European Journal of Operational Research 254 (2016) 113–126

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Risk measures and their application to staffing nonstationary service systems

Jamol Pender*

School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, United States

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 9 March 2016
Available online 20 April 2016

Keywords:
Queues and service systems
Risk measures
Healthcare
Time inhomogeneous markov processes
Staffing

ABSTRACT

In this paper, we explore the use of static risk measures from the mathematical finance literature to assess the performance of some standard nonstationary queueing systems. To do this we study two important queueing models, namely the infinite server queue and the multi-server queue with abandonment. We derive exact expressions for the value of many standard risk measures for the $M_t/M/\infty$, $M_t/G/\infty$, and $M_t/M_t/\infty$ queueing models. We also derive Gaussian based approximations for the value of risk measures for the Erlang-A queueing model. Unlike more traditional approaches of performance analysis, risk measures offer the ability to satisfy the unique and specific risk preferences or tolerances of service operations managers. We also show how risk measures can be used for staffing nonstationary systems with different risk preferences and assess the impact of these staffing policies via simulation.

1. Introduction

Time varying queueing models such as the $M_t/G/\infty$ queue and the $M_t/M/k_t + M$ queue are standard models for describing the dynamics of large scale service systems like telecommunication systems, call centers, and healthcare systems like hospitals. To get a good understanding of the wide variety of applications of nonstationary queueing models, see for example (Khudyakov, Feigin, & Mandelbaum, 2010) for applications to call centers with interactive voice response and Yom-Tov and Mandelbaum (2014) for application to healthcare systems. However, staffing these systems appropriately and stabilizing salient performance measures such as the probability of delay and waiting times for these stochastic systems has been a long standing problem in the queueing literature for many years.

One of the first solutions for stabilizing the delay probabilities for multiserver queues without abandonment was developed by Jennings, Mandelbaum, Massey, and Whitt (1996). Jennings et al. (1996) develop a novel square root staffing algorithm that uses the offered load of an infinite server queue and the square root of the offered load for refinements to stabilize the delay probabilities in multi-server queues. In the case of exponential service times, it only requires the solution to a simple ordinary differential equation to find the appropriate staffing level. However, as noted in Feldman, Mandelbaum, Massey, and Whitt (2008) and Liu and Whitt (2012) and Massey and Pender (2013), this algorithm for stabilizing the delay probabilities does not stabilize the abandonment probabilities and other performance measures. Thus, Liu and Whitt developed a new approach that stabilizes the abandonment probabilities and mean delay using the combination of two infinite server queues.

Nonetheless, these algorithms for performance stabilization are only useful for a few performance measures that are well-studied in the queueing literature and are especially tailored for applications in telecommunications where there is no extreme consequence if a customer waits a long time for service. For instance, in a call center it is considered good performance if 99 percent of customers are served within 2 minutes and we might not care about the 1 percent of customers who might have extremely long wait times. However, in a healthcare or emergency care setting, patients with extremely long waiting times can be very costly to the hospital, especially if their health deteriorates while waiting and subsequently they die before being seen, see for example (Castillo, 2014). Consequently, it is not sufficient to just make sure that waiting times are short, but it is also important to make sure that even excessive waiting times are short in the context of healthcare.

To address the difference between application settings like telecommunications and healthcare, in this paper we propose analyzing the new problems in applications like healthcare with new

* Tel.: +1 6464185950; fax: +1 646 418 5950.
E-mail address: jjp274@cornell.edu

http://dx.doi.org/10.1016/j.ejor.2016.03.012
0377-2217/

ideas, namely using static risk measures from the mathematical finance literature. The advantage of using static risk measures over traditional approaches of performance analysis, is that the risk measure approach can be adapted to a manager's risk preferences and the particular application context. The fact that the risk measure approach can be adapted to different applications and in different contexts within a particular application is quite useful for managers of service systems. One example in healthcare is that patients with shortness of breath might be less willing to tolerate long waits than patients with an ankle sprain so a different risk measure should be used for those patients. Thus, this risk measure approach allows the manager of a service center such as hospital to choose his or her own risk preferences for the overall performance of the system as well as the individual parts of the system.

In order to develop this risk measure approach for general service systems, we need to specify a stochastic model for the dynamics of our service systems. In this paper, we begin with the infinite server queueing model. This model is very natural as a start since its dynamics are tractable in the stationary and nonstationary setting. Not only are the mean and variance dynamics tractable, but also the entire distribution is known for the infinite server queue when initialized with a Poisson distribution or at zero. Besides the fact that the infinite server queue is a relatively simple model, it is also an offered load model. Thus, the infinite server dynamics represents the system when an unlimited number of resources are available and serves as an lower bound for the dynamics of finite server systems.

In addition to the infinite server queue, we also analyze the canonical nonstationary Erlang-A queueing model. The nonstationary Erlang-A model assumes the customer arrival process is a nonhomogenous Poisson process with nonstationary arrival rate $\lambda(t)$. We also have k servers with i.i.d. service times that are exponentially distributed with mean $1/\mu$. Finally, all the customers have i.i.d. abandonment times that are also exponentially distributed with mean $1/\beta$. Although the Erlang-A model is a simple model for some complex realities, it is also very hard to find closed form expressions for many of the performance measures of interest in the nonstationary setting. Thus, we must find approximations of the Erlang-A that are accurate and more tractable in terms of providing closed form expressions for performance measures of interest.

One standard method would be to use the fluid and diffusion limits of Mandelbaum, Massey, and Reiman (1998). However, it is well known that for small values of the scaling parameter η, the fluid and diffusion limits are not warranted. Moreover, when the mean queue length is near the number of servers, the fluid and diffusion limits are not Gaussian. Thus, in this work, we use another approximation to accurately estimate the queue length process. This approximations is known as the Gaussian variance approximation (GVA) of Massey and Pender (2011) and uses a Gaussian surrogate distribution to approximate the queue length dynamics. With this approximation for the queue length dynamics, we then approximate various risk measures for the queue length process and illustrate their performance as tools for staffing the system. We are not the first to study staffing issues in queues, see for example (Engblom & Pender, 2014; Pender, 2015; Stolletz, 2008; Tirdad, Grassmann, & Tavakoli, 2016; Yarmand & Down, 2013), however, we are the first to use risk measures in this context.

1.1. Contributions

To the best of our knowledge our contributions in this work are the following.

- We are the first to illustrate how static risk measures from the mathematical finance literature can be used in the context of server staffing and performance analysis in queueing theory.
- We derive explicit approximate staffing schedules for various risk measures that are widely used in the financial community and derive closed form expressions for the values of risk measures under Poisson and Gaussian distributional assumptions.
- We use the risk measures as staffing procedures and assess the results through comparing standard performance measures such as the probability of delay and abandonment probabilities.

1.2. Outline of paper

The rest of the paper is as follows. In Section 2, we introduce the concept of risk measures and provide several examples of risk measures. We also introduce the concept of functional risk measures, which will also be used throughout the rest of the paper. In Section 3, we start with the infinite server queue and derive closed form formulas for several risk measures for the queueing process. In Section 4, we introduce the Erlang-A model and several approximations for it. In Section 5, we use the approximations for the Erlang-A model queueing model and derive closed form expressions for the risk measures of the queueing model. In Section 6, we give numerical results and describe the impact of using the risk measures for staffing the system. We give examples of some extensions and conclude with final remarks in Section 7.

2. Static risk measures

One of the central goals in mathematical finance is to assess the risk of financial positions. The risk of a financial position may be seen as the capital reserves that a bank should hold in response to the risk it exposes itself to. Inspired by this notion of risk as a minimal capital reserve and by the shortcomings of *V@R*, Artzner et al. (1997,1999) introduced an axiomatic approach to coherent risk measures. The goal of a coherent risk measure is to quantify the risk of X by a number $\rho(X)$. It is our goal in this paper to introduce this notion of risk measures into the world of queueing theory where there are analogous notions of risk and reserves. In fact, in the context of queueing theory and staffing, the notions of risk and reservers can be viewed as the number of staff needed to maintain a specific quality of service level. Before we describe how various risk measures are related to various performance quantities in the service systems literature, we give a brief overview of risk measures to make the paper self-contained for the reader's convenience.

Definition 2.1. A mapping $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is called a monetary risk measure if $\rho(0)$ is finite and if ρ satisfies the following conditions for X,Y $\in \mathcal{X}$.

- Monotonicity: If X $\leq$ Y , then $\rho(X) \geq \rho(Y)$.
- Cash Invariance: If m $\in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$

These two conditions are very necessary to define risk measure. It is clear that if X is always smaller than Y under every scenario $(\forall\omega)$, then the risk associated with X should be higher than the risk associated with Y. Moreover, if we add cash to our position, it should reduce the risk of that position because cash is not a risky asset.

Definition 2.2. A monetary risk measure ρ is called a convex or quasi-convex risk measure if ρ satisfies the following condition for X,Y $\in \mathcal{X}$.

- Convex: If $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$ for all $\lambda \in$ [0,1].
- Quasi-Convex: If $\rho(\lambda X + (1-\lambda)Y) \leq \max\{\rho(X), \rho(Y)\}$ for all $\lambda \in$ [0,1]

From a financial perspective this is an important property for a risk measure since, it agrees with the theory of diversification in portfolio theory. Thus, if you invest in different assets that are not perfectly positively correlated, then you should naturally lower the risk of your overall portfolio.

Definition 2.3. A convex risk measure ρ is called a coherent risk measure if ρ satisfies the following condition for X,Y $\in \mathcal{X}$.

- Positive Homogeneity: If $\rho(\lambda X) = \lambda\rho(X)$ for all $\lambda \geq 0$.

Note that when we assume that the convex risk measure is positive homogenous, we get for free that the risk measure is sub-additive. Moreover, positive homogeneity and sub-additivity imply that the risk measure is convex. We now give some examples of more common risk measures that are used frequently in the finance literature.

2.1. Examples of risk measures

One of the most important and widely used risk measures is the Value at Risk risk measure (V@R).

2.1.1. Value at risk

The Value at Risk (V@R) is defined as:

$$V@R(X,\epsilon) = \inf\{y \mid \mathbb{P}\{X \leq y\} \geq \epsilon\}. \tag{2.1}$$

The $V@R(X, \epsilon)$ is interpreted as the minimal amount of money that an agent must add to a position X such that, with probability not greater than ϵ, he will not become bankrupt. However in the context of queueing theory and optimal server staffing the $V@R(X, \epsilon)$ can be interpreted as the number of agents you need to staff your system to keep the probability of waiting for service below some pre-specified tolerance value ϵ. Thus, there are important interpretations of risk measures from a queueing theory perspective.

Unfortunately, $V@R(X, \epsilon)$ does not satisfy sub-additivity and therefore is not a coherent risk measure unless the distributions involved are jointly elliptically distributed. One typical example of a coherent risk measure is the Average Value at Risk (also called Tail-V@R or AV@R).

2.1.2. Average value at risk

The Average Value at Risk (AV@R) is defined as:

$$AV@R(X,\epsilon) = \frac{1}{1-\epsilon}\int_{\epsilon}^{1} V@R(X,\nu)d\nu$$

and when the distribution of X is continuous, we have that

$$\begin{aligned} AV@R(X,\epsilon) &= \frac{1}{1-\epsilon}\int_{\epsilon}^{1} V@R(X,\nu)d\nu \\ &= \frac{1}{1-\epsilon}\int_{\epsilon}^{1} \inf\{y|\mathbb{P}\{X \leq y\} \geq \nu\}d\nu \\ &= \mathbb{E}[X|X \geq V@R(X,\epsilon)] \end{aligned}$$

Hence, AV@R takes the average over all Values at Risk between 1 and $1-\epsilon$. If the distribution of X is continuous, this is equivalent to conditional expectation of a drop in the market and then taking the average over all these bad scenarios. In this respect, the AV@R is more robust than V@R to changes in the distribution since the V@R is a point value of a quantile and AV@R is the average of the tail behavior of the V@R. It can also be shown that the AV@R is greater than V@R for the same value of ϵ and the same random variable. Moreover, AV@R can be seen as a compromise between the V@R risk measure and the maximal loss of the random variable since it is larger than the V@R, but smaller than the maximal loss.

Like the V@R, AV@R has important meaning in the context of queueing theory. In queueing theory, it is an important goal to understand the behavior of the queueing process when the queue length exceeds the currently available number of servers. The AV@R allows one to compute the mean of the queue length when the queue is overloaded. Understanding this risk measure for queueing processes will allow us to staff our system when it is overloaded with customers. Unlike *V@R* the *AV@R* is a convex risk measure. This means for queueing theory that adding two different queues together should only lower the total staffing needed to properly staff both. In finance, the convexity is motivated by diversification, which is meant to lower risk, however, for the perspective of a manager of a queueing system, it can be interpreted as economies of scale. As we add more queues to the network, we would hope that our optimal staffing would be no worse than staffing the two queues separately.

2.1.3. Entropic risk measure

The Entropic risk measure is defined as:

$$\rho(X,\gamma) = \frac{1}{\gamma}\log\left(\mathbb{E}[e^{-\gamma X}]\right). \tag{2.2}$$

The Entropic risk measure also has a dual representation as

$$\rho(X,\gamma) = \sup_{\tilde{P}\in\mathcal{M}_1}\left\{E^{\tilde{P}}[-X] - \frac{1}{\theta}H(\tilde{P}|P)\right\} \tag{2.3}$$

where

$$H(\tilde{P}|P) = E\left[\frac{d\tilde{P}}{dP}\log\frac{d\tilde{P}}{dP}\right] \tag{2.4}$$

is the relative entropy of measures $\tilde{P}$ and P and where $\tilde{P}$ is absolutely continuous with respect to P. This dual representation of the Entropic risk measure can be view as the worst case of the expected loss under measure $\tilde{P}$, corrected by a penalty term, where the probabilistic model $\tilde{P}$ is penalized proportional to the deviation of $\tilde{P}$ from P, measured by the relative entropy. Moreover, in the realm of mathematical finance the Entropic risk measure can be viewed as the indifference price of an investor with the constant risk aversion utility function $u(x) = 1 - e^{-x}$ and is also widely used when there is incomplete information or uncertainty about the models for the market dynamics. See for example (Rudloff, Sass, & Wunderlich, 2008). Although we do not see an immediate analogous connection to queueing systems like the other risk measures, it is nonetheless important since the Entropic risk measure is a scaled version of the cumulant generating function, which is an important probabilistic quantity to understand.

2.1.4. Mean-Variance risk measure

Another important risk measure that is popular in the financial literature is the Mean-Variance risk measure. The Mean-Variance risk measure is defined as:

$$\rho(X,\gamma) = E[X] + \gamma\cdot\mathrm{Var}[X]. \tag{2.5}$$

The Mean-Variance risk measure is quite popular since it is made up of the first two cumulants of the distribution of the random variable. It is also most notably used in the capital asset pricing model known as the (CAPM). It allows one to trade off expected return with the variance of that return in a simple and elegant manner. A slight modification of the Mean Variance risk measure leads to the following risk measure.

2.1.5. Mean-deviation risk measure of order p

The last risk measure that we describe here is the Mean-Deviation risk measure of order p. The Mean-Deviation risk measure of order p is defined as:

$$\rho(X,\gamma,p) = E[X] + \gamma\cdot(E[|X - E[X]|^p])^{1/p}. \tag{2.6}$$

This is somewhat of a generalization of the Mean-Variance risk measure since it allows different values of p other than 2. When $p = 2$, we have a similar risk measure like the Mean-Variance risk measure i.e.

$$\rho(X, \gamma, 2) = E[X] + \gamma \cdot (E[|X - E[X]|^2])^{1/2} \tag{2.7}$$

$$= E[X] + \gamma \cdot \sqrt{\text{Var}[X]}. \tag{2.8}$$

We will demonstrate later in the paper that this risk measure is important for staffing nonstationary systems. Currently, we limit the number of risk measures that we explain in detail, however, see (Cheridito & Li, 2009) for more examples of risk measures. To find out more about risk measures and their applications in finance or optimization, see (Cheridito & Li, 2008; 2009; Ruszczyński & Shapiro, 2006).

2.2. Functional risk measures

In addition to understanding the performance of risk measures with respect to random variables, it is also important to understand the performance with respect to functions of those random variables. When we consider a risk measure with respect to a function of a random variable, we call these *functional risk measures*.

2.2.1. Functional value at risk

The functional Value at Risk (V@R) is defined as:

$$V@R(f(X), \epsilon) = \inf\{y \mid \mathbb{P}\{f(X) \le y\} \ge \epsilon\} \tag{2.9}$$

$V@R(f(X), \epsilon)$ is interpreted as the minimal amount of money that an agent must add to a position $f(X)$ such that, with probability not greater than ϵ, he will not become bankrupt. However in the context of queueing theory and optimal staffing the $V@R(f(X), \epsilon)$ can be interpreted as the number of agents you need to staff your system to keep the probability of delay below some prespecified tolerance value ϵ at a particular time t.

2.2.2. Functional average value at risk

The functional Average Value at Risk (AV@R) is defined as:

$$\begin{aligned} AV@R(f(X), \epsilon) &= \frac{1}{1-\epsilon}\int_{\epsilon}^{1} V@R(f(X), \nu)d\nu \\ &= \frac{1}{1-\epsilon}\int_{\epsilon}^{1} \inf\{y|\mathbb{P}\{f(X) \le y\} \ge \nu\}d\nu \\ &= \mathbb{E}[f(X)|f(X) \ge V@R(f(X), \epsilon)] \end{aligned}$$

2.2.3. Functional mean-variance risk

The functional mean-variance risk measure is defined as:

$$\rho(f(X), \gamma) = E[f(X)] + \gamma \cdot \text{Var}[f(X)] \tag{2.10}$$

2.2.4. Functional mean-deviation risk measure of order p

The functional mean-variance risk measure of order p is defined as:

$$\rho(f(X), \gamma, p) = E[f(X)] + \gamma \cdot (E[|f(X) - E[f(X)]|^p])^{1/p} \tag{2.11}$$

2.2.5. Functional entropic risk

The functional entropic risk measure is defined as:

$$\rho(f(X), \gamma) = \frac{1}{\gamma}\log\left(\mathbb{E}[e^{-\gamma \cdot f(X)}]\right) \tag{2.12}$$

Now that we have define a general class of risk measures that may be applicable it is our hope that these risk measures can give insight into the performance of queueing systems under different types of management since each management style has a different appetite for risk. As an example, in healthcare it may not be sufficient for a manager to control the average time that customer is delayed for emergency care. In fact it is more reasonable to control average time that a customer is delayed given that they are delayed. This quantity is more realistic for hospitals because it is not a concern if a customer is taken into the ER immediately. It is only a concern of the manager how long customers will wait, given that they have to wait and there is no room for them when they are admitted. In this particular case, a manager would choose to staff the hospital using the AV@R since this risk measure has the ability to condition on the customers being delayed when they arrive to the ER. In the same context, a manager also may choose a different risk measure for different times as well. For instance, during a terrorist attack or catastrophic event, hospital managers may want to staff for a worse case scenario. In this case, one would use the entropic risk measure with a high value of γ. As we have seen the entropic risk measure has the interpretation of being the worst case scenario as $\gamma \to \infty$.

3. The infinite server queue

In this section, we give a brief introduction to queueing theory, its applications, and also describe some of the simple queueing models that we will analyze in this paper. Queueing theory has the beginnings of its history in the context of telecommunications. Queueing theory was invented by a Danish engineer, Agner Erlang, who worked for the Copenhagen Telephone Exchange. He published the first paper on what would now be called queueing theory in 1909 and this work developed stochastic models for callers that dropped due to frustration from waiting for an operator. Simple queueing models are often denoted by Kendall notation A/B/C/D/E, where A stands for the distribution of arrivals, B stands for the distribution of service times, C stands for the number of servers, D stands for the waiting room capacity of the queue, and finally E stands for the service discipline. For example, the M/G/1/∞/FIFO queue represents Poisson arrivals, general service times, one server, an infinite waiting room, and the first in first out service discipline. More recently, some queueing models include customer abandonment and these are often denoted by a $+G$ after the number of servers. Queues have a variety of application areas such as telecommunications, healthcare, finance (limit order books), transportation, and data centers just to name a few. Thus, queueing theory is an important area of research and we intend to connect this literature with the risk measure literature.

To begin our analysis of stochastic queueing models, we start with the $M_t/G/\infty$ queueing model. There are two main reasons to start with the $M_t/G/\infty$ queueing model. The first is that the $M_t/G/\infty$ queue is very tractable since the distribution is known in closed form. The second reason is that the $M_t/G/\infty$ queue is the best type of queue one can hope for where everyone is served immediately and no one ever waits for service. In this regard, the $M_t/G/\infty$ infinite server queue is a lower bound for queueing models with a finite number of servers and without abandonment since it represents the dynamics if the manager had access to an infinite amount of resources and is not resource constrained.

3.1. The $M_t/G/\infty$ queue

In this section, we derive closed form formulas for risk measures for the $M_t/G/\infty$ queueing model, which exploits the results of Eick, Massey, and Whitt (1993) for the time varying infinite server queue. In the paper of Eick et al. (1993), they use the properties of the Poisson arrival process and use Poisson random measure arguments to show that the $M_t/G/\infty$ queue $Q^{\infty}(t)$, has a Poisson distribution with time varying mean $q^{\infty}(t)$. The exact analysis of the infinite server queue is often useful since it represents the dynamics of the queueing process if there were an unlimited amount of resources to satisfy the nonstationary demand process. As observed

in Eick et al. (1993), $q^\infty(t)$ has the following integral representation

$$q^\infty(t) = E[Q^\infty(t)] \tag{3.13}$$

$$= \int_{-\infty}^{t} \overline{G}(t-u)\lambda(u)du \tag{3.14}$$

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] \tag{3.15}$$

$$= E[\lambda(t-S_e)] \cdot E[S] \tag{3.16}$$

where $\lambda(u)$ is the time varying arrival rate and S represents a service time with distribution G, $\overline{G} = 1 - G(t) = \mathbb{P}(S > t)$, and S_e is a random variable with distribution that follows the stationary excess of residual-lifetime cdf G_e, defined by

$$G_e(t) \equiv \mathbb{P}(S_e < t) = \frac{1}{E[S]}\int_0^t \overline{G}(u)du = \frac{1}{E[S]}\int_0^t \mathbb{P}(S > u)du, \quad t \geq 0. \tag{3.17}$$

When the service time distribution is exponential, we know that the mean queue length, $q_\infty(t)$, solves the autonomous differential equation

$$\dot{q}_\infty = \lambda(t) - \mu \cdot q_\infty(t), \tag{3.18}$$

which is very easy to solve numerically. Moreover, from the standard theory of infinite server queues, the distribution of the queue length process is Poisson with mean $q^\infty(t)$ when initialized with a Poisson distributed number customers or initialized at zero. Using this fact, we now compute several risk measures for the infinite server queue to get a better understanding of the impact of these risk measures in a relatively simple context.

3.2. The $M_t/M_t/\infty$ queue

Theorem 3.1. The solution to the mean and variance of the $M_t/M_t/\infty$ queue with initial values of Q_0 and V_0 is given by

$$E[Q_t] = Q_0 \cdot \exp\left\{-\int_0^t \mu(s)ds\right\} + \left(\exp\left\{-\int_0^t \mu(s)ds\right\} \cdot \left(\int_0^t \lambda(s)\exp\left\{-\int_0^s \mu(r)dr\right\}ds\right)\right) \tag{3.19}$$

$$\text{Var}[Q_t] = E[Q_t] + (V_0 - Q_0) \cdot \exp\left\{-2\int_0^t \mu(s)ds\right\}. \tag{3.20}$$

Proof. Using the functional forward equations for the mean and variance as in Pender (2014a), we know that the mean and variance of the infinite server queue with a time varying arrival rate and service rate solves the following non-homogeneous differential equations

$$\dot{E}[Q_t] = \lambda(t) - \mu(t) \cdot E[Q_t]$$
$$\dot{\text{Var}}[Q_t] = \lambda(t) + \mu(t) \cdot E[Q_t] - 2 \cdot \mu(t) \cdot \text{Var}[Q_t].$$

Thus, since the mean is independent of the variance, we can solve the mean equation by standard ordinary differential equation theory. Since there is a uniqueness theory for simple first order equations, it only follows to show that the solution above actually solves the differential equation. Although integrating factors are standard in any text on ordinary differential equations, we sketch the solution for $E[Q_t]$ using the integrating factor method for ordinary differential equations. From the functional forward equations, we know the mean satisfies the following equation

$$\frac{d}{dt}E[Q_t] + \mu(t) \cdot E[Q_t] = \lambda(t).$$

Now multiply both sides by an integrating factor $e^{\int \mu(u)du}$ to get that

$$e^{\int \mu(u)du}\left(\frac{d}{dt}E[Q_t] + \mu(t) \cdot E[Q_t]\right) = \lambda(t)e^{\int \mu(u)du}.$$

Now using the product rule and chain rule of differentiation, we can write the left hand side as

$$e^{\int \mu(u)du}\left(\frac{d}{dt}E[Q_t] + \mu(t) \cdot E[Q_t]\right) = \frac{d}{dt}\left(\lambda(t)e^{\int \mu(u)du}\right)$$

where

$$\frac{d}{dt}\left(\lambda(t)e^{\int_0^t \mu(u)du}\right) = \lambda(t)e^{\int_0^t \mu(u)du}.$$

Now by integrating both sides, we have that

$$\int e^{\int \mu(u)du}\left(\frac{d}{dt}E[Q_t] + \mu(t) \cdot E[Q_t]\right) = \int \frac{d}{dt}\left(\lambda(t)e^{\int \mu(u)du}\right).$$

Finally, using the fundamental theory of calculus and dividing by the integrating factor, we have that

$$E[Q_t] = Q_0 \cdot \exp\left\{-\int_0^t \mu(s)ds\right\} + \left(\exp\left\{-\int_0^t \mu(s)ds\right\} \cdot \left(\int_0^t \lambda(s)\exp\left\{-\int_0^s \mu(r)dr\right\}ds\right)\right).$$

For the variance, we observe the fact that

$$\begin{aligned}\dot{\text{Var}}[Q_t] - \dot{E}[Q_t] &= \lambda(t) + \mu(t) \cdot E[Q_t] - 2 \cdot \mu(t) \cdot \text{Var}[Q_t] \\ &\quad - \lambda(t) + \mu(t) \cdot E[Q_t] \\ &= 2 \cdot \mu(t) \cdot E[Q_t] - 2 \cdot \mu(t) \cdot \text{Var}[Q_t] \\ &= 2 \cdot \mu(t) \cdot (E[Q_t] - \text{Var}[Q_t]) \\ &= -2 \cdot \mu(t) \cdot (\text{Var}[Q_t] - E[Q_t]).\end{aligned}$$

Since the last equation describes a first order linear differential equation for $\text{Var}[Q_t] - E[Q_t]$, we know that its solution is

$$\text{Var}[Q_t] - E[Q_t] = (V_0 - Q_0) \cdot \exp\left\{-2\int_0^t \mu(s)ds\right\}.$$

Moving the mean to the righthand side yields the solution for the variance as

$$\text{Var}[Q_t] = E[Q_t] + (V_0 - Q_0) \cdot \exp\left\{-2\int_0^t \mu(s)ds\right\}.$$

□

From this theorem, it is immediately clear that the mean is very close to the variance and when the queue is initialized at zero or with a Poisson distribution, then the mean and variance are equal for all times. Moreover, the Poisson assumption in this case is not unrealistic when the initial mean and initial variance are close or when the time is large enough and $\mu(t)$ is bounded away from zero.

3.3. Risk measures for the $M_t/G/\infty$ and $M_t/M_t/\infty$ queues

However, before we prove the results, we provide a lemma that shows that the tail distribution of a Poisson distribution can be expressed in terms of the incomplete gamma function.

Lemma 3.2.

$$\Gamma(c,x) = \sum_{m=c}^{\infty} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(c)} \int_0^x e^{-y} y^{c-1} dy \tag{3.21}$$

$$\overline{\Gamma}(c,x) = \sum_{m=0}^{c-1} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(c)} \int_x^{\infty} e^{-y} y^{c-1} dy. \tag{3.22}$$

where

$$\Gamma(c,x) = \frac{1}{\Gamma(c)} \int_0^x e^{-y} y^{c-1} dy \quad \text{and}$$
$$\overline{\Gamma}(c,x) = \frac{1}{\Gamma(c)} \int_x^{\infty} e^{-y} y^{c-1} dy \tag{3.23}$$

are the lower and upper incomplete gamma functions respectively. Moreover, we define $\Gamma^{-1}(x,\epsilon,k)$ and $\overline{\Gamma}^{-1}(x,\epsilon,k)$ to be the functional inverses of $\Gamma(c+k,x)$ and $\overline{\Gamma}(c+k,x)$ respectively.

Proof. See (Janssen, Van Leeuwaarden, Zwart et al., 2008) or Pender (2014c). □

Moreover, the Chen–Stein method can aid us in our computation of various quantiles and expectations related to the risk measures. We include the Chen–Stein theorem below for the convenience of the reader.

Theorem 3.3 (Chen–Stein)**.** Let Q be a random variable with values in $\mathbb{N}$. Then, Q has the Poisson distribution with mean rate q if and only if, for every bounded function $f : \mathbb{N} \to \mathbb{N}$,

$$\mathbb{E}[Q \cdot f(Q)] = q \cdot \mathbb{E}[f(Q+1)] \tag{3.24}$$

Proof. See (Peccati & Taqqu, 2011). □

Proposition 3.4. The Value at Risk, Average Value at Risk, Mean-Variance Risk, and Entropic Risk for the nonstationary infinite server queue with mean q is given by the following formulas

$$V@R(Q,\epsilon) = \overline{\Gamma}^{-1}(q,\epsilon) \tag{3.25}$$

$$AV@R(Q,\epsilon) = \frac{q \cdot \Gamma(q,\nu_\epsilon - 1)]}{\Gamma(q,\nu_\epsilon)} \tag{3.26}$$

$$\text{Mean} - \text{Variance}(Q,\gamma,p) = q + \gamma \cdot q \tag{3.27}$$

$$\text{Entropic}(Q,\gamma) = \frac{q \cdot (e^{-\gamma} - 1)}{\gamma} \tag{3.28}$$

Proof. For the Value at Risk, we have that

$$V@R(Q,\epsilon) = \inf\{y | \mathbb{P}\{Q \le y\} \ge \epsilon\} \tag{3.29}$$

$$= \inf\{y | \overline{\Gamma}(q,y) \ge \epsilon\} \tag{3.30}$$

$$= \overline{\Gamma}^{-1}(q,\epsilon). \tag{3.31}$$

For the Average Value at Risk, we first let $\nu_\epsilon = V@R(Q,\epsilon)$ and thus we obtain the following

$$AV@R(Q,\epsilon) = E[Q | Q > \nu_\epsilon] \tag{3.32}$$

$$= \frac{E[Q \cdot \{Q > \nu_\epsilon\}]}{P(Q > \nu_\epsilon)} \tag{3.33}$$

$$= \frac{E[Q \cdot \{Q > \nu_\epsilon\}]}{\Gamma(q,\nu_\epsilon)} \tag{3.34}$$

$$= \frac{q \cdot E[\{Q+1 > \nu_\epsilon\}]}{\Gamma(q,\nu_\epsilon)} \quad \text{Chen} - \text{Stein Identity} \tag{3.35}$$

$$= \frac{q \cdot \Gamma(q,\nu_\epsilon - 1)]}{\Gamma(q,\nu_\epsilon)}. \tag{3.36}$$

For the Mean-Variance Risk, we exploit the fact that the Poisson distribution has all of its cumulant moments equal to its mean. Thus,

$$\text{Mean} - \text{Variance}(Q,\gamma,p) = E[Q] + \gamma \cdot \text{Var}[Q] \tag{3.37}$$

$$= q + \gamma \cdot q \tag{3.38}$$

Lastly, for the entropic risk measure, we have that

$$\text{Entropic} = \frac{1}{\gamma} \log\left(\mathbb{E}[e^{-\gamma \cdot Q}]\right) \tag{3.39}$$

$$= \frac{1}{\gamma} \log\left(\sum_{k=0}^{\infty} e^{-\gamma \cdot k} \cdot \frac{q^k}{k!} \cdot e^{-q}\right) \tag{3.40}$$

$$= \frac{1}{\gamma} \log\left(e^{q \cdot (e^{-\gamma} - 1)}\right) \tag{3.41}$$

$$= \frac{q \cdot (e^{-\gamma} - 1)}{\gamma}. \tag{3.42}$$

□

One thing to observe is that the Mean-Deviation of order 2 risk measure, which is similar in spirit to the Mean-Variance risk measure, is similar to the staffing level derived from the square root staffing formula of Jennings et al. (1996). Recall that in the paper of Jennings et al. (1996), they use the infinite server queue as the mean offered load and provide a square root safety factor, which turns out to be the square root of the mean queue length since they exploit the fact that the Poisson distribution has the unique property of having all of its cumulant moments equalling the mean. Thus, this implies that the mean is equivalent to the variance. Furthermore, if q_t^∞ is sufficiently large, then we can use a normal approximation to the Poisson distribution to gain insight in staffing. Using the infinite server approximation and then the normal approximation implies that

$$\mathbb{P}(Q(t) \ge c(t)) \approx \mathbb{P}(Q^\infty(t) \ge k(t)) \tag{3.43}$$

$$\approx = \mathbb{P}(q_t^\infty + \sqrt{q_t^\infty} \cdot X \ge k(t)) \tag{3.44}$$

$$= \mathbb{P}\left(X \ge \frac{k(t) - q_t^\infty}{\sqrt{q_t^\infty}}\right) \tag{3.45}$$

$$= \overline{\Phi}\left(\frac{k(t) - q_t^\infty}{\sqrt{q_t^\infty}}\right) \tag{3.46}$$

where X denotes a Gaussian random variable with mean 0 and variance 1 and $\overline{\Phi}(x)$ is the complement of the standard Gaussian cdf. Using this approximation, one can generate the following square root staffing formula

$$k^\infty(t) = \lceil q_t^\infty + \beta \cdot \sqrt{q_t^\infty} \rceil, \tag{3.47}$$

which should yield stable delay probabilities for our original system. Thus, in order to stabilize the delay probabilities of a multi-server queue at probability ϵ, one must replace γ in the risk measure expression with the inverse Gaussian cdf at ϵ or $\overline{\Phi}^{-1}(\epsilon)$.

On the left of Fig. 1 we plot the V@R for the infinite server queue for different values of ϵ. We see that our explicit formulas are quite accurate for estimating the value at risk for the $M_t/M/\infty$ queueing model. On the right of Fig. 1, we plot the AV@R for the $M_t/M/\infty$ queueing model and for different values of tolerance level ϵ. Like in the case of the V@R, we see that our explicit formulas are also quite good at estimating the simulated AV@R risk measure for the queueing model. On the left of Fig. 2, we plot the Entropic

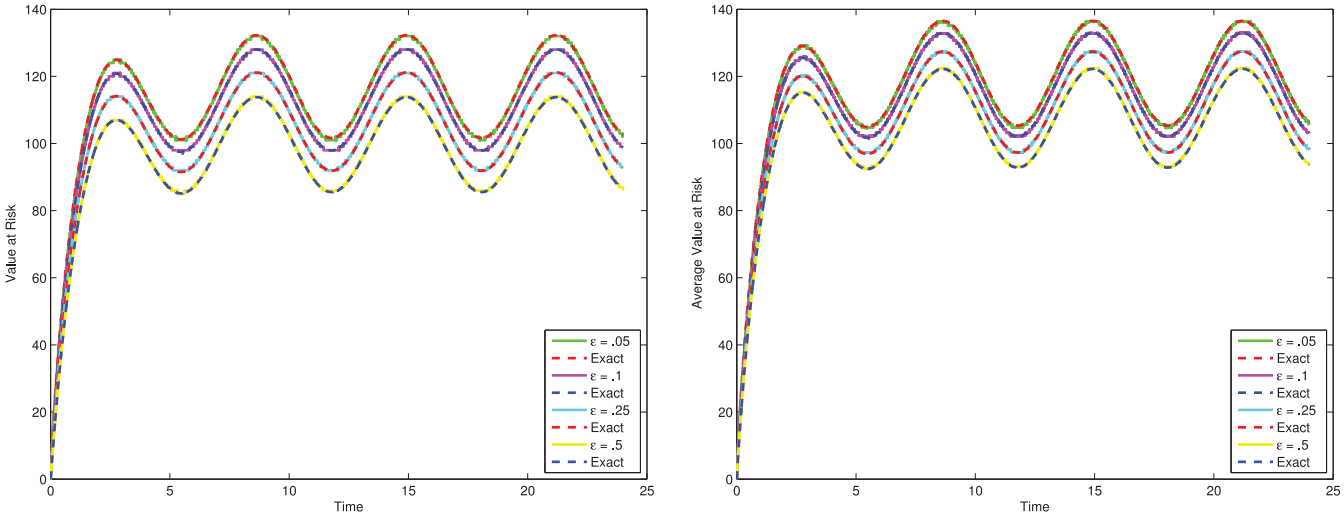

Fig. 1. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $q(0) = 0$. Value at Risk for M/M/∞ Queue. $\epsilon = \{.05, .1, .25, .5\}$ (Left). Average Value at Risk for M/M/∞ Queue. $\epsilon = \{.05, .1, .25, .5\}$ (Right). .

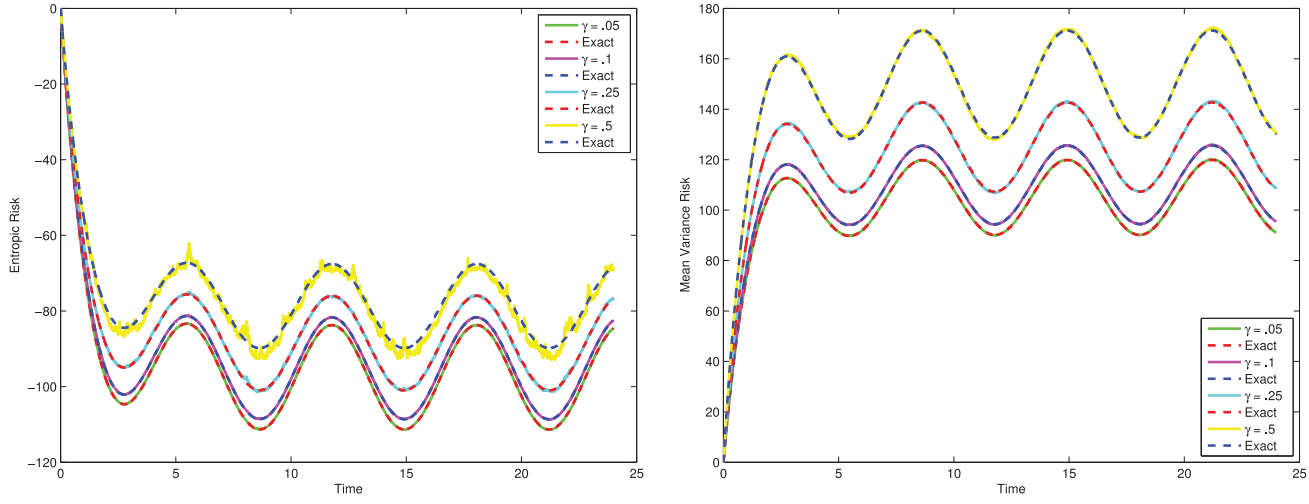

Fig. 2. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $q(0) = 0$. Entropic Risk for M/M/∞ Queue. $\epsilon = \{.05, .1, .25, .5\}$ (left). Mean-Variance Risk for M/M/∞ Queue. $\epsilon = \{.05, .1, .25, .5\}$ (right). .

risk measure for different values of γ. We also see that our explicit formulas are accurate at estimating the appropriate risk for the infinite server queue. We see however, that the case where $\gamma = .05$ or where γ is closest to 0, seems to be the least accurate case. On the right of Fig. 2, we plot the Mean-Variance risk measure for the queueing process and different values of γ. We also see that our explicit formulas accurately capture the dynamic behavior of the risk measure over time.

It is clear that the explicit formulas for the infinite server queue are quite accurate for the queue length process itself. However, in some settings is also interesting to derive risk measures for non-linear functions of the queue length process. In the sequel, we will derive explicit formulas for several functions of the queue length, which are important to the queueing and financial mathematics communities.

3.4. Functional risk measures for the $M_t/G/\infty$ and $M_t/M_t/\infty$ queues

In this section, we now derive some functional risk measures for the $M_t/G/\infty$ and $M_t/M_t/\infty$ queues. Two of the most important functions that appear in both the financial and queueing literature are $f(X) = (X-k)^+$ and $f(X) = (k-X)^+$. In the financial literature these functions respectively represent call and put option payoffs with respect to the random variable X and have strike price equal to k. In the infinite server queue context, k is viewed as a fictitious buffer where we can see how much the buffer is either being exceeded when $f(X) = (X-k)^+$ or how much the buffer is being underutilized when $f(X) = (k-X)^+$.

In the multi-server queueing literature, if k represents the number of servers that are providing service for the system then the two functions $f(X) = (X-k)^+$ and $f(X) = (k-X)^+$ respectively represent the number of customers that are waiting to engage in service with an agent and the number of agents that are currently idle and are not currently serving a customer. Moreover, from a manager's perspective, both of these functions represent inefficiencies in the system. When the function $f(X) = (X-k)^+$ is positive customers are waiting for service and the quality of service is perceived to be low. However, when the function $f(X) = (k-X)^+$ is positive, this means that the manager is staffing the system with too many agents and this is not cost effective from the manager's perspective.

Proposition 3.5. The Functional Value at Risk, Functional Average Value at Risk, Functional Mean-Variance Risk, and Functional Entropic Risk for the nonstationary infinite server queue with mean

q and function $f(Q) = (Q-k)^+$ is given by the following formulas

$$\text{Functional } V@R(f(Q),\epsilon) = \overline{\Gamma}^{-1}(q,\epsilon,k)$$

$$\text{Functional } AV@R(f(Q),\epsilon) = \frac{q\cdot\Gamma(q,\nu_\epsilon+k-1)-k\cdot\Gamma(q,\nu_\epsilon+k)}{\Gamma(q,\nu_\epsilon+k)}$$

$$\text{Functional Mean} - \text{Variance}(f(Q),\gamma,p) = E[(Q-k)^+]+\gamma\cdot\text{Var}[(Q-k)^+]$$

$$\text{Functional Entropic}(f(Q),\gamma) = \frac{1}{\gamma}\cdot\log\left(\overline{\Gamma}(q,k)+e^{\gamma\cdot k-q}\cdot\Gamma(e^{-\gamma}\cdot q,k)\right)$$

where

$$\begin{aligned}E[(Q-k)^+] &= q\cdot\Gamma(q,k-1)-k\cdot\Gamma(q,k)\\ \text{Var}[(Q-k)^+] &= q^2\cdot\Gamma(q,k-2)+q\cdot\Gamma(q,k-1)\\ &\quad -2\cdot k\cdot q\cdot\Gamma(q,k-1)+k^2\cdot\Gamma(q,k)2\\ &\quad \cdot q\cdot k\cdot\Gamma(q,k-1)\cdot\Gamma(q,k)\\ &\quad -k^2\cdot\Gamma^2(q,k)-q^2\cdot\Gamma^2(q,k-1).\end{aligned}$$

Moreover, for the function $f(Q) = (k-Q)^+$ we have that

$$\text{Functional } V@R(f(Q),\epsilon) = \Gamma^{-1}(q,\epsilon,k)$$

$$\text{Functional } AV@R(f(Q),\epsilon) = \frac{k\cdot\overline{\Gamma}(q,\nu_\epsilon+k)-q\cdot\overline{\Gamma}(q,\nu_\epsilon+k+1)}{\overline{\Gamma}(q,\nu_\epsilon+k)}$$

$$\text{Functional MV}(f(Q),\gamma,p) = E[(k-Q)^+]+\gamma\cdot\text{Var}[(k-Q)^+]$$

$$\text{Functional Entropic}(f(Q),\gamma) = \frac{1}{\gamma}\cdot\log\left(\Gamma(q,k)+e^{\gamma\cdot k-q}\cdot\overline{\Gamma}(e^{-\gamma}\cdot q,k)\right)$$

where

$$\begin{aligned}E[(k-Q)^+] &= k\cdot\overline{\Gamma}(q,k)-q\cdot\overline{\Gamma}(q,k-1)\\ \text{Var}[(k-Q)^+] &= k^2\cdot\overline{\Gamma}(q,k)-2\cdot k\cdot q\cdot\overline{\Gamma}(q,k-1)\\ &\quad +q\cdot\overline{\Gamma}(q,k-1)+q^2\cdot\overline{\Gamma}(q,k-2)\\ &\quad -k^2\cdot\overline{\Gamma}^2(q,k)+2\cdot\overline{\Gamma}(q,k)\cdot\overline{\Gamma}(q,k-1)\\ &\quad -q^2\cdot\overline{\Gamma}^2(q,k-1).\end{aligned}$$

Proof. See Appendix. □

4. Erlang-A queueing model

Now that we have addressed the infinite server queue, we would like to extend our risk measures to a more general queueing model that takes into account that most queues have a finite number of servers and that most customers are not infinitely patient when waiting for service from an agent. The Erlang-A model is the canonical choice when considering these new features of the queueing model. Since the Erlang-A has these new features, it is not as tractable as the infinite server queue. Thus, we will exploit new approximations for this queueing system that are accurate and derive risk measures for these approximations. However, now we give a brief overview of the Erlang-A model for the reader's convenience.

4.1. Stochastic analysis of Erlang-A model

In this section we introduce the $M_t/M/k_t+M$ queueing model that serves to describe the dynamics of our hospital dynamics. Mandelbaum et al. (1998), showed that the queueing system process $\{Q(t)|t\geq 0\}$ is represented by the following equation

$$\begin{aligned}Q(t) &= Q(0)+\Pi_1\left(\int_0^t\lambda(s)ds\right)-\Pi_2\left(\int_0^t\mu\cdot(Q(s)\wedge c(s))ds\right)\\ &\quad -\Pi_3\left(\int_0^t\beta\cdot(Q(s)-c(s))^+ds\right),\end{aligned}$$

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-homogenous Poisson arrival process with rate $\lambda(t)$. Subjecting Π_2 to random time change causes it to count the number of service departures from c servers and exponentially distributed service times function of mean $1/\mu$. Finally the random time changes of Π_3 cause it to count the number of abandonments from c servers and exponentially distributed abandonment times of mean $1/\beta$. With this representation of our queueing dynamics, this model is an example of a Markovian service network, which were extensively studied in Mandelbaum et al. (1998).

4.2. Fluid and diffusion limits

In Mandelbaum et al. (1998) it was shown that a Markovian service network always has a fluid and diffusion limits i.e.

$$\frac{1}{\eta}Q^\eta = q \;\; a.s \quad \text{and} \quad \sqrt{\eta}\cdot\left(\frac{1}{\eta}Q^\eta-q\right)\Rightarrow \hat{Q} \tag{4.49}$$

where the fluid mean is governed by the one dimensional dynamical system

$$\dot{q} = \lambda-\mu\cdot(q\wedge c)-\beta\cdot(q-c)^+ \tag{4.50}$$

and if the set $\{t|q(t)=c\}$ has measure zero then $\hat{Q}$ is a Gaussian diffusion whose variance combines with the fluid mean to form a 2-dimensional dynamical system given by (2.2) and

$$\begin{aligned}\dot{\text{Var}}[\hat{Q}] &= \lambda+\mu\cdot(q\wedge c)+\beta\cdot(q-c)^+\\ &\quad -2\cdot\text{Var}[\hat{Q}]\cdot(\mu\cdot\{q<c\}+\beta\cdot\{q\geq c\}).\end{aligned} \tag{4.51}$$

where $\{q<c\}$ denotes an indicator function equaling one if, $q<c$, and zero otherwise. However, for small systems like in hospitals, it was shown in Ko and Gautam (2013), Massey and Pender (2011, 2013), Pender (2014a,b) that these fluid and diffusion limits do not do a great job of characterizing the correct moment behavior. Therefore, we will use approximations for our queueing process that will serve to estimate the transient dynamics.

4.3. Gaussian variance approximation

The first approximation, which we call the Gaussian variance approximation was first developed by Ko and Gautam (2013) and further simplified and explained in Massey and Pender (2011). This approximation assumes a Gaussian distribution for the queueing model i.e

$$Q(t) \overset{d}{=} q(t)+X\cdot\sqrt{v(t)}. \tag{4.52}$$

for all $t\geq 0$, where $\{q(t),v(t)|t\geq 0\}$ is some two-dimensional dynamical system where the v process is always positive and X is a standard Gaussian random variable. We call this the *Gaussian variance approximation* (GVA). The forward equations for the mean and variance of Q are

$$\dot{E}[Q] = \lambda-\left(\mu\cdot E[Q\wedge c]+\beta\cdot E[(Q-c)^+]\right) \tag{4.53}$$

$$\begin{aligned}\dot{\text{Var}}[Q] &= \lambda+\mu\cdot E[Q\wedge c]+\beta\cdot E[(Q-c)^+]\\ &\quad -2\left(\mu\cdot\text{Cov}[Q,Q\wedge c]+\beta\cdot\text{Cov}[Q,(Q-c)^+]\right)\end{aligned} \tag{4.54}$$

Now if we define the following variable $\chi = \frac{c-q}{\sqrt{v}}$, we get the following differential equations for the mean and variance of the queueing process under the distributional assumptions of GVA

$$\dot{E}[Q] = \lambda - \left(\mu \cdot \sqrt{v} \cdot E[X \wedge \chi] + \beta \cdot \sqrt{v} \cdot E[(X-\chi)^+]\right)$$

$$\begin{aligned}\dot{\text{Var}}[Q] &= \lambda + \mu \cdot \sqrt{v} \cdot E[X \wedge \chi] + \beta \cdot \sqrt{v} \cdot E[(X-\chi)^+] \\ &\quad -2\left(\mu \cdot v \cdot \text{Cov}[X, X \wedge \chi] + \beta \cdot v \cdot \text{Cov}[X, (X-\chi)^+]\right)\end{aligned}$$

Thus, in order to understand the dynamics, it only remains to compute the expectations and covariance. Like Massey and Pender (2011), we resort to using Stein's lemma to derive the expectations and covariance terms. To do this we use the following Hermite polynomial generalization of Stein's lemma.

Lemma 4.1. *If X is a standard Gaussian random variable and $E[f^{(n)}(X)] < \infty$, then*

$$E[f(X) \cdot h_n(X)] = E[f^{(n)}(X)]$$

where f is any generalized function and $f^{(n)}$ is the n^{th} derivative of the function f.

The use of Stein's lemma yields the following equations for the mean and variance dynamics of our queueing process:

$$\begin{aligned}\dot{E}[Q] &= \lambda - \mu \cdot q + \sqrt{v} \cdot (\mu - \beta) \cdot \left(\varphi(\chi) - \chi \cdot \overline{\Phi}(\chi)\right) \\ \dot{\text{Var}}[Q] &= \lambda + \mu \cdot q - \sqrt{v} \cdot (\mu - \beta) \cdot \left(\varphi(\chi) - \chi \cdot \overline{\Phi}(\chi)\right) \\ &\quad - 2 \cdot v \cdot \left(\mu \cdot \Phi(\chi) + \beta \cdot \overline{\Phi}(\chi)\right)\end{aligned}$$

where we define φ and Φ to be the density and the cumulative distribution functions, respectively, for $X \sim \text{Normal}(0, 1)$, i.e.,

$$\varphi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) \equiv \int_{-\infty}^{x} \varphi(y)\, dy, \quad \text{and let}$$
$$\overline{\Phi}(x) \equiv \int_{x}^{\infty} \varphi(y)\ dy. \tag{4.55}$$

In the sequel, we will show how to use the Gaussian variance approximation in order to approximate various risk measures of the Erlang-A queueing model.

4.4. Risk measures for Erlang-A queue

In this section, we show that the GVA can be used to approximate various risk measures of the Erlang-A queue. However, we should mention that when $c = \infty$ or $\mu = \beta$, the Erlang-A queueing model reduces to the infinite server queue and we have already given exact results for this case in the previous sections of the paper.

Proposition 4.2. *Under the assumptions of the GVA for the queue length, then the Value at Risk, Average Value at Risk, Mean-Variance Risk, and Entropic Risk have the following expressions*

$$\begin{aligned}V@r(Q,\epsilon) &= q + \sqrt{v} \cdot \Phi^{-1}(\epsilon) \\ AV@r(Q,\epsilon) &= q + \sqrt{v} \cdot \frac{\varphi\left(\frac{\nu_\epsilon - q}{\sqrt{v}}\right)}{\overline{\Phi}\left(\frac{\nu_\epsilon - q}{\sqrt{v}}\right)} \\ \text{Mean} - \text{Variance}(Q,\gamma) &= q + \gamma \cdot v \\ \text{Entropic}(Q,\gamma) &= -q + \frac{\gamma \cdot v}{2}.\end{aligned}$$

Proof. From the definition of the Value at risk, we have that

$$V@r(Q,\epsilon) = \inf\{y : P\{Q \le y\} \ge \epsilon\}. \tag{4.56}$$

Thus, we we assume that the queue length is approximated by GVA, we get that

$$\begin{aligned}V@r(Q,\epsilon) &= \inf\{y : P\{Q \le y\} \ge \epsilon\} \\ &= \inf\left\{y : P\left\{q + \sqrt{v} \cdot X \le y\right\} \ge \epsilon\right\} \\ &= \inf\left\{y : P\left\{X \le \frac{y-q}{\sqrt{v}}\right\} \ge \epsilon\right\} \\ &= \inf\left\{y : \Phi\left(\frac{y-q}{\sqrt{v}}\right) \ge \epsilon\right\} \\ &= q + \sqrt{v} \cdot \Phi^{-1}(\epsilon).\end{aligned}$$

Now we show the exact formula for the AV@R. From the definition of the AV@R, we have that

$$AV@r(Q,\epsilon) = E[Q|Q > \nu_\epsilon] \tag{4.57}$$

where we define $\nu_\epsilon = V@r(Q,\epsilon)$. Thus, when we approximate the queue length distribution with the GVA and define $\chi = \frac{\nu_\epsilon - q}{\sqrt{v}}$, we have that

$$\begin{aligned}AV@r(Q,\epsilon) &= E[Q|Q > \nu_\epsilon] \\ &= E[q + \sqrt{v} \cdot X | q + \sqrt{v} \cdot X > \nu_\epsilon] \\ &= \frac{E[(q + \sqrt{v} \cdot X) \cdot \{q + \sqrt{v} \cdot X > \nu_\epsilon\}]}{E[\{q + \sqrt{v} \cdot X > \nu_\epsilon\}]} \\ &= \frac{q \cdot P\{q + \sqrt{v} \cdot X > \nu_\epsilon\} + \sqrt{v} \cdot E[X \cdot \{q + \sqrt{v} \cdot X > \nu_\epsilon\}]}{P\{q + \sqrt{v} \cdot X > \nu_\epsilon\}} \\ &= \frac{q \cdot P\{X > \chi\} + \sqrt{v} \cdot E[X \cdot \{X > \chi\}]}{P\{X > \chi\}} \\ &= \frac{q \cdot \overline{\Phi}(\chi) + \sqrt{v} \cdot \varphi(\chi)}{\overline{\Phi}(\chi)} \\ &= q + \sqrt{v} \cdot \frac{\varphi(\chi)}{\overline{\Phi}(\chi)}.\end{aligned}$$

Remark 4.3. Note that $q + \sqrt{v} \cdot \frac{\varphi(\chi)}{\overline{\Phi}(\chi)} > q + \chi \cdot \sqrt{v}$, which implies that the AV@R is larger than the V@R for the same value of ϵ.

For the Mean-Variance risk measure, which is defined as

$$\text{Mean} - \text{Variance}(Q,\gamma) = E[Q] + \gamma \cdot \text{Var}[Q], \tag{4.58}$$

we use the standard properties of the Gaussian distribution to conclude that

$$\text{Mean} - \text{Variance}(Q,\gamma) = q + \gamma \cdot v. \tag{4.59}$$

Lastly, we derive the exact expressions for the Entropic risk measure. For the Entropic risk measure we have that

$$\rho(Q,\gamma) = \frac{1}{\gamma} \cdot \log\left(E\left[e^{-\gamma \cdot Q}\right]\right). \tag{4.60}$$

Thus, when we assume that the queue length is approximated by GVA, we get that

$$\begin{aligned}\frac{1}{\gamma} \cdot \log\left(E\left[e^{-\gamma \cdot Q}\right]\right) &= \log\left(E\left[e^{-\gamma \cdot (q + \sqrt{v} \cdot X)}\right]\right) \\ &= \frac{1}{\gamma} \cdot \log\left(E\left[e^{-\gamma \cdot q}\right] \cdot E\left[e^{-\gamma \cdot \sqrt{v} \cdot X}\right]\right) \\ &= \frac{1}{\gamma} \cdot \log\left(e^{-\gamma \cdot q}\right) + \frac{1}{\gamma} \cdot \log\left(E\left[e^{-\gamma \cdot \sqrt{v} \cdot X}\right]\right) \\ &= -q + \frac{\gamma \cdot v}{2}\end{aligned}$$

□

On the left of Fig. 3 we plot the V@R for the Erlang-A queue for different values of ϵ. We see that our approximate formulas are quite accurate for estimating the V@R for the Erlang-A queueing model. On the right of Fig. 3, we plot the AV@R for the Erlang-A queueing model and for different values of tolerance level ϵ. Like in the case of the V@R, we see that our explicit formulas are also quite good at estimating the simulated AV@R risk measure for the queueing model. On the left of Fig. 4, we plot the Entropic risk

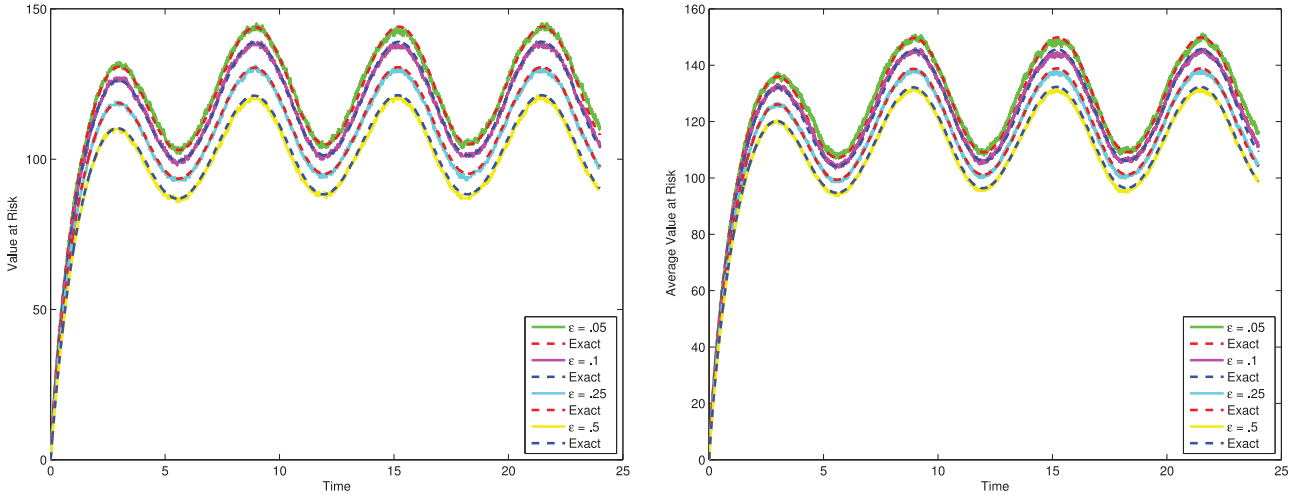

Fig. 3. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = .5$, $c = 100$, $q(0) = 0$. Value at Risk for Erlang-A Queue. $\epsilon = \{.05, .1, .25, .5\}$ (Left). Average Value at Risk for Erlang-A Queue. $\epsilon = \{.05, .1, .25, .5\}$ (Right).

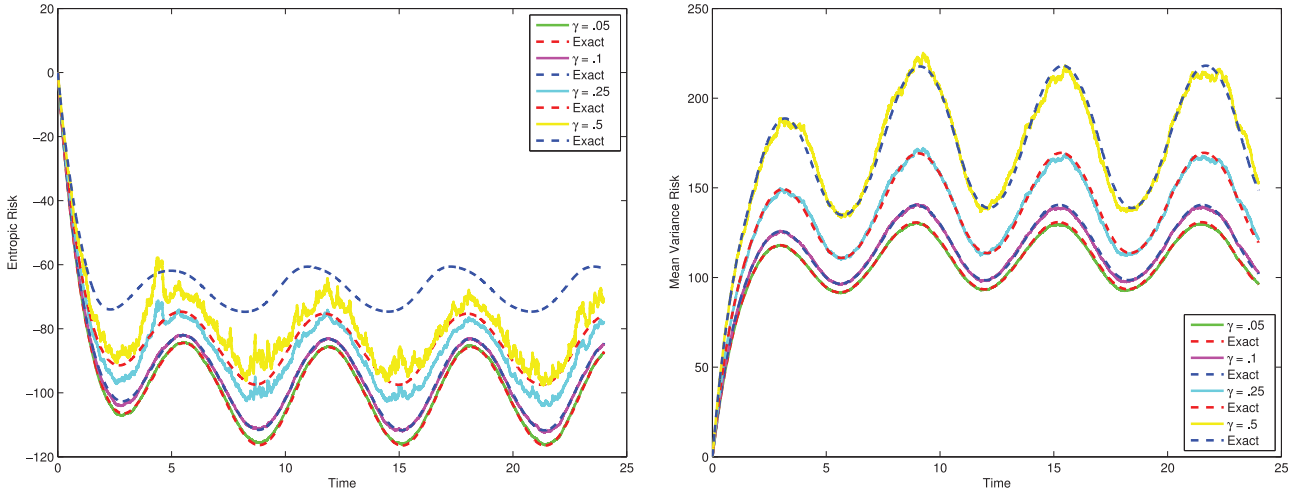

Fig. 4. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = .5$, $c = 100$, $q(0) = 0$. Entropic Risk for Erlang-A Queue. $\epsilon = \{.05, .1, .25, .5\}$ (left). Mean-Variance Risk for Erlang-A Queue. $\epsilon = \{.05, .1, .25, .5\}$ (right).

measure for different values of γ. We also see that our Gaussian approximations are accurate at estimating the appropriate risk for the Erlang-A queue. We see however, that the case where $\gamma = .05$ or where γ is closest to 0, seems to be the least accurate case. On the right of Fig. 4, we plot the Mean-Variance risk measure for the queueing process and different values of γ. We also see that our approximate formulas accurately capture the dynamic behavior of the risk measure over time.

It is clear that the explicit formulas for the Erlang-A queue are quite accurate for the queue length process itself. However, in some settings is also interesting to derive risk measures for non-linear functions of the queue length process. In the sequel, we will derive explicit formulas for several functions of the queue length, which are important to the queueing literature.

4.5. Functional risk measures

Proposition 4.4. *Under the GVA and when* $f(Q) = (Q - k)^+$, *the Functional Value at Risk, Functional Average Value at Risk, Functional Mean-Variance Risk, and Functional Entropic Risk have the following expressions*

$$\text{Functional } V@r((Q-k)^+, \epsilon) = q - k + \sqrt{v} \cdot \Phi^{-1}(\epsilon)$$

$$\text{Functional } AV@r((Q-k)^+, \epsilon) = \frac{\sqrt{v} \cdot \varphi(\chi_{k,c})}{\overline{\Phi}(\chi_{k,c})} - \sqrt{v} \cdot \chi_{k,c}$$

$$\begin{aligned}\text{Functional Mean} - \text{Var}((Q-k)^+, \epsilon) \\ = E[(Q-k)^+] + \gamma \cdot \text{Var}[(Q-k)^+]\end{aligned}$$

$$\begin{aligned}\text{Functional Entropic}((Q-k)^+, \epsilon) \\ = \frac{1}{\gamma} \log\left(\Phi(\chi_k) + e^{\gamma^2 \cdot v/2 + \gamma \cdot \sqrt{v} \cdot \chi_k} \cdot \overline{\Phi}(\chi_k + \gamma \cdot \sqrt{v})\right)\end{aligned}$$

where

$$E[(Q-k)^+] = \sqrt{v} \cdot \left(\varphi(\chi_k) - \chi_k \cdot \overline{\Phi}(\chi_k)\right)$$

$$\begin{aligned}\text{Var}[(Q-k)^+] = v \cdot \left(-\chi_k \cdot \varphi(\chi_k) + (\chi_k^2 + 1) \cdot \overline{\Phi}(\chi_k) - \varphi(\chi_k)^2\right) \\ + v \cdot \left(-\chi_k^2 \cdot \overline{\Phi}(\chi_k)^2 + 2 \cdot \chi_k \cdot \varphi(\chi_k) \cdot \overline{\Phi}(\chi_k)\right).\end{aligned}$$

Moreover, when $f(Q) = (k - Q)^+$, *the Functional Value at Risk, Functional Average Value at Risk, Functional Mean-Variance Risk, and Functional Entropic Risk have the following expressions*

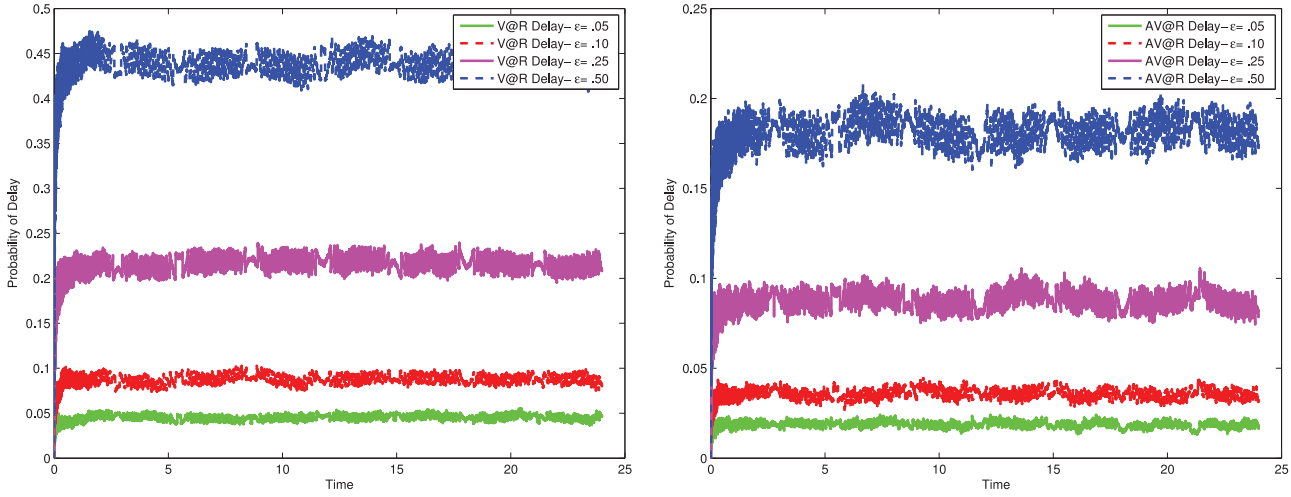

Fig. 5. Staffing the $M_t/M/c+M$ Queue. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = .5$, $c = 100$, $q(0) = 0$. Probability of Delay using V@R as staffing procedure. $\epsilon = \{.05, .1, .25, .5\}$ (left). Probability of Delay using AV@R as staffing procedure. $\epsilon = \{.05, .1, .25, .5\}$ (right).

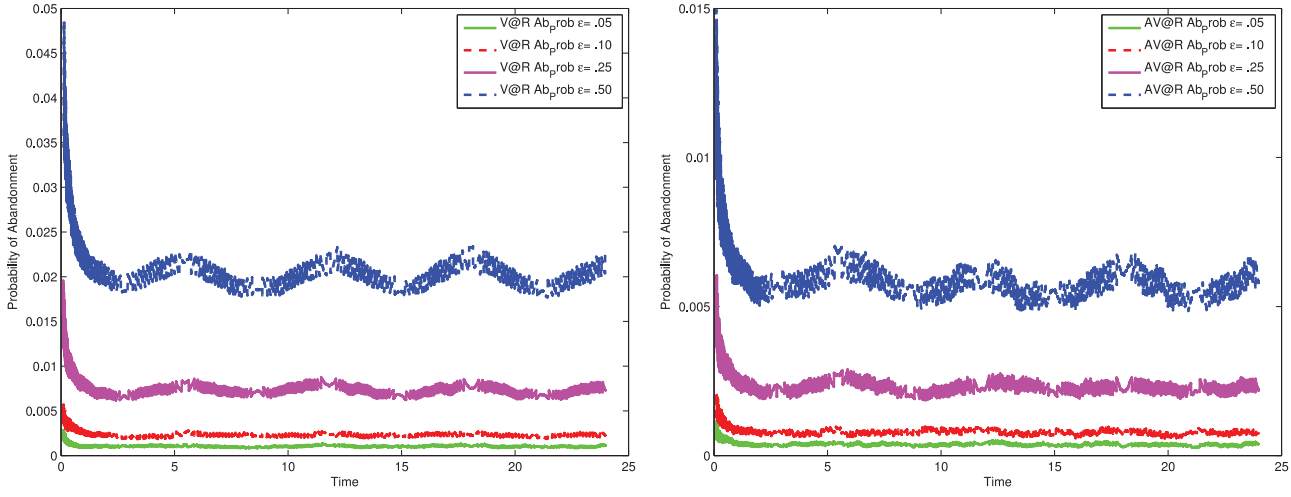

Fig. 6. Staffing the $M_t/M/c+M$ Queue. $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = .5$, $c = 100$, $q(0) = 0$. Abandonment probability using V@R as staffing procedure. $\epsilon = \{.05, .1, .25, .5\}$ (left). Abandonment probability using AV@R as staffing procedure. $\epsilon = \{.05, .1, .25, .5\}$ (right).

$$\text{Functional } V@r((k-Q)^+, \epsilon) = k - q + \sqrt{v} \cdot \overline{\Phi}^{-1}(\epsilon)$$

$$\text{Functional } AV@r((k-Q)^+, \epsilon) = k - q + \sqrt{v} \cdot \frac{\varphi(\chi_{k,c})}{\overline{\Phi}(\chi_{k,c})}$$

$$\text{Functional Mean} - \text{Var}((k-Q)^+, \epsilon) = E[(k-Q)^+] + \gamma \cdot \text{Var}[(k-Q)^+]$$

$$\text{Functional Entropic}((k-Q)^+, \epsilon) = \frac{1}{\gamma} \log\left(\overline{\Phi}(\chi_k) + e^{\gamma^2 \cdot v/2 - \gamma \cdot \sqrt{v} \cdot \chi_k} \cdot \Phi(\chi_k - \gamma \cdot \sqrt{v})\right).$$

where

$$E[(k-Q)^+] = (k-q) \cdot \Phi(\chi) + \sqrt{v} \cdot \varphi(\chi)$$

$$\text{Var}[(k-Q)^+] = v \cdot (\chi^2 + 1) \cdot \Phi(\chi) + \chi \cdot v \cdot \varphi(\chi) - ((k-q) \cdot \Phi(\chi) + \sqrt{v} \cdot \varphi(\chi))^2.$$

Proof. See Appendix. □

4.6. Staffing nonstationary queues with risk measures

In addition to the fact that the risk measures provide performance measures for queueing models, the risk measures derived for these nonstationary queues can also be used as staffing procedures. On the left of Fig. 5, we use the V@R as a staffing procedure for the Erlang-A queueing model. We see that the V@R somewhat stabilizes the probability of delay near the value of ϵ that is used for the V@R calculation. This is partly because the staffing level given by the V@R is similar to the offered load approach of Jennings et al. (1996). Moreover, on the right of Fig. 5, we use the AV@R as a staffing procedure for the Erlang-A queueing model. We also see this time that staffing with the AV@R also somewhat stabilizes the probability of delay. It should be noted that the probability of delay using AV@R is lower than the probability of delay when using V@R. This is because the staffing level derived from AV@R is larger since AV@R is the average of the tail values of the V@R. However, we should also mention that the variation that one sees for the probability of delay is due to the addition or removal of one server. This is partially observed when ϵ is the smallest since the variation is the smallest and the largest number of servers is used for staffing. Thus, the impact of adding or removing one server is much smaller.

On the left of Fig. 6, we use the V@R as a staffing procedure for the Erlang-A queueing model. We see that the V@R does not stabilizes the abandonment probability. This is consistent with Liu and Whitt (2012) where they show that it is impossible to stabilize both using the same staffing procedure, except when the

probabilities are very small. Moreover, on the right of Fig. 6, we use the AV@R as a staffing procedure for the Erlang-A queueing model. We also see this time that staffing with the AV@R does not stabilizes the abandonment probability. Similar to the previous figure, we also observe that noted that the resuting abandonment probability using AV@R as the staffing procedure is lower than the abandonment probability when using V@R.

4.7. Beyond the Erlang-A queueing model

This method of using the mean and variance of the queueing model can be used beyond the Erlang-A model. There are fluid and diffusion limits for queues with more general arrival, service times, and abandonment times. See for example , Liu and Whitt (2011); 2012); Liu, Whitt et al. (2014) and Jelenković, Mandelbaum, and Momčilović (2004); Mandelbaum and Momcilovic (2012); Zeltyn and Mandelbaum (2005) . The mean and variance of these fluid and diffusion models can be used as Gaussian approximations of the underyling stochastic process in the risk measure formulation. Using the same formulas, one can also derive approximations to the risk measures for more general queueing models. Although we do not consider these queueing models in this paper, it is perhaps interesting to see if the Gaussian fluid and diffusion approximations help in providing good approximations for the risk measures of the original queueing model.

5. Conclusions and future work

We have analyzed the problem of staffing queueing systems with risk measures. We have shown that many of the traditional staffing procedures like square root staffing and the modified offered load procedure can be derived from some of the standard risk measures. This paper introduces the concept of risk measures to the queueing theory community and shows how they can be relevant, especially in the context of healthcare systems. One extension worth pursing, is to extend our analysis to a multi-dimensional setting. To pursue this requires extensions of the approximations of the queueing systems to multi-dimensional settings and also extending the notion of risk measures to a multi-dimensional setting. Some progress has been made on the risk measure side see for example (Klyman, 2011). It would be of particular interest to apply it to variants of the Erlang-R model of Yom-Tov and Mandelbaum (2014). Moreover, it is also possible to analyze many other risk measures that are not presented in this work, see for example the list of risk measures in Cheridito and Li (2008); 2009). Lastly, the concept of a conditional risk measure has been developed recently in the work of Detlefsen and Scandolo (2005) and these types of risk measures are worth study given new applications such as queueing theory.

Another area of interest is to use the fluid and diffusion limits of Mandelbaum and Momcilovic (2012) to approximate the risk measure performance of more general queueing systems. This would allow managers to approximate risk measures for more general systems that model reality well. Another extension would be to construct risk measure approximations for the single server queue as well. In the stationary case, this analysis would involve the geometric distribution. Like the Chen–Stein theorem and Stein's lemma, the geometric distribution also has a characterizing operator and it is of the form

$$\mathcal{A}f(k) = (1-p)\cdot(f(k+1)-f(k)) - p\cdot f(k) + p\cdot f(0). \quad (5.61)$$

The same analysis used earlier can also be used to calculate risk measures for the single server queueing model and yield performance measures for managers with different risk profiles.

Acknowledgments

The author would like to thank Cornell University (ORIE) for its gracious support and Dr. Jared Klyman for his insight on the topic of risk measures. Lastly, the author would like to thank the two referees for their insightful comments and feedback.

Appendix A

A.1. Proofs of results

Proof of Proposition 3.5. For the Functional Value at Risk, we have that

$$V@R(Q,\epsilon) = \inf\{y \mid \mathbb{P}\{(Q-k)^+ \le y\} \ge \epsilon\} \quad (6.62)$$

$$= \inf\{y \mid \mathbb{P}\{Q \le y+k\} \ge \epsilon\} \quad (6.63)$$

$$= \inf\{y \mid \overline{\Gamma}(q, y+k) \ge \epsilon\} \quad (6.64)$$

$$= \overline{\Gamma}^{-1}(q,\epsilon,k). \quad (6.65)$$

For the Average Value at Risk, we first let $\nu_\epsilon = V@R((Q-k)^+,\epsilon)$ and thus we obtain the following

$$AV@R(Q,\epsilon) = E[(Q-k)^+ \mid (Q-k)^+ > \nu_\epsilon] \quad (6.66)$$

$$= \frac{E[(Q-k)^+\cdot\{(Q-k)^+ > \nu_\epsilon\}]}{P((Q-k)^+ > \nu_\epsilon)} \quad (6.67)$$

$$= \frac{E[(Q-k)^+\cdot\{Q > \nu_\epsilon + k\}]}{P(Q > \nu_\epsilon + k)} \quad (6.68)$$

$$= \frac{E[(Q-k)\cdot\{Q>k\}\cdot\{Q > \nu_\epsilon + k\}]}{P(Q > \nu_\epsilon + k)} \quad (6.69)$$

$$= \frac{E[(Q-k)\cdot\{Q > \nu_\epsilon + k\}]}{\Gamma(q, \nu_\epsilon + k)} \quad (6.70)$$

$$= \frac{E[Q\cdot\{Q > \nu_\epsilon + k\}] - k\cdot E[\{Q > \nu_\epsilon + k\}]}{\Gamma(q, \nu_\epsilon + k)} \quad (6.71)$$

$$= \frac{q\cdot\Gamma(q, \nu_\epsilon + k - 1) - k\cdot\Gamma(q, \nu_\epsilon + k)}{\Gamma(q, \nu_\epsilon + k)}. \quad (6.72)$$

For the Mean-Variance Risk, we just need to compute the mean and variance of the function $(Q-k)^+$.

$$\text{Mean} - \text{Variance}((Q-k)^+,\gamma,p) = E[(Q-k)^+] + \gamma\cdot\text{Var}[(Q-k)^+] \quad (6.73)$$

$$= q\cdot\Gamma(q,k-1) - k\cdot\Gamma(q,k) + \gamma\cdot q \quad (6.74)$$

$$\begin{aligned} E[(Q-k)^+] &= E[(Q-k)\cdot\{Q>k\}] \\ &= E[Q\cdot\{Q>k\}] - k\cdot E[\{Q>k\}] \\ &= q\cdot\Gamma(q,k-1) - k\cdot\Gamma(q,k) \end{aligned}$$

$$\begin{aligned} \text{Var}[(Q-k)^+] &= E[((Q-k)^+)^2] - E[(Q-k)^+]^2 \\ &= E[(Q^2 - 2\cdot Q\cdot k + k^2)\cdot\{Q>k\}] - E[(Q-k)^+]^2 \\ &= q^2\cdot\Gamma(q,k-2) + q\cdot\Gamma(q,k-1) \\ &\quad - 2\cdot k\cdot q\cdot\Gamma(q,k-1) + k^2\cdot\Gamma(q,k) \\ &\quad - (q\cdot\Gamma(q,k-1) - k\cdot\Gamma(q,k)) \\ &= q^2\cdot\Gamma(q,k-2) + q\cdot\Gamma(q,k-1) \end{aligned}$$

$$
\begin{aligned}
&- 2 \cdot k \cdot q \cdot \Gamma(q, k-1) + k^2 \cdot \Gamma(q, k) \\
&- q^2 \cdot \Gamma^2(q, k-1) - k^2 \cdot \Gamma^2(q, k) \\
&+ 2 \cdot q \cdot k \cdot \Gamma(q, k-1) \cdot \Gamma(q, k)
\end{aligned}
$$

Lastly, for the functional entropic risk measure, we have that

$$\text{Entropic} = \frac{1}{\gamma} \log\left(\mathbb{E}[e^{-\gamma \cdot (Q-k)^+}]\right) \tag{6.75}$$

$$= \frac{1}{\gamma} \log\left(\sum_{m=0}^{k-1} e^{-\gamma \cdot (m-k)^+} \cdot \frac{q^m}{m!} \cdot e^{-q} + \sum_{m=k}^{\infty} e^{-\gamma \cdot (m-k)^+} \cdot \frac{q^m}{m!} \cdot e^{-q}\right) \tag{6.76}$$

$$= \frac{1}{\gamma} \log\left(\sum_{m=0}^{k-1} \frac{q^m}{m!} \cdot e^{-q} + \sum_{m=k}^{\infty} e^{-\gamma \cdot (m-k)^+} \cdot \frac{q^m}{m!} \cdot e^{-q}\right) \tag{6.77}$$

$$= \frac{1}{\gamma} \log\left(\overline{\Gamma}(q, k) + \sum_{m=k}^{\infty} e^{-\gamma \cdot (m-k)^+} \cdot \frac{q^m}{m!} \cdot e^{-q}\right) \tag{6.78}$$

$$= \frac{1}{\gamma} \log\left(\overline{\Gamma}(q, k) + e^{\gamma \cdot k} \cdot \sum_{m=k}^{\infty} e^{-\gamma \cdot m} \cdot \frac{q^m}{m!} \cdot e^{-q}\right) \tag{6.79}$$

$$= \frac{1}{\gamma} \log\left(\overline{\Gamma}(q, k) + e^{\gamma \cdot k} \cdot \Gamma(e^{-\gamma} \cdot q, k)\right) \tag{6.80}$$

For the terms involving the function $f(Q) = (k - Q)^+$, we can use the same type of analysis as above. We do not derive these separately for brevity. □

Proof of Proposition 4.5.. We will provide proof of the all of the terms that involve the function $(Q - k)^+$ and not the terms involving $(k - Q)^+$ since the terms involving $(k - Q)^+$ can be derived in a similar manner. We define the functional Value at Risk as:

$$FV@r(f(Q), \epsilon) = \inf\{y : P\{f(Q) \leq y\} \geq \epsilon\} \tag{6.81}$$

where $f(Q) = (Q - k)^+$.

Using GVA as our approximation for the queueing dynamics and $f(Q) = (Q - k)^+$, we have that

$$
\begin{aligned}
FV@r(f(Q), \epsilon) &= \inf\left\{y : P\left\{(Q-k)^+ \leq y\right\} \geq \epsilon\right\} \\
&= \inf\left\{y : P\left\{\sqrt{v} \cdot (X - \chi)^+ \leq y\right\} \geq \epsilon\right\} \\
&= \inf\left\{y : P\left\{(X - \chi)^+ \leq \frac{y}{\sqrt{v}}\right\} \geq \epsilon\right\} \\
&= \inf\left\{y : P\left\{X - \chi \leq \frac{y}{\sqrt{v}}\right\} \geq \epsilon\right\} \\
&= \inf\left\{y : \Phi\left(\chi + \frac{y}{\sqrt{v}}\right) \geq \epsilon\right\} \\
&= \inf\left\{y : \Phi\left(\frac{k-q}{\sqrt{v}} + \frac{y}{\sqrt{v}}\right) \geq \epsilon\right\} \\
&= \inf\left\{y : \Phi\left(\frac{k-q+y}{\sqrt{v}}\right) \geq \epsilon\right\}.
\end{aligned}
$$

Now by inversion of the cdf, we finally get that

$$FV@r((Q-k)^+, \epsilon) = k - q + \sqrt{v} \cdot \Phi^{-1}(\epsilon).$$

which completes the proof for the functional value at risk.

Now for the AV@R we have that

$$
\begin{aligned}
E[(Q-k)^+ | (Q-k)^+ > \nu_\epsilon] &= E[(Q-k)^+ | Q > \nu_\epsilon + k] \\
&= \frac{E[(Q-k)^+ \cdot \{Q > \nu_\epsilon + k\}]}{P(Q > \nu_\epsilon + k)} \\
&= \frac{\sqrt{v} \cdot E\left[\left(X - \frac{k-q}{\sqrt{v}}\right)^+ \cdot \{\sqrt{v} \cdot X > \nu_\epsilon + k - q\}\right]}{\overline{\Phi}\left(\frac{\nu_\epsilon + k - q}{\sqrt{v}}\right)} \\
&= \frac{\sqrt{v} \cdot E\left[(X - \chi_k)^+ \cdot \{X > \chi_{k,\nu_\epsilon}\}\right]}{\overline{\Phi}\left(\chi_{k,\nu_\epsilon}\right)} \\
&= \frac{\sqrt{v} \cdot E\left[(X - \chi_k) \cdot \{X > \chi_k\} \cdot \{X > \chi_{k,\nu_\epsilon}\}\right]}{\overline{\Phi}\left(\chi_{k,\nu_\epsilon}\right)} \\
&= \frac{\sqrt{v} \cdot \left(\varphi(\chi_{k,\nu_\epsilon}) - \chi_{k,\nu_\epsilon} \cdot \overline{\Phi}\left(\chi_{k,\nu_\epsilon}\right)\right)}{\overline{\Phi}\left(\chi_{k,\nu_\epsilon}\right)} \\
&= \nu_\epsilon - k + \sqrt{v} \cdot \frac{\varphi(\chi_{k,\nu_\epsilon})}{\overline{\Phi}\left(\chi_{k,\nu_\epsilon}\right)}
\end{aligned}
$$

For the functional Mean-Variance risk measure it suffices to look at the paper of Pender (2014d).

Lastly, for the functional entropic risk measure, we have that

$$\text{Entropic} = \frac{1}{\gamma} \log\left(\mathbb{E}\left[e^{-\gamma \cdot (Q-k)^+}\right]\right) \tag{6.82}$$

$$= \frac{1}{\gamma} \log\left(\mathbb{E}\left[e^{-\gamma \cdot \sqrt{v} \cdot (X-\chi)^+}\right]\right) \tag{6.83}$$

$$= \frac{1}{\gamma} \log\left(\Phi(\chi) + \int_{\chi}^{\infty} e^{-\gamma \cdot \sqrt{v} \cdot (x-\chi)} \cdot \varphi(x) dx\right) \tag{6.84}$$

$$= \frac{1}{\gamma} \log\left(\Phi(\chi) + e^{\gamma \cdot \sqrt{v} \cdot \chi} \int_{\chi}^{\infty} e^{-\gamma \cdot \sqrt{v} \cdot x} \cdot \varphi(x) dx\right) \tag{6.85}$$

$$= \frac{1}{\gamma} \log\left(\Phi(\chi) + e^{\gamma \cdot \sqrt{v} \cdot \chi - \gamma \cdot v/2} \cdot \overline{\Phi}(\chi + \gamma \cdot \sqrt{v})\right). \tag{6.86}$$

□

References

Castillo, M. (2014). Man found dead in nyc hospital waiting room more than 8 hours after entering. http://www.cbsnews.com/news/man-found-dead-in-st-barnabas-hospital-waiting-room-8-hours/.

Cheridito, P., & Li, T. (2008). Dual characterization of properties of risk measures on orlicz hearts. *Mathematics and Financial Economics, 2*(1), 29–55.

Cheridito, P., & Li, T. (2009). Risk measures on orlicz hearts. *Mathematical Finance, 19*(2), 189–214.

Detlefsen, K., & Scandolo, G. (2005). Conditional and dynamic convex risk measures. *Finance and Stochastics, 9*(4), 539–561.

Eick, S. G., Massey, W. A., & Whitt, W. (1993). The physics of the mt/g/∞ queue. *Operations Research, 41*(4), 731–742.

Engblom, S., & Pender, J. (2014). Approximations for the moments of nonstationary and state dependent birth-death queues. arXiv preprint arxiv:1406.6164.

Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time–varying queues to achieve time-stable performance. *Management Science, 54*(2), 324–338.

Janssen, A., Van Leeuwaarden, J., Zwart, B., et al. (2008). Gaussian expansions and bounds for the poisson distribution applied to the erlang b formula. *Advances in Applied Probability, 40*(1), 122–143.

Jelenković, P., Mandelbaum, A., & Momčilović, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems, 47*(1–2), 53–69.

Jennings, O. B., Mandelbaum, A., Massey, W. A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science, 42*(10), 1383–1394.

Khudyakov, P., Feigin, P. D., & Mandelbaum, A. (2010). Designing a call center with an ivr (interactive voice response). *Queueing Systems, 66*(3), 215–237.

Klyman, J. (2011). *Systemic risk measures: DistVaR and other*. Princeton University.

Ko, Y. M., & Gautam, N. (2013). Critically loaded time-varying multiserver queues: computational challenges and approximations. *INFORMS Journal on Computing, 25*(2), 285–301.

Liu, Y., & Whitt, W. (2011). Large-time asymptotics for the g t/m t/s t+ gi t many-server fluid queue with abandonment. *Queueing systems, 67*(2), 145–182.

Liu, Y., & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations research, 60*(6), 1551–1564.
Liu, Y., Whitt, W., et al. (2014). Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability, 24*(1), 378–421.
Mandelbaum, A., Massey, W. A., & Reiman, M. I. (1998). Strong approximations for markovian service networks. *Queueing Systems, 30*(1–2), 149–201.
Mandelbaum, A., & Momcilovic, P. (2012). Queues with many servers and impatient customers. *Mathematics of Operations Research, 37*(1), 41–65.
Massey, W., & Pender, J. (2011). Skewness variance approximation for dynamic rate multi-server queues with abandonment. *Performance Evaluation Review, 39*. 74–74
Massey, W., & Pender, J. (2013). Approximation and Stabilizing Jackson Networks with Abandonment. *Technical Report*. Working Paper.
Massey, W., & Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment.. *Queueing Systems, 75*(2), 243–277.
Peccati, G., & Taqqu, M. S. (2011). *Wiener chaos: moments, cumulants and diagrams: a survey with computer implementation*: Vol. 1. Springer.
Pender, J. (2014a). Gram charlier expansions for time varying multi-server queues with abandonment. *SIAM Journal on Applied Mathematics, 74*(4), 1238–1265.
Pender, J. (2014b). Laguerre polynomial expansions for time varying multiserver queues with abandonment. Available at http://people.orie.cornell.edu/jpender/LSA.pdf.
Pender, J. (2014c). A Poisson-Charlier approximation for nonstationary queues. *Operations Research Letters, 42*(4), 293–298.
Pender, J. (2014d). Sampling the functional kolmogorov forward equations: applications to nonstationary queues. *INFORMS Journal on Computing*.
Pender, J. (2015). The truncated normal distribution: applications to queues with impatient customers. *Operations Research Letters, 43*(1), 40–45.
Rudloff, B., Sass, J., & Wunderlich, R. (2008). Entropic risk constraints for utility maximization. *Festschrift in celebration of Prof. Dr. Wilfried Grecksch?s 60th birthday. Aachen: Shaker Verlag. Berichte aus der Mathematik*, 149–180.
Ruszczyński, A., & Shapiro, A. (2006). Optimization of risk measures. In *Probabilistic and randomized methods for design under uncertainty* (pp. 119–157). Springer.
Stolletz, R. (2008). Approximation of the non-stationary m (t)/m (t)/c (t)-queue using stationary queueing models: the stationary backlog-carryover approach. *European Journal of operational research, 190*(2), 478–493.
Tirdad, A., Grassmann, W. K., & Tavakoli, J. (2016). Optimal policies of m (t)/m/c/c queues with two different levels of servers. *European Journal of Operational Research, 249*(3), 1124–1130.
Yarmand, M. H., & Down, D. G. (2013). Server allocation for zero buffer tandem queues. *European Journal of Operational Research, 230*(3), 596–603.
Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management, 16*(2), 283–299.
Zeltyn, S., & Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the m/m/n+ g queue. *Queueing Systems, 51*(3–4), 361–402.