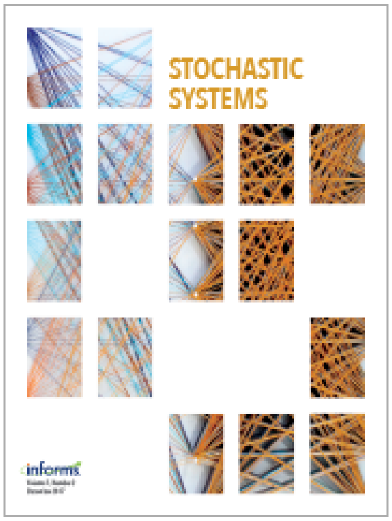

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Exact Asymptotics for a Multitimescale Model with Applications in Modeling Overdispersed Customer Streams

Mariska Heemskerk, Michel Mandjes

informs
http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 9, No. 3, September 2019, pp. 208–230
ISSN 1946-5238 (online)

Exact Asymptotics for a Multitimescale Model with Applications in Modeling Overdispersed Customer Streams

Mariska Heemskerk,[a] Michel Mandjes[a]

[a] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, 1098 XG Amsterdam, Netherlands

Contact: j.m.a.heemskerk@uva.nl, http://orcid.org/0000-0002-9784-3615 (MH);
m.r.h.mandjes@uva.nl, https://orcid.org/0000-0001-6783-4833 (MM)

Received: January 9, 2018
Revised: October 31, 2018; February 19, 2019
Accepted: February 21, 2019
Published Online in Articles in Advance: August 7, 2019

MSC2010 Subject Classification: Primary: 60F10, 60G51, 60K37; Secondary: 60K25, 68M20, 90B22

https://doi.org/10.1287/stsy.2019.0032

Abstract. In this paper, we study the asymptotics of the (scaled) exceedance probability of a multitimescale doubly-stochastic Lévy process. Two timescale regimes are distinguished: a fast regime in which one of the timescales is superlinear and a slow regime in which it is sublinear. We provide the exact asymptotics of the exceedance probability for both regimes, relying on change-of-measure arguments in combination with Edgeworth-type estimates. The asymptotics have an unconventional form: the exponent contains the commonly observed linear term but may also contain sublinear terms (the number of which depends on the precise form of the timescales chosen). To showcase the power of our results, we include two examples covering the cases where the multitimescale Lévy process is lattice and nonlattice. Finally, we present numerical experiments that demonstrate the importance of taking into account the doubly stochastic nature in a practical application related to customer streams in service systems. They show that the asymptotic results obtained yield highly accurate approximations also in scenarios in which there is no pronounced timescale separation.

History: Former designation of this paper was SSy-2018-005.R2.

Funding: The research of M. Heemskerk and M. Mandjes is partly funded by NWO Gravitation project NETWORKS [Grant 024.002.003].

Keywords: Lévy processes • multitimescale model • exact asymptotics • Edgeworth expansions • overdispersion • service systems

1. Introduction, Preliminaries, Notation, and Literature

Consider a scalar Lévy process $A(\cdot)$ and (independently of $A(\cdot)$) an increasing scalar Lévy process $B(\cdot)$. These are uniquely characterized by their *characteristic exponents*, which are defined as

$$\alpha(\vartheta) := \log \mathbb{E}e^{\vartheta A(1)}, \beta(\vartheta) := \log \mathbb{E}e^{\vartheta B(1)},$$

and are in fact the logarithmic moment-generating functions of $A(1)$ and $B(1)$. Targeting large-deviation asymptotics, we impose the assumption that both characteristic exponents are finite in an open neighborhood of the origin, implying that all moments of $A(t)$ and $B(t)$ exist. Let φ_n and ψ_n be nonnegative sequences such that $\varphi_n\psi_n = n$ and $\varphi_n \to \infty$ as $n \to \infty$. Our objective is to find the exact asymptotics of

$$\xi_n(u) := \mathbb{P}\big(A(\psi_n B(\varphi_n)) \geqslant un\big);$$

that is, we wish to find a sequence f_n such that $\xi_n(u)/f_n \to 1$ as $n \to \infty$. We assume that $u > ab$, with $a := \mathbb{E}\,A(1) > 0$ and $b := \mathbb{E}\,B(1) > 0$, such that the event under consideration becomes increasingly rare as $n \to \infty$. One special case has been studied in detail: the choices $\varphi_n = n$ and $\psi_n = 1$ reduce $\xi_n(u)$ to $\mathbb{P}(A(B(n)) \geqslant un)$, whose exact asymptotics follow from Bahadur and Ranga Rao (1960) (using that $A(B(\cdot))$ is a Lévy process). To the best of our knowledge, other cases have not been analyzed in the literature.

1.1. Effect of Multiple Timescales

To get some intuition for the behavior of $\xi_n(u)$, we write it as

$$\xi_n(u) = \mathbb{P}\left(A\left(n\frac{B(\varphi_n)}{\varphi_n}\right) \geqslant un\right).$$

We proceed by explaining that there are two timescale regimes. First, if φ_n is superlinear, it is anticipated that $B(\varphi_n)/\varphi_n$ is close to b such that $\xi_n(u)$ resembles $\mathbb{P}(A(bn) \geqslant un)$. We refer to this setting as the *fast regime* because the fluctuations of the process $B(\cdot)$ are so fast that it can be replaced by its mean value. Second, if, on the contrary, φ_n is sublinear (which we will refer to as the *slow regime* for obvious reasons), then one may expect that the event of interest roughly looks like $anB(\varphi_n)/\varphi_n \geqslant un$, and therefore, $\xi_n(u)$ essentially behaves as $\mathbb{P}(aB(\varphi_n) \geqslant u\varphi_n)$. The objective of this paper is to make these claims precise. Our main contribution is that we succeed in identifying the exact asymptotics of $\xi_n(u)$ as $n \to \infty$. These turn out to have a nonstandard form, in the sense that the exponent, in addition to the linear term that also appears in classical asymptotics (Dembo and Zeitouni 1998), may also contain sublinear terms.

To further investigate these two timescale regimes, it is instructive to calculate the variance of $C_n := A(\psi_n B(\varphi_n))$. To this end, we first express the log moment-generating function (l-mgf) $\gamma_n(\cdot)$ of C_n in terms of $\alpha(\cdot)$ and $\beta(\cdot)$. It requires a direct computation to verify that

$$\gamma_n(\vartheta) = \varphi_n \beta\big(\alpha(\vartheta)\,\psi_n\big).$$

Then it is direct that $\mathbb{E}\,C_n = n\alpha'(0)\beta'(0) = nab$, and

$$\mathbb{V}\mathrm{ar}\,C_n = \gamma_n''(0) = n\psi_n(\alpha'(0))^2\beta''(0) + n\alpha''(0)\beta'(0) = n\psi_n\sigma_-^2 + n\sigma_+^2, \tag{1}$$

with $\sigma_-^2 := a^2\beta''(0)$ and $\sigma_+^2 := \alpha''(0)b$. In this decomposition of the variance, the aforementioned regime dichotomy is nicely reflected, as can be seen as follows. If φ_n is superlinear (and hence ψ_n vanishes), then the first term on the right-hand side of Equation (1) is small relative to the second term, and $\mathbb{V}\mathrm{ar}\,C_n$ essentially behaves as $n\sigma_+^2$ (which is also the variance of $A(bn)$). By contrast, if φ_n grows sublinearly, then so does ψ_n, so in this case the first term of Equation (1) will dominate. As a consequence, $\mathbb{V}\mathrm{ar}\,C_n$ behaves as $n\psi_n\sigma_-^2$ (which equals the variance of $anB(\varphi_n)/\varphi_n$).

The given intuition can be translated in terms of a central limit theorem by following a classical approach. In the fast regime (in which φ_n is superlinear), direct computations reveal that the characteristic function of

$$D_n := \frac{C_n - (ab)n}{\sqrt{n}\sigma_+}$$

converges to that of a standard Normal random variable. Hence, as a direct application of Lévy's convergence theorem, D_n converges in distribution to a standard Normal random variable. Likewise, in the slow regime (in which φ_n is sublinear),

$$E_n := \frac{C_n - (ab)n}{\sqrt{n\psi_n}\sigma_-}$$

converges in distribution to a standard Normal random variable. We conclude that the dichotomy described earlier manifests in a central limit context.

1.2. Large Deviations, Contributions

In this paper, we assess to what extent this dichotomy carries over to a large-deviations setting. Based on the earlier reasoning, it is tempting to believe the following conjecture: in the fast regime, $\xi_n(u)/\mathbb{P}(A(bn) \geqslant un) \to 1$ as $n \to \infty$, and in slow regime, $\xi_n(u)/\mathbb{P}(aB(\varphi_n) \geqslant u\varphi_n) \to 1$ as $n \to \infty$. Our study, however, shows that this conjecture does not hold in general. In more detail, denoting by $k_n \sim \ell_n$ that $k_n/\ell_n \to 1$ as $n \to \infty$, the main result is that we find explicit sequences $\lambda_{+,n}$ and $\lambda_{-,n}$ such that in the fast regime, as $n \to \infty$,

$$\xi_n(u) \sim \mathbb{P}(A(bn) \geqslant un)\lambda_{+,n},$$

and in the slow regime, as $n \to \infty$,

$$\xi_n(u) \sim \mathbb{P}(aB(\varphi_n) \geqslant u\varphi_n)\lambda_{-,n}.$$

Here $\lambda_{+,n}$ and $\lambda_{-,n}$ do not necessarily equal 1, thus refuting the previous conjecture. Note that hereby the exact asymptotics of $\xi_n(u)$ are fully identified as sequences $\kappa_{+,n}$ and $\kappa_{-,n}$ such that $\mathbb{P}(A(bn) \geqslant un)\kappa_{+,n} \to 1$ and $\mathbb{P}(aB(\varphi_n) \geqslant u\varphi_n)\kappa_{-,n} \to 1$ are given in Bahadur and Ranga Rao (1960). (As an aside, we mention that in specific situations, $\lambda_{+,n}$ and $\lambda_{-,n}$ do equal 1 so that in those situations the conjecture applies; we return to this later.)

The resulting exact asymptotics of $\xi_n(u)$ can be written as the product of a polynomial and an exponential part. In the fast regime, the polynomial part is inversely proportional to $\sqrt{n}$, as was found previously in various related settings [such as the one studied in Bahadur and Ranga Rao (1960)]. The exponential part, however, has a rather unusual shape. Not only does the exponent contain a commonly observed term that is linear in n, but it also consists of finitely many sublinear terms (the number of which depends on the specific form of φ_n and ψ_n). In the slow regime, similar results apply, but with the role of n taken over by φ_n, meaning that the polynomial part is inversely proportional to $\sqrt{\varphi_n}$, and the exponent is the sum of a term that is linear in φ_n and finitely many terms that are $o(\varphi_n)$.

An immediate consequence of our result is that the conjecture we stated earlier is true if the timescales of both Lévy processes are sufficiently separated; then there are no sublinear terms in the exponent (where sublinear means $o(n)$ in the fast regime and $o(\varphi_n)$ in the slow regime). More specifically, the conjecture holds in the fast regime if $n\psi_n \to 0$ (i.e., then $\lambda_{+,n} = 1$) and in the slow regime if $\varphi_n/\psi_n \to 0$ (i.e., then $\lambda_{-,n} = 1$). It is further remarked that on a logarithmic scale, the conjecture always holds, albeit in the following (weaker) sense: we show that in the fast regime

$$\lim_{n\to\infty}\frac{1}{n}\log\xi_n(u) = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(A(bn)\geqslant un),$$

whereas in the slow regime

$$\lim_{n\to\infty}\frac{1}{\varphi_n}\log\xi_n(u) = \lim_{n\to\infty}\frac{1}{\varphi_n}\log\mathbb{P}(aB(\varphi_n)\geqslant u\varphi_n).$$

Example 1. The leading example, which we will visit several times in this paper, is that of $\varphi_n \sim n^f$ and $\psi_n \sim n^{1-f}$ for some $f > 0$. The fast regime corresponds to $f > 1$ and the slow regime to $f \in (0,1)$. The criteria just described yield that $\lambda_{+,n} = 1$ if $f > 2$ and $\lambda_{-,n} = 1$ if $f \in (0,\frac{1}{2})$.

Remark 1. We mentioned earlier that the case $\varphi_n = n$ can be dealt with in a straightforward way. For the sake of completeness, we include this argumentation here. In this case, $\gamma_n(\vartheta) = n\beta(\alpha(\vartheta))$, which effectively means that C_n can be written as the sum of n independent and identically distributed (i.i.d.) random variables, bringing us directly into the setting of Bahadur and Ranga Rao (1960). With $\vartheta^\star$ solving $\beta'(\alpha(\vartheta))\alpha'(\vartheta) = u$, one obtains

$$\xi_n(u) \sim \frac{1}{\vartheta^\star\sigma_0\sqrt{2\pi n}}\exp\big((\beta(\alpha(\vartheta^\star)) - \vartheta^\star u)n\big),$$

where $\sigma_0 := \beta''(\vartheta^\star)(\alpha'(\vartheta^\star))^2 + \beta'(\alpha(\vartheta^\star))\alpha''(\vartheta^\star)$.

We proceed with a few words on the approach followed. In the large-deviations analysis, a crucial role is played by the twisting factor, that is, the solution $\vartheta = \vartheta_n$ of the equation $un = \gamma_n'(\vartheta)$, or

$$u = \beta'\big(\alpha(\vartheta)\,\psi_n\big)\,\alpha'(\vartheta). \tag{2}$$

Because $u > ab$, it follows that ϑ_n is positive. In the fast regime, ϑ_n is close to the solution $\vartheta^\star$ of $u = b\alpha'(\vartheta^\star)$; in the slow regime, ϑ_n resembles $\tau^\star/\psi_n$, where $\tau^\star$ solves $u = \beta'(a\tau^\star)a$. Our proofs rely on a change of measure via an exponential twist that is based on the solution of Equation (2). By and large, the proof presented in Dembo and Zeitouni (1998) underlying the Bahadur–Ranga Rao result (Bahadur and Ranga Rao 1960) is followed: first, we capture the exponential part of $\xi_n(u)$ by applying a change of measure, after which the polynomial part of $\xi_n(u)$ is identified using delicate Berry–Esséen-based calculations (which are considerably more involved than the ones needed in the classical Bahadur–Ranga Rao case). In line with the proof in Dembo and Zeitouni (1998), the case where the underlying Lévy processes are lattice has to be dealt with separately.

Concluding the technical part of the paper, we want to apply the gained insights in a practical example, thus demonstrating their use in operations research. In this part, we (1) provide more background on the rationale behind the use of these kinds of models when analyzing service systems, (2) point out how the asymptotics identified in our paper can be translated into approximations (for an unscaled model), and (3) assess the accuracy of such approximations.

1.3. Literature and Motivation

Our work fits in the tradition of large-deviation asymptotics of sample-mean-related quantities in the light-tailed setting. Without attempting to provide an exhaustive overview, we give a number of key references.

In the classical framework, S_n is defined as the sum of i.i.d. random variables $X_1,\ldots,X_n$, where the X_i are assumed to have a finite moment-generating function in a neighborhood of the origin. The exceedance probability $\chi_n(u) := \mathbb{P}(S_n \geqslant un)$ is the object of study, with $u > \mathbb{E}X_1$. Cramér (1938) characterized a function $I(u)$ such that $\frac{1}{n}\log\chi_n(u) \to -I(u)$ as $n\to\infty$. This type of result is commonly referred to as *logarithmic asymptotics*.

The proof technique used relied on change-of-measure argumentation that has been applied extensively since then.

Cramér's seminal work was extended in several directions. In Chernoff (1952), a uniform upper bound on $\chi_n(u)$, generally known as the *Chernoff bound*, was derived. We refer to Bahadur and Zabell (1979) for a generalization of Cramér's logarithmic asymptotics to infinite-dimensional topological vector spaces. Cramér's theorem was also extended to sums of dependent random variables (Plachky and Steinebach 1975) and vectors (Gärtner 1977, Ellis 1984).

Whereas the earlier results focus on logarithmic asymptotics, another strand of research addresses exact asymptotics. In this respect, we mention the pioneering paper by Bahadur and Ranga Rao (1960) that shows that, under the conditions of Cramér's result, $\chi_n(u)\,\mathrm{e}^{nI(u)}\sqrt{n}$ converges to a positive constant as $n\to\infty$. In Blackwell and Hodges (1959), a similar result had already been derived for the case that the X_i are lattice. We refer to, for example, Chaganty and Sethuraman (1996) for exact asymptotics in the vector-valued case and to Höglund (1979) for a uniform framework covering both the central-limit regime and large deviations.

Our investigations were motivated by models recently suggested to incorporate overdispersion. The reason for developing such models was the observation that in various types of service systems (Whitt et al. 2007, Bassamboo et al. 2010), the customer arrival process is intrinsically more variable than the traditionally used Poisson process. An approach to overcome the lack of overdispersion was proposed in Heemskerk et al. (2017b): extra variability is produced by periodically resampling the Poisson arrival rate. As a consequence, when resampling every unit of time, the number of arrivals in $[0,m]$, denoted by $C(m)$ (for $m\in\mathbb{N}$), is Poisson distributed with (random) parameter $\sum_{i=1}^m \Lambda_i$, where the Λ_i are i.i.d. nonnegative random variables. Assuming for simplicity that the Λ_i are integer valued as well, this means that

$$C(m) \stackrel{\mathrm{d}}{=} \sum_{j=1}^{\sum_{i=1}^m \Lambda_i} X_j, \tag{3}$$

with X_j a sequence of i.i.d. Poisson random variables with parameter 1. We thus have introduced overdispersion, as desired. More specifically, we have

$$\mathbb{V}\mathrm{ar}\,C(m) = m\,\mathbb{E}\Lambda\,\mathbb{V}\mathrm{ar}X + m\,(\mathbb{E}X)^2\mathbb{V}\mathrm{ar}\Lambda \geqslant m\,\mathbb{E}\Lambda\,\mathbb{V}\mathrm{ar}X = m\,\mathbb{E}\Lambda = \mathbb{E}\,C(m),$$

making the process more variable than the Poisson process. Observe that the two-timescale random walk in Equation (3) can be seen as the discrete counterpart of the two-timescale Lévy processes considered in this paper. In Heemskerk et al. (2017b), this two-timescale random walk model is studied under certain scalings (comparable to the scalings with φ_n and ψ_n that are imposed in this paper). The results in Heemskerk et al. (2017b) include logarithmic asymptotics, which are for special cases refined to exact asymptotics in Heemskerk et al. (2017a).

1.4. Organization

The fast regime (in which φ_n grows superlinearly) is covered by Section 2, followed by the slow regime (in which φ_n grows sublinearly) in Section 3. Examples are presented in Section 4. Section 5 describes how our findings can be applied when modeling overdispersed customer streams. We conclude in Section 6 by putting our results into perspective and by identifying directions for follow-up research.

2. Fast Regime

In this section, we analyze the case that φ_n is superlinear, entailing that $\psi_n\to 0$ as $n\to\infty$. This corresponds to $f>1$ in the context of Example 1. The approach followed echoes the proof of, for example, Dembo and Zeitouni (1998, theorem 3.7.4) and comprises the following elements:

• We first identify in Section 2.1 the twisting factor ϑ_n (i.e., the solution of Equation (2)). It means that twisting C_n by ϑ_n leads to a random variable with mean un. More concretely, we define a measure $\mathbb{Q}_n$ through (in self-evident notation)

$$\mathbb{Q}_n(C_n\in \mathrm{d}x) = \mathbb{P}(C_n\in \mathrm{d}x)\frac{\exp(\vartheta_n x)}{\exp(\gamma_n(\vartheta_n))},$$

where it holds that (in self-evident notation)

$$\mathbb{E}_{\mathbb{Q}_n} C_n = \int_{-\infty}^{\infty} x\, \mathbb{Q}_n(C_n \in \mathrm{d}x) = un.$$

• Then we rewrite in Section 2.2 the probability $\xi_n(u)$ using the ϑ_n-twisted version of C_n. As in Asmussen (2003, chapter XIII), we can express $\xi_n(u)$ into a mean under $\mathbb{Q}_n$:

$$\xi_n(u) = \mathbb{P}(C_n \geqslant un) = \mathbb{E}_{\mathbb{Q}_n}(L(C_n) 1\{C_n \geqslant un\}), \tag{4}$$

for an appropriately chosen $L(\cdot)$ (which can be interpreted as a likelihood ratio).

The right-hand side of Equation (4) turns out to consist of an exponential part and a polynomial part (in n, that is). Section 2.3 analyzes the exponential part. As opposed to the standard Bahadur–Ranga Rao result (Dembo and Zeitouni 1998, theorem 3.7.4), the exponent potentially also contains sublinear terms.

The polynomial part of the right-hand side of Equation (4) is technically the most demanding element of the proof and relies on Berry–Esséen-based arguments (see Section 2.4 and Appendix A).

In Section 2.5, all elements are put together, and our result is stated.

The formal assumption we impose on ψ_n is the following.

Assumption 1. *The sequence* ψ_n *satisfies*

$$\limsup_{n\to\infty} \frac{\log \psi_n}{\log n} < 0.$$

This assumption means that there is an $\varepsilon > 0$ such that $\psi_n < n^{-\varepsilon}$ and hence $\varphi_n > n^{1+\varepsilon}$. We observe that φ_n is superlinear.

In the main theorem of this section, that is, Theorem 1, we assume that $A(\cdot)$ is nonlattice. To establish the result for the case where $A(\cdot)$ is lattice (covering, e.g., Poisson processes), a minor adaptation needs to be made. More specifically, the steps of Sections 2.1–2.3 apply to the lattice as well as nonlattice case, whereas in Section 2.4 in the lattice case slightly different bounds have to be used [fully analogously to the treatment of the lattice and nonlattice cases in the proof of Dembo and Zeitouni (1998, theorem 3.7.4)]. The result for the lattice case is discussed separately in Remark 2. We refer to Sections 4 and 5 for illustrative examples covering both the lattice and nonlattice cases.

2.1. Analysis of Twisting Factor

Because $\psi_n \to 0$ as $n \to \infty$, the twisting factor ϑ_n converges to the solution $\vartheta^\star$ of $b\alpha'(\vartheta) = u$, where $b = \mathbb{E}B(1) = \beta'(0)$. To establish the exact asymptotics of $\xi_n(u)$, it turns out that we need an expansion of ϑ_n. We start by arguing that ϑ_n has the form

$$\vartheta_n = \vartheta^\star + \sum_{k=1}^{\infty} v_k \psi_n^k. \tag{5}$$

and then we point out how to determine the coefficients v_k.

• The main idea is that one can (implicitly) define a function $\vartheta_+(x)$ as the solution of the equation

$$\beta'(\alpha(\vartheta)x)\alpha'(\vartheta) = u.$$

Such a solution is unique because the left-hand side of the previous display is the derivative of a convex function (hence increasing). Observe that ϑ_n (solving $\gamma_n'(\vartheta_n) = u$) equals $\vartheta_+(\psi_n)$, and recall that $\psi_n \to 0$ as $n \to \infty$.

Observe that $\vartheta_+(0) = \vartheta^\star$ (with $\vartheta^\star$, as defined previously, solving $b\alpha'(\vartheta^\star) = u$). Writing down the Taylor expansion (at $x = 0$) of the implicitly defined function $\vartheta_+(x)$, we obtain

$$\vartheta_+(x) = \sum_{k=0}^{\infty} v_k x^k,$$

with $v_0 = \vartheta^\star$. This means that the v_k can be determined in terms of the derivatives of $\alpha(\cdot)$ and $\beta(\cdot)$ (which are well defined). Next, we provide a constructive procedure that identifies all these coefficients. Upon combining the previous elements, we obtain the expansion of Equation (5)

$$\vartheta_n = \vartheta_+(\psi_n) = \sum_{k=0}^{\infty} v_k \psi_n^k.$$

• We present a constructive procedure that yields the coefficients v_k. To this end, first observe that by evaluating $\alpha(\cdot)$ as a Taylor series around $\vartheta^\star$,

$$\beta'\left(\sum_{\ell=0}^{\infty}\frac{\alpha^{(\ell)}(\vartheta^\star)}{\ell!}\left(\sum_{k=1}^{\infty} v_k\psi_n^k\right)^{\ell}\psi_n\right)\cdot\sum_{\ell=0}^{\infty}\frac{\alpha^{(\ell+1)}(\vartheta^\star)}{\ell!}\left(\sum_{k=1}^{\infty} v_k\psi_n^k\right)^{\ell} = u,$$

where, by evaluating $\beta(\cdot)$ as a Taylor series around 0,

$$\beta'(\vartheta) = \sum_{m=0}^{\infty}\frac{\beta^{(m+1)}(0)}{m!}\vartheta^m.$$

This equation effectively determines all v_k, which can be found by grouping together the appropriate terms (i.e., terms with the same power). We demonstrate this procedure for v_1. Up to terms that are $o(\psi_n)$,

$$\big(\beta'(0) + \beta''(0)\alpha(\vartheta^\star)\psi_n\big)\big(\alpha'(\vartheta^\star) + \alpha''(\vartheta^\star)v_1\psi_n\big) = u.$$

Using that $\beta'(0)\alpha'(\vartheta^\star) = b\alpha'(\vartheta^\star) = u$, we obtain

$$v_1 = -\frac{\alpha(\vartheta^\star)\,\alpha'(\vartheta^\star)}{\alpha''(\vartheta^\star)}\frac{\beta''(0)}{\beta'(0)}.$$

The v_i for $i \in \{2, 3\ldots\}$ can be computed in the exact same way, but this does not lead to clean expressions; for $k = 2$, we find

$$v_2 = -\frac{1}{2}\frac{\beta^{(1)}(0)\alpha^{(3)}(\theta^*)v_1^2 + \beta^{(2)}(0)(\alpha(\theta^*)\alpha^{(2)}(\theta^*) + (\alpha^{(1)}(\theta^*))^2)v_1}{\beta^{(1)}(0)\alpha^{(2)}(\theta^*)}.$$

2.2. Change of Measure

The next step is to rewrite the probability of interest, relying on the usual change-of-measure procedure. This effectively means that we let $\mathbb{Q}_n$ correspond to twisting the distribution of C_n by the solution ϑ_n of Equation (2). The l-mgf of C_n under $\mathbb{Q}_n$ thus can be expressed in terms of the l-mgf of C_n under the original measure:

$$\gamma_n^{\mathbb{Q}}(\vartheta) := \gamma_n(\vartheta + \vartheta_n) - \gamma_n(\vartheta_n).$$

Later on we need the mean and variance of C_n under $\mathbb{Q}_n$. From the definition of ϑ_n, it follows that $\mathbb{E}_{\mathbb{Q}_n} C_n = un$. Differentiating once more, we obtain

$$\mathbb{V}\mathrm{ar}_{\mathbb{Q}_n} C_n = n\psi_n\beta''\big(\alpha(\vartheta_n)\psi_n\big)(\alpha'(\vartheta_n))^2 + n\beta'\big(\alpha(\vartheta_n)\psi_n\big)\alpha''(\vartheta_n).$$

The next step is to use the change of measure to rewrite $\xi_n(u)$. By applying the definition of $\mathbb{Q}_n$, we find the identity

$$\xi_n(u) = \mathbb{E}_{\mathbb{Q}_n}\left(\mathrm{e}^{\gamma_n(\vartheta_n)-\vartheta_n C_n}\,1\{C_n \geqslant un\}\right),$$

realizing that $\mathrm{e}^{\gamma_n(\vartheta_n)-\vartheta_n C_n}$ plays the role of the likelihood ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}_n$. On the event $C_n \geqslant un$, C_n will typically be relatively close to un under the new measure $\mathbb{Q}_n$ (recall that $\mathbb{E}_{\mathbb{Q}_n} C_n = un$). To exploit this, we define, with $\sigma_+^{\mathbb{Q}} := \sqrt{b\alpha''(\vartheta^\star)}$,

$$\overline{D}_n := \frac{C_n - un}{\sqrt{n}\sigma_+^{\mathbb{Q}}},$$

which is a random variable that has, by construction, mean 0 under $\mathbb{Q}_n$. (As a result of the $\sqrt{n}$ in the denominator, it is anticipated that under $\mathbb{Q}_n$, $\overline{D}_n$ can be approximated by a standard Normal random variable. We come back to this idea later.) We thus obtain that for all n,

$$\xi_n(u) = \mathrm{e}^{\gamma_n(\vartheta_n)-\vartheta_n un}\,\Delta_n, \text{ with } \Delta_n := \mathbb{E}_{\mathbb{Q}_n}\left(\mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}\sqrt{n}\,\overline{D}_n}\,1\{\overline{D}_n \geqslant 0\}\right). \tag{6}$$

Consequently, we are left with analyzing the exponential factor $\delta_n := \exp(\gamma_n(\vartheta_n) - \vartheta_n un)$ and the expectation Δ_n as n grows large.

2.3. Analysis of δ_n as $n \to \infty$

We proceed by analyzing the exponent in the expansion of Equation (6), that is, $\gamma_n(\vartheta_n) - \vartheta_n un$. Define

$$m_+ := \sup\left\{k \in \mathbb{N} : \liminf_{n\to\infty} \varphi_n \psi_n^k > 0\right\};$$

note that $m_+ \geqslant 1$. Our claim is the following.

Lemma 1. *As $n \to \infty$, for constants $\overline{v}_k$,*

$$\gamma_n(\vartheta_n) - \vartheta_n un = \big(b\alpha(\vartheta^\star) - \vartheta^\star u\big)n + \sum_{k=2}^{m_+} \overline{v}_k \varphi_n \psi_n^k + o(1),$$

where the empty sum is defined as 0.

The validity of this claim (as well as the procedure to determine the coefficients $\overline{v}_k$) can be demonstrated as follows. Relying on Taylor expansions, we obtain

$$\gamma_n(\vartheta_n) - \vartheta_n un = \varphi_n \beta\left(\sum_{\ell=0}^{\infty} \frac{\alpha^{(\ell)}(\vartheta^\star)}{\ell!}\left(\sum_{k=1}^{\infty} v_k \psi_n^k\right)^{\ell} \psi_n\right) - \left(\vartheta^\star + \sum_{k=1}^{\infty} v_k \psi_n^k\right)un.$$

The claim for $m_+ = 1$ can be directly verified. Now consider the case $m_+ = 2$. Any higher value of m can be dealt with fully analogously. For $m_+ = 2$, the sequence $\varphi_n\,\psi_n^k$ converges to 0 when $k > 2$, whereas $\varphi_n\,\psi_n^2 = n\psi_n$ stays away from 0; think of, for example, $\psi_n = n^{-2/3}$. Hence we obtain that as $n \to \infty$,

$$\gamma_n(\vartheta_n) - \vartheta_n un = \big(b\alpha(\vartheta^\star) - \vartheta^\star u\big)n + v_1\left(\frac{\beta''(0)}{2}\alpha(\vartheta^\star) + b\alpha'(\vartheta^\star) - u\right)\varphi_n \psi_n^2 + o(1).$$

It is clear that for $m_+ = 3$, an additional term needs to be included. In general, by using this procedure, any value of m_+ can be dealt with.

2.4. Analysis of Δ_n as $n \to \infty$

We are left with analyzing Δ_n for n large. Our objective is to prove that $\sqrt{n}\Delta_n$ converges to a positive constant as $n \to \infty$. Mimicking the line of reasoning of Dembo and Zeitouni (1998, equation (3.7.7)), we apply integration by parts:

$$\begin{aligned}\Delta_n &= \int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}\sqrt{n}x}\mathbb{Q}_n(\overline{D}_n \in \mathrm{d}x) = \sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}\sqrt{n}x}\big(\mathbb{Q}_n(\overline{D}_n \leqslant x) - \mathbb{Q}_n(\overline{D}_n \leqslant 0)\big)\mathrm{d}x \\ &= \vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\big(\mathbb{Q}_n(\overline{D}_n \leqslant x/\sqrt{n}) - \mathbb{Q}_n(\overline{D}_n \leqslant 0)\big)\mathrm{d}x.\end{aligned} \tag{7}$$

Because we will intensively rely on this representation of Δ_n, an important role in our argumentation is played by the probability distribution

$$\mathbb{Q}_n(\overline{D}_n \leqslant x) = \mathbb{Q}_n\left(\frac{A(\psi_n B(\varphi_n)) - un}{\sqrt{n}\sigma_+^{\mathbb{Q}}} \leqslant x\right),$$

where it is noted that $\mathbb{E}_{\mathbb{Q}_n} A(\psi_n B(\varphi_n)) = un$. Estimates for this distribution can be established, essentially as variants of the classical Edgeworth expansion. Because such expansions follow by applying well-established techniques, we restrict ourselves to providing the main steps of the derivation in Appendix A. An excellent introduction to the Edgeworth expansion is provided in Hall (1992, chapter II); see also Barndorff-Nielsen and Cox (1989, chapter IV).

• In the case where $\lim_{n\to\infty}\psi_n\sqrt{n}=0$ (which corresponds to $f>\frac{3}{2}$ in the context of Example 1), we have, as pointed out in Appendix A, as $n\to\infty$,

$$\sqrt{n}\sup_x\left(\mathbb{Q}_n(\overline{D}_n\leqslant x)-\Phi(x)+\phi(x)H_2(x)\frac{\kappa_+}{\sqrt{n}}\right)\to 0, \kappa_+:=\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}; \tag{8}$$

cf. Equation (A.2), where $H_2(x)=x^2-1$, $\phi(\cdot)$ denotes the probability density function of a standard Normal random variable and $\Phi(\cdot)$ the corresponding cumulative distribution function. We are now in a position to prove that $\sqrt{n}\Delta_n$ converges to a constant. We provide the upper bound, but the lower bound follows fully analogously. By virtue of Equation (8), for any $\varepsilon>0$ and n sufficiently large,

$$\begin{aligned}\sqrt{n}\Delta_n\leqslant{}&\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\left(\Phi(\frac{x}{\sqrt{n}})-\Phi(0)\right)dx-\\&\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\left(\phi(\frac{x}{\sqrt{n}})H_2(\frac{x}{\sqrt{n}})-\phi(0)H_2(0)\right)\frac{\kappa_+}{\sqrt{n}}dx\;+\;\varepsilon\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}dx.\end{aligned}$$

Let us start by evaluating the first term on the right-hand side. It can be rewritten as

$$\begin{aligned}\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\int_0^{x/\sqrt{n}}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy\,dx&=\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\int_{y\sqrt{n}}^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}dx\,dy\\&=\sqrt{n}\int_0^\infty\frac{1}{\sqrt{2\pi}}e^{-y^2/2}e^{-\vartheta_n\sigma_+^{\mathbb{Q}}\sqrt{n}y}dy\\&=\sqrt{n}\exp\left(\frac{1}{2}(\vartheta_n\sigma_+^{\mathbb{Q}})^2n\right)\left(1-\Phi(\vartheta_n\sigma_+^{\mathbb{Q}}\sqrt{n})\right).\end{aligned}$$

Using the known limit $x(1-\Phi(x))/\phi(x)\to 1$ as $x\to\infty$ and $\vartheta_n\to\vartheta^\star$, we have that the expression in the previous display converges to

$$\left(\vartheta^\star\sigma_+^{\mathbb{Q}}\sqrt{2\pi}\right)^{-1}. \tag{9}$$

Using that $H_2(x)=x^2-1$, the second term can be written as the sum of

$$\begin{aligned}t_{+,n}^{(1)}&:=\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\left(\phi(0)-\phi\left(\frac{x}{\sqrt{n}}\right)\right)\frac{\kappa_+}{\sqrt{n}}dx,\\t_{+,n}^{(2)}&:=\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\phi\left(\frac{x}{\sqrt{n}}\right)\frac{x^2}{n}\frac{\kappa_+}{\sqrt{n}}dx.\end{aligned}$$

From (1) $\sqrt{n}\,(\phi(0)-\phi(x/\sqrt{n}))\to x\phi'(0)=0$, (2) $\vartheta_n\to\vartheta^\star$ as $n\to\infty$, and (3) the fact that $\phi'(\cdot)$ is bounded, applying the dominated convergence theorem yields that $t_{+,n}^{(1)}\to 0$. Owing to (1) $\phi(x/\sqrt{n})\leqslant 1/\sqrt{2\pi}$, (2) $\vartheta_n\to\vartheta^\star$ as $n\to\infty$, and (3)

$$\limsup_{n\to\infty}\int_0^\infty e^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}x^2\,dx=\limsup_{n\to\infty}\frac{2}{(\vartheta_n\sigma_+^{\mathbb{Q}})^3}<\infty,$$

we also have $t_{+,n}^{(2)}\to 0$.

The third term equals ε, which can be made arbitrarily small. As mentioned earlier, by the same token, it can be verified that the corresponding lower bound applies as well. We thus conclude that $\sqrt{n}\Delta_n$ converges to the constant in Equation (9).

We proceed with the case $\liminf_{n\to\infty}\psi_n\sqrt{n}>0$ (which simplifies to $f\in(1,\frac{3}{2}]$ in the context of Example 1). From Equation (A.3) in Appendix A, we have

$$\sqrt{n}\sup_x\left(\mathbb{Q}_n(\overline{D}_n\leqslant x)-\Phi(x)+\phi(x)\left(H_1(x)\sum_{k=1}^{m_+}c_k\psi_n^k+H_2(x)\frac{\kappa_+}{\sqrt{n}}\right)\right)\to 0. \tag{10}$$

By Equation (10), for any $\varepsilon > 0$ and n sufficiently large, using that $H_1(x) = x$, and hence $H_1(0) = 0$,

$$\begin{aligned}\sqrt{n}\Delta_n \leqslant{} & \sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\left(\Phi\left(\frac{x}{\sqrt{n}}\right) - \Phi(0)\right)\mathrm{d}x \\ & -\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\phi\left(\frac{x}{\sqrt{n}}\right)\frac{x}{\sqrt{n}}\left(\sum_{k=1}^{m_+} c_k\psi_n^k\right)\mathrm{d}x \\ & -\sqrt{n}\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\left(\phi\left(\frac{x}{\sqrt{n}}\right)H_2\left(\frac{x}{\sqrt{n}}\right) - \phi(0)H_2(0)\right)\frac{\kappa_+}{\sqrt{n}}\mathrm{d}x + \varepsilon\vartheta_n\sigma_+^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}\mathrm{d}x.\end{aligned}$$

The first, third, and fourth terms can be dealt with as in the case $\psi_n\sqrt{n} \to 0$. Owing to (1) $\phi(x) \leqslant 1/\sqrt{2\pi}$ for all x, (2) $\vartheta_n \to \vartheta^\star$ as $n \to \infty$, (3) $\psi_n^k \to 0$ as $n \to \infty$ for any $k \in \{1, \ldots, m_+\}$, and (4)

$$\limsup_{n\to\infty}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_+^{\mathbb{Q}}x}x\,\mathrm{d}x = \limsup_{n\to\infty}\frac{1}{(\vartheta_n\sigma_+^{\mathbb{Q}})^2} < \infty,$$

the second term vanishes. It follows that again $\sqrt{n}\Delta_n$ converges to the constant in Equation (9). We have thus established the following lemma.

Lemma 2. *As* $n \to \infty$,

$$\sqrt{n}\Delta_n \to \left(\vartheta^\star\sigma_+^{\mathbb{Q}}\sqrt{2\pi}\right)^{-1}.$$

2.5. Result

Upon combining Lemmas 1 and 2, as presented in the previous subsections, we have arrived at the following result.

Theorem 1. *As* $n \to \infty$, *under Assumption* 1, *for nonlattice* $A(\cdot)$,

$$\xi_n(u) \sim \frac{1}{\vartheta^\star\sigma_+^{\mathbb{Q}}\sqrt{2\pi n}}\exp\left(\left(b\alpha(\vartheta^\star) - \vartheta^\star u\right)n + \sum_{k=2}^{m_+}\overline{v}_k\varphi_n\psi_n^k\right).$$

An immediate consequence of this theorem is that $\xi_n(u)$ behaves as $\mathbb{P}(A(bn) \geqslant un)$ when $\varphi_n\psi_n^2 = n\psi_n \to 0$. This is, for instance, the case when $\psi_n = n^\zeta$ for $\zeta < -1$ or, in the setting of Example 1, $f > 2$. It reflects that the timescale of $B(\cdot)$ is so much faster than that of $A(\cdot)$ that it can be replaced by its mean. In addition, it implies that the rough (logarithmic) asymptotics are not affected by the choice of ψ_n (as long as Assumption 1 is fulfilled). These findings are summarized in the following corollary.

Corollary 1. *If* $\varphi_n\psi_n^2 = n\psi_n \to 0$ *as* $n \to \infty$, *then, under Assumption* 1, *for nonlattice* $A(\cdot)$,

$$\xi_n(u) \sim \mathbb{P}(A(bn) \geqslant un) \sim \frac{1}{\vartheta^\star\sigma_+^{\mathbb{Q}}\sqrt{2\pi n}}\exp\left(\left(b\alpha(\vartheta^\star) - \vartheta^\star u\right)n\right).$$

As $n \to \infty$, *under Assumption* 1, *for nonlattice* $A(\cdot)$,

$$\frac{1}{n}\log\xi_n(u) \to b\alpha(\vartheta^\star) - \vartheta^\star u.$$

Remark 2. So far we have assumed throughout that the Lévy processes $A(\cdot)$ are nonlattice, thus ruling out, for example, Poisson processes. With a minor adaptation, however, the lattice case can be dealt with as well. Let, for some x_0 and d and any $t \geqslant 0$, the random variable $d^{-1}(A(t) - x_0)$ be integer almost surely, and let d be the largest number with this property; for example, in the Poisson process case, $d = 1$. Then, following the proof in, for example, Dembo and Zeitouni (1998, theorem II.7.4.b), under Assumption 1, we find the following counterpart of Theorem 1:

$$\xi_n(u) \sim \frac{d}{1 - \mathrm{e}^{-\vartheta^\star d}}\frac{1}{\sigma_+^{\mathbb{Q}}\sqrt{2\pi n}}\exp\left(\left(b\alpha(\vartheta^\star) - \vartheta^\star u\right)n + \sum_{k=2}^{m_+}\overline{v}_k\varphi_n\psi_n^k\right),$$

as $n \to \infty$. This behavior is consistent with that of Theorem 1 when taking $d \downarrow 0$.

3. Slow Regime

In this section, we analyze the case that φ_n grows sublinearly, implying that $\psi_n \to \infty$ as $n \to \infty$ (also sublinearly). As before, we first characterize the solution ϑ_n of Equation (2), then we rewrite $\xi_n(u)$ using the ϑ_n-twisted version of C_n, and finally, we determine the corresponding exact asymptotics. The formal assumption we impose on ψ_n in this section is the following.

Assumption 2. *The sequence ψ_n satisfies*

$$0 < \liminf_{n\to\infty} \frac{\log \psi_n}{\log n} \leqslant \limsup_{n\to\infty} \frac{\log \psi_n}{\log n} < 1.$$

The first inequality of this assumption ensures that there is an $\varepsilon \in (0,1)$ such that $\psi_n > n^{\varepsilon}$, and hence $\varphi_n < n^{1-\varepsilon}$, so that φ_n is sublinear. In addition, the second inequality entails that ψ_n is sublinear, too. For the moment, we assume that $B(\cdot)$ is nonlattice; see Remark 3 for the corresponding result in the lattice case.

The procedure we follow to prove the exact asymptotics in the slow regime is in line with the one we developed for the fast regime: we perform a change of measure, take out the exponential factor δ_n, and analyze the remainder term Δ_n. Because this procedure echoes the one developed in the preceding section, we only include the main steps-.

The change of measure is again based on the solution ϑ_n that solves Equation (2). In this case, as we argue later, ϑ_n obeys the expansion

$$\vartheta_n = \sum_{k=1}^{\infty} w_k \psi_n^{-k}, \tag{11}$$

and the coefficients w_k can be recursively determined. The reasoning is analogous to the argumentation followed in the fast regime.

• To show the validity of the expansion in Equation (11), we write $\vartheta_-(x)$ as the solution of the equation

$$\beta'\left(\frac{\alpha(\vartheta)}{x}\right)\alpha'(\vartheta) = u,$$

so $\vartheta_n = \vartheta_-(1/\psi_n)$. Recall that in this regime, $\psi_n \to \infty$ as $n \to \infty$.

Observe that $\vartheta_-(x)/x \to \tau^\star$ as $x \to 0$, where $\tau^\star$ solves $a\beta'(a\tau^\star) = u$. We write $\vartheta_-(\cdot)$ as a Taylor expansion around $x = 0$:

$$\vartheta_-(x) = \sum_{k=1}^{\infty} w_k x^k,$$

with $w_1 = \tau^\star$. As in the fast regime, the coefficients allow expressions in terms of the derivatives of $\alpha(\cdot)$ and $\beta(\cdot)$. Combining these, we thus obtain the expansion in Equation (11):

$$\vartheta_n = \vartheta_-(1/\psi_n) = \sum_{k=1}^{\infty} w_k \psi_n^{-k}.$$

• The procedure to identify the coefficients w_k in this slow regime is analogous to the procedure to identify the coefficients v_k in this fast regime. In particular, w_1 solves $a\beta'(aw_1) = u$. In what follows, we refer to w_1 by $\tau^\star$.

We define $\sigma_-^{\mathbb{Q}} := a\sqrt{\beta''(a\tau^\star)}$ (cf. the decomposition in Equation (1)). For this regime, we define

$$\overline{E}_n := \frac{C_n - un}{\psi_n \sqrt{\varphi_n}\sigma_-^{\mathbb{Q}}},$$

which is a random variable that has, by construction, mean 0 and variance converging to 1 under $\mathbb{Q}_n$. Comparing $\overline{E}_n$ with $\overline{D}_n$, observe that there is a factor $\psi_n\sqrt{\varphi_n}$ rather than $\sqrt{n}$ in the denominator. Again, we obtain, for all n, the factorization

$$\xi_n(u) = \mathrm{e}^{\gamma_n(\vartheta_n) - \vartheta_n un}\,\Delta_n, \text{ with } \Delta_n := \mathbb{E}_{\mathbb{Q}_n}\left(\mathrm{e}^{-\vartheta_n \sigma_-^{\mathbb{Q}} \psi_n \sqrt{\varphi_n}\,\overline{E}_n}\,1\{\overline{E}_n \geqslant 0\}\right). \tag{12}$$

As before, it remains to analyze the exponential factor $\delta_n := \exp(\gamma_n(\vartheta_n) - \vartheta_n un)$ and the expectation Δ_n, in the regime of n growing large.

The analysis of δ_n precisely follows the corresponding step in the fast regime. Define

$$m_- := \sup\Big\{k \in \mathbb{N} : \liminf_{n\to\infty} \varphi_n \psi_n^{-k} > 0\Big\}.$$

It thus follows that as $n \to \infty$, for constants $\overline{w}_k$, with the empty sum being defined as 0,

$$\delta_n = \gamma_n(\vartheta_n) - \vartheta_n u n = \big(\beta(a\tau^\star) - \tau^\star u\big)\varphi_n + \sum_{k=1}^{m_-} \overline{w}_k \varphi_n \psi_n^{-k} + o(1).$$

For the analysis of Δ_n, we refer to Appendix B. Specifically, there it is proven that $\sqrt{\varphi_n}\Delta_n$ converges to an (explicitly given) constant as $n \to \infty$. We thus arrive at the following result.

Theorem 2. *As $n \to \infty$, under Assumption 2, for nonlattice $B(\cdot)$,*

$$\xi_n(u) \sim \frac{1}{\tau^\star \sigma_-^{\mathbb{Q}} \sqrt{2\pi\varphi_n}} \exp\left(\big(\beta(a\tau^\star) - \tau^\star u\big)\varphi_n + \sum_{k=1}^{m_-} \overline{w}_k \varphi_n \psi_n^{-k}\right).$$

Corollary 2. *If $\varphi_n \psi_n^{-1} = n/\psi_n^2 \to 0$ as $n \to \infty$, then, under Assumption 2, for nonlattice $B(\cdot)$,*

$$\xi_n(u) \sim \mathbb{P}(aB(\varphi_n) \geqslant u\varphi_n) \sim \frac{1}{\tau^\star \sigma_-^{\mathbb{Q}} \sqrt{2\pi\varphi_n}} \exp\big(\big(\beta(a\tau^\star) - \tau^\star u\big)\varphi_n\big).$$

As $n \to \infty$, under Assumption 2, for nonlattice $B(\cdot)$,

$$\frac{1}{\varphi_n} \log \xi_n(u) \to \beta(a\tau^\star) - \tau^\star u.$$

Remark 3. In the case where $B(\cdot)$ is lattice, the result of Theorem 2 has to be slightly adjusted (cf. Remark 2). Let, for some x_0 and d and any $t \geqslant 0$, the random variable $d^{-1}(B(t) - x_0)$ be almost surely integer, and let d be the largest number with this property. Then, under Assumption 2, we obtain

$$\xi_n(u) \sim \frac{d}{1 - \mathrm{e}^{-a\tau^\star d}} \frac{1}{\sigma_-^{\mathbb{Q}}/a \cdot \sqrt{2\pi\varphi_n}} \exp\left(\big(\beta(a\tau^\star) - \tau^\star u\big)\varphi_n + \sum_{k=1}^{m_-} \overline{w}_k \varphi_n \psi_n^{-k}\right),$$

as $n \to \infty$. Analogously with what we observed in the fast regime, this behavior is consistent with that of Theorem 2 when taking $d \downarrow 0$.

Remark 4. In our analysis, we have assumed that $a > 0$. One may wonder whether our findings extend to general a. When picking $a \leqslant 0$, however, complications arise in the slow regime. The constant $\tau^\star$, which plays a crucial role in this regime, should solve the equation $u = \beta'(a\tau^\star)a$, but observe that $\beta'(\cdot)$ is positive (because it is the characteristic exponent of a subordinator); as a consequence, $u = \beta'(a\tau^\star)a$ has no solution for $a \leqslant 0$. This fact implies that the case $a \leqslant 0$ should be dealt with using an entirely different technique, in the sense that the change-of-measure technique (as was used earlier) cannot work.

4. Examples

In this section, we include two examples that demonstrate how the asymptotic expansion can be evaluated.

4.1. Example 1: $A(\cdot)$ Is a Poisson Process and $B(\cdot)$ a Gamma Process

Let $A(\cdot)$ be a Poisson process. We assume that its rate is $\lambda > 0$, so $\alpha(\vartheta) = \lambda(\mathrm{e}^\vartheta - 1)$ for $\vartheta \in \mathbb{R}$. Let $B(\cdot)$ be a Gamma process. We call the parameters $r > 0$ (shape) and $\mu > 0$ (rate), so $\beta(\vartheta) = r \log \mu - r \log(\mu - \vartheta)$, where we require $\vartheta < \mu$. Observe that $A(t)$ has a Poisson distribution with parameter λt and that $B(t)$ has a Gamma distribution with parameters rt and μ. In particular, in the terminology of our paper, $a = \lambda$ and $b = r/\mu$. To make the event of interest rare, we assume that $u > ab$. Writing $\varrho := \lambda r/(\mu u)$, this translates to assuming that $\varrho < 1$. Note that we have that $A(\cdot)$ is lattice but $B(\cdot)$ is not, so we should apply Remark 2 in the fast regime and Theorem 2 in the slow regime.

We start our computations by providing the l-mgf of $A(\psi_n B(\phi_n))$ for this special case:

$$\gamma_n(\vartheta) = \varphi_n \,\beta\big(\alpha(\vartheta)\,\psi_n\big) = r\,\varphi_n\,\log\left(\frac{\mu}{\mu - \lambda(e^{\vartheta} - 1)\psi_n}\right). \tag{13}$$

Because ϑ_n satisfies the first-order condition $\gamma_n'(\vartheta_n) = un$, we are to solve

$$\beta'(\alpha(\vartheta_n)\psi_n)\,\alpha'(\vartheta_n) = \frac{r\lambda\, e^{\vartheta_n}}{\mu - \lambda(e^{\vartheta_n} - 1)\psi_n} = u. \tag{14}$$

We thus find

$$\vartheta_n = \log\left(\frac{\mu u + \lambda\psi_n u}{\lambda r + \lambda\psi_n u}\right) = \log\left(1 + \frac{\lambda}{\mu}\psi_n\right) - \log\left(\varrho\,(1 + \frac{u}{r}\psi_n)\right).$$

We distinguish between the fast regime and the slow regime. In the fast regime, in which $\psi_n \to 0$ as $n \to \infty$, applying the Taylor expansion of the logarithm yields an expression for the coefficients v_k. With $\zeta_1 := \lambda/\mu$ and $\zeta_2 := u/r$,

$$\vartheta_n = \vartheta^\star + \sum_{k=1}^{\infty} v_k\psi_n^k,\ \vartheta^\star = \log\frac{1}{\varrho},\ v_k := \frac{(-1)^{k+1}}{k}\left(\zeta_1^k - \zeta_2^k\right).$$

We compute the coefficients $\overline{v}_k$ featuring in Lemma 1. To this end, note that by inserting the first-order condition of Equation (14) into Equation (13),

$$\gamma_n(\vartheta_n) = -r\,\varphi_n\,\log(\varrho e^{\vartheta_n}) = -r\,\varphi_n\,(\vartheta_n - \vartheta^\star) = -r\sum_{k=1}^{\infty} v_k\,\varphi_n\psi_n^k,$$

so

$$\delta_n = \gamma_n(\vartheta_n) - \vartheta_n un = -r\,\varphi_n\sum_{k=1}^{\infty} v_k\psi_n^k - \vartheta^\star un - un\sum_{k=1}^{\infty} v_k\psi_n^k.$$

It follows that $-r\,v_1 = (1-\varrho)u = b\alpha(\vartheta^\star)$. We thus observe that δ_n indeed has the form that was established in Lemma 1, with, for $k \in \{2, 3, \ldots\}$, $\overline{v}_k = -(r\,v_k + u\,v_{k-1})$. More explicitly,

$$\overline{v}_k = (-1)^k\left(\frac{r\left(\zeta_1^k - \zeta_2^k\right)}{k} - \frac{u\left(\zeta_1^{k-1} - \zeta_2^{k-1}\right)}{k-1}\right).$$

Noting that $(\sigma_+^{\mathbb{Q}})^2 = b\alpha''(\vartheta^\star) = u$, and bearing in mind Remark 2, we conclude that

$$\xi_n(u) \sim \frac{1}{1-\varrho}\frac{1}{\sqrt{2\pi\, un}}\exp\left(\left(1 - \varrho + \log\varrho\right)un + \sum_{k=2}^{m_+}\overline{v}_k\varphi_n\psi_n^k\right), \tag{15}$$

as $n \to \infty$.

The computations pertaining to the slow regime work very similarly. The crucial step is that we rewrite ϑ_n by “Tayloring” with respect to ψ_n^{-1} (rather than to ψ_n), with $\overline{\zeta}_1 := \mu/\lambda$ and $\overline{\zeta}_2 := r/u$,

$$\vartheta_n = \log\left(1 + \frac{\mu}{\lambda}\psi_n^{-1}\right) - \log\left(1 + \frac{r}{u}\psi_n^{-1}\right) = \sum_{k=1}^{\infty} w_k\psi_n^{-k},\ w_k = \frac{(-1)^{k+1}}{k}\left(\overline{\zeta}_1^k - \overline{\zeta}_2^k\right).$$

Then $\tau^\star = w_1 = \overline{\zeta}_1 - \overline{\zeta}_2 = \overline{\zeta}_1(1-\varrho) = \frac{r}{u}\left(\frac{1}{\varrho} - 1\right)$, and $\beta(a\tau^\star) = r\,\log\frac{1}{\varrho}$. This gives

$$\gamma_n(\vartheta_n) = -r\,\varphi_n\log(\varrho\, e^{\theta_n}) = \beta(a\tau^\star)\varphi_n - r\,\varphi_n\sum_{k=1}^{\infty} w_k\psi_n^{-k},$$

so

$$\delta_n = \gamma_n(\vartheta_n) - \vartheta_n un = (\beta(a\tau^\star) - \tau^\star u)\varphi_n - \sum_{k=1}^{\infty}(r\,w_k + u\,w_{k+1})\varphi_n\psi_n^{-k}.$$

Defining $(\sigma_-^{\mathbb{Q}})^2 = a^2\beta''(a\tau^\star) = u^2/r$, application of Theorem 2 leads, after minor calculations, to

$$\xi_n(u) \sim \frac{1}{\frac{1}{\varrho}-1}\frac{1}{\sqrt{2\pi\, r\varphi_n}}\exp\left(\left(1-\tfrac{1}{\varrho}+\log\tfrac{1}{\varrho}\right)r\varphi_n + \sum_{k=1}^{m_-}\overline{w}_k\varphi_n\psi_n^{-k}\right), \tag{16}$$

as $n \to \infty$, where $\overline{w}_k = -(r\,w_k + u\,w_{k+1})$ for $k \in \{1,2,\ldots\}$; observe the similarity with Equation (15).

Remark 5. It is noted that the expressions encountered in this example align with those featured in Heemskerk et al. (2017a, section 3) that deal with a case in which $\xi_n(u)$ could be evaluated explicitly. It can be checked that C_n has a negative binomial distribution, as follows. Observe that

$$\gamma_n(\vartheta) = r\varphi_n\log\left(\frac{\mu}{\mu - \lambda(\mathrm{e}^{\vartheta}-1)\psi_n}\right).$$

If a random variable has a negative binomial distribution with parameters k and p, then its l-mgf is of the form

$$k\log\left(\frac{p}{1-(1-p)\mathrm{e}^{\vartheta}}\right).$$

It thus follows that in the setting considered in this subsection, we have that C_n is negative binomial, with success probability equal to $p = \mu/(\mu + \lambda\psi_n)$ and the allowed number of failures $k = r\varphi_n$. Further intuition behind the appearance of the negative binomial distribution has been provided in Heemskerk et al. (2017a, remark 6).

4.2. Example 2: $A(\cdot)$ Is a Gamma Process and $B(\cdot)$ a Poisson Process

In our second example, the Gamma process and the Poisson process swap roles. In other words, we have $\alpha(\vartheta) = r\log\mu - r\log(\mu-\vartheta)$ for $\vartheta < \mu$ and $\beta(\vartheta) = \lambda(\mathrm{e}^{\vartheta}-1)$ for $\vartheta \in \mathbb{R}$. Again, to make the event of interest rare, we assume that $u > ab$ or, alternatively, $\varrho := \lambda r/(\mu u) < 1$ (where $a = r/\mu$ and $b = \lambda$). In this case, $A(\cdot)$ is nonlattice, whereas $B(\cdot)$ is, so we have to use Theorem 1 in the fast regime and Remark 3 in the slow regime.

Remark 6. Note that the random variable $A(\psi_n B(\varphi_n))$ is compound Poisson in this case, with Gamma distributed jumps. More precisely, the jumps are generated according to a Poisson process with rate $\varphi_n\lambda$, where the jumps are Gamma, with parameters $r\psi_n$ and μ.

In this case, the l-mgf is given by

$$\gamma_n(\vartheta) = \varphi_n\lambda\left(\left(\frac{\mu}{\mu-\vartheta}\right)^{r\psi_n} - 1\right). \tag{17}$$

Again, we have to distinguish between the fast regime and the slow regime. As before, we start with the fast regime. From Equation (17) it follows, by solving the first-order condition, that

$$\vartheta_n = \mu\left(1-\varrho^{\eta_n}\right), \eta_n := \frac{1}{1+r\psi_n},$$

so

$$\gamma_n(\vartheta_n) = \varphi_n\lambda\left(\tfrac{1}{\varrho}\cdot\varrho^{\eta_n} - 1\right).$$

To find the v_k, we write ϑ_n as a Taylor series in ψ_n. It is seen directly that $\vartheta^\star = \mu(1-\varrho)$. Some elementary calculus thus yields for the first coefficients $v_1 = -\mu r\varrho\,\log\frac{1}{\varrho}$ and

$$v_2 = \mu r^2\varrho\left(\log\tfrac{1}{\varrho}\right)\left(1-\tfrac{1}{2}\log\tfrac{1}{\varrho}\right).$$

Higher coefficients can be found in the same way.

The coefficients $\overline{v}_k$ can also be found:

$$\begin{aligned}\delta_n &= \gamma_n(\vartheta_n) - \vartheta_n un = \varphi_n\lambda\left(\tfrac{1}{\varrho}\cdot\varrho^{\eta_n}-1\right) - \mu(1-\varrho^{\eta_n})un\\ &= \frac{\lambda}{\varrho}(\varrho^{\eta_n}-1+1-\varrho)\varphi_n - \mu(1-\varrho^{\eta_n})un = \frac{\lambda}{\varrho}(1-\varrho)\varphi_n - \mu u\left(\tfrac{1}{r}\varphi_n+n\right)(1-\varrho^{\eta_n})\\ &= \frac{\lambda}{\varrho}(1-\varrho)\varphi_n - u\left(\tfrac{1}{r}\varphi_n+n\right)\left(\vartheta^\star+\sum_{k=1}^{\infty}v_k\psi_n^k\right)\\ &= \left(\frac{\lambda}{\varrho}-\frac{\mu u}{r}\right)(1-\varrho)\varphi_n - (1-\varrho)\mu un - \frac{u}{r}v_1 n - u\sum_{k=2}^{\infty}\left(\tfrac{1}{r}v_k+v_{k-1}\right)\varphi_n\psi_n^k\\ &= \left(1-\tfrac{1}{\varrho}+\log\tfrac{1}{\varrho}\right)\lambda rn - u\sum_{k=2}^{\infty}\left(\tfrac{1}{r}v_k+v_{k-1}\right)\varphi_n\psi_n^k,\end{aligned}$$

so $\overline{v}_k = -u\,(\frac{1}{r}v_k+v_{k-1})$. Note that the first term equals $(b\alpha(\vartheta^\star)-\vartheta^\star u)n$, in accordance with the result in Lemma 1. Theorem 1, with $(\sigma_+^{\mathbb{Q}})^2 = u^2/(\lambda r)$, yields

$$\xi_n(u) \sim \frac{1}{\left(\frac{1}{\varrho}-1\right)}\frac{1}{\sqrt{2\pi\,\lambda rn}}\exp\left(\left(1-\tfrac{1}{\varrho}+\log\tfrac{1}{\varrho}\right)\lambda rn + \sum_{k=2}^{m_+}\overline{v}_k\varphi_n\psi_n^k\right).$$

We proceed with the slow case. The w_k follow by expanding ϑ_n in terms of a Taylor series with respect to ψ_n^{-1} (recalling that $\psi_n\to\infty$):

$$\vartheta_n = \mu\left(1-\varrho\cdot\varrho^{\overline{\eta}_n}\right),\ \overline{\eta}_n := -\frac{r}{r+1/\psi_n}.$$

Routine calculations show that $\tau^\star = w_1 = (\mu/r)\log\frac{1}{\varrho}$, and

$$w_2 = -\frac{\mu}{r^2}\left(\log\tfrac{1}{\varrho}\right)\left(1+\tfrac{1}{2}\log\tfrac{1}{\varrho}\right),$$

where higher coefficients follow along the same lines. The $\overline{w}_k$ can be determined as before, leaving out a few intermediate steps:

$$\begin{aligned}\delta_n &= \gamma_n(\vartheta_n) - \vartheta_n un = \varphi_n\lambda(\varrho^{\overline{\eta}_n}-1) - \mu(1-\varrho\cdot\varrho^{\overline{\eta}_n})un\\ &= \left(1-\varrho+\log\varrho\right)\frac{\lambda}{\varrho}\,\varphi_n - u\sum_{k=1}^{\infty}\left(\tfrac{1}{r}w_k+w_{k+1}\right)\varphi_n\psi_n^{-k}.\end{aligned}$$

By applying Remark 3, with $(\sigma_-^{\mathbb{Q}})^2 = (r/\mu)^2\cdot\lambda/\varrho = ru/\mu$ and $\overline{w}_k = -u\,(\frac{1}{r}w_k+w_{k+1})$,

$$\xi_n(u) \sim \frac{1}{1-\frac{1}{\varrho}}\frac{1}{\sqrt{2\pi\,(\lambda/\varrho)\,\varphi_n}}\exp\left(\left(1-\varrho+\log\varrho\right)(\lambda/\varrho)\,\varphi_n + \sum_{k=1}^{m_-}\overline{w}_k\varphi_n\psi_n^{-k}\right).$$

5. Applications in Modeling Overdispersed Customer Streams

In operations research, performance evaluation and design of service systems are key topics of interest. Traditionally, the dominant assumption when modeling customer arrival processes is that of Poisson arrivals. As we mentioned in the Introduction, however, measurements indicate that the level of variation observed in practice may significantly exceed that predicted by the Poisson process (Whitt et al. 2007, Bassamboo et al. 2010, Liu and Whitt 2014, Mathijsen et al. 2018). More concretely, where for Poisson arrivals the mean and variance of the number of arrivals in a given time interval coincide, in practice, one typically observes that the variance is larger than the mean; this phenomenon is typically referred to as *overdispersion*.

In Heemskerk et al. (2017b), a mechanism is proposed that produces overdispersion. The main idea is that the arrival rate is random rather than deterministic; this is achieved by resampling the arrival rate periodically. The main underlying idea behind the resampling is that there is a (perhaps unobservable) environmental process that influences all potential customers in the same way. Think, for instance, of the weather conditions or the occurrence of specific events. As we argued in the Introduction, our two-timescale framework can be used to represent this resampled Poisson process.

Renormalizing time such that the resampling intervals have length 1, considering K such intervals, and letting the samples be i.i.d. nonnegative random variable $\Lambda_1, \dots, \Lambda_K$, then the number of arrivals in there K intervals is distributed as

$$C(K) = \text{Pois}\left(\sum_{i=1}^{K} \Lambda_i\right).$$

When designing service centers (e.g., when making staffing decisions), one would like to control the probability of excessive waiting times. For this reason, it could be informational to have a handle on the tail distribution corresponding to the number of arrivals in a certain time window.

The remainder of this section has two objectives. First, we quantify the error made by neglecting overdispersion (i.e., the situation in which the number of arrivals in the K slots has a Poisson distribution with parameter $K\,\mathbb{E}\Lambda$). Second, we compare our refined asymptotics with the crude asymptotics that were found in Heemskerk et al. (2017b), just involving the dominant term in the exponent (as featuring in the second statements of Corollaries 1 and 2).

5.1. Setup

In our experiments, the following example is studied. Let $\Lambda_1, \dots, \Lambda_K$ be i.i.d. samples from an exponential distribution with mean $1/\overline{\mu}$. We are interested in using our expansions to approximate the probability

$$\Pi(K, \overline{u}, \overline{\mu}) := \mathbb{P}(C(K) \geqslant \overline{u}), \tag{18}$$

for some $\overline{u} > K/\overline{\mu}$.

• Importantly, as we have seen in Remark 5, this example allows an explicit solution: here $C(K)$ has a negative binomial distribution with parameters K and success probability $\overline{\mu}/(1+\overline{\mu})$. We thus have, in self-evident notation,

$$\Pi(K, \overline{u}, \overline{\mu}) = \mathbb{P}\left(\text{NegBin}\left(K, \frac{\overline{\mu}}{1+\overline{\mu}}\right) \geqslant \overline{u}\right).$$

We chose the example of exponentially distributed Λ_i because the explicit expressions allow us to compare the various approximations with the corresponding exact values.

• When neglecting the overdispersion, one does not work with sampled values $\Lambda_1, \dots, \Lambda_K$ but rather with a deterministic arrival rate $\mathbb{E}\Lambda$ in any interval. In other words, when one would ignore the overdispersion, the probability $\Pi(K, \overline{u}, \overline{\mu})$ would be approximated by

$$\Pi^{(\text{Pois})}(K, \overline{u}, \overline{\mu}) := \mathbb{P}(\text{Pois}(K\,\mathbb{E}\Lambda) \geqslant \overline{u}).$$

This approximation, corresponding with the fast regime, is expected to work well if K is large relative to $\overline{u}$, and the contributions of the individual Λ_i are relatively small (i.e., $\overline{\mu}$ is relatively large).

• Along the same lines, one could consider the regime that K is substantial, but not in the order of $\overline{u}$, and the contributions of the individual Λ_i are relatively large (i.e., $\overline{\mu}$ is relatively small). Here the slow regime is expected to apply, suggesting that one ignore the Poisson sampling. Bearing in mind that the sum of exponentially distributed random variables (with equal parameter) has an Erlang distribution, we end up with (in self-evident notation)

$$\Pi^{(\text{Gamma})}(K, \overline{u}, \overline{\mu}) := \mathbb{P}\big(\text{Erl}\big(K, \overline{\mu}\big) \geqslant \overline{u}\big).$$

The crude approximations $\Pi^{(\text{Pois})}(K, \overline{u}, \overline{\mu})$ and $\Pi^{(\text{Gamma})}(K, \overline{u}, \overline{\mu})$ we wish to compare with approximations based on our two-timescale model. To this end, we convert the asymptotics that we found in Section 4 into approximations. We fix a number $n > 1$. This scaling parameter can be considered artificial in the sense that the choice of n will not affect the approximation (as will turn out later). Then we pick f such that $n^f = K$ (i.e., $f = \log K / \log n > 0$), and we set $u := \overline{u}/n$ and $\mu := \overline{\mu} n^{1-f}$. We thus obtain

$$\sum_{i=1}^{K} \Lambda_i \overset{\text{d}}{=} n^{1-f} \sum_{i=1}^{n^f} \Lambda_i^{\circ},$$

where Λ_i° is exponentially distributed with parameter μ. We have thus rewritten the probability in Equation (18) as

$$\mathbb{P}\left(\text{Pois}\left(n^{1-f}\sum_{i=1}^{n^f}\Lambda_i^\circ\right) \geqslant un\right).$$

Notice that this representation falls in the framework of Section 4.1, with $\varphi_n = n^f$. More specifically, we are in the situation that $A(\cdot)$ is a Poisson process with rate $\lambda = 1$, and $B(\cdot)$ is a Gamma process with shape parameter $r = 1$ and rate parameter μ.

The next step is to translate the asymptotics that we identified for $\xi_n(u)$, which are in terms of the parameters of the scaled model (i.e., μ, f, and u), into approximations for $\Pi(K,\overline{u},\overline{\mu})$, which are in terms of the original parameters (i.e., $\overline{\mu}$, K, and $\overline{u}$).

5.2. Fast Regime

We start by considering the regime in which K is large relative to $\overline{u}$ so that it makes sense to work with the asymptotics pertaining to the case $f > 1$ (i.e., the fast regime). First, observe that

$$\varrho = \frac{\lambda r}{\mu u} = \frac{K}{\overline{\mu}\,\overline{u}}, \tag{19}$$

which we assumed to be smaller than 1. In addition,

$$\zeta_1 = \frac{1}{\overline{\mu}\,n^{1-f}} = \frac{1}{\overline{\mu}\psi_n}, \zeta_2 = \frac{\overline{u}}{n} = \frac{\overline{u}}{K\psi_n},$$

so (after some elementary computations)

$$\overline{v}_k\varphi_n\psi_n^k = (-1)^k\left(\frac{K\,t_k^+}{k} - \frac{\overline{u}\,t_{k-1}^+}{k-1}\right), t_k^+ := \left(\frac{1}{\overline{\mu}}\right)^k - \left(\frac{\overline{u}}{K}\right)^k. \tag{20}$$

Also,

$$\vartheta^\star = \log\frac{\overline{\mu}\,\overline{u}}{K}.$$

Based on Equation (15), we eventually obtain the following approximation for the probability in Equation (18):

$$\Pi^{(\text{fast})}(K,\overline{u},\overline{\mu}) := \frac{1}{1-\varrho}\frac{1}{\sqrt{2\pi\overline{u}}}\exp\left((1-\varrho+\log\varrho)\,\overline{u} + \sum_{k=2}^{M_+}(-1)^k\left(\frac{K\,t_k^+}{k} - \frac{\overline{u}\,t_{k-1}^+}{k-1}\right)\right), \tag{21}$$

with ϱ given by Equation (19) and t_k^+ by Equation (20). In practice, one should choose the threshold M_+ such that adding more terms in the sum does not affect the outcome; realize that the object m_+ is not well defined in the prelimit context. Observe that as announced, the scaling parameter n does not appear in the approximation.

In Heemskerk et al. (2017b), logarithmic asymptotics of $\xi_n(u)$ were found. These lead to (crude) approximations that only cover the dominant term in the exponent (as in the second statement of Corollary 1). More concretely, one would approximate the probability of our interest by

$$\hat{\Pi}^{(\text{fast})}(K,\overline{u},\overline{\mu}) = \exp\big((1-\varrho+\log\varrho)\,\overline{u}\big).$$

5.3. Slow Regime

Now consider the regime in which K is small relative to $\overline{u}$, in which we opt for relying on the asymptotics corresponding to $f < 1$. Then

$$\overline{\zeta}_1 = \overline{\mu}\,n^{1-f} = \overline{\mu}\,\psi_n,\ \ \overline{\zeta}_2 = \frac{n}{\overline{u}} = \frac{K\psi_n}{\overline{u}},$$

implying that

$$\overline{w}_k\varphi_n\psi_n^{-k} = (-1)^k\left(\frac{K\,t_k^-}{k} - \frac{\overline{u}\,t_{k+1}^+}{k+1}\right),\ \ t_k^- := \overline{\mu}^k - \left(\frac{K}{\overline{u}}\right)^k. \tag{22}$$

Table 1. Approximations for Fast Regime

$(K,\bar{\mu},\bar{u})$	Π	$\Pi^{(\text{Pois})}$	$\hat{\Pi}^{(\text{fast})}$	$\Pi^{(\text{fast},0)}$	$\Pi^{(\text{fast},1)}$
$(100{,}000, 1{,}000, 150)$	1.90×10^{-6}	1.88×10^{-6}	2.00×10^{-5}	1.95×10^{-6}	1.97×10^{-6}
$(50{,}000, 500, 150)$	1.93×10^{-6}	1.88×10^{-6}	2.00×10^{-5}	1.95×10^{-6}	2.00×10^{-6}
$(10{,}000, 100, 150)$	2.13×10^{-6}	1.88×10^{-6}	2.00×10^{-5}	1.95×10^{-6}	2.21×10^{-6}
$(5{,}000, 50, 150)$	2.41×10^{-6}	1.88×10^{-6}	2.00×10^{-5}	1.95×10^{-6}	2.51×10^{-6}
$(1{,}000, 10, 150)$	5.89×10^{-6}	1.88×10^{-6}	2.00×10^{-5}	1.95×10^{-6}	6.82×10^{-6}

In addition,

$$\tau^\star = K^{1/f}\left(\frac{\bar{\mu}}{K}-\frac{1}{\bar{u}}\right).$$

Based on Equation (16), we thus obtain the following approximation for Equation (18):

$$\Pi^{(\text{slow})}(K,\bar{u},\bar{\mu}) := \frac{1}{\frac{1}{\varrho}-1}\frac{1}{\sqrt{2\pi K}}\exp\left(\left(1-\tfrac{1}{\varrho}+\log\tfrac{1}{\varrho}\right)K+\sum_{k=1}^{M_-}(-1)^k\left(\frac{K\,t_k^-}{k}-\frac{\bar{u}\,t_{k+1}^-}{k+1}\right)\right), \tag{23}$$

with ϱ given by Equation (19) and t_k^- by Equation (22). As before, the truncation level M_- is chosen such that including additional terms would not have any impact on the approximation. Again, we observe that the scaling parameter n does not appear in the approximation.

We also compare with the approximation that only covers the dominant term in the exponent [as in the second statement of Corollary 2, in line with the decay rate identified in Heemskerk et al. (2017b)]. This approximation reads

$$\hat{\Pi}^{(\text{slow})}(K,\bar{u},\bar{\mu}) = \exp\left(\left(1-\tfrac{1}{\varrho}+\log\tfrac{1}{\varrho}\right)K\right).$$

5.4. Numerical Experiments

We conclude this section by presenting a set of examples that illustrate the use of the developed approximations. We subsequently consider an example that corresponds to the fast regime and one that corresponds to the slow regime. They provide indications of the accuracy that can be achieved by Equations (21) and (23) that were based on our refined asymptotics relative to more crude approximations.

We start with an example corresponding to the fast regime. Recalling Equation (21), notice that our approximation $\Pi^{(\text{fast})}(K,\bar{u},\bar{\mu})$ is parameterized by the threshold M_+. In Table 1, the column with superscript 0 corresponds to neglecting the sum in the exponent of Equation (21) altogether, which could be seen as choosing $M_+ := 1$, whereas the column with superscript 1 corresponds to including one term in the sum in the exponent of Equation (21), that is, $M_+ := 2$. Throughout this example, $\varrho = \frac{1}{3}$ and $\bar{u} = 150$ are held fixed, explaining that the values in the third, fourth, and fifth columns are constant (as can be seen from the expressions for the approximations $\Pi^{(\text{Pois})}, \hat{\Pi}^{(\text{fast})}$, and $\Pi^{(\text{fast},0)}$).

The following observations can be made from Table 1. First, in the two top rows, there is so much timescale separation that $\Pi^{(\text{fast},0)}$ provides highly accurate results. In the other rows, the timescale separation becomes less pronounced, which leads to $\Pi^{(\text{fast},1)}$ becoming the preferred approximation. Here it is noted that including two or more terms in the exponent hardly improves the approximation in the third and fourth rows. In the last row, the timescales are so poorly separated that $\Pi^{(\text{fast},1)}$ is still relatively far off. We mention that the performance is significantly improved by using $M_+ = 3$, which leads to the highly accurate approximation $5.90\cdot10^{-6}$. Second, as anticipated, the Poisson approximation (neglecting the overdispersion) performs well if there is a substantial degree of timescale separation but leads to substantial underestimation if the timescales are relatively close together.

Table 2. Approximations for Slow Regime

$(K,\bar{\mu},\bar{u})$	Π	$\Pi^{(\text{Gamma})}$	$\hat{\Pi}^{(\text{fast})}$	$\Pi^{(\text{fast},0)}$	$\Pi^{(\text{fast},1)}$
$(100, 0.001, 150{,}000)$	5.98×10^{-6}	5.93×10^{-6}	7.84×10^{-5}	6.26×10^{-6}	6.31×10^{-6}
$(100, 0.05, 30{,}000)$	6.20×10^{-6}	5.93×10^{-6}	7.84×10^{-5}	6.26×10^{-6}	6.52×10^{-6}
$(100, 0.01, 15{,}000)$	6.48×10^{-6}	5.95×10^{-6}	7.84×10^{-5}	6.26×10^{-6}	6.80×10^{-6}
$(100, 0.05, 3{,}000)$	9.11×10^{-6}	6.03×10^{-6}	7.84×10^{-5}	6.26×10^{-6}	9.49×10^{-6}
$(100, 0.1, 1{,}500)$	1.35×10^{-5}	6.14×10^{-6}	7.84×10^{-5}	6.26×10^{-6}	1.44×10^{-5}

In the slow regime, our approximation $\Pi^{(\text{fast})}(K,\overline{u},\overline{\mu})$, as given by Equation (23), is parameterized by the threshold M_-. Analogously to the notation used in Table 1, in Table 2 the column with superscript 0 corresponds to neglecting the sum in the exponent of Equation (23), which one could identify with the choice $M_- := 0$, whereas the column with superscript 1 corresponds to including one term in the sum in the exponent of Equation (23), that is, $M_- := 1$. We chose scenarios in which $\varrho = \frac{2}{3}$ and $K = 100$ are held fixed, implying that the values in the fourth and fifth columns are constant (as can be seen from the expressions for $\hat{\Pi}^{(\text{fast})}$ and $\Pi^{(\text{fast},0)}$). Note that in this slow regime the values in the third column (i.e., $\Pi^{(\text{Gamma})}$) are not constant, in that they (very slowly) increase.

The conclusions from Table 2 are as follows. First, again, in the top rows (with a strong timescale separation), $\Pi^{(\text{fast},0)}$ performs well. This approximation, however, degrades in the lower rows, where the accuracy substantially improves if $M_- = 1$ (i.e., the approximation $\Pi^{(\text{fast},1)}$) is used. We mention that for the parameters in the last row, the accuracy is further improved by taking $M_- = 2$, yielding $1.34 \cdot 10^{-5}$. Second, the Gamma approximation (assuming full separation of time scales) provides accurate approximations in the top row but performs substantially worse in the bottom rows (where there is less timescale separation), as expected.

6. Discussion and Concluding Remarks

Motivated by recent developments in arrival process modeling, this paper presents tail asymptotics for a multitimescale model. The focus has been on approximating $\xi_n(u) = \mathbb{P}(A(\psi_n B(\varphi_n) \geqslant un)$, with $A(\cdot)$ a Lévy process and $B(\cdot)$ a Lévy subordinator. Our analysis shows that subtle analysis is required to identify the exact asymptotics. The structure found is considerably richer than in the model's classical (single-timescale) counterpart; in particular, the exponent includes additional terms.

6.1. Effect of Leaving Customers

In the case where $A(\cdot)$ is a Poisson process, we recover the mixed Poisson model proposed in Heemskerk et al. (2017b), which can be thought of as a Poisson process with periodically resampled rate (so as to generate overdispersion). An interesting extension of the results developed in this paper could concern the model in which the arrived clients leave after a service time distributed as the random variable J with $F_J(t) := \mathbb{P}(J \leqslant t)$. Such a setting could be considered as an infinite-server queue with overdispersed input. Using the insights from Heemskerk et al. (2017b), one thus obtains the following expression for the number of customers $\overline{C}_n$ present at time φ_n, under the scaling considered in this paper:

$$\overline{C}_n = A\left(\psi_n \int_0^{\varphi_n} (1 - F_J(t))\, \mathrm{d}B(t)\right).$$

Using calculation rules for Lévy processes, one obtains

$$\log \mathbb{E}e^{\vartheta \overline{C}_n} = \int_0^{\varphi_n} \beta\big(\alpha(\vartheta)\psi_n(1 - F_J(t))\big)\mathrm{d}t.$$

Plugging in $F_J(t) = 0$ for all $t \in [0, \varphi_n]$, we recover $\varphi_n \beta(\alpha(\vartheta)\psi_n)$ (which makes sense because this choice, customers do not leave the system). As a topic for further research, one could pursue deriving the exact asymptotics of $\mathbb{P}(\overline{C}_n \geqslant un)$ (which may require scaling the service times J). These asymptotics can be converted into approximations to be used as the basis of refined staffing rules.

6.2. Follow-Up Research

Other directions for future work include

- In the first place, multivariate extensions could be considered. One could, for instance, consider the probability that the vector

$$\big(A_1(\psi_n B(\varphi_n)), A_2(\psi_n B(\varphi_n))\big),$$

with $A(\cdot) = (A_1(\cdot), A_2(\cdot))$ a bivariate Lévy processes, attains a value in the set $[u_1 n, \infty) \times [u_2 n, \infty)$. The components of $A(\cdot)$ could be assumed dependent, but observe that $A_1(\psi_n B(\varphi_n))$ and $A_2(\psi_n B(\varphi_n))$ are dependent anyway (because the same $B(\varphi_n)$ is used). The multivariate single-timescale results of Chaganty and Sethuraman (1996) show that one should expect that various cases arise. There will be cases in which one component exceeding its threshold (with high probability) implies that the other component exceeds its threshold as well but also cases in which both constraints are tight.

• In Remark 4, we pointed out that in our setup it is required to assume that $a>0$, particularly in the slow regime. For $a\leqslant 0$, one may anticipate that $\xi_n(u)$ decays exponentially in n (unlike in the case of $a>0$, where $\xi_n(u)$ decays exponentially in φ_n), based on the following reasoning. The starting point is the large-deviations heuristic

$$\xi_n(u)\approx\max_{x>0}\mathbb{P}(B(\varphi_n)\approx x\varphi_n)\,\mathbb{P}(A(xn)\approx un). \tag{24}$$

The main insight is that taking into account only the leading term of the asymptotics, the first probability on the right-hand side of Equation (24) decays exponentially in φ_n, but the second decays exponentially in n so that their product decays exponentially in n. This should be contrasted with the case $a>0$ studied in this paper: here the maximum will be attained (roughly) at $x=u/a>0$, so $\{A(xn)\approx un\}$ is no rare event, and hence $\mathbb{P}(aB(\varphi_n)\approx u\varphi_n)$ dictates the tail behavior (thus leading to exponential decay in φ_n). A comparable situation has been considered in Heemskerk et al. (2017a, section 2.2). We stress, however, that relative to the $a>0$ case, this $a\leqslant 0$ case has a considerably lower practical relevance.

Appendix A. Edgeworth Expansions

In this appendix, we present the Berry-Esséen based calculations needed. More specifically, we establish Edgeworth expansions for C_n. We successively address the fast and slow regimes.

A.1. Fast Regime

The goal is to develop an expansion for $\mathbb{Q}_n(\overline{D}_n\leqslant x)$. The proof is a variation of that for sums of i.i.d. random variables (Feller 1971), and therefore, we only provide the main steps. First, observe that

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\overline{D}_n}=\left(\mathrm{e}^{\Gamma_n(\vartheta/\sigma_+^{\mathbb{Q}}\sqrt{n})}\right)^{\varphi_n}, \tag{A.1}$$

with $\Gamma_n(\vartheta):=\beta(\alpha(\vartheta+\vartheta_n)\psi_n)-\beta(\alpha(\vartheta_n)\psi_n)-\vartheta u\psi_n$. Obviously,

$$\mathrm{e}^{\Gamma_n(\vartheta/(\sigma_+^{\mathbb{Q}}\sqrt{n}))}=\sum_{k=0}^{\infty}\frac{1}{k!}\omega_n^{(k)},\ \omega_n^{(k)}:=\left(\Gamma_n\left(\frac{\vartheta}{\sigma_+^{\mathbb{Q}}\sqrt{n}}\right)\right)^k.$$

Owing to the definition of ϑ_n, $\Gamma_n'(0)=0$. It requires some elementary calculus to verify that with $\gamma_+^\circ:=\beta''(0)(\alpha'(\vartheta^\star))^2+\beta''(0)\alpha(\vartheta^\star)\alpha''(\vartheta^\star)+\beta'(0)\alpha'''(\vartheta^\star)v_1$,

$$\Gamma_n''(0)=\beta''(\alpha(\vartheta_n)\psi_n)(\alpha'(\vartheta_n))^2\psi_n^2+\beta'\big(\alpha(\vartheta_n)\psi_n\big)\alpha''(\vartheta_n)\psi_n=b\alpha''(\vartheta^\star)\psi_n+\gamma_+^\circ\psi_n^2+o(\psi_n^2),$$

and $\Gamma_n'''(0)=b\alpha'''(\vartheta^\star)\psi_n+o(\psi_n)$. Upon combining these results, we obtain

$$\mathrm{e}^{\Gamma_n(\vartheta/(\sigma_+^{\mathbb{Q}}\sqrt{n}))}=1+\frac{\vartheta^2}{2\varphi_n}+\frac{1}{2}\frac{\gamma_+^\circ}{(\sigma_+^{\mathbb{Q}})^2}\frac{\vartheta^2\psi_n}{\varphi_n}+\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{\vartheta^3}{\varphi_n\sqrt{n}}+o\left(\max\left\{\frac{\psi_n}{\varphi_n},\frac{1}{\varphi_n\sqrt{n}}\right\}\right).$$

Note that we have to distinguish between two cases: $\lim_{n\to\infty}\psi_n\sqrt{n}=0$ and $\liminf_{n\to\infty}\psi_n\sqrt{n}>0$.

In the case where $\lim_{n\to\infty}\psi_n\sqrt{n}=0$, we have, as a result of Equation (A.1),

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\overline{D}_n}=\left(1+\frac{\vartheta^2}{2\varphi_n}+\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{\vartheta^3}{\varphi_n\sqrt{n}}+o\left(\frac{1}{\varphi_n\sqrt{n}}\right)\right)^{\varphi_n},$$

which can be rewritten as

$$\left(1+\frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n}+\left(1+\frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n-1}\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{\vartheta^3}{\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right).$$

By a direct computation, we find that

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\overline{D}_n}=\exp\left(\frac{1}{2}\vartheta^2\right)\left(1+\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{\vartheta^3}{\sqrt{n}}\right)+o\left(\frac{1}{\sqrt{n}}\right).$$

Using the familiar inversion procedure for characteristic functions, we thus obtain, with $\phi(\cdot)$ denoting the probability density function of a standard Normal random variable,

$$\mathbb{Q}_n(\overline{D}_n\in\mathrm{d}x)=\phi(x)\left(1+H_3(x)\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{1}{\sqrt{n}}\right)+o\left(\frac{1}{\sqrt{n}}\right),$$

with $H_k(\cdot)$ the Hermite polynomial of degree k. This leads to, with $\Phi(\cdot)$ denoting the cumulative distribution function of a standard Normal random variable,

$$\mathbb{Q}_n(\overline{D}_n \leqslant x) = \Phi(x) - \phi(x)\left(H_2(x)\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{1}{\sqrt{n}}\right) + o\left(\frac{1}{\sqrt{n}}\right). \tag{A.2}$$

With the same reasoning as in the standard Edgeworth expansion (i.e., that for sums of i.i.d. random variables), the error term (being small relative to $1/\sqrt{n}$) is uniform in x.

In the case where $\liminf_{n\to\infty}\psi_n\sqrt{n} > 0$, by inserting our expansion into Equation (A.1),

$$\mathbb{E}_{\mathbb{Q}_n}\,e^{\vartheta\overline{D}_n} = \left(1 + \frac{\vartheta^2}{2\varphi_n} + \frac{1}{2}\frac{\gamma_+^\circ}{(\sigma_+^{\mathbb{Q}})^2}\frac{\vartheta^2\psi_n}{\varphi_n} + o\left(\frac{\psi_n}{\varphi_n}\right)\right)^{\varphi_n},$$

which can be rewritten as

$$\left(1 + \frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n} + \left(1 + \frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n - 1}\frac{1}{2}\frac{\gamma_+^\circ}{(\sigma_+^{\mathbb{Q}})^2}\vartheta^2\psi_n + o(\psi_n).$$

Using the same procedure as previously yields, uniformly in x,

$$\mathbb{Q}_n(\overline{D}_n \leqslant x) = \Phi(x) - \phi(x)\left(H_1(x)\frac{1}{2}\frac{\gamma_+^\circ}{(\sigma_+^{\mathbb{Q}})^2}\psi_n\right) + o(\psi_n).$$

In our setting, however, we need an error that is small relative to $1/\sqrt{n}$, which can be achieved by expanding $\Gamma_n''(0)$ further (because it can be verified that the contributions of the higher derivatives $\Gamma_n^{(k)}(0)$ for $k \geqslant 3$ are $o(1/\sqrt{n})$). Define

$$k_+ := \sup\left\{k \in \mathbb{N} : \liminf_{n\to\infty}\psi_n^k\sqrt{n} > 0\right\};$$

As a result of Assumption 1, this is a finite constant. Using the earlier type of reasoning, we conclude that there are constants $c_1, \ldots, c_{k_+}$ such that

$$\mathbb{Q}_n(\overline{D}_n \leqslant x) = \Phi(x) - \phi(x)\left(H_1(x)\sum_{k=1}^{k_+} c_k\psi_n^k + H_2(x)\frac{1}{6}\frac{b\alpha'''(\vartheta^\star)}{(\sigma_+^{\mathbb{Q}})^3}\frac{1}{\sqrt{n}}\right) + o\left(\frac{1}{\sqrt{n}}\right). \tag{A.3}$$

In the boundary case, where $\psi_n\sqrt{n}$ converges to a constant, we have $k_+ = 1$; the sum in Equation (A.3) consists of only one term.

A.2. Slow Regime

Here our objective is to find an expansion for $\mathbb{Q}_n(\overline{E}_n \leqslant x)$; the reasoning is analogous to that for the fast regime. The starting point is the mgf of $\overline{E}_n$ under the twisted distribution:

$$\mathbb{E}_{\mathbb{Q}_n}\,e^{\vartheta\overline{E}_n} = \left(e^{\Gamma_n(\vartheta/\sigma_-^{\mathbb{Q}}\sqrt{n\psi_n})}\right)^{\varphi_n}, \tag{A.4}$$

with, as before, $\Gamma_n(\vartheta) := \beta(\alpha(\vartheta + \vartheta_n)\psi_n) - \beta(\alpha(\vartheta_n)\psi_n) - \vartheta u\psi_n$. We get

$$\frac{1}{2}\left(\frac{\vartheta}{\sigma_-^{\mathbb{Q}}\sqrt{n\psi_n}}\right)^2\Gamma_n''(0) = \frac{1}{2}\frac{\vartheta^2}{(\sigma_-^{\mathbb{Q}})^2}\left(\beta''(\alpha(\vartheta_n)\psi_n)(\alpha'(\vartheta_n))^2\frac{1}{\varphi_n} + \beta'(\alpha(\vartheta_n)\psi_n)\alpha''(\vartheta_n)\frac{1}{n}\right). \tag{A.5}$$

Using the expansion of ϑ_n, the right-hand side of the previous display reads

$$\frac{1}{2}\frac{\vartheta^2}{(\sigma_-^{\mathbb{Q}})^2}\left(\beta''(a\tau^\star)a^2\frac{1}{\varphi_n} + \frac{\gamma_-^\circ}{n} + o\left(\frac{1}{n}\right)\right) = \frac{\vartheta^2}{2\varphi_n} + \frac{1}{2}\frac{\gamma_-^\circ}{(\sigma_-^{\mathbb{Q}})^2}\frac{\vartheta^2}{n} + o\left(\frac{1}{n}\right),$$

where $\gamma_-^\circ := \beta'(a\tau^\star)\alpha''(0) + 2a\,\alpha''(0)\beta''(a\tau^\star)\tau^\star + \frac{1}{2}a^2\alpha''(0)\beta'''(a\tau^\star)(\tau^\star)^2 + a^3\beta'''(a\tau^\star)w_2$. In addition,

$$\frac{1}{6}\left(\frac{\vartheta}{\sigma_-^{\mathbb{Q}}\sqrt{n\psi_n}}\right)^3\Gamma_n'''(0) = \frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{\vartheta^3}{\varphi_n^{3/2}} + o\left(\frac{1}{\varphi_n^{3/2}}\right).$$

We distinguish between the cases $\lim_{n\to\infty}\varphi_n^{3/2}/n = 0$ and $\liminf_{n\to\infty}\varphi_n^{3/2}/n > 0$. Mimicking the reasoning used for the fast regime, in the case where $\lim_{n\to\infty}\varphi_n^{3/2}/n = 0$,

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\bar{E}_n} = \left(1+\frac{\vartheta^2}{2\varphi_n}+\frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{\vartheta^3}{\varphi_n^{3/2}}+o\left(\frac{1}{\varphi_n^{3/2}}\right)\right)^{\varphi_n},$$

which can be rewritten as

$$\left(1+\frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n}+\left(1+\frac{\vartheta^2}{2\varphi_n}\right)^{\varphi_n-1}\frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{\vartheta^3}{\sqrt{\varphi_n}}+o\left(\frac{1}{\sqrt{\varphi_n}}\right).$$

This leads to, inserting our expansion into Equation (A.4),

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\bar{E}_n} = \exp\left(\frac{1}{2}\vartheta^2\right)\left(1+\frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{\vartheta^3}{\sqrt{\varphi_n}}\right)+o\left(\frac{1}{\sqrt{\varphi_n}}\right),$$

and, after inversion, to the Edgeworth expansion

$$\mathbb{Q}_n(\bar{E}_n \leqslant x) = \Phi(x)-\phi(x)\left(H_2(x)\frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{1}{\sqrt{\varphi_n}}\right)+o\left(\frac{1}{\sqrt{\varphi_n}}\right). \tag{A.6}$$

Finally, we focus on $\liminf_{n\to\infty}\varphi_n^{3/2}/n > 0$. Performing the same steps,

$$\mathbb{E}_{\mathbb{Q}_n}\,\mathrm{e}^{\vartheta\bar{E}_n} = \exp\left(\frac{1}{2}\vartheta^2\right)\left(1+\frac{1}{2}\frac{\gamma_-^\circ}{(\sigma_-^{\mathbb{Q}})^2}\frac{\vartheta^2}{\psi_n}+o\left(\frac{1}{\psi_n}\right)\right),$$

leading to, uniformly in x,

$$\mathbb{Q}_n(\bar{E}_n \leqslant x) = \Phi(x)-\phi(x)\left(H_1(x)\frac{1}{2}\frac{\gamma_-^\circ}{(\sigma_-^{\mathbb{Q}})^2}\frac{1}{\psi_n}\right)+o\left(\frac{1}{\psi_n}\right).$$

Our objective, however, is to obtain an error that is $o(1/\sqrt{\varphi_n})$. This is achieved by expanding Equation (A.5) further. Again, the contributions of the higher derivatives (i.e., derivatives of order 3 and higher) can be verified to be negligible. Following the line of reasoning of the fast regime, we define

$$k_- := \sup\left\{k\in\mathbb{N}: \liminf_{n\to\infty}\frac{\sqrt{\varphi_n}}{\psi_n^k} > 0\right\};$$

as a result of Assumption 2, this is a finite constant. Using the previous type of reasoning, we conclude that there are constants $c_1',\ldots,c_{k_-}'$ such that

$$\mathbb{Q}_n(\bar{E}_n \leqslant x) = \Phi(x)-\phi(x)\left(H_1(x)\sum_{k=1}^{k_-}\frac{c_k'}{\psi_n^k}+H_2(x)\frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}\frac{1}{\sqrt{\varphi_n}}\right)+o\left(\frac{1}{\sqrt{\varphi_n}}\right). \tag{A.7}$$

Again, in the boundary case where $\varphi_n^{3/2}/n$ converges to a constant, we have $k_- = 1$, and the sum in Equation (A.7) consists of only one term.

Appendix B. Analysis of Δ_n in the Slow Regime

Analogously to our analysis in the fast regime, applying integration by parts, we write Δ_n as

$$\begin{aligned}\Delta_n &= \int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_-^{\mathbb{Q}}\psi_n\sqrt{\varphi_n}\,x}\,\mathbb{Q}_n(\bar{E}_n\in\mathrm{d}x)\\ &= \psi_n\sqrt{\varphi_n}\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\vartheta_n\sigma_-^{\mathbb{Q}}\psi_n\sqrt{\varphi_n}\,x}\big(\mathbb{Q}_n(\bar{E}_n\le x)-\mathbb{Q}_n(\bar{E}_n\le 0)\big)\mathrm{d}x\\ &= \psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty \mathrm{e}^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\big(\mathbb{Q}_n(\bar{E}_n\leqslant x/\sqrt{\varphi_n})-\mathbb{Q}_n(\bar{E}_n\leqslant 0)\big)\mathrm{d}x.\end{aligned} \tag{B.1}$$

The goal of this appendix is to prove that $\sqrt{\varphi_n}\Delta_n$ converges to an (explicitly given) constant as $n\to\infty$. We again proceed by using the Edgeworth expansion presented in Appendix A.

In the case $\lim_{n\to\infty}\varphi_n^{3/2}/n = 0$ (which corresponds to $f < \frac{2}{3}$ in the context of Example 1), we have, as pointed out in Equation (A.6) in Appendix A, as $n \to \infty$,

$$\sqrt{\varphi_n}\sup_x\left(\mathbb{Q}_n(\overline{E}_n \leqslant x) - \Phi(x) + \phi(x)H_2(x)\frac{\kappa_-}{\sqrt{\varphi_n}}\right) \to 0,\ \kappa_- := \frac{1}{6}\frac{\beta'''(a\tau^\star)a^3}{(\sigma_-^{\mathbb{Q}})^3}. \tag{B.2}$$

Our objective is to prove that $\sqrt{\varphi_n}\Delta_n$ converges to a constant; as in the fast regime, we provide the upper bound, but the lower bound follows fully analogously. By Equations (B.1) and (B.2), for any $\varepsilon > 0$ and n sufficiently large,

$$\begin{aligned}\sqrt{\varphi_n}\Delta_n \leqslant{}& \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\left(\Phi\left(\frac{x}{\sqrt{\varphi_n}}\right) - \Phi(0)\right)dx\\ &- \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\left(\phi\left(\frac{x}{\sqrt{\varphi_n}}\right)H_2\left(\frac{x}{\sqrt{\varphi_n}}\right) - \phi(0)H_2(0)\right)\frac{\kappa_-}{\sqrt{\varphi_n}}dx + \varepsilon.\end{aligned}$$

The first term on the right-hand side equals

$$\sqrt{\varphi_n}\exp\left(\frac{1}{2}(\vartheta_n\sigma_-^{\mathbb{Q}}\psi_n)^2\varphi_n\right)(1 - \Phi(\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\sqrt{\varphi_n})),$$

which, using $x(1-\Phi(x))/\phi(x) \to 1$ and $\psi_n\vartheta_n \to \tau^\star$, converges to

$$\left(\tau^\star\sigma_-^{\mathbb{Q}}\sqrt{2\pi}\right)^{-1}. \tag{B.3}$$

As before, we split the second term into (recalling that $H_2(x) = x^2 - 1$)

$$\begin{aligned}t_{-,n}^{(1)} &:= \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\left(\phi(0) - \phi\left(\frac{x}{\sqrt{\varphi_n}}\right)\right)\frac{\kappa_-}{\sqrt{\varphi_n}}dx,\\ t_{-,n}^{(2)} &:= \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\phi\left(\frac{x}{\sqrt{\varphi_n}}\right)\frac{x^2}{\varphi_n}\frac{\kappa_-}{\sqrt{\varphi_n}}dx.\end{aligned}$$

Mimicking the reasoning used in the fast regime, it can be shown that $t_{-,n}^{(1)} \to 0$ and $t_{-,n}^{(2)} \to 0$ as $n \to \infty$. The third term equals ε, which can be made arbitrarily small. Combining this with the corresponding lower bound, we find that $\sqrt{\varphi_n}\Delta_n$ converges to Equation (B.2).

Considering $\liminf_{n\to\infty}\varphi_n^{3/2}/n > 0$ (which simplifies to $f \in [\frac{2}{3}, 1)$ in the context of Example 1), we obtain from Equation (A.7),

$$\sqrt{\varphi_n}\sup_x\left(\mathbb{Q}_n(\overline{E}_n \leqslant x) - \Phi(x) + \phi(x)\left(H_1(x)\sum_{k=1}^{k_-}c_k'\psi_n^{-k} + H_2(x)\frac{\kappa_-}{\sqrt{\varphi_n}}\right)\right) \to 0. \tag{B.4}$$

By Equation (B.4), for any $\varepsilon > 0$ and n sufficiently large, using that $H_1(x) = x$ and hence $H_1(0) = 0$,

$$\begin{aligned}\sqrt{\varphi_n}\Delta_n \leqslant{}& \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\left(\Phi\left(\frac{x}{\sqrt{\varphi_n}}\right) - \Phi(0)\right)dx\\ &- \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\phi\left(\frac{x}{\sqrt{\varphi_n}}\right)\frac{x}{\sqrt{\varphi_n}}\left(\sum_{k=1}^{k_-}c_k'\psi_n^{-k}\right)dx\\ &- \sqrt{\varphi_n}\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}\int_0^\infty e^{-\psi_n\vartheta_n\sigma_-^{\mathbb{Q}}x}\left(\phi\left(\frac{x}{\sqrt{\varphi_n}}\right)H_2\left(\frac{x}{\sqrt{\varphi_n}}\right) - \phi(0)H_2(0)\right)\frac{\kappa_-}{\sqrt{\varphi_n}}dx + \varepsilon.\end{aligned}$$

The first, third, and fourth terms can be dealt with as in the case $\varphi_n^{3/2}/n \to 0$. Analogously to the reasoning used for the fast regime, the second term vanishes. The corresponding lower bound is established in the same way. We conclude that also in this case $\sqrt{\varphi_n}\Delta_n$ converges to Equation (B.2).

References

Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).
Bahadur RR, Rao RR (1960) On deviations of the sample mean. *Ann. Math. Statist.* 31(4):1015–1027.
Bahadur RR, Zabell SL (1979) Large deviations of the sample mean in general vector spaces. *Ann. Probab.* 7(4):587–621.
Barndorff-Nielsen O, Cox D (1989) *Asymptotic Techniques for Use in Statistics* (Chapman & Hall, London)
Bassamboo A, Ramandeep S, Zeevi A (2010) Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* 56(10):1668–1686.
Blackwell D, Hodges J (1959) The probability in the extreme tail of a convolution. *Ann. Math. Statist.* 30(4):1113–1120.
Chaganty N, Sethuraman J (1996) Multidimensional strong large deviation theorems. *J. Statist. Planning Inference* 55(3):265–280.

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* 23(4): 493–507.

Cramér H (1938) Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736:5–23.

Dembo A, Zeitouni O (1998) *Large Deviations Techniques and Applications*, 2nd ed. (Springer, New York)

Ellis R (1984) Large deviations for a general class of random vectors. *Ann. Probab.* 12(1):1–12.

Feller W (1971) *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd ed. (John Wiley & Sons, New York)

Gärtner J (1977) On large deviations from the invariant measure. *Theory Probab. Appl.* 22(1):24–39.

Hall P (1992) *The Bootstrap and Edgeworth Expansion* (Springer, New York).

Heemskerk M, Kuhn J, Mandjes M (2017b) Rare-event analysis of mixed Poisson random variables, and applications in staffing. *Performance Evaluation* 114:56–77.

Heemskerk M, van Leeuwaarden J, Mandjes M (2017a) Scaling limits for infinite-server systems in a random environment. *Stochastic Systems* 7(1): 1–31.

Höglund T (1979) A unified formulation of the central limit theorem for small and large deviations from the mean. *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete* 49(1):105–117.

Liu Y, Whitt W (2014) Stabilizing performance in networks of queues with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 28(4):419–449.

Mathijsen B, Janssen A, van Leeuwaarden J, Zwart B (2018) Robust heavy-traffic approximations for service systems facing overdispersed demand. *Queueing Systems* 90(3/4):257–289.

Plachky D, Steinebach J (1975) A theorem about probabilities of large deviations with an application to queuing theory. *Periodica Mathematica Hungaria* 6(4):343–345.

Whitt W, Green L, Kolesar P (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Productions Oper. Management* 16(1):13–39.