Singular perturbation for a two-class Processor-Sharing queue with impatience

Alain Simonian, Florian Simatos, Ridha Nasri

Singular perturbation for a two-class Processor-Sharing queue with impatience

R. Nasri* F. Simatos† A. Simonian‡

May 7, 2021

Abstract

A two-class Processor-Sharing queue with one impatient class is studied. Local exponential decay rates for its stationary distribution $(N(\infty), M(\infty))$ are established in the heavy traffic regime where the arrival rate of impatient customers grows proportionally to a large factor A. This regime is characterized by two time-scales, so that no general Large Deviations result is applicable. In the framework of singular perturbation methods, we instead assume that an asymptotic expansion of the solution of associated Kolmogorov equations exists for large A and derive it in the form

$$\mathbb{P}(N(\infty) = Ax, M(\infty) = Ay) \sim \frac{g(x,y)}{2\pi A} \cdot e^{-A\,H(x,y)}, \qquad x > 0,\ y > 0,$$

with explicit functions g and H.

This result is then applied to the model of mobile networks proposed in [15] and accounting for the spatial movement of users. We give further evidence of a unusual growth behavior in heavy traffic in that the stationary mean queue length $\mathbb{E}(N_{\mathtt{mob}}(\infty))$ and $\mathbb{E}M_{\mathtt{mob}}(\infty))$ of each customer-class increases proportionally to

$$\mathbb{E}(N_{\mathtt{mob}}(\infty)) \propto \mathbb{E}(M_{\mathtt{mob}}(\infty)) \propto \log\left(\frac{1}{1-\varrho_{\mathtt{tot}}}\right)$$

with system load $\varrho_{\mathtt{tot}}$ tending to 1, instead of the usual $1/(1-\varrho_{\mathtt{tot}})$ growth behavior.

1 Queuing Model and Main Results

We describe the addressed queuing system and the specific asymptotic regime considered to evaluate its stationary occupancy distribution. We then state our main mathematical results and apply them to the account of spatial user movement in mobile networks.

*Orange Labs, OLN/GDM, Orange Gardens, 44 avenue de la République, CS 50010, 92326 Chatillon Cedex (ridha.nasri@orange.com).

†ISAE-SUPAERO, Université de Toulouse, 10 avenue Edouard Belin, 31055 Toulouse, France, (florian.simatos@isae.fr).

‡Orange Labs, OLN/GDM, Orange Gardens, 44 avenue de la République, CS 50010, 92326 Chatillon Cedex (alain.simonian@orange.com).

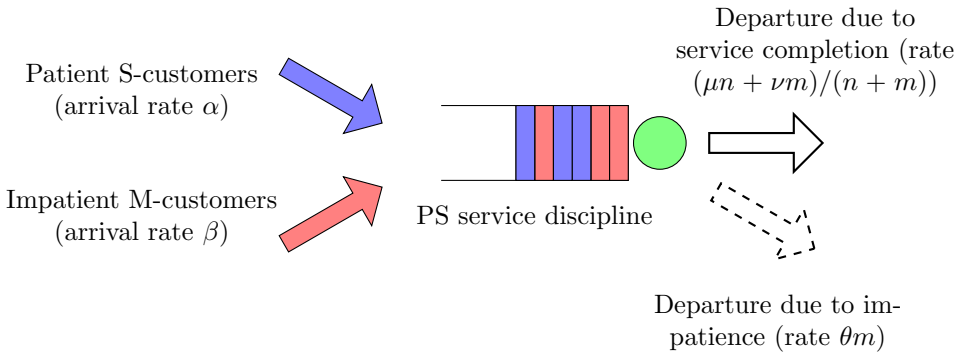

Figure 1: Multi-class PS queue with impatience.

1.1 Two-class Processor-Sharing queue with one impatient class

In this paper, we consider a two-class Markovian Processor-Sharing (PS) queue where one class of users are impatient and leave the system at rate $\theta > 0$. This queuing system is depicted in Figure 1 and can be described as follows:

- the arrival process of patient (resp. impatient) customers entering the queue is Poisson with rate α (resp. β);
- service requirements for successive patient (resp. impatient) customers are i.i.d. and exponentially distributed with mean $1/\mu$ (resp. $1/\nu$);
- server capacity is normalized to unity and customers are served according to the PS service discipline, that is, when there are $k \geqslant 1$ customers in service, each one is served instantaneously at rate $1/k$;
- the sojourn times of impatient customers in queue (before possible service completion) are i.i.d. and exponentially distributed with mean $1/\theta$.

This defines a birth-and-death process $(N, M) = ((N(t), M(t)), t \geqslant 0)$ with values in $\mathbb{N}^2$ and whose infinitesimal generator Ω is given by

$$\Omega(f)(n,m) = \alpha\left(f(n+1,m) - f(n,m)\right) + \beta\left(f(n,m+1) - f(n,m)\right)$$
$$+ \frac{\mu n}{n+m}\left(f(n-1,m) - f(n,m)\right) + \left(\frac{\nu m}{n+m} + \theta m\right)\left(f(n,m-1) - f(n,m)\right)$$

for $f : \mathbb{N}^2 \to \mathbb{R}$ and $(n, m) \in \mathbb{N}^2$ (with the convention $0/0 = 0$). This process has a stationary distribution $(N(\infty), M(\infty))$ if and only if the stability condition

$$\varrho = \frac{\alpha}{\mu} < 1 \tag{1}$$

holds [14, Sect.12.2, Prop.12.1], where ϱ denotes the load of patient customers offered to the system. Note that this stability condition only involves the load of patient customers (through their arrival rate α and service requirement μ) and not that of impatient ones, as the latter can always leave the system in a finite time whatever the system load.

1.2 Two time scales in the heavy traffic regime

In this queue, we are interested in the heavy traffic regime where β tends to infinity, while the four other parameters α, μ, ν and θ remain fixed. We will consider $A = \beta/\theta$ as our scaling parameter and write $A \to \infty$ to mean that $\beta \to \infty$ with all other parameters kept fixed. In this regime, both processes N and M become of the order of A but evolve on different time scales as can be observed when considering their fluid behavior.

As A becomes large, M becomes large and so departures are mostly due to the impatience term $\theta \cdot m$, given $M = m$ and $N = n$, since the service term $\nu m/(n+m)$ remains bounded. If this service term could be neglected, then M would be equal to M', the $M/M/\infty$ queue length with input rate $\beta = A\theta$ and service rate θ. As specified below, M and M' indeed behave very similarly in the considered heavy traffic regime. In fact, a simple coupling argument between M and M' makes it possible to transfer to M the well-known heavy traffic behavior of M', namely, to show that the process $(M(t)/A, t \geqslant 0)$ scaled only in space converges (weakly, in a functional sense) to the deterministic solution $(y(t), t \geqslant 0)$ to the ordinary differential equation (ODE)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \theta - \theta\, y$$

and that its stationary distribution $M(\infty)/A$ converges to the unique stable point

$$y^* = 1 \tag{2}$$

of this ODE.

On the other hand, arrival and service rates of N remain bounded: they are respectively equal to α and $\mu n/(n+m) \in [0, \mu]$. As defined by this service rate, component N needs to become commensurate with component M in order to obtain some service and so it will also live on the $O(A)$ space scale. But since its arrival rate is bounded, it needs a time of order $O(A)$ to reach such values and it is indeed on this time scale that it evolves. On this time scale, however, M evolves very rapidly and so an averaging behavior is to be expected, whereby N and M would interact through the mean value of M which, as argued above, is close to A. In other words, the asymptotic behavior of N is expected to be close to that of N', the length of the single-server PS queue with A permanent customers. In fact, standard methods could be used to prove that N and N' have the same fluid limit; specifically, the process $(N(At)/A, t \geqslant 0)$ scaled both in time and space converges to the deterministic solution $(x(t), t \geqslant 0)$ to the

ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha - \mu \frac{x}{x+1},$$

and its stationary distribution $N(\infty)/A$ converges to the unique stable point

$$x^* = \frac{\varrho}{1-\varrho} \tag{3}$$

of this ODE, with again $\varrho = \alpha/\mu < 1$.

In other words, in the heavy traffic regime when $A \to \infty$, the fluid behavior of (N, M) is the same as that of (N', M') and the main goal of this paper is to investigate to which extent this approximation holds in a Large Deviations setting.

1.3 Main results

In order to emphasize the dependency with respect to the scaling parameter A, let us denote by $\mathbf{\Pi}_A$ the stationary distribution of (N, M) when $\beta/\theta = A$ (recall that we let $A \to \infty$ while the four parameters α, μ, ν and θ remain fixed). It follows from the above discussion that the mass of distribution $\mathbf{\Pi}_A$ is essentially concentrated around (Ax^*, Ay^*) in the sense that $\mathbf{\Pi}_A([A\underline{x}, A\bar{x}] \times [A\underline{y}, A\bar{y}]) \to 1$ when $A \to \infty$, for any $\underline{x} < x^* < \bar{x}$ and $\underline{y} < y^* < \bar{y}$. This regime therefore defines a Large Deviations setting for $\mathbf{\Pi}_A$, whereby probabilities $\mathbf{\Pi}_A(Ax, Ay)$ decrease exponentially for increasing A and fixed $x \geqslant 0$, $y \geqslant 0$.

The main result of the present paper is to establish sharp local asymptotics using the singular perturbation method, as discussed in more detail in Section 2 below. In this framework, it is admitted that an expansion of the form

$$\mathbf{\Pi}_A(Ax, Ay) = \frac{1}{2\pi A} \times \exp\left[-A \cdot H(x,y) - h_0(x,y) - \frac{h_1(x,y)}{A} - \frac{h_2(x,y)}{A^2} + O\left(\frac{1}{A^3}\right)\right], \quad x,\ y > 0, \tag{4}$$

exists for functions H and h_0, h_1, h_2 satisfying some specific smoothness assumptions; these functions are then successively determined via the Kolmogorov equations. Note that H in expansion (4) is the usual decay function of the Large Deviations theory, defined by

$$H(x,y) = -\lim_{A\to\infty} \frac{1}{A} \log \mathbf{\Pi}_A(Ax, Ay).$$

Our main result involves the functions Φ, Ψ and g that will appear repeatedly in the sequel, and which are respectively defined by

$$\Phi(x) = x \log\left(\frac{x}{\varrho}\right) - (x+1)\log(x+1) - \log(1-\varrho), \qquad x \geqslant 0, \tag{5}$$

$$\Psi(y) = y \log y - y + 1, \qquad y \geqslant 0, \tag{6}$$

and

$$g(x,y) = (1-\varrho)\sqrt{\frac{x+1}{x\,y}}\left(\frac{x+1}{x+y}\right)^{\nu/\theta} \exp\left[\frac{\mu}{\theta}(1-\varrho)\left(\frac{x-x^*}{x+1}\right)\log\left(\frac{x+1}{x+y}\right)\right]$$

for $x,\ y > 0$ (recall that x^* and y^* have been defined in (3) and (2)).

Theorem 1. *Beside stability condition $\varrho < 1$, assume further that an asymptotic expansion of the form (4) exists and satisfies the following smoothness conditions:*

1. *the functions H, h_0, h_1 and h_2 are respectively of class $\mathscr{C}^3$, $\mathscr{C}^2$, $\mathscr{C}^1$ and $\mathscr{C}^0$ in the open quarter-plane $\mathbb{R}^{+*} \times \mathbb{R}^{+*}$;*
2. *the decay function H is non negative, continuous over the closed quarter plane $\mathbb{R}^+ \times \mathbb{R}^+$, and satisfies $H(x^*, y^*) = 0$.*

Then as $A \to \infty$, we have

$$\mathbf{\Pi}_A(Ax, Ay) \sim \frac{g(x,y)}{2\pi A}\, e^{-A\cdot(\Phi(x)+\Psi(y))} \tag{7}$$

for any $x,\ y > 0$.

The assumption on the continuity of H over $\mathbb{R}^+ \times \mathbb{R}^+$ is motivated by the fact that these properties hold in the case when a large deviations principle (LDP) exists (this results from the lower semi-continuity of H [8, Chap.7, Sect.6], together with the existence of an attained infimum for the action functional on any closed subset [8, p.81]). Focusing on the decay function H, Theorem 1 has the following consequence.

Theorem 2. *Under the same assumptions as that of Theorem 1, the decay rate H of distribution $\mathbf{\Pi}_A$ equals the sum*

$$H(x,y) = \Phi(x) + \Psi(y)$$

for all $x,\ y \geqslant 0$.

For illustration (see Figure 2), the convex surface $z = H(x,y)$ in the (x,y,z)-space is plotted for $\varrho = 0.5$. We then have $(x^*, y^*) = (1,1)$ and, in particular, $H(0,0) \approx 1.69$, $H(3,0) \approx 1.52$, $H(0,3) \approx 1.99$. The level curves $H(x,y) =$ constant in the positive quadrant are also depicted.

Theorem 2 thus asserts that distribution $\mathbf{\Pi}_A$ is asymptotically the product of two marginal distributions in the logarithmic order. Actually, the components Φ and Ψ that appear are exactly those of the processes N' and M' introduced earlier, that is, Φ is the decay rate of the single-server PS queue with input rate α, A permanent customers and service rate μ (this will be proved in Appendix A), and Ψ is the decay rate of the $M/M/\infty$ queue with input rate $A\theta$ and service rate θ [8, Chap.5, p.160]. This result therefore shows that the approximation $(N, M) \approx (N', M')$ remains accurate in the logarithmic order for

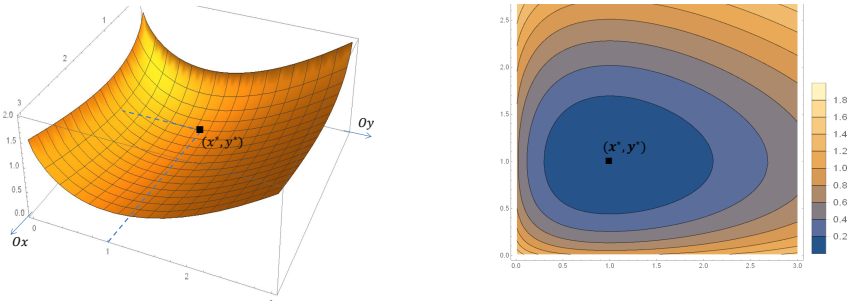

Figure 2: Surface $z = H(x, y)$ and level curves $H(x, y) = \text{constant}$.

large deviations. However, Theorem 1 shows that this approximation breaks down in the usual, say $O(1)$, order because function g in (7) does not factorize into the product of two functions of x and y.

The next result shows that this independence property in the logarithmic order is enough to imply independence of centered and scaled stationary distributions.

Theorem 3. *Under the same assumptions as that of Theorem 1, the centered pair*

$$(\xi_A, \eta_A) = \sqrt{A}\left(\frac{N(\infty)}{A} - x^*, \frac{M(\infty)}{A} - y^*\right)$$

converges weakly as $A \to \infty$ towards the centered Gaussian variable (ξ, η) with covariance structure

$$\mathbb{E}(\xi^2) = \frac{\varrho}{(1-\varrho)^2}, \quad \mathbb{E}(\eta^2) = 1, \quad \mathbb{E}(\xi\eta) = 0.$$

Moreover, we have

$$\mathbb{E}(N(\infty)) \sim A\,x^*, \qquad \mathbb{E}(M(\infty)) \sim A$$

for large A.

Theorem 3 implies, in particular, that the scaled pair $(N(\infty)/A, M(\infty)/A)$ converges weakly to the deterministic point (x^*, y^*), as was alluded to before. Besides, the asymptotic distribution of (ξ, η) has zero covariance, so that its components are asymptotically independent, although $N(\infty)$ and $M(\infty)$ are dependent for finite A. In fact, (ξ, η) is the limit of the centered and scaled stationary distribution of (N', M'), showing that the approximation $(N, M) \approx (N', M')$ still holds for fluctuations of stationary distributions around their deterministic limits (this is verified in Appendix A, Remark 5, for the variable ξ; on the other hand, this readily follows from the Gaussian approximation of the Poisson distribution of the $M/M/\infty$ queue for the variable η).

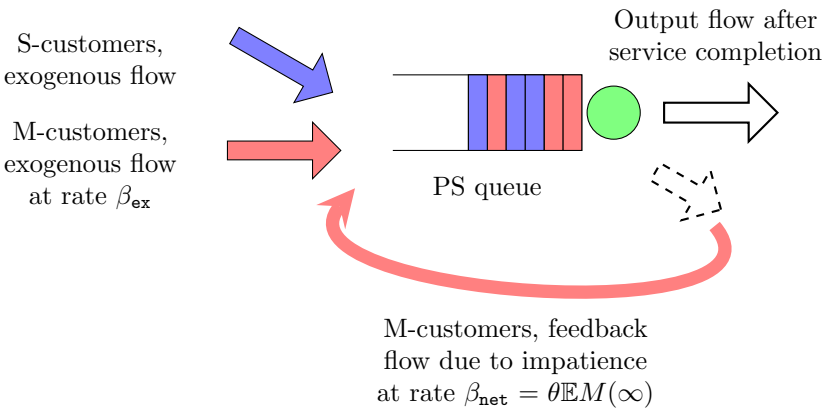

Figure 3: Multi-class closed-loop PS queue.

1.4 Application to mobile networks

Numerous models of multi-class PS queues with impatience have been investigated in the queueing literature. The single-class PS queue with impatience has been early dealt with to derive asymptotics for the stationary queue distribution [4]. In the context of radio communication networks, the multi-class case when all classes are impatient has been addressed for the control of early customer departure in the overload regime [9]. More recently, this multi-class queue has been invoked for the performance of radio networks when accounting for spatial mobility [14, 15, 19]. In this context, impatience is used to model mobility, as both impatience and mobility make customers leave the system independently of the service received.

This last stream of results is actually one of our motivation for investigating Theorem 2. These papers consider the process (N, M) above, with N the number of Static (patient) customers, and M the number of Moving (impatient) customers in the considered radio cell: a departure of an M-customer is thus either due to a service completion, or to a spatial movement to another neighboring cell of the network, the latter happening in stationarity at rate $\theta\, \mathbb{E}(M(\infty))$.

In order to account for possible reverse movements of users to the considered cell in the network but outside the cell, the authors of [14, 15, 19] consider the so-called *closed-loop* Processor-Sharing queue (see Figure 3). In the latter, the arrival rate of M-customers is decomposed as

$$\beta = \beta_{\mathtt{ex}} + \beta_{\mathtt{net}},$$

$\beta_{\mathtt{ex}}$ representing the rate of exogenous arrivals and $\beta_{\mathtt{net}}$ the rate of arrivals within the network. The rate $\beta_{\mathtt{ex}}$ is fixed, while the rate $\beta_{\mathtt{net}}$ is obtained by imposing a balance condition. In fact, the authors consider the case of a balanced cell

where the movements of mobile users within the network from and to the cell balance each other, that is, $\beta_{\mathtt{net}}$ is equal to the rate $\theta\,\mathbb{E}(M(\infty))$ of customers moving out of the cell. This balance condition is captured by the equation

$$\theta \cdot \mathbb{E}(M(\infty)) = \beta_{\mathtt{net}}. \tag{8}$$

Since $\mathbb{E}(M(\infty))$ is itself a function of $\beta_{\mathtt{net}}$, (8) is a Fixed-Point equation. It has been proved [14, Proposition 3.1] that this Fixed-Point equation has a unique solution if and only if

$$\varrho_{\mathtt{tot}} := \frac{\alpha}{\mu} + \frac{\beta_{\mathtt{ex}}}{\nu} < 1,$$

meaning that the total load imposed by exogenous arrivals is smaller than the cell capacity. When it is enforced, this defines a Markov process $(N_{\mathtt{mob}}, M_{\mathtt{mob}})$ which is a particular case of the above (N, M) process with a parameter β specifically chosen as an implicit function of other parameters $\alpha, \beta_{\mathtt{ex}}, \mu$ and ν, that is, $\beta = \beta_{\mathtt{ex}} + \beta_{\mathtt{net}}$ with $\beta_{\mathtt{net}}$ determined by Fixed-Point equation (8).

In [19], this Markov process $(N_{\mathtt{mob}}, M_{\mathtt{mob}})$ is studied in the heavy traffic regime $\varrho_{\mathtt{tot}} \uparrow 1$: this makes the rate $\beta_{\mathtt{net}}$ of inner movements grow large and it thus amounts to studying the (N, M) process in the regime $A \to \infty$. It is proved there, in particular, that the stationary distribution remarkably grows as the logarithm of $1/(1-\varrho_{\mathtt{tot}})$, a very peculiar result in sharp contrast with the usual $1/(1-\varrho_{\mathtt{tot}})$ growth in heavy traffic. More precisely, the authors show that the random sequence $(N_{\mathtt{mob}}(\infty), M_{\mathtt{mob}}(\infty))/\log(1/(1-\varrho_{\mathtt{tot}}))$ is tight when $\varrho_{\mathtt{tot}} \uparrow 1$, that any accumulation point is larger than (x^*, y^*) and they conjecture that this lower bound is actually the exact limit. As argued in [19], proving this requires to prove that

$$-\frac{1}{A}\log \mathbf{\Pi}_A(0,0) \longrightarrow H(0,0) = 1 - \log(1-\varrho)$$

when $A \to \infty$, which is a direct consequence of Theorem 2 in the framework of the present singular perturbation setting.

It is proved in [19] that $\mathbb{P}(N_{\mathtt{mob}}(\infty) = M_{\mathtt{mob}}(\infty) = 0) = 1 - \varrho_{\mathtt{tot}}$ so that as $\varrho_{\mathtt{tot}} \uparrow 1$, we have $\beta \to \infty$ in such a way that $A \sim -\log(1-\varrho_{\mathtt{tot}})/H(0,0)$. A direct application of Theorem 2 to the $(N_{\mathtt{mob}}, M_{\mathtt{mob}})$ process then enables us to state the following.

Theorem 4. *Suppose $\varrho < 1$ and that the assumptions of Theorem 1 hold. We further let*

$$A_{\mathtt{mob}} = -\frac{\log(1-\varrho_{\mathtt{tot}})}{H(0,0)}$$

with $H(0,0) = 1 - \log(1-\varrho)$. As $\varrho_{\mathtt{tot}} \uparrow 1$, the centered pair

$$\sqrt{A_{\mathtt{mob}}}\left(\frac{N_{\mathtt{mob}}(\infty)}{A_{\mathtt{mob}}} - x^*, \frac{M_{\mathtt{mob}}(\infty)}{A_{\mathtt{mob}}} - y^*\right)$$

converges weakly to the same Gaussian variable (ξ, η) as that of Theorem 3. Moreover, the mean queue occupancies grow logarithmically as

$$\mathbb{E}(N_{\mathtt{mob}}(\infty)) \sim A_{\mathtt{mob}}\, x^* \qquad \textit{and} \qquad \mathbb{E}(M_{\mathtt{mob}}(\infty)) \sim A_{\mathtt{mob}}$$

when $\varrho_{\mathtt{tot}} \uparrow 1$*, with again* $x^* = \varrho/(1-\varrho)$ *and* $y^* = 1$*.*

The latter estimates of the mean queue occupancies enable us to derive asymptotics for the average throughput of each customer class. Seeing the workload brought by each arriving customer as a data volume to be transferred through a communication link (server) with total transmission capacity C, the mean throughput can be defined as the ratio of the mean volume of transferred data to the mean transfer time of a given customer [15, Section 2.1]. Normalizing the server capacity C to unity and using the general expressions of [15, Prop.2.2], the efficient throughputs γ and Γ of class S (Static) and M (Moving) customer flows can then be readily expressed by

$$\gamma = \frac{\varrho}{\mathbb{E}(N_{\mathtt{mob}}(\infty))}, \qquad \Gamma = \frac{1}{\mathbb{E}(M_{\mathtt{mob}}(\infty))}\left(\varrho_{\mathtt{tot}} - \varrho + \frac{\beta_{\mathtt{net}}}{\nu}\right) - \frac{\theta}{\nu},$$

respectively, where rate $\beta_{\mathtt{net}}$ is defined by (8). As $\varrho_{\mathtt{tot}} \uparrow 1$, the estimates of $\mathbb{E}(N_{\mathtt{mob}}(\infty))$ and $\mathbb{E}(M_{\mathtt{mob}}(\infty))$ provided by Theorem 4 then yield

$$\gamma \sim -(1-\log(1-\varrho))\,\frac{1-\varrho}{\log(1-\varrho_{\mathtt{tot}})}, \qquad \Gamma \sim -(1-\log(1-\varrho))\,\frac{\varrho_{\mathtt{tot}}-\varrho}{\log(1-\varrho_{\mathtt{tot}})} \quad (9)$$

for each customer class of the closed-loop queue.

1.5 Organization of paper

Before presenting the proofs of the latter results, we first discuss in Section 2 their relevance compared to the current literature on both Large Deviations and Singular Perturbation methods. Section 3 contains preliminary technical results. Although Theorem 2 above was claimed as a consequence of Theorem 1, the proof proceeds by first proving Theorem 2 in Section 4, and then iterating the argument to prove Theorem 1 in Section 5. Section 5 also presents a direct Corollary to Theorem 1 concerning the asymptotic behavior of the marginal distributions of $N(\infty)$ and $M(\infty)$ (Corollary 1). The proof of Theorem 3 is then given in Section 6; it essentially relies on the asymptotics that Theorem 1 enables us to obtain for the generating function of distribution $\mathbf{\Pi}_A$. Appendix A establishes that function Φ is the decay rate of the single-server PS-queue with A permanent customers; Appendices B and C provide the proofs of two intermediate results that intervene in the proof of Theorem 3.

2 Asymptotics of stationary distributions

Prior to proceeding to the detailed proofs of our main results, we first review previous works addressing asymptotics for the stationary distribution of Markov jump processes.

2.1 Large Deviations Principles

Consider a scaled jump process $\mathbf{Z}_A$ in some subset of the lattice $\mathbb{Z}^d/A$, $d \geqslant 2$. The scaling applied to $\mathbf{Z}_A$ is said *regular* if all transition rates are proportional to parameter A. Assume then that an LDP can be stated for $\mathbf{Z}_A$, with an action functional S_T defined on the metric space $\mathscr{C}_T(\mathbb{R}^d)$ of continuous $\mathbb{R}^d$-valued functions on interval $[0,T]$, $T \geqslant 0$. If process $\mathbf{Z}_A$ has a stationary distribution $\mathbf{\Pi}_A$, its decay rate

$$H(\mathbf{z}) = -\lim_{A\uparrow+\infty} \frac{1}{A} \cdot \log \mathbf{\Pi}_A(A\mathbf{z}), \qquad \mathbf{z} = (z_1, \ldots, z_d) \in \mathbb{R}^d, \tag{10}$$

is then obtained [8, Chap.5, 6] by minimizing functionals S_T, $T \geqslant 0$, on the whole union $\bigcup_{T\geqslant 0} \mathscr{C}_T(\mathbb{R}^d)$.

The scaling presently envisaged for process (N, M), however, is not regular since only $\beta = O(A)$ grows to infinity while α is kept fixed. This amounts to squeezing the time scale of the impatient customers arrival process, while keeping the initial time scale for the patient customers arrival flow. For this *singular* scaling, N is thus seen as a slow process driven by the fast variations of M.

Similar settings have been investigated in previous work, but none seems to directly apply to our problem. Given a homogeneous Markov chain $\mathbf{Y}$ with finite state space $\Gamma \subset \mathbb{N}$, consider the pair $\mathbf{Z}_A = (\mathbf{X}_A, \mathbf{Y}_A)$ where the fast process $\mathbf{Y}_A(t) = \mathbf{Y}(At)$, $t \geqslant 0$, drives the slow process $\mathbf{X}_A$ via the differential equation

$$\frac{\mathrm{d}\mathbf{X}_A}{\mathrm{d}t}(t) = \mathbf{b}(\mathbf{X}_A(t), \mathbf{Y}(At)), \qquad t \geqslant 0, \tag{11}$$

for a drift $\mathbf{b} : \mathbb{R}^{d-1} \times \Gamma \to \mathbb{R}^{d-1}$. Then:

- an LDP can be stated [8, Chap.7, Section 4] for the slow component $\mathbf{X}_A$ of $\mathbf{Z}_A = (\mathbf{X}_A, \mathbf{Y}_A)$, with an action functional S_T defined on space $\mathscr{C}_T(\mathbb{R}^{d-1})$;
- consider further the set $\mathscr{L}^{\Gamma}_{0,T}$ of mappings $\mathfrak{p} : (t, y) \in [0,T] \times \Gamma \mapsto \mathfrak{p}(t, y)$ such that $\mathfrak{p}(\cdot, y)$ is Borelian on $[0,T]$ for each $y \in \Gamma$, and the vector $(\mathfrak{p}(t, y))_{y\in\Gamma}$ is a probability on Γ for each $t \in [0,T]$. Let then the process $\mathfrak{P}_A$ be the random element of $\mathscr{L}^{\Gamma}_{0,T}$ defined by

  $$\mathfrak{P}_A(t, y) = \mathbf{1}_{Y_A(t)=y}, \qquad t \in [0,T],\ y \in \Gamma.$$

  An LDP for the pair $(\mathbf{X}_A, \mathfrak{P}_A)$ is then stated in [5, Theorem 2.3], with an action functional $\mathfrak{S}_T$ now defined on the product space $\mathscr{C}_T(\mathbb{R}^{d-1}) \times \mathscr{L}^{\Gamma}_{0,T}$.

The case where $\mathbf{X}_A$ is a diffusion process has also received attention. In [17], a general LDP is derived when both $\mathbf{X}_A$ and $\mathbf{Y}_A$ are diffusion processes while, closer to our case, [10] considers the case where $\mathbf{X}_A$ is a diffusion process and $\mathbf{Y}_A$ a finite-state space Markov chain. To our knowledge, however, no general LDP is known in the case when the slow component $\mathbf{X}_A$ is itself a Markov chain depending on the evolution of the fast driving chain $\mathbf{Y}_A$, both evolving with

increments of order $O(1/A)$, all the more since all previous results assume the finiteness of the state space Γ of the fast process $\mathbf{Y}$, which assumption fails for the process M presently considered.

2.2 Sharp Asymptotics via singular perturbation methods

LDP's concern the asymptotic behavior of stationary distributions on the logarithmic scale. In order to derive sharp (that is, not only logarithmic) asymptotics, we will now invoke singular perturbation methods. These methods have been justified for specific classes of problems:

- both a classification and rigorous foundation are established in [6, Chap.6] for some classical families of partial differential equations;

- the present case of jump processes has been considered in [20, Chap.4, 6] where asymptotics of the solutions of transient backward or forward Kolmogorov equations at finite time t are stated, but for a regular scaling only (in a different meaning to that introduced in Section 1.2 above, the two-time scales in [20] refer to either small $t = O(\varepsilon)$ or large $t = O(1/\varepsilon)$);

- asymptotic expansion for Laplace transforms have also been proven in the following context [7]. Consider a Markov process $\mathbf{Z}_A$ in $\mathbb{R}^d$, moving with a deterministic drift $\mathbf{b}$ and perturbed by a jump process with jump rates $O(A)$ and increments $O(1/A)$. Given a function $f \in \mathscr{C}^\infty(\mathbb{R}^d;\mathbb{R})$, let

$$\mathrm{F}_A(\mathbf{z},t) = \mathbb{E}\left(e^{-A\cdot f(\mathbf{Z}_A(T))} \,|\, \mathbf{Z}_A(t) = \mathbf{z}\right), \qquad \mathbf{z} \in \mathbb{R}^d,\ t \in [0,T].$$

Provided that the drift $\mathbf{b}$ belongs to $\mathscr{C}^\infty(\mathbb{R}^d;\mathbb{R}^d)$ and the transition distribution of $\mathbf{Z}_A$ satisfies boundedness and non degeneracy conditions, it is then shown [7, Theorem 5.1] that function F_A is $\mathscr{C}^\infty$ on $\mathbb{R}^d \times [0,T]$ and has the asymptotic expansion

$$\mathrm{F}_A(\mathbf{z},t) = \exp\left[-A\cdot G(\mathbf{z},t) - G_0(\mathbf{z},t) - \cdots - \frac{G_k(\mathbf{z},t)}{A^k} + O\left(\frac{1}{A^{k+1}}\right)\right] \quad (12)$$

for any $k \in \mathbb{N}$. Functions G and G_k, $k \geqslant 0$, are locally $\mathscr{C}^\infty$ and recursively obtained by solving partial differential equations. Expansion (12) applies, in particular, to the Laplace transform of $\mathbf{Z}_A(T)$ at finite time T by choosing $f(\mathbf{z}) = \langle \mathbf{u}, \mathbf{z}\rangle$ for given $\mathbf{u} \in \mathbb{R}^d$. If process $\mathbf{Z}_A$ has a stationary distribution, the Laplace transform of $\mathbf{Z}_A(\infty)$ is then deduced from (12) by letting $t \uparrow +\infty$.

Such expansions have been invoked and applied in other contexts, even if their existence is not formally stated. The analysis of coupled queuing systems is, in particular, one of the application fields of these perturbation methods (see [12, 13], [18, Chap.9] and references therein). In this framework, expansions of the form

$$\mathbf{\Pi}_A(A\mathbf{z}) = \frac{1}{(2\pi A)^{d/2}}\ \exp\left[-A\cdot H(\mathbf{z}) - h_0(\mathbf{z}) - \frac{h_1(\mathbf{z})}{A} - \cdots\right], \qquad \mathbf{z} \in \mathbb{R}^d, \quad (13)$$

for the stationary distribution $\mathbf{\Pi}_A$ are assumed to hold, with the decay rate H and functions h_0, h_1, ... successively determined via the Kolmogorov equations. While the existence of expansion (13) is admitted, the actual determination of unknown functions H, h_0, h_1, ... is considered as a consistent argument for its validity. To illustrate simply the approach developed in the latter references, consider the one-dimensional processes ($d = 1$) on the half-line $[0, +\infty)$. Using (13), the asymptotics of $\mathbf{\Pi}_A(Ax)$ for fixed $x > 0$ and for small $x = n/A$, $n = O(1)$, are shown to differ by some unknown multiplying constant; this constant is determined through the "Asymptotic Matching" principle [1, Chap.7, 7.4] which consists in identifying asymptotics of $\mathbf{\Pi}_A(Ax)$ and $\mathbf{\Pi}_A(n)$ when making x tend to 0 and n tend to $+\infty$, respectively.

To summarize this review, we can thus assert that LDP's with singular scaling are known for some specific classes of Markov processes, although not including the case of the birth-and-death process (N, M) presently considered. On the other hand, the analytical approach developed in the Singular Perturbation framework can be applied for classes of processes for which no LDP is known; assuming the existence of an asymptotic expansion of the form (13), this analytical approach then brings more precise information on the asymptotic behavior of their distribution. In this paper, admitting the existence of expansions such as (13), the analytical approach will thus be chosen to obtain the desired asymptotics for the stationary distribution $\mathbf{\Pi}_A$ of process (N, M) in the regime where A grows to infinity.

3 Preliminary results

In this section, we introduce a scaled version of distribution $\mathbf{\Pi}_A$ and explicit the Kolmogorov equation it satisfies. We also recall a Laplace expansion that will be used repeatedly.

3.1 Kolmogorov equations and scale change

Recall that $\mathbf{\Pi}_A$ denotes the stationary distribution of (N, M) when the stability condition (1) holds. By definition of its dynamics, it satisfies the associated set of Kolmogorov equations

$$\left[\alpha + \beta + \left(\frac{\mu\, n}{n+m} + \frac{\nu\, m}{n+m}\right)\mathbf{1}_{n+m>0} + \theta\, m\right]\mathbf{\Pi}_A(n,m) \;= \tag{14}$$

$$\alpha\,\mathbf{\Pi}_A(n-1,m)\mathbf{1}_{n>0} + \beta\,\mathbf{\Pi}_A(n,m-1)\mathbf{1}_{m>0} + \frac{\mu(n+1)}{n+m+1}\,\mathbf{\Pi}_A(n+1,m) +$$

$$(m+1)\left(\frac{\nu}{n+m+1} + \theta\right)\mathbf{\Pi}_A(n,m+1), \qquad (n,m) \in \mathbb{N}^2.$$

As explained in Section 1, in the heavy traffic regime $A \to \infty$, N and M become of the order of A and we will consequently study them on this scale. More precisely, we define the function $\mathbf{p}_A$ by

$$\mathbf{p}_A(x,y) = A^2 \cdot \mathbf{\Pi}_A([Ax],[Ay]), \qquad x,\ y \geqslant 0, \tag{15}$$

$[x] \in \mathbb{N}$ denoting the integer part of $x \in \mathbb{R}^+$. Linear system (14) translates for $\mathbf{p}_A$ into the following functional equations on the open quarter-plane $(0,+\infty)\times(0,+\infty)$ and its boundary $\{(x,0), x \geqslant 0\} \cup \{(0,y), y \geqslant 0\}$, namely

$$\begin{aligned}
&\left[\alpha + A\theta + \frac{\mu\, x}{x+y} + \frac{\nu\, y}{x+y} + A\theta\, y\right] \mathbf{p}_A(x,y) \;= \\
&\alpha\, \mathbf{p}_A\left(x - \frac{1}{A}, y\right) + A\theta\, \mathbf{p}_A\left(x, y - \frac{1}{A}\right) + \frac{\mu(Ax+1)}{A(x+y)+1}\, \mathbf{p}_A\left(x + \frac{1}{A}, y\right) \\
&+ (Ay+1)\left(\frac{\nu}{A(x+y)+1} + \theta\right) \mathbf{p}_A\left(x, y + \frac{1}{A}\right), \qquad\qquad x > 0,\;\; y > 0,
\end{aligned} \tag{16}$$

in the interior quarter-plane,

$$\begin{cases}
[\alpha + A\theta + \mu]\, \mathbf{p}_A(x,0) = \alpha\, \mathbf{p}_A\left(x - \dfrac{1}{A}, 0\right) + \mu\, \mathbf{p}_A\left(x + \dfrac{1}{A}, 0\right) \\
\qquad\qquad + \left(\dfrac{\nu}{Ax+1} + \theta\right) \mathbf{p}_A\left(x, \dfrac{1}{A}\right), \qquad x > 0, \\
[\alpha + A\theta + \nu + A\theta y]\, \mathbf{p}_A(0,y) = A\theta\, \mathbf{p}_A\left(0, y - \dfrac{1}{A}\right) + \dfrac{\mu}{Ay+1} \times \\
\quad \mathbf{p}_A\left(\dfrac{1}{A}, y\right) + (Ay+1)\left(\dfrac{\nu}{Ay+1} + \theta\right) \mathbf{p}_A\left(0, y + \dfrac{1}{A}\right), \qquad y > 0,
\end{cases} \tag{17}$$

on the boundary and

$$(\alpha + A\,\theta)\, \mathbf{\Pi}_A(0,0) = \frac{\mu}{A^2}\, \mathbf{p}_A\left(\frac{1}{A}, 0\right) + \frac{\nu + \theta}{A^2}\, \mathbf{p}_A\left(0, \frac{1}{A}\right), \tag{18}$$

at the origin, together with the normalization condition

$$\iint_{\mathbb{R}^{2+}} \mathbf{p}_A(x,y)\, \mathrm{d}x\, \mathrm{d}y = 1. \tag{19}$$

In the rest of the paper, we assume that the assumptions of Theorem 1 hold, that is, $\varrho < 1$ and the expansion (4) holds with functions H, h_0, h_1 and h_2 respectively of class $\mathscr{C}^3$, $\mathscr{C}^2$, $\mathscr{C}^1$ and $\mathscr{C}^0$ in the open quarter-plane $\mathbb{R}^{+*} \times \mathbb{R}^{+*}$, and function H being non-negative, continuous over the closed quarter plane $\mathbb{R}^+ \times \mathbb{R}^+$ and with $H(x^*, y^*) = 0$. In terms of density function $\mathbf{p}_A$ introduced in (15), the expansion (4) equivalently reads

$$\begin{aligned}
\mathbf{p}_A(x,y) = \frac{A}{2\pi} \times \\
\exp\left[-A \cdot H(x,y) - h_0(x,y) - \frac{h_1(x,y)}{A} - \frac{h_2(x,y)}{A^2} + O\left(\frac{1}{A^3}\right)\right]
\end{aligned} \tag{20}$$

for all $x,\; y > 0$.

Remark 1. *An explicit solution to system* (14) *seems out of reach for an arbitrary set of parameters α, μ, β, ν and θ. To obtain an efficient approximation for this stationary distribution $\mathbf{\Pi}_A$, a heuristic framework has been developed in [15] on the basis of the so-called "Quasi-Stationary approximation". Specifically, for any state $N = n \geqslant 0$ of the number of patient customers, this approximation assumes that the conditional distribution $\mathbf{D}(m \,|\, n) = \mathbb{P}(M(\infty) = m \,|\, N(\infty) = n)$, $m \in \mathbb{N}$, of $M(\infty)$, given $N(\infty) = n$, is evaluated by considering that the dynamics of process M is described by keeping the value of N constant in time. The Quasi-Stationary approximation proves, in particular, more robust than the direct numerical resolution of infinite system* (14). *This numerical stability is beneficial, in particular, in the high load regime when $\varrho = \alpha/\mu$ tends to 1.*

Remarkably, the functional $\mathfrak{S}_T$ arising in the LDP of Section 2.1 involves the Quasi-Stationary distribution $\mathbf{D}(\cdot \,|\, \mathbf{x})$ of $\mathbf{Y}$, when fixing the state $\mathbf{x}$ of the slow process $\mathbf{X}_A$.

3.2 Laplace expansion

We finally recall a classical Laplace expansion for an integral with exponential integrand and a large parameter A, which will be repeatedly used in the forthcoming sections. Given

- a real (possibly infinite) interval $[a, b]$,
- real-valued functions g and h on $[a, b]$ such that $h \in \mathscr{C}^2[a, b]$ has a unique minimum at the interior point $r^* \in (a, b)$ with $h''(r^*) \neq 0$,
- and $g \in \mathscr{C}^0[a, b]$ with $g(r^*) \neq 0$,

then [2, Section 5.3, Equ.(5.3.9)]

$$\int_a^b e^{-A\cdot h(r)} g(r)\,\mathrm{d}r = e^{-A\cdot h(r^*)} \sqrt{\frac{2\pi}{A\,h''(r^*)}}\, g(r^*) \left[1 + O\left(\frac{1}{A}\right)\right]. \tag{21}$$

Similar asymptotics hold for complex-valued integrals with the same conditions for both functions g and h, namely

$$\begin{aligned}\int_a^b & e^{-A\cdot h(r)+iA\zeta r} g(r)\mathrm{d}r \\ &= e^{-A\cdot h(r^*)+iA\zeta r^*} \cdot \exp\left(\frac{-A\zeta^2}{2h''(r^*)}\right) \sqrt{\frac{2\pi}{A\,h''(r^*)}}\, g(r^*) \left[1 + O\left(\frac{1}{A}\right)\right]\end{aligned} \tag{22}$$

for large A and any real constant ζ. When either function g or h depends smoothly on a real parameter σ, the $O(1/A)$ remainder in (21) or (22) tends to 0 when $A \uparrow +\infty$, uniformly with respect to σ pertaining to a given compact interval.

4 Proof of Theorem 2

Consider functional equation (16) for large A. Fix the point (x, y) with $x > 0$ and $y > 0$; expansion (20) applied at neighboring point $(x - 1/A, y)$ yields

$$\mathbf{p}_A\left(x - \frac{1}{A}, y\right) = \exp\left[-A \cdot H\left(x - \frac{1}{A}, y\right) - h_0\left(x - \frac{1}{A}, y\right) - \frac{1}{A} h_1\left(x - \frac{1}{A}, y\right) + \cdots\right]$$

(up to factor $A/2\pi$). Using the assumed smoothness of H, h_0 and h_1, Taylor expansions at first order in $1/A$ near point (x, y) give

$$\mathbf{p}_A\left(x - \frac{1}{A}, y\right) = \exp\left[-A\, H(x,y) - h_0(x,y) - \frac{1}{A} h_1(x,y) + \cdots\right] \times \exp\left[\frac{\partial H}{\partial x}(x,y) - \frac{1}{2A}\frac{\partial^2 H}{\partial x^2}(x,y) + \frac{1}{A}\frac{\partial h_0}{\partial x}(x,y) + \cdots\right],$$

dots denoting $O(1/A^2)$ terms. By (20) again, the first exponential factor in the right-hand side of the latter relation equals $\mathbf{p}_A(x, y)$; expanding the second exponential term at first order in $1/A$ then gives

$$\frac{\mathbf{p}_A(x - 1/A, y)}{\mathbf{p}_A(x, y)} = e^{+\partial_x H}\left(1 - \frac{1}{A}\left[\frac{1}{2}\frac{\partial^2 H}{\partial x^2} - \frac{\partial h_0}{\partial x}\right] + \cdots\right), \tag{23}$$

all derivatives being taken at point (x, y) ($\partial_x H$, $\partial_y H$ denote derivatives $\partial H/\partial x$ and $\partial H/\partial y$ for short, respectively). In a similar manner, we obtain the expansions of function $\mathbf{p}_A$ at neighboring points $(x, y - 1/A)$, $(x + 1/A, y)$ and $(x, y + 1/A)$ in the form

$$\begin{cases} \dfrac{\mathbf{p}_A(x, y - 1/A)}{\mathbf{p}_A(x, y)} = e^{+\partial_y H}\left(1 - \dfrac{1}{A}\left[\dfrac{1}{2}\dfrac{\partial^2 H}{\partial y^2} - \dfrac{\partial h_0}{\partial y}\right] + \cdots\right), \\ \dfrac{\mathbf{p}_A(x + 1/A, y)}{\mathbf{p}_A(x, y)} = e^{-\partial_x H}\left(1 - \dfrac{1}{A}\left[\dfrac{1}{2}\dfrac{\partial^2 H}{\partial x^2} + \dfrac{\partial h_0}{\partial x}\right] + \cdots\right), \\ \dfrac{\mathbf{p}_A(x, y + 1/A)}{\mathbf{p}_A(x, y)} = e^{-\partial_y H}\left(1 - \dfrac{1}{A}\left[\dfrac{1}{2}\dfrac{\partial^2 H}{\partial y^2} + \dfrac{\partial h_0}{\partial y}\right] + \cdots\right). \end{cases} \tag{24}$$

Inserting expressions (23)–(24) into equation (16) and dividing throughout by factor $\mathbf{p}_A(x,y)$, we then obtain

$$\alpha + A\,\theta + \frac{\mu\,x}{x+y} + \frac{\nu\,y}{x+y} + A\,\theta y = \alpha\cdot e^{\partial_x H}\left(1 - \frac{1}{A}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial x^2} - \frac{\partial h_0}{\partial x}\right] + \cdots\right) +$$
$$A\,\theta\cdot e^{\partial_y H}\left(1 - \frac{1}{A}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2} - \frac{\partial h_0}{\partial y}\right] + \cdots\right) +$$
$$\left[\frac{\mu\,x}{x+y} + \frac{\mu\,y}{A(x+y)^2} + \cdots\right] e^{-\partial_x H}\left(1 - \frac{1}{A}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial x^2} + \frac{\partial h_0}{\partial x}\right] + \cdots\right) +$$
$$\left[\frac{\nu\,y}{x+y} + \frac{\nu\,x}{A(x+y)^2} + \cdots + \theta(Ay+1)\right] e^{-\partial_y H}\left(1 - \frac{1}{A}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2} + \frac{\partial h_0}{\partial y}\right] + \cdots\right).$$

At order $O(A)$ and $O(1)$ for large A, the latter relation then entails

$$\theta + \theta\,y = \theta\cdot e^{\partial_y H} + \theta\,y\cdot e^{-\partial_y H} \tag{25}$$

and

$$\alpha + \frac{\mu\,x}{x+y} + \frac{\nu\,y}{x+y} = \alpha\cdot e^{\partial_x H} - \theta\,e^{\partial_y H}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2} - \frac{\partial h_0}{\partial y}\right]$$
$$+\ \frac{\mu\,x}{x+y}\cdot e^{-\partial_x H} + \left[\frac{\nu\,y}{x+y} + \theta\right] e^{-\partial_y H} - \theta\,y e^{-\partial_y H}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2} + \frac{\partial h_0}{\partial y}\right] \tag{26}$$

respectively. We then successively observe that

(A) Relation (25) is a quadratic equation for $e^{\partial_y H}$. We can exclude the trivial solution $e^{\partial_y H(x,y)} = 1$ which would give $\partial_y H(x,y) = 0$ and a solution H depending on variable x only. We are thus left with the other solution $e^{\partial_y H(x,y)} = y$, that is, $\partial_y H(x,y) = \log y$ for $y > 0$. Integrating with respect to variable y, the latter relation provides

$$H(x,y) = \widetilde{\Phi}(x) + \Psi(y), \qquad x > 0,\, y > 0, \tag{27}$$

for a function $\widetilde{\Phi}$ to be determined and with function Ψ given by (6).

(B) After (27), we have $\partial H/\partial y = \log y$ and $\partial^2 H/\partial y^2 = 1/y$ at point (x,y), $x > 0$, $y > 0$. Carrying over these values into equation (26), the latter solves for the first derivative $\partial h_0/\partial y$ into

$$\theta\,\frac{\partial h_0}{\partial y}(x,y) = \frac{1}{y-1}\left[\alpha(1 - e^{\widetilde{\Phi}'(x)}) + \frac{\mu\,x}{x+y}(1 - e^{-\widetilde{\Phi}'(x)})\right] + \frac{\nu}{x+y} + \frac{\theta}{2y}.$$

Integrating the latter equality with respect to variable y then yields

$$\theta\,h_0(x,y) = \theta\,\Omega(x) + \left[\alpha(1 - e^{\widetilde{\Phi}'(x)}) + \frac{\mu\,x}{x+1}(1 - e^{-\widetilde{\Phi}'(x)})\right]\log(y-1)$$
$$-\ \frac{\mu\,x}{x+1}(1 - e^{-\widetilde{\Phi}'(x)})\log(x+y) + \nu\log(x+y) + \frac{\theta}{2}\,\log y \tag{28}$$

for all $x > 0$, $y > 0$ and some unknown function Ω. By assumption, h_0 is continuously differentiable in the open quarter-plane and, in particular, on the vertical line $y = 1$. After relation (28), this implies that the coefficient of $\log(y-1)$ should vanish identically, hence

$$\alpha(1 - e^{\widetilde{\Phi}'(x)}) + \frac{\mu\, x}{x+1}(1 - e^{-\widetilde{\Phi}'(x)}) = 0$$

or, equivalently,

$$\alpha e^{2\,\widetilde{\Phi}'(x)} - \left(\alpha + \frac{\mu\, x}{x+1}\right) e^{\widetilde{\Phi}'(x)} + \frac{\mu\, x}{x+1} = 0, \qquad x > 0.$$

This quadratic equation for $e^{\widetilde{\Phi}'(x)}$ has the non-constant ($\neq 1$) solution

$$e^{\widetilde{\Phi}'(x)} = \frac{\mu x}{\alpha(x+1)} = \frac{x}{\varrho(x+1)} \tag{29}$$

which differential equation readily integrates for $\widetilde{\Phi}$ into

$$\widetilde{\Phi}(x) = x\log x - (x+1)\log(x+1) - x\log\varrho + C_0, \qquad x > 0, \tag{30}$$

for some constant C_0. As $\Psi(y^*) = \Psi(1) = 0$, the assumption $H(x^*, y^*) = 0$ on H then implies that $\widetilde{\Phi}(x^*) = 0$ with x^* introduced in (3); this readily determines the value $C_0 = -\log(1-\varrho)$. The latter and (30) thus entirely determine the function $\widetilde{\Phi}$, which is thus equal to Φ defined by (5). The final expression of decay rate $H = \Phi + \Psi$ in the open quarter-plane $\mathbb{R}^{+*} \times \mathbb{R}^{+*}$ follows. Since H is assumed to be continuous on the closed quarter plane, this expression extends by continuity to $\mathbb{R}^+ \times \mathbb{R}^+$, which concludes the proof of Theorem 2.

Remark 2. *Equation* (25) *is the so-called Hamilton-Jacobi equation for the component M [8, Chap.5, Theorem 4.3] which determines the partial derivative $\partial H/\partial y$ only. In the present singular Large Deviations setting, however, the full derivation of function H requires another partial differential equation for the next function h_0, together with its smoothness across the line $y = y^* = 1$.*

5 Proof of Theorem 1

We now determine the prefactor h_0 in the expansion (20) of density $\mathbf{p}_A$. Given the expression (30) of function $\tilde{\Phi}$, formula (28) for function h_0 now easily reduces to

$$h_0(x,y) = \Omega(x) + \frac{\mu}{\theta}(1-\varrho)\left(\frac{x-x^*}{x+1}\right)\log(x+y) + c\log(x+y) + \frac{\log y}{2} \tag{31}$$

for $x > 0$, $y > 0$, with $c = \nu/\theta$, x^* introduced in (3) and some unknown function Ω. In order to specify Ω, we evaluate terms of subsequent order $O(1/A)$ in the functional equation (16) for $x > 0$ and $y > 0$.

To this end, expansions (24) for both $\mathbf{p}_A(x, y-1/A)$ and $\mathbf{p}_A(x, y+1/A)$ have to be extended up to order $O(1/A^2)$. Besides, the expansions for $\mathbf{p}_A(x\pm 1/A, y)$ at order $O(1/A)$ only are still sufficient. Applying then (20) at point $(x, y-1/A)$, we have

$$\mathbf{p}_A\left(x, y-\frac{1}{A}\right) = \exp\Big[\, -A\cdot H\left(x, y-\frac{1}{A}\right) - h_0\left(x, y-\frac{1}{A}\right) - \frac{1}{A}\, h_1\left(x, y-\frac{1}{A}\right) - \frac{1}{A^2}\, h_2\left(x, y-\frac{1}{A}\right) + \cdots\Big];$$

(up to multiplying factor $A/2\pi$). Writing Taylor expansions at second order in $1/A$ for functions H, h_0, h_1, h_2, ... near point (x, y), we then easily obtain

$$\mathbf{p}_A\left(x, y-\frac{1}{A}\right) = \exp\left[-A\, H(x,y) - h_0(x,y) - \frac{1}{A}\, h_1(x,y) - \frac{1}{A^2}\, h_2\,(x,y) + \cdots\right] \times$$
$$e^{\partial_y H} \exp\left[-\frac{1}{2A}\frac{\partial^2 H}{\partial y^2} + \frac{1}{6A^2}\frac{\partial^3 H}{\partial y^3} + \cdots + \frac{1}{A}\frac{\partial h_0}{\partial x} - \frac{1}{2A^2}\frac{\partial^2 h_0}{\partial y^2} + \cdots + \frac{1}{A^2}\frac{\partial h_1}{\partial y} + \cdots\right],$$

all derivatives being taken at point (x, y) and dots denoting $O(1/A^3)$ terms. By expansion (20) again, the first exponential factor in the right-hand side of the latter equality equals $\mathbf{p}_A(x, y)$ (up to $A/2\pi$). Expanding the second exponential term in the right-hand side at second order in $1/A$ then provides

$$\frac{\mathbf{p}_A(x, y-1/A)}{\mathbf{p}_A(x, y)} = e^{+\partial_y H}\Big(1 - \frac{1}{A}\left[\frac{1}{2}\frac{\partial^2 H}{\partial y^2} - \frac{\partial h_0}{\partial y}\right] + \frac{1}{A^2}\left\{\frac{1}{2}\left[\frac{1}{2}\frac{\partial^2 H}{\partial y^2} - \frac{\partial h_0}{\partial y}\right]^2 + \frac{1}{6}\frac{\partial^3 H}{\partial y^3} - \frac{1}{2}\frac{\partial^2 h_0}{\partial y^2} + \frac{\partial h_1}{\partial y}\right\} + \cdots\Big). \quad (32)$$

At neighboring point $(x, y+1/A)$, a similar calculation yields

$$\frac{\mathbf{p}_A(x, y+1/A)}{\mathbf{p}_A(x, y)} = e^{-\partial_y H}\Big(1 - \frac{1}{A}\left[\frac{1}{2}\frac{\partial^2 H}{\partial y^2} + \frac{\partial h_0}{\partial y}\right] + \frac{1}{A^2}\left\{\frac{1}{2}\left[\frac{1}{2}\frac{\partial^2 H}{\partial y^2} + \frac{\partial h_0}{\partial y}\right]^2 - \frac{1}{6}\frac{\partial^3 H}{\partial y^3} - \frac{1}{2}\frac{\partial^2 h_0}{\partial y^2} - \frac{\partial h_1}{\partial y}\right\} + \cdots\Big). \quad (33)$$

Inserting then expansions (32), (33) and retaining terms of order $1/A$ in the

identity following (24) in the proof of Theorem 2, we then obtain the equation

$$\begin{aligned} 0 = & -\alpha\, e^{\partial_x H}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial x^2}-\frac{\partial h_0}{\partial x}\right] \\ & +\theta\, e^{\partial_y H}\left\{\frac{1}{2}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2}-\frac{\partial h_0}{\partial y}\right]^2+\frac{1}{6}\,\frac{\partial^3 H}{\partial y^3}-\frac{1}{2}\,\frac{\partial^2 h_0}{\partial y^2}+\frac{\partial h_1}{\partial y}\right\} \\ & -\frac{\mu\, x}{x+y}\,e^{-\partial_x H}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial x^2}+\frac{\partial h_0}{\partial x}\right]+\frac{\mu\, y}{(x+y)^2}\,e^{-\partial_x H} \\ & -\left(\frac{\nu\, y}{x+y}+\theta\right)e^{-\partial_y H}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2}+\frac{\partial h_0}{\partial y}\right]+\frac{\nu\, x}{(x+y)^2}\,e^{-\partial_y H} \\ & +\theta y\, e^{-\partial_y H}\left\{\frac{1}{2}\left[\frac{1}{2}\,\frac{\partial^2 H}{\partial y^2}+\frac{\partial h_0}{\partial y}\right]^2-\frac{1}{6}\,\frac{\partial^3 H}{\partial y^3}-\frac{1}{2}\,\frac{\partial^2 h_0}{\partial y^2}-\frac{\partial h_1}{\partial y}\right\} \end{aligned} \tag{34}$$

involving $\partial h_0/\partial x$ and $\partial h_1/\partial y$. The derivative $\partial h_0/\partial x$ intervenes in (34) in the first and third brackets only, with multiplying coefficient

$$K(x,y)=\alpha\, e^{\partial_x H}-\frac{\mu\, x}{x+y}\,e^{-\partial_x H}=\mu\,\frac{x(x+y)-\varrho(x+1)^2}{(x+1)(x+y)}, \tag{35}$$

after using expression (29) for $\widetilde{\Phi}'(x)=\partial H(x,y)/\partial x$. Calculating the derivatives $\partial H/\partial y=\log y$, $\partial^2 H/\partial y^2=1/y$, $\partial^3 H/\partial y^3=-1/y^2$ together with

$$\frac{\partial h_0}{\partial y}(x,y)=\frac{c}{x+y}+\frac{1}{2y},\quad \frac{\partial^2 h_0}{\partial y^2}(x,y)=-\frac{c}{(x+y)^2}-\frac{1}{2y^2}$$

after (31), equation (34) then reads

$$K(x,y)\frac{\partial h_0}{\partial x}(x,y)-L(x,y)=\theta(1-y)\frac{\partial h_1}{\partial y}(x,y),\qquad x>0,\ y>0, \tag{36}$$

when isolating each derivative $\partial h_0/\partial x$, $\partial h_1/\partial y$ and setting

$$\begin{aligned} L(x,y)= & \frac{\mu}{2(x+1)^2}+\frac{\alpha}{2x(x+y)}-\frac{\alpha(x+1)y}{x(x+y)^2}-\frac{\nu\, x}{(x+y)^2 y}-\frac{c(1+c)\theta y}{2(x+y)^2} \\ & -\frac{\theta}{12y}-\theta\left[\frac{c(1+c)}{2(x+y)^2}+\frac{c}{(x+y)y}+\frac{11}{12\,y^2}\right]+\frac{1}{y}\left(\frac{\nu\, y}{x+y}+\theta\right)\left(\frac{1}{y}+\frac{c}{x+y}\right). \end{aligned} \tag{37}$$

By assumption, h_1 is of class $\mathscr{C}^1$ in the open quarter-plane and, in particular, on the vertical line $y=y^*=1$. In view of functional relation (36), this implies that its left-hand side should identically vanish for $y=y^*=1$, that is,

$$\forall\, x>0,\qquad \frac{\partial h_0}{\partial x}(x,1)=\frac{L(x,1)}{K(x,1)}. \tag{38}$$

By expressions (35) and (37) of $K(x,y)$ and $L(x,y)$, elementary algebra provides

$$K(x,1) = \mu(1-\varrho)\,\frac{x-x^*}{x+1}, \qquad L(x,1) = \mu(1-\varrho)\,\frac{x-x^*}{2x(x+1)^2}$$

(note that both rational fractions $K(x,1)$ and $L(x,1)$ have a simple zero at $x = x^*$ so that the ratio $L(x,1)/K(x,1)$ is well-defined for all $x > 0$). Using the latter, (38) then gives $\partial h_0(x,1)/\partial x = 1/\left[2x(x+1)\right]$, $x > 0$ which readily integrates to

$$h_0(x,1) = C_0 + \log\sqrt{\frac{x}{x+1}}, \qquad x > 0, \tag{39}$$

for some constant C_0. Besides, expression (31) for $h_0(x,y)$ readily shows that the difference $h_0(x,y) - h_0(x,1)$ is independent of the function Ω and equals

$$h_0(x,y) - h_0(x,1) = \frac{\log y}{2} + \left[c + \frac{\mu}{\theta}(1-\varrho)\,\left(\frac{x-x^*}{x+1}\right)\right]\log\left(\frac{x+y}{x+1}\right). \tag{40}$$

Using relation (40), we thus deduce that $h_0(x,y) = h_0(x,1) + (h_0(x,y) - h_0(x,1))$ eventually equals

$$h_0(x,y) = C_0 + \log\sqrt{\frac{x}{x+1}} + \frac{\log y}{2} + \left[c + \frac{\mu}{\theta}(1-\varrho)\,\left(\frac{x-x^*}{x+1}\right)\right]\log\left(\frac{x+y}{x+1}\right)$$

for $x > 0$, $y > 0$. At first order in $1/A$, the expansion (20) for density $\mathbf{p}_A$ in the interior quarter plane therefore reads

$$\begin{aligned}\mathbf{p}_A(x,y) \sim\ & \frac{A\,e^{-C_0}}{2\pi\sqrt{y}}\,e^{-A\cdot H(x,y)}\,\sqrt{\frac{x+1}{x}} \\ & \times\ \exp\left[\left\{c + \frac{\mu}{\theta}(1-\varrho)\left(\frac{x-x^*}{x+1}\right)\right\}\log\left(\frac{x+1}{x+y}\right)\right], \qquad x > 0,\ y > 0,\end{aligned} \tag{41}$$

which determines $\mathbf{p}_A$ in the interior quarter plane, up to constant e^{-C_0}. The latter is determined by condition (19), once written as $\int_{\mathbb{R}^{+*}\times\mathbb{R}^{+*}} \mathbf{p}_A(x,y)\mathrm{d}x\,\mathrm{d}y \sim 1$; using (41) and applying asymptotics (21) successively to the integral with respect to variable x and to variable y in the latter, we obtain $e^{-C_0} = 1-\varrho$. After the definition (15) of $\mathbf{\Pi}_A$ in terms of $\mathbf{p}_A$, expression (7) eventually follows. This concludes the proof of Theorem 1, from which we can deduce the following corollary.

Corollary 1. *Given the assumptions of Theorem 1, the marginal stationary distributions $N(\infty)$ and $M(\infty)$ are respectively asymptotic to*

$$\begin{cases} \mathbb{P}(N(\infty) = Ax) \sim \dfrac{1-\varrho}{\sqrt{2\pi A}}\,\sqrt{\dfrac{x+1}{x}}\,e^{-A\cdot\Phi(x)}, & x > 0, \\[2ex] \mathbb{P}(M(\infty) = Ay) \sim \dfrac{1}{\sqrt{2\pi A\,y}}\,\dfrac{e^{-A\cdot\Psi(y)}}{(1-\varrho)^c(x^*+y)^c}, & y > 0, \end{cases} \tag{42}$$

for large A.

Proof. Integrating expression (7) with respect to variable $y > 0$ and applying the Laplace expansion (21) at the unique minimum of function Ψ at point $y = y^*$, asymptotics (42) for $N(\infty)$ follows. Integrating in turn (7) with respect to variable $x > 0$ and applying Laplace expansion (21) at the unique minimum of function Φ at point $x = x^*$, asymptotics (42) for $M(\infty)$ is similarly derived from Theorem 1. □

Remark 3. *In a way similar to that used in this Section, sharp asymptotics of density* $\mathbf{p}_A$ *on the boundary* $\{(x,0), x \geqslant 0\} \cup \{(0,y), y \geqslant 0\}$ *could be derived from equations* (17)–(18)*. Such evaluations are not needed in the present study and we only sketch the resolution procedure. For* $x > 0$ *and* $y > 0$*, asymptotic matching arguments can be first invoked to set*

$$\mathbf{p}_A(x,0) \sim \frac{A^{\frac{3}{2}}}{2\pi} \cdot \varphi(x)\, e^{-A(\Phi(x)+\Psi(0))}, \quad \mathbf{p}_A(0,y) \sim \frac{A^{\frac{3}{2}}}{2\pi} \cdot \psi(x)\, e^{-A(\Phi(0)+\Psi(y))}$$

for some functions φ*,* ψ*, together with*

$$\mathbf{p}_A\left(x, \frac{1}{A}\right) \sim \frac{A^{\frac{5}{2}}}{2\pi} \cdot \varphi_1(x)\, e^{-A(\Phi(x)+\Psi(0))}, \quad \mathbf{p}_A\left(\frac{1}{A}, y\right) \sim \frac{A^{\frac{5}{2}}}{2\pi} \cdot \psi_1(y)\, e^{-A(\Phi(0)+\Psi(y))}$$

where φ_1*,* ψ_1 *can be derived from* (7)*. Each equation* (17) *then provides the respective solution for* φ *and* ψ *by identifying* $O(1)$ *terms for large* A*. The last equation* (18) *gives the final asymptotics for* $\mathbf{\Pi}_A(0,0)$*.*

6 Proof of Theorem 3

Define the generating function F_A of the pair $(N(\infty), M(\infty))$ by

$$\mathrm{F}_A(u,v) = \mathbb{E}\left(u^{N(\infty)} v^{M(\infty)}\right), \qquad (u,v) \in \mathbb{D} \times \mathbb{D}, \tag{43}$$

where $\mathbb{D}$ is the open unit disk. The sharp asymptotics for $\mathbf{\Pi}_A$ stated in Theorem 1 in the interior quarter-plane $\mathbb{R}^{+*} \times \mathbb{R}^{+*}$ are now applied to obtain estimates for generating function F_A in a relevant domain. As a preamble, we first show that F_A has an analytic continuation from the product $\mathbb{D} \times \mathbb{D}$ to a larger domain containing a neighborhood of point $(u,v) = (1,1)$.

Lemma 1. *Given* $\varrho < 1$*, the generating function* F_A *can be analytically extended to the product domain*

$$\mathbf{\Omega} = \mathbb{D}\left(0, \frac{1}{\varrho}\right) \times \mathbb{C} \tag{44}$$

where $\mathbb{D}(0, \frac{1}{\varrho})$ *is the open disk centered at* $u = 0$ *and with radius* $1/\varrho$*.*

The proof is detailed in Appendix B. The main steps sum up as follows: a sample path property of process M first ensures the existence of $F_A(u,v)$ for all $(u,v) \in \mathbb{D} \times \mathbb{C}$; an estimate of the marginal distribution of $N(\infty)$ justifies in

turn the existence for $(u,v) \in \mathbb{D}\,(0,1/\varrho) \times \mathbb{D}$; finally, a convexity property of the convergence domain of power series $\mathrm{F}_A(u,v)$ concludes for its finiteness over $\mathbf{\Omega}$.

Now, consider the open subset $\mathbf{\Omega'} \subset \mathbf{\Omega}$ defined by

$$\mathbf{\Omega'} = \mathbf{\Omega} \setminus \{u,\ u \in (-1/\varrho, 0]\} \times \{v,\ v \in\ (-\infty, 0]\}$$

(we have thus excluded the non positive real points (u,v) from $\mathbf{\Omega}$). Using Theorems 2 and 1, we can then assert the following.

Proposition 1. *Given $\varrho < 1$ and the assumptions of Theorem 1, the generating function F_A of the pair $(N(\infty), M(\infty))$ is asymptotic for large A to*

$$\mathrm{F}_A(u,v) \sim \left(\frac{1-\varrho}{1-\varrho r}\right)^A e^{A(s-1)} \exp\left[iA\left(\frac{\varrho r\zeta}{1-\varrho r} + s\eta\right)\right] \times \exp\left[-\frac{A}{2}\left(\frac{\varrho r\zeta^2}{(1-\varrho r)^2} + s\eta^2\right)\right] G_0(u,v) \quad (45)$$

for $(u,v) \in \mathbf{\Omega'}$, where we set $u = r\,e^{i\zeta}$, $0 < r < 1/\varrho$, $\zeta \in\ (-\pi, \pi)$, and $v = s\,e^{i\eta}$, $s > 0$, $\eta \in\ (-\pi, \pi)$, respectively and where the continuous function G_0 is given by

$$G_0(u,v) = \left(\frac{1-\varrho}{1-\varrho r}\right)\left[s + \varrho r(1-s)\right]^{\frac{\alpha}{\theta}(1-r)-c}$$

with $G_0(1,1) = 1$.

The proof is detailed in Appendix C.

Remark 4. *After the general result* (22)*, asymptotics* (45) *can be also specified by stating a remainder term of order $O(1/A)$ for large A which tends to 0, uniformly with respect to variable (u,v) pertaining to any compact subset of domain $\mathbf{\Omega'}$.*

We can now proceed with the proof of Theorem 3.

Proof of Theorem 3. First address the weak convergence of the pair (ξ_A, η_A). Let L_A be the characteristic function of random variable (ξ_A, η_A); we have

$$\begin{aligned}\mathrm{L}_A(\sigma,\tau) &= \mathbb{E}\left(\exp\left[i\sigma\sqrt{A}\left(\frac{N(\infty)}{A} - x^*\right) + i\tau\sqrt{A}\left(\frac{M(\infty)}{A} - 1\right)\right]\right) \quad (46)\\ &= e^{-i\sqrt{A}(\sigma\,x^* + \tau)}\,\mathrm{F}_A(e^{\frac{i\sigma}{\sqrt{A}}}, e^{\frac{i\tau}{\sqrt{A}}})\end{aligned}$$

for all $(\sigma,\tau) \in \mathbb{R}^2$, with $i^2 = -1$. Apply then Proposition 1 to the point (u,v) with $u = \exp(i\sigma/\sqrt{A})$ and $v = \exp(i\tau/\sqrt{A})$, pertaining to a neighborhood of $(1,1)$ for large enough A. We clearly have $r = |u| = 1$ and $s = |v| = 1$, so that $G_0(u,v) = 1$ and asymptotics (45) presently reduces to

$$\begin{aligned}&\mathrm{F}_A(e^{\frac{i\sigma}{\sqrt{A}}}, e^{\frac{i\tau}{\sqrt{A}}}) =\\ &\exp\left[iA\,(x^*\sigma + \tau)\,\frac{1}{\sqrt{A}}\right] \cdot \exp\left[-\frac{A}{2}\left(\frac{\sigma^2}{A}\,\frac{\varrho}{(1-\varrho)^2} + \frac{\tau^2}{A}\right)\right] \times \left[1 + O\left(\frac{1}{A}\right)\right]\end{aligned}$$

where, after Remark 4, the remainder term $O(1/A)$ tends to 0 uniformly in variables (σ, τ) (in fact, the pair $(u, v) = (e^{i\sigma/\sqrt{A}}, e^{i\tau/\sqrt{A}})$ pertains to a compact neighborhood of point $(1, 1) \in \mathbf{\Omega}'$ for large enough A). It then follows that

$$\mathrm{F}_A(e^{\frac{i\sigma}{\sqrt{A}}}, e^{\frac{i\tau}{\sqrt{A}}}) \sim e^{+i\sqrt{A}(\sigma\, x^* + \tau)} \, \exp\left[-\frac{1}{2}\left(\frac{\varrho\sigma^2}{(1-\varrho)^2} + \tau^2\right)\right] \tag{47}$$

for large A and any given $(\sigma, \tau) \in \mathbb{R}^2$. By equality (46) and estimate (47), we thus derive that

$$\lim_{A\uparrow+\infty} \mathrm{L}_A(\sigma, \tau) = \exp\left[-\frac{1}{2}\left(\frac{\varrho\sigma^2}{(1-\varrho)^2} + \tau^2\right)\right], \qquad (\sigma, \tau) \in \mathbb{R},$$

which limit defines the characteristic function of the Gaussian distribution with covariance matrix given in Theorem 3. By Lévy's continuity Theorem [11, Chap.19, Theorem 19.1], we conclude that the scaled random variable (ξ_A, η_A) converges weakly towards this Gaussian distribution.

Finally consider the estimation of expectations $\mathbb{E}(N(\infty))$ and $\mathbb{E}(M(\infty))$. Note that, in general, the latter weak convergence of (ξ_A, η_A) does not necessarily imply that $\mathbb{E}(\xi_A) \to \mathbb{E}(\xi)$ and $\mathbb{E}(\eta_A) \to \mathbb{E}(\eta)$. Presently, however, we can directly rely on the asymptotics of Corollary (1) for the marginal distributions of $N(\infty)$ to write

$$\mathbb{E}(N(\infty)) = \sum_{n\geqslant 0} n\, \mathbb{P}(N(\infty) = n) \sim \int_0^{+\infty} (A\, x)\, \mathbb{P}(N(\infty) = A\, x)\, A\, \mathrm{d}x$$

for large A, after estimating the discrete sum by a Riemann integral with integral step $1/A$; using asymptotics (42) for $\mathbb{P}(N(\infty) = Ax)$, the latter then entails

$$\mathbb{E}(N(\infty)) \sim A^2 \left(\frac{1-\varrho}{\sqrt{2\pi A}}\right) \int_0^{+\infty} \sqrt{x(x+1)}\, e^{-A\cdot\Phi(x)}\, \mathrm{d}x.$$

Applying the Laplace asymptotics (21) to the latter integral, with the minimum of Φ located at $x = x^*$ with $\Phi(x^*) = \Phi'(x^*) = 0$ and $\Phi''(x^*) = (1-\varrho)^2/\varrho$, we readily obtain $\mathbb{E}(N(\infty)) \sim A\, x^*$ as claimed. A similar calculation provides $\mathbb{E}(M(\infty)) \sim A\, y^*$ for large A. □

7 Conclusion

In this paper, sharp large deviations asymptotics and limit theorems for the stationary queue occupancy distribution have been derived for the Processor-Sharing queue with both patient and impatient customers, in the case when the normalized arrival rate A of impatient customers grows to infinity. On mathematical ground, the asymptotic setting is a new case of *singular pertubation* for the underlying bi-dimensional birth-and-death process where the time scale of one component is accelerated while that of the other component is kept fixed. As

no general large deviations principle is available for such a Markov process with discrete state space, the sharp asymptotics have been obtained by assuming an expansion of the form

$$\mathbf{p}_A = \frac{A}{2\pi}\, e^{-A\cdot H}\, \left(g + \frac{g_1}{A} + \cdots\right), \qquad A \uparrow +\infty,$$

for the scaled solution $\mathbf{p}_A$ to Kolmogorov equations. We have shown how unknown functions H, g, ... can be iteratively determined.

These results have been applied to the *closed-loop* PS queue fed back by the flow of impatient customers with still uncompleted service. Unlike the common queueing systems with growth $1/(1-\varrho_{\mathtt{tot}})^{\alpha}$ in high load condition for some $\alpha > 0$, this closed-loop PS queue has been shown to exhibit a slower logarithmic growth $-\log(1-\varrho_{\mathtt{tot}})$ in the high load regime. In performance terms, the account of the so-called moving users is beneficial to the system behavior and the throughput of each user class decays less fast in case of congestion, as per estimates (9).

The present approach offers generalizations when extended to queuing systems with a state space with higher dimension. Specifically, consider the PS queue with a number K of patient or impatient customer classes, with arrival rate α_k, service rate μ_k and impatient rate $\theta_k \geqslant 0$ for class $k \in \{1,\ldots,K\}$. This system should be amenable to the techniques applied in the present paper when the arrival rate α_k, with $\theta_k \neq 0$, of some class k of impatient customers tends to infinity proportionally to a dimensionless parameter A. While the present approach has directly considered asymptotics for the solution of the Kolmogorov equations in dimension $K = 2$, an alternative approach for $K > 2$ consists in deriving asymptotics for the generating function F_A of the queue occupancy $(N_1,\ldots,N_K)$. In fact, it can be easily shown from system (14) that F_A verifies the integro-differential equation

$$\left[\sum_{k=1}^{K} \alpha_k(1-u_k)\right] \mathrm{F}_A(\mathbf{u}) + \sum_{k=1}^{K} \theta_k\,(u_k-1)\frac{\partial \mathrm{F}_A}{\partial u_k}(\mathbf{u}) \;=$$
$$\int_0^1 \left[\sum_{k=1}^{K} \mu_k(1-u_k)\frac{\partial \mathrm{F}_A}{\partial u_k}(t\mathbf{u})\right] \mathrm{d}t, \qquad \mathbf{u} = (u_1,\ldots,u_K) \in \mathbb{D}^K,$$

with $\mathrm{F}_A(1,\ldots,1) = 1$. In some extended analyticity domain $\boldsymbol{\Omega} \supset \mathbb{D}^K$, an expansion

$$\mathrm{F}_A = e^{-A\cdot G}\, \left(G_0 + \frac{G_1}{A} + \cdots\right)$$

for F_A could then be determined through the latter equation and provide general information on the corresponding multivariate queue distribution.

A Derivation of decay rate K

In this appendix, we prove that the component Φ of H is the decay rate related to the single-server PS queue with A permanent customers, arrival rate α and

service rate μ, as was claimed in Section 1. More generally, assume $\varrho = \alpha/\mu < 1$ and let $\mathbf{E}_m$ denote the stationary distribution of the single-server PS queue with a fixed number m of permanent customers in queue, arrival rate α and service rate μ.

Lemma 2. *Consider $x = O(1)$ and $y = O(1)$. We then have*

$$\lim_{A\uparrow+\infty} \frac{1}{A} \cdot \log \mathbf{E}_{Ay}(Ax) = -K(x,y) \tag{48}$$

where

$$K(x,y) = x \log\left(\frac{x}{\varrho}\right) + y \log y - (x+y)\log(x+y) - y\log(1-\varrho).$$

Since $K(x,1) = \Phi(x)$ for $y = 1$, this indeed shows that Φ is the decay rate of the single-server PS queue with A permanent customers.

Proof of Lemma 2. By a simple reversibility argument for the Markov chain representing the queue occupancy, we first have

$$\mathbf{E}_m(n) = \varrho^n \prod_{k=1}^{n} \left(1 + \frac{m}{k}\right) \times \mathbf{E}_m(0), \qquad n \in \mathbb{N}, \tag{49}$$

with $\mathbf{E}_m(0)$ given by the normalization condition. More precisely, $\mathbf{E}_m(0) = 1/R_m(\varrho)$ where

$$R_m(z) = \sum_{n\geqslant 0} \frac{z^n}{n!} \prod_{k=1}^{n} (k+m) = \frac{1}{(1-z)^{m+1}}, \qquad 0 < z < 1, \tag{50}$$

hence $\mathbf{E}_m(0) = (1-\varrho)^{m+1}$. Now address the estimation of $\mathbf{E}_{Ay}(Ax)$ for large A and fixed $x > 0$, $y > 0$. The logarithm of the product

$$\mathbf{W}_{Ay}(Ax) = \prod_{1\leqslant k\leqslant Ax} \left(1 + \frac{Ay}{k}\right)$$

involved in expression (49) where $n = Ax$ and $m = Ay$, with $x = O(1)$ and $y = O(1)$, can be written as the sum

$$\log \mathbf{W}_{Ay}(Ax) = T(Ay) + Ay \sum_{k=1}^{Ax} \frac{1}{k} - \sum_{k=Ax+1}^{+\infty} \mathbf{g}_{Ay}(k) \tag{51}$$

where we set

$$\mathbf{g}_m(u) = \log\left(1 + \frac{m}{u}\right) - \frac{m}{u}, \qquad T(z) = \sum_{j\geqslant 1} \mathbf{g}_z(j)$$

for $u \in [1, +\infty)$ and $z > 1$, respectively. We successively evaluate each term of the right-hand side of (51) for large A:

a) by the Weierstrass product formula [16, Sect. 5.8.2], the sum $T(z)$ can be first made explicit in terms of the Γ function only, namely

$$T(z) = -\log\Gamma(z) - \gamma z - \log z, \qquad z > 0,$$

γ denoting the Euler constant. Using this expression of $T(z)$ and the Stirling's asymptotic formula $\log\Gamma(z) = z\log z - z - (\log z)/2 + \log\sqrt{2\pi} + o(1)$ for large positive z [16, Sect. 5.11.1], we thus obtain

$$T(Ay) = -Ay\cdot\log(Ay) + (1-\gamma)Ay - \frac{1}{2}\log(Ay) - \log\sqrt{2\pi} + o(1); \tag{52}$$

b) besides, the second term in the right-hand side of (51) is proportional to the harmonic sum, which is known to expand as [16, Sect. 2.10.8]

$$\sum_{j=1}^{Ax}\frac{1}{j} = \log(Ax) + \gamma + \frac{1}{2Ax} + o(1); \tag{53}$$

c) finally, the last sum in the right-hand side of (51) can be written via the Euler-MacLaurin formula [16, Sect. 2.10.1] in the form

$$\sum_{k=Ax+1}^{+\infty}\mathbf{g}_{Ay}(k) = \int_{Ax+1}^{+\infty}\mathbf{g}_{Ay}(u)\mathrm{d}u + \frac{\mathbf{g}_{Ay}(+\infty)+\mathbf{g}_{Ay}(Ax+1)}{2} + \frac{1}{12}\left(\mathbf{g}'_{Ay}(+\infty) - \mathbf{g}'_{Ay}(Ax+1)\right) + \cdots \tag{54}$$

From the derivative $\mathbf{g}'_m(u) = m^2/u^2(u+m)$, $u \geqslant 1$, we have $\mathbf{g}'_{Ay}(+\infty) = 0$ and $\mathbf{g}'_{Ay}(Ax+1) = o(1)$. Besides, calculating the integral in the right-hand side of (54) gives

$$\begin{aligned}\int_{Ax+1}^{+\infty}\mathbf{g}_{Ay}(u)\mathrm{d}u &= \Big[u\,\mathbf{g}_{Ay}(u)\Big]_{u=Ax+1}^{+\infty} - \int_{Ax+1}^{+\infty}u\,\mathbf{g}'_{Ay}(u)\mathrm{d}u \\ &= -(Ax+1)\mathbf{g}_{Ay}(Ax+1) - Ay\cdot\log\left(\frac{Ax+Ay+1}{Ax+1}\right)\end{aligned} \tag{55}$$

by using an integration by parts along with the previous expression of $\mathbf{g}'_m(u)$; furthermore, the factor $\mathbf{g}_{Ay}(Ax+1)$ in the right-hand side of (55) expands as

$$\mathbf{g}_{Ay}(Ax+1) = \log\left(1+\frac{y}{x}\right) - \frac{y}{x} + \frac{y^2}{x^2(x+y)A} + o\left(\frac{1}{A}\right);$$

gathering expression (55) and the former results, the sum (54) can consequently be evaluated as

$$\begin{aligned}\sum_{k=Ax+1}^{+\infty}\mathbf{g}_{Ay}(k) = {} & \left\{-A\left[x\log\left(1+\frac{y}{x}\right) - y\right] - \left[\log\left(1+\frac{y}{x}\right) - \frac{y}{x}\right]\right\} \\ & - \frac{y^2}{x(x+y)} - \left\{Ay\log\left(1+\frac{y}{x}\right) - \frac{y^2}{x(x+y)}\right\} + \frac{1}{2}\left[\log\left(1+\frac{y}{x}\right) - \frac{y}{x}\right] + o(1)\end{aligned} \tag{56}$$

after expanding all contributing terms up to order $1/A$. After (52), (53) and (56), we conclude that the logarithm (51) expands as

$$\begin{aligned}\log \mathbf{W}_{Ay}(Ax) = \; & -A(y\log y - (x+y)\log(x+y) + x\log x) \\ & - \log(\sqrt{2\pi Ay}) \; + \frac{1}{2}\log\left(1+\frac{y}{x}\right) + o(1). \quad (57)\end{aligned}$$

Coming back to the expression (49) of probability $\mathbf{E}_{Ay}(Ax)$, and using the value of $\mathbf{E}_{Ay}(0)$ obtained after (50) yields

$$\begin{aligned}\mathbf{E}_{Ay}(Ax) &= \varrho^{Ax} \cdot \mathbf{W}_{Ay}(Ax) \cdot \mathbf{E}_{Ay}(0) \quad (58)\\ &= (1-\varrho)\, \exp\left[A\, x \log \varrho + \log \mathbf{W}_{Ay}(Ax) + A\, y \log(1-\varrho)\right];\end{aligned}$$

inserting the expansion (57) for $\log \mathbf{W}_{Ay}(Ax)$ into equality (58) then provides the limit (48) with the expected decay rate $K(x,y)$. □

Remark 5. *Note for completeness that a sharp asymptotics for $\mathbf{E}_{Ay}(Ax)$ also readily follows from* (57)–(58)*, giving*

$$\mathbf{E}_{Ay}(Ax) \sim \frac{1-\varrho}{\sqrt{2\pi A}} \sqrt{\frac{x+y}{x\,y}}\; \exp(-A\cdot K(x,y)) \quad (59)$$

for large A. For any real $r = O(1)$, in particular, write $x = x^* + r/\sqrt{A}$ and $y = 1$; a Taylor expansion then gives $K(x,1) = \Phi(x) = \Phi''(x^*)r^2/2A + o(1/A)$ so that, after asymptotics (59),

$$\begin{aligned}\mathbf{E}_A(Ax^* + r\sqrt{A}) &\sim \frac{1-\varrho}{\sqrt{2\pi A}} \sqrt{\frac{x^*+1}{x^*\cdot 1}} \times e^{-A\cdot K(x,y)} \\ &\sim \frac{1-\varrho}{\sqrt{\varrho}}\; \frac{1}{\sqrt{2\pi A}} \cdot \exp\left[-\frac{(1-\varrho)^2}{\varrho}\; \frac{r^2}{2}\right], \qquad r\in\mathbb{R}. \quad (60)\end{aligned}$$

We thus conclude from (60) that the probability $\mathbb{P}(N'(\infty) = Ax^* + r\sqrt{A})$ is asymptotic to $\mathbb{P}(\xi = r)/\sqrt{A}$, where $\mathbb{P}(\xi = r)$ denotes the value of the density function of Gaussian variable ξ at point r. This confirms the fact that the centered variable

$$\sqrt{A}\left(\frac{N'(\infty)}{A} - x^*\right),$$

like ξ_A, also converges in distribution towards Gaussian variable ξ.

B Proof of Lemma 1

The proof of Lemma 1 proceeds in three steps.

(a) We first prove the analytic continuation of F_A to the product $\mathbb{D}\times\mathbb{C}$. As in Section 1, let $M'(t)$ denote the number of customers in the $M/M/\infty$ queue with Poisson arrival process of rate β and i.i.d. "service times" with exponential

distribution of parameter θ. Departures for the occupancy process M for M-customers stem from both service completion and departures due to impatience, while the departures for process M' come from impatience only. The random processes (N, M) and M' can be coupled in such a way that N and M' are independent, $M'(0) = M(0)$ and

$$M(t) \leqslant M'(t), \quad t \geqslant 0.$$

Letting $t \to \infty$, this readily entails that

$$\mathbb{P}(N(\infty) = n, M(\infty) = m) \leqslant \mathbb{P}(M'(\infty) \geqslant m), \qquad (n, m) \in \mathbb{N}^2,$$

where $M'(\infty)$ has a Poisson distribution with parameter $A = \beta/\theta$. The latter inequality ensures that $\mathbb{P}(N(\infty) = n, M(\infty) = m) = O(A^m/m!)$ for large m and the power series defining $\mathrm{F}_A(u, v)$ is thus convergent for all $u \in \mathbb{D}$ and $v \in \mathbb{C}$. Function F_A is therefore analytically defined in $\mathbb{D} \times \mathbb{C}$.

(b) We now consider the extension of F_A to the product $\mathbb{D}(0, 1/\varrho) \times \mathbb{D}$. Summing all equations (14) with respect to index $m \geqslant 0$, we have

$$\alpha\, \mathbf{Q}_A(n) + \mu\, n \sum_{m \geqslant 0} \frac{\mathbf{\Pi}_A(n, m)}{n + m} = \alpha\, \mathbf{Q}_A(n-1) + \mu\, (n+1) \sum_{m \geqslant 0} \frac{\mathbf{\Pi}_A(n+1, m)}{n + m + 1}$$

where we set $\mathbf{Q}_A(n) = \sum_{m \geqslant 0} \mathbf{\Pi}_A(n, m)$, hence

$$\alpha\, \mathbf{Q}_A(n) = \mu(n+1) \sum_{m \geqslant 0} \frac{\mathbf{\Pi}_A(n+1, m)}{n + m + 1}, \qquad n \in \mathbb{N}. \tag{61}$$

Let

$$\boldsymbol{\varepsilon}_A(n+1) = \frac{1}{\mathbf{Q}_A(n+1)} \left| (n+1) \sum_{m \geqslant 0} \frac{\mathbf{\Pi}_A(n+1, m)}{n + m + 1} - \mathbf{Q}_A(n+1) \right|.$$

Considering the right-hand side of (61) for large n, we calculate

$$\begin{aligned} \boldsymbol{\varepsilon}_A(n+1) &= \frac{1}{\mathbf{Q}_A(n+1)} \sum_{m \geqslant 0} \frac{m}{n + m + 1} \cdot \mathbf{\Pi}_A(n+1, m) \qquad (62) \\ &\leqslant \frac{1}{n+1} \cdot \frac{\mathbb{E}(M(\infty); N(\infty) = n+1)}{\mathbf{Q}_A(n+1)} \end{aligned}$$

hence

$$\boldsymbol{\varepsilon}_A(n+1) \leqslant \frac{1}{n+1} \cdot \mathbb{E}(M(\infty)\,|N(\infty) = n+1). \tag{63}$$

Since $M \leqslant M'$ and by the independence of random variables M' and N, we have $\mathbb{E}(M(\infty)\,|N(\infty) = n+1) \leqslant \mathbb{E}(M'(\infty)) = A$. It thus follows from upper bound (63) that $\boldsymbol{\varepsilon}_A(n+1) \to 0$ when $n \uparrow +\infty$ hence, after equality (61), $\alpha\, \mathbf{Q}_A(n) \sim \mu \mathbf{Q}_A(n+1)$, that is,

$$\frac{\mathbf{Q}_A(n+1)}{\mathbf{Q}_A(n)} \sim \varrho$$

for large n. We conclude that the power series $\mathrm{F}_A(u,v)$ converges for all $u \in \mathbb{D}(0, \frac{1}{\varrho})$ and $v \in \mathbb{D}$.

(c) Let $\overline{S}$ be the set of points $(u,v) \in \mathbb{C}^2$ where the power series $\mathrm{F}_A(u,v)$ converges absolutely and S the interior of $\overline{S}$. We let

$$S_0 = \{(u,v) \in S,\ uv \neq 0\}$$

and consider the mapping $\lambda : (u,v) \in S_0 \mapsto (\log|u|, \log|v|) \in \mathbb{R}^2$. By [3, Chap.I, Théorème 3], it is known that the set S is logarithmically convex, that is, the image $\lambda(S_0)$ is convex in $\mathbb{R}^2$.

Now, by Item **(a)** above, S contains $\mathbb{D} \times \mathbb{C}$ hence the image $\lambda(S_0)$ contains the square $(-\infty, 0) \times (-\infty, +\infty)$ in $\mathbb{R}^2$. Similarly, by Item **(b)** above, S contains $\mathbb{D}(0, 1/\varrho) \times \mathbb{D}$, hence the image $\lambda(S_0)$ contains the square $(-\infty, -\log\varrho) \times (-\infty, 0)$. By the convexity property of $\lambda(S_0)$, we then deduce that $\lambda(S_0)$ contains its convex envelope and thus also the complementary square $(0, -\log\varrho) \times (0, +\infty)$, so that we eventually have

$$\lambda(S_0) \supset (-\infty, -\log\varrho) \times (-\infty, +\infty).$$

The convergence domain S of power series $\mathrm{F}_A(u,v)$ therefore contains the product $\mathbf{\Omega} = \mathbb{D}(0, 1/\varrho) \times \mathbb{C}$. Function F_A is thus analytically defined in $\mathbf{\Omega}$, as claimed.

C Proof of Proposition 1

In the following, we further assume that $u \notin (-\infty, 0]$, $v \notin (-\infty, 0]$ and log denotes the principal determination of the logarithm over the cut plane $\mathbb{C} \setminus (-\infty, 0]$. By definition of F_A, write

$$\mathrm{F}_A(u,v) = I + J + K + L$$

where

$$I = \sum_{n \geqslant 1, m \geqslant 1} \mathbb{P}(N(\infty) = n, M(\infty) = m)\, u^n v^m,$$

$$J = \sum_{m \geqslant 1} \mathbb{P}(N(\infty) = 0, M(\infty) = m)\, v^m, \quad K = \sum_{n \geqslant 1} \mathbb{P}(N(\infty) = n, M(\infty) = 0)\, u^n$$

and $L = \mathbf{\Pi}_A(0,0)$.

(a) First consider the sum I. With the change scale $n = Ax$ and $m = Ay$, $x > 0$, $y > 0$ and by Theorem 2, we have

$$\mathbb{P}(N(\infty) = Ax, M(\infty) = Ay) u^{Ax} v^{Ay} \asymp \exp(-A\, h_{u,v}(x,y))$$

for large A ($f \asymp g$ meaning that $-\log f/A \sim -\log g/A$ when $A \uparrow +\infty$) where

$$\begin{aligned} h_{u,v}(x,y) &= \Phi(x) - x\, \log u + \Psi(y) - y\, \log v \\ &= \Phi(x) - x\, \log r + \Psi(y) - y\, \log s - i(x\zeta + y\eta) \end{aligned} \tag{64}$$

with $u = r\,e^{i\zeta}$ and $v = s\,e^{i\eta}$ as in the statement of Proposition 1. For given real $r = |u| > 0$, $s = |v| > 0$, and after the respective definitions (5) and (6) of functions Φ and Ψ, the real-valued function $h_{r,s} : (x, y) \in \mathbb{R}^{+*2} \mapsto h_{r,s}(x, y)$ is easily shown to have a unique minimum at point $(X_u, Y_v) \in \mathbb{R}^{+*2}$ given by

$$X_u = \frac{\varrho\, r}{1 - \varrho r}, \quad Y_v = s. \tag{65}$$

The corresponding value $h_{r,s}(X_u, Y_v) = \Phi(X_u) - X_u \log r + \Psi(Y_v) - Y_v \log s$ is easily calculated as $h_{r,s}(X_u, Y_v) = -\log(1 - \varrho) + \log(1 - \varrho r) - (s - 1)$ so that

$$\exp\left(-A\, h_{r,s}(X_u, Y_v)\right) = \left(\frac{1 - \varrho}{1 - \varrho r}\right)^A e^{A(s-1)} \tag{66}$$

while the second-order derivatives of $h_{r,s}$ at (X_u, Y_v) are given by

$$\frac{\partial^2 h_{r,s}}{\partial x^2}(X_u, Y_v) = \frac{(1 - \varrho r)^2}{\varrho r} := a_r, \qquad \frac{\partial^2 h_{r,s}}{\partial y^2}(X_u, Y_v) = \frac{1}{s} := b_s$$

and $\partial^2_{xy} h_{r,s}(X_u, Y_v) = 0$. Estimating the discrete sum I over the lattice $\mathbb{N}^{*2}$ by a Riemann integral over $\mathbb{R}^{+*2}$ with integration step $1/A$ and using asymptotics (7) for $\mathbf{\Pi}_A(Ax, Ay)$, we further have

$$\begin{aligned} I &\sim \int_{0^+}^{+\infty} A\,\mathrm{d}x \int_{0^+}^{+\infty} A\,\mathrm{d}y\, \mathbf{\Pi}_A(Ax, Ay) \cdot u^{Ax} v^{Ay} \\ &\sim \frac{A}{2\pi} \int_{0^+}^{+\infty} \int_{0^+}^{+\infty} g(x, y)\, e^{-A\, h_{u,v}(x,y)} \mathrm{d}x\mathrm{d}y \end{aligned} \tag{67}$$

with function $h_{u,v}$ introduced in (64). Now applying the asymptotics (22) to evaluate the (complex-valued) integral (67) with help of (66), we then obtain

$$\begin{aligned} I \sim\ & \frac{A}{2\pi} \times g(X_u, Y_v) \cdot e^{-A\, h_{r,s}(X_u,Y_v) + iA(\zeta X_u + \eta Y_v)} \times \\ & \exp\left[-\frac{A\zeta^2}{2\, a_r}\right] \exp\left[-\frac{A\eta^2}{2\, b_s}\right] \sqrt{\frac{2\pi}{A\, a_r}} \sqrt{\frac{2\pi}{A\, b_s}}. \end{aligned}$$

Using successively the expressions (66) for $e^{-A\, h_{r,s}(X_u,Y_v)}$ and the second-order derivatives a_r and b_s of $h_{r,s}$ at point (X_u, Y_v), together with the definition of g in (7) eventually reduces the latter estimate of I to

$$\begin{aligned} I \sim\ & \left(\frac{1 - \varrho}{1 - \varrho r}\right)^A e^{A(s-1)} \exp\left[iA\left(\frac{\varrho r \zeta}{1 - \varrho r} + s\eta\right)\right] \times \\ & \exp\left[-\frac{A}{2}\left(\frac{\varrho r \zeta^2}{(1 - \varrho r)^2} + s\eta^2\right)\right] G_0(u, v) \end{aligned} \tag{68}$$

with coefficient

$$G_0(u, v) = \frac{g(X_u, Y_v)}{\partial^2_{xx} h_{r,s}(X_u, Y_v)\, \partial^2_{yy} h_{r,s}(X_u, Y_v)} = \left(\frac{1 - \varrho}{1 - \varrho r}\right) \left[s + \varrho r(1 - s)\right]^{\frac{\alpha}{\theta}(1-r)-c}.$$

To specify the definition domain of function G_0, first assume $s \leqslant 1$; then the argument $s + \varrho r(1-s) \geqslant s > 0$ for all r (and thus also for $r < 1/\varrho$); now if $s > 1$, $s + \varrho r(1-s)$ is positive if and only if $\varrho r < 1 + 1/(s-1)$, which is fulfilled if $\varrho r < 1$. We thus conclude that function G_0 is well-defined and continuous over $\mathbf{\Omega}$, and thus also in the subset $\mathbf{\Omega}'$. At point $(u, v) = (1, 1)$, in particular, we clearly have $r = s = 1$ so that $G_0(1, 1) = 1$.

(b) Now address the second term J. By Theorem 2 again, we can write

$$\mathbb{P}(N(\infty) = 0, M(\infty) = Ay) \asymp e^{-A\,H(0,y)} v^{Ay} = e^{-A\,h_{1,v}(0,y)}$$

with the notation (64) for function $h_{1,v}$. For $s = |v| > 0$, the real-valued function $y \in \mathbb{R}^{+*} \mapsto h_{1,s}(0, y)$ has a unique minimum at $y = Y_s = s > 0$, with value

$$h_{1,s}(0, Y_v) = -\log(1-\varrho) - (s-1).$$

The module of the sum J is therefore of order $|J| \asymp e^{-A\,h_{1,s}(0,Y_v)} = (1-\varrho)^A e^{A(s-1)}$ and, after the estimate (68) of I, the ratio I/J is of order

$$\frac{|I|}{|J|} \asymp \frac{1}{(1-\varrho r)^A}$$

and thus tends to 0 when $A \uparrow +\infty$ and for $(u, v) \in \mathbf{\Omega}'$. We conclude that I dominates J for large A.

(c) As to the third term K, it is similarly verified that $|I|/|K| \asymp e^{As}$ with $s = |v| > 0$. As the latter ratio tends to $+\infty$ when $A \uparrow +\infty$, I also dominates K for large A. Finally,

$$\frac{|I|}{|L|} \asymp \frac{e^{As}}{(1-\varrho r)^A}$$

and I also dominates L for large A.

After (68) and the latter discussion, asymptotics (45) for $\mathrm{F}_A(u, v)$, $(u, v) \in \mathbf{\Omega}'$ eventually follows.

References

[1] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I, Asymptotic Methods and Perturbation Theory*, Springer, New York, 1999.

[2] N. Bleistein and R. A. Handelsman, *Asymptotic expansions of integrals*, Dover Publications, Inc., New York, second ed., 1986.

[3] B. Chabat, *Introduction à l'Analyse Complexe, Tome 2, Fonctions de Plusieurs Variables*, Mir, 1990.

[4] E. G. Coffman, Jr., A. A. Puhalskii, M. I. Reiman, and P. E. Wright, *Processor-shared buffers with reneging*, Performance Evaluation, 19 (1994), pp. 25–46, https://doi.org/10.1016/0166-5316(94)90053-1.

[5] M. R. Crivellari, A. Faggionato, and D. Gabrielli, *Averaging and Large Deviation Principles for Fully-Coupled Piecewise Deterministic Markov Processes and Applications to Molecular Motors*, Markov Process. Related Fields, 16 (2010), pp. 497–548.

[6] W. Eckhaus, *Asymptotic analysis of singular perturbations*, vol. 9 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam-New York, 1979.

[7] W. H. Fleming and H. M. Soner, *Asymptotic expansions for Markov processes with Lévy generators*, Appl. Math. Optim., 19 (1989), pp. 203–223, https://doi.org/10.1007/BF01448199.

[8] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, vol. 260 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer, Heidelberg, third ed., 2012, https://doi.org/10.1007/978-3-642-25847-3. Translated from the 1979 Russian original by Joseph Szücs.

[9] F. Guillemin, S. Elayoubi, P. Robert, C. Fricker, and B. Sericola, *Controlling impatience in cellular networks using qoe-aware radio resource allocation*, in Teletraffic Congress (ITC 27), 2015 27th International, Sept 2015, pp. 159–167.

[10] G. Huang, M. Mandjes, and P. Spreij, *Large deviations for Markov-modulated diffusion processes with rapid switching*, Stochastic Process. Appl., 126 (2016), pp. 1785–1818, https://doi.org/10.1016/j.spa.2015.12.005, https://doi.org/10.1016/j.spa.2015.12.005.

[11] J. Jacod and P. Protter, *Probability essentials*, Universitext, Springer-Verlag, Berlin, second ed., 2003, https://doi.org/10.1007/978-3-642-55682-1.

[12] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier, *On the performance of state-dependent single server queues*, SIAM J. Appl. Math., 46 (1986), pp. 657–697, https://doi.org/10.1137/0146045.

[13] C. Knessl and C. Tier, *Applications of singular perturbation methods in queueing*, in Advances in queueing, Probab. Stochastics Ser., CRC, Boca Raton, FL, 1995, pp. 311–336.

[14] P. Olivier, F. Simatos, and A. Simonian, *Performance analysis of data traffic in small cells networks with user mobility, in "systems modeling: Methodologies and tools", chap.12*, (2019), pp. 181–197, https://doi.org/10.1007/978-3-319-92378-9_12.

[15] P. Olivier and A. Simonian, *Performance of data traffic in small cells networks with inter-cell mobility*, in 10th EAI International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2016,

Taormina, Italy, 25th-28th Oct 2016, A. Puliafito, K. S. Trivedi, B. Tuffin, M. Scarpa, F. Machida, and J. Alonso, eds., ACM, 2016, https://doi.org/10.4108/eai.25-10-2016.2266520.

[16] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, eds., *NIST handbook of mathematical functions*, U.S. Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge University Press, Cambridge, 2010. With 1 CD-ROM (Windows, Macintosh and UNIX).

[17] A. A. Puhalskii, *On large deviations of coupled diffusions with time scale separation*, Ann. Probab., 44 (2016), pp. 3111–3186, https://doi.org/10.1214/15-AOP1043, https://doi.org/10.1214/15-AOP1043.

[18] Z. Schuss, *Theory and applications of stochastic processes*, vol. 170 of Applied Mathematical Sciences, Springer, New York, 2010, https://doi.org/10.1007/978-1-4419-1605-1. An analytical approach.

[19] F. Simatos and A. Simonian, *Mobility can drastically improve the heavy traffic performance from* $\frac{1}{1-\varrho}$ *to* $\log(\frac{1}{1-\varrho})$, Queueing Syst., 95 (2020), pp. 1–28, https://doi.org/10.1007/s11134-020-09652-0.

[20] G. G. Yin and Q. Zhang, *Continuous-time Markov chains and applications*, vol. 37 of Stochastic Modelling and Applied Probability, Springer, New York, second ed., 2013, https://doi.org/10.1007/978-1-4614-4346-9. A two-time-scale approach.