Many-Server Queues with Autoregressive Inputs

Xu Sun

Received: date / Accepted: date

Abstract Recent studies have revealed the presence of significant autocorrelation and overdispersion in arrival data at large call centers. Motivated by these findings, we study a class of queueing systems where customers arrive according to a doubly stochastic Poisson point process whose intensities are driven by a time-dependent Cox-Ingersoll-Ross (CIR) process. The nonnegativity and autoregressive feature of the CIR process makes it a good candidate for modeling temporary dips and surges in arrivals. We conduct performance analysis of such systems. In particular, we study asymptotic performances such as the queue length and customer delays. The results acknowledge the presence of autoregressive structure in arrivals and produce operational insights into staffing decisions.

Keywords heavy-traffic approximations · autocorrelation · overdispersion · mean-reverting process · parameter uncertainty · outsourcing

1 Introduction

Poisson arrival is one of the most prevalent assumptions in queueing theory. Evidence that supports its validity is provided by [7,19]. A Poisson input facilitates the mathematical analysis and produces insights on capacity planning for service systems. For example, the performance of an $Mt/G/\infty$ model (with a nonhomogeneous Poisson arrival process and infinite service capacity) has simple expressions (see [9,26]) which gives rise to the celebrated square-root-staffing rule. Although natural and convenient from a mathematical point of view, the Poisson assumption does not always align well with real-life data. A growing body of empirical research has shown that the variance of the arrival count over a fixed time period tends to dominate its
mean, a common feature known as overdispersion; see e.g., [18,41,35]. This phenomenon violates the fundamental property that underpins the Poisson input assumption and can potentially affect performance evaluation and choice of staffing rule. In addition, arrival counts over successive periods during the day are observed to possess strong autocorrelation (see [17,34,37]), so that the demand volume may temporarily dip or surge. These non-Poisson features, such as overdispersion and autocorrelation, serve as the primary motivation for this work.

The central theme of this paper is to investigate how the autocorrelation structure of an arrival process can affect the performance of service system by considering an autoregressive input model and analyzing the corresponding queueing system in a heavy-traffic environment. Specifically, we model customer arrivals as a doubly stochastic Poisson process (DSPP) of which the intensities are driven by a time-dependent Cox-Ingersoll-Ross (CIR) process. The CIR process is typically used to model short rates (instantaneous interest rates), but they have some nice properties that can be leveraged to model the randomly varying intensities of a counting process. First, the CIR process will never become negative under appropriate regularity condition (known as the Feller condition). Second, the process periodically spikes but has a tendency to return to its mean level. Lastly, the process is very amenable to statistical inference. For example, the method of maximum likelihood estimation can be easily implemented for this process ([20, 29]). These properties make the CIR process a natural candidate for modeling the unpredictable and temporary fluctuations observed in arrival data at modern call centers.

We are by no means the first to use the CIR process to model the intensities of a counting process. Zhang, Hong and Zhang [41] propose to use a stationary CIR process to model the intensity of call arrivals and show that the proposed traffic model can faithfully reproduce the behavior of interest observed in practice. Moreover, they derive the scaling limit for the arrival process concerning “convergence in marginal distribution”. Here, we consider a more general CIR model by allowing the reversion level to be time-dependent and show that the arrival process (with proper scaling) converges weakly to a limit in the Skorokhod topology, a stronger version of Theorem 2 in [41]. We then rely on the resulting approximation to study many-server queues under suitable heavy-traffic regimes.

Our paper relates to a growing body of research that considers random arrival rate. Brown, Zhang and Zhao [8] develop an autoregressive model for the arrival rate that can capture the correlation across successive time periods. Avramidis, Deslauriers and L’Ecuyer [4] propose several arrival process models and examine their fit to call center data. They show that the performance strongly depends on the form of the arrival-rate process. Built upon the work of Harrison and Zeevi [15], Whitt [38] uses fluid-model analysis to derive staffing solutions for a call center with uncertain arrival rate and employee absenteeism. More recently, Koçağa, Armony and Ward [21] develop an economic model to aid staffing decisions in the presence of random arrival rates with a co-sourcing option. We too consider staffing problems in the context of call center con-sourcing, but we do not consider economic models; instead, we work directly with performance measures associated with customer delays and abandonments. The two works most closely related to ours are Gao and Zhu [12] and Koops, Boxma and Mandjes [22]. We discuss each in turn.

The authors of [12] model customer arrivals as a stationary Hawkes process and derive heavy-traffic approximations for an infinite-server queue fed by a Hawkes process. Their model is especially attractive when demand exhibits clustering and self-exciting features. In [23], the
authors analyze an infinite-server queue fed by a Cox process of which the intensities are driven by a short-rate process. They illustrate via both exact and asymptotic analysis that such traffic models can indeed capture the overdispersed demand observed in practice. There is one key difference between the modeling approaches of these two papers and the one adopted in the present paper. The intensity process in [12] and [23] has jumps that decay exponentially over time whereas our intensity process is continuous, positive and mean-reverting. These properties are especially suitable for modeling traffic sources with smoothly changing intensity. Moreover, the CIR process is more parsimonious as the main features are characterized by fewer parameters, each having clear physical interpretation.

Unlike [12] and [23] that study infinite-server models, we deal with constrained systems in which customer delay and abandonment can actually occur. Admittedly, an infinite-server queue can sometimes be used to approximate queueing systems with many servers, but a model with infinite capacity provides little indication of how the temporary dips and surges in arrivals can affect delay-related metrics. On the other hand, real-world service systems tend to operate in a resource-constrained environment, which is especially true for modern call centers and hospitals where the challenge is to translate service-quality metrics (expressed in terms of delay statistics) into concrete staffing decisions. An analysis of large-scale constrained systems, as we do here, sheds light on how the autocorrelation structure of the arrival process affects key performance indicators such as queue length and customer waiting time. Ultimately, we hope that our findings can help call center managers make informed staffing decisions in the presence of stochastic demand fluctuations.

The contribution and organization of this paper are as follows. We formally introduce the DSPP with CIR-driven intensities and summarize some of its key properties in §2.1. We describe the corresponding $M_{CIR}/M/s_t + G$ model and specify two operational regimes, namely, the rationalized and the efficiency-driven (ED) regime based on fluid model analysis in §2.3. Our first main result is Lemma 2, stated in §3.1, whose proof relies on the key Lemma 1. In Lemma 1, we establish what we believe the first FCTL for CIR processes which can be of independent interest. We present the FCLT results for the $M_{CIR}/M/s_t + G$ system in §3.2 and §3.3 under the rationalized and ED regimes, respectively. For the former, we show that the limit of the number-in-service process is a piecewise-linear diffusion driven by a superposition of Brownian motion and an integrated Ornstein-Uhlenbeck (OU) process, similar to that in [23]. Our analysis of the ED regime uses the framework introduced by [3]. In particular, we focus on the waiting time of the head-of-line (HoL) customer and show that this waiting-time process (with proper scaling) converges to a diffusion limit; we then establish the convergence of other relevant processes by invoking the continuous mapping theorem. Here, unlike [3] in which the authors base their proof of the FCLT on a separate proof of a functional weak law of large numbers (FWLLN), we establish the FCLT directly by applying a novel change of variables in a stochastic setting to derive a stochastic integral equation for the \textit{prelimit} HoL waiting-time process. Because our proof of the FCLT does not require proving a separate FWLLN, we are able to write a much shorter proof. In §4, we discuss implications to call center staffing and outsourcing. We demonstrate the operational benefits of demand forecasting in the presence of arrival rate uncertainty and propose a novel modeling framework that allows for real-time updating of staffing levels. We also take into account forecasting lead times and consider their effect on long-run performance.
measures. We show that given a fixed performance target, the choice of the lead time parameter delicately balances the outsourced capacity level and the number of in-house agents.

2 Preliminaries

2.1 Notation and conventions

We denote by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$, respectively, the sets of all real numbers, non-negative reals and nonnegative integers. We use $\lceil a \rceil$ to denote the least integer that is greater than or equal to $a$. For a real-value function $f$, we write $f[x_1, x_2]$ as shorthand for $f(x_2) - f(x_1)$. We use $z_\alpha$ to denote the quantile value from a standard normal distribution at $\alpha$. Let $(\mathcal{D}([0, \infty), \mathbb{R}), J_1)$ denote the space of càdlàg (right continuous with left limits) functions equipped with the Skorokhod $J_1$ topology, and write “$\Rightarrow$” for weak convergence. All random entities introduced in this paper are supported by a complete probability space $(\Omega, \mathcal{F}, P)$.

2.2 Doubly stochastic Poisson process with CIR intensities

We model the arrival process $A(t)$ as a DSPP with CIR-driven intensities, i.e.,

$$A(t) \equiv \Pi_a \left( \int_0^t \lambda(u)du \right) \quad t \geq 0,$$

where $\Pi_a(\cdot)$ is a unit-rate Poisson point process and $\lambda(t)$ is a stochastic process satisfying the following stochastic differential equation (SDE)

$$d\lambda(t) = \kappa(\alpha(t) - \lambda(t))dt + \sigma \sqrt{\lambda(t)}dB(t)$$

for $\kappa, \sigma$ and $\alpha(\cdot)$ being model parameters and $B(\cdot)$ being a standard Brownian motion that is independent of the point process $\Pi(\cdot)$. In fact, (2) represents the dynamics of a CIR process with mean-reversion speed $\kappa$, volatility rate $\sigma$, and shift $\alpha(t)$ at time $t$. In the context of call centers, the shift function $\alpha(\cdot)$ can be used to model the predictable time-varying patterns such as the time-of-day effect in arrivals. Throughout, we assume that the model parameters $\kappa, \sigma, \alpha(\cdot)$ satisfied the generalized Feller condition as specified by [27], so that $\lambda(\cdot)$ is always positively.

To facilitate the presentation, we summarize below some key properties of the DSPP that will prove useful in the subsequent analysis.

• (Markov property) The process $(\lambda(t), A(t))$ is Markovian with respect to the natural filtration $(\mathcal{F}_t)_{t \geq 0}$, and the intensity process $\lambda(t)$ itself is also Markovian. For all time intervals $(t_1, t_2]$,

$$\mathbb{E}[A(t_1, t_2)|\mathcal{F}_{t_1}] \stackrel{a.s.}{=} \mathbb{E} \left[ \int_{t_1}^{t_2} \lambda(u)du \bigg| \mathcal{F}_{t_1} \right],$$

where we have defined $A(t_1, t_2) \equiv A(t_2) - A(t_1)$. 

b (Martingale property) By the definition of the intensity process $\lambda(\cdot)$ in (2), we have that $A(t) - \int_0^t \lambda(u)du$ is a square integrable martingale with quadratic variable given by $\int_0^t \lambda(u)du$, so that

$$\left( A(t) - \int_0^t \lambda(u)du \right)^2 - \int_0^t \lambda(u)du$$

is also a martingale. We will apply this martingale property in the proofs of Lemma 2 and the main theorems.

c (First-order behavior) Let $\bar{\lambda}(t) \equiv \mathbb{E}[\lambda(t)]$ for each $t \geq 0$. Taking expectation on both sides of (2), we obtain

$$\bar{\lambda}(t) = \bar{\lambda}(0) + \int_0^t \kappa(\alpha(u) - \bar{\lambda}(u))du,$$

whereupon we see

$$\bar{\lambda}(t) = e^{-\kappa t} \bar{\lambda}(0) + \int_0^t \kappa e^{-\kappa(t-u)}\alpha(u)du.$$

Below, we will be using the function $\bar{\lambda}(\cdot)$ in characterizing the process limits of the arrival process.

As indicated earlier, the CIR process permits versatile correlation structure while also maintaining analytical tractability. Moreover, it is very easy to simulate; for an illustration of the sample path, see Fig. 1. Given a sample path of the intensity process $\lambda(t)$, one can then generate a sample of the arrival process $A(t)$ using the thinning method for nonhomogeneous Poisson process; see Chap. 2 of Ross [33].
2.3 The $M_{\text{CIR}}/M/s_t + G$ model

We are now ready to move to the corresponding queueing model. To start, let customers arrive to the system accordingly to a DSPP with CIR-driven intensity process $\lambda(\cdot)$ given by (2). Service times are assumed to be i.i.d. exponential random variables with rate $\mu$. We allow waiting customers to abandon the queue, and assume the abandonment times of successive arrivals to be i.i.d. random variables with cumulative distribution function (CDF) $F$ and probability density function (PDF) $f$. Moreover, we stipulate that service times and abandonment times are mutually independent, independent of the arrival processes.

The system adopts a work-conserving policy, i.e., no customers wait in queue if there is an available server. Let $Q(t)$ denote the number of customers in queue at time $t$. Furthermore, we use $E(t)$ and $R(t)$ to represent the number of customers that have entered service and the number of abandonments, all up to time $t$, respectively. By flow conservation,

$$Q(t) = Q(0) + A(t) - E(t) - R(t).$$

In addition, let $B(t)$ be the number of busy servers at time $t$ and $D(t)$ be the cumulative number of customer that have departed due to service completion up to time $t$. Again by flow conservation, we have

$$B(t) = B(0) + E(t) - D(t).$$

Finally, let $X(t)$ denote the head-count process recording the total number of customers in the system (both in queue and in service). Adding up (4) and (5) yields

$$X(t) = Q(t) + B(t) = X(0) + A(t) - D(t) - R(t).$$

Alternatively, one can derive (6) directly from flow conservation.

We now introduce two waiting-time processes that we will exploit heavily in the subsequent analysis. Let $H(t)$ denote the head-of-line waiting time (HWT) at time $t$, i.e., the waiting time of the customer who has been waiting the longest (if there is any); $H(t) = 0$ if there is no customer waiting in queue. Let $V(t)$ represent the potential waiting time (PWT) at time $t$, i.e., the waiting time of an arriving customer at time $t$ assuming the customer has infinite patience. With the newly introduced processes, namely, $H(t)$ and $V(t)$, we can conveniently express the enter-service and queue-length processes in the following way:

$$E(t) = \sum_{i=1}^{A(t-H(t))} \mathbb{1}_{\{\gamma_i > V(\tau_i)\}}$$

and

$$Q(t) = \sum_{i=A(t-H(t))}^{A(t)} \mathbb{1}_{\{\tau_i + \gamma_i > t\}} \text{ for } t \geq 0,$$

where the random variables $0 \leq \tau_1 \leq \tau_2 \leq \cdots$ denote arrival epochs, and $\gamma_1, \gamma_2, \ldots$ represent the abandonment times of successive customers that arrived to the system. As will become clear in the subsequent analysis, these representations are especially useful in deriving FCLT results for the overloaded $M_{\text{CIR}}/M/s_t + G$ system.
It remains to specify the staffing levels (number of servers). In practice, staffing levels are selected to trade off operational efficiency and service quality. Here we follow a constraint-satisfaction approach; i.e., the system operator (later referred to as the service provider) specifies performance metric(s) and then assigns the least staffing level that satisfies the target. Of particular interest is a constraint on the tail probability of delay (TPoD)

$$P(V(t) > w) \leq \varrho,$$

where $w$ is a delay target and $\varrho$ is a constant probability target. Common choices for $w$ are 0 and 20 s. When $w = 0$, the TPoD reduces to the delay probability. It has long been known that to stabilize the delay probability, the system would have to be critically loaded and has negligible delay ($V(t) \approx 0$); see, e.g., [11,13]. It is also known that with $w > 0$ the constraint (9) entails an overloaded system, in which case customer waiting times are comparable to the service times; see e.g., [25,28]. Because each condition gives rise to a different operational regime and produces different asymptotic results, we present them separately in the ensuing sections.

To formally define these two regimes, we look at the associated fluid model which can be regarded as a first-order approximation of the original stochastic system. The fluid model is obtained by replacing the discrete stochastic arrival process by its mean arrival-rate function $\bar{\lambda}(t)$. Because there is no stochastic element to the fluid model, one can staff to achieve perfect stabilization, i.e., each fluid atom that does no abandon waits exactly $w$ time units before entering service. Suppose for the moment that a staffing process has been selected to stabilize the fluid model. Let $\varepsilon(t)$ and $q(t)$ be the corresponding enter-service process (the cumulative amount of fluid that enters service up to the current time) and queue length (the amount of fluid content in queue) at time $t$, respectively. Just like how we derive (7) and (8), we get

$$\varepsilon(t) = \int_0^{t-w} F^c(w)\bar{\lambda}(u)du,$$

$$q(t) = \int_{t-w}^t F^c(t-u)\bar{\lambda}(u)du.$$  

Thus far we have implicitly assumed that the system starts empty at time 0 and a staffing process $m(\cdot)$ begins at time $w$. With this assumption, the policy should provide no staffing at all and thus allows no fluid to enter service until time $w$. On the other hand, fluid enters service at the rate $\dot{\varepsilon}(t) = \dot{m}(t) + \mu m(t)$. Combining with (10), we derive

$$m(t) = F^c(w) \int_w^t e^{-\mu(t-u)}\bar{\lambda}(u-w)du,$$

The result of (12) was previously derived by Liu and Whitt [26]; see eq. (13) there. Finally, let $\alpha^*$ be the long-run average mean-reversion level, namely,

$$\alpha^* = \lim_{T \to \infty} T^{-1} \int_0^T \alpha(u)du.$$  

We can then spell out our staffing formula for the $M_{cr}/M/s_t + G$ model

$$s(t) = m(t) + \sqrt{\alpha^*} c(t),$$

where $c(t)$ is the customer arrival process.
where $c(t)$ is a design function to be specified to meet the performance target. When all servers are busy but the staffing level is forced to decrease, we push the customer who entered service the latest back into the queue, as in [31]. If the shift function $\alpha(\cdot)$ is constant, i.e., $\alpha(t) \equiv \alpha$ for some $\alpha \in \mathbb{R}$, and further $c(t) \equiv c^*/\sqrt{\mu}$ for some $c^* \in \mathbb{R}$, then in view of (12), we have $m(t) = m(\infty) = \alpha/\mu$ for all $t \geq 0$, and so

$$s(t) = s \equiv \alpha/\mu + c^* \sqrt{\alpha/\mu}.$$ 

We thus recover the classical square-root staffing rule for the time-stationary model.

We say that a system resides in the rationalize regime if and (i) the staffing function $s(t)$ satisfies (14) and (ii) $m$ follows (12) for $w = 0$, in which case the fluid model will allow each (customer) fluid atom to enter service upon arrival. It is readily checked that the function $m$ coincides with the offered load process for the infinite-server model. That is, $m$ solves the differential equation

$$\dot{m}(t) = \bar{\lambda}(t) - \mu m(t) \quad \text{or} \quad m(t) - m(0) = \int_0^t \bar{\lambda}(u)du - \mu \int_0^t m(u)du. \quad (15)$$

Similarly, a system is said to operate in the ED regime if (i) the staffing function $s(t)$ satisfies (14) and (ii) $m$ follows (12) for $w > 0$, in which case each fluid atom waits exactly $w$ time units before entering service.

### 3 Heavy-traffic analysis

The presence of a stochastic arrival rate makes an exact analysis of the queueing system extremely difficult. This leads us to apply fairly standard approximation techniques used in the extant literature. In particular, we assume that the system is facing high demand volume and has a large number of servers. Below, we formally introduce our asymptotic framework and perform some preliminary analysis in §3.1. The main results are presented in §3.2 and §3.3 for the rationalized and ED regimes, respectively.

#### 3.1 Asymptotic framework

We consider an asymptotic framework in which the long-run average demand volume grows to infinity, i.e., $\alpha^* \to \infty$ for $\alpha^*$ given by (13). Following the convention in the literature, we will use $n$ in place of $\alpha^*$ as the scaling parameter. More precisely, we apply a linear scaling to the shift function, namely,

$$\alpha_n(t) \equiv n\alpha(t) \quad \text{for} \quad t \geq 0, \quad (16)$$

where, by slight abuse of notation, we used $\alpha(t)$ to denote the baseline mean-reversion level at time $t$. We subscript all relevant notation with $n$ to capture the dependence on this scaling parameter $n$. For example, $A_n$ denotes a DSPP with intensity process $\lambda_n$ satisfying

$$d\lambda_n(t) = \kappa(\alpha_n(t) - \lambda_n(t))dt + \sigma\sqrt{\lambda_n(t)}dB(t). \quad (17)$$
Note that $\kappa$ and $\sigma$ are fixed. That is, this sequence of CIR processes indexed by the scaling parameter $n$ shares common mean-reversion speed and volatility rate. It is readily checked that with this scaling, the mean value of the intensity process $\lambda_n(\cdot)$ and the number of arrivals over any fixed time period blow up linearly by a factor of $n$. The remainder of this section is devoted to showing that a sequence of properly scaled intensity processes converges weakly to a Gaussian process. For that purpose, we define

$$\bar{\lambda}_n(t) \equiv \lambda_n(t)/n \quad \text{and} \quad \hat{\lambda}_n(t) \equiv \sqrt{n} \left( \bar{\lambda}_n(t) - \bar{\lambda}(t) \right).$$

The result below establishes the FCLT for the sequence of intensity processes.

**Lemma 1 (FWLLN and FCLT for the arrival intensity process)** Suppose that the intensity process for the $n$-th model follows (17). If, in addition, there is convergence of the initial distribution at time 0, i.e., if

$$(\bar{\lambda}_n(0), \hat{\lambda}_n(0)) \Rightarrow (\bar{\lambda}(0), \hat{\lambda}(0)) \quad \text{in} \quad \mathbb{R}^2 \quad \text{as} \quad n \to \infty,$$

then we have the joint convergence

$$(\bar{\lambda}_n(t), \hat{\lambda}_n(t)) \Rightarrow (\bar{\lambda}(t), \hat{\lambda}(t)) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \quad (18)$$

where $\bar{\lambda}(\cdot)$ follows (3) and $\hat{\lambda}(\cdot)$ satisfies the stochastic integral equation

$$\hat{\lambda}(t) = \hat{\lambda}(0) - \kappa \int_0^t \hat{\lambda}(u)du + \sigma \int_0^t \sqrt{\bar{\lambda}(u)}dB(u). \quad (19)$$

Hence the diffusion limit of the arrival-rate process is an OU process whose solution admits a closed-form expression:

$$\hat{\lambda}(t) = e^{-\kappa t} \hat{\lambda}(0) + \sigma \int_0^t e^{-\kappa(t-u)} \sqrt{\bar{\lambda}(u)}dB(u). \quad (20)$$

Next define $A(t) \equiv \int_0^t \bar{\lambda}(u)du$ and the scaled versions of the arrival process:

$$\bar{A}_n(t) \equiv A_n(t)/n, \quad \hat{A}_n(t) \equiv n^{-1/2} (A_n(t) - n\Lambda(t)). \quad (21)$$

Per our previous discussion, it is natural to center $A_n(t)$ around $n\Lambda(t)$. We will show in Lemma 2 below that this centering indeed gives rise to meaningful limit.

**Lemma 2 (FWLLN and FCLT for the arrival process)** The centered and normalized version of the arrival process $\hat{A}_n$ satisfies an FCLT:

$$\hat{A}_n(t) \Rightarrow \hat{A}(t) \equiv B_\lambda \left( \int_0^t \bar{\lambda}(u)du \right) + \mathcal{K}(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \quad (22)$$

for

$$\mathcal{K}(t) \equiv \int_0^t \hat{\lambda}(u)du, \quad (23)$$
where $\hat{\lambda}(\cdot)$ is given in (19) and $B_{\lambda}(\cdot)$ is a standard Brownian motion independent of $K(\cdot)$. As an immediate consequence, we have the WFLLN

$$ \hat{A}_n(t) \Rightarrow \Lambda(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \rightarrow \infty, $$

jointly with (22).

Remark 1  Accordingly to Lemma 2, the diffusion-scaled arrival process $\hat{A}_n(\cdot)$ converges to a Gaussian process which is characterized by two independent terms. The first term is a time-changed Brownian motion that arises from the inherent variability in the Poisson process $\Pi_a$, while the second term is an integrated OU process that stems from the stochasticity of the intensity process. One can solve $K(\cdot)$ in closed form. Indeed, substituting (20) into (19) and rearranging, one gets

$$ K(t) = \frac{1}{\kappa} \left( 1 - e^{-\kappa t} \right) \hat{\lambda}(0) + \frac{\sigma}{\kappa} \int_0^t \left( 1 - e^{-\kappa(t-u)} \right) \sqrt{\hat{\lambda}(u)} dB(u). $$

If the shift function of the CIR process is constant, i.e., $\alpha_n(t) \equiv n$, then Lemma 2 applies to the time-stationary model as an important special case.

Corollary 1 (the stationary case) If the CIR-driven intensity $\lambda_n$ mean-reverts to a constant $n$ and $\hat{\lambda}(0) = 0$, then the scaled arrival process $\hat{A}_n$ satisfies an FCLT:

$$ \hat{A}_n(t) \Rightarrow \hat{A} \equiv B_{\lambda}(t) + K(t), $$

where $B_{\lambda}$ is a standard Brownian motion and

$$ K(t) \equiv \frac{\sigma}{\kappa} \int_0^t \left( 1 - e^{-\kappa(t-u)} \right) dB(u). $$

for $B$ being a standard Brownian motion independent of $B_{\lambda}$. In addition, for $s \leq t$ for covariance between $\hat{A}(s)$ and $\hat{A}(t)$ can be computed using the following formula

$$ \text{Cov}(\hat{A}(s), \hat{A}(t)) = s + \frac{\sigma^2}{\kappa^3} \left( \kappa s - 1 + e^{-\kappa s} + e^{-\kappa t} \right) - \frac{\sigma^2}{2\kappa^3} \left( e^{-\kappa(t-s)} + e^{-\kappa(t+s)} \right). $$

In particular, the formula for the variance is given by

$$ \text{Var}(\hat{A}(t)) = \left( 1 + \frac{\sigma^2}{\kappa^2} \right) t - \frac{3\sigma^2}{2\kappa^3} + \frac{2\sigma^2}{\kappa^3} e^{-\kappa t} - \frac{\sigma^2}{2\kappa^3} e^{-2\kappa t}. $$

Note that when $\sigma = 0$, $\text{Var}(A_n(t)) = nt$. So with deterministic arrival rate, the variance of the number of arrivals up to time $t$ is equal to its mean, and this is the level of variability that staffing levels are typically chosen to handle. If, however, $\sigma = 2$ and $\kappa = 1$, then from (25) it follows that the variance can be five times the mean. It is apparent that such a high degree of variability can exert considerable influence on the system performance; that is the theme of the remainder of the the article.
To proceed, it is convenient (and natural) to consider a sequence of queueing systems, indexed by the scaling parameter $n$. In the $n$-th model, customers arrive to the system according to $A_n$, i.e., $A_n(t)$ represents the number of arrivals over $[0, t]$. The service rate $\mu$ and the abandonment-time distribution $F$ are held fixed, but the staffing process is allowed to grow linear with $n$, so that the corresponding staffing function satisfies

$$s_n(t) = nm(t) + \sqrt{n}c(t),$$

(26)

where the base-line staffing process $m$ is in the form of (12) and $c$ is a control function to be determined to meet the performance target (9).

At this point we make the key assumption to obtain our heavy-traffic approximations for the $M_{cIR}/M/s + G$ model.

**Assumption 1** The function $\bar{\lambda}(t)$ is bounded away from zero and has bounded first-order derivative.

**Assumption 2** The PDF $f$ of the abandonment time is bounded and the complementary CDF $F^c(x) \equiv 1 - F(x)$ is bounded away from zero on any compact interval.

### 3.2 Rationalized regime

**Heavy-traffic scalings** For the headcount and queue-length processes, define their centered and normalized version as

$$\hat{X}_n(\cdot) \equiv n^{-1/2}(X_n(\cdot) - s_n(\cdot)) \quad \text{and} \quad \hat{Q}_n(\cdot) \equiv n^{-1/2}Q_n(\cdot),$$

where the centering component $s_n$ is given by (26) for $m$ there being in the form of (15). For the waiting-time processes, define their scaling version as follows

$$\hat{H}_n(\cdot) \equiv \sqrt{n}H_n(\cdot) \quad \text{and} \quad \hat{V}_n(\cdot) \equiv \sqrt{n}V_n(\cdot).$$

It is well known that, for a many-server queue having customer abandonment and operating in the rationalized regime, only the density of the abandonment-time distribution at the origin plays a role in the corresponding diffusion limit; see Zeltyn and Mandelbaum [40] and Reed and Tezcan [32]. Therefore, without loss we may assume the abandonment times to be exponentially distributed with rate $\theta$, in which case Assumption 2 is trivially satisfied.

The theorem below establishes the FCLT results showing that the above diffusion-scaled processes converge weakly to their corresponding limits.

**Theorem 1** (FCLT for the critically-loaded $M_{cIR}/M/s + M$ model) Suppose customers arrive according to the DSPP $A_n(\cdot)$ with intensity process $\lambda_n(\cdot)$ given by (17), the system is staffed according to (26). Then we have the joint convergence with (18)

$$(\hat{X}_n, \hat{Q}_n, \hat{H}_n, \hat{V}_n) \Rightarrow (\hat{X}, \hat{Q}, \hat{H}, \hat{V}) \quad \text{in} \quad D^4 \quad \text{as} \quad n \to \infty$$
where the limiting process $\hat{X}$ satisfies stochastic integral equation

$$
\hat{X}(t) = \hat{X}(0) - c[0,t] - \mu \int_0^t c(u)du - \int_0^t \Psi(\hat{X}(u))du + \mathcal{K}(t) + \int_0^t \vartheta(u)d\mathcal{B}_2(u),
$$

for $\Psi(x) \equiv \mu x^- + \theta x^+$, $\mathcal{K}$ given by (23), $\vartheta(t) \equiv \sqrt{\lambda(t) + \mu m(t)}$, and $\mathcal{B}_2$ being a standard Brownian motion, independent of the process $\mathcal{K}$; in addition, the limits of queue-length and waiting-time processes satisfy

$$
\hat{Q}(t) = [\hat{X}(t)]^+ \quad \text{and} \quad \hat{H}(t) = \hat{V}(t) = \hat{Q}(t)/\bar{\lambda}(t).
$$

Theorem 1 applies to the stationary model as an important special case when the mean-reversion level is constant. In particular, the mean-reversion level is equal to $n$ for the $n$-th stochastic model.

**Corollary 2 (the stationary case)** Suppose that the conditions in Theorem 1 are satisfied. In addition, $\alpha_n \equiv n$ in equation (17) and $c(t) \equiv c$ for some constant $c$. Then

$$
(\hat{X}_n, \hat{Q}_n, \hat{H}_n, \hat{V}_n) \Rightarrow (\hat{X}, \hat{Q}, \hat{H}, \hat{V}) \quad \text{in} \quad \mathcal{D}^4 \quad \text{as} \quad n \to \infty
$$

where the limiting process $\hat{X}$ satisfies stochastic integral equation

$$
\hat{X}(t) = \hat{X}(0) - \mu ct - \int_0^t \Psi(\hat{X}(u))du + \mathcal{K}(t) + \sqrt{2B_2(t)},
$$

for $\mathcal{K}$ specified by (24) and $\mathcal{B}_2$ being a standard Brownian motion, independent of $\mathcal{K}$; in addition, the limits of queue-length and waiting-time processes are in the form of (27) with $\bar{\lambda}(t)$ there being replace by 1.

Corollary 2 is in agreement with Theorem 2 of Garnett et al. [13] and Theorem 7.1 of Pang et al. [30] except for the additional term $\mathcal{K}(\cdot)$ that arises naturally from the autoregressive assumption of the arrival-rate process. It is clear that the randomly varying arrival rates introduce additional variability into the system which needs to be handled by additional safety capacity. We will explore this issue further in §4.

3.3 ED regime

**Heavy-traffic scalings** For the headcount process $X_n$ and the queue-length process $Q_n$, we define their centered and normalized version as

$$
\hat{X}_n(\cdot) \equiv n^{-1/2}(X_n(\cdot) - s_n(\cdot)) \quad \text{and} \quad \hat{Q}_n(\cdot) \equiv n^{-1/2}(Q_n(\cdot) - nq(\cdot)),
$$

where $s_n(\cdot)$ is specified by (26) for $m$ there being in the form of (12) and $q(\cdot)$ is given by (11). Next we define the diffusion-scaled HWT and PWT processes

$$
\hat{H}_n(\cdot) \equiv \sqrt{n} (H_n(\cdot) - w) \quad \text{and} \quad \hat{V}_n(\cdot) \equiv \sqrt{n} (V_n(\cdot) - w).
$$

Our next theorem establishes the FCLT result for the above centered and normalized versions.
Theorem 2 (FCLT for the overloaded $M_{c_{d}}/M/s_{t} + G$ model) Suppose customers arrive according to the DSPP $A_{n} (\cdot)$ with intensity process $\lambda_{n} (\cdot)$ given by (17) and the system is staffed according to (26). In addition, Assumptions 1 and 2 hold. Then we have the joint convergence

$$(\hat{X}_{n}, \hat{Q}_{n}, \hat{H}_{n}, \hat{V}_{n}) \Rightarrow (\hat{Q}, \hat{H}, \hat{V}) \text{ in } \mathcal{D}^{4},$$

where $\hat{H}$ satisfies the stochastic integral equation

$$F^{c}(w)\bar{\lambda}(t-w)\hat{H}(t) = -c[w, t] - \mu \int_{w}^{t} c(u)du - \int_{w}^{t} f(w)\bar{\lambda}(u-w)\hat{H}(u)du + \mathcal{G}(t),$$

for the noise term $\mathcal{G}(t)$ given by

$$\mathcal{G}(t) \equiv F^{c}(w) \int_{0}^{t-w} \sqrt{\bar{\lambda}(u)}dB_{\bar{\lambda}}(u) + F^{c}(w) \int_{0}^{t-w} \bar{\lambda}(u)du$$

$$- \int_{w}^{t} \sqrt{\mu m(u)}dB_{\mu}(u) + \sqrt{F^{c}(w)F(w)} \int_{0}^{t-w} \sqrt{\bar{\lambda}(u)}dB_{\theta}(u)$$

where $B_{\bar{\lambda}}, B_{\mu}, B_{\theta}$ are mutually independent Brownian motions. The limit for the PWT process is a deterministic functional of $\hat{H}$ and satisfies $\hat{V}(t) = \hat{H}(t+w)$; the FCLT limit for queue length $\hat{Q}$ is the sum of three terms, namely,

$$\hat{Q}(t) \equiv \hat{Q}_{1}(t) + \hat{Q}_{2}(t) + \hat{Q}_{3}(t),$$

where

$$\hat{Q}_{1}(t) \equiv \int_{t-w}^{t} F^{c}(t-u)\sqrt{\bar{\lambda}(u)}dB_{\bar{\lambda}}(u) + \int_{t-w}^{t} F^{c}(t-u)\bar{\lambda}(u)du,$$

$$\hat{Q}_{2}(t) \equiv \int_{t-w}^{t} \sqrt{F^{c}(t-u)F(t-u)\bar{\lambda}(u)}dB_{\mu}(u), \quad \hat{Q}_{3}(t) \equiv F^{c}(w)\bar{\lambda}(t-w)\hat{H}(t).$$

Remark 2 (Separation of variability) The noise term $\mathcal{G}$ in (29) is characterized by four independent terms. The first term captures the inherent variability in the Poisson process $\Pi_{\alpha}$ given in (1); the second term stems from the stochasticity of the arrival-rate process; the third term comes from the variability of the service-completion process; and finally the fourth term accounts for the randomness of customer abandonment.

Just like what we did in Corollary 2, we obtain an important special case for the overloaded $M_{c_{d}}/M/s_{t} + G$ system when the intensity process mean reverts to a constant level.

Corollary 3 (the stationary case) Suppose that the conditions in Theorem 2 are satisfied. In addition, $\alpha_{n}(t) \equiv n$, and $c(t) \equiv c$ for some constant $c$. Then

$$(\hat{H}_{n}, \hat{V}_{n}) \Rightarrow (\hat{H}, \hat{V}) \text{ in } \mathcal{D}^{2},$$

where $\hat{V}(t) = \hat{H}(t+w)$ and $\hat{H}$ satisfies the stochastic integral equation

$$\hat{H}(t) = -\frac{\mu c}{F^{c}(w)} \int_{w}^{t} du - h_{F} \int_{w}^{t} \hat{H}(u)du + \tilde{\mathcal{G}}(t),$$

(30)
for $h_F \equiv h_F(w)$ being the value of the hazard rate function of the distribution $F$ evaluated at the delay target $w$ and the noise term $\tilde{G}(t)$ given by

$$\tilde{G}(t) \equiv \frac{1}{\sqrt{F^c(w)}} \left( \int_0^{t-w} dB_2(u) - \int_w^t dB_\mu(u) \right) + \int_0^{t-w} \lambda(u) du,$$

(31)

where $B_2$ and $B_\mu$ are two independent Brownian motions, independent of $\hat{\lambda}$.

When the arrival rate is deterministic (i.e., $\hat{\lambda}$ degenerates to zero), the last term of (31) disappears and we recover the FCLT results of Aras, Chen and Liu [3] for the overloaded $M/M/s+G$ system as an important special case.

4 Implications on call center staffing

One of the main challenges in operating a telephone call center is finding correct staffing levels to meet desired performance target. A common approach for achieving prescribed performance target is to use the square root staffing law to estimate the amount of capacity needed assuming that call arrivals follow a Poisson process with a fixed rate parameter (see [6,13,28]). We will argue that, when the arrival rates themselves are modeled as a random process, a naive application of the square-root staffing rule can fail to achieve the desired levels of service quality. Thus, it would be worthwhile to investigate techniques for selecting staffing levels in the context of stochastic arrival rates.

Here we restrict our discussion to the rationalized regime only, but a similar analysis and the operational insights carry over to the overloaded case. Further, for clarity of exposition, we take the mean-reversion level of the CIR-driven intensity process to be a constant so that $\alpha(t) \equiv \alpha$.

To gain greater simplicity and tractability, we stipulate that the service rate and the abandonment rate are equal, in which case we would get $\Psi(x) \equiv \mu x - \theta x^+ = \mu x$. By Corollary 2, the limiting number-in-system process $\hat{X}$ satisfies

$$\hat{X}(t) = e^{-\mu t} \hat{X}(0) - c \int_0^t \mu e^{-\mu(t-u)} du + \int_0^t e^{-\mu(t-u)} \hat{\lambda}(u) du + \sqrt{2} \int_0^t e^{-\mu(t-u)} dB_2(u).$$

(32)

From (32) we can directly compute the mean

$$E[\hat{X}(t)|\hat{X}(0)] = e^{-\mu t} \hat{X}(0) - c \int_0^t \mu e^{-\mu(t-u)} du \rightarrow -c \text{ as } t \rightarrow \infty.$$

To derive the variance formula for $\hat{X}(t)$, we apply Fubini’s theorem to obtain

$$\int_0^t e^{-\mu(t-u)} \hat{\lambda}(u) du = \sigma \int_0^t e^{-\mu(t-u)} e^{\kappa s} \left( \int_s^t e^{\mu u - \kappa u} du \right) dB(s)$$

$$= \frac{\sigma}{\mu - \kappa} \int_0^t \left( e^{-\kappa(t-s)} - e^{-\mu(t-s)} \right) dB(s),$$
so that the variance of this term can be derived using Itō isometry, yielding
\[
\text{Var} \left( \int_0^t e^{-\mu(t-u)} \lambda(u) \, du \right) = \frac{\sigma^2}{(\mu - \kappa)^2} \int_0^t \left( e^{-\kappa(t-s)} - e^{-\mu(t-s)} \right)^2 \, ds
\]
\[
= \frac{\sigma^2}{(\mu - \kappa)^2} \left[ \frac{1}{2\kappa} \left( 1 - e^{-2\kappa t} \right) - \frac{2}{\kappa + \mu} \left( 1 - e^{-(\kappa+\mu)t} \right) + \frac{1}{2\mu} \left( 1 - e^{-2\mu t} \right) \right].
\]
On the other hand, applying Itō isometry again allows us to conclude
\[
\text{Var} \left( \int_0^t e^{-\mu(t-u)} \, dB_2(u) \right) = \int_0^t e^{-2\mu(t-u)} \, du = \frac{1}{2\mu} \left( 1 - e^{-2\mu t} \right).
\]
Thus, by combining above and sending \( t \to \infty \), we obtain the following result about the long-term behavior of the limiting process \( \hat{X} \).

**Corollary 4** As \( t \to \infty \), the sequence of random variables \( \hat{X}(t) \) in (32) converges weakly to \( \hat{X}(\infty) \) which is a Gaussian random variable with mean \(-c\) and variance
\[
\text{Var}(\hat{X}(\infty)) = \frac{\sigma^2}{2\kappa\mu(\kappa+\mu)} + \frac{1}{\mu}.
\]

Corollary 4 states that when the mean-reversion level \( \alpha \) is large, we can heuristically approximate the steady-state distribution of the number of customers in the \( M_{\text{CIR}}/M/s + M \) model as follows:
\[
X(\infty) \approx \frac{\alpha}{\mu} + \sqrt{\alpha k} \cdot \mathcal{N} \quad \text{for} \quad k \equiv k(\mu, \kappa, \sigma) \equiv \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\kappa\mu(\kappa+\mu)}}
\]
and \( \mathcal{N} \) being a standard normal random variable. Recall the number of servers was selected according to the square-root staffing formula, which we reproduce here for convenience
\[
s \equiv \alpha/\mu + \sqrt{\alpha c}.
\]
Combining (33) and (34) yields
\[
\mathbb{P}(V(\infty) > 0) = \mathbb{P}(X(\infty) > s) \approx \mathbb{P}(\mathcal{N} > c/k).
\]
From this normal approximation we immediately understand that, in order to stabilize the delay probability at the target value \( \varphi \), one ought to choose
\[
c \equiv z_{1-\varphi} \cdot \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\kappa\mu(\kappa+\mu)}}
\]
in the square-root staffing formula (34). The result of (35) demonstrates that ignoring a randomly varying arrival rate in making staffing decisions can result in severe under-staffing. In other words, in the presence of stochastic arrival rates, the use of square-root staffing rule would lead to a higher safety staffing level compared to the case where the arrival rate is deterministic. This happens because of a mismatch between the realized arrival rate and the number of servers.
available to handle those demands. As a result, the service provider needs to hire additional staff (corresponding to the second term under the square root) to ensure that the system can handle a larger-than-foreseen demand volume without jeopardizing the quality of service. However, hiring extra staff as “insurance” for high-volume periods can be very expensive while also resulting in excessive server idleness over low-volume periods. This suggests that contracting some of the call center services, also commonly referred to as “co-sourcing”, could be an effective strategy for managing demand fluctuations.

Two forms of outsourcing contracts have been considered in the literature; see e.g., [1]. The first type is in a form of capacity reservation whereby the service provider reserves certain amount of capacity with the contractor and pays a reservation fee regardless whether the reserved capacity will be used up or not. The second type, on the other hand, allows the service provider to answer calls in house up to a predetermined level, beyond which calls are handed over to the contract. In other words, overflow calls are diverted, in which case the contractor charges a fee per call outsourced. Here we consider the first form of sharing and leave the second type for future studies.

Below, we propose a flexible modeling framework in the context of call center co-sourcing that allows for real-time updating of demand forecasts and staffing levels. To keep the program simple, we assume that the instantaneous arrival rate, once realized, is revealed to the service provider. Although this assumption is a bit unrealistic, as the arrival rates are often not directly observable, the service provide can rely on historical data and implement effective algorithms (e.g., exponential smoothing and Bayesian techniques) to gain reliable estimate of the materialized arrival rates. Thus, we feel that the assumption is not too outlandish.

To make the model more realistic (and more general), we incorporate forecasting lead times (denoted by $\tau$) and consider their effect on prediction error. We will show that the forecasting error influences the system dynamics in a rather explicit fashion, and so the choice of the lead time parameter can indirectly affect system performance. The inclusion of forecasting lead times and a real-time updating of staffing level in queueing analysis is relatively nonstandard. The only previous examples we know of are [2] and [17], both of which consider the effect of lead times on prediction accuracy. But they do not explore in detail the implications on queueing performance.

To proceed, we assume that, at each time $t$, the service provider observes the materialized arrival rate $\lambda(t)$ and computes its deviate from the long-run average $\Delta\lambda(t) \equiv \lambda(t) - \alpha$. The forecast of the arrival rate at the future time $t + \tau$, denoted by $\hat{\lambda}(t + \tau)$, utilizing the information up to time $t$ is given by

$$\hat{\lambda}(t + \tau) = \alpha + e^{-\kappa \tau} \Delta\lambda(t).$$

(36)

The forecasting formula (36) is inspired by the established heavy-traffic approximations and derived based on an intuitive argument: Given that the limiting process $\hat{\lambda}(t)$ in (19) satisfies

$$\mathbb{E}\left[\hat{\lambda}(t + \tau) | \mathcal{F}_t\right] = e^{-\kappa \tau} \hat{\lambda}(t) \quad \text{for} \quad t \geq 0,$$

one would expect

$$\mathbb{E}\left[\Delta\lambda(t + \tau) | \mathcal{F}_t\right] \approx e^{-\kappa \tau} \Delta\lambda(t)$$

and so

$$\mathbb{E}\left[\lambda(t + \tau) | \mathcal{F}_t\right] \approx \alpha + e^{-\kappa \tau} \Delta\lambda(t).$$
In the meantime the service provider uses a modified square-root staffing rule to decide the staffing level by letting

\[ s(t + \tau) = \frac{\alpha}{\mu} + \sqrt{\alpha c} + e^{-\kappa \tau} \int_0^t e^{-\mu(t-u)} \Delta \lambda(u) \, du, \]

or

\[ s(t) = \frac{\alpha}{\mu} + \sqrt{\alpha c} + e^{-\kappa \tau} \int_0^t e^{-\mu(t-u)} \Delta \lambda(u - \tau) \, du, \]  \hspace{1cm} (37)

where, by default, \( \Delta \lambda(s) = 0 \) for all \( s \leq 0 \). The first two terms can be thought of as the desired long-run service capacity that the service provider would like to maintain over time. The remaining term can be viewed as flexible staffing that evolves in accordance with the updated forecast of arrival rates. Note that this term can be either positive or negative. Positive values are naturally interpreted as the outsourced service capacity; a negative value, on the other hand, could mean arranging agents to come off duty, having them take training sessions or work on alternative tasks. The ability to adjust the staffing levels may also follow from calling in part-time or work-from-home agents, as has been discussed by Ibrahim [16]. As before, we treat \( \tilde{c} \) as a design parameter to be determined based on the performance target and the choice of the lead time parameter \( \tau \), which we will explore in greater detail.

The above derivation may appear somewhat heuristic in nature. It turns out that the predictor obtained in (36) is asymptotically unbiased. As before, we consider a sequence of models, indexed by \( n \), where the mean-reversion level of the CIR-driven intensity equals \( n \) for the \( n \)-th stochastic model. The following result is proved in the appendix.

**Proposition 1** Suppose that the conditions in Corollary 2 are satisfied except that the staffing function now follows the modified square-root staffing rule specified by (37). Then

\[ \hat{X}_n \Rightarrow \hat{X} \text{ in } \mathcal{D} \text{ as } n \to \infty \]

where the limiting process \( \hat{X} \) satisfies stochastic integral equation

\[ \hat{X}(t) = \hat{X}(0) - \mu \hat{c} t - \int_0^t \Psi(\hat{X}(u)) \, du + \sigma \int_0^t \int_{u-\tau}^u e^{-\kappa(u-s)} \, dB(s) \, du + \sqrt{2} B_2(t), \]  \hspace{1cm} (38)

for \( B \) and \( B_2 \) being two independent standard Brownian motions.

Proposition 1 explicitly quantifies the impact of forecasting errors on the system dynamics. More specifically, the first noise term on the right-hand side of (38) arises from the staffing formula (37), and can be viewed as the aggregate forecasting error up to time \( t \). Consistent with our intuition, the magnitude of this error is an increasing function of the lead time parameter \( \tau \). The second noise term, on the other hand, captures the remaining variability coming from the Poisson point process \( \Pi_a \) and the service-completion process.

We see that the choice of the lead time parameter \( \tau \) delicately balances the prediction accuracy and the long-run outsourced capacity level. To elaborate, by using longer lead times, the
service provider can effectively reduce the number of calls to be diverted to the contractor, as can be glimpsed through (37); however the long-run performance can be seriously impacted by the greater variability caused by the bigger forecasting errors. Similarly, shorter lead times can improve the prediction accuracy, but at the cost of having more calls handed over to the contractor. The rest of the section is concerned with finding the correct quality-of-service (QoS) parameter $\tilde{c}$ in (37) to achieve desired long-run performance target for each fixed lead time parameter.

Driven by tractability concerns, we again require the abandonment rate to be equal to the service rate, in which case we can solve (38) explicitly, yielding

$$\hat{X}(t) = e^{-\mu t} \hat{X}(0) - \tilde{c} \int_0^t \mu e^{-\mu(t-u)} du + \sqrt{2} \int_0^t e^{-\mu(t-u)} dB_2(u) + \sigma \Theta(t)$$

where we defined

$$\Theta(t) \equiv \int_0^t e^{-\mu(t-u)} \left( \int_{u-\tau}^u e^{-\kappa(u-s)} dB(s) \right) du.$$

Combining above and sending $t \to \infty$, we obtain the following result about the long-term behavior of the limiting process $\hat{X}$ given by (39).

**Corollary 5** As $t \to \infty$, the sequence of random variables $\hat{X}(t)$ in (39) converges weakly to $\hat{X}(\infty)$ which is a Gaussian random variable with mean $-\tilde{c}$ and variance

$$\text{Var}(\hat{X}(\infty)) = \frac{\sigma^2}{2\kappa \mu (\kappa + \mu)} + \frac{1}{\mu} + \frac{\sigma^2}{\mu (\kappa - \mu)} \left( \frac{1}{2\kappa} e^{-2\kappa \tau} - \frac{1}{\kappa + \mu} e^{-(\kappa + \mu) \tau} \right).$$

It is readily checked that the steady-state variance, as a function of $\tau$, is monotonically increasing. Its minimum value is attained at $\tau = 0$ and equal to $1/\mu$, in which case the forecasting error vanishes entirely. The maximum value equals

$$\frac{\sigma^2}{2\kappa \mu (\kappa + \mu)} + \frac{1}{\mu}$$

and is achieved at $\tau = \infty$. We find that the maximum value coincides with that in Corollary 4. The results of Corollary 5 suggest an approximation for the steady-state distribution of the number of customers in the $M_{CIR}/M/s + M$ queue operating under the modified square-root staffing rule, which is given by

$$X(\infty) \approx \frac{\alpha}{\mu} + \sqrt{\alpha \tilde{k}} \cdot \mathcal{N}$$

for

$$\tilde{k} \equiv \tilde{k}(\mu, \kappa, \sigma) \equiv \sqrt{\frac{\sigma^2}{2\kappa \mu (\kappa + \mu)} + \frac{1}{\mu} + \frac{\sigma^2}{\mu (\kappa - \mu)} \left( \frac{1}{2\kappa} e^{-2\kappa \tau} - \frac{1}{\kappa + \mu} e^{-(\kappa + \mu) \tau} \right)}$$

and $\mathcal{N}$ being a standard normal random variable.
Hence, in order to stabilize the probability of delay at the target $\rho$, it suffices to choose $\tilde{c}$ in (37) such that
\[ P(N > \tilde{c}/\tilde{k}) = \rho, \]
from which we conclude
\[ \tilde{c} = z_{1-\rho} \tilde{k}. \]
We see that the value of $\tilde{k}$ increases with the lead time parameter $\tau$. So we can think of the QoS coefficient $\tilde{c}$ as an increasing function in $\tau$.

To summarize, when the performance target is fixed, the choice of the lead time parameter $\tau$ balances the in-house staffing level and the outsourced capacity: Longer lead times would require the service provider to hire more in-house agents whereas shorter lead times would require the service provider to reserve more capacity with the contractor. The optimal combination will be determined by economic implications, which we do not pursue further in this paper. We leave that for future research.

5 Proofs

Proof of Lemma 1. Dividing both sides of (17) by $n$ and recalling (16), we obtain
\[ \bar{\lambda}_n(t) = \bar{\lambda}_n(0) + \kappa \int_0^t (\alpha(u) - \bar{\lambda}_n(u))du + \sigma_n \int_0^t \sqrt{\bar{\lambda}_n(u)}dB(u), \] (41)
where we have defined $\sigma_n \equiv \sigma/n.$ We prove the WFLLN by arguing that the volatility term vanishes as $n \to \infty.$ Because $\sigma_n \to 0$ as $n \to \infty,$ it suffices to argue that the sequence $\{\bar{\lambda}_n; n \in \mathbb{N}\}$ is stochastically bounded. Here we appeal to Lemma 3.9 of [39]. In particular, if the sequence has continuous sample paths, then the proof of stochastic boundedness amounts to verifying the modulus of continuity condition (MCC); we refer the reader to Theorem 3.2 in [39] or Theorem 16.8 in Billingsley [5] for a formal definition of MCC. Towards that end, we verify the (sufficient) moment condition laid out in Lemma 3.11 (ii.b) of [39]. By (41), we have
\[
\mathbb{E} \left[ (\bar{\lambda}_n(t+u) - \bar{\lambda}_n(t))^2 \middle| \mathcal{F}_t \right] \\
\leq 2\mathbb{E} \left[ \left( \int_t^{t+u} \kappa (\alpha(s) - \bar{\lambda}_n(s)) ds \right)^2 \right] + 2\sigma_n^2 \mathbb{E} \left[ \left( \int_t^{t+u} \sqrt{\bar{\lambda}_n(s)}dB(s) \right)^2 \right] \\
\leq 2u \mathbb{E} \left[ \int_t^{t+u} \kappa^2 (\alpha(s) - \bar{\lambda}_n(s))^2 ds \right] + 2\sigma_n^2 C_p \mathbb{E} \left[ \int_t^{t+u} \bar{\lambda}_n(s) ds \right] \\
\leq 2u \int_0^T \kappa^2 \left( (\alpha(s))^2 + 2\alpha(s)\mathbb{E} [\bar{\lambda}_n(s)] + \mathbb{E} [\bar{\lambda}_n^2(s)] \right) ds \\
+ \sigma_n^2 C_p u^{1/4} \left( 1 + \int_0^T \mathbb{E} [\bar{\lambda}_n^2(s)] ds \right),
\]
where the first inequality follows from (41) plus the basic algebraic relation $(a + b)^2 \leq 2(a^2 + b^2)$, the second inequality follows by applying the Cauchy-Schwartz inequality to the first term.
and the Burkholder-Davis-Gundy inequality to the second term, and the third inequality follows by appealing to (6.1) in [14]. It is immediate by Lemma 3 that both integrals on the right-hand side approach zero as \( u \to 0 \), uniformly over all \( t \) and \( n \). This shows that the MCC stated in Lemma 3.11 (ii.b) of [39] is indeed satisfied. Then applying Lemma 3.9 in [39], we conclude the stochastic boundedness of \( \{\bar{\lambda}_n; n \in \mathbb{N}\} \) from which the desired WFLLN follows, namely,

\[
\bar{\lambda}_n \Rightarrow \bar{\lambda} \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{42}
\]

Towards proving the FCLT, we subtract (3) from (41) and then scale up both sides of the resulting equation by \( \sqrt{n} \) to obtain

\[
\hat{\lambda}_n(t) = \hat{\lambda}_n(0) - \kappa \int_0^t \hat{\lambda}_n(u)du + \sigma \int_0^t \sqrt{\hat{\lambda}_n(u)}dB(u). \tag{43}
\]

The desired FCLT for the sequence \( \{\hat{\lambda}_n; n \in \mathbb{N}\} \) follows immediately by applying the continuous mapping theorem (see Theorem 4.1 in [30]) together with the established WFLLN result in (42). \( \square \)

**Proof of Lemma 2.** The proof of this lemma is standard. First use (21) to write

\[
\hat{A}_n(t) = \hat{A}_{n,1}(t) + \hat{A}_{n,2}(t)
\]

where we have defined

\[
\hat{A}_{n,1}(t) \equiv n^{-1/2} \left( A_n(t) - \int_0^t \lambda_n(u)du \right) \quad \text{and} \quad \hat{A}_{n,2}(t) \equiv \int_0^t \hat{\lambda}_n(u)du. \tag{44}
\]

The first term is a square integrable martingale with quadratic variation \( \int_0^t \bar{\lambda}(u)du \) converging to \( \int_0^t \bar{\lambda}(u)du \) as \( n \to \infty \). Appealing to the martingale FLCT, we obtain

\[
\hat{A}_{n,2}(t) \Rightarrow B_{\lambda} \left( \int_0^t \bar{\lambda}(u)du \right) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.
\]

For the second term, we obtain the convergence \( \hat{A}_{n,2}(t) \Rightarrow K(t) \) by applying the continuous mapping theorem with (18). Combining the above gives the desired result. \( \square \)

**Proof of Theorem 1.** The proof follows the standard martingale method to establish the FCLT for many-server queueing systems; see, e.g. [11, 31]. As the proof is fairly standard (see, e.g. [11, 31]), we choose to present the key steps only.

First, multiple both sides of (15) by \( n \) and subtract it from (6) to get

\[
X_n(t) - nm(t) = X_n(0) - nm(0) + A_n(t) - n\Lambda(t)
\]

\[
- \left( D_n(t) - \mu \int_0^t B_n(u)du \right) - \mu \int_0^t (B_n(u) - nm(u))du
\]

\[
- \left( R_n(t) - \theta \int_0^t Q_n(u)du \right) - \theta \int_0^t Q_n(u)du.
\]
Dividing both sides by $\sqrt{n}$ and in view of the relations

$$B_n(\cdot) = X_n(\cdot) \wedge s_n(\cdot) \quad \text{and} \quad Q_n(\cdot) = [X_n(\cdot) - s_n(\cdot)]^+,$$

we derive the stochastic integral equation satisfied by the diffusion-scaled headcount process $\hat{X}_n$

$$\dot{X}_n(t) = \hat{X}_n(0) - c[0,t] - \mu \int_0^t c(u)du - \int_0^t \Psi(\hat{X}_n(u))du + \hat{A}_n(t) - \hat{D}_n(t) - \hat{R}_n(t),$$

where we recall $\Psi \equiv \mu x^- + \theta x^+$, $\hat{A}_n(t)$ is defined by (21), and

$$\hat{D}_n(t) \equiv n^{-1/2} \left( D_n(t) - \mu \int_0^t B_n(u)du \right),$$

$$\hat{R}_n(t) \equiv n^{-1/2} \left( R_n(t) - \theta \int_0^t Q_n(u)du \right)$$

are two square-integrable martingales; see [30]. An application of the Gronwall’s inequality to (45) gives the stochastic boundedness of $\{\hat{X}_n; n \in \mathbb{N}\}$. As a consequence, the quadratic variation of $\hat{D}_n$ and $\hat{R}_n$ converges to the corresponding limit, which in turn implies

$$\hat{D}_n(t) \Rightarrow \hat{D}(t) \equiv \int_0^t \sqrt{\mu m(u)}d\mathcal{B}_\mu(u) \quad \text{in} \quad \mathcal{D} \quad \text{and} \quad \hat{R}_n(t) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad (46)$$

for $\mathcal{B}_\mu$ being a standard Brownian motion and 0 being the zero function. Appealing to Theorem 4.1 in [30] with (22) and (46) yields the desired FCLT result for $\{\hat{X}_n; n \in \mathbb{N}\}$. The assertion $\hat{Q}_n \Rightarrow \hat{Q}$ is fairly straightforward and follows from the simple relation between $\dot{X}_n$ and $\hat{Q}_n$. Finally, the FCLT for $\{\hat{V}_n; n \in \mathbb{N}\}$ follows by applying the two-parameter version of Puhalskii’s invariance principle for first passage time; see Theorem 2.9 in [36]. \hfill \Box

**Proof of Theorem 2.** We start by observing the relation

$$V_n(t - H_n(t)) = H_n(t) + O(n^{-1} \log n). \quad (47)$$

The above relation holds because the PWT at the time of arrival of the HoL customer is the HoL customer’s elapsed waiting time (i.e., the HWT) at time $t$ plus the additional time until one of busy servers becomes idle.

Define the diffusion- and fluid-scaled empirical processes

$$\hat{U}_n(t, x) \equiv \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} \mathbb{I}_{\{X_k \leq x\}} \quad \text{for} \quad t \geq 0, \quad 0 \leq x \leq 1,$$

$$\hat{U}_n(t, x) \equiv \sqrt{n} \left( \bar{U}_n(t, x) - \mathbb{E} \left[ \bar{U}_n(t, x) \right] \right) = \frac{1}{\sqrt{n}} \left( \mathbb{I}_{\{X_k \leq x\}} - x \right),$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0, 1]$. Krichagina and Puhalskii [24] have shown that $\hat{U}_n \Rightarrow \bar{U}$ in $\mathcal{D}_\mathbb{D}$ as $n \to \infty$, where $\bar{U}$ is the standard Kiefer
process. Paralleling (3.3) - (3.6) in Aras, Chen and Liu [3], the enter-service process $E_n(t)$ in (7) can be decomposed into the sum of three terms:

$$E_n(t) = E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t),$$  \hspace{1cm} (49)

where

$$E_{n,1}(t) \equiv \sqrt{n} \int_0^{t-H_n(t)} F^c(V_n(u))d\hat{A}_n(u), \quad t \geq 0,$$  \hspace{1cm} (50)

$$E_{n,2}(t) \equiv \sqrt{n} \int_0^{t-H_n(t)} \int_0^1 \mathbb{1}_{y>F^c(V_n(u))} d\hat{U}_n(\hat{A}_n(u), y) \quad t \geq 0,$$  \hspace{1cm} (51)

$$E_{n,3}(t) \equiv n \int_0^{t-H_n(t)} F^c(V_n(u))\tilde{\lambda}(u)du \quad t \geq 0,$$  \hspace{1cm} (52)

for $\hat{A}_n, \hat{A}_n$ given in (21) and $\hat{U}_n$ specified by (48).

Next, we define the diffusion-scaled enter-service process

$$\hat{E}_n(t) \equiv n^{-1/2} \left( E_n(t) - n \int_0^{t-w} F^c(w)\tilde{\lambda}(u)du \right).$$  \hspace{1cm} (53)

Following the decomposition given in (49) - (52), we can write

$$\hat{E}_n(t) = \hat{E}_{n,1}(t) + \hat{E}_{n,2}(t) + \hat{E}_{n,3}(t),$$  \hspace{1cm} (54)

where

$$\hat{E}_{n,1}(t) \equiv n^{-1/2} E_{n,1}(t) = \int_0^{t-H_n(t)} F^c(V_n(u))d\hat{A}_n(u),$$  \hspace{1cm} (55)

$$\hat{E}_{n,2}(t) \equiv n^{-1/2} E_{n,2}(t) = \int_0^{t-H_n(t)} \int_0^1 \mathbb{1}_{y>F^c(V_n(u))} d\hat{U}_n(\hat{A}_n(u), y),$$  \hspace{1cm} (56)

$$\hat{E}_{n,3}(t) \equiv n^{-1/2} \left( E_{n,3}(t) - n \int_0^{t-w} F^c(w)\tilde{\lambda}(u)du \right).$$  \hspace{1cm} (57)

For (57) we further deduce

$$\hat{E}_{n,3}(t) = \sqrt{n} \left( \int_0^{t-H_n(t)} F^c(V_n(u))\tilde{\lambda}(u)du - \int_0^{t-w} F^c(w)\tilde{\lambda}(u)du \right)$$

$$= \sqrt{n} \int_w^t F^c(H_n(u))\tilde{\lambda}(u-H_n(u))du - \sqrt{n} \int_w^t F^c(w)\tilde{\lambda}(u-w)du$$

$$- \int_w^t F^c(H_n(u))\tilde{\lambda}(u-H_n(u))d\hat{H}_n(u) + O(n^{-1/2} \log n)$$  \hspace{1cm} (58)

$$= - \int_w^t \{ f(\eta_n(u))\tilde{\lambda}(u-\eta_n(u)) + F^c(\eta_n(u))\tilde{\lambda}'(u-\eta_n(u)) \} \hat{H}_n(u)du$$

$$- \int_w^t F^c(H_n(u))\tilde{\lambda}(u-H_n(u))d\hat{H}_n(u) + O(n^{-1/2} \log n),$$
where the second equality follows by a change of variables \( t \to t - H_n(t) \) plus the use of relation (47), while the third equality follows by applying the mean-value theorem with \( \eta_n(t) \) satisfying
\[
H_n(t) \land w \leq \eta_n(t) \leq H_n(t) \lor w. \tag{59}
\]
Note that the law of conservation of flow implies
\[
E_n(t) = B_n(t) + D_n(t). \tag{60}
\]
Moreover, from (10) it follows
\[
\varepsilon(t) = \int_w^t (\dot{m}(u) + \mu m(u)) \, du = m(t) + \mu \int_w^t m(u) \, du. \tag{61}
\]
Scaling both sides of (61) by \( n \) and subtracting it from (60) yields
\[
E_n(t) - n \varepsilon(t) = B_n(t) - nm(t) + D_n(t) - \mu \int_w^t nm(u) \, du.
\]
Then divide both sides by \( \sqrt{n} \) and make use of the relation \( B_n(t) = s_n(t) \) (which holds with arbitrarily high probability for \( n \) sufficiently large) to obtain
\[
\hat{E}_n(t) = c[w, t] + \mu \int_w^t c(u) \, du + \hat{D}_n(t), \tag{62}
\]
where we defined centered and normalized version of the departure process
\[
\hat{D}_n(t) \equiv n^{-1/2} \left( D_n(t) - \mu \int_w^t B_n(u) \, du \right). \tag{63}
\]
Upon plugging (55) - (56) and (58) into (62), we establish a stochastic integral equation for the diffusion-scaled process \( \hat{H}_n \). We can then apply the Gronwall’s inequality to resulting equation with the stochastic boundedness of \( \hat{E}_{n,1}, \hat{E}_{n,2} \) and \( \hat{D}_n \) plus the assumed properties of \( \tilde{\lambda} \) and \( f \) to conclude the stochastic boundedness of the sequence \( \{\hat{H}_n; n \in \mathbb{N}\} \). Here we choose not to go through all the details because we regard this step as standard; see e.g., (5.14) in [3] and the inequality that ensued. Then the FWLLN for the HWT and PWT processes follows as a directly consequence of the established stochastic boundedness of the sequence \( \{\hat{H}_n; n \in \mathbb{N}\} \), namely,
\[
(H_n, V_n) \Rightarrow (w_e, w_e) \in \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \tag{64}
\]
where the joint convergence holds by applying the converging-together lemma with (47). Having established the FWLLN for \( (H_n, V_n) \), we can invoke the continuous mapping theorem with (22) and (64) to derive
\[
\hat{E}_{n,1}(t) \Rightarrow \hat{E}_1(t) \equiv F^c(w) \int_0^{t-w} \sqrt{\lambda(u)} \, dB_\lambda(u) + F^c(w) \int_0^{t-w} \lambda(u) \, du. \tag{65}
\]
To show the convergence of the sequence \( \{ \hat{E}_{n,2}; n \in \mathbb{N} \} \), we invoke Theorem 7.1.4 in [10] to obtain
\[
\hat{E}_{n,2}(t) \Rightarrow \hat{E}_2(t) \equiv \sqrt{F^c(w)F(w)} \int_0^{t-w} \sqrt{\lambda(u)}dB_\theta(u),
\]
where \( B_\theta \) denotes a standard Brownian motion, independent of \( B_\lambda \) and \( \hat{\lambda} \). The detailed proof follows closely the steps in A.7.2 of [3] and is hence omitted. Towards obtaining a clean expression for \( \hat{E}_{n,3} \), we seek to simplify the right-hand side of (58). In view of (59) and (64), we have
\[
\eta_n(t) = H_n(t) + O(n^{-1/2} \log n) = w + O(n^{-1/2} \log n).
\]
Applying integration by parts, we deduce
\[
\int_w^t F^c(\eta_n(u))\lambda'(u - \eta_n(u))\dot{H}_n(u)du + \int_w^t F^c(H_n(u))\dot{\lambda}(u - H_n(u))d\dot{H}_n(u)
\]
\[
= F^c(\eta_n(t))\lambda(t - \eta_n(t))\dot{H}_n(t) - \int_w^t \lambda(u - \eta_n(u))\dot{H}_n(u)dF^c(\eta_n(u))
\]
\[- \int_w^t \left\{ F^c(\eta_n(u))\dot{\lambda}(u - \eta_n(u)) - F^c(H_n(u))\dot{\lambda}(u - H_n(u)) \right\} d\dot{H}_n(u)
\]
\[
= F^c(w)\lambda(t - w)\dot{H}_n(t) + O(n^{-1/2} \log n),
\]
where the last equality is due to (67). Combining (58) and (68) yields
\[
\hat{E}_{n,3}(t) = - \int_w^t f(w)\lambda(u - w)\dot{H}_n(u)du - F^c(w)\lambda(t - w)\dot{H}_n(t) + O(n^{-1/2} \log n).
\]
Finally, by a standard random-time-change argument, we get
\[
\dot{D}_n(t) \Rightarrow B_\mu \left( \mu \int_w^t m(u)du \right) \text{ in } \mathcal{D} \text{ as } n \to \infty,
\]
where \( B_\mu \) denotes a standard Brownian motion independent of anything else. The FCLT for the HWT process then follows by first substituting (65) - (66) and (69) - (70) into (62) and then applying Theorem 4.1 of [30].

Having established the FCLT for the sequence \( \{ \hat{H}_n; n \in \mathbb{N} \} \), the convergence of \( \{ \hat{V}_n; n \in \mathbb{N} \} \) follows by applying convergence-together lemma with (47). To establish the claimed FCLT for the queue-length process, we break the right-hand side of (8) into three pieces to get
\[
Q_n(t) = Q_{n,1}(t) + Q_{n,2}(t) + Q_{n,3}(t),
\]
where
\[
Q_{n,1}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^t F^c(t-u)d\hat{A}_n(u),
\]
\[
Q_{n,2}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^t \int_0^1 1_{\{x>F(t-u)\}}d\hat{U}_n(\hat{A}_n(u), x),
\]
\[
Q_{n,3}(t) \equiv n \int_{t-H_n(t)}^t F^c(t-u)\hat{\lambda}(u)du.
\]
Similar to what we did in (54) for the centered and normalized HWT process, we can decompose the diffusion-scaled queue-length process into three components

\[ \hat{Q}_n(t) = \hat{Q}_{n,1}(t) + \hat{Q}_{n,2}(t) + \hat{Q}_{n,3}(t), \]

where

\[
\begin{align*}
\hat{Q}_{n,1}(t) &\equiv \int_{t-H_n(t)}^t F^c(t-u) d\hat{A}_n(u) \Rightarrow \int_{t-w}^t F^c(t-u) d\hat{A}(u), \\
\hat{Q}_{n,2}(t) &\equiv \int_{t-H_n(t)}^t \int_{0}^{1} \mathbb{1}_{\{x>F(t-w)\}} d\hat{U}_n(A_n(u), x) \Rightarrow \int_{t-w}^t \sqrt{F^c(t-u) F(t-u) \bar{\lambda}(u)} dB\theta(u), \\
\hat{Q}_{n,3}(t) &\equiv \sqrt{n} \int_{t-H_n(t)}^{t-w} F^c(t-u) \bar{\lambda}(u) du \Rightarrow F^c(w) \bar{\lambda}(t-w) \tilde{H}(t).
\end{align*}
\]  

Here the proof for the convergence in (71) and (72) is very similar to the proof for (65) and (66), and the proof for (73) is fairly straightforward. \qed

A Additional Proofs

Lemma 3  If \( \bar{\lambda}_n \equiv \lambda_n/n \). Then for any finite \( T > 0 \) and \( k \in \mathbb{N} \), we have

\[ \sup_n \sup_{0 \leq t \leq T} \mathbb{E} \left[ \bar{\lambda}^k_n(t) \right] < \infty. \]

Proof of Lemma 3. Recall that

\[ d\bar{\lambda}_n(t) = \kappa(\alpha(t) - \bar{\lambda}_n(t)) dt + \sigma_n \sqrt{\bar{\lambda}_n(t)} dB(t), \]

where \( \sigma_n \equiv \sigma/n \). Applying Itô’s formula to the smooth function \( f(x) = x^k \), we obtain

\[ \bar{\lambda}^k_n(t) = \bar{\lambda}^k_n(0) + k\kappa \int_0^t \alpha(u) \bar{\lambda}^{k-1}_n(u) du - k\kappa \int_0^t \bar{\lambda}^k_n(u) du + k\sigma_n \int_0^t \bar{\lambda}^{k-1/2}_n(u) dB(u). \]  

An application of Young’s inequality yields

\[ \bar{\lambda}^{k-1}_n(u) \leq (k - 1) \bar{\lambda}^k_n(u)/k + 1/k. \]

Substituting the above into (74) gives

\[ \bar{\lambda}^k_n(t) \leq \bar{\lambda}^k_n(0) + \kappa \int_0^t [(k - 1) \alpha(u) - k] \bar{\lambda}^k_n(u) du + \kappa \int_0^t \alpha(u) du + k\sigma_n \int_0^t \bar{\lambda}^{k-1/2}_n(u) dB(u). \]

Taking expectation on both sides, we get

\[ \mathbb{E} \left[ \bar{\lambda}^k_n(t) \right] \leq \mathbb{E} \left[ \bar{\lambda}^k_n(0) \right] + \kappa \int_0^t [(k - 1) \alpha(u) - k] \mathbb{E} \left[ \bar{\lambda}^k_n(u) \right] du + \kappa \int_0^t \alpha(u) du. \]

An application of the Gronwall’s inequality allows us to conclude

\[ \mathbb{E} \left[ \bar{\lambda}^k_n(t) \right] \leq C(k, T)e^{C(k, T)T}. \]

The result immediately follows due to fact that the bound on the right hand side is independent of both \( t \) and \( n \). \qed
Proof of Corollary 1. That \( \hat{\Lambda}_n \to \hat{\Lambda} \) follows directly from Lemma 2. It remains to derive the covariance formula. For that purpose, note that

\[
\text{Cov}[K(s),K(t)] = E \left[ \int_0^s \hat{\lambda}(u) du \int_0^t \hat{\lambda}(v) dv \right] = \int_0^s \int_0^t E \left[ \hat{\lambda}(u) \hat{\lambda}(v) \right] du dv
\]

\[
= \int_0^s \left( \int_0^u E \left[ \hat{\lambda}(u) \hat{\lambda}(v) \right] dv + \int_u^t E \left[ \hat{\lambda}(u) \hat{\lambda}(v) \right] dv \right) du
\]

\[
= \frac{\sigma^2}{2\kappa} \int_0^s \left( \int_0^u \left( e^{-\kappa(u-v)} - e^{-\kappa(u+v)} \right) dv + \int_u^t \left( e^{-\kappa(v-u)} - e^{-\kappa(u+v)} \right) dv \right) du
\]

\[
= \frac{\sigma^2}{\kappa^3} (\kappa s - 1 + e^{-\kappa s} + e^{-\kappa t}) - \frac{\sigma^2}{2\kappa^3} \left( e^{-\kappa(t-s)} + e^{-\kappa(t+s)} \right)
\]

(75)

On the other hand, we have

\[
\text{Cov}[B_\lambda(s),B_\lambda(t)] = s.
\]

(76)

The desired conclusion then follows by combining (75) and (76) plus the independence between the process \( K \) and the Brownian motion \( B_\lambda \). \( \square \)

Proof of Proposition 1. Most of the proof follows that of Theorem 1. The exception is the function \( c(t) \) in (45) which will be replaced by the random process

\[
\hat{c}(t) \equiv c + e^{-\kappa t} \int_0^t e^{-\mu(t-u)} \hat{\lambda}_n(u) du,
\]

so that the diffusion-scaled headcount process \( \hat{X}_n \) satisfies

\[
\hat{X}_n(t) = \hat{X}_n(0) - \hat{c}[0,t] - \mu \int_0^t \hat{c}(u) du - \int_0^t \Psi(\hat{X}_n(u)) du + \hat{\Lambda}_n(t) - \hat{D}_n(t) - \hat{R}_n(t).
\]

It follows that limiting headcount process \( \hat{X} \) satisfies

\[
\hat{X}(t) = \hat{X}(0) - \mu \hat{c}t - e^{-\kappa t} \int_0^t \hat{\lambda}(u) du - \int_0^t \Psi(\hat{X}(u)) du + K(t) + \sqrt{2}B_2(t).
\]

The proof ends by noting that

\[
K(t) = \int_0^t \left( e^{-\kappa \tau} \hat{\lambda}(u - \tau) + \sigma \int_{u-\tau}^u e^{-\kappa(u-s)} B(s) \right) du.
\]

\( \square \)

References