Queueing Syst (2013) 75:243–277
DOI 10.1007/s11134-012-9340-8

Gaussian skewness approximation for dynamic rate multi-server queues with abandonment

William A. Massey · Jamol Pender

Received: 10 October 2011 / Revised: 15 November 2012 / Published online: 21 February 2013

Abstract The multi-server queue with non-homogeneous Poisson arrivals and customer abandonment is a fundamental dynamic rate queueing model for large scale service systems such as call centers and hospitals. Scaling the arrival rates and number of servers arises naturally when a manager updates a staffing schedule in response to a forecast of increased customer demand. Mathematically, this type of scaling ultimately gives us the fluid and diffusion limits as found in Mandelbaum et al., Queueing Syst 30:149–201 (1998) for Markovian service networks. The asymptotics used here reduce to the Halfin and Whitt, Oper Res 29:567–588 (1981) scaling for multi-server queues. The diffusion limit suggests a Gaussian approximation to the stochastic behavior of this queueing process. The mean and variance are easily computed from a two-dimensional dynamical system for the fluid and diffusion limiting processes. Recent work by Ko and Gautam, INFORMS J Comput, to appear (2012) found that a modified version of these differential equations yield better Gaussian estimates of the original queueing system distribution. In this paper, we introduce a new three-dimensional dynamical system that is based on estimating the mean, variance, and third cumulant moment. This improves on the previous approaches by fitting the distribution from a quadratic function of a Gaussian random variable.

Keywords Multi-server queues · Abandonment · Dynamical systems · Asymptotics · Time-varying rates · Time inhomogeneous Markov processes · Hermite polynomials · Fluid and diffusion limits · Skewness · Cumulant moments

W. A. Massey · J. Pender (✉)
Department of Operations Research and Financial Engineering, Princeton University,
Princeton, NJ, USA
e-mail: jamol.pender@gmail.com; jpender@princeton.edu

W. A. Massey
e-mail: wmassey@princeton.edu

Mathematics Subject Classification 33C45 · 60K25

1 Introduction

Large scale service systems, such as telephone call centers or healthcare centers such as hospitals, have customer inflow–outflow dynamics with many common features. For example:

- The customer arrival patterns have both time of day and seasonal effects.
- Customer population sizes are large where the individual customer actions are mutually independent.
- Multiple service agents allow many customers simultaneous access to services in parallel.

In a call center, these service agents are the telephone operators. For a hospital, these service agents can be the beds, attendant nurses, or doctors. Other common features for these service systems include:

- Arriving customers wanting to engage in service are delayed if all the available service agents are busy.
- Waiting customers may decide to leave the system if their delay is excessively long.

In the case of a hospital emergency room, this can be a patient "leaving without being seen" and deciding to go to a second hospital.

To understand and predict the customer dynamics of these service systems, we need to construct a *queueing model*. The resulting queueing analysis, which is a probabilistic study of the number of customers currently engaged in or requesting services, gives us various performance measurements. One such quality of service metric is the average number of customers in the system at any given time.

Inputs to this queue are due to customer arrivals. Outputs to this queue are due to either customers who complete service or ones who leave prematurely. Let us assume that all customers have a personal expectation of how long their queueing delay should be. It then follows that this latter group departs the system when those expectations are *not* met. This phenomenon is referred to in the queueing literature as *abandonment*.

For a large population of customers with independent actions, we model the arrival pattern of customers as a simple (single jump), random counting process with *independent increments*. The latter statement means that the number of customer arrivals during disjoint time intervals are statistically independent. According to Prékopa [18], this is equivalent to saying that our customer arrival model is a *non-homogeneous Poisson process*. In practice, we can use historical customer arrival data to infer some average arrival rate function $\{ \lambda(t) \mid t \geq 0 \}$ forecast, which parameterizes this process.

Now we model both the service and abandonment times for each arriving customer as two independent sequences of i.i.d. random variables. For mathematical simplicity, we assume that the service time distributions are exponentially distributed with mean $1/\mu$. We also assume that the abandonment time distributions are exponential with mean $1/\beta$.

Letting c equal the deterministic process of $c(t)$ service agents at time t, we define our service model to be an $M(t)/M/c(t) + M$ queue where "$M(t)$" denotes a nonhomogeneous Poisson arrival process and "$+M$" denotes an exponential distribution for the abandonment time. We also refer to this as an *Erlang-A* queueing model [4]. This is a discrete state space Markov process that is time inhomogeneous or has *dynamic rates*. It is the canonical Markovian queueing process for multi-agent service center models that we study in this paper.

The standard analysis for the time homogeneous or constant rate version of this queueing system process is to compute the steady state distribution as $t \to \infty$. For the dynamic rate case, this is a static equilibrium analysis that may not be applicable and gives no insight into how the system evolves over time. The probability distribution for the queue can be studied as a dynamical system but it is infinite dimensional. At first glance, the most viable alternative would be to apply Monte-Carlo simulation methods directly to the sample paths of this Markov process.

Mandelbaum et al. [13] used the theory of strong approximations to develop fluid and diffusion limit theorems for the stochastic sample path behavior of *Markovian service networks*, where the Erlang-A model is the special case of a single node network. The resulting fluid and diffusion limits are characterized by low-dimensional dynamical systems that approximate the stochastic evolution of the queueing process. Moreover, the precise meaning of "low" for multi-server queueing networks is that the dynamical system dimension is *only* a quadratic function of the number of nodes in the network and *not* the number of agents at any of the service centers.

Moreover, Mandelbaum et al. [14] illustrate that these approximations for the mean and variance of the Erlang-A model are at their best when the scaling parameter η is large and the queue does not "linger" through periods of "critical loading." This corresponds to times when the mean number of customers in the queueing system nearly matches the number of agents. Ko and Gautam [11] extended the applicability of these approximations by using Gaussian convolutions to smooth the rates as nonsmooth functions of the states. They obtained better approximations for the mean and variance dynamics which suggest that there is still room for improvement.

The contribution of this paper is to present a simple method that generates new algorithms that are successively better approximations of the stochastic dynamics for the $M(t)/M/c(t) + M$ model. This is achieved by computing a low-dimensional, deterministic, dynamical system. Computing a single run of such a deterministic process is a significant savings in computational time, compared to averaging over the typical 10,000 runs of a Monte Carlo simulation. Papers by Hampshire [6] and with co-authors in [7] and [8] show how low-dimensional dynamical systems can approximate the statistics of the stochastic evolution for many different types of fundamental queueing models. We plan to show in a subsequent paper how to extend this method to a similar approximate analysis for all Markovian service networks.

The rest of the paper is as follows. In Sect. 2, we review the approximation methods for the mean and variance of our queueing process as derived from papers [13] and [14]. In Sect. 3, we introduce our method. We show how its one-dimensional case reduces to the fluid model and its two-dimensional case reduces to the Gaussian smoothing methods of Ko and Gautam [11]. In Sect. 4, we introduce our new method, the *Gaussian skewness approximation*. This improves on the one- and two-dimensional cases of the method to offer a three-dimensional approximating dynamical system. We use this technique to improve our estimation of the distribution for our queueing process. We compare simulated histograms of the queueing system distribution at specific time points with the corresponding density distributions constructed by our algorithm. Section 6 concludes the paper with insights and possible extensions. Finally, all the necessary technical proofs and calculations to produce the simple formulas for the algorithm appear in Appendix A. A detailed sketch of our simulation algorithm for multi-server queues is contained in Appendix B.

2 Stochastic analysis of the queueing model

Mandelbaum et al. [13] shows that the queueing system process $Q \equiv \{Q(t) | t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

$$Q(t) = Q(0) + \Pi_1\left(\int_0^t \lambda(s)ds\right) - \Pi_2\left(\int_0^t \mu \cdot (Q(s) \wedge c(s))ds\right) - \Pi_3\left(\int_0^t \beta \cdot (Q(s) - c(s))^+ ds\right),$$

where $\Pi_i \equiv \{\Pi_i(t) | t \geq 0\}$ for $i = 1, 2, 3$ are i.i.d. standard (rate 1) Poisson processes. The deterministic time change for Π_1 transforms it into a non-homogeneous Poisson arrival process with rate $\lambda(t)$ that counts the customer arrivals. Subjecting Π_2 to a random time change rate $\mu \cdot (Q(t) \wedge c(t))$, at time t, gives us a departure process that counts the number of serviced customers. Here we assume that there are a deterministic number of $c(t)$ servers, at time t, and i.i.d. exponentially distributed service times of mean $1/\mu$. Similarly the random time change of Π_3 gives us a counting process for the number of queueing abandonments from $c(t)$ servers and i.i.d. exponentially distributed abandonment times of mean $1/\beta$. When the mean number of the system $E[Q(t)]$ is less than the number of servers $c(t)$ or $E[Q(t)] < c(t)$, we say that the system is *underloaded*. Conversely, when $E[Q(t)] > c(t)$, we say that the system is *overloaded*. Finally, when $E[Q(t)] = c(t)$, we say that the system is *critically loaded*.

The numerical example that we consider in this paper to illustrate our approximation methods evolves over the time interval (0, 40). Moreover, we assume an arrival rate function $\lambda(t) = 10 + 5 \sin t$, a constant service rate $\mu = 1.0$, a constant abandonment rate $\beta = 0.25$, and the number of servers is constant with $c = 10$. We summarize this in Table 1. All our simulations are averaged over 10,000 runs with an initial load of zero or $Q(0) = 0$. Figure 1 is a plot over time of the simulated estimations for the actual mean $E[Q(t)]$, variance $\mathrm{Var}[Q(t)]$, and third cumulant moment $C^{(3)}[Q(t)]$, where the first two cumulant moments are given by the mean and variance. The second and third cumulant moments are given by the formulas

$$\mathrm{Var}\,[Q(t)] \equiv E\left[(Q(t) - E[Q(t)])^2\right] \quad \text{and} \quad C^{(3)}\,[Q(t)] \equiv E\left[(Q(t) - E[Q(t)])^3\right]. \tag{2.1}$$

Table 1 Main numerical example

Parameter	Value (at time t)
λ	$10.0 + 5.0 * \sin t$
μ	1.0
c	10
β	0.25

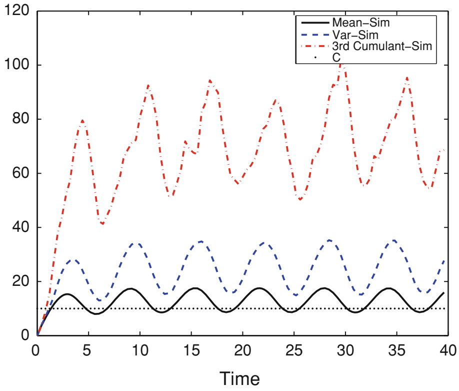

Fig. 1 Simulation of mean, variance, and third cumulant moment of the queueing process

When the $M(t)/M/c(t) + M$ queueing system is underloaded for all t, then it behaves more like an infinite server queue or $c = \infty$. Even with a dynamic arrival rate, when initialized by a Poisson distribution, such a queue always has a Poisson transient distribution. A more detailed exploration of infinite server queueing dynamics with non-homogeneous input can be found in the works of Palm [17], Khintchine [10], and Eick et al. [1].

Under general conditions, the Poisson distribution is uniquely characterized by having all its cumulant moments equal to its mean [10]. During the initial period of underloaded behavior, starting at the trivial Poisson distribution of $Q(0) = 0$, all the curves in Fig. 1 appear to be identical during the initial underloaded period. The divergence of these three curves starts during the overloaded period. Note that during the recurring, but brief, periods of underloading the three curves attempt to reconverge. The longer that the number in the queueing system stays below the number of servers c, the more its infinite server behavior forgets about its initial conditions. At this point, the system distribution more closely resembles a Poisson distribution, whose three cumulant moments all equal each other.

To gain a better understanding of the dynamics of the mean, variance, and third cumulant moment for our queueing process, we need to study their rates of change over time. This leads us to the *functional version* of the Kolmogorov forward equations for the $M(t)/M/c(t) + M$ queue, which is of the form

$$\begin{aligned}\overset{\bullet}{E}\left[f(Q)\right] &= \lambda \cdot E\left[f(Q+1) - f(Q)\right] + \mu \cdot E\left[(Q \wedge c) \cdot (f(Q-1) - f(Q))\right] \\ &\quad + \beta \cdot E\left[(Q-c)^{+} \cdot (f(Q-1) - f(Q))\right], \qquad (2.2)\end{aligned}$$

for all appropriate functions f. We always assume, for this paper, that quantities such as β and μ are constant. To simplify our notation, time dependent quantities such as $Q(t)$, $\lambda(t)$, and $c(t)$ are denoted in this paper as Q, λ, and c, with their time dependence suppressed. For an expression like $E\left[f(Q(t))\right]$ we use the "dot" notation of physics to denote its time derivative when we do not make time explicit or

$$\overset{\bullet}{E}\left[f(Q)\right] \equiv \frac{d}{dt} E\left[f(Q(t))\right]. \tag{2.3}$$

Using special cases of f, such as $f(Q)$ equalling Q, Q^2, or Q^3, we can then obtain the following set of cumulant moment, Kolmogorov forward equations:

$$\begin{aligned}\overset{\bullet}{E}[Q] &= \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E\left[(Q-c)^+\right] \\ \overset{\bullet}{\mathrm{Var}}[Q] &= \lambda + \mu \cdot E\left[Q \wedge c\right] + \beta \cdot E\left[(Q-c)^+\right] \\ &\quad -2\left(\mu \cdot \mathrm{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[Q, (Q-c)^+\right]\right),\end{aligned}$$

and

$$\begin{aligned}\overset{\bullet}{C}{}^{(3)}[Q] &= \lambda - \mu \cdot E\left[Q \wedge c\right] - \beta \cdot E\left[(Q-c)^+\right] \\ &\quad +3\left(\mu \cdot \mathrm{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[Q, (Q-c)^+\right]\right) \\ &\quad -3\left(\mu \cdot \mathrm{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right),\end{aligned}$$

where $\overline{Q} \equiv Q - E[Q]$. We can write the final equations more compactly as

$$\overset{\bullet}{E}[Q] = \lambda - \mu \cdot E\left[Q \wedge c\right] - \beta \cdot E\left[(Q-c)^+\right], \tag{2.4}$$

$$\frac{\overset{\bullet}{E}[Q] + \overset{\bullet}{\mathrm{Var}}[Q]}{2} = \lambda - \mu \cdot \mathrm{Cov}\left[Q, Q \wedge c\right] - \beta \cdot \mathrm{Cov}\left[Q, (Q-c)^+\right], \tag{2.5}$$

and

$$\frac{\overset{\bullet}{E}[Q]}{6} + \frac{\overset{\bullet}{\mathrm{Var}}[Q]}{2} + \frac{\overset{\bullet}{C}{}^{(3)}[Q]}{3} = \lambda - \mu \cdot \mathrm{Cov}\left[\overline{Q}^2, Q \wedge c\right] - \beta \cdot \mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right]. \tag{2.6}$$

Since the $M(t)/M/c(t) + M$ queueing process is a special case of a single node *Markovian service network*, we can also construct an associated, *uniformly accelerated* queueing process where both the new arrival rate $\eta \cdot \lambda$ and the new number of servers $\eta \cdot c$ are both scaled by the same factor $\eta > 0$. A call center interpretation of this is called *resource pooling*. We are scaling up simultaneously the customer demand (arrival rate) and the customer resource supply (number of service agents). Taking the following limits gives us the *fluid* and *diffusion* models of [13], i.e.,

$$\lim_{\eta\to\infty} \frac{Q^\eta}{\eta} = q \quad \text{a.s.} \quad \text{and} \quad \lim_{\eta\to\infty} \sqrt{\eta}\cdot\left(\frac{Q^\eta}{\eta} - q\right) \overset{d}{=} \hat{Q}, \tag{2.7}$$

where the deterministic process q, the *fluid mean*, is governed by the one-dimensional dynamical system

$$\overset{\bullet}{q} = \lambda - \mu\cdot(q \wedge c) - \beta\cdot(q-c)^+. \tag{2.8}$$

Moreover, as pointed out in [13], if the set of time points $\{ t \mid q(t) = c \}$ has measure zero, then $\hat{Q}$ is a *Gaussian* diffusion process (with mean zero when $Q^\eta(0)$ is only a constant scaled by η) whose variance combines with the fluid mean to form a two-dimensional dynamical system given by (2.8) and

$$\overset{\bullet}{v} = \lambda + \mu\cdot(q\wedge c) + \beta\cdot(q-c)^+ - 2\cdot(\mu\cdot\{q<c\} + \beta\cdot\{q\geq c\})\cdot v, \tag{2.9}$$

where $v \equiv \mathrm{Var}[\hat{Q}]$ and $\{q < c\}$ denotes an *indicator function* that equals one if the statement is true, i.e., if $q < c$, and zero if the statement is false. We can write these equations more compactly as

$$\overset{\bullet}{q} = \lambda - \mu\cdot(q\wedge c) - \beta\cdot(q-c)^+ \tag{2.10}$$

$$\frac{\overset{\bullet}{q} + \overset{\bullet}{v}}{2} = \lambda - (\mu\cdot\{q<c\} + \beta\cdot\{q\geq c\})\cdot v. \tag{2.11}$$

In Fig. 2, we compare simulations of the mean and variance for the $M(t)/M/c(t)+M$ queueing process to our dynamical system fluid and diffusion estimates given by (2.10) and (2.11). We do this using the fluid limit as an approximation for the mean of queueing process, as shown in the left hand graph. We then use the diffusion limit variance as an approximation to the variance of the queueing process, as shown in the right hand graph. In this example, the fluid limit works everywhere as a "reasonable" approximation (i.e., relative error of 10 %) to the dynamics of the mean. There are only a few time periods where the diffusion variance is a reasonable approximation of the queueing process variance, such as during the initial period of underloading. In the next section, we discuss new sets of low-dimensional dynamical systems that yield better approximations of the relevant stochastic queueing behavior.

3 Deterministic and Gaussian approximations

From a computational perspective, we want the ensemble of formulas for the time derivatives of the mean, variance, and third cumulant moment, as summarized in (2.4)–(2.6), to be an *autonomous* set of differential equations. This means that their current behavior should be some integral functional of their past behavior. We can achieve this by making a *closure approximation* in the same spirit as Rothkopf and Oren [20]. The philosophy that they give for this technique is as follows (see page 524 of [20]):

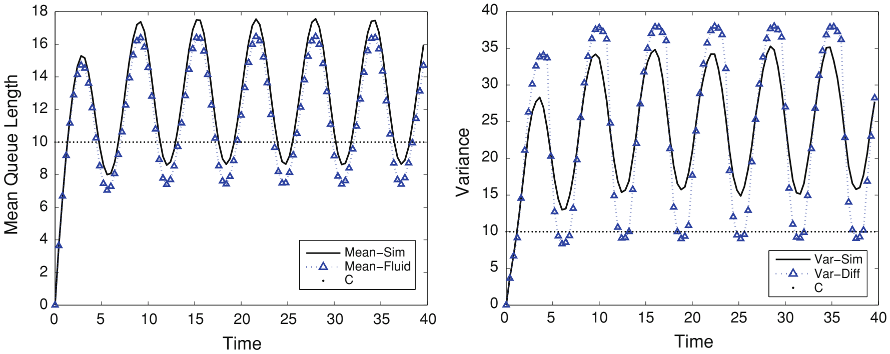

Fig. 2 Comparisons of simulated mean and its fluid limit (*left*). Comparisons of simulated variance and diffusion limit (*right*)

> … The basic strategy of a closure technique is to reduce an infinite system of equations to a finite system by making a "closure assumption" in the form of a functional relationship between the variables of the system.

Similar techniques for non-stationary (dynamic rate) queueing models are also used in Taaffe and Ong [23].

In general, we start by assuming that our underlying closure distribution for the queueing process is uniquely defined by a finite set of parameters. Next, we assume that these parameters are uniquely defined by the same number of expectations of some distinct functions of the queueing process. The forward equations for these functional expectations then form a finite dimensional, dynamical system for these parameters. Whereas [20] and [23] assume an underlying discrete distribution for their closure assumptions, our underlying distribution is continuous and based on polynomials of Gaussian random variables.

For example, we can define a *deterministic mean approximation* for our queueing model by assuming that some underlying deterministic process $q \equiv \{q(t) | t \geq 0\}$ approximates our Markovian queueing process. If we replace Q by q in the Kolmogorov forward equation for the mean of Q as given by (2.4), then q solves the resulting autonomous, one-dimensional, dynamical system.

$$\overset{\bullet}{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^+, \tag{3.1}$$

where we set $q(0) = Q(0)$. This method, however, takes us right back to the dynamical system characterization of the fluid limit given by (2.8).

Let us now extend this method to the two-dimensional, dynamical system case. Inspired by our diffusion limit being Gaussian, suppose that we approximate the dynamics of the mean and variance of Q by a random process $\mathcal{Q} \equiv \{\mathcal{Q}(t) | t \geq 0\}$ such that

$$\mathcal{Q}(t) \overset{d}{=} q(t) + X \cdot \sqrt{v(t)}. \tag{3.2}$$

for all $t \geq 0$, where $\{q(t), v(t) | t \geq 0\}$ is some two-dimensional, deterministic, dynamical system, where the v process is always positive. In this paper, we always define X to be a *standard* Gaussian random variable or Gaussian(0, 1). Either one is shorthand for a Gaussian distribution with zero mean and unit variance. We define φ and Φ to be the density and the cumulative distribution functions, for X, respectively, where

$$\varphi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) \equiv \int_{-\infty}^{x} \varphi(y)\, dy, \quad \text{and} \quad \overline{\Phi}(x) \equiv 1 - \Phi(x) = \int_{x}^{\infty} \varphi(y)\, dy. \tag{3.3}$$

Now we make this substitution of q, v, and X into the forward equations for the mean and the variance of Q, i.e., (2.4) and (2.5). We can call the resulting two-dimensional dynamical system the *Gaussian variance approximation* (GVA). The new autonomous differential equations for q and v are

$$\overset{\bullet}{q} = \lambda - \mu \cdot q - \big(\mu \cdot E[X \wedge \chi] + \beta \cdot E[(X - \chi)^+]\big) \cdot \sqrt{v} \tag{3.4}$$

and

$$\frac{\overset{\bullet}{q} + \overset{\bullet}{v}}{2} = \lambda - \big(\mu \cdot \mathrm{Cov}[X, X \wedge \chi] + \beta \cdot \mathrm{Cov}[X, (X - \chi)^+]\big) \cdot v, \tag{3.5}$$

where

$$E[Q] = q, \quad \mathrm{Var}[Q] = v, \quad \text{and} \quad \chi \equiv \frac{c - q}{\sqrt{v}}. \tag{3.6}$$

To solve these equations numerically, we need to compute the expectation and covariance terms involving functions of the standard Gaussian random variable X. The final results yield generic functions of χ, which is a simple function of q and v. We compute these Gaussian terms using the following lemma:

Lemma 3.1 (Stein [21]) *The random variable X is Gaussian*(0, 1) *if and only if*

$$E\left[X \cdot f(X)\right] = E\left[\frac{d}{dX} f(X)\right], \tag{3.7}$$

for all generalized functions f.

For example, since $(X - \chi)^+ = (X - \chi) \cdot \{X \geq \chi\}$, then Stein's lemma can be used to obtain

$$E\left[(X - \chi)^+\right] = E\left[X \cdot \{X \geq \chi\}\right] - \chi \cdot P\{X \geq \chi\} = \varphi(\chi) - \chi \cdot \overline{\Phi}(\chi). \tag{3.8}$$

Observe that as a function of X, $\{X \geq \chi\}$ is an increasing unit, single step function. Moreover, the density φ is an infinitely differentiable function. Finally, the derivative of $\{X \geq \chi\}$, as a generalized function evaluated at X, is the delta function δ_χ centered at χ. As a result, Stein's lemma gives us

$$E\left[X \cdot \{X \geq \chi\}\right] = E\left[\delta_\chi(X)\right] = \int_{-\infty}^{\infty} \delta_\chi(y) \cdot \varphi(y)\, dy = \varphi(\chi). \tag{3.9}$$

Moreover, since $(X - \chi)^+ = X - X \wedge \chi$, we have

$$E\left[X \wedge \chi\right] = -E\left[(X - \chi)^+\right] = \chi \cdot \overline{\Phi}(\chi) - \varphi(\chi). \tag{3.10}$$

Similar arguments give us

$$\operatorname{Cov}\left[X, X \wedge \chi\right] = E\left[X \cdot (X \wedge \chi)\right] = E\left[\{X \leq \chi\}\right] = P\{X \leq \chi\} = \Phi(\chi) \tag{3.11}$$

and

$$\operatorname{Cov}\left[X, (X - \chi)^+\right] = 1 - \operatorname{Cov}\left[X, X \wedge \chi\right] = \overline{\Phi}(\chi). \tag{3.12}$$

These positive covariances are in keeping with the FKG inequality by Fortuin et al. [3]. This theorem states that increasing functions of the same random variable are always positively correlated.

When we make the substitutions into (3.4) and (3.5), the GVA dynamical system reduces to

$$\overset{\bullet}{q} = \lambda - \mu \cdot q - (\mu - \beta) \cdot \left(\chi \cdot \overline{\Phi}(\chi) - \varphi(\chi)\right) \cdot \sqrt{v} \tag{3.13}$$

and

$$\frac{\overset{\bullet}{q} + \overset{\bullet}{v}}{2} = \lambda - \left(\mu \cdot \Phi(\chi) + \beta \cdot \overline{\Phi}(\chi)\right) \cdot v. \tag{3.14}$$

These equations are the same as the ones with the "g functions" as used in Ko and Gautam [11]. In Fig. 3, the GVA estimation for the mean (given by the square plots) yields a better approximation to the simulated mean than the fluid model (given by the triangular plots) in the left hand graph. For the variance, however, the GVA plot does not match the simulation as well, but it is a significant improvement over the diffusion variance approximation in the right hand graph. In the next section, we introduce a new approximation that is a refinement to the GVA estimate just as GVA is a refinement to the fluid model.

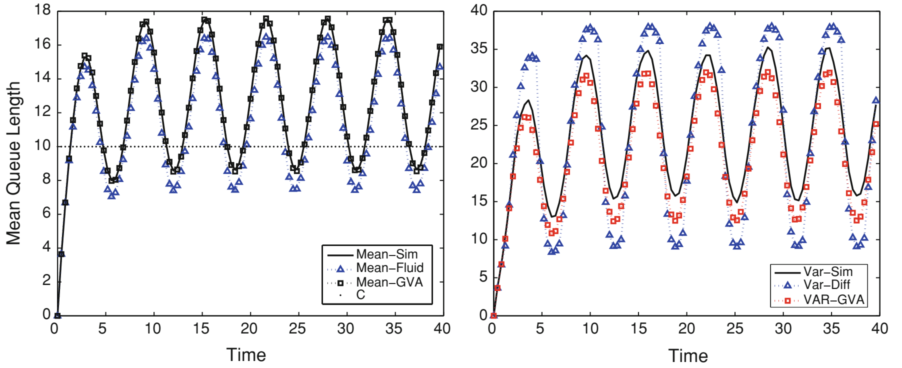

Fig. 3 Comparison of simulation, fluid, and GVA means (*left*). Comparison of simulation, diffusion, and GVA variances (*right*)

4 Gaussian skewness approximation

First, recall that the *skewness* of a random variable Z is defined to be

$$\mathrm{Skew}[Z] \equiv \frac{C^{(3)}[Z]}{\sqrt{\mathrm{Var}[Z]^3}}. \tag{4.1}$$

Skewness captures an intrinsic property of the distribution that is invariant with respect to deterministic translations and positive scalings of the underlying random variable. It follows from these invariance properties that any Gaussian distribution has the same skewness as the standard Gaussian distribution for X, which is zero. We can show that under general conditions, a Gaussian distribution is uniquely characterized by having its third and all higher degree cumulant moments equal to zero. Thus we can use skewness informally as a metric for how "close" a distribution is to being Gaussian.

The graphs of Fig. 4 compare the simulated mean (top graph) to the skewness (bottom graph) as they both evolve over time. We see that the skewness here is locally the furthest from zero, more precisely when it is positive and a local maximum, exactly when the queueing system is nearly critically loaded or its mean is close to the number of servers $c = 10$. This suggests that critical loading times occur precisely at the times when a Gaussian approximation for the queue length distribution is the least effective.

Our diffusion limit theorem suggests that a Gaussian distribution is a good first order approximation for the queueing process distribution. However, for *any* given distribution we can always find some function of a standard Gaussian random variable whose distribution matches it. This is due to the inverse transform method (see Ross [19], p 67) and the fact that $\Phi(X)$ has a uniform distribution on the interval $(0, 1)$. If F is the cumulative distribution function of the queueing process at some fixed time, then the random variable $F^{-1} \circ \Phi(X)$ has the same distribution function defined by F, where

$$F^{-1}(y) \equiv \inf \{ x \mid F(x) \geq y \}.$$

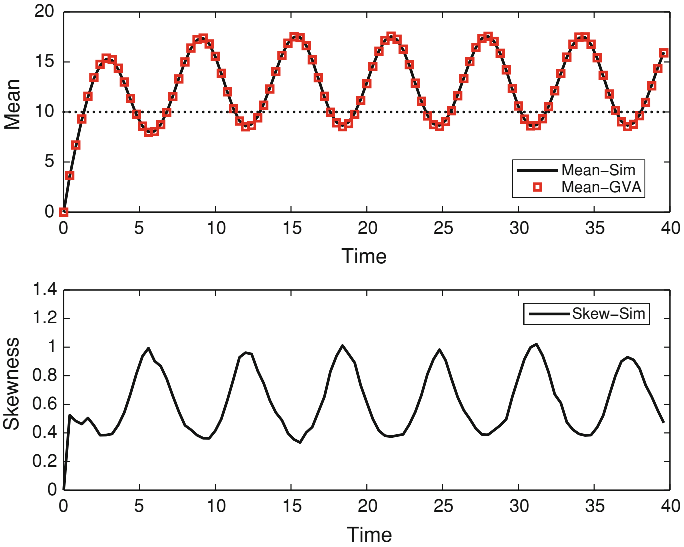

Fig. 4 Comparison of simulation and GVA means (*top graph*) with skewness (*bottom graph*)

Since all polynomials of X are square integrable random variables, it follows that the distribution for any square integrable random variable can be approximated to arbitrary precision by such a polynomial. All linear functions are polynomials of degree 1 and they generate random variables for every possible Gaussian distribution when applied to X. This suggests that the next step in approximating the queueing process distribution uses quadratic functions (degree 2 polynomials) of X.

The family of all square integrable functions of X forms a Hilbert space with respect to an inner product defined as the expectation of the product of two such random variables. The fluid approximation evolves within the one-dimensional Hilbert subspace of constants for our queueing process approximation where the unit basis vector is the constant 1. The GVA method evolves within the two-dimensional subspace for linear functions of X where the orthonormal basis vectors 1 and X generate random variables for every Gaussian distribution. The three-dimensional subspace for quadratic polynomials of X leads to a new unique orthogonal component in the direction of $X^2 - 1$ (we can divide it by $\sqrt{2}$ to make it orthonormal). Inductively, this leads us to the family of *Hermite polynomials*, see Fedoryuk [2]. As functions of X, they form an orthogonal family of polynomials and are a basis for square integrable functions of X.

We now introduce a new approximation method and call it the *Gaussian skewness approximation* (GSA). For some three-dimensional, dynamical system $\{q(t), v(t), \sigma_\theta(t) | t \geq 0\}$, we assume that

$$\mathcal{Q}(t) \overset{d}{=} q(t) + \left(X \cos\theta(t) + \frac{X^2 - 1}{\sqrt{2}} \sin\theta(t) \right) \cdot \sqrt{v(t)}, \tag{4.2}$$

for all $t \geq 0$, where X is a Gaussian(0,1) random variable. Since 1, X, and $X^2 - 1$ are orthogonal vectors, they are all uncorrelated as random variables. This means that $E[\mathcal{Q}] = q$ and $\text{Var}[\mathcal{Q}] = v$. We also define $\sigma_\theta(t)$ to be the skewness of $\mathcal{Q}(t)$ or

$$\sigma_\theta(t) \equiv \text{Skew}[\mathcal{Q}(t)]. \tag{4.3}$$

This approximation is a natural extension of the GVA method since we are using the next orthogonal Hermite polynomial of X. Notice that neither the mean nor the variance are a function of θ. Moreover, the second and third orthonormal components of $\mathcal{Q}$, i.e., X and $(X^2 - 1)/\sqrt{2}$, are multiplied by $\sqrt{v}\,\cos\theta$ and $\sqrt{v}\,\sin\theta$, respectively. This is the general representation for any two-dimensional rectangular coordinate system in terms of its polar coordinates, i.e., the "radius" $\sqrt{v}$ and the "angle" θ. Any two-dimensional vector in this subspace with mean squared norm $\sqrt{v}$ would have this form for some value of θ.

Now observe that the one-dimensional distributions of the process $\mathcal{Q}$ are uniquely parameterized by the deterministic processes given by q, v, and $\sin\theta$ where $v \geq 0$. This is due to the fact that the squares of either $\cos\theta$ or $-\cos\theta$ plus the square of $\sin\theta$ equals 1. Moreover, X and $-X$ are identically distributed since X is a standard Gaussian random variable. Hence $X\cos\theta$ and $-X\cos\theta$ are identically distributed with $X^2 = (-X)^2$.

The next theorem shows that σ_θ is an invertible function of $\sin\theta$. This means that the one-dimensional distributions of $\mathcal{Q}$ are also uniquely parameterized by the deterministic processes q, v, and σ_θ. Now we state and prove our main result.

Theorem 4.1 *Suppose that the one-dimensional distributions of the process $\mathcal{Q}$ are given by*

$$\mathcal{Q} \overset{d}{=} q + Y_\theta \cdot\sqrt{v}, \quad \textit{where} \quad Y_\theta \equiv X\cos\theta + \frac{1}{\sqrt{2}}\left(X^2 - 1\right)\sin\theta \tag{4.4}$$

and the time dependent parameters q and v, combined with σ_θ or $\sin\theta$*, form a three-dimensional dynamical system.*

If $\mathcal{Q}$ solves the same moment forward equations as our queueing process Q for its mean, variance, and third cumulant moment, then we must have

$$E[\mathcal{Q}] = q, \quad Var[\mathcal{Q}] = v, \quad \textit{and} \quad Skew[\mathcal{Q}] = C^{(3)}[Y_\theta] = \sigma_\theta, \tag{4.5}$$

where

$$\sigma_\theta \equiv \sqrt{2}\cdot\left(3 - \sin^2\theta\right)\cdot\sin\theta, \tag{4.6}$$

which holds if and only if

$$\sin\theta = 2\cdot\sin\left(\frac{1}{3}\sin^{-1}\left(\frac{\sigma_\theta}{2\sqrt{2}}\right)\right). \tag{4.7}$$

Moreover, the differential equations are

$$\overset{\bullet}{q} = \lambda - \mu \cdot q - \big(\mu \cdot E[Y_\theta \wedge \chi] + \beta \cdot E[(Y_\theta - \chi)^+]\big) \cdot \sqrt{v}, \tag{4.8}$$

$$\frac{\overset{\bullet}{q} + \overset{\bullet}{v}}{2} = \lambda - \big(\mu \cdot Cov[Y_\theta, Y_\theta \wedge \chi] + \beta \cdot Cov[Y_\theta, (Y_\theta - \chi)^+]\big) \cdot v, \tag{4.9}$$

and

$$\frac{\overset{\bullet}{q}}{6} + \frac{\overset{\bullet}{v}}{2} + \frac{\overset{\bullet}{(\sigma_\theta \cdot \sqrt{v^3})}}{3} = \lambda - \Big(\mu \cdot Cov[Y_\theta^2, Y_\theta \wedge \chi] + \beta \cdot Cov[Y_\theta^2, (Y_\theta - \chi)^+]\Big) \cdot \sqrt{v^3}, \tag{4.10}$$

with $\chi \equiv (c - q)/\sqrt{v}$ *or equivalently* $c = q + \chi \cdot \sqrt{v}$.

Proof We defer this to Appendix A. □

Requiring that the closure approximation version of the Kolmorgov forward equations for the mean, variance, and third cumulant moment all hold is what finally dictates the values of q, v, and $\sin\theta$, or equivalently, q, v, and σ_θ.

To compute the expectations involving functions of Y_θ, we first define $z_+(\theta, \chi)$ and $z_-(\theta, \chi)$ to be the two roots of the quadratic polynomial

$$z \cdot \cos\theta + (z^2 - 1) \cdot \frac{\sin\theta}{\sqrt{2}} = \chi. \tag{4.11}$$

As functions of χ and θ, where $\sin\theta \neq 0$, they have the following form

$$z_+(\theta, \chi) = \frac{\sqrt{2} \cdot \sin\theta + 2\chi}{\cos\theta + \sqrt{1 + 2\sqrt{2} \cdot \chi \sin\theta + \sin^2\theta}} \tag{4.12}$$

and

$$z_-(\theta, \chi) = \frac{\cos\theta + \sqrt{1 + 2\sqrt{2} \cdot \chi \sin\theta + \sin^2\theta}}{-\sqrt{2} \cdot \sin\theta}. \tag{4.13}$$

Now we can characterize the distribution for Y_θ.

Theorem 4.2 *If* $0 < \theta < \pi/2$, *then the cdf of* Y_θ *is for all* $a > -\ell(\theta)$

$$P\{Y_\theta \le a\} = \Phi(z_+(\theta, a)) - \Phi(z_-(\theta, a))$$

with density function

$$\frac{\partial}{\partial a} P\{Y_\theta \le a\} = \frac{\varphi(z_+(\theta, a)) - \varphi(z_-(\theta, a))}{\sqrt{1 + 2 \cdot \sqrt{2} \cdot a \cdot \sin\theta + \sin^2\theta}},$$

where

$$\ell(\theta) \equiv \frac{1+\sin^2\theta}{2\sqrt{2}\cdot\sin\theta}. \tag{4.14}$$

Moreover, both the cdf and density functions are zero when $a \le -\ell(\theta)$.

Proof First, we observe that the event $\{Y_\theta \le a\}$ occurs with probability zero when we have $a \le -\ell(\theta)$. Second, for all $a > -\ell(\theta)$, we have

$$\{Y_\theta \le a\} = \{z_-(\theta,a) \le X \le z_+(\theta,a)\} \tag{4.15}$$

Moreover, since the sum of $z_+(\theta,a) + z_-(\theta,a)$ is *not* a function of a, we have

$$\frac{\partial z_+}{\partial a}(\theta,a) = -\frac{\partial z_-}{\partial a}(\theta,a) = \frac{1}{\sqrt{1+2\cdot\sqrt{2}\cdot a\cdot\sin\theta+\sin^2\theta}}. \tag{4.16}$$

□

Here are simple formulas for the covariance terms of Y_θ used in Theorem 3.1. We prove these results in the Appendix A using Hermite polynomials.

Theorem 4.3 *For all θ and χ, we have*

$$\begin{aligned} E\left[(Y_\theta-\chi)^+\right] &= -\chi\cdot P\{Y_\theta \ge \chi\} + \cos\theta\cdot(\varphi(z_+)-\varphi(z_-)) \\ &\quad + \frac{1}{\sqrt{2}}\cdot\sin\theta\cdot(z_+\cdot\varphi(z_+) - z_-\cdot\varphi(z_-)), \\ Cov\left[Y_\theta,(Y_\theta-\chi)^+\right] &= P\{Y_\theta \ge \chi\} + \frac{3\sqrt{2}}{2}\cdot\sin\theta\cdot\cos\theta\cdot(\varphi(z_+)-\varphi(z_-)) \\ &\quad + \sin^2\theta\cdot(z_+\cdot\varphi(z_+) - z_-\cdot\varphi(z_-)), \end{aligned}$$

and

$$\begin{aligned} Cov\left[Y_\theta^2,(Y_\theta-\chi)^+\right] = {} & \left(3-\sin^2\theta\right)\cdot\sqrt{2}\cdot\sin\theta\cdot P\{Y_\theta \ge \chi\} \\ & + \left(1+\frac{3\sqrt{2}}{2}\cdot\chi\cdot\sin\theta + 8\sin^2\theta\right) \\ & \cdot\cos\theta\cdot(\varphi(z_+)-\varphi(z_-)) \\ & + \left(\frac{3\sqrt{2}}{2} + \chi\cdot\sin\theta + 2\sqrt{2}\sin^2\theta\right) \\ & \cdot\sin\theta\cdot(z_+\cdot\varphi(z_+) - z_-\cdot\varphi(z_-)), \end{aligned}$$

where $z_+ \equiv z_+(\theta,\chi)$ and $z_- \equiv z_-(\theta,\chi)$.

Observe that $Y_\theta \wedge \chi = Y_\theta - (Y_\theta - \chi)^+$, so we also have

$$E\left[(Y_\theta \wedge \chi)\right] = -E\left[(Y_\theta - \chi)^+\right],$$
$$\operatorname{Cov}\left[Y_\theta, (Y_\theta \wedge \chi)\right] = 1 - \operatorname{Cov}\left[Y_\theta, (Y_\theta - \chi)^+\right],$$

and

$$\operatorname{Cov}\left[Y_\theta^2, (Y_\theta \wedge \chi)\right] = \sigma_\theta - \operatorname{Cov}\left[Y_\theta^2, (Y_\theta - \chi)^+\right].$$

5 Numerical examples

Now let $f(q, v, \sigma_\theta)$, $g(q, v, \sigma_\theta)$, and $h(q, v, \sigma_\theta)$, respectively, represent the right hand sides of (4.8)–(4.10). We can numerically solve these equations by applying the standard Euler scheme (see Strogatz [22] on page 32), which implies that if Δt is the time step, then we make the following substitutions

$$q(t+\Delta t) \longleftarrow q(t) + \Delta t \cdot f(q(t), v(t), \sigma_\theta(t)),$$
$$\frac{q(t+\Delta t) + v(t+\Delta t)}{2} \longleftarrow \frac{q(t) + v(t)}{2} + \Delta t \cdot g(q(t), v(t), \sigma_\theta(t)),$$

and

$$\left(\frac{q}{6} + \frac{v}{2} + \frac{\left(\sigma_\theta \cdot \sqrt{v^3}\right)}{3}\right)(t+\Delta t) \longleftarrow \left(\frac{q}{6} + \frac{v}{2} + \frac{\left(\sigma_\theta \cdot \sqrt{v^3}\right)}{3}\right)(t)$$
$$+\Delta t \cdot h(q(t), v(t), \sigma_\theta(t)).$$

If $v = 0$, then the GVA and GSA equations are discontinuous. We can avoid this case, by assigning the non-zero initial value of v to be $v(0) = 10^{-5}$. The error for such a forward Euler scheme is of order Δt as it becomes small. For our numerical examples, we set $\Delta t = 10^{-3}$.

For our simulations, the time step here is also $\Delta t = 10^{-3}$. The sample variance averages around 25.0 and the number of total runs is 10,000. This gives us a 99 % confidence interval close to plus or minus 0.15 for a mean queue number that averages over time around 12.0. For the histograms of the distribution at given time points, since the queueing process is integer valued, we always set the bin size to be 1. In Appendix B we describe, with a simple (i.e., abandonment rate zero) multi-server example, the queueing simulation algorithm that we used.

5.1 General numerical case and histogram plots

In Fig. 5, we see that the estimate for the mean behavior is nearly identical for GVA and GSA. This give us confidence that our method at least gives us good estimates for the mean behavior. A more refined picture of the difference of GVA and GSA is

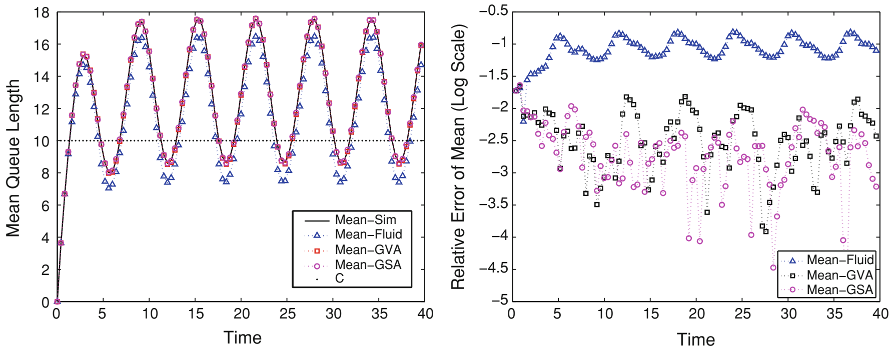

Fig. 5 Comparison of simulated, GVA, and GSA means (*left*). Relative error of means using GVA and GSA approximations (*right*)

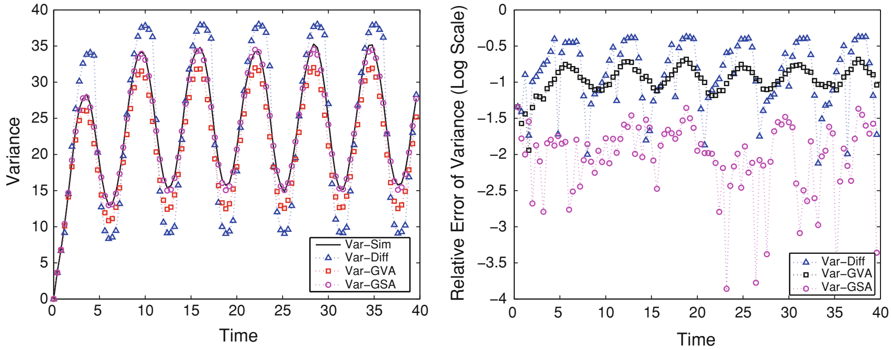

Fig. 6 Comparison of simulated, GVA, and GSA variances (*left*). Relative error of variances using GVA and GSA approximations (*right*)

given in the right hand graph of Fig. 5. This shows that GSA is slightly outperforming the GVA estimate for the mean behavior since the $\log_{10}$ or common logarithm of the relative error for GSA is smaller. In Fig. 6 we see that the variance estimate for the GSA is much better than the estimate by GVA. The curves for the simulated variance and the GSA variance are nearly identical. When one refines the picture on the right of 6 one notices that GSA is outperforming GVA by one order of magnitude. Finally, in Fig. 7 we compare the estimates of the skewness from the GSA method with the simulated skewness of the queueing process. A more refined look reveals a worst case accuracy of 10 % relative error.

In addition to exploring the mean, variance, and skewness of the queueing process, we also compared the queue length distributions for fixed time points. The set of points that we considered were $\{1.0, 1.4, 2.75, 4.90\}$. These points represent the following different queueing time periods: underloaded ($t = 1.0$), underloaded to critically loaded ($t = 1.4$), fully overloaded ($t = 2.75$), and overloaded to critically loaded ($t = 4.90$). These periods correspond to the ones considered in Massey [15] as well as Mandelbaum and Massey [12]. Figure 8 shows the queue length distributions for each

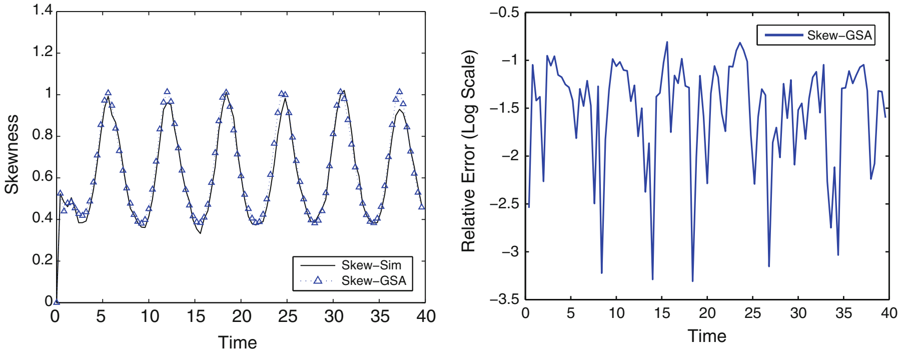

Fig. 7 Comparison of simulated and ODE approximation of skewness (*left*). Relative error of GSA approximation (*right*)

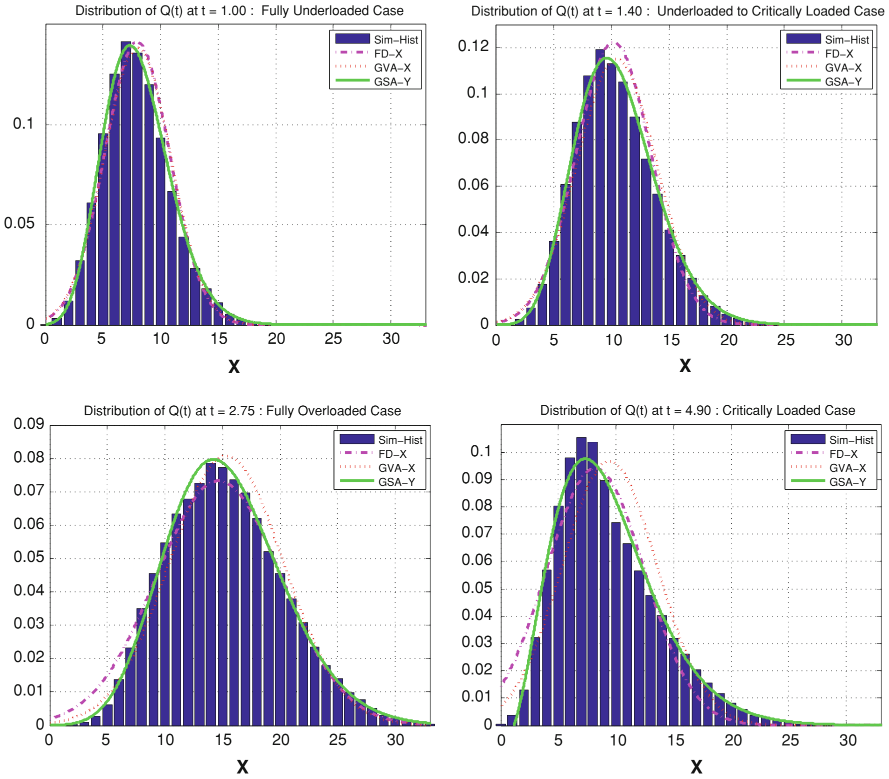

Fig. 8 Comparison of fluid/diffusion, GVA, and GSA densities to histograms of simulated queue length distributions

of these time points. As we can see in the bottom two histograms that accommodating distributions with asymmetric tails is a key feature of the GSA method, over both GVA and a Gaussian distribution with a fluid mean and a diffusion variance.

5.2 Square wave arrival rate

Now consider the case of discontinous arrival rate as given by Table 2. In the top left of Fig. 9 we see that the GSA and GVA are doing a better job of approximating the mean behavior of the queueing process than the fluid limit. This is also confirmed in the bottom right of Fig. 9 where we see that GSA is doing slightly better than GVA. In the top right of Fig. 9 we also see that the GSA is estimating the variance better than in the case of GVA and the diffusion limit. Finally, we see in the bottom left of Fig. 9 that the GSA provides a good estimate of the skewness of the queueing system, when the arrival rate is a discontinuous function of time. This provides some more evidence that the GSA method can handle rather arbitrary arrival rate functions that are not just sinusoidal. In this and all the subsequent numerical examples, we deliberately plot the GSA estimate graph twice on the bottom left and right. This makes it easier to compare the approximations to the mean and variance with the times that the skewness has a locally extreme value.

5.3 Dynamic staffing

This numerical example, given by Table 3, is for the case of the staffing function varying over time. In the top left of Fig. 10 we see that the GSA and GVA are doing a better job of approximating the mean behavior of the queueing process than the fluid limit. This is also confirmed in the bottom right of Fig. 10, where we see that GSA is doing moderately better than GVA. In the top right of Fig. 10 we also see that the GSA is estimating the variance better than in the case of GVA and the diffusion limit. Finally, we see in the bottom left of Fig. 10 that the GSA provides a quite good estimate of the skewness of the queueing system, when the arrival rate is a discontinuous function of time. This provides some more evidence that the GSA method can handle rather arbitrary staffing functions that are not just constant since our staffing function is piecewise constant over time.

5.4 High arrival rate and large number of servers

This numerical example, given by Table 4, shows that the GSA method can handle a large number of servers. On top left of Fig. 11, the fluid limit is comparable to GSA and GVA in estimating the mean behavior of the queueing process. This is also further supported in the bottom right of Fig. 11 where we see that it is achieving 10^{-2}

Table 2 Square wave parameters

Parameter	Value (at time t)
λ	17.0 if $2k \leq t/\pi < 2k+1$, otherwise 7.0
μ	1.0
c	10
β	0.25

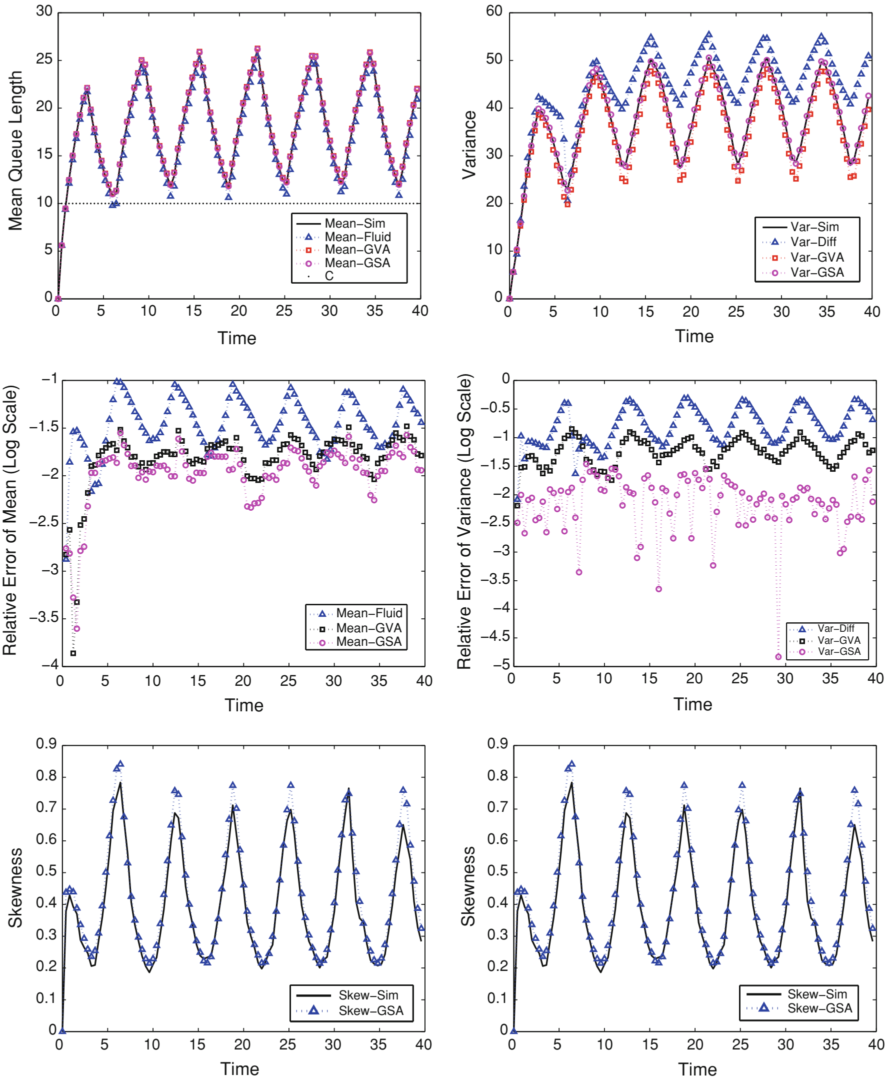

Fig. 9 Square wave arrival rate and constant staffing schedule $c = 10$

Table 3 Dynamic staffing parameters

Parameter	Value (at time t)
λ	$10.0 + 5.0 * \sin t$
μ	1.0
c	$\lceil \lambda(t) * 1.5 \rceil$
β	0.25

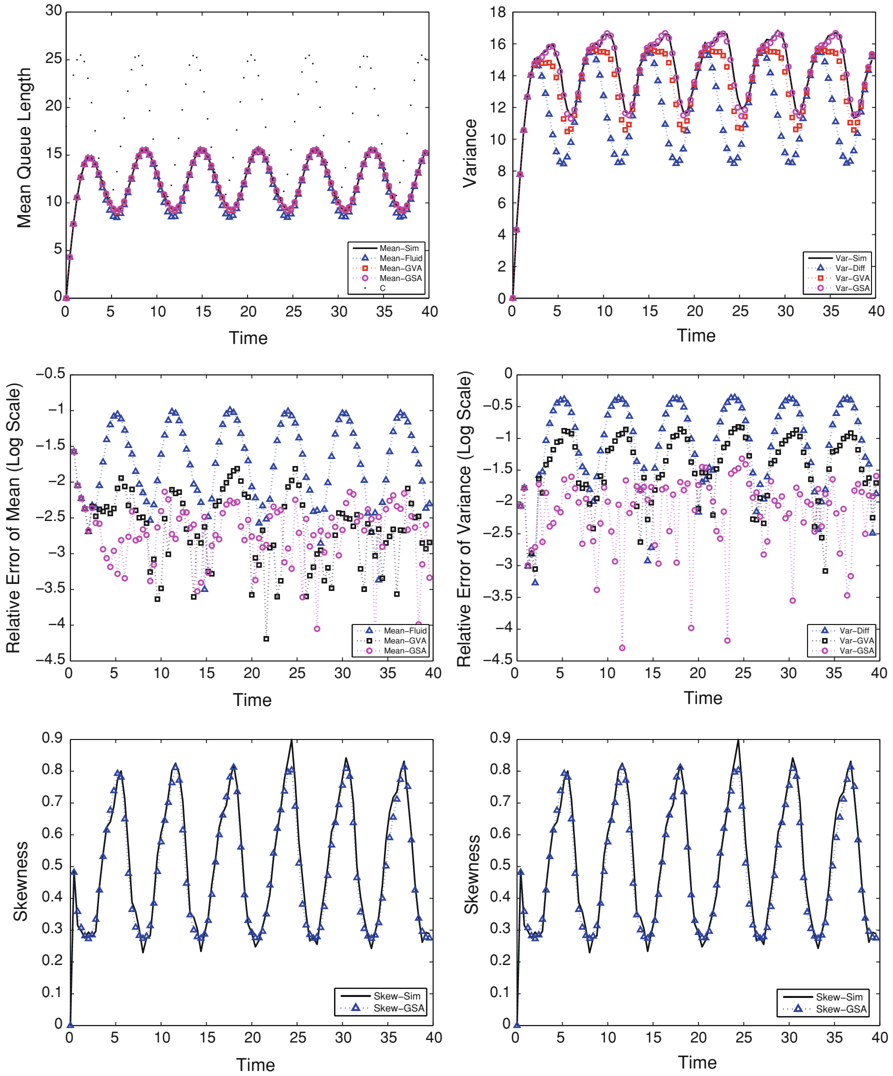

Fig. 10 Sinusoidal arrival rate and sinusoidal staffing schedule

Table 4 High arrival rate parameters

Parameter	Value (at time t)
λ	$100.0 + 50.0 * \sin t$
μ	1.0
c	100
β	0.25

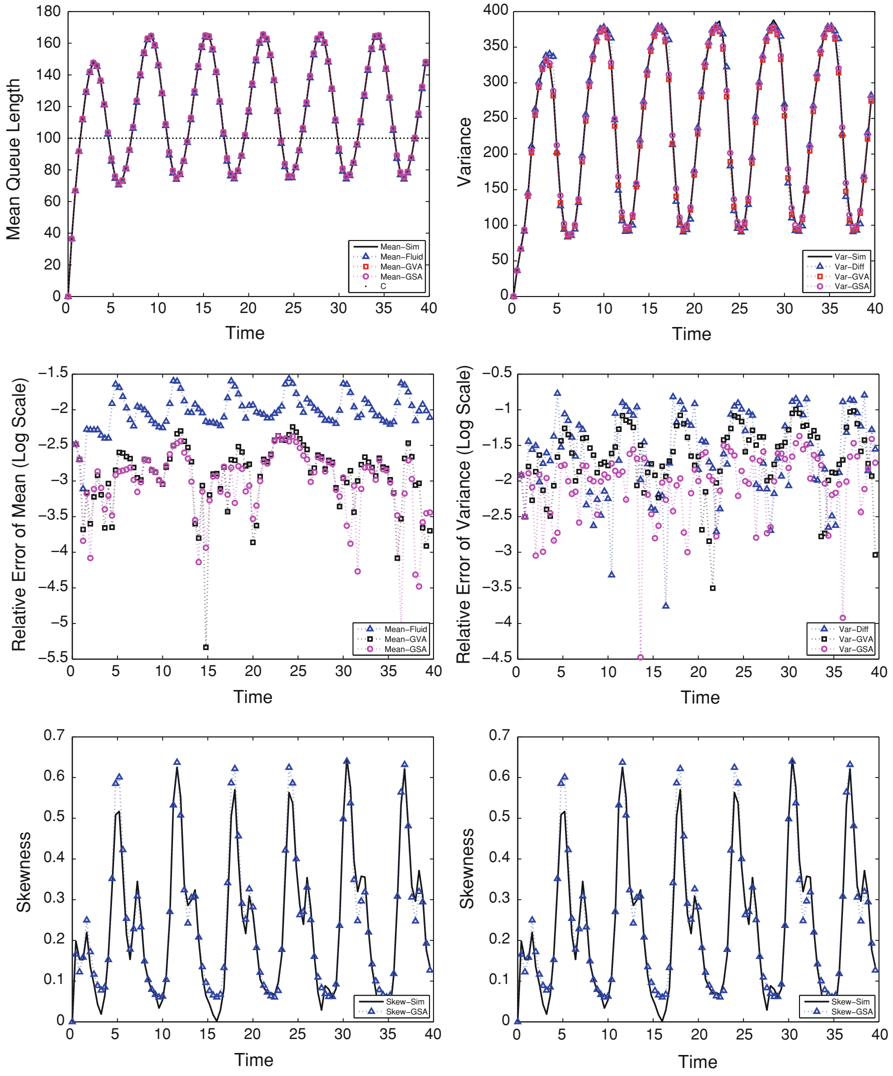

Fig. 11 Sinusoidal arrival rate and constant staffing schedule $c = 100$

accuracy. In the top right of Fig. 11 we also see that the GSA estimate of the variance is slightly better than both GVA and the diffusion limit. Finally, in the bottom left of Fig. 11 the GSA estimates the skewness of the queueing system by a 10 percent error at worst. Note that the skewness peaks when the queueing process is critically loaded. Moreover, the skewness values are about half the size when compared to the two server case. We expect this because as the number of arrivals and servers tends to infinity, the diffusion should be Gaussian, which implies a negligible skewness value.

Table 5 Two server parameters

Parameter	Value (at time t)
λ	$2.0 + \sin t$
μ	1.0
c	2
β	0.25

5.5 Two servers

The purpose of this numerical example, given by Table 5, is to show how the GSA method handles a small number of servers, i.e., $c = 2$. On top left of Fig. 12 we see that the GSA and GVA are doing a better job of estimating the mean behavior of the queueing process than the fluid limit. Moreover, in the bottom right of Fig. 12, we see that GSA is doing slightly better than GVA. In the top right graph, we also see that GSA is estimating the variance better than GVA or the diffusion limit. Finally, we see in the bottom left graph that the GSA provides a decent estimate of the skewness of the queueing system; however, it may need some improvement. It is also important to note that the system has a larger skewness value. One should expect that the skewness should be higher as one moves away from the limiting regime of large arrival rates and large number of servers.

5.6 Single server

Our last numerical example, given by Table 6, explores the case of a single server queue. Unlike some of earlier examples, the GSA does not provide a good estimate of the mean. The GSA estimate is better than GVA and the fluid limit; however, the GSA estimate is close to 10 percent relative error, as shown on the top left of Fig. 13. This is also further supported in the bottom right of Fig. 13 where we see that the fluid, GVA, and GSA equations are all achieving similar results. In the top right of Fig. 13, we also see that the GSA estimates the variance better than with GVA or the diffusion limit and this improvement is very large. Finally, we see in the bottom left of Fig. 11 that the GSA provides a decent estimate of the skewness of the queueing system; however, this estimate also has the potential for improvement. Furthermore, like in the two server case, the skewness value is quite high and for the most part is above 1, which suggests that the queueing process is less Gaussian here for the entire duration of time. This example suggest that the GSA is limited for the single server queue and might benefit from a kurtosis expansion in this case.

6 Conclusion and final remarks

The results we have discussed can be extended to any Markovian service network as we will discuss in a future paper. Their fluid and diffusion limits suggest that a Gaussian approximation captures a significant component of their stochastic behavior.

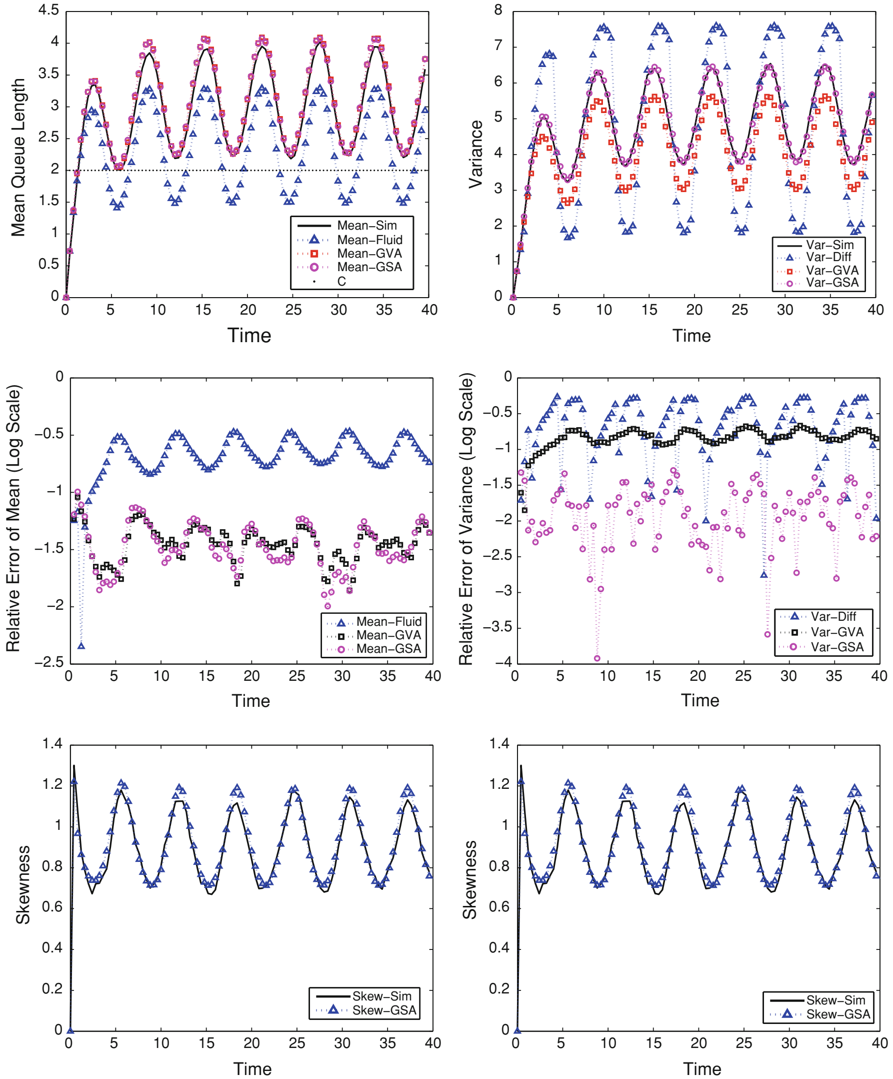

Fig. 12 Sinusoidal arrival rate and constant staffing schedule $c = 2$

Table 6 Single server parameters

Parameter	Value (at time t)
λ	$1.0 + 0.5 * \sin t$
μ	1.0
c	1
β	0.25

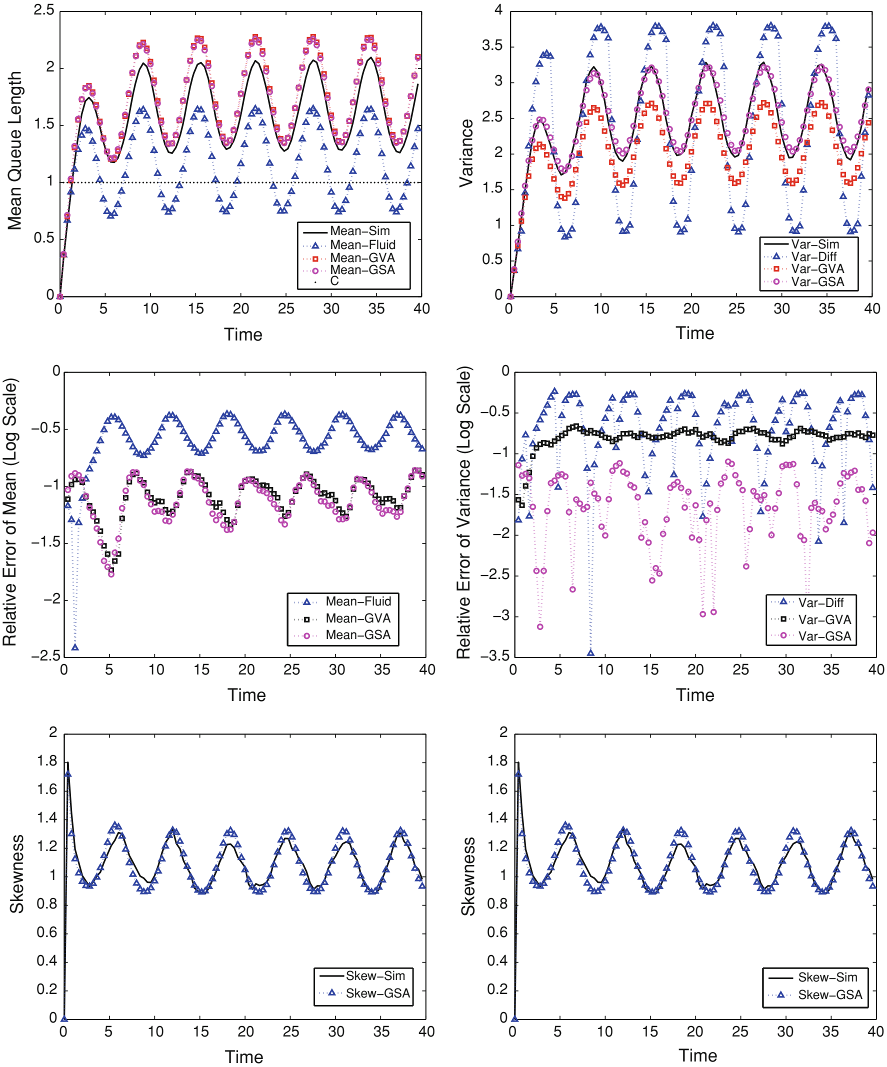

Fig. 13 Sinusoidal arrival rate and constant staffing schedule $c = 1$.

This makes the case for a Hermite polynomial Gaussian expansion of the queueing process distribution as a natural basis for an approximation. Another future paper will address the equations needed to estimate kurtosis. By GSA, we have constructed a "non-Gaussian" approximation of our queueing system that is a quadratic function of a standard Gaussian random variable X, using only the first three Hermite–Gaussian basis elements. GSA generates a finite dimensional dynamical system that improves our estimation of both the mean and variance of the original queueing process. This is especially needed during the times of critical loading, where the queueing distribution is less "Gaussian." GSA also helps to capture random phenomena like the asymmetry in the tail behavior of the queue. This method can be naturally extended to develop higher dimensional dynamical systems as closure approximations.

Acknowledgements This work was partially supported by NSF grant DMS-0807440.

Appendix A: Hermite polynomials and derivation of equations

The probabilistic Hermite polynomials as described in Nualart [16] are defined as:

$$h_n(x) = \frac{1}{\varphi(x)} \cdot \left(-\frac{d}{dx}\right)^n \varphi(x). \tag{6.1}$$

The first four Hermite polynomials are

$$h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = x^2 - 1, \quad h_3(x) = x^3 - 3x, \tag{6.2}$$

and in general they solve the recurrence relation

$$h_{n+1}(x) = x \cdot h_n(x) - n \cdot h_{n-1}(x). \tag{6.3}$$

We have the following Hermite polynomial generalization of Stein's lemma.

Theorem 6.1 *If X is a standard Gaussian random variable, then*

$$E\left[f(X) \cdot h_n(X)\right] = E\left[\frac{d^n}{dX^n} f(X)\right]$$

where f is any generalized function.

This follows from induction and integration by parts, since the Gaussian density is smooth (infinitely differentiable). From this result follows the orthogonality property of Hermite polynomials, namely

$$E\left[h_n(X) \cdot h_m(X)\right] = \begin{cases} m! & \text{if } n = m, \\ 0 & \text{if } n \neq m. \end{cases}$$

This follows from the fact that m derivatives of a degree n polynomial is always zero when $m > n$. Moreover, we get $m!$ when $m = n$ since $h_m(x)$ is always a monic polynomial. Since $h_0(X) = 1$, it now follows that all random variables of the form $h_n(X)$ have expectation zero, when $n \geq 1$.

If $f(X)$ is square integrable, then it can be written as an infinite sum of Hermite polynomials of X, i.e.,

$$f(X) = \sum_{n=0}^{\infty} \frac{1}{n!} E\left[\frac{d^n}{dX^n} f(X)\right] \cdot h_n(X),$$

where the convergence is with respect to the mean square (or L^2) norm. From this orthogonal series expansion, it follows that

$$E\left[f(X)\cdot g(X)\right]=\sum_{n=0}^{\infty}\frac{1}{n!}\cdot E\left[\frac{d^n}{dX^n}f(X)\right]\cdot E\left[\frac{d^n}{dX^n}g(X)\right]$$

or equivalently

$$\mathrm{Cov}\left[f(X),g(X)\right]=\sum_{n=1}^{\infty}\frac{1}{n!}\cdot E\left[\frac{d^n}{dX^n}f(X)\right]\cdot E\left[\frac{d^n}{dX^n}g(X)\right].$$

Proof of Theorem 4.1 This proof follows from the forward equations for the first, second, and third cumulant moments or Eq. 2.2 and our assumption that $\mathcal{Q}=q+\sqrt{v}\cdot Y_\theta$. We show it for two of the terms. All remaining expressions are derived similarly.

$$E\left[(\mathcal{Q}-c)^+\right]=E\left[(Y_\theta-\chi)^+\right]\cdot\sqrt{v}. \tag{6.4}$$

Using the property that the covariance of any random variable with a constant is zero, then we have

$$\mathrm{Cov}\left[\mathcal{Q},(\mathcal{Q}-c)^+\right]=\mathrm{Cov}\left[Y_\theta,(Y_\theta-\chi)^+\right]\cdot v. \tag{6.5}$$

□

Recall that for our GSA algorithm, we need to find simple expressions for terms like

$$\mathrm{Cov}\left[Y_\theta,(Y_\theta-\chi)^+\right]\quad\text{and}\quad\mathrm{Cov}\left[Y_\theta^2,(Y_\theta-\chi)^+\right].$$

Observing that $(Y_\theta-\chi)^+=(Y_\theta-\chi)\cdot\{Y_\theta\geq\chi\}$, motivates the next two lemmas.

Lemma 6.2 *For all $n\geq 1$, we have*

$$E\left[\frac{d^n}{dX^n}\{Y_\theta\geq\chi\}\right]=h_{n-1}(z_+)\cdot\varphi(z_+)-h_{n-1}(z_-)\cdot\varphi(z_-)$$

Proof We use the identity $\{Y_\theta\geq\chi\}=\overline{\{z_-\leq X\leq z_+\}}$, where we use the overline of an event to denote its complement. □

Lemma 6.3 *For all $n \geq 3$, we have*

$$\begin{aligned}
&E\left[\frac{d^n}{dX^n}(Y_\theta - \chi)^+\right] \\
&\quad = \cos\theta \cdot (h_{n-2}(z_+)\cdot\varphi(z_+) - h_{n-2}(z_-)\cdot\varphi(z_-)) \\
&\qquad + \sqrt{2}\cdot\sin\theta\cdot((z_+\cdot h_{n-2}(z_+) + h_{n-3}(z_+))\cdot\varphi(z_+) \\
&\qquad - (z_-\cdot h_{n-2}(z_-) + h_{n-3}(z_-))\cdot\varphi(z_-)).
\end{aligned}$$

with

$$\begin{aligned}
&E\left[\frac{d^2}{dX^2}(Y_\theta - \chi)^+\right] \\
&\quad = \cos\theta\cdot(\varphi(z_+) - \varphi(z_-)) + \sqrt{2}\cdot\sin\theta\cdot((z_+\cdot\varphi(z_+) - z_-\cdot\varphi(z_-)) + P\{Y_\theta \geq \chi\}).
\end{aligned}$$

and

$$E\left[\frac{d}{dX}(Y_\theta - \chi)^+\right] = \cos\theta\cdot P\{Y_\theta \geq \chi\} + \sqrt{2}\cdot\sin\theta\cdot(\varphi(z_+) - \varphi(z_-)). \quad (6.6)$$

Proof We prove the first equality to establish the idea and omit the proof of the latter two terms as they are proved in an identical fashion. Define $f(z_\pm) = f(z_+) - f(z_-)$ and $\zeta(z) \equiv z$. We then have

$$\begin{aligned}
&E\left[\frac{d^n}{dX^n}(Y_\theta - \chi)^+\right] \\
&\quad = E\left[h_{n-1}(X)\cdot\frac{d}{dX}(Y_\theta - \chi)^+\right] \\
&\quad = E\left[h_{n-1}(X)\cdot(Y_\theta - \chi)\cdot\frac{d}{dX}\{Y_\theta \geq \chi\}\right] \\
&\qquad + E\left[h_{n-1}(X)\cdot\{Y_\theta \geq \chi\}\cdot\frac{d}{dX}(Y_\theta - \chi)\right] \\
&\quad = E\left[h_{n-1}(X)\cdot(Y_\theta - \chi)\cdot\delta_{z\pm}(X)\right] + E\left[h_{n-1}(X)\cdot\{Y_\theta \geq \chi\}\cdot Y_\theta'\right] \\
&\quad = \cos\theta\cdot E\left[h_{n-1}(X)\cdot\{Y_\theta \geq \chi\}\right] + \sqrt{2}\cdot\sin\theta\cdot E\left[h_{n-1}(X)\cdot X\cdot\{Y_\theta \geq \chi\}\right] \quad (6.7) \\
&\quad = \cos\theta\cdot E\left[\frac{d^{n-1}}{dX^{n-1}}\cdot\{Y_\theta \geq \chi\}\right] + \sqrt{2}\cdot\sin\theta\cdot E\left[\frac{d^{n-1}}{dX^{n-1}}\cdot X\cdot\{Y_\theta \geq \chi\}\right] \\
&\quad = \cos\theta\cdot E\left[(h_{n-2}\cdot\delta_{z\pm})(X)\right] \\
&\qquad + \sqrt{2}\cdot\sin\theta\cdot\left(E\left[X\cdot\frac{d^{n-1}}{dX^{n-1}}\{Y_\theta \geq \chi\}\right] + (n-1)\cdot E\left[\frac{d^{n-2}}{dX^{n-2}}\{Y_\theta \geq \chi\}\right]\right) \\
&\quad = \cos\theta\cdot E\left[(h_{n-2}\cdot\delta_{z\pm})(X)\right] \\
&\qquad + \sqrt{2}\cdot\sin\theta\cdot\left(E\left[(h_{n-1}\cdot\delta_{z\pm})(X)\right] + (n-1)\cdot E\left[(h_{n-3}\cdot\delta_{z\pm})(X)\right]\right) \\
&\quad = \cos\theta\cdot E\left[(h_{n-2}\cdot\delta_{z\pm})(X)\right]
\end{aligned}$$

$$+\sqrt{2}\cdot\sin\theta\cdot\left(E\left[\left(\left(\zeta\cdot h_{n-2}-(n-2)\cdot h_{n-3}\right)\cdot\delta_{z\pm}\right)(X)\right]\right.$$
$$\left.+(n-1)\cdot E\left[\left(h_{n-3}\cdot\delta_{z\pm}\right)(X)\right]\right)=\cos\theta\cdot E\left[\left(h_{n-2}\cdot\delta_{z\pm}\right)(X)\right]$$
$$+\sqrt{2}\cdot\sin\theta\cdot E\left[\left(\left(\zeta\cdot h_{n-2}+h_{n-3}\right)\cdot\delta_{z\pm}\right)(X)\right]. \tag{6.8}$$

□

Now that we have established our main lemmas, we use them to obtain the following simple expressions for the covariance terms that are used in the forward equations for the mean, variance, and skewness of the GSA.

Proof of Theorem 4.3

$$\begin{aligned}\operatorname{Cov}\left[Y_\theta,(Y_\theta-\chi)^+\right] &= E\left[\frac{d}{dX}Y_\theta\right]\cdot E\left[\frac{d}{dX}(Y_\theta-\chi)^+\right]\\ &\quad+\frac{1}{2}\cdot E\left[\frac{d^2}{dX^2}Y_\theta\right]\cdot E\left[\frac{d^2}{dX^2}(Y_\theta-\chi)^+\right]\\ &=\cos\theta\cdot E\left[\frac{d}{dX}(Y_\theta-\chi)^+\right]\\ &\quad+\frac{1}{2}\cdot\sqrt{2}\cdot\sin\theta\cdot E\left[\frac{d^2}{dX^2}(Y_\theta-\chi)^+\right]\\ &=\cos\theta\cdot\Big(\cos\theta\cdot P\left\{Y_\theta\geq\chi\right\}+\sqrt{2}\cdot\sin\theta\cdot\varphi(z_\pm)\Big)\\ &\quad+\frac{\sin\theta}{\sqrt{2}}\cdot\Big(\cos\theta\cdot\varphi(z_\pm)+\sqrt{2}\cdot\sin\theta\cdot((\zeta\cdot\varphi)(z_\pm)\\ &\quad+P\{Y_\theta\geq\chi\})\Big)\\ &=P\left\{Y_\theta\geq\chi\right\}+\frac{3\sqrt{2}}{2}\cdot\sin\theta\cdot\cos\theta\cdot\varphi(z_\pm)\\ &\quad+\sin^2\theta\cdot(\zeta\cdot\varphi)(z_\pm).\end{aligned}$$

For the next set of calculations, we define

$$Y'_\theta\equiv\frac{d}{dX}Y_\theta \quad\text{and}\quad Y''_\theta\equiv\frac{d^2}{dX^2}Y_\theta \tag{6.9}$$

and use them when convenient. Moreover, for the second covariance term, we have

$$\begin{aligned}&\operatorname{Cov}\left[Y_\theta^2,(Y_\theta-\chi)^+\right]\\ &=E\left[\frac{d}{dX}Y_\theta^2\right]\cdot E\left[\frac{d}{dX}(Y_\theta-\chi)^+\right]+\frac{1}{2}\cdot E\left[\frac{d^2}{dX^2}Y_\theta^2\right]\cdot E\left[\frac{d^2}{dX^2}(Y_\theta-\chi)^+\right]\\ &\quad+\frac{1}{6}\cdot E\left[\frac{d^3}{dX^3}Y_\theta^2\right]\cdot E\left[\frac{d^3}{dX^3}(Y_\theta-\chi)^+\right]\end{aligned}$$

$$\begin{aligned}
&+\frac{1}{24} \cdot E\left[\frac{d^4}{dX^4} Y_\theta^2\right] \cdot E\left[\frac{d^4}{dX^4}\left(Y_\theta - \chi\right)^+\right] \\
&= 2\sqrt{2}\sin\theta\cos\theta \cdot E\left[\frac{d}{dX}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\frac{1}{2} \cdot 2\left(1+\sin^2\theta\right) \cdot E\left[\frac{d^2}{dX^2}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\frac{1}{6} \cdot 6\sqrt{2}\sin\theta\cos\theta \cdot E\left[\frac{d^3}{dX^3}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\frac{1}{24} \cdot 12\sin^2\theta \cdot E\left[\frac{d^4}{dX^4}\left(Y_\theta - \chi\right)^+\right] \\
&= 2\sqrt{2}\sin\theta\cos\theta \cdot E\left[\frac{d}{dX}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\left(1+\sin^2\theta\right) \cdot E\left[h_1(X) \cdot \frac{d}{dX}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\sqrt{2}\sin\theta\cos\theta \cdot E\left[h_2(X) \cdot \frac{d}{dX}\left(Y_\theta - \chi\right)^+\right] \\
&\quad +\frac{\sin^2\theta}{2} \cdot E\left[h_3(X) \cdot \frac{d}{dX}\left(Y_\theta - \chi\right)^+\right] \\
&= 2\sqrt{2}\sin\theta\cos\theta \cdot E\left[\{Y_\theta \geq \chi\} \cdot Y_\theta'\right] + \left(1+\sin^2\theta\right) \cdot E\left[h_1(X) \cdot \{Y_\theta \geq \chi\} \cdot Y_\theta'\right] \\
&\quad +\sqrt{2}\sin\theta\cos\theta \cdot E\left[h_2(X) \cdot \{Y_\theta \geq \chi\} \cdot Y_\theta'\right] + \frac{\sin^2\theta}{2} \cdot E\left[h_3(X) \cdot \{Y_\theta \geq \chi\} \cdot Y_\theta'\right] \\
&= 2\sqrt{2}\sin\theta\cos^2\theta \cdot P\left\{Y_\theta \geq \chi\right\} + \left(1+\sin^2\theta\right) \cdot E\left[\delta_{z\pm}(X) \cdot Y_\theta'\right] \\
&\quad +4\sin^2\theta\cos\theta \cdot E\left[\{Y_\theta \geq \chi\} \cdot X\right] + \left(1+\sin^2\theta\right) \cdot E\left[\{Y_\theta \geq \chi\} \cdot Y_\theta''\right] \\
&\quad +\sqrt{2}\sin\theta\cos\theta \cdot E\left[\left(h_1 \cdot \delta_{z\pm}\right)(X) \cdot Y_\theta'\right] + \frac{\sin^2\theta}{2} \cdot E\left[\left(h_2 \cdot \delta_{z\pm}\right)(X) \cdot Y_\theta'\right] \\
&\quad +\sqrt{2}\sin\theta\cos\theta \cdot E\left[h_1(X) \cdot \{Y_\theta \geq \chi\} \cdot Y_\theta''\right] + \frac{\sin^2\theta}{2} \cdot E\left[h_2(X) \cdot \{Y_\theta \geq \chi\} \cdot Y_\theta''\right] \\
&= 2\sqrt{2}\sin\theta\cos^2\theta \cdot P\left\{Y_\theta \geq \chi\right\} + \left(1+\sin^2\theta\right) \cdot E\left[\delta_{z\pm}(X) \cdot Y_\theta'\right] \\
&\quad +4\sin^2\theta\cos\theta \cdot E\left[\delta_{z\pm}(X)\right] + \sqrt{2}\sin\theta\left(1+\sin^2\theta\right) \cdot E\left[\{Y_\theta \geq \chi\}\right] \\
&\quad +\sqrt{2}\sin\theta\cos\theta \cdot E\left[\left(h_1 \cdot \delta_{z\pm}\right)(X) \cdot Y_\theta'\right] + \frac{\sin^2\theta}{2} \cdot E\left[\left(h_2 \cdot \delta_{z\pm}\right)(X) \cdot Y_\theta'\right] \\
&\quad +2\sin^2\theta\cos\theta \cdot E\left[h_1(X) \cdot \{Y_\theta \geq \chi\}\right] + \frac{\sin^2\theta}{2}\sqrt{2}\sin\theta \cdot E\left[h_2(X) \cdot \{Y_\theta \geq \chi\}\right] \\
&= 2\sqrt{2}\sin\theta\cos^2\theta \cdot P\left\{Y_\theta \geq \chi\right\} + \left(1+\sin^2\theta\right) \cdot E\left[\delta_{z\pm}(X) \cdot Y_\theta'\right] \\
&\quad +4\sin^2\theta\cos\theta \cdot E\left[\delta_{z\pm}(X)\right] + \sqrt{2}\sin\theta\left(1+\sin^2\theta\right) \cdot P\left\{Y_\theta \geq \chi\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{2}\sin\theta\cos\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]+\frac{\sin^2\theta}{2}\cdot E\left[\left(h_2\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]\\
&+2\sin^2\theta\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]+\frac{\sin^2\theta}{2}\sqrt{2}\sin\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\operatorname{Cov}\left[Y_\theta^2,(Y_\theta-\chi)^+\right]\\
&=\left(2\sqrt{2}\sin\theta-2\sqrt{2}\sin^3\theta+\sqrt{2}\sin\theta+\sqrt{2}\sin^3\theta\right)\cdot P\left\{Y_\theta\geq\chi\right\}\\
&\quad+\left(1+\sin^2\theta\right)\cdot E\left[\delta_{z\pm}(X)\cdot Y'_\theta\right]+6\sin^2\theta\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&\quad+\frac{\sqrt{2}}{2}\sin\theta\cos\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]+\frac{\sin^2\theta}{2}\cdot E\left[\left(h_2\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]\\
&\quad+\frac{\sqrt{2}}{2}\sin\theta\cos\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]+\frac{\sin^2\theta}{2}\sqrt{2}\sin\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&=\left(3-\sin^2\theta\right)\sqrt{2}\sin\theta\cdot P\left\{Y_\theta\geq\chi\right\}+\left(1+\sin^2\theta\right)\cdot E\left[\delta_{z\pm}(X)\cdot Y'_\theta\right]\\
&\quad+6\sin^2\theta\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&\quad+\frac{\sin\theta}{\sqrt{2}}\cdot E\left[\left(\left(\cos\theta\cdot h_1+\frac{\sin\theta}{\sqrt{2}}\cdot h_2\right)\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]\\
&\quad+\frac{\sqrt{2}}{2}\sin\theta\cos\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]+\frac{\sin^2\theta}{2}\sqrt{2}\sin\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&=\left(3-\sin^2\theta\right)\sqrt{2}\sin\theta\cdot P\left\{Y_\theta\geq\chi\right\}+\left(1+\sin^2\theta\right)\cdot E\left[\delta_{z\pm}(X)\cdot Y'_\theta\right]\\
&\quad+6\sin^2\theta\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]+\frac{\sin\theta}{\sqrt{2}}\cdot E\left[\left(\chi\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]\\
&\quad+\frac{\sin\theta}{\sqrt{2}}\cdot\cos\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\cdot Y'_\theta\right]+\frac{\sin^3\theta}{\sqrt{2}}\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&=\left(3-\sin^2\theta\right)\sqrt{2}\sin\theta\cdot P\left\{Y_\theta\geq\chi\right\}+\left(1+\sin^2\theta\right)\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&\quad+\left(1+\sin^2\theta\right)\sqrt{2}\sin\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]+6\sin^2\theta\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&\quad+\frac{\sin\theta}{\sqrt{2}}\cos\theta\cdot E\left[\left(\chi\cdot\delta_{z\pm}\right)(X)\right]+\chi\cdot\frac{\sin\theta}{\sqrt{2}}\sqrt{2}\sin\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&\quad+\frac{\sin\theta}{\sqrt{2}}\cdot\cos^2\theta\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&\quad+\frac{\sin\theta}{\sqrt{2}}\cdot\cos\theta\cdot\sqrt{2}\sin\theta\cdot E\left[\left(\left(h_2+1\right)\cdot\delta_{z\pm}\right)(X)\right]\\
&\quad+\frac{\sin^3\theta}{\sqrt{2}}\cdot E\left[\left(h_1\cdot\delta_{z\pm}\right)(X)\right]\\
&=\left(3-\sin^2\theta\right)\sqrt{2}\sin\theta\cdot P\left\{Y_\theta\geq\chi\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\left(1+\chi\cdot\frac{\sin\theta}{\sqrt{2}}+8\sin^2\theta\right)\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&+\left(\frac{3}{2}+\sin^2\theta\right)\sqrt{2}\sin\theta\cdot E\left[(h_1\cdot\delta_{z\pm})(X)\right]\\
&+\chi\cdot\frac{\sin\theta}{\sqrt{2}}\sqrt{2}\sin\theta\cdot E\left[(h_1\cdot\delta_{z\pm})(X)\right]\\
&+\cos\theta\sin\theta\sqrt{2}\cdot E\left[((\chi-\cos\theta\cdot h_1)\cdot\delta_{z\pm})(X)\right]\\
=&\left(3-\sin^2\theta\right)\sqrt{2}\sin\theta\cdot P\{Y_\theta\geq\chi\}\\
&+\left(1+\frac{3\sqrt{2}}{2}\cdot\chi\sin\theta+8\sin^2\theta\right)\cos\theta\cdot E\left[\delta_{z\pm}(X)\right]\\
&+\left(\frac{\sqrt{2}}{2}+\chi\cdot\sin\theta+2\sqrt{2}\sin^2\theta\right)\sin\theta\cdot E\left[(h_1\cdot\delta_{z\pm})(X)\right].
\end{aligned}
$$

□

Appendix B: Monte Carlo queueing simulation

To simulate the mean of our queueing process, we implement an algorithm similar to the one illustrated below for a Monte Carlo simulation of the transient mean for an $M/M/c$ queue. As shown in Fig. 14, the inner loop is for each i.i.d. Markov sample path labeled by "run." In theory, we are summing up all the sample path realizations of the number in the queue at a fixed "time." In practice, however, the sum of the previous instance, "time" minus "tick," is known. Determining which "next event time" for a given "run" exceeds the current "time" is now a "run" whose queue must be updated. Applying this same update of a customer arrival or departure to the total sum, or "queue run sum," updates the total sum. After we are done with all the "total runs," we then move forward in time by a small amount "tick." Care must be taken to make sure that the size tick is smaller than any of the average holding times for any state.

Using this approach, we do not have to store the discretized version of each sample path. For our numerical example, we would have to store 10,000 vectors of dimension 40,000. The latter number is the length of the time interval 40.0 divided by $\Delta t = 10^{-3}$. We only need to store a single vector of this type which corresponds to the sample mean of the number in the queue as it evolves over time.

Figure 15 illustrates what is going on inside the "Update" subroutine in more detail. At state n, the holding time for our $M/M/c$ queue has an exponential distribution with rate $\lambda + \mu * \min(n, c)$. The next transition state is $n + 1$ (or using the programming language of C, n++, see Kernighan and Richie for details [9]) with probability λ divided by the holding time rate. Otherwise, the next transition state is $n - 1$ (or n-- when programming in C), if this is possible, or nothing happens. We refer to the state as "queue state[run]" and the holding time rate as "event rate[run]."

Using a pseudo random number generator, we simulate a random variable U that is uniformly distributed on the interval (0, 1] (see Ross for more details [19]). The

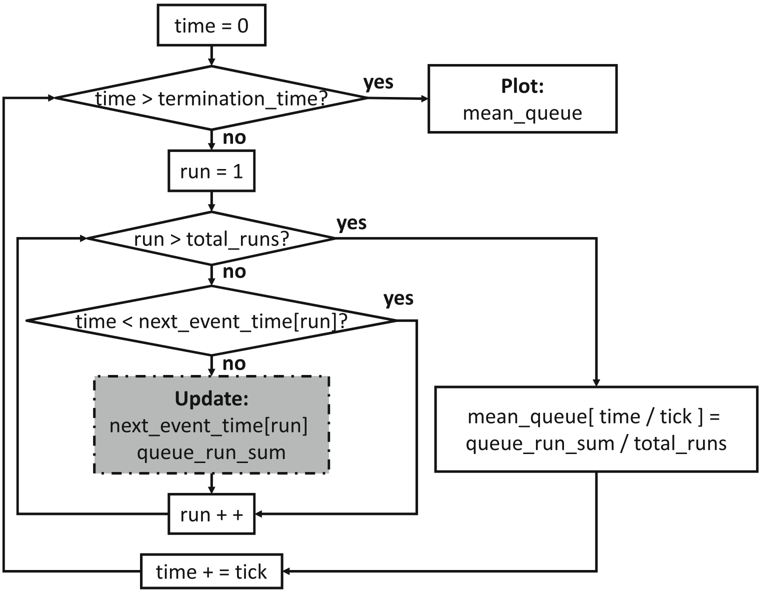

Fig. 14 Flow chart diagram for simulation algorithm

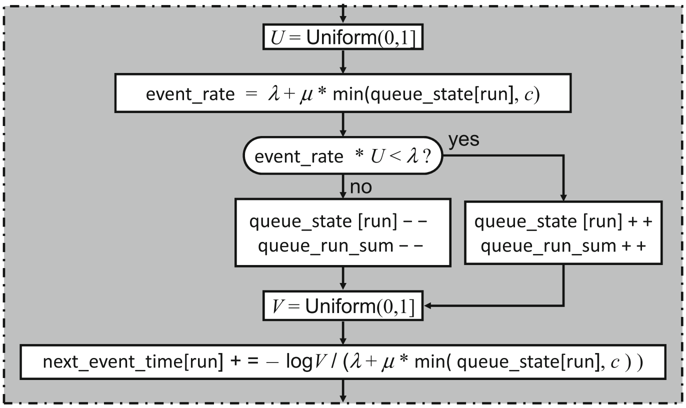

Fig. 15 Flow chart diagram for update subroutine of simulation algorithm

random event of the holding time rate times U being less than λ has the desired probability of λ divided by the holding time rate. When this event occurs, we then execute the arrival simulation as “queue state[run]++” and “queue run sum++”. Now that we are in our new state, we compute its holding time rate. If we generate another uniform random variable V, then $-\log V$ divided by the next holding time rate gives us an exponentially distributed random variable with the holding rate, again see Ross [19]. Adding this to the current time gives us the time of the next event or updates “next event time[run].”

Finally, to simulate this model with a time varying arrival rate function, according to a Poisson thinning method as found in Ross [19], two things need to be done. First, change the previous use of λ to the maximum possible value for the arrival rate over the given, finite time interval. Second, after a positive test for the product of U times the holding rate being less than the maximum arrival rate occurs, now test for almost the same event except that the maximum possible arrival rate is replaced by the arrival rate that happens at the current "time." If this is also true, then the arriving customer event is still valid. Otherwise, the arrival event does not happen.

For this given "time," the average of value for "mean queue" is "queue run sum" divided by "total runs." The confidence interval for simulating the mean is plus or minus 3 times the square root of the ratio for the simulation of the variance divided by the number of "total runs."

References

1. Eick, S., Massey, W.A., Whitt, W.: The physics of the $M(t)/G/\infty$ queue. Oper. Res. **41**, 400–408 (1993)
2. Fedoryuk, M.V.: Hermite polynomials. In: Hazewinkel, M. Encyclopaedia of mathematics. Springer. ISBN 978-1556080104. http://eom.springer.de/H/h046980.htm (2001)
3. Fortuin, C.M., Kasteleyn, P.N., Ginibre, J.: Correlation inequalities for some partially ordered sets. Commun. Math. Phys. **22**(2), 89–103 (1971)
4. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review and research prospects. Manuf. Serv. Oper. Manag. **5**(2), 79–141 (2003)
5. Halfin, S., Whitt, W.: Heavy-traffic limit theorems for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
6. Hampshire, R.C.: Dynamic queueing models for the operations management of communication services. Ph.D. thesis, Princeton University (2007)
7. Hampshire, R.C., Jennings, O.B., Massey, W.A.: A time varying call center design with Lagrangian mechanics. Probab. Eng. Inf. Sci. **23**(2), 231–259 (2009)
8. Hampshire, R.C., Massey, W.A.: A tutorial on dynamic optimization and applications to queueing systems with time-varying rates. Tutor. Oper. Res. **23**(2), 231–259 (2010)
9. Kernighan, B.W., Ritchie, D.M.: C programming language. PTR Prentice Hall, Englewood Cliffs (1988)
10. Khintchine, A.Y.: Mathematical methods in the theory of queueing (in Russian), Trudy Mat Inst. Steklov Vol. 49 (1955) (English translation by Charles Griffin and Co., London, 1960)
11. Ko, Y.M., Gautam, N.: Critically loaded time-varying multiserver queues: computational challenges and approximations. INFORMS J. Comput. (2012) to appear
12. Mandelbaum, A., Massey, W.A.: Strong approximations for time-dependent queues. Math. Oper. Res. **20**(1), 33–64 (1995)
13. Mandelbaum, A., Massey, W.A., Reiman, M.: Strong approximations for Markovian service networks. Queueing Syst. **30**, 149–201 (1998)
14. Mandelbaum, A., Massey, W.A., Reiman, M., Rider, B., Stolyar, A.: Queue lengths and waiting times for multi-server queues with abandonment and retrials. Telecommun. Syst. **21**, 149–172 (2002)
15. Massey, W.A.: Asymptotic analysis of the time dependent M/M/1 queue. Math. Oper. Res. **54**(2), 324–338 (1985)
16. Nualart, D.: The Malliavin calculus and related topics. Springer, New York (1995)
17. Palm, C.: Intensity variations in telephone traffic. Ericsson Tech. **44**, 1–189 (1943)
18. Prékopa, A.: On Poisson and composed Poisson stochastic set functions. Stud. Math. **16**, 142–155 (1957)
19. Ross, S.: Simulation, 4th edn. Elsevier Academic Press, Amsterdam (2006)
20. Rothkopf, M.H., Oren, S.S.: A closure approximation for the nonstationary $M/M/s$ queue. Manag. Sci. **25**(6), 522–534 (1979)

21. Stein, C.M.: Approximate computation of expectations. Lecture notes monograph series, vol. 7. Institute of Mathematical Statistics, Hayward (1986)
22. Strogatz, S.: Nonlinear dynamics and chaos. Westview Press, Boulder (1994)
23. Taaffe, M.R., Ong, K.L.: Approximating nonstationary $Ph(t)/M(t)/s/c$ queueing systems. Ann. Oper. Res. **8**, 103–116 (1987)