Queueing Systems
https://doi.org/10.1007/s11134-021-09704-z

Infinite server queues in a random fast oscillatory environment

Yiran Liu[2] · Harsha Honnappa[1] · Samy Tindel[2] · Nung Kwan Yip[2]

Received: 9 April 2020 / Revised: 7 March 2021 / Accepted: 24 March 2021

Abstract

In this paper, we consider a $\mathrm{Cox}/G_t/\infty$ infinite server queueing model in a random environment. More specifically, the arrival rate in our server is modeled as a highly fluctuating stochastic process, which arguably takes into account some small timescale variations often observed in practice. We prove a homogenization property for this system, which yields an approximation by an $M_t/G_t/\infty$ queue with some effective parameters. Our limiting results include the description of the number of active servers, the total accumulated input and the solution of the storage equation. Hence, in the fast oscillatory context under consideration, we show how the queuing system in a random environment can be approximated by a more classical Markovian system.

Keywords Infinite server queue · Random environment · Homogenization

Mathematics Subject Classification 60K25 · 60G55 · 60F15 · 90B22

S. Tindel is supported by the NSF Grant DMS-1952966. H. Honnappa is supported by the NSF Grant CMMI-1636069.

✉ Harsha Honnappa
honnappa@purdue.edu

Yiran Liu
liu387@purdue.edu

Samy Tindel
stindel@purdue.edu

Nung Kwan Yip
yipn@purdue.edu

[1] School of Industrial Engineering, Purdue University, 315 N. Grant Street, West Lafayette, IN 47907, USA

[2] Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907, USA

Published online: 20 April 2021

1 Introduction

Nonstationary models have been extensively studied in the literature on queues, particularly in the Markovian setting. A typical assumption in this setting is that the arrival and service intensities are deterministic time-varying functions. In Markovian settings, it is also natural to assume that the intensity functions are smooth [1]. However, in practice, queueing systems are often subject to "environmental" noise: for instance, while arrival intensities to call centers and hospitals display time-of-day (or "diurnal") effects, the intensity functions also vary based on the day-of-week and seasonal effects. In other queueing systems, particularly those with high-intensity arrivals such as computer networks or cloud service systems, there is also intra-day and intra-hour stochastic variation in the intensity process. The performance of these queueing systems is therefore affected by both the smaller timescale stochastic variations and (relatively) longer timescale time-of-day effects.

Our aim is to show that in a setting where the stochastic fluctuations are strongly ergodic (in a general sense), a separation of short- and long-term variations naturally leads to a simpler model description. While infinite server queues are an approximation in the real world, they are a useful vehicle to address the questions of interest to us. To this end, we study a $\mathrm{Cox}/G_t/\infty$ infinite server queueing model imbedded in a random environment, wherein we assume a doubly stochastic Poisson (DSPP), or Cox, process traffic model, and conditional on the stochastic intensity the cumulative number of arrivals in a fixed time interval is Poisson-distributed. We assume that the stochastic intensity is modeled as $\mu(s) = \Psi(s, Z_s)$, where Z_s is an ergodic stochastic process. (As we will see later on, a typical example is an Ornstein–Uhlenbeck (OU)-type process.) We will develop and concentrate much of our theory under the special case of a "separable" stochastic intensity function, denoted by μ^ε and defined by

$$\mu^\varepsilon(\mathrm{d}s) = \lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s, \quad \varepsilon > 0,\ s > 0 \tag{1}$$

though a more general form is certainly possible. In the above, $\lambda(s)$ is a deterministic function of time modeling time-of-day effects. It is multiplied by some positive function of a stochastic process Z (modeling fluctuations). Note the timescale $\varepsilon^{-1}t$ associated with Z. The constant ε in this paper is intended to be a small parameter reflecting the fast oscillatory nature of the fluctuations. Coming to the service model, we consider a general setting where the parameters of the service time distribution functions are assumed to vary temporally with the long timescale variations in the traffic intensity. An example is given by Pareto service times, with temporally varying scale coefficients that depend on the arrival epoch.

Performance analysis of non-Markovian queueing models is in general rather difficult. Consequently, we focus on developing stochastic process approximations as the parameter $\varepsilon \to 0$. Such a limiting procedure is often called *homogenization* in the sense that the fast oscillating process is averaged out to produce an effective description which is usually much easier to analyze. In studying homogenization limits of stochastic processes in random environments, one may choose to fix a realization of the random environment and then take the homogenization limit, or to first average out the random environment and then take the averaged system to its homogenization

limit. The former approach is typically termed the *quenched* regime, while the latter is the *annealed* regime. In this paper, we present homogenization limits for both, though we primarily focus on the quenched scenario.

The literature on infinite server queues in random environments [2–16] has focused on the annealed regime. A natural question to ask is the relevance of the quenched regime to performance analysis of queues. As [17] observes in the context of Markov models in random environments, *"A reasonable first answer to this question could simply be that the quenched approach solves the true question but the averaged* [annealed] *approach, being often much simpler, has the merit of being the first possibility to understand a hard problem..."*. Performance analysis is typically an *ex ante* exercise, using a stochastic model that predicts the future evolution of the queue. Roughly speaking, an annealed approach to performance analysis permits the prediction of the behavior of the queue in the *typical* random environment. The quenched analysis, on the other hand, makes predictions for *any (almost sure)* realization of the random environment. In this sense, our main result in Theorem 3.11 showing that, in the limit, the $\mathrm{Cox}/G_t/\infty$ queue is closely approximated by an $M_t/G_t/\infty$ queue, is a result that is "robust" to the specific realization of the random environment. As Ben-Arous et al. observe [17], this is in some sense the "true question" in that, *ex ante*, it is unclear what sample path of the random environment will be encountered by the queue, and any analysis should be agnostic to this.

Another perspective is that which of the annealed or quenched regimes is more indicative will certainly depends on the *variability* of the system or more precisely the *fluctuation* around the typical (annealed) description. Furthermore, there are ample examples in probability and stochastic analysis in which there are interesting transitions and connections between the two regimes, and the results intimately depend on the how the averaging procedures are performed. For this, besides the work [17], we also refer the reader to [18,19].

With the above observation in mind, we will provide results for both the quenched and annealed regimes. In Theorem 3.11, we show that N^ε, the *quenched* stochastic process representing the state of the $\mathrm{Cox}/G_t/\infty$ queue, converges weakly to a Poisson point process N that is the state of an $M_t/G_t/\infty$ queue. We write this limit as

$$\{N^\varepsilon(t) : t \in [0,T]\} \xrightarrow{\text{(d)}} \{N(t) : t \in [0,T]\},\ \mathbb{P}_Z - \text{a.s.}, \tag{2}$$

where $T > 0$ is an arbitrary time horizon and $\mathbb{P}_Z$ denotes the quenched probability for a fixed environment sample path Z. The proof of our main result crucially depends on the assumption that the short-term stochastic fluctuation model mixes rapidly enough (see Hypothesis 2.1 for a more precise statement). Consequently, the stochastic fluctuations can reach a steady state within the relatively longer timescale of the time-of-day effects; thereby, the short timescale fluctuations are "averaged out." The proof relies on first showing that the mean measure of the queue state satisfies a strong law-type limit. Namely it can be seen that the mean measure of the queue state is an additive functional of the stochastic fluctuations, and when the latter reach steady-state fast enough, the mean measure converges to a deterministic limit. This type of result is well known for general Markov processes [20]. Our homogenization limit follows by leveraging the

convergence of mean measure to show that the finite-dimensional distributions of the Poisson random measure corresponding to the (quenched) state N^ε of the $\mathrm{Cox}/G_t/\infty$ queue converge to those of the state N of an $M_t/G_t/\infty$ queue. A further tightness condition leads to the convergence of the whole process. As the reader can see, the main technical novelty in our paper consists in combining those homogenization results with some considerations about limits for queues.

In Theorem 4.1 we also provide a homogenization analysis in the annealed regime. Specifically, we show that the $\mathrm{Cox}/G_t/\infty$ queue converges to the same $M_t/G_t/\infty$ queue in Theorem 4.1 even in the annealed regime, under a more relaxed ergodicity condition than that implied by Hypothesis 2.1. This result clearly shows that the quenched analysis is different, requiring stronger ergodicity conditions to ensure *almost sure* convergence.

Our main results Theorems 3.11 and 4.1 yield two crucial insights into the performance analysis of queueing models imbedded in random environments. First, it shows that ergodic properties of the underlying stochastic fluctuations play a critical role in separating the fast stochastic fluctuations from the longer timescale temporal variations. Second, it shows that, under this asymptotic timescale separation and rapid averaging of the stochastic fluctuations, the state of the infinite server queue in a random environment can be approximated by that of a time-varying infinite server queue. The latter model has been extensively studied [21–23], and there is a substantial literature available on its properties, particularly with stationary service.

The aforementioned homogenization phenomenon put forward in Theorem 3.11 is similar to the "rapid fluctuation" analysis in [24]. In that work, the weak convergence of a general point process (for example, a DSPP) to a constant rate Poisson process (under the assumption that the compensator of the point process satisfies a strong law) was used to approximate the state distribution of a $G_t/G/\infty$ queue by that of an $M_t/G/\infty$ queue. The analysis in [24] crucially used Taylor expansions of the state probability distribution at a fixed time t in terms of the scaling parameter. The approach in the current paper is completely different. Indeed, we mostly exploit the Poisson random measure representation of the state process and then establish the process-level stochastic approximation limit for this object. Our paper is also closely related to [25], where an infinite server queue with "extremely" heavy-tailed Pareto service times is studied in a time homogeneous setting in order to explain network self-similarity effects. The relatively simple homogeneous setting allowed the authors to establish not only a functional strong law of large numbers (FSLLN), but also a functional central limit theorem (FCLT). Our results in this paper substantially generalize the FSLLN result to a queueing model imbedded in a random environment. However, the analysis to establish the corresponding FCLT in our setting is significantly more complicated; see our conclusion section.

1.1 Literature survey

We provide a brief survey of literature relevant to this paper, placing it in context. Indeed, our paper can be related to multiple threads of research. First, our main theorems rely on the establishment of a stochastic averaging/homogenization result for the

state of the queue, both conditioned and averaged over the underlying stochastic environment model. Homogenization/stochastic averaging has been studied extensively in the literature; see [24,26–29] for a small sample of this literature. Typically, results on stochastic averaging principles in the form of a law of large number (LLN) involve two processes, one evolving on a fast timescale and reaching an equilibrium quickly and another one on a slower timescale that only experiences the former in equilibrium. In the limit the slow process is (typically) approximated by a process with "averaged" coefficients. Averaging principles have been used in finance [30–32], as well as in heavy-traffic analyses of certain controlled stochastic network models [33–35]. The averaging principle in our setting is established under the presumption that the process on the fast timescale is an external stochastic environment, while in [33–35] the process on the fast timescale is usually the state of one of the queues in the network that influences the state of other slowly varying queues. More closely related is the work in [24] where the rapidly fluctuating process is an external stochastic environment process.

We also allow for time-dependent service and traffic intensity in our model, and therefore, the literature on the analysis of time-varying queues and stochastic models is highly relevant. In the context of the performance analysis of deterministically time-varying queues (not necessarily in a random environment), there is a significant body of work developing both uniform acceleration [36–39] and many-server heavy-traffic limit theorems [40–42] to time-varying queues. In much of this literature, the limit processes are shown to be (reflected) fluid or diffusion limit processes. Note that all of this work assumes that the nonstationarity manifests as a deterministic temporal variation. There is also a growing body of work developing asymptotic expansions [5,24,43–47] of performance metrics. It is well known, however, that traffic arriving at call centers and hospitals displays significant over-dispersion relative to a Poisson process with deterministic intensity [48], implying that a DSPP is an appropriate model of the traffic in these systems. Much of this literature assumes that either the traffic and/or service processes are Markov modulated, where the underlying stochastic environment process is a finite state Markov chain; the vast majority of the related literature focuses on characterizing stationary behavior, but [45,49] exhibit a couple of examples where asymptotic limit theorems and expansions can be established. The plethora of methods for analyzing time-varying queues will, of course, be crucial for further analyzing the limit $M_t/G_t/\infty$ queue. We do not address this fact explicitly in this paper, however.

Most relevant to our current setting, of course, is the expanding literature on infinite server queues in random environments [2–4,6–16,50]. More specifically, there are extensive studies on infinite server queues with Markov-modulated input, in which the arrival intensity of the input is λ_i with i being the state of the latent environment modeled by a Markov jump process. Results established in this realm touch upon steady-state, functional central limit theorems and large deviations where the latent Markov process is sped up appropriately resulting in a homogenization effect [7,9–16]. In all of these papers, the objective is an annealed analysis in the sense described above. The more general setting of input models where the arrival intensity changes continuously as a function of the latent environment process has also been extensively studied [2–4,6,8], including homogenization results [3,4] showing that the state of the

infinite server queue asymptotically aligns with that of a time-homogeneous $M/G/\infty$ queue, steady-state analysis [2,6] and large deviation analysis [8,51,52]. [50] studies a so-called $M_S/G/\infty$ queue with a Cox arrival process with shot noise intensity and independent and identically distributed (i.i.d.) service times. This intensity model is a special case of the general setting we consider. This paper establishes both exact and asymptotic results; in the former case, the paper derives an expression for the joint moment generating function of the number in system and arrival intensity process, while in the latter a functional central limit theorem (FCLT) is established in the limit of a large arrival intensity scale. Our annealed results (and, obviously, the quenched results) are fundamentally different since our analysis is one of stochastic homogenization, as opposed to a large-scale asymptotic. Note that all of these cited results are in the annealed regime, and to the best of our knowledge the quenched regime has not been studied before for infinite server queues. As our results show, under suitable conditions on the random environment process, homogenization of the quenched state process aligns with the results on the annealed regime.

We note in particular the work [3] where the effect of over-dispersed traffic on the performance of an infinite server queue is studied. Paralleling our findings, this paper shows that a sufficiently rapidly fluctuating environment (relative to a slowly changing arrival intensity) will, in an appropriate asymptotic regime, ensure that the infinite server system behaves like a "standard" infinite server queue in steady state. On the other hand, in [3] the traffic intensity does not have an explicit time-of-day component and, for analytical reasons, the random environment is formulated in a somewhat "stylized" fashion. Our statements complement these results and, more significantly, show that the standard infinite server queue behavior is preserved even with explicit time-of-day effects in the traffic and service processes. In addition, let us observe again that our results are obtained in the quenched and annealed regimes (as opposed to only the annealed regime in [3]). Otherwise stated, the main limit result Theorem 3.11 is valid for almost any realization of the environment Z.

The rest of this paper is organized as follows: Section 2 introduces the notation that will be used throughout this paper and constructs the random arrival model of interest with appropriate hypotheses. Section 3 analyzes our queueing model in the quenched regime, with Sect. 3.1 deriving a formula for the mean measure m^ε of the Poisson random variable N^ε, Sect. 3.2 analyzing the limiting behavior of m^ε and Sect. 3.3 for proving our main theorem which is the convergence of the overall process N^ε. Section 3.4 provides an extension to more general arrival density. In Sect. 4, we prove a corresponding result but in the annealed regime. In Sect. 5, we provide some numerical examples to illustrate the theory. Finally, in Sect. 6, we give some conclusion and future perspectives.

2 Basic notation and Poisson-based model

We model the $\mathrm{Cox}/G_t/\infty$ queue using a Poisson point process imbedded in a random environment. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with respect to which we define all random elements to follow. The expectation with respect to $\mathbb{P}$ is denoted as $\mathbb{E}(\cdot)$.

2.1 Model for the random environment

As mentioned in the introduction, we incorporate fast oscillations modeled by an ergodic stochastic process Z into the arrival rate of our queueing system. In this section, we proceed to describe such a process. The underlying assumption about the process will clearly depend on the type of limiting results we want to establish. In particular, we will introduce two hypotheses, one for the quenched and one for the annealed consideration.

Hypothesis 2.1 Let $Z = \{Z_t\,;\, t \geq 0\}$ be a $\mathbb{R}^d$-valued stochastic process with sample paths that are right continuous with left limits (RCLL) defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The initial distribution $\mathcal{L}(Z_0)$ of Z is denoted by ρ_0, and we suppose that Z possesses a unique invariant probability measure π. We also assume that Z is strongly ergodic with rate $\kappa > 0$ in the following sense: for any regular enough function $\psi : \mathbb{R}^d \to \mathbb{R}$, there exists a finite random variable $C = C_\psi(\omega) > 0$ such that $\mathbb{P}$-almost surely we have

$$\left|\frac{1}{t}\int_0^t \psi(Z_u)\,\mathrm{d}u - \bar{\psi}\right| \leq \frac{C}{(1+t)^\kappa}\,, \quad \text{with} \quad \bar{\psi} = \int_{\mathbb{R}^d} \psi(z)\,\pi(\mathrm{d}z). \tag{3}$$

Remark 2.2 Note that the above hypotheses gives a convergence rate in the law of large number-type statement. The constant C in general can depend on the realization of the random process Z. There exists an abundant literature about results of the form (3) for Markov chains. A general framework is developed in [53], which yields the following particular case: Set $Z_u = X_{[u]}$, where $[u]$ denotes the integer part of u and $\{X_j\,;\, j \geq 0\}$ is a reversible ergodic Markov chain on a countable state space E. Let $\psi : E \to \mathbb{R}$ be such that $\sigma^2(\psi) < \infty$, where

$$\sigma^2(\psi) = \lim_{N\to\infty} \frac{1}{N}\,\mathrm{Var}\left(\sum_{j=0}^{N-1} \psi(X_j)\right).$$

Then, under some additional moment condition, it can be established that for any $0 < \kappa < \frac{1}{2}$, the following holds almost surely:

$$\lim_{t\to\infty} (1+t)^\kappa \left[\frac{1}{t}\int_0^t \psi(Z_u)\mathrm{d}u - \bar{\psi}\right] = 0, \tag{4}$$

which immediately implies relation (3). The bound $1/2$ on κ is also consistent with the Law of the Iterated Logarithm. Other examples of Markov processes (more specifically Harris chains) satisfying (3) are provided in [54,55], based on the law of the iterated logarithm. Notice that [55] handles directly some continuous time Markov processes.

Remark 2.3 Hypothesis 2.1 can also be fulfilled in some non-Markovian contexts. Indeed, consider an $\mathbb{R}^d$-valued fractional Brownian motion B with Hurst parameter

$H \in (0,1)$. Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be a function such that the following inward property is satisfied for a constant $a > 0$:

$$\langle b(x) - b(y), x - y \rangle_{\mathbb{R}^d} \leq -a \|x - y\|^2 .$$

We consider the process Z which solves the following stochastic differential equation:

$$Z_t = a + \int_0^t b(Z_s)\, \mathrm{d}s + B_t ,$$

where $a \in \mathbb{R}^d$. Then, combining [56] and [57], one can prove that Z satisfies Hypotheses 2.1 (details are omitted since this result is unrelated to the main message of the current paper). Observe that the case of a Brownian motion B with $b = -a\,\mathrm{Id}_{\mathbb{R}^d}$ corresponds to the classical Ornstein-Uhlenbeck case, for which Hypothesis 2.1 thus hold.

The consideration of the annealed case allows us to use a weaker ergodicity condition. More precisely, we introduce

Hypothesis 2.4 Let Z be the $\mathbb{R}^d$-valued process introduced in Hypothesis 2.1. We assume that

$$\lim_{t\to\infty} \left\| \frac{1}{t} \int_0^t \psi(Z_r)\, \mathrm{d}r - \bar{\psi} \right\|_{L^1(\Omega)} = 0, \quad \text{with} \quad \bar{\psi} = \int_{\mathbb{R}^d} \psi(z)\, \pi(\mathrm{d}z). \tag{5}$$

Observe that the above condition will certainly follow if we assume some uniform integrability of the constant C (which in fact is a random variable) in Hypothesis 2.1. As noted in the introduction, since the annealed analysis demonstrates the limiting behavior of the infinite server queue in a "typical" stochastic environment, it is somewhat unsurprising that a weaker ergodicity condition can be assumed.

2.2 Model for the system state

Having specified our random environment, we now describe our model for the system state. It is determined by the arrival and service times that we proceed to define below.

Our random environment will enter into the intensity of the arrival process. Namely, the arrival process is conceived as follows.

Hypothesis 2.5 Given $\varepsilon > 0$, the sequence of arrival times $\{\Gamma_k^\varepsilon ; k \geq 1\}$ is distributed as the jump times of a Poisson process with nonhomogeneous intensity $\{\lambda(s)\psi(Z_{s/\varepsilon}) ; s \geq 0\}$, where Z fulfills Hypothesis 2.1 and $\lambda, \psi : \mathbb{R}^d \to \mathbb{R}_+$ are positive Lipschitz (or C^1-) functions.

Next we turn our attention to the service process. It is given as a family $\{L_k^\varepsilon ; k \geq 1\}$ of random variables which are independent conditional on the arrival process $\{\Gamma_k^\varepsilon ; k \geq 1\}$. Their distribution is defined by $\mathcal{L}(L_k^\varepsilon | \Gamma_k^\varepsilon) = \nu(\Gamma_k^\varepsilon, \mathrm{d}r)$ for all k, where we consider

$$\nu = \{\nu(s, \mathrm{d}r) ;\ s \geq 0, r \geq 0\} \tag{6}$$

as a family of conditional regular laws. In the sequel, we will also resort to the following notation for the tail of ν:

$$\bar{F}_s(r) := \int_r^\infty \nu(s, \mathrm{d}\tau), \quad r \geq 0. \tag{7}$$

The main assumption on the measures ν is given below.

Hypothesis 2.6 Let ν be the family of measures defined by (6). We suppose that every $\nu(s, \cdot)$ admits a density ℓ_s, that is, $\nu(s, \mathrm{d}r) = \ell_s(r)\mathrm{d}r$. Moreover, there exist $\alpha > 0$ and a constant $c > 0$ such that the family ℓ_s verifies

$$\ell_s(r) \leq c\left(\frac{1}{r^{1+\alpha}} \wedge 1\right), \quad \text{for all} \quad r, s > 0\,; \tag{8}$$

$$\frac{\partial \ell_s(r)}{\partial s} \leq c\left(\frac{1}{r^{1+\alpha}} \wedge 1\right), \quad \text{for all} \quad r, s > 0. \tag{9}$$

Remark 2.7 Hypothesis 2.6 covers a large variety of service time distributions, including both "light-tailed" models such as the Gamma distribution and "heavy-tailed" models such as the Pareto-like distributions. The latter is of particular interest, as attested by Resnick and Rootzén [25], for example. Therefore, a typical example the reader might have in mind is given by

$$\ell_s(r) = k_s(r)\, \mathbb{1}_{[0,d_s]}(r) + \frac{c_s}{r^{1+\alpha}}\, \mathbb{1}_{(d_s,\infty)}(r)\,, \tag{10}$$

where for each $s > 0$, k_s is a smooth function, d_s and c_s are positive constants, and $\alpha > 0$. Notice that a positive random variable X whose density ℓ satisfies (10) has the property that $E[X^\beta] < \infty$ for $\beta < \alpha$ and $E[X^\beta] = \infty$ for $\beta \geq \alpha$. Hence, relation (10) allows a good calibration of the boundedness of moments for the service time.

Remark 2.8 The function $s \mapsto c_s$ in (10) can be thought of as a smooth and bounded slowly-varying function which modulates the service according to the arrival time. A specific example is given by the following oscillating function:

$$c_s = 1 + \beta \sin(ks), \quad \text{with} \quad \beta \in (0, 1) \quad \text{and} \quad k \geq 0. \tag{11}$$

With the arrival and service times in hand, our queueing system is classically described by a point process. Namely for $\varepsilon > 0$, we consider the following counting measure on $\mathbb{R}_+ \times \mathbb{R}_+$:

$$M^\varepsilon := \sum_{k=1}^\infty \delta_{(\Gamma_k^\varepsilon, L_k^\varepsilon)}. \tag{12}$$

Then, our main variable of interest is the number of active jobs in the infinite server queue at time t, which can be expressed as

$$N^{\varepsilon}(t) = \sum_{k=1}^{\infty} \mathbb{1}_{\{\Gamma_k^{\varepsilon} < t < \Gamma_k^{\varepsilon} + L_k^{\varepsilon}\}} = M^{\varepsilon}\{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+, x < t < x + y\}. \quad (13)$$

Our main aim in this paper is to derive a limit theorem for the process $N^{\varepsilon} = \{N^{\varepsilon}(t); t \in [0, T]\}$ as $\varepsilon \to 0$, for an arbitrary time horizon $T > 0$.

Remark 2.9 As the reader might have seen, there are two levels of randomness in our model. The first level corresponds to the random environment Z described in Sect. 2.1, while the second source of randomness is embodied in the Poisson point process N^{ε} given by (13). As in most of the literature on random environments, we shall play with the notion of quenched and annealed probabilities. The quenched probability corresponds to conditioning on the process Z. This probability will be denoted by $\mathbb{P}_Z$, with a corresponding expectation $\mathbb{E}_Z$. The annealed probability, given and denoted by $\mathbb{P}$, represents the global probability taking into account all the randomness involved in our system, in particular from Z and the arrival and service times T_k^{ε} and L_k^{ε} for $k = 1, 2, 3, \ldots$. The relation between quenched and annealed probabilities is summarized as

$$\mathbb{P}_Z(\cdot) = \mathbb{P}(\cdot|Z), \quad \text{and} \quad \mathbb{E}_Z[\cdot] = \mathbb{E}[\cdot|Z]. \quad (14)$$

3 Quenched analysis

Here, we proceed to analyze the Poisson point process N^{ε} defined in the previous section. We first consider the quenched regime, i.e., conditioning on a given realization of the random environment Z. The key technical device is the analysis of the mean measure m^{ε} and its asymptotics as $\varepsilon \to 0$.

3.1 Mean measure of N^{ε}

This section is devoted to a full description of the law of N^{ε}. The main result is summarized in the following proposition giving the conditional law of $N^{\varepsilon}(t)$. We want to emphasize that the quantities $\tilde{\nu}^{\varepsilon}$, μ^{ε} and m^{ε} below are all functions of Z, or its realization.

Proposition 3.1 *Let M^{ε} and $\{N^{\varepsilon}(t) : t \geq 0\}$ be defined by (12) and (13), respectively. Then, under the quenched probability $\mathbb{P}_Z$, M^{ε} is a Poisson random measure with mean measure given by*

$$\tilde{\nu}^{\varepsilon}(\mathrm{d}x, \mathrm{d}y) = \nu(x, \mathrm{d}y)\mu^{\varepsilon}(\mathrm{d}x), \quad \textit{with} \quad \mu^{\varepsilon}(\mathrm{d}s) = \lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s, \quad (15)$$

where we recall that ν is introduced in (6). Furthermore, we have that for any $t > 0$, $N^\varepsilon(t)$ is a Poisson random variable with parameter

$$m^\varepsilon(t) = \int_{\{(x,y):x<t<x+y\}} \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x) = \int_0^t \int_{t-x}^\infty \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x). \tag{16}$$

Proof First, we will show that M^ε is a Poisson random measure. To this aim, recall Remark 2.9 for the definition of the quenched probability $\mathbb{P}_Z$. Then, under $\mathbb{P}_Z$ and according to (12), the point process M^ε is of the form $\sum_{k\geq 1} \delta_{(\Gamma_k^\varepsilon, L_k^\varepsilon)}$, where $\{\Gamma_k^\varepsilon;\ k \geq 1\}$ is a Poisson process (see Hypothesis 2.5). Thanks to [58, Proposition 2.2], we get that M^ε is a Poisson point process under $\mathbb{P}_Z$, whose mean measure $\tilde{\nu}^\varepsilon$ can be decomposed as

$$\tilde{\nu}^\varepsilon(\mathrm{d}x, \mathrm{d}y) = \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x),$$

where ν is the measure featuring in (7) and μ^ε is defined by (15).

Therefore, according to [59, Chapter VI Theorem 2.9], the quenched Laplace transform of M^ε is given for all measurable and positive functions $f : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ by

$$\mathbb{E}_Z\left[e^{-M^\varepsilon f}\right] = e^{-\tilde{\nu}^\varepsilon(1-e^{-f})}, \tag{17}$$

where the notation $\tilde{\nu}^\varepsilon(1-e^{-f})$ in the above stands for the integral of $(1-e^{-f})$ with respect to the measure $\tilde{\nu}^\varepsilon = \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x)$. Hence, we can rewrite (17) as

$$\begin{aligned}&\mathbb{E}_Z\left[\exp\left\{-\int_{\mathbb{R}_+\times\mathbb{R}_+} M^\varepsilon(\mathrm{d}x, \mathrm{d}y) f(x, y)\right\}\right] \\ &\quad = \exp\left\{-\int_{\mathbb{R}_+\times\mathbb{R}_+} \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x)(1 - e^{-f(x,y)})\right\}.\end{aligned} \tag{18}$$

In addition, according to (13) we have $N^\varepsilon(t) = M^\varepsilon f$ with $f(x, y) = \mathbb{1}_{(x<t<x+y)}$. Plugging this expression into (18), we immediately get relation (16), completing the proof. □

3.2 Limit for $m^\varepsilon(t)$

According to relation (16) in Proposition 3.1 and the fact that $\mu^\varepsilon(\mathrm{d}s) = \lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s$, the (quenched) mean of the random variable $N^\varepsilon(t)$ defined in (13) is given by

$$m^\varepsilon(t) = \int_0^t \left(\lambda(s)\psi(Z_{s/\varepsilon}) \int_{t-s}^\infty \nu(s, \mathrm{d}r)\right) \mathrm{d}s = \int_0^t \lambda(s)\,\psi(Z_{s/\varepsilon})\,\bar{F}_s(t-s)\,\mathrm{d}s, \tag{19}$$

where $\bar{F}_s$ is the tail function given by (7). The following theorem gives a full description of the almost sure asymptotic behavior of $m^\varepsilon(t)$.

Theorem 3.2 *We assume the same setup and notation as in Proposition 3.1. Furthermore, suppose Hypotheses 2.1, 2.5 and 2.6 hold. Then, under the quenched probability $\mathbb{P}_Z$, for any $t > 0$ we have almost surely that*

$$\lim_{\varepsilon \to 0} m^\varepsilon(t) = \bar{m}(t), \tag{20}$$

where $\bar{m}(t)$ is given by

$$\bar{m}(t) = \sigma(t)\bar{\psi}\,, \tag{21}$$

and the quantities $\sigma(t)$ and $\bar{\psi}$ are, respectively, defined by

$$\sigma(t) = \int_0^t \lambda(s)\bar{F}_s(t-s)\mathrm{d}s, \quad \textit{and} \quad \bar{\psi} = \int_{\mathbb{R}^d} \psi(z)\,\pi(\mathrm{d}z). \tag{22}$$

In relation (22), π is the invariant measure of the process Z introduced in Hypothesis 2.1.

Here, we provide two specific examples of the process Z.

Example 3.3 Suppose Z is an Ornstein–Uhlenbeck (OU) process satisfying the SDE $dZ_t = -\theta(\mu - Z_t)dt + \Sigma dW_t$ and $Z_0 = z_0$. It is well known that if $\theta > 0$, the OU process is stationary, ergodic with Gaussian invariant distribution with mean μ and variance $\Sigma^2/2\theta$.

Example 3.4 Suppose Z is a Cox–Ingersoll–Ross (CIR) process satisfying the SDE $dZ_t = \theta(\mu - Z_t)dt + \Sigma\sqrt{Z_t}dW_t$ and $Z_0 = z_0 > 0$. In this case the invariant distribution is Gamma with shape parameter $\alpha = \frac{2\theta\mu}{\Sigma^2}$ and scale parameter $\beta = \frac{2\theta}{\Sigma^2}$.

Proof of Theorem 3.2 Starting from (19) and upon introducing the notation $h(s,t) = \lambda(s)\bar{F}_s(t-s)$, we write $m^\varepsilon(t)$ as

$$m^\varepsilon(t) = \int_0^t h(s,t)\psi(Z_{s/\varepsilon})\,\mathrm{d}s. \tag{23}$$

We then compute, making use of a simple integration by parts,

$$\begin{aligned} m^\varepsilon(t) &= \int_0^t h(s,t)\frac{\mathrm{d}}{\mathrm{d}s}\left(\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r\right)\mathrm{d}s \\ &= h(s,t)\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r\Big|_{s=0}^{s=t} - \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\left(\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r\right)\mathrm{d}s \\ &= A_1^\varepsilon(t) - A_2^\varepsilon(t), \end{aligned} \tag{24}$$

where we have set

$$A_1^\varepsilon(t) = h(t,t)\int_0^t \psi(Z_{r/\varepsilon})\,\mathrm{d}r, \quad \text{and} \quad A_2^\varepsilon(t) = \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\left(\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r\right)\mathrm{d}s. \tag{25}$$

We now treat the limits of $A_1^\varepsilon(t)$ and $A_2^\varepsilon(t)$ separately.

The term $A_1^\varepsilon(t)$ can be analyzed as follows: The elementary change of variables $r := r/\varepsilon$ yields

$$A_1^\varepsilon(t) = h(t,t)t \lim_{\varepsilon\to 0} \frac{\varepsilon}{t}\int_0^{t/\varepsilon} \psi(Z_r)\,\mathrm{d}r.$$

Hence, invoking Hypothesis 2.1 we get

$$\lim_{\varepsilon\to 0} A_1^\varepsilon(t) = h(t,t)t\bar{\psi}, \qquad \mathbb{P}_Z - \text{a.s.} \tag{26}$$

For $A_2^\varepsilon(t)$, we add and subtract $\bar{\psi}$ to get

$$\begin{aligned} A_2^\varepsilon(t) &= \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right) s\left[\frac{1}{s}\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r - \bar{\psi} + \bar{\psi}\right]\mathrm{d}s \\ &= \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right) s\bar{\psi}\,\mathrm{d}s + \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right) s\left[\frac{1}{s}\int_0^s \psi(Z_{r/\varepsilon})\,\mathrm{d}r - \bar{\psi}\right]\mathrm{d}s \\ &\equiv A_{2,1}^\varepsilon(t) + A_{2,2}^\varepsilon(t). \end{aligned} \tag{27}$$

We now proceed to bound the term $A_{2,2}^\varepsilon(t)$ in relation (27). Namely, a trivial integral bound and the same change of variables as for $A_1^\varepsilon(t)$ enable us to write

$$\left|A_{2,2}^\varepsilon(t)\right| \le \int_0^t s\left|\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right|\left|\frac{\varepsilon}{s}\int_0^{s/\varepsilon}\psi(Z_r)\,\mathrm{d}r - \bar{\psi}\right|\mathrm{d}s.$$

Hence, invoking Hypothesis 2.1 we get

$$\left|A_{2,2}^\varepsilon(t)\right| \le C\int_0^t s\left|\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right|(1+s/\varepsilon)^{-\kappa}\,\mathrm{d}s.$$

Next recall that $h(s,t) = \lambda(s)\bar{F}_s(t-s)$. Owing to Hypothesis 2.6, the measure $\nu(s,\cdot)$ admits a density ℓ_s, and thus,

$$\frac{\mathrm{d}}{\mathrm{d}s}h(s,t) = \lambda'(s)\bar{F}_s(t-s) + \lambda(s)\int_{t-s}^\infty \frac{\partial \ell_s(r)}{\partial s}\,\mathrm{d}r + \lambda(s)\ell_s(t-s). \tag{28}$$

Invoking the fact that λ is a Lipschitz function (see Hypothesis 2.5), together with relations (8) and (9), it is readily checked that $|\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)|$ is uniformly bounded by a constant C. Hence, a straightforward application of the dominated convergence theorem yields

$$\lim_{\varepsilon\to 0} A_{2,2}^\varepsilon(t) = 0. \tag{29}$$

We also notice that the term $A^{\varepsilon}_{2,1}(t)$ introduced in (27) can be simplified thanks to an elementary integration by parts. We get

$$A^{\varepsilon}_{2,1}(t) = \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s}h(s,t)\right) s\bar{\psi}\,\mathrm{d}s = h(t,t)t\bar{\psi} - \bar{\psi}\int_0^t h(s,t)\,\mathrm{d}s, \tag{30}$$

the right-hand side of which is finite. Therefore, plugging (30) and (29) into relation (27), we have obtained

$$\lim_{\varepsilon\to 0} A^{\varepsilon}_2(t) = h(t,t)t\bar{\psi} - \bar{\psi}\int_0^t h(s,t)\,\mathrm{d}s. \tag{31}$$

We can now conclude as follows: Gathering (26) and (31) into (24), we have

$$\lim_{\varepsilon\to 0} m^{\varepsilon}(t) = \bar{\psi}\int_0^t h(s,t)\,\mathrm{d}s, \quad \mathbb{P}_Z - \text{a.s.},$$

which is exactly our claim (20). This completes the proof. □

Remark 3.5 A brief remark on the proof. Observe that the whole set of assumptions in Hypotheses 2.1, 2.5 and 2.6 is effectively invoked in order to achieve the proof of Theorem 3.2. Indeed, the ergodic properties of Z contained in Hypothesis 2.1 are appealed to, for example, in equation (26). The ergodic character of Z is crucial for our homogenization-type approach. Next, the regularity we suppose for λ and ℓ (respectively, Hypotheses 2.5 and 2.6) is used for the analysis of (28). One way to weaken some of those regularity assumptions would be to avoid the integration by parts method employed for Theorem 3.2. However, this would come at the price of stronger (and probably not realistic) ergodic hypotheses on Z. This point will be detailed in Sect. 3.4.

3.3 Homogenized process

With the limiting behavior of m^{ε} in hand, we are now ready to give the asymptotic description of the process N^{ε}. As usual, we will decompose this analysis into a study of the convergence of the finite dimensional distributions and a tightness result.

3.3.1 Limit for the finite-dimensional distributions

In order to simplify our presentation, we will first derive the limit of bivariate quantities of the form $(N^{\varepsilon}(t_1), N^{\varepsilon}(t_2))$ for two instants $t_1 < t_2$. To this aim, inside the quadrant $\mathbb{R}_+ \times \mathbb{R}_+$, we will consider three disjoint regions $\{A_i,\ i = 1, 2, 3\}$ defined as follows (see Fig. 1):

$$A_1 = \{(\gamma, l) \in \mathbb{R}_+ \times \mathbb{R}_+ : \gamma \le t_1 \text{ and } t_1 < \gamma + l \le t_2\}; \tag{32}$$

$$A_2 = \{(\gamma, l) \in \mathbb{R}_+ \times \mathbb{R}_+ : \gamma \le t_1 \text{ and } t_2 < \gamma + l\}; \tag{33}$$

$$A_3 = \{(\gamma, l) \in \mathbb{R}_+ \times \mathbb{R}_+ : t_1 < \gamma \le t_2 \text{ and } t_2 < \gamma + l\}. \tag{34}$$

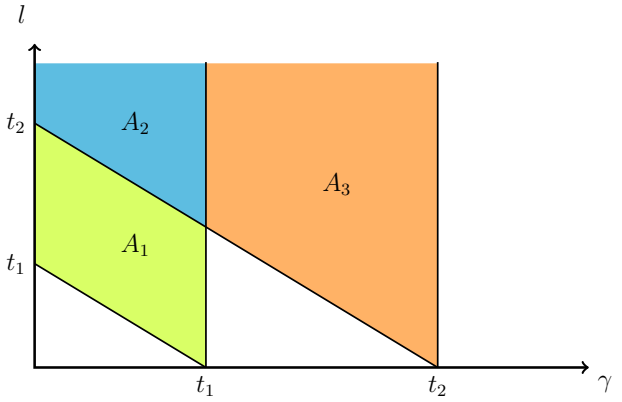

Fig. 1 Three disjoint regions used for the limit of bivariate quantities. Figure adopted from [21]

Notice that since the A_i's are disjoint, the quantities $\{M^\varepsilon(A_i);\ i = 1, 2, 3\}$ are independent Poisson random variables. Similar to the proof of (16), their respective quenched means are given by (see Proposition 3.1)

$$m_i^\varepsilon := \mathbb{E}_Z[M^\varepsilon(A_i)] = \int_{A_i} \nu(s, \mathrm{d}r)\lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s, \quad \text{for } i = 1, 2, 3. \tag{35}$$

Now observe that the vector $(N^\varepsilon(t_1), N^\varepsilon(t_2))$ can be decomposed as

$$N^\varepsilon(t_1) = M^\varepsilon(A_1) + M^\varepsilon(A_2), \qquad N^\varepsilon(t_2) = M^\varepsilon(A_2) + M^\varepsilon(A_3). \tag{36}$$

As a consequence, the means $m^\varepsilon(t_1)$, $m^\varepsilon(t_2)$ can also be written in terms of the m_i^ε's given by (35):

$$m^\varepsilon(t_1) = m_1^\varepsilon + m_2^\varepsilon, \qquad m^\varepsilon(t_2) = m_2^\varepsilon + m_3^\varepsilon. \tag{37}$$

We now state a proposition giving the quenched limit in law for $(N^\varepsilon(t_1), N^\varepsilon(t_2))$.

Proposition 3.6 *Let M^ε be the Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}_+$ defined by (12), with mean measure $\tilde{\nu}^\varepsilon(\mathrm{d}s, \mathrm{d}r) = \nu(s, \mathrm{d}r)\mu^\varepsilon(\mathrm{d}s)$, as given in (15). Assume that the conditions in Hypotheses 2.1, 2.5 and 2.6 are met. Then, for any two fixed time points $0 \le t_1 < t_2$, we have the following statements:*

(i) *$\mathbb{P}_Z$-almost surely we have that, for all $\xi_1, \xi_2 > 0$,*

$$\lim_{\varepsilon \to 0} \mathbb{E}_Z[e^{-(\xi_1 N^\varepsilon(t_1) + \xi_2 N^\varepsilon(t_2))}] = \mathbb{E}[e^{-(\xi_1 N(t_1) + \xi_2 N(t_2))}]. \tag{38}$$

In the right-hand side of Eq. (38), the process $\{N(t);\ t \ge 0\}$ is independent of N^ε and is defined similarly to that of (13), albeit in a nonrandom environment. More

precisely, N can be expressed as

$$N(t) = \sum_{k=1}^{\infty} \mathbb{1}_{\{\Gamma_k < t < \Gamma_k + L_k\}} = M\{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+, x < t < x + y\}, \quad (39)$$

with M being a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}_+$ of the form $M = \sum_{k=1}^{\infty} \delta_{(\Gamma_k, L_k)}$. The mean measure of M is given by

$$\tilde{\nu}(\mathrm{d}s, \mathrm{d}r) = \bar{\psi}\lambda(s)\nu(s, \mathrm{d}r)\mathrm{d}s, \quad (40)$$

where we recall that $\bar{\psi}$ is defined by (22).

(ii) *$\mathbb{P}_Z$-almost surely we have the following limit in law as $\varepsilon \to 0$:*

$$(N^\varepsilon(t_1), N^\varepsilon(t_2)) \xrightarrow{(\mathrm{d})} (N(t_1), N(t_2)).$$

Remark 3.7 In order to alleviate notation, we have assumed that our underlying probability space carries the family $\{N^\varepsilon;\ \varepsilon \geq 0\}$ as well as the process N defined by (39). This explains why we have expressed (38) with the same expectation $\mathbb{E}$ on both sides of the relation.

Proof of Proposition 3.6 We first introduce the notation that will be used in the proof. For $i = 1, 2, 3$, and $\lambda_i > 0$, we define the functions $f_i := \lambda_i \mathbb{1}_{A_i}$, where the sets A_i are given by (32)-(34). We note that $M^\varepsilon(f_i) = \lambda_i M^\varepsilon(A_i)$. By considering the Laplace transform of $M^\varepsilon(A_i)$, we can see that Theorem 2.9 in Chapter VI of [59] yields

$$\mathbb{E}_Z[e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}] = \mathbb{E}_Z[e^{-\sum_{i=1}^3 M^\varepsilon(f_i)}] = e^{-\sum_{i=1}^3 \tilde{\nu}^\varepsilon(1-e^{-f_i})}, \quad (41)$$

where $\tilde{\nu}^\varepsilon$ is the measure defined by (15). In equation (41), we specify again that $\tilde{\nu}^\varepsilon(1 - e^{-f_i})$ stands for the integral of $(1 - e^{-f_i})$ with respect to the measure $\tilde{\nu}^\varepsilon$, as with (17). We now split the analysis of (41) into several steps.

Step 1: Decomposition of the Laplace transform. Taking into account the expression (15) for $\tilde{\nu}^\varepsilon$, the right-hand side of (41) can be rewritten as

$$e^{-\sum_{i=1}^3 \tilde{\nu}^\varepsilon(1-e^{-f_i})} = \prod_{i=1}^3 \exp\left\{-\tilde{\nu}^\varepsilon(1 - e^{-f_i})\right\} = \prod_{i=1}^3 \exp\left\{-G_i^\varepsilon\right\}, \quad (42)$$

where each function G_i^ε, for $i = 1, 2, 3$, is given as the following integral:

$$\begin{aligned} G_i^\varepsilon &= \int_{\mathbb{R}_+ \times \mathbb{R}_+} (1 - e^{-f_i})\nu(s, \mathrm{d}r)\mu^\varepsilon(\mathrm{d}s) \\ &= \int_{\mathbb{R}_+ \times \mathbb{R}_+} (1 - e^{-f_i})\nu(s, \mathrm{d}r)\lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s. \end{aligned} \quad (43)$$

In the sequel, we shall characterize the limit of each G_i^ε. Note first that, owing to the relation $f_i = \lambda_i \mathbb{1}_{A_i}$, we have

$$1 - e^{-f_i(s,r)} = (1 - e^{-\lambda_i})\mathbb{1}_{A_i}(s,r).$$

Therefore, one can recast the term G_i^ε as

$$G_i^\varepsilon = (1 - e^{-\lambda_i})m_i^\varepsilon, \quad \text{where we recall that } m_i^\varepsilon = \int_{A_i} \nu(s, \mathrm{d}r)\lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s. \tag{44}$$

In summary, substituting (44) into (42) and then (41), we have shown that

$$\mathbb{E}_Z[e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}] = \prod_{i=1}^3 \exp\left\{-(1 - e^{-\lambda_i})m_i^\varepsilon\right\}. \tag{45}$$

We are now reduced to an examination of $\lim_{\varepsilon\to 0} m_i^\varepsilon$ in the right-hand side of equation (45).

Step 2: Analysis of an integral with fast oscillatory integrand. In order to handle the terms m_i^ε in (44), we will generalize slightly the analysis of integral expressions like (23). Namely, consider a continuously differentiable function $g : \mathbb{R}_+ \to \mathbb{R}$ and for $\varepsilon > 0$, let $I^\varepsilon(\tau_1, \tau_2)$ be given in the following integral:

$$I^\varepsilon(\tau_1, \tau_2) = \int_0^{\tau_1} g(s, \tau_2)\psi\left(Z_{s/\varepsilon}\right) \mathrm{d}s, \tag{46}$$

for $0 \le \tau_1 \le \tau_2$. Then, following the same integration by parts procedure as for (24) in the proof of Theorem 3.2, we have

$$\begin{aligned} I^\varepsilon(\tau_1, \tau_2) &= \int_0^{\tau_1} g(s, \tau_2)\left(\frac{\mathrm{d}}{\mathrm{d}s}\int_0^s \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r\right) \mathrm{d}s \\ &= g(s, \tau_2)\int_0^s \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r\Big|_{s=0}^{s=\tau_1} - \int_0^{\tau_1} \frac{\mathrm{d}}{\mathrm{d}s} g(s, \tau_2)\left(\int_0^s \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r\right) \mathrm{d}s \\ &= g(\tau_1, \tau_2)\int_0^{\tau_1} \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r - \int_0^{\tau_1} \frac{\mathrm{d}}{\mathrm{d}s} g(s, \tau_2)\left(\int_0^s \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r\right) \mathrm{d}s. \end{aligned} \tag{47}$$

Then, following the same steps as for the analysis of $A_1^\varepsilon(t)$ and $A_2^\varepsilon(t)$ in the proof of Theorem 3.2 (see, respectively, (26) and (31)), we compute the limit of $I^\varepsilon(\tau_1, \tau_2)$ as

$\varepsilon \to 0$:

$$\begin{aligned}\lim_{\varepsilon\to 0} I^{\varepsilon}(\tau_1,\tau_2) &= g(\tau_1,\tau_2)\tau_1\bar{\psi} - \int_0^{\tau_1}\left(\frac{\mathrm{d}}{\mathrm{d}s}g(s,\tau_2)\right)s\bar{\psi}\,\mathrm{d}s\\ &= \left[g(\tau_1,\tau_2)\tau_1 - g(s,\tau_2)s\Big|_{s=0}^{s=\tau_1} + \int_0^{\tau_1} g(s,\tau_2)\,\mathrm{d}s\right]\bar{\psi}\\ &= \left[\int_0^{\tau_1} g(s,\tau_2)\,\mathrm{d}s\right]\bar{\psi}, \end{aligned} \tag{48}$$

where the second equality is obtained thanks to another use of the integration by parts formula.

Step 3: Analysis of m_i^{ε}, for $i = 1, 2, 3$. We will now resort to the general identity (48) in order to analyze the terms m_i^{ε} in (44). We start by recasting m_1^{ε} as an expression involving (46). Namely, invoking the definition (7) of $\bar{F}_s$, we write

$$\begin{aligned} m_1^{\varepsilon} &= \int_0^{t_1}\left(\lambda(s)\psi(Z_{s/\varepsilon})\int_{t_1-s}^{t_2-s}\nu(s,\mathrm{d}r)\right)\mathrm{d}s\\ &= \int_0^{t_1}\lambda(s)\psi(Z_{s/\varepsilon})\Big(\bar{F}_s(t_1-s) - \bar{F}_s(t_2-s)\Big)\mathrm{d}s\\ &= \int_0^{t_1}\lambda(s)\bar{F}_s(t_1-s)\psi(Z_{s/\varepsilon})\,\mathrm{d}s - \int_0^{t_1}\lambda(s)\bar{F}_s(t_2-s)\psi(Z_{s/\varepsilon})\,\mathrm{d}s. \end{aligned} \tag{49}$$

Upon setting

$$g(s,t) = \lambda(s)\bar{F}_s(t-s), \tag{50}$$

and recalling (46), we can rewrite the expression (51) of m_1^{ε} as follows:

$$m_1^{\varepsilon} = \int_0^{t_1} g(s,t_1)\psi(Z_{s/\varepsilon})\mathrm{d}s - \int_0^{t_1} g(s,t_2)\psi(Z_{s/\varepsilon})\mathrm{d}s = I^{\varepsilon}(t_1,t_1) - I^{\varepsilon}(t_1,t_2). \tag{51}$$

Due to the fact that λ and $\bar{F}_s$ are continuous and bounded functions, we can apply directly the result (48) from Step 2. We get the following $\mathbb{P}_Z$-almost sure limit for m_1^{ε}:

$$\begin{aligned}\lim_{\varepsilon\to 0} m_1^{\varepsilon} &= \left[\int_0^{t_1} g(s,t_1)\,\mathrm{d}s\right]\bar{\psi} - \left[\int_0^{t_1} g(s,t_2)\,\mathrm{d}s\right]\bar{\psi}\\ &= \left[\int_0^{t_1}\left[g(s,t_1) - g(s,t_2)\right]\,\mathrm{d}s\right]\bar{\psi} \equiv m_1, \end{aligned} \tag{52}$$

where we recall that the function g is given by (50).

The analysis of m_2^{ε} and m_3^{ε} is obtained along similar lines. Hence, we will just write down the main steps and invite the patient reader to fill in the corresponding details.

First, we have the following expressions:

$$m_2^\varepsilon = \int_0^{t_1} \lambda(s)\bar{F}_s(t_2 - s)\psi(Z_{s/\varepsilon})\mathrm{d}s\,;$$
$$m_3^\varepsilon = \int_0^{t_2} \lambda(s)\bar{F}_s(t_2 - s)\psi(Z_{s/\varepsilon})\mathrm{d}s - \int_0^{t_1} \lambda(s)\bar{F}_s(t_2 - s)\psi(Z_{s/\varepsilon})\mathrm{d}s. \quad (53)$$

Then, we can obtain a $\mathbb{P}_Z$-almost sure limit of the form

$$\lim_{\varepsilon\to 0} m_2^\varepsilon = \left[\int_0^{t_1} g(s,t_2)\,\mathrm{d}s\right]\bar{\psi} \equiv m_2\,; \quad (54)$$
$$\lim_{\varepsilon\to 0} m_3^\varepsilon = \left[\int_0^{t_2} g(s,t_2)\,\mathrm{d}s\right]\bar{\psi} - \left[\int_0^{t_1} g(s,t_2)\,\mathrm{d}s\right]\bar{\psi} = \left[\int_{t_1}^{t_2} g(s,t_2)\,\mathrm{d}s\right]\bar{\psi} \equiv m_3. \quad (55)$$

Hence, gathering (52), (54) and (55) into relation (45), we have obtained

$$\lim_{\varepsilon\to 0} \mathbb{E}_Z[e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}] = \prod_{i=1}^3 \exp\left\{-(1-e^{-\lambda_i})m_i\right\}. \quad (56)$$

Step 4: Final concluding steps. Recall that our aim is to analyze the left-hand side of relation (38). To this aim, notice that, owing to relation (36), we have

$$\mathbb{E}_Z[e^{-(\xi_1 N^\varepsilon(t_1)+\xi_2 N^\varepsilon(t_2))}] = \mathbb{E}_Z[e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}],$$

where we have set

$$\lambda_1 = \xi_1, \quad \lambda_2 = \xi_1 + \xi_2, \quad \text{and} \quad \lambda_3 = \xi_2. \quad (57)$$

Therefore, an immediate application of (56) yields

$$\lim_{\varepsilon\to 0} \mathbb{E}_Z\left[e^{-(\xi_1 N^\varepsilon(t_1)+\xi_2 N^\varepsilon(t_2))}\right] = \prod_{i=1}^3 \exp\left\{-(1-e^{-\lambda_i})m_i\right\}. \quad (58)$$

By the definition $f_i := \lambda_i \mathbb{1}_{A_i}$ given at the beginning of our proof and the expressions (52), (54) and (55) for m_1, m_2 and m_3, respectively, relation (58) can be rewritten as

$$\lim_{\varepsilon\to 0} \mathbb{E}_Z\left[e^{-(\xi_1 N^\varepsilon(t_1)+\xi_2 N^\varepsilon(t_2))}\right] = e^{-\sum_{i=1}^3 \tilde{\nu}(1-e^{-f_i})}, \quad (59)$$

where $\tilde{\nu}$ stands for the mean measure of the point process M (see equation (40)). Taking into account the values (57) for $\lambda_1, \lambda_2, \lambda_3$ and the definition (39) of the process N, then by Chapter VI Theorem 2.9 in [59] we end up with

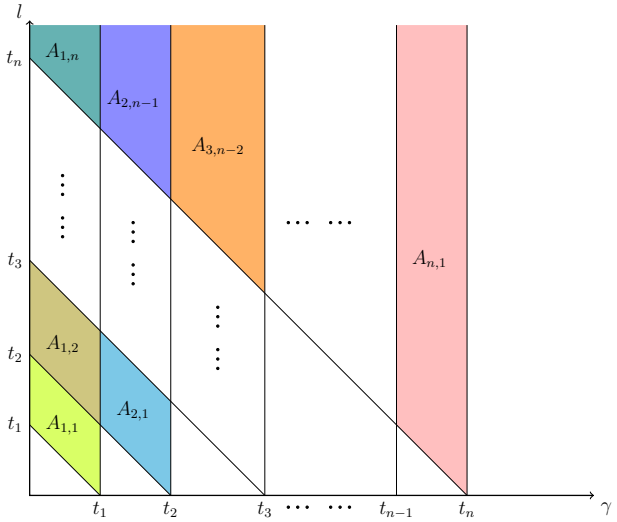

Fig. 2 Disjoint regions used for the limit of finite dimensional distributions

$$e^{-\sum_{i=1}^{3}\tilde{\nu}(1-e^{-f_i})} = \mathbb{E}\left[e^{-\sum_{i=1}^{3}M(f_i)}\right] = \mathbb{E}\left[e^{-\sum_{i=1}^{3}\lambda_i M(A_i)}\right] = \mathbb{E}\left[e^{-(\xi_1 N(t_1)+\xi_2 N(t_2))}\right],$$

which leads to our claim (38). This proves statement *(i)*.

Finally, in light of statement *(i)* and Laplace transform properties, statement *(ii)* follows. □

Once the limit for the bivariate vector $(N^\varepsilon(t_1), N^\varepsilon(t_2))$ is obtained, the extension to the multivariate case can be done through routine (though tedious) considerations. We state this generalization and a sketch of its proof below.

Corollary 3.8 *With the assumptions in Proposition 3.6, let $t_1 < t_2 < \cdots < t_n$ be fixed. Then, $\mathbb{P}_Z$-almost surely we have that*

$$(N^\varepsilon(t_1), N^\varepsilon(t_2), \ldots, N^\varepsilon(t_n)) \xrightarrow{\text{(d)}} (N(t_1), N(t_2), \ldots, N(t_n)).$$

Proof Following the same procedures as in the proof of Proposition 3.6, we first decompose the quadrant $\mathbb{R}_+ \times \mathbb{R}_+$ into several disjoint regions (see Figure 2 for a depiction of the sets $A_{i,j}$):

$$\{A_{i,j} : \text{for } 1 \le i \le n \text{ and } 1 \le j \le n-i+1\},$$

where for each i and j, the region is defined as

$$A_{i,j} = \left\{(\gamma, l) \in \mathbb{R}_+ \times \mathbb{R}_+ : t_{i-1} \le \gamma \le t_i \text{ and } t_j \le \gamma + l \le t_{j+1}\right\},$$

with the additional convention $t_0 = 0$ and $t_{n+1} = \infty$. Since the $A_{i,j}$'s are disjoint regions, the quantities $\{M^\varepsilon(A_{i,j}) : i, j = 1, 2, ..., n\}$ are independent Poisson random variables. Similar to (35), their respective quenched means are given by

$$\mathbb{E}_Z[M^\varepsilon(A_{i,j})] = \int_{A_{i,j}} \nu(s, \mathrm{d}r)\lambda(s)\psi(Z_{s/\varepsilon})\mathrm{d}s.$$

Furthermore, as with (36) and (37), the quantity $N^\varepsilon(t_k)$ can be written as

$$N^\varepsilon(t_k) = \sum_{i=1}^{k} \sum_{j=k-i+1}^{n-i+1} M^\varepsilon(A_{i,j}),$$

whose mean can be expressed as

$$\mathbb{E}_Z[N^\varepsilon(t_k)] = m^\varepsilon(t_k) = \sum_{i=1}^{k} \sum_{j=k-i+1}^{n-i+1} \mathbb{E}_Z[M^\varepsilon(A_{i,j})].$$

Starting from this set of relations, the rest of the proof will be a repetition of that for Proposition 3.6. We omit the details for the sake of conciseness. □

3.3.2 Tightness and homogenization results

In this section, we shall summarize our previous considerations about the limiting behavior of our queueing system. This will yield the homogenization results stated in the introduction. The next natural step in establishing our limiting description of the family $\{N^\varepsilon; \varepsilon > 0\}$ is a tightness result. Due to the expression (13) for N^ε, it is natural to consider this process (restricted on the interval $[0, T]$) as an element of the following space:

$$D_T = \{f : [0, T] \to \mathbb{R}_+;\ \ f \text{ right-continuous with left limits}\}. \tag{60}$$

The Borel σ-field of D_T defined with respect to the Skorokhod topology will be denoted by $\mathcal{D}$. Then, according to [60, Proposition 4.2], the tightness of $\{N^\varepsilon\ ;\ \varepsilon > 0\}$ in the space D_T stems from the following criterion.

Proposition 3.9 *Let X and X_n, $n \in \mathbb{N}$, be random variables in $(D_T, \mathcal{D})$. Suppose that*

(1) *for any k and instants $t_1 < \cdots < t_k$, $(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{(\mathrm{d})} (X(t_1), \ldots, X(t_k))$,*
(2) *X has jumps of size ± 1, and*
(3) *X_n has integer-valued jumps.*

Then, the sequence $\{X_n : n \geq 1\}$ is tight.

Now we state a proposition about the tightness of $\{N^\varepsilon; \varepsilon > 0\}$ in the space D_T.

Proposition 3.10 *Let $\{N^\varepsilon\ ;\ \varepsilon > 0\}$ be the sequence of Poisson processes defined by (13), which belongs to the space D_T. Then, $\mathbb{P}_Z$-almost surely $\{N^\varepsilon\ ;\ \varepsilon > 0\}$ is tight.*

Proof We observe that $N^\varepsilon(t)$ defined by (13) and $N(t)$ defined by (39) have intrinsically jumps of size ± 1. In the light of Corollary 3.8, the tightness of the sequence $\{N^\varepsilon\,;\,\varepsilon > 0\}$ is a direct consequence of Proposition 3.9. □

We can now state our main quenched limit theorem for the process N^ε. Given the preparation already done, it is an immediate consequence of Corollary 3.8 and Proposition 3.10.

Theorem 3.11 *Consider an arbitrary time horizon $T > 0$. We assume that Hypotheses 2.1, 2.5 and 2.6 are verified. Recall that the processes N^ε and N are, respectively, defined by (13) and (39), and the functional space D_T is introduced in (60). Then, as $\varepsilon \to 0$, $\mathbb{P}_Z$-almost surely the following limit in distribution holds true in D_T:*

$$\{N^\varepsilon(t) : t \in [0,T]\} \xrightarrow{\text{(d)}} \{N(t) : t \in [0,T]\}.$$

3.3.3 Limiting description of associated processes

We now state some consequences of Theorem 3.11 that are of interest in practice.

We first consider the total accumulated input on the interval $[0,t]$, defined by

$$A^\varepsilon(t) = \int_0^t N^\varepsilon(s)\mathrm{d}s. \tag{61}$$

The process A^ε is a continuous function on $[0,T]$, due to the fact that $N^\varepsilon \in D_T$. This continuous quantity models a stochastic fluid input to a queueing system. The limiting behavior of A^ε is summarized in the following proposition.

Proposition 3.12 *Let C_T be the space of continuous functions from $[0,T]$ to $\mathbb{R}$. With the same assumptions of Theorem 3.11, the input process $A^\varepsilon(t)$ defined by (61) converges in distribution. More precisely, $\mathbb{P}_Z$-almost surely we have the following limit in law in the space C_T as $\varepsilon \to 0$:*

$$\left\{A^\varepsilon(t) = \int_0^t N^\varepsilon(s)\mathrm{d}s : t \in [0,T]\right\} \xrightarrow{\text{(d)}} \left\{A := \int_0^t N(s)\mathrm{d}s : t \in [0,T]\right\},$$

where N is the process given by (39).

Proof Let $\phi : D_T \to C_T$ be defined by

$$[\phi(f)]_t = \int_0^t f(s)\mathrm{d}s, \quad \text{for} \quad f \in D_T.$$

It is readily checked that ϕ is a continuous function. Since $N^\varepsilon \xrightarrow{\text{(d)}} N$ in D_T, we get that

$$A^\varepsilon = \phi(N^\varepsilon) \xrightarrow{\text{(d)}} \phi(N) =: A, \quad \text{in the space } C_T.$$

This completes our proof. □

Another useful corollary of our main Theorem 3.11 is the following. The quantity $A^\varepsilon(t)$ defined by (61) may be treated as a Cox/G_t/∞-input for another single-server queue. The state of the single server queue, $X^\varepsilon(t)$, satisfies the following storage equation driven by $dA^\varepsilon(t) = N^\varepsilon(t)dt$:

$$\mathrm{d}X^\varepsilon(t) = N^\varepsilon(t)\mathrm{d}t - r\mathbb{1}_{\{X^\varepsilon(t)>0\}}\mathrm{d}t, \quad X^\varepsilon(0) = 0,$$

where we assume that the server works at constant rate r. Thanks to Skorohod's lemma on reflected processes (see, for example, [61, Theorem 6.1]), the application $N^\varepsilon \to X^\varepsilon$ is continuous from D_T to C_T. Therefore, we obtain a limiting behavior for X^ε as follows:

Proposition 3.13 *Let the assumptions of Theorem 3.11 prevail. Then, $\mathbb{P}_Z$-almost surely, X^ε converges in distribution to a process X, such that X solves the following equation driven by N:*

$$\mathrm{d}X(t) = N(t)\mathrm{d}t - r\mathbb{1}_{\{X(t)>0\}}\mathrm{d}t,$$

with the initial condition $X(0) = 0$.

3.4 A more general stochastic intensity model

Thus far, we have assumed that the stochastic intensity model is of product form: $\lambda(s)\psi(Z_{s/\varepsilon})$. This made our presentation simpler and covers most cases of interest in practice. We now demonstrate that our main result Theorem 3.11 can be extended to more general stochastic intensity models of the form $\Psi(s, Z_s)$, with $\Psi : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}_+$. For simplicity and just to illustrate the idea, we still assume Ψ to be bounded and uniformly continuous in both variables.

It is fairly straightforward to show that an analogue of Proposition 3.1 holds in our context with a general stochastic intensity. We thus state the following result without proof for further use.

Proposition 3.14 *Let M^ε and $\{N^\varepsilon(t) : t \geq 0\}$ be defined by (12) and (13), respectively. Instead of Hypothesis 2.5, we assume that the stochastic intensity of the arrival times $\{\Gamma_k^\varepsilon ; k \geq 1\}$ takes the form*

$$\Psi(s, Z_s), \quad \textit{with} \quad \Psi : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}_+ \quad \textit{uniformly continuous.}$$

Then, under the quenched probability $\mathbb{P}_Z$, M^ε is a Poisson random measure with mean measure given by

$$\tilde{\nu}^\varepsilon(\mathrm{d}x, \mathrm{d}y) = \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x), \quad \textit{with} \quad \mu^\varepsilon(\mathrm{d}s) = \Psi(s, Z_{s/\varepsilon})\mathrm{d}s,$$

where we recall that ν is introduced in (6). Furthermore, we have that for any $t > 0$, $N^\varepsilon(t)$ is a Poisson random variable with parameter

$$m^\varepsilon(t) = \int_{\{(x,y):x<t<x+y\}} \nu(x, \mathrm{d}y)\mu^\varepsilon(\mathrm{d}x) = \int_0^t \Psi(s, Z_{s/\epsilon})\bar{F}_s(t-s)\mathrm{d}s.$$

With Proposition 3.14 in hand, we can now state the analogue of Theorem 3.2 within our generalized framework.

Theorem 3.15 *Let the assumptions of Proposition 3.14 prevail. We also suppose that Hypothesis 2.6 holds without change, while Hypothesis 2.1 holds with $\psi(z) = \Psi(t, z)$ for each $t \geq 0$ but the constant C in (3) does not depend on the initial condition of Z. Then, under the quenched probability $\mathbb{P}_Z$, for any $t > 0$ we have almost surely*

$$\lim_{\varepsilon\to 0} m^\varepsilon(t) = \bar{m}(t), \quad \textit{where} \quad \bar{m}(t) = \int_0^t \mathbb{E}[\Psi(s, \bar{Z})]\bar{F}_s(t-s)\mathrm{d}s. \tag{62}$$

Proof Looking carefully at the proof of Theorem 3.2, it is easily conceived [see, for example, (25)] that the key to obtaining our results lies in the analysis of

$$\lim_{\varepsilon\to 0} \frac{1}{t}\int_0^t \Psi(s, Z_{s/\varepsilon})\mathrm{d}s. \tag{63}$$

The following argument shows how the results obtained in the previous sections can be replicated in this general case.

We will employ the Riemann sum approximation of the integral in (63) with uniform partition $0 = t_0 < t_1 < \cdots < t_n = t$ as follows:

$$\begin{aligned}
\int_0^t \Psi(s, Z_{s/\varepsilon})\mathrm{d}s &= \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} \Psi(s, Z_{s/\varepsilon})\mathrm{d}s \\
&= \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} \Big[\Psi(s, Z_{s/\varepsilon}) - \Psi(t_i, Z_{s/\varepsilon})\Big]\mathrm{d}s \\
&\quad + \sum_{i=0}^{n-1}\left[\int_{t_i}^{t_{i+1}} \Psi(t_i, Z_{s/\varepsilon})\mathrm{d}s - (t_{i+1}-t_i)\mathbb{E}\big[\Psi(t_i, \bar{Z})\big]\right] \\
&\quad + \sum_{i=0}^{n-1}(t_{i+1}-t_i)\mathbb{E}\big[\Psi(t_i, \bar{Z})\big],
\end{aligned}$$

where $\bar{Z}$ is the stationary solution corresponding to the process Z. The above procedure essentially freezes the (slow) time variable s to the discrete epochs t_i. We now analyze the three summations in order to obtain the limit in (63).

For the first summation, by the uniform continuity of Ψ in the first variable, we have for each ε fixed that

$$\lim_{n\to\infty}\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}}\Big(\Psi(s,Z_{s/\varepsilon})-\Psi(t_i,Z_{s/\varepsilon})\Big)\mathrm{d}s=0.$$

For the second summation, we compute for each n fixed that

$$\begin{aligned}&\lim_{\varepsilon\to 0}\sum_{i=0}^{n-1}\left[\int_{t_i}^{t_{i+1}}\Psi(t_i,Z_{s/\varepsilon})\mathrm{d}s-(t_{i+1}-t_i)\mathbb{E}\big[\Psi(t_i,\bar{Z})\big]\right]\\&=\lim_{\varepsilon\to 0}\sum_{i=0}^{n-1}(t_{i+1}-t_i)\left[\frac{1}{\frac{t_{i+1}-t_i}{\varepsilon}}\int_{\frac{t_i}{\varepsilon}}^{\frac{t_{i+1}}{\varepsilon}}\Psi(t_i,Z_s)\mathrm{d}s-\mathbb{E}\big[\Psi(t_i,\bar{Z})\big]\right].\end{aligned}\tag{64}$$

The right-hand side above is easily shown to converge to 0, $\mathbb{P}_Z$-almost surely, thanks to our ergodic assumptions in Hypothesis 2.1. Summarizing our considerations so far, the limit in (63) can be written as

$$\begin{aligned}&\lim_{\varepsilon\to 0}\int_0^t\Psi(s,Z_{s/\varepsilon})\mathrm{d}s\\&=\lim_{n\to\infty}\sum_{i=0}^{n-1}(t_{i+1}-t_i)\mathbb{E}\big[\Psi(t_i,\bar{Z})\big]=\int_0^t\mathbb{E}\big[\Psi(s,\bar{Z})\big]\mathrm{d}s\\&\equiv\int_0^t\left(\int\Psi(s,y)\pi(\mathrm{d}y)\right)\mathrm{d}s.\end{aligned}$$

This proves the equivalent of (26)–(30)–(31) in our context. We can now show our claim (62) easily, following the steps of Theorem 3.2. □

Remark 3.16 On top of providing a more general result than Theorem 3.2, it seems at first sight that Theorem 3.15 comes with a simpler proof avoiding integration by parts procedures (see Remark 3.5). Also notice that we have not imposed differentiability conditions on Ψ, which is another contrast with Theorem 3.2. However, we have hidden a subtle point in the analysis of (64) for the sake of clarity. Specifically, in order to get the relation

$$\lim_{\varepsilon\to 0}\left[\frac{1}{\frac{t_{i+1}-t_i}{\varepsilon}}\int_{\frac{t_i}{\varepsilon}}^{\frac{t_{i+1}}{\varepsilon}}\Psi(t_i,Z_s)\mathrm{d}s-\mathbb{E}\big[\Psi(t_i,\bar{Z})\big]\right]=0,$$

uniformly in $n\geq 1$ and $i\leq n-1$, we need a strengthened version of Hypothesis 2.1. Namely, we have to impose that relation (3) holds true with a random variable C which does not depend on the initial condition of Z. This extra assumption is not satisfied in the cases mentioned in Remarks 2.2 and 2.3, which justifies the integration by parts method in the proof of Theorem 3.2.

Remark 3.17 Starting from Theorem 3.2, the asymptotic convergence result in Theorem 3.11 can also be generalized so that the corresponding limit process $\{N(t) : t \in [0, T]\}$ is a Poisson point process with mean intensity $\bar{m}(t)$. More importantly, observe that yet again, there is a well-defined homogenization limit due to a timescale separation.

4 Annealed analysis

As mentioned above, our main focus in the current contribution is the quenched regime. Nevertheless, an annealed analysis of our system (that is averaging over both the Poisson and the environment randomness) is also in order. We will give some details about the annealed regime in this section. Notice that for the annealed probability we will follow the notation introduced in Remark 2.9, in particular (14).

In the annealed situation, we will be able to relax somewhat the hypotheses on the environment process Z, allowing the convergence in (3) to hold in $L^1(\Omega)$ only. This is summarized in Hypothesis 2.4. As a prelude to the precise statement of result and its proof, we point out that due to our ergodicity assumption on the process Z, we obtain a *deterministic limit* for the quenched regime. Then, it is natural that under appropriate integrability condition on Z again, the "typical" description of the queue obtained in the annealed setting is the same as the quenched one.

Under Hypothesis 2.4, we will now show the following annealed convergence in distribution for the process N^ε.

Theorem 4.1 *Let T be a positive number, and suppose Hypothesis 2.5 to 2.6 hold true. Then, for the processes N^ε and N, respectively, defined by (13) and (39), the following limit in distribution holds true in D_T with respect to the annealed probability $\mathbb{P}$:*

$$\left\{N^\varepsilon(t)\,;\ t \in [0, T]\right\} \xrightarrow{\text{(d)}} \{N(t)\,;\ t \in [0, T]\}\,.$$

Proof For the sake of conciseness, we will only focus on the convergence of bivariate distributions. Namely, we will show that for two fixed time points $0 \le t_1 \le t_2 \le T$, we have that

$$\left(N^\varepsilon(t_1), N^\varepsilon(t_2)\right) \xrightarrow{\text{(d)}} (N(t_1), N(t_2))\,. \tag{65}$$

The steps that allow us to go from (65) to a convergence of finite-dimensional distributions and then to a convergence of processes is very similar to what we did in Sect. 3.3. Details are thus left to the patient reader.

In order to prove (65), we will take advantage of our computations for the bivariate quenched case that has been presented in Sect. 3.3.1. Let us refer to Fig. 1 and consider the same three disjoint regions $\{A_i,\ i = 1, 2, 3\}$ defined in the relations (32)–(34). We basically follow the same strategy as in Proposition 3.6. Hence, similarly as (41), the

annealed Laplace transform of $M^\varepsilon(A_i)$ is then given by

$$\begin{aligned}\mathbb{E}\left\{e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}\right\} &= \mathbb{E}\left\{\mathbb{E}_Z[e^{-\sum_{i=1}^3 M^\varepsilon(f_i)}]\right\} \\ &= \mathbb{E}\left[e^{-\sum_{i=1}^3 \tilde{\nu}^\varepsilon(1-e^{-f_i})}\right], \quad \text{for} \quad f_i = \lambda_i \mathbb{1}_{A_i}.\end{aligned}$$

We now look into the Laplace transform. Following our computations in the quenched case and recalling expression (35), we get

$$\mathbb{E}\left\{e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}\right\} = \mathbb{E}\left\{\prod_{i=1}^3 \exp(-\eta_i m_i^\varepsilon)\right\} = \mathbb{E}\left[\exp\left(-\sum_{i=1}^3 \eta_i m_i^\varepsilon\right)\right], \quad (66)$$

with $\eta_i = 1 - e^{-\lambda_i}$.

In the forthcoming Proposition 4.2, we will prove that the following convergence holds in $L^1(\Omega)$ using Hypothesis 2.4:

$$\lim_{\varepsilon\to 0} \sum_{i=1}^3 \eta_i m_i^\varepsilon = \sum_{i=1}^3 \eta_i m_i. \quad (67)$$

Now for all $\lambda_1, \lambda_2, \lambda_3 \geq 0$, we set

$$\Delta_\varepsilon(\lambda_1, \lambda_2, \lambda_3) \equiv \left|\mathbb{E}\left[e^{-\sum_{i=1}^3 \lambda_i M^\varepsilon(A_i)}\right] - \mathbb{E}\left[e^{-\sum_{i=1}^3 \lambda_i M(A_i)}\right]\right|.$$

Due to the fact that $|e^{-x} - e^{-y}| \leq |x - y|$ for all $x, y \geq 0$, we easily get that

$$\begin{aligned}\Delta_\varepsilon(\lambda_1, \lambda_2, \lambda_3) &= \left|\mathbb{E}\left[e^{-\sum_{i=1}^3 \eta_i m_i^\varepsilon}\right] - \mathbb{E}\left[e^{-\sum_{i=1}^3 \eta_i m_i}\right]\right| \\ &\leq \mathbb{E}\left[\left|\sum_{i=1}^3 \eta_i (m_i^\varepsilon - m_i)\right|\right].\end{aligned}$$

Hence, resorting to (67), we end up with

$$\lim_{\varepsilon\to 0} \Delta_\varepsilon(\lambda_1, \lambda_2, \lambda_3) = 0. \quad (68)$$

Relation (68) being true for all $\lambda_1, \lambda_2, \lambda_3 \geq 0$, this proves the desired convergence in distribution. □

We finally prove the following key proposition, which yields relation (67) above.

Proposition 4.2 *For $i = 1, 2, 3$, let m_i^ε be the random variable defined by (35). Also recall that m_1, m_2 and m_3 are, respectively, defined by (52), (54) and (55). Let η_1, η_2,*

η_3 be positive numbers. We assume that Hypothesis 2.4 holds true. Then, we have

$$L^1(\Omega) - \lim_{\varepsilon \to 0} \sum_{i=1}^{3} \eta_i m_i^\varepsilon = \sum_{i=1}^{3} \eta_i m_i. \tag{69}$$

Proof It suffices to show that

$$L^1(\Omega) - \lim_{\varepsilon \to 0} m_i^\varepsilon = m_i, \tag{70}$$

for $i = 1, 2, 3$. In the sequel we will prove relation (70) for $i = 1$, the other cases being handled similarly. In order to achieve this, we proceed along the lines of relations (46)–(52) in Proposition 3.6. Namely, it is enough to establish L^1-convergence for the following integral defined for $\varepsilon > 0$ and $0 \leq \tau_1 \leq \tau_2$:

$$I^\varepsilon(\tau_1, \tau_2) = \int_0^{\tau_1} g(s, \tau_2) \psi\left(Z_{s/\varepsilon}\right) \mathrm{d}s,$$

where $g : \mathbb{R}_+ \to \mathbb{R}$ is a continuously differentiable function. The remainder of the proof is devoted to an L^1-analysis of $I^\varepsilon(\tau_1, \tau_2)$.

Recall that in (47) we have obtained the following decomposition for $I^\varepsilon(\tau_1, \tau_2)$:

$$I^\varepsilon(\tau_1, \tau_2) = I_1^\varepsilon(\tau_1, \tau_2) - I_2^\varepsilon(\tau_1, \tau_2), \tag{71}$$

with

$$I_1^\varepsilon(\tau_1, \tau_2) = g(\tau_1, \tau_2) \int_0^{\tau_1} \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r,$$
$$I_2^\varepsilon(\tau_1, \tau_2) = \int_0^{\tau_1} \frac{\mathrm{d}}{\mathrm{d}s} g(s, \tau_2) \left(\int_0^s \psi\left(Z_{r/\varepsilon}\right) \mathrm{d}r \right) \mathrm{d}s.$$

Furthermore, as in (25)-(26), we have

$$I_1^\varepsilon(\tau_1, \tau_2) = g(\tau_1, \tau_2)\, \tau_1\, \frac{\varepsilon}{\tau_1} \int_0^{\tau_1/\varepsilon} \psi\left(Z_r\right) \mathrm{d}r. \tag{72}$$

Plugging (5) into (72), we get

$$\lim_{\varepsilon \to 0} \left\| I_1^\varepsilon(\tau_1, \tau_2) - g(\tau_1, \tau_2)\, \tau_1 \bar{\psi} \right\|_{L^1(\Omega)} = 0. \tag{73}$$

Similarly to (73), it is also readily seen that

$$\lim_{\varepsilon \to 0} \left\| I_2^\varepsilon(\tau_1, \tau_2) - \int_0^{\tau_1} \left(\frac{\mathrm{d}}{\mathrm{d}s} g(s, \tau_2) \right) s \bar{\psi}\, \mathrm{d}s \right\|_{L^1(\Omega)} = 0. \tag{74}$$

Now gathering (73) and (74) into (71) and applying integration by parts again similarly to (48), we obtain the following limit in $L^1(\Omega)$:

$$\lim_{\varepsilon\to 0} I^{\varepsilon}(\tau_1,\tau_2) = \left[\int_0^{\tau_1} g(s,\tau_2)\,\mathrm{d}s\right]\bar{\psi}.$$

As mentioned before relation (71), this is enough to complete our proof. □

Remark 4.3 Proposition 4.2 asserts that the limit of m_i^{ε} is m_i, which is the same quantity as in the quenched case. This implies that the annealed queue converges to the same limit as the quenched queue.

4.1 Discussion

In this section we discuss briefly the hypotheses used in the current work, in particular, Hypotheses 2.1 and 2.4. We believe that the approach of combining queueing analysis and the framework of homogenization with separation of temporal (and spatial) scales is robust enough, so that same or similar results will still hold under more general or relaxed assumptions. But instead of opting for optimal conditions, we feel that the two hypotheses we used are quite natural.

Note that Hypothesis 2.1 is an *almost sure condition with a specific rate*. It is used in Theorem 3.2, to justify the limit in (29) for the term $A_{2,2}^{\varepsilon}$. In contrast, Hypothesis 2.4 is expressed in terms of the L^1-distance *without a rate*. It is used in Proposition 4.2 to handle the terms I_1^{ε} and I_2^{ε} in (73) and (74). Depending on the regularity and boundedness assumptions we impose on the function ψ, certainly different criteria can be used. Note that for fairly general processes, the rate in Hypothesis 2.1 is almost sharp with $\kappa < \frac{1}{2}$ due to the Law of the Iterated Logarithm. For Hypothesis 2.4, a specific rate can also be imposed if the underlying process has good spectral properties. Such information will be useful if we want to investigate the next order information or questions on the convergence rate or fluctuation phenomena.

5 Numerical examples

We now illustrate our main results in Theorems 3.11 and 3.15 on a couple of specific examples. Following Example 3.3, we model the random environment by an Ornstein–Uhlenbeck (OU) diffusion process satisfying the SDE $dZ_t = \theta(\mu - Z_t)dt + \Sigma dW_t$, where $(W_t : t \geq 0)$ is a standard Brownian motion. We fix the mean $\mu = 0.5$ and the diffusion coefficient $\Sigma = 1.0$, and vary the 'damping' parameter θ. Each value of θ represents a different model of the random environment. The intensity function is set to $(\lambda(s) = 5 + 2\sin(10s) : s \geq 0)$, and the function $\psi(x) = \exp(0.1x)$. Since the invariant distribution of the OU process is Gaussian with mean μ and variance Σ^2, it follows that $\bar{\psi} = \exp(0.1\mu + 0.005\Sigma^2) = e^{0.055} = 1.056$.

The following tables present the estimated sup-norm distance between m^{ϵ} and m, $\|m^{\epsilon} - m\| = \sup_{t\in[0,T]} |m^{\epsilon}(t) - m(t)|$ with $T = 10$, for both the quenched case (Tables 1 and 3) and the annealed case (Tables 2 and 4). In Tables 1 and 2, the service times

Table 1 (A) and (B): Sup-norm distance between quenched m^ϵ and m for different damping factors θ and two independent sample paths of an OU environment process, with exponential service times

θ	ϵ				
	.01	.1	.5	1.0	10.0
(A)					
1.0	17.978	0.813	0.211	0.157	0.068
0.5	14.517	0.710	0.219	0.150	0.067
6.25e-2	6.31e+6	13.079	0.308	0.156	0.070
3.90e-3	5.51e+7	5.281	0.182	0.097	0.069
9.76e-4	5.37e+9	1.464	0.154	0.108	0.071
6.10e-5	6.87e+12	1.358	0.139	0.093	0.067
(B)					
1.0	0.640	0.402	0.258	0.185	0.073
0.5	1.234	0.498	0.237	0.147	0.071
6.25e-2	0.731	1.719	0.303	0.159	0.069
3.90e-3	4.99e+5	1.428	0.135	0.094	0.068
9.76e-4	2.31e+10	2.000	0.153	0.103	0.070
6.10e-5	5.37e+11	1.853	0.149	0.108	0.068

are assumed to be exponential with parameter $\alpha = 10.0$ (or mean 0.1), and in Tables 3 and 4 Pareto with shape parameter $\alpha = 10.0$ (and with mean $10.0/9.0 = 1.11$). Tables 1 and 2 (or 3 and 4, resp.) for two different sample paths of the random environment.

We observe the following qualitative effects from this experiment: First, observe that smaller damping factors imply a larger normed distance $\|m^\epsilon - m\|$. This is true, both in the quenched and annealed cases, as well as for the exponential and Pareto service times. Furthermore, the damping factor cannot be too small either, as the first column in Tables 1 and 2 show. In this case the relaxation time of the OU process is too large to admit any averaging. This is illustrative of the necessity of Hypotheses 2.1 and 2.4.

Second, comparing Tables 1 and 3, observe that the heavier tails imply a larger normed distance (for given ϵ and θ). However, for a fixed θ, for both the heavy- and light-tailed settings, we see that the normed distance decreases with ϵ (excluding $\theta = 0.01$), thereby suggesting that the homogenization effects are apparent. This is particularly true for larger θ. We observe similar effects between Tables 2 and 4.

Third, comparing Tables 1 and 2 (or 3 and 4, resp.), observe that for larger damping factors the normed distance is consistent between the quenched and annealed cases, decreasing with ϵ. This is consistent with our observation in Theorem 3.15 that the annealed process converges weakly to the same the limit as the quenched process, under weaker conditions. The OU process can be shown to satisfy these conditions.

For Pareto services, the annealed results seem to be slightly better than the quenched ones. Therefore, we might be empirically witnessing some fluctuation effects.

Table 2 Sup-norm distance between (estimated) annealed m^ϵ and m for different damping factors θ of an OU environment process, with exponential service times

θ	ϵ				
	.01	.1	.5	1.0	10.0
1.0	52.785	1.196	0.257	0.181	0.074
0.5	1.99e+5	1.700	0.328	0.189	0.073
6.25e-2	7.90e+9	4.641	0.310	0.159	0.069
3.90e-3	1.93e+14	3.563	0.187	0.103	0.069
9.76e-4	2.09e+17	2.055	0.153	0.100	0.069
6.10e-5	1.12e+17	1.839	0.152	0.099	0.069

Table 3 (A) and (B): Sup-norm distance between quenched m^ϵ and m for different damping factors θ and two independent sample paths of an OU environment process, with Pareto service times

θ	ϵ				
	.01	.1	.5	1.0	10.0
(A)					
1.0	5.659	3.836	2.109	1.390	0.667
0.5	7.681	4.252	1.962	1.299	0.650
6.25e-2	5.819	5.844	1.356	0.870	0.620
3.90e-3	9.08e+5	4.749	0.875	0.705	0.615
9.76e-4	4.09E+10	7.763	1.025	0.781	0.629
6.10e-5	6.64E+11	7.360	1.022	0.756	0.619
(B)					
1.0	103.168	6.477	1.751	1.155	0.639
0.5	67.348	5.162	1.468	1.013	0.627
6.25e-2	2.002e+7	86.459	2.749	1.278	0.627
3.90e-3	7.492e+7	9.434	0.969	0.746	0.620
9.76e-4	8.231e+9	7.031	0.963	0.762	0.626
6.10e-5	8.405e+12	5.615	0.825	0.702	0.619

Table 4 Sup-norm distance between annealed m^ϵ and m (estimated over 35 sample paths) for different damping factors θ of an OU environment process, with Pareto service times

θ	ϵ				
	.01	.1	.5	1.0	10.0
1.0	34.286	2.185	0.711	0.533	0.613
0.5	195.333	2.932	0.968	0.593	0.609
6.25e-2	8.598e+5	6.166	1.087	0.725	0.610
3.90e-3	1.653e+11	5.804	0.805	0.670	0.611
9.76e-4	5.881e+12	5.780	0.778	0.647	0.611
6.10e-5	1.082e+14	4.996	0.743	0.646	0.611

6 Conclusion and perspectives

Our primary results in Theorems 3.11 and 3.15 demonstrate that the state of a $\mathrm{Cox}/G_t/\infty$ queue in a random fast oscillatory environment is, in the homogenization limit, closely approximated by that of an $M_t/G_t/\infty$ queue with nonhomogeneous

Poisson traffic. More precisely, the rapid fluctuations of the stochastic intensity are averaged out in the limit. These results could prove useful for performance analysis in some circumstances, such as bounded performance metric functions, since much is known about the properties of the $M_t/G/\infty$ queue (and this is also a much simpler object to simulate). Our model assumes a two timescale structure, wherein time-of-day effects in the traffic intensity are modeled by a smoothly varying function, and stochastic fluctuations are modeled by a strongly ergodic stochastic process. We also assume a very general model of time-varying service wherein the service time distribution itself depends on the arrival epoch. A crucial insight that emerges from our analysis is the fact that we do not require exponential ergodicity of the stochastic environment, though we do require a strong sense of ergodicity to be satisfied. Furthermore, our analysis permits both light- and heavy-tailed service time distributions. Therefore, for a rather broad range of infinite server queueing models, the system state is well approximated by a much simpler $M_t/G/\infty$ queue, under the homogenization scaling considered in this paper.

Of course, these insights are greatly facilitated by the fact that we study an infinite server queue, allowing us to leverage the properties of Poisson point processes. In this setting we anticipate proving a functional central limit theorem (FCLT) for the system state in the homogenization limit, complementing Theorems 3.11 and 3.15 with a rate of convergence. In a stationary setting where the arrival intensity $\lambda(\cdot)$ is a constant, it is well known that a FCLT holds and that the approximating process is O-U; see [3, Section 3]. In our setting, however this analysis is complicated by the fact that the centering is by a time-varying function $\bar{m}(\cdot)$, and the analysis appears to require some further technical development which is outside the scope of this paper. Second, while our results are in the homogenization limit as $\epsilon \to 0$, it would be interesting to consider the large time behavior of the process N^ϵ for a fixed $\epsilon > 0$, when $\lambda(\cdot)$ is a constant. Given the more general setting we are studying, this type of result would expand on the results in [3, Section 3]. It should also be noted that in [3] the traffic model is a special stationary DSPP where the stochastic intensity process is constructed as $\Lambda(t) = \sum_j \Lambda_j 1_{\{[j\Delta,(j+1)\Delta)\}}(t)$, where $\{\Lambda_j\}$ are i.i.d. random variables and $1_{\{\cdot\}}$ represents an indicator function. That paper establishes a rather interesting "trichotomy" result, in particular showing that if the sampling is "rapid" then the traffic process is Poisson-like, reflecting an averaging effect; on the other hand, they also find that if the sampling is "slow," then the over-dispersed nature of the DSPP is maintained in the limit, and consequently, the limit process is *not* a $M_t/G/\infty$ queue. The small ϵ setting in this paper is, in a sense, a more general "rapid" sampling procedure. In this context, it is reassuring to see that we are able to recover the Poisson-like structure in the limit, even with heavy-tailed service and polynomial ergodicity of the underlying stochastic intensity. We do not, however, have a result that parallels the "slow" sampling result in [3]. This suggests that there are regimes where timescale separation between the time-of-day effects and the stochastic intensity are not manifested in the limit. This appears to require a more refined CLT-type analysis. We will address this interesting phenomenon in a future paper.

References

1. Whitt, W.: Time-varying queues. Queueing models and service management **1**(2), (2018)
2. O'Cinneide, C.A., Purdue, P.: The $M/M/\infty$ queue in a random environment. J. Appl. Probab. **23**(1), 175–184 (1986)
3. Heemskerk, M., van Leeuwaarden, J., Mandjes, M.: Scaling limits for infinite-server systems in a random environment. Stochast. Syst. **7**(1), 1–31 (2017)
4. Heemskerk, M., Mandjes, M.: Exact asymptotics in an infinite-server system with overdispersed input. Oper. Res. Lett. **47**(6), 513–520 (2019)
5. Pender, J., Ko, Y.M.: Approximations for the queue length distributions of time-varying many-server queues. INFORMS J. Comput. **29**(4), 688–704 (2017)
6. Boxma, O., Kella, O., Mandjes, M.: Infinite-server systems with Coxian arrivals. Queueing Syst. **92**(3–4), 233–255 (2019)
7. Jansen, H.M., Mandjes, M., De Turck, K., Wittevrongel, S.: Diffusion limits for networks of Markov-modulated infinite-server queues. Perform. Evaluat. **135**, 102039 (2019)
8. Dean, J., Ganesh, A., Crane, E.: Functional large deviations for Cox processes and $\mathrm{Cox}/G/\infty$ queues, with a biological application. Ann. Appl. Probab. **30**(5), 2465–2490 (2020)
9. Anderson, D., Blom, J., Mandjes, M., Thorsdottir, H., De Turck, K.: A functional central limit theorem for a Markov-modulated infinite-server queue. Methodol. Comput. Appl. Probab. **18**(1), 153–168 (2016)
10. Blom, J., Mandjes, M.: A large-deviations analysis of Markov-modulated infinite-server queues. Oper. Res. Lett. **41**(3), 220–225 (2013)
11. Blom, J., Mandjes, M., Thorsdottir, H.: Time-scaling limits for Markov-modulated infinite-server queues. Stochastic Models **29**(1), 112–127 (2013)
12. Blom, J., Kella, O., Mandjes, M., Thorsdottir, H.: Markov-modulated infinite-server queues with general service times. Queueing Syst. **76**(4), 403–424 (2014)
13. Blom, J., De Turck, K., Mandjes, M.: Functional central limit theorems for Markov-modulated infinite-server systems. Math. Methods Oper. Res. **83**(3), 351–372 (2016)
14. Fralix, B.H., Adan, I.J.B.F.: An infinite-server queue influenced by a semi-Markovian environment. Queueing Syst. **61**(1), 65–84 (2009)
15. Hellings, T., Mandjes, M., Blom, J.: Semi-Markov-modulated infinite-server queues: approximations by time-scaling. Stochastic Models **28**(3), 452–477 (2012)
16. Mandjes, M., De Turck, K.: Markov-modulated infinite-server queues driven by a common background process. Stochastic Models **32**(2), 206–232 (2016)
17. Arous, G.B., Molchanov, S., Ramírez, A.F.: Transition from the annealed to the quenched asymptotics for a random walk on random obstacles. Ann. Probab. **33**(6), 2149–2187 (2005)
18. Arous, G.B., Bogachev, L.V., Molchanov, S.A.: Limit theorems for sums of random exponentials. Probab. Theory Relat. Fields **132**(4), 579–612 (2005)
19. Schlather, M.: Limit distributions of norms of vectors of positive IID random variables. Ann. Probab. **29**(2), 862–881 (2001)
20. Kontoyiannis, I., Meyn, S.P.: Spectral theory and limit theorems for geometrically ergodic Markov processes. Ann. Appl. Probab. **13**(1), 304–362 (2003)
21. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_t/G/\infty$ queue. Oper. Res. **41**(4), 731–742 (1993)
22. Eick, S.G., Massey, W.A., Whitt, W.: $M_t/G/\infty$ queues with sinusoidal arrival rates. Manage. Sci. **39**(2), 241–252 (1993)
23. Massey, W.A., Whitt, W.: Networks of infinite-server queues with nonstationary poisson input. Queueing Syst. **13**(1–3), 183–250 (1993)
24. Zheng, Z., Honnappa, H., Glynn, P.W.: Approximating systems fed by Poisson processes with rapidly changing arrival rates. arXiv preprint, arXiv:1807.06805 (2018)
25. Resnick, S., Rootzén, H.: Self-similar communication models and very heavy tails. Ann. Appl. Probab. **10**, 753–778 (2000)
26. Fibich, G., Gavious, A., Solan, E.: Averaging principle for second-order approximation of heterogeneous models with homogeneous models. Proc. Nat. Acad. Sci. **109**(48), 19545–19550 (2012)
27. Khasminskii, R.Z., Yin, G.: On averaging principles: an asymptotic expansion approach. SIAM J. Math. Anal. **35**(6), 1534–1560 (2004)

28. Khasminskii, R.Z.: On the principle of averaging the Ito's stochastic differential equations. Kybernetika **4**(3), 260–279 (1968)
29. Kurtz, T.G.: Averaging for martingale problems and stochastic approximation. In: Applied Stochastic Analysis, pp. 186–209. Springer (1992)
30. Fouque, J.-P., Papanicolaou, G., Sircar, K.R.: Financial modeling in a fast mean-reverting stochastic volatility environment. Asia-Pacific Finance Markets **6**(1), 37–48 (1999)
31. Fouque, J.-P., Ruimeng, H.: Optimal portfolio under fast mean-reverting fractional stochastic environment. SIAM J. Financ. Math. **9**(2), 564–601 (2018)
32. Blanchet, J., Chen, X.: Continuous-time modeling of bid-ask spread and price dynamics in limit order books. arXiv preprint, arXiv:1310.1103 (2013)
33. Coffman Jr., E.G., Puhalskii, A.A., Reiman, M.I.: Polling systems with zero switchover times: a heavy-traffic averaging principle. Ann. Appl. Probab. **5**, 681–719 (1995)
34. Hunt, P.J., Kurtz, T.G.: Large loss networks. Stochastic Processes Appl. **53**(2), 363–378 (1994)
35. Perry, O., Whitt, W.: An ODE for an overloaded X model involving a stochastic averaging principle. Stochastic Syst. **1**(1), 59–108 (2011)
36. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**(1–2), 149–201 (1998)
37. Honnappa, H., Jain, R., Ward, A.R.: A queueing model with independent arrivals, and its fluid and diffusion limits. Queueing Syst. **80**(1–2), 71–103 (2015)
38. Mandelbaum, A., Massey, W.A.: Strong approximations for time-dependent queues. Math. Oper. Res. **20**(1), 33–64 (1995)
39. Spreij, P., Storm, J.: Diffusion limits for a Markov modulated counting process. arXiv preprint arXiv:1801.03682 (2018)
40. Liu, Y., Whitt, W., et al.: Many-server heavy-traffic limit for queues with time-varying parameters. Ann. Appl. Probab. **24**(1), 378–421 (2014)
41. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. Oper. Res. Lett. **40**(5), 307–312 (2012)
42. Chakraborty, P., Honnappa, H.: A many-server functional strong law for a non-stationary loss model. Oper. Res. Lett. 49(3), 338-344 (2021)
43. Zheng, Z., Honnappa, H., Glynn, P.W.: Approximating performance measures for slowly changing non-stationary Markov chains. arXiv preprint, arXiv:1805.01662 (2018)
44. Massey, W.A., Whitt, W.: Uniform acceleration expansions for markov chains with time-varying rates. Ann. Appl. Probab. **5**, 1130–1155 (1998)
45. Pender, J.: Nonstationary loss queues via cumulant moment approximations. Probab. Eng. Inf. Sci. **29**(1), 27–49 (2015)
46. Khashinskii, R.Z., Yin, G., Zhang, Q.: Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains. SIAM J. Appl. Math. **56**(1), 277–293 (1996)
47. Khasminskii, R.Z., Yin, G., Zhang, Q.: Constructing asymptotic series for probability distributions of Markov chains with weak and strong interactions. Q. Appl. Math. **55**(1), 177–200 (1997)
48. Kim, S.-H., Whitt, W.: Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf. Service Oper. Manag. **16**(3), 464–480 (2014)
49. Choudhury, G.L., Mandelbaum, A., Reiman, M.I., Whitt, W.: Fluid and diffusion limits for queues in slowly changing environments. Stochastic Models **13**(1), 121–146 (1997)
50. Koops, D.T., Boxma, O.J., Mandjes, M.R.H.: Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. Queueing Syst. **86**(3), 301–325 (2017)
51. De Turck, K.E.E.S., Mandjes, M.R.H.: Large deviations of an infinite-server system with a linearly scaled background process. Perform. Evaluat. **75**, 36–49 (2014)
52. Jansen, H.M., Mandjes, M.R.H., De Turck, K., Wittevrongel, S.: A large deviations principle for infinite-server queues in a random environment. Queueing Syst. **82**(1–2), 199–235 (2016)
53. Komorowski, T., Landim, C., Olla, S.: Fluctuations in Markov Processes: Time Symmetry and Martingale Approximation. Springer, Berlin (2012)
54. Chen, X.: The law of the iterated logarithm for functionals of Harris recurrent Markov chains: self normalization. J. Theor. Probab. **12**(2), 421–445 (1999)
55. Löcherbach, E., Loukianova, D.: The law of iterated logarithm for additive functionals and martingale additive functionals of Harris recurrent Markov processes. Stochastic Processes Appl. **119**, 2312–2335 (2009)

56. Saussereau, B.: Transportation inequalities for stochastic differential equations driven by a fractional Brownian motion. Bernoulli **18**(1), 1–23 (2012)
57. Garrido-Atienza, M.J., Kloeden, P.E., Neuenkirch, A.: Discretization of stationary solutions of stochastic systems driven by fractional Brownian motion. Appl. Math. Optim. **60**(2), 151–172 (2009)
58. Resnick, S.I.: Point processes, regular variation and weak convergence. Adv. Appl. Probab. **18**(1), 66–138 (1986)
59. Çinlar, E.: Probability and Stochastics. Springer, Berlin (2011)
60. Ferger, D., Vogel, D.: Weak convergence of the empirical process and the rescaled empirical distribution function in the Skorokhod product space. Theory Probab. Appl. **54**(4), 609–625 (2010)
61. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, vol. 46. Springer, Berlin (2013)