1/2]^2$, $i = 1, \ldots, \ell$. Some ambulances may operate from the same base, in which case some of the d_is take the same value. The dynamics of the system are as follows. Calls are answered in first-in first-out order. When a call is received, a dispatcher assigns the call to the closest available ambulance. If no ambulances are available, then the first one that becomes available is selected. Ties are broken through random uniform selection. This dispatching policy does not necessarily minimize response times because the selected ambulance may be far from the call, and a closer ambulance that will soon be free might get to the call sooner. The details of exactly which ambulance is selected are not too important from our standpoint, so long as a sensible rule is used that spreads the workload among the available ambulances.

After traveling to the call location, the ambulance spends some time at the scene after which, with probability p, the patient is transported to the hospital, which is again at the point $(0, 0)$. In this case, after reaching the hospital the patient is instantaneously dropped off and the ambulance is freed for other work, typically returning to its base. If the patient does not require hospitalization, then after the scene time is complete the ambulance is freed for other work, again typically returning to its base. We allow redirection, where the ambulance may be redirected to a new call before it reaches its base.

A natural Markov chain that models this process is $Z = (Z_n\colon n \geqslant 0)$, where $Z_n = (W_n(i), \beta_n(i)\colon i = 1, \ldots, \ell)$. Here $W_n(i)$ gives the workload for ambulance i associated with all calls that have been received up to, and including, call n. Notice that the workload will be associated with only a subset of the first

n calls since there are multiple ambulances. The vector $\beta_n(i) \in [-1/2, 1/2]^2$ gives the location in the unit square where ambulance i will be located when the workload $W_n(i)$ is first cleared. This is the hospital if its last patient is hospitalized, and the location of its last call if not.

Under what conditions is the chain stable? Consider the work associated with a single call. The ambulance first needs to travel to the call from its current location, taking at most 2 time units. It then tends to the patient at the scene, taking, on average, EU time units. It may then transport the patient to the hospital, taking at most 1 time unit. Therefore, a bound on the expected amount of work brought in by each call is again $3 + EU$. Since there are ℓ ambulances we expect that the system will be stable if

$$3 + EU < \ell E\tau,$$

where τ is a random variable representing the time between calls. To verify this belief we can appeal to the FLC.

As already mentioned, the FLC is actually stronger than required to establish stability. For example, we have seen that it is also useful for proving that certain steady-state expectations are finite. A "tighter" condition is the following one.

The Second Lyapunov Condition (SLC). There exists a nonempty $B \subseteq \mathcal{S}$, positive scalars ε, b and δ, an integer $m \geqslant 1$, a probability distribution φ on $\mathcal{S}$, and a function $V : \mathcal{S} \to [0, \infty)$ such that

(1) $P(X_m \in \cdot | X_0 = z) \geqslant \delta\varphi(\cdot)$ for all $z \in B$, and
(2) $E(V(X_1)|X_0 = z) \leqslant V(z) - \varepsilon + bI(z \in B)$ for all $z \in \mathcal{S}$.

The only change in this definition from the previous one is requirement (2). Here the nonnegative function V again represents energy, and requirement (2) states that the energy tends to decrease when the chain is outside the set B. If the Markov chain satisfies the SLC, then it is again positive recurrent in a certain precise sense; see Meyn and Tweedie (1993, Theorem 13.0.1).

The FLC implies the SLC. To see why, notice that if the chain satisfies requirement (2) of the FLC, then

$$\begin{aligned} E\big(V(X_1)|X_0 = z\big) &\leqslant aV(z) + bI(z \in B) \\ &= V(z) - (1 - a)V(z) + bI(z \in B) \\ &\leqslant V(z) - (1 - a) + bI(z \in B), \end{aligned}$$

where the final inequality follows since $V(z) \geqslant 1$ in the FLC.

So now let us turn to finding a function V that satisfies the SLC for the multiple-ambulance case.

First consider requirement (1). Suppose the workloads $w(i)$, $i = 1, \ldots, \ell$, are all at most K, say. If $\tau_1 > K + 2$, then when the next call is received, all of the ambulances will be at their bases. Therefore, when $\tau_1 > K + 2$, Z_1 has a certain distribution φ and is independent of Z_0. (In fact, the chain

regenerates at such times.) Now τ_1 is exponentially distributed, and therefore $\mathrm{P}(\tau_1 > K + 2) > 0$. Hence, requirement (1) of the SLC is satisfied when $B = \{z\colon w(i) \leqslant K\}$ for any $K > 0$ where $z = (w(i), \beta(i)\colon 1 \leqslant i \leqslant \ell)$.

Next we turn to requirement (2). It is natural to try $V(z) = w(1) + \cdots + w(\ell)$, the sum of the workloads of the ambulances. Consider what happens on a single step of the Markov chain. Let D_i be the index of the ambulance that responds to the ith call, and Q_i denote the time required for this ambulance to travel to the scene, treat the patient at the scene and, if necessary, transport the patient to the hospital. Then

$$\begin{aligned}\mathrm{E}\big[V(Z_1)|Z_0 = z\big] &= \mathrm{E}\sum_{i=1}^{\ell}\big(\big[w(i) - \tau_1\big]^+ + Q_1 I(D_1 = i)\big)\\ &= \mathrm{E}Q_1 + \sum_{i=1}^{\ell}\mathrm{E}\big[w(i) - \tau_1\big]^+. \end{aligned} \tag{29}$$

If all of the $w(i)$s are large, then $\mathrm{E}[w(i) - \tau_1]^+ \approx w(i) - \mathrm{E}\tau_1$ for each i. So then

$$\mathrm{E}\big[V(Z_1)|Z_0 = z\big] - V(z) \approx \mathrm{E}Q_1 - \ell\,\mathrm{E}\tau_1$$

which is negative under our conjectured stability condition as desired. But when one or more of the $w(i)$s is "small", this heuristic argument breaks down. In fact, what can happen in this case is that the overall workload as measured by V increases! The problem is that while the overall work in the system may be high, the work is not evenly shared by the ambulances, and so some may be idle when they are greatly needed.

Perhaps we chose a poor Lyapunov function V? An alternative is $V(z) = \max_i w(i)$, the maximum workload of an ambulance. One finds that it works well when one of the workloads is large and the rest are small, but it fails when all of the workloads are large and roughly equal.

Both of our choices of Lyapunov function have failed. However, both *can* be made to work. We will show how with our first choice, the sum of the workloads of the ambulances. In order to gain some insight it is helpful to consider the drift of the Markov chain. To visualize this consider the case $\ell = 2$, so there are only 2 ambulances. Consider the workloads of the ambulances as a point in the plane, and compute the expected change in workload levels in a single transition as a vector. This can be visualized as in Figure 3. The drift arrows have been normalized to ensure that they do not cross, to ensure that the figure remains uncluttered. The diagonal lines represent lines of constant workload.

Notice that away from the boundary where both workloads are large, the drift is toward lower workload, as desired. But near the boundaries, the drift is toward higher workload, at least for a short time. This plot suggests that the one-step drift of the sum of the workloads is negative "away from the boundary", but not so near the boundary. It also suggests that if we are near the

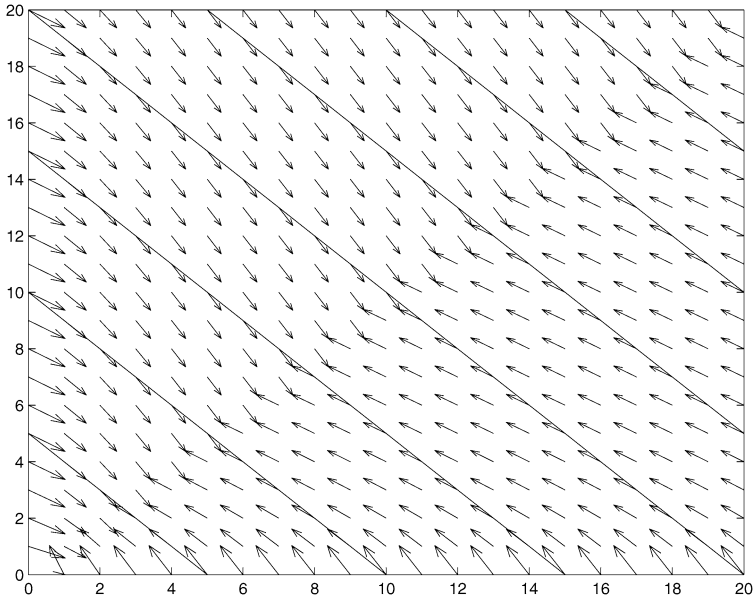

Fig. 3. The workload drift.

boundary, we should wait for a few transitions to see negative drift. The appropriate concept here is "state-dependent drift". A state-dependent version of the SLC (Meyn and Tweedie, 1993, Theorem 19.1.2) is also sufficient to ensure stability, and can be stated as follows.

The Third Lyapunov Condition (TLC). There exists a nonempty $B \subseteq \mathcal{S}$, positive scalars ε, b and δ, an integer $m \geqslant 1$, a probability distribution φ on $\mathcal{S}$, and a function $V : \mathcal{S} \to [0, \infty)$ such that

(1) $\mathrm{P}(X_m \in \cdot | X_0 = z) \geqslant \delta\varphi(\cdot)$ for all $z \in B$, and
(2) $\mathrm{E}(V(X_{n(z)}) | X_0 = z) \leqslant V(z) - n(z)\varepsilon + bI(z \in B)$ for some integer $n(z) \geqslant 1$, for all $z \in \mathcal{S}$.

We are now in a position to state and prove a stability result. (The proof given here is more involved than I would prefer, and so is deferred to the Appendix. There may be a simpler proof but I could not find it.)

Proposition 15. *Suppose that* $\mathrm{E}Q_1 < 2\mathrm{E}\tau_1$. *Then V satisfies the TLC, and so the two-ambulance model is stable.*

So under a natural condition the two-ambulance model is stable. The proof uses the TLC, which is a state-dependent generalization of the SLC. There is also a state-dependent version of the FLC that allows us to prove that certain expectations are finite (Meyn and Tweedie, 1993, Theorem 19.1.3), but space reasons prevent us from discussing that further.

In this section we have tried to demonstrate the power of Lyapunov function methods in simulation analysis. There is some art involved in finding a Lyapunov function and proving that it works, but the benefits are great from a theoretical point of view. These techniques have been used in, for example, establishing asymptotic properties of gradient estimators (Glynn and L'Ecuyer, 1995), proving that standardized time series procedures are valid in discrete-event simulation (Haas, 1999), analyzing stochastic approximation algorithms (Bhatnagar et al., 2001), showing that certain variance reduction methods are valid in Markov chain simulations (Henderson and Glynn, 2002; Henderson and Meyn, 2003), and establishing moment conditions used to establish consistency of quantile estimators in Markov chain simulations (Henderson and Glynn, 2004).

Acknowledgements

This work was partially supported by National Science Foundation Grants DMI 0230528, DMI 0224884 and DMI 0400287.

Appendix Proof of Proposition 15

We showed above that if $B = \{z\colon\ w(i) \leqslant K\}$, where $z = (w(i), \beta(i)\colon\ i = 1, 2)$ and $K > 0$ is arbitrary, then requirement (1) of the TLC was satisfied. It remains to establish requirement (2).

First consider the case where $\mathrm{E}Q_1 < \mathrm{E}\tau_1$. Let $\varepsilon_1 > 0$ be such that $\mathrm{E}Q_1 < \mathrm{E}\tau_1 - \varepsilon_1$. Now, $[x-\tau_1]^+ - x \to -\tau_1$ a.s. as $x \to \infty$. Furthermore, $|[x-\tau_1]^+ - x| = \min(\tau_1, x)^+ \leqslant \tau_1$ and $\mathrm{E}\tau_1 < \infty$. Therefore, by dominated convergence, there exists an $x^* > 0$ such that $\mathrm{E}[x-\tau_1]^+ - x \leqslant -\mathrm{E}\tau_1 + \varepsilon_1$ for $x > x^*$. Take $K = x^*$. Then for $z \notin B$, at least one of $w(1)$ and $w(2)$ exceeds K. Suppose, without loss of generality, that $w(1) > K$. From (29),

$$\begin{aligned}
&\mathrm{E}\big[V(Z_1)|Z_0 = z\big] - V(z)\\
&\quad = \mathrm{E}Q_1 + \sum_{i=1}^{2}\big(\mathrm{E}\big[w(i) - \tau_1\big]^+ - w(i)\big)\\
&\quad \leqslant \mathrm{E}Q_1 + (-\mathrm{E}\tau_1 + \varepsilon_1) + \big(\mathrm{E}\big[w(2) - \tau_1\big]^+ - w(2)\big)\\
&\quad \leqslant \mathrm{E}Q_1 - \mathrm{E}\tau_1 + \varepsilon
\end{aligned}$$

which is the required negative drift for requirement (2) of the TLC.

So now suppose that $\mathrm{E}\tau_1 \leqslant \mathrm{E}Q_1 < 2\mathrm{E}\tau_1$. Let $\tau_i' = \tau_i \wedge K_2$ for some $K_2 > 0$ chosen so that $\mathrm{E}\tau_1' < \mathrm{E}Q_1 < 2\mathrm{E}\tau_1'$. Suppose that we replace the interarrival times τ_i by their truncated versions τ_i' for all i. If we show that the system with truncated interarrival times has negative drift, then so does the system with the

untruncated interarrival times. So let us now assume that $\tau_i \leqslant K_2 < \infty$ and $\mathrm{E}\tau_1 < \mathrm{E}Q_1 < 2\mathrm{E}\tau_1$ for all i.

Let B be of the form specified above. We will specify the constant $K > K_2$ soon. Fix $z \notin B$ and assume, without loss of generality, that $w(1) > K$. If $w(2) > K_2$ then the one-step drift is exactly $\mathrm{E}Q_1 - 2\mathrm{E}\tau_1 < 0$ as required, since $\tau_1 \leqslant K_2$. So assume that $w(2) \leqslant K_2$, so that it is "near the boundary". The remainder of the proof is essentially a formalization of the following observation about the dynamics of the chain. For K large enough, and as long as the incoming jobs do not require a huge amount of work, all of the incoming work will be assigned to the second ambulance for some time. The workload of the second ambulance will increase, and after a while it will be far enough from the boundary that its mean drift will be $\mathrm{E}Q_1 - \mathrm{E}\tau_1 > 0$ on each step. Meanwhile, the workload of the first ambulance decreases at rate $\mathrm{E}\tau_1$, so that the overall workload decreases once the workload of the second ambulance is large enough. This will then supply the needed negative drift. Our last few definitions will appear somewhat cryptic, but hopefully their choice will make more sense shortly. Select $k \geqslant 1$ so that $k(\mathrm{E}Q_1 - 2\mathrm{E}\tau_1) + C < 0$, where the constant C does not depend on z, and will be specified below. Choose $K_3 > 0$ so that

$$k\big[\mathrm{E}Q_1 - \mathrm{E}\tau_1\big(1 + \mathrm{P}(Q_1 \leqslant K_3)^k\big)\big] + C < 0.$$

Finally, choose K large enough that $K - kK_2 > K_2 + kK_3$. We now show that after k transitions, the expected change in workload is negative.

Over the first k transitions, the total inflow of work is $Q_1 + \cdots + Q_k$. Furthermore, the workload of ambulance 1 is so large that it decreases by $\tau_1 + \cdots + \tau_k$. It may also increase if some of the Q_is are very large. Let $\mathcal{E}$ be the event that $Q_i \leqslant K_3$ for all $i = 1, \ldots, k$, and let $\mathcal{E}^c$ denote its complement. The event $\mathcal{E}$ occurs if the first k jobs are all not too large. On the event $\mathcal{E}$, the first k jobs are all assigned to the second ambulance, and so the second ambulance's workload follows a Lindley recursion, as described next.

Let Y_i denote the waiting time in the queue (not counting service) for the ith job in a single-server queue with interarrival times $(\tau_j\colon\ j \geqslant 1)$ and service times $(Q_j\colon\ j \geqslant 1)$. Then $Y_1 = [w(2) - \tau_1]^+$ and for $i \geqslant 1$, $Y_{i+1} = [Y_i + Q_i - \tau_{i+1}]^+$. For $i \geqslant 2$ define $S_i = \sum_{j=2}^{i}(Q_{j-1} - \tau_j)$. Then (Asmussen, 2003, p. 94), for $i \geqslant 2$,

$$Y_i = S_i - \min\{-Y_1, S_2, S_3, \ldots, S_i\}.$$

So we can now write

$$\begin{aligned} &V(Z_k) - V(z) \\ &\quad \leqslant -\sum_{i=1}^{k} \tau_i + I\big(\mathcal{E}^c\big) \sum_{i=1}^{k} Q_i + \big(W_k(2) - w(2)\big) I(\mathcal{E}). \end{aligned} \tag{30}$$

The last term in (30) can be written as

$$\begin{aligned}
&\bigl[Y_k + Q_k - w(2)\bigr]I(\mathcal{E}) \\
&\quad = \bigl[S_k - \min\{-Y_1, S_2, S_3, \ldots, S_k\} + Q_k - w(2)\bigr]I(\mathcal{E}) \\
&\quad = I(\mathcal{E}) \sum_{i=1}^{k} Q_i - I(\mathcal{E}) \sum_{i=2}^{k} \tau_i - w(2)I(\mathcal{E}) \\
&\qquad - I(\mathcal{E}) \min\{-Y_1, S_2, S_3, \ldots, S_k\} \\
&\quad \leqslant I(\mathcal{E}) \sum_{i=1}^{k} Q_i - I(\mathcal{E}) \sum_{i=2}^{k} \tau_i - I(\mathcal{E}) \min\{-Y_1, S_2, S_3, \ldots, S_k\}.
\end{aligned} \tag{31}$$

From (30) and (31) we see that

$$\begin{aligned}
&\mathrm{E}\bigl[V(Z_k)|Z_0 = z\bigr] - V(z) \\
&\quad \leqslant -k\,\mathrm{E}\tau_1 + k\,\mathrm{E}Q_1 \\
&\qquad - \mathrm{E}I(\mathcal{E}) \sum_{i=2}^{k} \tau_i - \mathrm{E}\bigl[I(\mathcal{E}) \min\{-Y_1, S_2, S_3, \ldots, S_k\}|Z_0 = z\bigr] \\
&\quad = k(\mathrm{E}Q_1 - \mathrm{E}\tau_1) - \mathrm{P}(Q_1 \leqslant K_3)^k (k-1)\,\mathrm{E}\tau_1 \\
&\qquad - \mathrm{E}\bigl[I(\mathcal{E}) \min\{-Y_1, S_2, S_3, \ldots, S_k\}|Z_0 = z\bigr] \\
&\quad \leqslant k\bigl(\mathrm{E}Q_1 - \mathrm{E}\tau_1\bigl(1 + \mathrm{P}(Q_1 \leqslant K_3)^k\bigr)\bigr) + \mathrm{E}\tau_1 \\
&\qquad - \mathrm{E}\bigl[\min\{-Y_1, S_2, S_3, \ldots\}|Z_0 = z\bigr].
\end{aligned}$$

But $Y_1 \leqslant w(2) \leqslant K_2$, and so

$$\begin{aligned}
&\mathrm{E}\bigl[V(Z_k)|Z_0 = z\bigr] - V(z) \\
&\quad \leqslant k\bigl(\mathrm{E}Q_1 - \mathrm{E}\tau_1\bigl(1 + \mathrm{P}(Q_1 \leqslant K_3)^k\bigr)\bigr) + C < 0,
\end{aligned}$$

where

$$C = \mathrm{E}\tau_1 - \mathrm{E}\min\{-K_2, S_2, S_3, \ldots\}$$

does not depend on z. The constant C is finite since the random walk $S_2, S_3, \ldots$ has positive drift and the increments $Q_i - \tau_{i+1}$ have bounded negative part; see Asmussen (2003, p. 270). We have shown that after k steps the drift is negative, and this establishes Condition 2 as required.

References

Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edition. *Applications of Mathematics: Stochastic Modeling and Applied Probability*, vol. 51. Springer-Verlag, New York.

Avramidis, A.N., Wilson, J.R. (1996). Integrated variance reduction strategies for simulation. *Operations Research* 44, 327–346.

Bhatnagar, S., Fu, M.C., Marcus, S.I., Bhatnagar, S. (2001). Two-timescale algorithms for simulation optimization of hidden Markov models. *IIE Transactions* 33, 245–258.
Billingsley, P. (1986). *Probability and Measure*, 2nd edition. Wiley, New York.
Cheng, R.C.H., Holland, W. (1998). Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation* 60, 183–205.
Dieudonné, J.A. (1960). *Foundations of Modern Analysis*. Academic Press, New York.
Fu, M.C., Hu, J.-Q. (1997). *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer, Boston.
Glasserman, P. (1991). *Gradient Estimation via Perturbation Analysis*. Kluwer, Dordrecht, The Netherlands.
Glasserman, P. (1993). Filtered Monte Carlo. *Mathematics of Operations Research* 18, 610–634.
Glynn, P.W. (1995). Some new results on the initial transient problem. In: Alexopoulos, C., Kang, K., Lilegdon, W.R., Goldsman, D. (Eds.), *Proceedings of the 1995 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 165–170.
Glynn, P.W., Heidelberger, P. (1990). Bias properties of budget constrained simulations. *Operations Research* 38, 801–814.
Glynn, P.W., Iglehart, D.L. (1990). Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15, 1–16.
Glynn, P.W., L'Ecuyer, P. (1995). Likelihood ratio gradient estimation for stochastic recursions. *Advances in Applied Probability* 27, 1019–1053.
Glynn, P.W., Whitt, W. (1992). The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability* 2, 180–198.
Haas, P.J. (1999). On simulation output analysis for generalized semi-Markov processes. *Communications in Statistics: Stochastic Models* 15, 53–80.
Henderson, S.G. (2000). Mathematics for simulation. In: Joines, J.A., Barton, R.R., Kang, K., Fishwick, P.A. (Eds.), *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 137–146.
Henderson, S.G. (2001). Mathematics for simulation. In: Peters, B.A., Smith, J.S., Medeiros, D.J., Rohrer, M.W. (Eds.), *Proceedings of the 2001 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 83–94.
Henderson, S.G., Glynn, P.W. (2001). Computing densities for Markov chains via simulation. *Mathematics of Operations Research* 26, 375–400.
Henderson, S.G., Glynn, P.W. (2002). Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research* 27, 253–271.
Henderson, S.G., Glynn, P.W. (2004). A central limit theorem for empirical quantiles in the Markov chain setting, working paper.
Henderson, S.G., Meyn, S.P. (2003). Variance reduction for simulation in multiclass queueing networks. *IIE Transactions*, in press.
Irizarry, M.A., Kuhl, M.E., Lada, E.K., Subramanian, S., Wilson, J.R. (2003). Analyzing transformation-based simulation metamodels. *IIE Transactions* 35, 271–283.
Law, A.M., Kelton, W.D. (2000). *Simulation Modeling and Analysis*, 3rd edition. McGraw-Hill, New York.
Meyn, S.P., Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Academic Press, Orlando, FL.
Rice, J.A. (1988). *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
Schruben, L.W. (1983). Confidence interval estimation using standardized time series. *Operations Research* 31, 1090–1108.
Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
Shedler, G.S. (1987). *Regeneration and Networks of Queues*. Springer-Verlag, New York.
Wand, M.P., Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.