CrossMark

Dominant poles and tail asymptotics in the critical Gaussian many-sources regime

A. J. E. M. Janssen[1] · J. S. H. van Leeuwaarden[1]

Abstract The dominant pole approximation (DPA) is a classical analytic method to obtain from a generating function asymptotic estimates for its underlying coefficients. We apply DPA to a discrete queue in a critical many-sources regime, in order to obtain tail asymptotics for the stationary queue length. As it turns out, this regime leads to a clustering of the poles of the generating function, which renders the classical DPA useless, since the dominant pole is not sufficiently dominant. To resolve this, we design a new DPA method, which might also find application in other areas of mathematics, like combinatorics, particularly when Gaussian scalings related to the central limit theorem are involved.

1 Introduction

Probability generating functions (PGFs) encode the distributions of discrete random variables. When PGFs are considered analytic objects, their singularities or poles contain crucial information about the underlying distributions. Asymptotic expressions

✉ J. S. H. van Leeuwaarden
j.s.h.v.leeuwaarden@tue.nl

A. J. E. M. Janssen
a.j.e.m.janssen@tue.nl

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

for the tail distributions, related to large deviations events, can typically be obtained in terms of the so-called dominant singularities, or dominant poles. The dominant pole approximation (DPA) for the tail distribution is then derived from the partial fraction expansion of the PGF and maintaining from this expansion the dominant fraction related to the dominant pole. Dominant pole approximations have been applied in many branches of mathematics, including analytic combinatorics [7] and queueing theory [19]. We apply DPA to a discrete queue that has an explicit expression for the PGF of the stationary queue length. Additionally, this queue is considered in a many-sources regime, a heavy-traffic regime in which both the demand on and the capacity of the systems grow large, while their ratio approaches one. This many-sources regime combines high system utilization and short delays, due to economies of scale. The regime is similar in flavor to the QED (quality and efficiency driven) regime for many-server systems [8], although an important difference is that our discrete queue fed by many-sources falls in the class of single-server systems and therefore leads to a manageable closed form expression for the PGF of the stationary queue length Q. Denote this PGF by $Q(z) = \mathbb{E}(z^Q)$.

PGFs can be represented as power series around $z = 0$ with non-negative coefficients (related to the probabilities). We assume that the radius of convergence of $Q(z)$ is larger than one (in which case all moments of Q exist). This radius of convergence is in fact determined by the dominant singularity Z_0, the singularity in $|z| > 1$ closest to the origin. For PGFs, due to Pringsheim's theorem [7], Z_0 is always a positive real number larger than one. Then DPA leads to the approximation

$$\mathbb{P}(Q > N) \approx \frac{c_0}{1 - Z_0} \left(\frac{1}{Z_0}\right)^{N+1} \quad \text{for large } N, \tag{1.1}$$

with $c_0 = \lim_{z \to Z_0} (z - Z_0) Q(z)$. In many cases, the approximation (1.1) can be turned into a more rigorous asymptotic expansion (for N large) for the tail probabilities $\mathbb{P}(Q > N)$. We shall now explain in more detail the many-sources regime, the discrete queue, and when combining both, the mathematical challenges that arise when applying DPA.

1.1 Many sources and a discrete queue

Consider a stochastic system in which demand per period is given by some random variable A, with mean μ_A and variance σ_A^2. For systems facing large demand, one can set the capacity according to the rule $s = \mu_A + \beta \sigma_A$, which consists of a minimally required part μ_A and a variability hedge $\beta \sigma_A$. Such a rule can lead to economies of scale, as we will now describe in terms of a setting in which the demand per period is generated by many sources. Consider a system serving n independent sources and let X denote the generic random variable that describes the demand per source per period, with mean μ and variance σ^2. Denote the service capacity by s_n, so that the system utilization is given by $\rho_n = n\mu/s_n$, where the index n expresses the dependence on the scale at which the system operates. The traditional capacity sizing rule would then be

$$s_n = n\mu + \beta \sigma \sqrt{n} \tag{1.2}$$

with β some positive constant. The standard heavy-traffic paradigm [11,15,16], which builds on the central limit theorem, then prescribes considering a sequence of systems indexed by n with associated loads ρ_n such that (also using that $s = s_n \sim n\mu$)

$$\rho_n = \frac{n\mu}{s_n} \sim 1 - \frac{\beta\sigma}{\mu\sqrt{n}} = 1 - \frac{\gamma}{\sqrt{s_n}} + O(n^{-1}), \quad \text{as } n \to \infty, \qquad (1.3)$$

where $\gamma = \beta\sigma/\sqrt{\mu}$. We shall apply the many-sources regime given by (1.2) and (1.3) to a discrete queue, in which we divide time into periods of equal length and model the net input in consecutive periods as i.i.d. samples from the distribution of A, with mean $n\mu$ and variance $n\sigma^2$. The capacity per period s_n is fixed and integer valued. The scaling rule in (1.3) thus specifies how the mean and variance of the demand per period, and simultaneously s_n, will all grow to infinity as functions of n. Many-sources scaling became popular through the Anick–Mitra–Sondhi model [1], as one of the canonical models for modern telecommunications networks, in which a switch may have hundreds of different input flows. But apart from communication networks, the concept of many sources can apply to any service system in which demand can be regarded as coming from many different inputs (see, for example, [4,6,13,17,18] for specific applications).

1.2 How to adapt classical DPA?

As it turns out, the many-sources regime changes drastically the nature of the DPA. When the queue is pushed into the many-sources regime for $n \to \infty$, the dominant pole becomes barely dominant, in the sense that all the other poles (the dominated ones) of the PGF are approaching the dominant pole. For the partial fraction expansion of the PGF, this means that it becomes hard, or impossible even, to simply discard the contributions of the fractions corresponding to what we call *dominated* poles: all poles other than the dominant pole. Moreover, the dominant pole itself approaches 1 according to

$$Z_0 \sim 1 + \frac{2\beta}{\sqrt{n\sigma^2}}, \quad \text{as } n \to \infty. \qquad (1.4)$$

This implies that in (1.1) the factor $c_0/(1 - Z_0)$ potentially explodes, while without imposing further conditions on N, the factor Z_0^{-N-1} goes to the degenerate value 1. The many-sources regime thus has a fascinating effect on the location of the poles that renders a standard DPA useless for multiple reasons. We shall therefore adapt the DPA in order to make it suitable to deal with the complications that arise in the many-sources regime, with the goal to again obtain an asymptotic expansion for the tail distribution. First observe that the term Z_0^{-N-1} in (1.1) becomes non-degenerate when we impose that $N \sim K\sqrt{n\sigma^2}$, with K some positive constant, in which case

$$\left(\frac{1}{Z_0}\right)^{N+1} \sim \left(1 + \frac{2\beta}{\sqrt{n\sigma^2}}\right)^{-K\sqrt{n\sigma^2}} \to e^{-2\beta K} \in (0,1) \quad \text{as } n \to \infty. \qquad (1.5)$$

The condition $N \sim K\sqrt{n\sigma^2}$ is natural, because the fluctuations of our stochastic system are of the order $\sqrt{n\sigma^2}$. Of course, there are many ways in which N and n can be coupled, but due to (1.4), only couplings for which N is proportional to $\sqrt{n}$ lead to a non-degenerate limit for Z_0^{-N-1}. Now let us turn to the other two remaining issues: the fact that $c_0/(1 - Z_0)$ potentially explodes and that the dominated poles converge to the dominant pole.

The approach in this paper to deal with these two issues uses the integral representation of the tail probabilities involving the PGF $Q(z)$ of Q along a circle $|z| = 1 + \varepsilon < Z_0$. By Cauchy's theorem, the tail probabilities can be written as a residue $c_0/(1 - Z_0)Z_0^{N+1}$ at $z = Z_0$ and an integral along a contour shifted beyond Z_0. Using the product expansion of $Q(z)$, involving the zeros of a characteristic equation in $|z| \leq 1$, the quantity $c_0/(1 - Z_0)$ can be approximated in terms of an integral of the Pollaczek type that has been considered in [11]. In [11], a dedicated saddle point method has been developed to approximate Pollaczek-type integrals for the mass at zero, the mean, and the variance of Q in the many-sources regime. This dedicated saddle point method can also be used to find a many-sources approximation for $c_0/(1 - Z_0)$. The remaining challenge is then to bound the contribution of the contour integral shifted beyond the dominant pole Z_0. This depends on how far the integration path can be shifted, and this is determined by the positions of the first dominated poles. It turns out that, in the many-sources regime, the dominated poles have approximations of the type (1.4) as well. The integration path is then chosen, roughly, as the circle that passes halfway between Z_0 and the first dominated pole. This, together with the dedicated saddle point method to approximate $c_0/(1 - Z_0)$, then provides a fully rigorous derivation of the asymptotic expression for $\mathbb{P}(Q > N)$ that is of the form

$$\mathbb{P}(Q > K\sqrt{n\sigma^2}) \sim h(\beta) \cdot e^{-2\beta K}, \quad \text{as } n \to \infty. \tag{1.6}$$

The function $h(\beta)$ in this asymptotic expression involves infinite series and Riemann zeta functions that are reminiscent of the reflected Gaussian random walk [5,9,10]. Indeed, it follows from [16, Theorem 3] that our rescaled discrete queue converges under (1.3) to a reflected Gaussian random walk. Hence, the tail distribution of our system in the regime (1.3) should for large n be well approximated by the tail distribution of the reflected Gaussian random walk. We return to this connection in Sect. 5.

Our approach thus relies on detailed knowledge about the distribution of all the poles of the PGF of Q, and in particular how this distribution scales with the asymptotic regime (1.2)–(1.3). As it turns out, in contrast with classical DPA, this many-sources regime means that all poles contribute to the asymptotic characterization of the tail behavior. Our saddle point method leads to an asymptotic expansion for the tail probabilities, of which the limiting form corresponds to the heavy-traffic limit, and pre-limit forms present refined approximations for pre-limit systems ($n < \infty$) in heavy traffic. Such refinements to heavy-traffic limits are commonly referred to as *corrected diffusion approximations* [2,3,14]. Compared with the studies that directly analyzed the Gaussian random walk [5,9,10], which is the scaling limit of our queue in the many-sources regime, we start from the pre-limit process description and establish an asymptotic result which is valuable for a queue with a finite yet large number of

sources. Starting this asymptotic analysis from the actual pre-limit process description is mathematically more challenging than directly analyzing the process limit, but in return gives valuable insights into the manner and speed at which the system starts displaying its limiting behavior.

1.3 Outline of the paper

In Sect. 2 we describe the discrete queue in more detail and present some preliminary results for its stationary queue length distribution. In Sect. 3 we give an overview of the results and the contour integration representation for the tail distribution. In Sect. 4 we give a rigorous proof of the main result for the leading-order term using the dedicated saddle point method (Subsect. 4.1), and we bound the contour integral with integration path shifted beyond the dominant pole (Subsect. 4.2). In Sect. 5 we elaborate on the connection between the discrete queue and the Gaussian random walk, and we present an asymptotic series for $\mathbb{P}(Q > N)$ comprising not only the dominant poles but also the dominated poles.

2 Model description and preliminaries

We consider a discrete stochastic model in which time is divided into periods of equal length. At the beginning of each period $k = 1, 2, 3, \ldots$ new demand A_k arrives to the system. The demands per period $A_1, A_2, \ldots$ are assumed independent and equal in distribution to some non-negative integer-valued random variable A. The system has a service capacity $s \in \mathbb{N}$ per period, so that the recursion

$$Q_{k+1} = \max\{Q_k + A_k - s, 0\}, \qquad k = 1, 2, \ldots, \qquad (2.1)$$

assuming $Q_1 = 0$, gives rise to a Markov chain $(Q_k)_{k \geq 1}$ that describes the congestion in the system over time. The PGF

$$A(z) = \sum_{j=0}^{\infty} \mathbb{P}(A = j) z^j \qquad (2.2)$$

is assumed analytic in a disk $|z| < r$ with $r > 1$, which implies that all moments of A exist.

We assume that A_k is in distribution equal to the sum of work generated by n sources, $X_{1,k} + \ldots + X_{n,k}$, where the $X_{i,k}$ are, for all i and k, i.i.d. copies of a random variable X, of which the PGF $X(z) = \sum_{j=0}^{\infty} \mathbb{P}(X = j) z^j$ has radius of convergence $r > 1$, and

$$0 < \mu_A = n\mu = n X'(1) < s. \qquad (2.3)$$

Under the assumption (2.3), the stationary distribution $\lim_{k \to \infty} \mathbb{P}(Q_k = j) = \mathbb{P}(Q = j)$, $j = 0, 1, \ldots$ exists, with the random variable Q defined as having this stationary distribution. We let

$$Q(z) = \sum_{j=0}^{\infty} \mathbb{P}(Q = j) z^j \tag{2.4}$$

be the PGF of the stationary distribution.

It is a well-known consequence of Rouché's theorem that under (2.3) $z^s - A(z)$ has precisely s zeros in $|z| \leq 1$, one of them being $z_0 = 1$. We proceed in this paper under the same assumptions as in [11]. Thus, we assume that $|X(z)| < X(r_1)$, $|z| = r_1$, $z \neq r_1$, for any $r_1 \in (0, r)$; see [11], end of Sect. 2 for a discussion of this condition. Finally, we assume that the degree of $X(z)$ is larger than s/n. Under these conditions, z_0 is the only zero of $z^s - A(z)$ on $|z| = 1$, and all others in $|z| \leq 1$, denoted by $z_1, z_2, ..., z_{s-1}$, lie in $|z| < 1$. Furthermore, there are at most countably many zeros Z_k of $z^s - A(z)$ in $1 < |z| < r$, and there is precisely one, denoted by Z_0, with minimum modulus. The zeros Z_k are indexed using integer k; they come in conjugate pairs and we let $Z_k = Z_{-k}^*$ where Z_k is in the (closed) lower half plane for non-negative k. For our analysis, it is crucial that $r > 1$, and so certain heavy-tailed distributions for X (with $r = 1$) are excluded.

There is the product form representation [4,17]

$$Q(z) = \frac{(s - \mu_A)(z - 1)}{z^s - A(z)} \prod_{j=1}^{s-1} \frac{z - z_j}{1 - z_j} \tag{2.5}$$

that gives the analytic extension of $Q(z)$ to all z, $|z| < r$ and z not a zero of $z^s - A(z)$, where the right-hand side of (2.5) is analytic in $|z| < Z_0$ and has a first-order pole at $z = Z_0$. We have, for the tail probability (using that $Q(1) = 1$), for $N = 0, 1, ...$

$$P(Q > N) = \sum_{i=N+1}^{\infty} P(Q = i) = C_{z^N}\left[\frac{1 - Q(z)}{1 - z}\right], \tag{2.6}$$

where $C_{z^N}[f(z)]$ denotes the coefficient of z^N of the function $f(z)$. By contour integration, Cauchy's theorem and $Q(1) = 1$, we then get for $0 < \varepsilon < Z_0 - 1$

$$\begin{aligned} P(Q > N) &= \frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \frac{1}{z^{N+1}} \frac{1 - Q(z)}{1 - z} dz \\ &= \frac{1}{2\pi i} \int_{|z|=R} \frac{1}{z^{N+1}} \frac{1 - Q(z)}{1 - z} dz + \frac{c_0}{Z_0^{N+1}(1 - Z_0)}, \end{aligned} \tag{2.7}$$

where $c_0 = \text{Res}_{z=Z_0}[Q(z)]$ and R is any number between Z_0 and $\min_{k \neq 0} |Z_k|$. When n and s are fixed, we have that the integral on the second line of (2.7) is $O(R^{-N})$, and so there is the DPA

$$P(Q > N) = \frac{c_0}{Z_0^{N+1}(1 - Z_0)}(1 + \text{exponentially small}), \quad N \to \infty. \tag{2.8}$$

In this paper we consider the heavy-traffic setting described by (1.2) and (1.3) with $n \to \infty$. In this setting R and Z_0 in (2.7) tend to 1, and thus both terms on the second line of (2.7) need special attention. The second term here is approximated, using the product form expansion in (2.5), in terms of a contour integral of the Pollaczek type to which the dedicated saddle point method of [11], Sect. 3, can be applied. The method has been developed in [11] from Pollaczek's integral representation of the PGF of Q,

$$Q(w) = \exp\left(\frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \ln\left(\frac{w-z}{1-z}\right) \frac{(z^s - A(z))'}{z^s - A(z)} dz\right), \qquad (2.9)$$

which holds when $|w| < 1 + \varepsilon < Z_0$, to get limiting expressions for moments of Q in the generalized heavy-traffic regimes $n/s = 1 - \gamma n^{-\alpha}$, $\alpha > 0$, and $n \to \infty$. To this end, the contour integrals are deformed so as to pass through the saddle point $z_{sp} \in (1, Z_0)$ given as the zero z of $g'(z)$, where $g(z) = -\ln z + \frac{1}{s} \ln A(z) = -\ln z + \frac{n}{s} \ln X(z)$. A crucial ingredient here is the substitution $z = z(v)$, $-\frac{1}{2}\delta \le v \le \frac{1}{2}\delta$ for some $\delta > 0$, such that

$$\exp(sg(z(v))) = B \exp(-\tfrac{1}{2} s\eta v^2), \quad -\tfrac{1}{2}\delta \le v \le \tfrac{1}{2}\delta, \qquad (2.10)$$

where $B = \exp(sg(z_{sp}))$ and $\eta = g''(z_{sp})$. It is shown in [11], p. 796, using Lagrange's inversion theorem, that in the heavy-traffic regime $n/s = 1 - \gamma/\sqrt{n}$, for a $\delta > 0$ independent of s one has

$$z(v) = z_{sp} + iv + \sum_{j=2}^{\infty} c_j(iv)^j, \quad |v| < \delta, \qquad (2.11)$$

with real coefficients c_j, while (2.10) holds with B positive and bounded away from 0 and 1, and η positive and bounded away from 0.

3 Overview and results

In order to force the discrete queue to operate in the critical many-sources regime, we shall assume throughout the paper the following relation between the number of sources n and the capacity $s = s_n(\gamma)$:

$$\frac{n\mu}{s} = 1 - \frac{\gamma}{\sqrt{s}}, \qquad (3.1)$$

with $\gamma > 0$ bounded away from 0 and ∞ as $s \to \infty$. In this scaling regime, the zeros z_j and Z_k of $z^s - A(z) = 0$ start clustering near $z = 1$, as described in the next lemma (proved in the appendix). Let z_j^* and Z_k^* denote the complex conjugates of z_j and Z_k, respectively.

Lemma 3.1 *For finite $j, k = 1, 2, \ldots$ and $s \to \infty$,*

$$z_0 = 1, \qquad Z_0 = 1 + \frac{2a_0 b_0}{\sqrt{s}} + O(s^{-1}), \tag{3.2}$$

$$z_j = z_{s-j}^* = 1 + \frac{a_0}{\sqrt{s}}(b_0 - \sqrt{b_0^2 - 2\pi i j}) + O(s^{-1}), \tag{3.3}$$

$$Z_k = Z_{-k}^* = 1 + \frac{a_0}{\sqrt{s}}(b_0 + \sqrt{b_0^2 - 2\pi i k}) + O(s^{-1}), \tag{3.4}$$

with

$$a_0 = \frac{\sqrt{2\mu}}{\sigma}, \qquad b_0 = \frac{\gamma\sqrt{\mu}}{\sigma\sqrt{2}} \tag{3.5}$$

and principal roots in (3.3)–(3.5).

Due to this clustering phenomenon, the main reasoning that underpins classical DPA cannot be carried over. Starting from the expression (2.8), we need to investigate what becomes of the term $c_0/(1 - Z_0)$, and moreover, the validity of the exponentially small phrase in (2.8) and the actual N-range both become delicate matters that need detailed information about the distribution of the zeros as in Lemma 3.1.

Let us first present a result that identifies the relevant N-range:

Proposition 3.2

$$\frac{1}{Z_0^{N+1}} = \exp\left(\frac{-2L\gamma\mu}{\sigma^2}\right)(1 + O(s^{-1/2})) \tag{3.6}$$

when $N + 1 = L\sqrt{s}$ with $L > 0$ bounded away from 0 and ∞.

Proof We have from (3.2) and (3.5) that $Z_0 = 1 + \frac{2\gamma\mu}{\sigma^2\sqrt{s}} + O\left(\frac{1}{s}\right)$. Hence

$$\frac{1}{Z_0^{N+1}} = \exp\left(-L\sqrt{s}\ln\left(1 + \frac{2\gamma\mu}{\sigma^2\sqrt{s}} + O(s^{-1})\right)\right) = \exp\left(\frac{-L\gamma\mu}{\sigma^2} + O(s^{-1/2})\right) \tag{3.7}$$

when L is bounded away from 0 and ∞, and this gives the result. $\qquad\square$

From (2.5), we obtain the representation

$$\frac{c_0}{1 - Z_0} = -\frac{s - \mu_A}{sZ_0^{s-1} - A'(Z_0)} \prod_{j=1}^{s-1} \frac{Z_0 - z_j}{1 - z_j}. \tag{3.8}$$

The next result will be proved in Sect. 4.

Lemma 3.3

$$-\frac{s - \mu_A}{sZ_0^{s-1} - A'(Z_0)} = \left(\frac{1}{Z_0}\right)^{s-1}\left(1 + O(s^{-1/2})\right). \tag{3.9}$$

We thus get from (3.8) and Lemma 3.3

$$\frac{c_0}{1 - Z_0} = \frac{P(Z_0)}{P(1)}\left(1 + O(s^{-1/2})\right), \tag{3.10}$$

where

$$P(Z) = \prod_{j=1}^{s-1}(1 - z_j/Z) = \exp\left(\sum_{j=1}^{s-1}\ln(1 - z_j/Z)\right) \tag{3.11}$$

for $Z \in \mathbb{C}$, $|Z| \geq 1$ (principal logarithm). To handle the product $P(z)$, in Lemma 3.4 below we evaluate $\ln P(Z)$ for $|Z| \geq 1$ in terms of the contour integral

$$I(Z) = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon}\frac{\ln(1 - z^{-s}A(z))}{Z - z}dz, \tag{3.12}$$

where $\varepsilon > 0$ is such that $1 < 1 + \varepsilon < Z_0$.

Lemma 3.4 *Let* $\varepsilon > 0$, $1 < 1 + \varepsilon < Z_0$ *and* $|Z| \geq 1$. *Then*

$$\ln P(Z) = \begin{cases} -\ln(1 - Z^{-1}) + I(Z), & 1 + \varepsilon < |Z| < r, \\ -\ln(1 - Z^{-1}) + \ln(1 - Z^{-s}A(Z)) + I(Z), & 1 < |Z| < 1 + \varepsilon, \\ \ln(\gamma\sqrt{s}) + I(1), & Z = 1. \end{cases} \tag{3.13}$$

The dedicated saddle point method, as considered in [11], applied to $I(Z)$, with saddle point $z_{\text{sp}} = 1 + \varepsilon$ of the function $g(z) = -\ln z + \frac{n}{s}\ln(X(Z))$, yields

$$I(1) = -\ln[Q(0)] + O(s^{-1/2}), \quad I(Z_0) = \ln[Q(0)] + O(s^{-1/2}). \tag{3.14}$$

Combining (3.2), (3.5), (3.8), (3.9), and (3.14) then gives one of our main results:

Proposition 3.5

$$\ln\left(\frac{c_0}{1 - Z_0}\right) = -\ln(4b_0^2) + 2\ln[Q(0)] + O(s^{-1/2}). \tag{3.15}$$

The next step consists of bounding the integral on the second line of (2.7) that can be written as

$$\frac{-1}{2\pi i}\int_{|z|=R}\frac{Q(z)}{z^{N+1}(1 - z)}dz, \tag{3.16}$$

by choosing R appropriately. To do this, we consider the product representation (2.5) of $Q(z)$, and we want to choose R such that $|z^s - A(z)| \geq C|z|^s$, $|z| = R$, for some $C > 0$ independent of s. It will be shown in Sect. 4 that this is achieved by taking R such that the curve $|z^s| = |A(z)|$, on which Z_0 and $Z_{\pm 1}(= Z_{+1}, Z_{-1})$ lie, is crossed near a point z (also referred to as $Z_{\pm 1/2}$) where z^s and $A(z)$ have opposite sign. A

further analysis, using again the dedicated saddle point method to bound the product $\prod_{j=1}^{s-1}$ in (2.5), then yields that the integral in (3.16) decays as R^{-N}. Finally, using the asymptotic information in (3.2)–(3.4) for Z_0 and $Z_{\pm 1}$, with $Z_{\pm 1/2}$ lying midway between Z_0 and $Z_{\pm 1}$, the integral on the second line of (2.7) can be shown to have relative order $\exp(-DN/\sqrt{s})$, for some $D > 0$ independent of s, compared to the dominant pole term in (2.8).

To summarize, we have now that

$$\mathbb{P}(Q > N) = \frac{c_0}{1 - Z_0}\Big(\frac{1}{Z_0}\Big)^{N+1}\Big(1 + O(e^{-DN/\sqrt{s}})\Big), \tag{3.17}$$

for some $D > 0$ independent of s, $N = 1, 2, \ldots$. The DPA $c_0/(1 - Z_0)^{-1}Z_0^{-N-1}$ of $\mathbb{P}(Q > N)$ thus has a relative error that decays exponentially fast.

In Subsect. 5.1, the stationary queue length Q, considered in the many-sources regime, is shown to be connected to the Gaussian random walk. This connection will imply that the front factor of the DPA in (3.17) satisfies

$$\frac{c_0}{1 - Z_0} = H(b_0)\Big(1 + O(s^{-1/2})\Big), \tag{3.18}$$

where $\ln H(b_0)$ has a power series in b_0 with coefficients that can be expressed in terms of the Riemann zeta function. Combining this with Propositions 3.2 and 3.17 yields

$$\mathbb{P}(Q > N) = H(b_0)\exp\Big(-\frac{2L\gamma\mu}{\sigma^2}\Big)\Big(1 + O(s^{-1/2})\Big) \tag{3.19}$$

when $N + 1 = L\sqrt{s}$ with L bounded away from 0 and ∞. The leading term in (3.19) agrees with (1.6) when we identify

$$L = \sigma K/\sqrt{\mu}, \quad \gamma = \beta\sigma/\sqrt{\mu}, \quad s = n\mu + \beta\sigma\sqrt{n} \approx n\mu, \quad b_0 = \beta/\sqrt{2} \tag{3.20}$$

and $H(b_0) = h(\beta)$.

In Subsect. 5.2, we extend for a fixed $M = 1, 2, \ldots$ the approach in Sect. 4 by increasing the radius R of the integration contour in (3.16) to R_M such that the poles $Z_0, Z_{\pm 1}, \ldots, Z_{\pm M}$ are inside $|z| = R_M$. This leads to

$$P(Q > N) = \text{Re}\left[\frac{c_0}{(1 - Z_0)Z_0^{N+1}} + 2\sum_{k=1}^{M}\frac{c_k}{(1 - Z_k)Z_k^{N+1}}\right] + O\Big(|Z_{M+1}|^{-N}\Big). \tag{3.21}$$

The front factors $c_k/(1 - Z_k)$ in the series in (3.21) satisfy

$$\frac{c_k}{1 - Z_k} = H_k(b_0)\Big(1 + O(s^{-1/2})\Big), \tag{3.22}$$

with $H_k(b_0)$ some explicitly defined integral. When $N + 1 = L\sqrt{s}$ with L bounded away from 0 and ∞, we find from (3.22) and Proposition 3.2 that

$$\frac{c_k}{1 - Z_k} = H(b_k) \exp\left(- La_0(\sqrt{b_0^2 - 2\pi i k} + b_0)\right)\left(1 + O(s^{-1/2})\right); \quad (3.23)$$

compare with (3.18), and it can be shown that this gives rise to an $\exp(-DL/\sqrt{k})$ decay of the right-hand side of (3.23). The results in (3.19), (3.21) and (3.23) together give precise information as to how the DPA arises, with leading behavior from the dominant pole, and lower order refinements coming from the dominated poles.

4 DPA through contour integration

In this section we present the details of getting approximations of the tail probabilities using a contour integration approach as outlined in Sect. 3. In Subsect. 4.1 we concentrate on approximation of the front factor $c_0/(1 - Z_0)$ and the dominant pole Z_0, and combine these to obtain an approximation of the leading-order term in (2.8). This gives Lemmas 3.3 and 3.4, and Proposition 3.5.

In Subsect. 4.2 we assess and bound the integral on the second line of (2.7) and thereby make precise what exponentially small in (2.8) means in the present setting.

4.1 Approximation of the leading-order term

4.1.1 Proof of Lemma 3.3

From

$$Z_0^s = A(Z_0) = X^n(Z_0), \quad \mu_A = n\mu = s\left(1 - \frac{\gamma}{\sqrt{s}}\right), \quad A'(Z_0) = nX'(Z_0)X^{n-1}(Z_0), \quad (4.1)$$

we compute

$$\frac{s - \mu_A}{sZ_0^{s-1} - A'(Z_0)} = \frac{\gamma/\sqrt{s}}{1 - (1 - \gamma/\sqrt{s})\dfrac{X'(Z_0)Z_0}{X'(1)X(Z_0)}} \frac{1}{Z_0^{s-1}}. \quad (4.2)$$

With the approximation (3.2), written as

$$Z_0 = 1 + \frac{d_0}{\sqrt{s}} + O(s^{-1}), \quad d_0 = \frac{2\gamma\mu}{\sigma^2}, \quad (4.3)$$

we get

$$X'(Z_0) = X'(1) + X''(1)(Z_0 - 1) + O(s^{-1}) = \mu + \frac{X''(1)d_0}{\sqrt{s}} + O(s^{-1}) \quad (4.4)$$

and

$$X(Z_0) = X(1) + X'(1)(Z_0 - 1) + O(s^{-1}) = 1 + \frac{\mu d_0}{\sqrt{s}} + O(s^{-1}). \quad (4.5)$$

Hence, by (4.3–4.5) and (5.32),

$$
\begin{aligned}
1 - \left(1 - \frac{\gamma}{\sqrt{s}}\right) &\frac{X'(Z_0)Z_0}{X'(1)X(Z_0)} \\
&= 1 - \left(1 - \frac{\gamma}{\sqrt{s}}\right) \frac{\left(\mu + \dfrac{X''(1)d_0}{\sqrt{s}} + O(s^{-1})\right)\left(1 + \dfrac{d_0}{\sqrt{s}} + O(s^{-1})\right)}{\mu\left(1 + \dfrac{\mu d_0}{\sqrt{s}} + O(s^{-1})\right)} \\
&= 1 - \left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(1 + \frac{d_0}{\sqrt{s}}\left(\frac{X''(1)}{\mu} + 1 - \mu\right) + O(s^{-1})\right) \\
&= 1 - \left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(1 + \frac{d_0}{\sqrt{s}}\frac{\sigma^2}{\mu} + O(s^{-1})\right) = -\frac{\gamma}{\sqrt{s}} + O(s^{-1}), \quad (4.6)
\end{aligned}
$$

where we have used d_0 of (4.3) in the last step. This gives (3.9). □

4.1.2 Proof of Lemma 3.4

We have $|A(z)| < |z^s|$ when $z \neq 1$, $1 < |z| < Z_0$, and so $\ln(1 - A(z)/z^s)$ is analytic in $z \neq 1$, $1 < |z| < Z_0$. When $|Z| > 1 + \varepsilon$, we have by partial integration, using that $d[\ln(1 - z/Z)] = -dz/(Z - z)$ and Cauchy's theorem,

$$
\begin{aligned}
I(Z) &= \frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \ln\left(1 - \frac{z}{Z}\right)\frac{(1 - A(z)/z^s)'}{1 - A(z)/z^s}dz \\
&= \sum_{j=0}^{s-1}\ln\left(1 - \frac{z_j}{Z}\right) = \ln\left(1 - \frac{1}{Z}\right) + \ln P(Z). \quad (4.7)
\end{aligned}
$$

This gives the upper-case formula in (3.13), and the middle case follows in a similar manner by taking the residue at $z = Z$ inside $|z| = 1 + \varepsilon$ into account. For the lower case in (3.13), we use the result of the middle case, in which we take $1 < Z < 1 + \varepsilon$, $Z \downarrow 1$. We have $I(Z) \to I(1)$ as $Z \downarrow 1$, and

$$
\begin{aligned}
\lim_{Z \downarrow 1}&\left[-\ln\left(1 - \frac{1}{Z}\right) + \ln\left(1 - \frac{X^n(Z)}{Z^s}\right)\right] \\
&= \ln[(Z^s - X^n(Z))'|_{Z=1}] = \ln(s - n\mu) = \ln(\gamma\sqrt{s}), \quad (4.8)
\end{aligned}
$$

and this completes the proof. □

4.1.3 Proof of Proposition 3.5

By (3.10) and Lemma 3.4, we have

$$\ln\left(\frac{c_0}{1 - Z_0}\right) = \ln P(Z_0) - \ln P(1) + O(s^{-1/2})$$

$$= -\ln\left(1 - \frac{1}{Z_0}\right) - \ln(\gamma\sqrt{s}) + I(Z_0) - I(1) + O(s^{-1/2}). \quad (4.9)$$

From (3.2), it follows that

$$-\ln\left(1 - \frac{1}{Z_0}\right) - \ln(\gamma\sqrt{s}) = -\ln\left(\frac{2\gamma^2\mu}{\sigma^2}\right) + O(s^{-1/2}) = -\ln(4b_0^2) + O(s^{-1/2}), \quad (4.10)$$

and so

$$\ln\left(\frac{c_0}{1 - Z_0}\right) = I(Z_0) - I(1) - \ln(4b_0^2) + O(s^{-1/2}). \quad (4.11)$$

Next, we consider the integral representation (3.12) of $I(Z)$, where we take ε such that

$$1 + \varepsilon = z_{\text{sp}} = 1 + \frac{\gamma\mu}{\sigma^2\sqrt{s}} + O(s^{-1}), \quad (4.12)$$

with z_{sp}, see [11, Sect. 3], the unique point $z \in (1, Z_0)$ such that $g'(z) = 0$, where

$$g(z) = -\ln z + \frac{n}{s}\ln X(z). \quad (4.13)$$

We have $g(1) = 0 = g(Z_0)$, and so $z = z_{\text{sp}}$ for which $g'(z) = 0$ lies about midway between 1 and Z_0. More specifically, by expanding $g'(z) = g'(1) + (z - 1)g''(1) + O(s^{-1})$, $1 \le z \le Z_0$, and expressing $g'(1)$, $g''(1)$ in terms of γ, μ, σ, s, while using (4.3), it follows that

$$z_{\text{sp}} = \tfrac{1}{2}(1 + Z_0) + O(s^{-1}), \qquad Z_0 - z_{\text{sp}} = z_{\text{sp}} - 1 + O(s^{-1}). \quad (4.14)$$

This suggests that $I(Z_0) \approx -I(1)$, a statement made precise below, since the main contribution to $I(Z)$ comes from the z's in (3.12) close to z_{sp}.

We have, see [11],

$$\ln[P(Q = 0)] = \frac{1}{2\pi i}\int\limits_{|z| = z_{\text{sp}}} \frac{\ln\left(1 - z^{-s}A(z)\right)}{z(z - 1)}dz. \quad (4.15)$$

Now

$$\frac{1}{z - 1} = \frac{1}{z(z - 1)} + \frac{1}{z}, \quad (4.16)$$

and

$$\int\limits_{|z| = z_{\text{sp}}} \left|\ln\left(1 - z^{-s}A(z)\right)\right||dz| = O(s^{-1/2}); \quad (4.17)$$

see [11, Subsect. 5.3]. Hence,

$$I(1) = \frac{-1}{2\pi i} \int\limits_{|z|=z_{\mathrm{sp}}} \frac{\ln\left(1 - z^{-s} A(z)\right)}{z - 1} dz = \ln[P(Q = 0)] + O(s^{-1/2}). \quad (4.18)$$

As for $I(Z_0)$, we observe that, see (4.14),

$$\frac{1}{Z_0 - z} = \frac{1}{z - 1} + 2 \frac{z - \frac{1}{2}(1 + Z_0)}{(Z_0 - z)(z - 1)} = \frac{1}{z - 1} + 2 \frac{z - z_{\mathrm{sp}}}{(Z_0 - z)(z - 1)} + O(1). \quad (4.19)$$

Thus,

$$I(Z_0) = -I(1) + 2 \int\limits_{|z|=z_{\mathrm{sp}}} \frac{z - z_{\mathrm{sp}}}{(Z_0 - z)(z - 1)} \ln\left(1 - z^{-s} A(z)\right) dz + O(s^{-1/2}). \quad (4.20)$$

We next estimate the remaining integral in (4.20). With the substitution $z = z(v)$, $-\frac{1}{2}\delta \leq v \leq \frac{1}{2}\delta$, as explained at the end of Sect. 3, we have $A(z(v))/z^s(v) = B \exp(-\frac{1}{2} s\eta v^2)$ with $0 < B < 1$ and $\eta > 0$ bounded away from 1 and 0, respectively, and

$$z(v) = z_{\mathrm{sp}} + iv + \sum_{j=2}^{\infty} c_n (iv)^j, \qquad -\tfrac{1}{2}\delta \leq v \leq \tfrac{1}{2}\delta, \quad (4.21)$$

where c_j are real. Then we get with exponentially small error

$$\int\limits_{|z|=z_{\mathrm{sp}}} \frac{z - z_{\mathrm{sp}}}{(Z_0 - z)(z - 1)} \ln\left(1 - z^{-s} A(z)\right) dz$$

$$= \int\limits_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{(z(v) - z_{\mathrm{sp}})z'(v)}{(Z_0 - z(v))(z(v) - 1)} \ln(1 - B e^{-\frac{1}{2} s\eta v^2}) dv. \quad (4.22)$$

Now we get from (4.14) and (4.21) that

$$\begin{aligned}
(Z_0 - z(v))(z(v) - 1) &= (z_{\mathrm{sp}} - 1 - iv + O(v^2))(z_{\mathrm{sp}} - 1 + iv + O(v^2)) \\
&\quad + O\left(s^{-1}(z_{\mathrm{sp}} - 1 + iv + O(v^2))\right) \\
&= |z_{\mathrm{sp}} - 1|^2 + v^2 + O\left(\left(s^{-1} + v^2\right)^{3/2}\right). \quad (4.23)
\end{aligned}$$

Furthermore,

$$(z(v) - z_{\mathrm{sp}})z'(v) = -v + O(v^2). \quad (4.24)$$

Thus

$$\frac{(z(v) - z_{\text{sp}})z'(v)}{(Z_0 - z(v))(z(v) - 1)} = \frac{-v + O(v^2)}{|z_{\text{sp}} - 1|^2 + v^2 + O\left(\left(s^{-1} + v^2\right)^{3/2}\right)}. \tag{4.25}$$

Inserting this into the integral on the second line of (4.22), we see that the $-v$ in (4.25) cancels upon integration. Also

$$\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{v^2}{|z_{\text{sp}} - 1|^2 + v^2} \ln(1 - Be^{-\frac{1}{2}s\eta v^2})dv = O(s^{-1/2}), \tag{4.26}$$

and this finally shows that the integral in (4.22) is $O(s^{-1/2})$. Then combining (4.11), (4.18), and (4.20), we get the result. □

4.2 Bounding the remaining integral

We have from (2.7)

$$P(Q > N) = \frac{c_0}{Z_0^{N+1}(1 - Z_0)} - \frac{1}{2\pi i} \int_{|z|=R} \frac{Q(z)}{z^{N+1}(1 - z)}dz, \tag{4.27}$$

where $R \in (Z_0, |Z_{\pm 1}|)$, and we intend to bound the integral on the right-hand side of (4.27). We use in (4.27) the $Q(z)$ as represented by the right-hand side of (2.5) which is defined and analytic in z, $|z| < r$, $z \neq Z_k$. We write for $|z| < r$, $z \neq Z_k$

$$\frac{Q(z)}{(1 - z)z^{N+1}} = \frac{-1}{z^{N+2-s}} \frac{s - \mu_A}{\prod_{j=1}^{s-1}(1 - z_j)} \frac{1}{z^s - A(z)} \prod_{j=1}^{s-1}(1 - z_j/z). \tag{4.28}$$

Now $s - \mu_A = \gamma\sqrt{s}$, and by Lemma 3.4 and (4.18), we have $\prod_{j=1}^{s-1}(1 - z_j) = P(1) \geq C\gamma\sqrt{s}$ for some $C > 0$ independent of s. Hence $(s - \mu_A)/\prod_{j=1}^{s-1}(1 - z_j)$ is bounded in s. Next, for $|z| \geq Z_0$, we have by Lemma 3.4

$$\prod_{j=1}^{s-1}(1 - z_j/z) = \frac{z}{z - 1} \exp(I(z)), \tag{4.29}$$

with $I(z)$ given by (3.12) and admitting an estimate

$$|I(z)| = O\left(|z - z_{\text{sp}}|^{-1} \int_{-\infty}^{\infty} \ln(1 - Be^{-\frac{1}{2}s\eta t^2})dt\right) = O(1) \tag{4.30}$$

since $\sqrt{s}|z - z_{\mathrm{sp}}|$, $B \in (0, 1)$ and $\eta > 0$ are all bounded away from 0. Therefore, there remains to be considered $(z^s - A(z))^{-1}$. We show below that there is a $C > 0$, independent of s, such that

$$|z^s - A(z)| \geq C|z|^s \tag{4.31}$$

when z is on a contour K as in Fig. 1, consisting of a straight line segment

$$z = \xi + i\eta, \quad \xi = \mathrm{Re}\left[\hat{Z}(\pm \tfrac{1}{2})\right], \quad -\frac{1}{\sqrt{s}}y_0 \leq \eta \leq \frac{1}{\sqrt{s}}y_0, \tag{4.32}$$

and a portion of the circle

$$|z| = R = \sqrt{\mathrm{Re}^2\left[\hat{Z}(\pm \tfrac{1}{2})\right] + \frac{1}{s}y_0^2} \tag{4.33}$$

that are joined at the points $(\mathrm{Re}\left[\hat{Z}(\pm \tfrac{1}{2})\right], \pm\frac{1}{\sqrt{s}}y_0)$. Here

$$\hat{Z}(t) = 1 + \frac{a_0}{\sqrt{s}}((b_0^2 - 2\pi i t)^{1/2} + b_0), \tag{4.34}$$

with $a_0, b_0 > 0$ given in (5.36) and independent of s, approximates the solution $z = Z(t)$, for real t small compared to s, of the equation

$$\frac{n}{s} \ln X(z) - \ln z = \frac{2\pi i t}{s} \tag{4.35}$$

outside the unit disk, according to

$$Z(t) = \hat{Z}(t) + O\left(\frac{t}{s}\right). \tag{4.36}$$

Thus on K, we have from (4.31)

$$\left|\frac{Q(z)}{(1 - z)z^{N+1}}\right| = O\left(\frac{1}{(z - 1)z^{N+1}}\right), \tag{4.37}$$

and we estimate

$$
\left|\frac{1}{2\pi i} \int\limits_{|z|=R} \frac{Q(z)}{z^{N+1}(1 - z)} dz\right| = \left|\frac{1}{2\pi i} \int\limits_{z \in K} \frac{Q(z)}{z^{N+1}(1 - z)} dz\right|
$$
$$
= O\left(\left(\frac{1}{\mathrm{Re}\left[\hat{Z}(\pm \tfrac{1}{2})\right]}\right)^{N+1} \int\limits_{z \in K} \frac{|dz|}{|z - 1|}\right) = O\left(\ln s \left(\frac{1}{\mathrm{Re}[\hat{Z}(\pm \tfrac{1}{2})]}\right)^{N+1}\right). \tag{4.38}
$$

Here we have used that $|z - 1| \geq \mathrm{Re}\left[\hat{Z}(\pm\frac{1}{2})\right] - 1 \geq E/\sqrt{s}, z \in K$, for some $E > 0$ independent of s. Observing that

$$\hat{Z}(0) = 1 + \frac{2a_0 b_0}{\sqrt{s}}, \qquad (4.39)$$

$$\mathrm{Re}\left[\hat{Z}(\pm\tfrac{1}{2})\right] = 1 + \frac{a_0}{\sqrt{s}}\left[(\tfrac{1}{2}b_0^2 + \tfrac{1}{2}(b_0^4 + \pi^2)^{1/2})^{1/2} + b_0\right], \qquad (4.40)$$

we see that

$$\left(\frac{\hat{Z}(0)}{\mathrm{Re}[\hat{Z}(\pm\frac{1}{2})]}\right)^N = \left(1 - \frac{\frac{a_0}{\sqrt{s}}\left[(\tfrac{1}{2}b_0^2 + \tfrac{1}{2}(b_0^4 + \pi^2)^{1/2})^{1/2} - b_0\right]}{1 + \frac{a_0}{\sqrt{s}}[(\tfrac{1}{2}b_0^2 + \tfrac{1}{2}(b_0^4 + \pi^2)^{1/2})^{1/2} + b_0]}\right)^N$$

$$= O(\exp(-\hat{D}N/\sqrt{s})) \qquad (4.41)$$

for some $\hat{D} > 0$ independent of s. Hence, by (4.36), we see that the relative error in (4.27) due to ignoring the integral on the right-hand side is of order $\exp(-DN/\sqrt{s})$ with some $D > 0$, independent of s.

We show the inequality (4.31) for $z \in K$ using the following property of X: there is a $\delta > 0$ and a $\vartheta_1 \in (0, \pi/2)$ such that for any $R \in [1, 1+\delta]$ the function $|X(Re^{i\vartheta})|$ is decreasing in $|\vartheta| \in [0, \vartheta_1]$ while

$$\vartheta_1 \leq |\vartheta| \leq \pi \Rightarrow |X(Re^{i\vartheta})| \leq |X(Re^{i\vartheta_1})|. \qquad (4.42)$$

This property follows from strict maximality of $|X(e^{i\vartheta})|$ in $\vartheta \in [-\pi, \pi]$ at $\vartheta = 0$ and analyticity of $X(z)$ in the disk $|z| < r$ (with $r > 1$).

For the construction of the contour K in (4.31–4.32), we consider the quantity

$$n \ln[X(1 + v)] - \ln(1 + v), \qquad (4.43)$$

where v is of the form

$$v = \frac{2\gamma\mu}{\sigma^2\sqrt{s}} + \frac{x_0 + iy}{\sqrt{s}} = \hat{Z}(0) - 1 + \frac{x_0 + iy}{\sqrt{s}} \qquad (4.44)$$

with $x_0 > 0$ fixed and varying $y \in \mathbb{R}$. We choose x_0 such that the outer curve $Z(t)$ is crossed by $z = 1 + v$ near the points $Z(\pm\frac{1}{2})$, where $z^s - A(z)$ equals $2z^s$. Thus, we choose

$$x_0 = \mathrm{Re}[\sqrt{s}(\hat{Z}(\pm\tfrac{1}{2}) - \hat{Z}_0)] = \mathrm{Re}[a_0((b_0^2 + \pi i)^{1/2} - b_0)]. \qquad (4.45)$$

We have, as in the analysis in the appendix, that

$$n \ln[X(1 + v)] - s \ln(1 + v) = \frac{2\gamma\mu}{\sigma^2}x_0 + (x_0^2 - y^2)$$

$$+ 2i\left(\frac{\gamma\mu}{\sigma^2} - x_0\right)y + O(sv^3) + O(v^2\sqrt{s}). \qquad (4.46)$$

With $x_0 > 0$ fixed and independent of s, see (4.45), the leading part of the right-hand side in (4.46) is independent of s and describes, as a function of the real variable y, a parabola in the complex plane with real part bounded from above by its real value at $y = 0$ and that passes the imaginary axis at the points $\pm \pi i$. Therefore, this leading part has a positive distance to all points $2\pi i k$, integer k. Now take y_0 such that

$$\frac{2\gamma\mu}{\sigma^2} + (x_0^2 - y_0^2) = -\left(\frac{2\gamma\mu}{\sigma^2} + x_0^2\right). \tag{4.47}$$

In Fig. 1, we show the curve K (bold), the approximation $\hat{Z}(t)$ of the outer curve, and the choice $y_0 = \eta_0\sqrt{s}$ for the case that $\gamma = 1$, $\mu/\sigma^2 = 2$, $s = 100$.

It follows from the above analysis, with v as in (4.44), that

$$1 - \frac{X^n(1+v)}{(1+v)^s}, \qquad -y_0 \leq y \leq y_0, \tag{4.48}$$

is bounded away from 0 and has a value $1 - c$, $1 - c^*$ at $y = \pm y_0$, where c is bounded away from 1 and $|c| < 1$. Now write

$$Re^{i\vartheta_0} = \hat{Z}(0) + \frac{x_0 + iy_0}{\sqrt{s}} = 1 + v_0. \tag{4.49}$$

When s is large enough, we have that $R \in [1, 1 + \delta]$ and $0 \leq \vartheta_0 \leq \vartheta_1$, where δ and ϑ_1 are as above in (4.42). We have

$$|X^n(1+v_0)| \leq |c| \, |1 + v_0|^s, \tag{4.50}$$

and by (4.42) and monotonicity of $|X(Re^{i\vartheta})|$, $\vartheta_0 \leq |\vartheta| \leq \vartheta_1$,

$$|X^n(Re^{i\vartheta})| \leq |X^n(Re^{i\vartheta_0})| \leq |c|R^s, \qquad \vartheta_0 \leq |\vartheta| \leq \pi. \tag{4.51}$$

Therefore, (4.31) holds on K with

$$C = \min\left\{1 - |c|, \min_{|y| \leq y_0}\left|1 - \frac{X^n(1+v)}{(1+v)^s}\right|\right\} \tag{4.52}$$

positive and bounded away from 0 as s gets large.

5 Correction terms and asymptotic expansion

In this section, we give a series expansion for the leading term in (3.15) involving the Riemann zeta function. We also show how to find an asymptotic series for $P(Q > N)$ as $N \to \infty$ of which the term involving the dominant pole is the leading term. Before we do so, we first discuss how this leading term is related to the Gaussian random walk and a result of Chang and Peres [5].
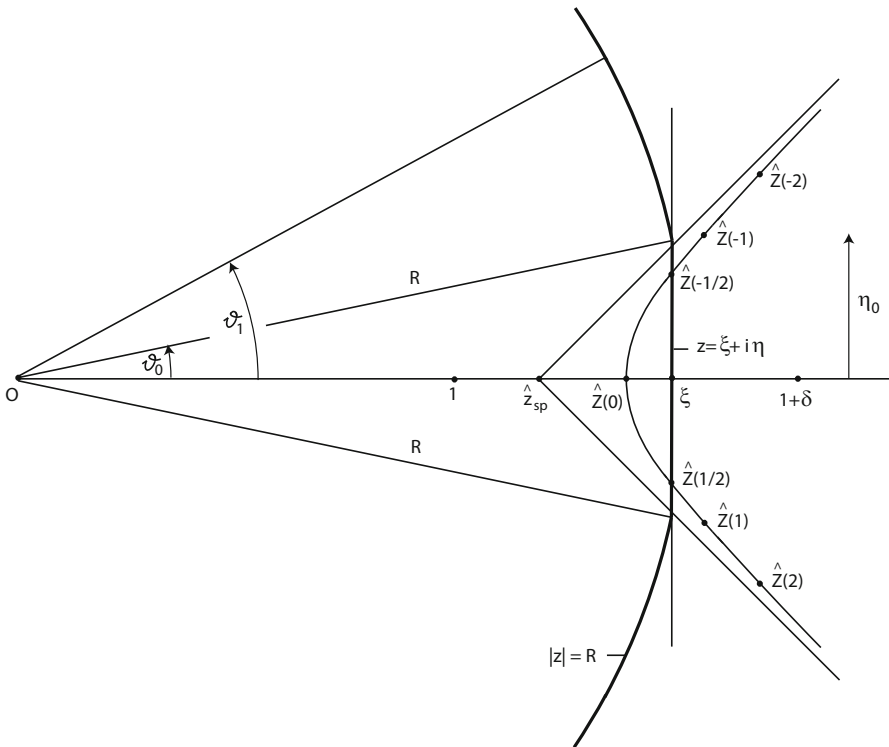
Fig. 1 Integration curve K (*bold*) consisting of line segment $z = \xi + i\eta$, $-\eta_0 \leq \eta \leq \eta_0$, where $\xi = \text{Re}[\hat{Z}(\pm\frac{1}{2})]$ and $\eta_0 = y_0/\sqrt{s}$, and portion of the circle $|z| = R$ with $R = (\xi^2 + \eta_0^2)^{1/2}$. Choice of parameters: $\gamma = 1$, $\mu/\sigma^2 = 2$, $s = 100$

5.1 Connection with Gaussian random walk

We know from [16, Theorem 3] that under the critical many-sources scaling, the rescaled queueing process converges to a reflected Gaussian random walk. The latter is defined as $(S_\beta(k))_{k \geq 0}$ with $S_\beta(0) = 0$ and

$$S_\beta(k) = Y_1 + \ldots + Y_k, \tag{5.1}$$

with $Y_1, Y_2, \ldots$ i.i.d. copies of a normal random variable with mean $-\beta$ and variance 1. Assume $\beta > 0$ (negative drift), and denote the all-time maximum of this random walk by M_β.

Denote by $Q^{(s)}$ the stationary congestion level for a fixed s [that arises from taking $k \to \infty$ in (2.1)]. Then, using $\rho_s = 1 - \gamma/\sqrt{s}$, with

$$\gamma = \frac{\beta \sigma \sqrt{s}}{\mu \sqrt{n}}, \tag{5.2}$$

the spatially scaled stationary queue length reaches the limit $Q^{(s)}/(\sigma\sqrt{n}) \xrightarrow{d} M_\beta$ as $s, n \to \infty$ (see [12,15,16]).

The random variable M_β was studied in [5,9,10]. In particular, [9, Theorem 1] yields, for $\beta < 2\sqrt{\pi}$,

$$\mathbb{P}(M_\beta = 0) = \sqrt{2}\beta \exp\left\{ \frac{\beta}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(\frac{1}{2}-r)(-\frac{1}{2}\beta^2)^r}{r!(2r+1)} \right\}, \tag{5.3}$$

and from [5], we have $\mathbb{P}(M_\beta > K) = h(\beta, K)e^{-2\beta K}$ with

$$h(\beta, K) \to h(\beta) = \exp\left\{ \frac{\beta\sqrt{2}}{\sqrt{\pi}} \sum_{r=0}^{\infty} \frac{\zeta(\frac{1}{2}-r)(-\frac{1}{2}\beta^2)^r}{r!(2r+1)} \right\}, \tag{5.4}$$

exponentially fast as $K \to \infty$.

Hence, there are the approximations

$$\mathbb{P}(Q > K\sqrt{n\sigma^2}) \approx \mathbb{P}(M_\beta > K) \approx h(\beta) \cdot e^{-2\beta K}, \quad \text{as } n \to \infty, \tag{5.5}$$

where the second approximation holds for small values of β. We will now show how this second approximation in (5.5) follows from our leading term in the expansion.

Proposition 5.1

$$\ln\left(\frac{c_0}{1-Z_0} \right) = \frac{2b_0}{\sqrt{\pi}} \sum_{r=0}^{\infty} \frac{\zeta(\frac{1}{2}-r)(-b_0^2)^r}{r!(2r+1)} + O(s^{-1/2}), \quad 0 < b_0 < \sqrt{2\pi}. \tag{5.6}$$

Proof It is shown in [11, Subsect. 5.3] that

$$\ln[Q(0)] = \ln[P(M_\beta = 0)] + O(s^{-1/2}), \tag{5.7}$$

in which we take the drift parameter β according to

$$\beta = b_0\sqrt{2} = \gamma\sqrt{\frac{\mu}{\sigma^2}}. \tag{5.8}$$

From [10], we have

$$\ln[P(M_\beta = 0)] = \ln(2b_0) + \frac{b_0}{\sqrt{\pi}} \sum_{r=0}^{\infty} \frac{\zeta(\frac{1}{2}-r)(-b_0^2)^r}{r!(2r+1)}, \quad 0 < b_0 < \sqrt{2\pi}. \tag{5.9}$$

Then from Proposition 3.5, (5.7) and (5.9), we get the results in (5.6).   □

5.2 Asymptotic series for $P(Q > N)$ as $N \to \infty$

When inspecting the argument that leads to (2.7), it is obvious that one can increase the radius R of the integration contour to values R_M between $|Z(\pm M)|$ and $|Z(\pm(M+1))|$ when $M = 1, 2, ...$ is fixed. Here it must be assumed that s is so large that $|Z_k|$ increases in $k = 0, 1, ..., M+1$. Then, the poles of $Q(z)$ at $z = Z_{\pm k}, k = 0, 1, ..., M$, are inside $|z| = R_M$, and we get

$$P(Q > N) = \frac{c_0}{(1 - Z_0)Z_0^{N+1}} + 2\sum_{k=1}^{M} \text{Re}\left[\frac{c_k}{(1 - Z_k)Z_k^{N+1}}\right]$$
$$-\frac{1}{2\pi i}\int\limits_{|z|=R_M} \frac{Q(z)}{(1 - z)z^{N+1}}dz. \qquad (5.10)$$

As in Subsect. 4.2, one can argue that the integral on the second line of (5.10) is relatively small compared to $|Z_M|^{-N-1}$ when R_M is chosen between but away from $|Z_M|$ and $|Z_{M+1}|$.

We now need the following result.

Lemma 5.2

$$-\frac{s - \mu_A}{sZ_k^{s-1} - A'(Z_k)} = \frac{b_0}{\sqrt{b_0^2 - 2\pi ik}} \frac{1}{Z_k^{s-1}}\left(1 + O\left(\frac{1 + |k|}{\sqrt{s}}\right)\right) \qquad (5.11)$$

when $k = o(s)$.

Proof This follows from the appendix with a similar argument to the proof of Lemma 3.3. □

As to the terms in the series in (5.10), we have for bounded k, see Lemma 5.2,

$$\frac{c_k}{1 - Z_k} = -\frac{s - \mu_A}{sZ_k^{s-1} - A'(Z_k)} \prod_{j=1}^{s-1} \frac{Z_k - z_j}{1 - z_j}$$
$$= \frac{b_0}{\sqrt{b_0^2 - 2\pi ik}} \cdot \prod_{j=1}^{s-1} \frac{1 - z_j/Z_k}{1 - z_j} \cdot \left(1 + O(s^{-1/2})\right). \qquad (5.12)$$

Furthermore, according to Lemma 3.4,

$$\ln\left[\prod_{j=1}^{s-1} \frac{1 - z_j/Z_k}{1 - z_j}\right] = I(Z_k) - I(1) - \ln\left(1 - \frac{1}{Z_k}\right) - \ln(\gamma\sqrt{s})$$
$$= I(Z_k) - I(1) - \ln\left[2b_0\left(b_0 + \sqrt{b_0^2 - 2\pi ik}\right)\right] + O(s^{-1/2}).$$
$$(5.13)$$

Thus, we get the following result.

Proposition 5.3 *For bounded* $k \in \mathbb{Z}$,

$$\frac{c_k}{1 - Z_k} = \frac{\exp(I(Z_k) - I(1))}{2\sqrt{b_0^2 - 2\pi ik}\left(b_0 + \sqrt{b_0^2 - 2\pi ik}\right)}\left(1 + O(s^{-1/2})\right). \tag{5.14}$$

We aim at approximating $I(Z_k)$, showing, in particular, that $c_k/(1 - Z_k) \neq 0$ is bounded away from 0 for bounded k and large s. To that end, we conduct the dedicated saddle point analysis for $I(Z_k)$. We have for $|Z| \geq Z_0$, $\mathrm{Re}(Z) > z_{sp}$,

$$\begin{aligned}
I(Z) &= \frac{1}{2\pi i} \int\limits_{|z| = z_{sp}} \frac{\ln\left(1 - z^{-s}A(z)\right)}{Z - z} dz \\
&= \frac{1}{2\pi i} \int\limits_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{z(v)}{Z - z(v)} \ln(1 - Be^{-\frac{1}{2}s\eta v^2}) dv,
\end{aligned} \tag{5.15}$$

with exponentially small error in the last identity as $s \to \infty$. With $g(z) = -\ln z + \frac{n}{s}\ln X(z)$, we let

$$B = \exp(sg(z_{sp})) = e^{-b_0^2}\left(1 + O(s^{-1/2})\right), \quad \eta = g''(z_{sp}) = \frac{\sigma^2}{\mu} + O(s^{-1/2}), \tag{5.16}$$

and $z(v)$ as in (4.21) and defined implicitly by $g(z(v)) = g(z_{sp}) - \frac{1}{2}v^2 g''(z_{sp})$. We then find, by using $z(v) = z_{sp} + iv + O(v^2)$ and $z'(v) = i + O(v)$, that

$$\begin{aligned}
I(Z) &= \frac{-1}{2\pi i} \int\limits_{-\infty}^{\infty} \frac{\ln(1 - Be^{-\frac{1}{2}s\eta v^2})}{v + i(Z - z_{sp})} dv + O(s^{-1/2}) \\
&= \frac{-1}{2\pi i} \int\limits_{-\infty}^{\infty} \frac{\ln(1 - Be^{-t^2})}{t + i(Z - z_{sp})\sqrt{s\eta/2}} dt + O(s^{-1/2}),
\end{aligned} \tag{5.17}$$

where in the last step the substitution $t = v\sqrt{s\eta/2}$ has been made. Combining in the last integrand in (5.17) the values at t and $-t$ for $t \geq 0$, we get the following result.

Proposition 5.4 *For* $|Z| \geq Z_0$, $\mathrm{Re}(Z) > z_{sp}$,

$$I(Z) = J(d) + O(s^{-1/2}), \tag{5.18}$$

where $d = (Z - z_{\mathrm{sp}})\sqrt{s\eta/2}$, and

$$J(d) = \frac{1}{\pi} \int\limits_0^\infty \frac{d}{t^2 + d^2} \ln(1 - Be^{-t^2})dt. \tag{5.19}$$

In the context of Proposition 5.3, we consider

$$d = d_k = (Z_k - z_{\mathrm{sp}})\sqrt{s\eta/2}, \tag{5.20}$$

with

$$z_{\mathrm{sp}} = 1 + \frac{a_0 b_0}{\sqrt{s}} + O(s^{-1}), \quad Z_k = 1 + \frac{a_0}{\sqrt{s}}\left(\sqrt{b_0^2 - 2\pi i k} + b_0\right) + O(s^{-1}). \tag{5.21}$$

Using the definitions of a_0, b_0 in (5.36) and η in (5.16), we get

$$d_k = \hat{d}_k + O(s^{-1/2}); \quad \hat{d}_k = (b_0^2 - 2\pi i k)^{1/2}. \tag{5.22}$$

We have that
$$|\hat{d}_k| \geq b_0, \quad \arg(\hat{d}_k) \in (-\tfrac{1}{4}\pi, \tfrac{1}{4}\pi), \quad k \in \mathbb{Z}. \tag{5.23}$$

Since for $t \geq 0$ and $\arg(d) \in (-\tfrac{1}{4}\pi, \tfrac{1}{4}\pi)$ we have

$$\ln(1 - Be^{-t^2}) < 0, \quad \arg\left(\frac{d}{t^2 + d^2}\right) \in (-\tfrac{1}{4}\pi, \tfrac{1}{4}\pi), \tag{5.24}$$

we see that we have complete control on the quantities $J(\hat{d}_k)$ (also note (5.16) for this purpose). Using that $-I(1) = I(Z_0) + O(s^{-1/2})$, see (4.20), we get the following result.

Proposition 5.5 *For bounded $k \in \mathbb{Z}$,*

$$\frac{c_k}{1 - Z_k} = \frac{\exp(J(\hat{d}_k) + J(\hat{d}_0))}{2\hat{d}_k(\hat{d}_k + \hat{d}_0)} + O(s^{-1/2}), \tag{5.25}$$

where the leading quantity in (5.25) $\neq 0, \infty$, $\hat{d}_k$ is given in (5.22) with $\hat{d}_0 = b_0$, and J is given in (5.19).

Theorem 5.6 *There is the asymptotic series*

$$P(Q > N) \sim \mathrm{Re}\left[\frac{c_0}{(1 - Z_0)Z_0^{N+1}} + 2\sum_{k=1}^\infty \frac{c_k}{(1 - Z_k)Z_k^{N+1}}\right], \tag{5.26}$$

where the ratio of the terms in the series with index M and $M-1$ is $O(|Z_{M-1}/Z_M|^N)$.

Proof This follows from (5.10), in which the integral is $o(|Z_M|^{-N})$ and the term with $k = M$ is $O(|Z_M|^{-N})$, while the reciprocal of the term with $k = M - 1$ is $O(|Z_{M-1}|^{-N})$ by Proposition 5.5. In the consideration of the terms with $k = M - 1, M$, it is tacitly assumed that s is so large that $|Z_k|, k = 0, 1, ..., M$ is a strictly increasing sequence. $\qquad\square$

Appendix: Proof of Lemma 3.1

We consider the zeros z_j, $j = 0, 1, ..., s - 1$, and $Z_k, k \in \mathbb{Z}$, of the function $z^s - A(z)$ in the unit disk $|z| \leq 1$ and in the annulus $1 < |z| < r$, respectively, in particular those that are relatively close to 1. These zeros are elements of the set $S_{A,s} = \{z \in \mathbb{C}||z| < r, |z^s| = |A(z)|\}$. For $z \in S_{A,s}$, we have that $\ln(z^s X^{-n}(z))$ is purely imaginary. We thus consider the equation

$$s \ln z = n \ln X(z) + 2\pi i t \tag{5.27}$$

with z near 1 and t small compared to s. Writing

$$u = 2\pi t, \quad z = 1 + v, \tag{5.28}$$

we get by Taylor expansion around $z = 1$ the equation

$$s(v - \tfrac{1}{2}v^2 + O(v^3)) = n \ln(1 + X'(1)v + \tfrac{1}{2}X''(1)v^2 + O(v^3)) + iu. \tag{5.29}$$

Dividing by s and using that $\frac{n}{s}X'(1) = 1 - \gamma/\sqrt{s}$ yields

$$v - \tfrac{1}{2}v^2 + O(v^3) = \left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(v + \frac{X''(1) - (X'(1))^2}{2X'(1)}v^2 + O(v^3)\right) + i\frac{u}{s}, \tag{5.30}$$

i.e.,

$$-\frac{\gamma}{\sqrt{s}}v + \frac{\sigma^2}{2\mu}v^2 + i\frac{u}{s} = O\left(\frac{v^2}{\sqrt{s}}\right) + O(v^3), \tag{5.31}$$

where we have used that

$$\mu = X'(1), \quad \sigma^2 = X''(1) - (X'(1))^2 + X'(1) > 0. \tag{5.32}$$

Dividing in (5.31) by $\sigma^2/2\mu$ and completing a square, we get

$$\Big(v - \frac{\gamma\mu}{\sigma^2\sqrt{s}}\Big)^2 = \Big(\frac{\gamma\mu}{\sigma^2\sqrt{s}}\Big)^2\Big(1 - \frac{2iu\sigma^2}{\gamma^2\mu}\Big) + O\Big(\frac{v^2}{\sqrt{s}}\Big) + O(v^3). \qquad (5.33)$$

Taking square roots on both sides of (5.33) and using that the leading term on the right-hand side of (5.33) has order $(1 + |u|)/s$, we get

$$v = \frac{\gamma\mu}{\sigma^2\sqrt{s}} \pm \frac{\gamma\mu}{\sigma^2\sqrt{s}}\Big(1 - \frac{2iu\sigma^2}{\gamma^2\mu}\Big)^{1/2} + O\Big(\frac{|v|^2 + |v|^3\sqrt{s}}{\sqrt{1 + |u|}}\Big). \qquad (5.34)$$

Irrespective of the $\pm$-sign, the leading part of the right-hand side of (5.34) has order $((1 + |u|)/s)^{1/2}$ (and even $(|u|/s)^{1/2}$ in the case of the $-$-sign), and the O-term has order $(1 + |u|)/s$ which is $o(((1 + |u|)/s)^{1/2})$ as long as $|u|/s = o(1)$. Thus, in that regime of u, we have

$$\begin{aligned} z = 1 + v &= 1 + \frac{\gamma\mu}{\sigma^2\sqrt{s}}\Big(1 \pm \Big(1 - \frac{2iu\sigma^2}{\gamma^2\mu}\Big)^{1/2}\Big) + O\Big(\frac{1 + |u|}{s}\Big) \\ &= 1 \pm \frac{a_0}{\sqrt{s}}((b_0^2 - iu)^{1/2} \pm b_0) + O\Big(\frac{1 + |u|}{s}\Big), \end{aligned} \qquad (5.35)$$

where we have inserted

$$a_0 = \Big(\frac{2\mu}{\sigma^2}\Big)^{1/2}, \qquad b_0 = \Big(\frac{\gamma^2\mu}{2\sigma^2}\Big)^{1/2} = \tfrac{1}{2}\gamma a_0. \qquad (5.36)$$

In the case of the minus sign in (5.35), the O-term may be replaced by $O(|u|/s)$. Choosing $u = 2\pi j$ with $j = 0, 1, ...,$ and $j = o(s)$, we get from (5.35) with the minus sign, (3.3). Choosing $u = 2\pi k$ with $k \in \mathbb{Z}$ and $k = o(s)$, we get from (5.35) with the plus sign, (3.4).

References

1. Anick, D., Mitra, D., Sondhi, M.M.: Stochastic theory of a data-handling system with multiple sources. Bell Syst. Tech. J. **61**(8), 1871–1894 (1982)
2. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
3. Blanchet, J., Glynn, P.: Complete corrected diffusion approximations for the maximum of a random walk. Ann. Appl. Probab. **16**(2), 951–983 (2006)
4. Bruneel, H., Kim, B.G.: Discrete-Time Models for Communication Systems Including ATM. Kluwer Academic Publishers, Boston (1993)
5. Chang, J.T., Peres, Y.: Ladder heights, Gaussian random walks and the Riemann zeta function. Ann. Probab. **25**(2), 787–802 (1997)
6. Dai, J.G., Shi, P.: A two-time-scale approach to time-varying queues for hospital inpatient flow management. Soc. Sci. Res. Network (2015). doi:10.2139/ssrn.2489533
7. Flajolet, P., Sedgewick, R.: Analytic Combinatorics, 1st edn. Cambridge University Press, New York (2009)
8. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)

9. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: On Lerch's transcendent and the Gaussian random walk. Ann. Appl. Probab. **17**, 421–439 (2006)
10. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Cumulants of the maximum of the Gaussian random walk. Stoch. Process. Appl. **117**(12), 1928–1959 (2007)
11. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Mathijsen, B.W.J.: Novel heavy-traffic regimes for large-scale service systems. SIAM J. Appl. Math. **75**, 212–787 (2015)
12. Jelenković, P., Mandelbaum, A., Momčilovic, P.: Heavy traffic limits for queues with many deterministic servers. Queueing Syst. **47**, 53–69 (2004)
13. Newell, G.F.: Queues for a fixed-cycle traffic light. Ann. Math. Stat. **31**, 589–597 (1960)
14. Siegmund, D.: Sequential Analysis. Springer Series in Statistics. Springer, New York (1985)
15. Sigman, K., Whitt, W.: Heavy-traffic limits for nearly deterministic queues. J. Appl. Probab. **48**(3), 657–678 (2011)
16. Sigman, K., Whitt, W.: Heavy-traffic limits for nearly deterministic queues: stationary distributions. Queueing Syst. Theory Appl. **69**(2), 145–173 (2011)
17. van Leeuwaarden, J.S.H.: Queueing models for cable access networks. Ph.D. thesis, Eindhoven University of Technology (2005)
18. van Leeuwaarden, J.S.H.: Delay analysis for the fixed-cycle traffic light queue. Transp. Sci. **40**, 189–199 (2006)
19. Van Mieghem, P.: The asymptotic behavior of queueing systems: large deviations theory and dominant pole approximation. Queueing Syst. Theory Appl. **23**(1–4), 27–55 (1996)