

Staffing many-server systems with admission control and retrials (report version)

Citation for published version (APA):

Janssen, A. J. E. M., & Leeuwaarden, van, J. S. H. (2013). *Staffing many-server systems with admission control and retrials (report version)*. (arXiv.org; Vol. 1302.3006 [math.PR]). s.n.

Document status and date:

Published: 01/01/2013

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Staffing many-server systems with admission control and retrials (report version)

A.J.E.M. Janssen*

J.S.H. van Leeuwen†

January 30, 2014

Abstract

In many-server systems it is crucial to staff the right number of servers so that targeted service levels are met. These staffing problems typically lead to constraint satisfaction problems that are hard to solve. During the last decade, a powerful many-server asymptotic theory has been developed to solve such problems and optimal staffing rules are known to obey the square-root staffing principle. This paper develops many-server asymptotics in the so-called QED regime, and presents refinements to many-server asymptotics and square-root staffing for a Markovian queueing model with admission control and retrials.

Keywords: Many-server systems, QED regime, Halfin-Whitt regime, heavy traffic, diffusion limits, admission control, square-root staffing, optimality gap, asymptotic dimensioning

AMS 2000 Subject Classification: 60K25, 68M10, 41A60.

1 Introduction

A key challenge in large many-server systems is to balance the trade-off between operational costs and quality-of-service offered to customers. There is by now a vast literature on the asymptotic analysis of many-server systems. Many papers describe limiting behavior of a sequence of queues, which is then used to approximate the characteristics of a member of the sequence, i.e., the performance of a finite-sized queueing system. Depending on how this sequence is parameterized, its limiting behavior is different, giving rise to different approximations (see [3] and [12]). One of the most popular approximations arises in the Quality-and-Efficiency-Driven (QED) regime, also known as the Halfin-Whitt regime [8], in which the number of servers s and the offered workload λ are related according to a square-root principle $s = \lambda + \beta\sqrt{\lambda}$, for a constant β , and s (and λ) are taken to infinity.

In this paper we consider a Markovian many-server system with admission control in the QED regime. We consider an admission control policy that lets an arriving customer enter the system according to a probability depending on the queue length. In particular, a customer meeting upon arrival k other customers is admitted with probability p_k , and we shall allow for a wide range

*Eindhoven University of Technology and Eurandom, Department of Mathematics and Computer Science and Department of Electrical Engineering, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail a.j.e.m.janssen@tue.nl

†Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail j.s.h.v.leeuwen@tue.nl

of such admission policies characterized by $(p_k)_{k \geq s}$. For such admission-controlled systems, we consider the QED regime and the square-root staffing rule $\lambda = s - \gamma\sqrt{s}$ for some fixed constant γ . This is asymptotically (as $s \rightarrow \infty$), and for all practical purposes, equivalent with setting $s = \lambda + \beta\sqrt{\lambda}$ for some fixed constant β . The reason for staffing via λ instead of s is that it allows for mathematically more elegant derivations. All results in this paper regarding $\lambda = s - \gamma\sqrt{s}$ can be converted into results for $s = \lambda + \beta\sqrt{\lambda}$ at the cost of increased notational burden.

Since certain customers are rejected upon arrival, it seems natural to extend the model with the feature that rejected customers can reattempt. The modeling of reattempts or retrials is known to be challenging [4, 6], which is why one often resorts to computational approaches [1]. These numerical approaches face increasing numerical difficulties when the number of servers becomes large, which is precisely the regime we are interested in. Therefore, we combine the QED regime with a limiting regime for slow retrials, meaning that rejected customers reattempt after a relatively (compared to the time scale of the system) long time. The combination of these two asymptotic regimes leads to a tractable model with a closed-form solution, for which we are able to derive QED approximations for some of the relevant performance measures.

We leverage these QED approximations to obtain results for staffing problems. The QED regime is particularly suited for staffing many-server systems, as it combines large capacity with high utilization. This appealing feature was exploited in [3] that introduced a rigorous asymptotic framework for applying the square-root staffing principle to two classes of problems: constraint satisfaction and cost minimization. In [3] it was observed that square-root staffing based on QED approximations is effective and accurate over a wide range of system parameters for the Erlang C model. We apply the same approach to develop square-root staffing rules for several constraint satisfaction problems for the system at hand.

Square-root staffing. The core of staffing problems in many-server systems is to determine the right trade-off between quality and capacity. Quality is formulated in terms of some targeted service level. Take as an example the delay probability $D_F(s, \lambda)$ (see (2.8)). A large delay probability is perceived as negative, and the targeted service level could be to keep the delay probability below some value ε . The smaller ε , the better the offered service. Once the targeted service level is set, the objective from the system's perspective is to determine the largest load λ (or the lowest staffing level s) such that the target $D_F(s, \lambda) \leq \varepsilon$ is met. This is what we have referred to as a constraint satisfaction problem.

The delay probability is a function of the two model parameters s and λ , and of the admission policy. Because the delay probability is a continuous and monotone increasing function in λ , the constraint satisfaction problem is equivalent to finding the λ_{opt} such that $D_F(s, \lambda_{\text{opt}}) = \varepsilon$. To solve this inverse problem, we shall invoke the theory of asymptotic dimensioning introduced in [3] and extended in the first part of this paper to admission control and retrials. This theory fully exploits the QED regime for large systems, in a way that reduces considerably the complexity of the inverse problem. That is, in the QED regime (with $\lambda = s - \gamma\sqrt{s}$, $s \rightarrow \infty$, γ not scaling with s), the performance measures in our model can be approximated by their limiting counterparts. For instance, $D_F(s, \lambda)$ can be approximated by some function $D_*(\gamma)$ that only depends on γ (and no longer on s or λ). Hence, the inverse problem can then be approximatively solved by searching for the γ_* such that $D_*(\gamma_*) = \varepsilon$, and then setting the load according to $\lambda_* = s - \gamma_*\sqrt{s}$. We refer to this rule as *conventional square-root staffing*. In this asymptotic approach, one expects that the better the approximation $D_F(s, \lambda) \approx D_*(\gamma)$, the smaller the error $|\lambda_{\text{opt}} - \lambda_*|$. Based on the QED regime, one

also expects the approximation λ_* to be accurate for large values of s , i.e., for large-scale systems with many servers. Indeed, we prove that $|\lambda_{\text{opt}} - \lambda_*| = O(1)$, where a function $f(\lambda) = O(g(\lambda))$ if $\limsup_{\lambda \rightarrow \infty} |f(\lambda)/g(\lambda)| < \infty$.

Refined staffing. We shall also derive refined staffing rules, for which we first develop refined QED approximations for the objective function, and then characterize the approximative solutions to the constraint satisfaction problems. The refined staffing rules are of the form

$$\lambda_{\bullet} = s - \gamma_* \sqrt{s} + r_{\bullet}, \quad (1.1)$$

with r_{\bullet} a simple function of γ_* , s and ε . We shall uniquely identify r_{\bullet} (for each considered constraint satisfaction problem), and prove that the refined staffing level in (1.1) satisfies

$$\lambda_{\text{opt}} - \lambda_{\bullet} = O(s^{-1/2}), \quad (1.2)$$

We refer to the order term that expresses the difference between the exact optimal staffing level and the approximate staffing level as the *optimality gap*. Hence, the optimality gap of λ_{\bullet} is $O(s^{-1/2})$, which suggests that the staffing level λ_{\bullet} becomes more accurate as s increases. Note that $\lambda_{\bullet} = \lambda_* + \gamma_{\bullet}$. Since the optimality gap of the conventional staffing level λ_* equals $O(1)$, we can expect that λ_{\bullet} should be a more accurate prescription than λ_* . In addition, because γ_{\bullet} in fact describes the optimality gap of λ_* , or more precisely,

$$\lambda_{\text{opt}} - \lambda_* = \gamma_{\bullet} + O(s^{-1/2}), \quad (1.3)$$

it allows us to perform an analytical assessment of the accuracy of both conventional and refined square-root staffing.

The influence of admission policies and retrials. Our main results are captured in Theorems 7 and 8, which formally establish the necessary results for both conventional staffing rules and their refinements. These two theorems establish that as the system size s increases, the difference between the true maximal load that adheres to the constraint λ_{opt} , and the conventional square-root staffing prescription λ_* , converges to the real number γ_{\bullet} , see (1.3). We stress that these results hold across the large class of admission control policies considered in this paper, and for the systems with and without retrials.

We obtain the explicit expression of γ_{\bullet} in each case, and show that it is always a simple function of γ_* . We can therefore investigate numerically the impact of the refinements in comparison with conventional square-root staffing. It turns out that, unlike in the classical Erlang C model [3], the refinement γ_{\bullet} is significant in many cases, due to different system parameters. Our findings further suggest that in the presence of admission control, more care needs to be taken in applying many-server asymptotic results to small or moderate size systems. In particular, when the admission control becomes more lenient, and hence on average more customers are admitted, our numerical findings suggest that the refinements to the conventional square-root staffing are necessary.

Overview of the paper. In this paper we apply the QED regime to a Markovian many-server system with admission control and retrials. We analyze this system in the limiting regime, in a similar spirit as was done for the Erlang C model [8, 11], the Erlang B model [10] and the Erlang A model (with abandonments) [14]. In the recent paper [2] on a many-server system with retrials, the

admission control policy that rejects all delayed customers (loss system) was analyzed. The current paper strongly builds on several results obtained in [2]. Compared to earlier studies [8, 11, 10, 14, 2], the system in this paper brings about additional mathematical challenges, because of the effects of rejection and reattempts. In short, we make the following contributions:

- (i) We consider two stationary performance measures: The probability that an arriving customer finds all servers occupied $D_F(s, \lambda)$, and the probability that an arriving customer is rejected $D_F^R(s, \lambda)$. For both performance measures we derive in Theorems 3 and 4 the limiting expressions in the QED regime, for the cases with and without retrials.
- (ii) We next consider dimensioning problems of the type: For fixed s , find λ for which the rejection probability $D_F^R(s, \lambda)$ has some prescribed value. We solve this inverse problem both for finite s and for $s \rightarrow \infty$ in the QED regime. The inverse problem in the QED regime is easier to solve, and provides a leading-order estimate, in terms of the hazard rate of the standard normal distribution, for the finite- s inverse problem. We characterize explicitly the error made by replacing the finite- s inverse problem by its QED counterpart, by deriving correction terms in Theorems 7 and 8. These correction terms are then also used to develop the refined staffing rules and establish the optimality gaps.
- (iii) We then consider dimensioning problems of the type: For fixed s , find λ for which the carried traffic $\lambda(1 - D_F^R(s, \lambda))$ has a prescribed value. These dimensioning problems are shown to have typically two solutions: A small λ leads to a large probability of being admitted $1 - D_F^R(s, \lambda)$, while a large λ leads to a small admittance probability; see Theorems 9 and 11. This phenomenon of two solutions has also been observed in certain loss networks with alternative two-link routing [7].
- (iv) The more technical part of this paper deals with incorporating the effect of retrials, for which an essential role is played by the generalized Cohen equation $\Omega = (\lambda + \Omega)D_F^R(s, \lambda + \Omega)$ and its solution Ω representing the average rate of retrials. In the next section we explain in more detail the fundamental role of this equation in our study.

The paper is structured as follows. In Section 2 we describe the model. In Section 3 we present the limiting expressions for the performance measures in the QED regime, and we use these results to deal with the dimensioning problems in Section 4. Finally, we present in Section 5 results for dimensioning problems with multiple solutions. The proofs of the main results are given in Sections 6-8, while the proofs of supporting results are deferred to the appendix.

2 Model description

The basic model described in this section is taken from [5]. Consider a system with s parallel servers to which customers arrive according to a Poisson process with rate λ . The service times of customers are exponentially distributed with mean one. An *admission policy* dictates whether a customer is admitted to the system or rejected. A customer that finds upon arrival a free server is immediately assigned to that server, and leaves the system after service. A customer that finds upon arrival k other customers in the system, $k \geq s$, is allowed to join the queue with probability p_k and is rejected with probability $1 - p_k$. In this way, the sequence $(p_k)_{k \geq s}$ defines the admission policy. Since we are interested in large many-server systems working at critical load, and hence serving

many customers, the parameter p_k should be interpreted as the fraction of customers admitted, instead of the probability that determines the fate of just one single customer. For the results presented in this paper we impose only mild conditions on the sequence $(p_k)_{k \geq s}$, allowing for a wide range of admission policies to be considered.

Under the above Markovian assumptions, and assuming that all interarrival times and service times are mutually independent, the system can be described as a birth-and-death process $(C(t))_{t \geq 0}$ with $C(t)$ the number of customers in the system at time t . The birth and death rates from state k are λp_k (with $p_k = 1$ for $k = 0, \dots, s-1$) and $\min(k, s)$, respectively. Assuming the stationary distribution exists, with $\pi_k = \lim_{t \rightarrow \infty} \mathbb{P}(C(t) = k)$, it readily follows from solving the detailed balance equations that

$$\pi_k = \begin{cases} \pi_0 \frac{\lambda^k}{k!}, & 1 \leq k \leq s, \\ \pi_0 \frac{\lambda^k}{s! s^{k-s}} \prod_{j=s}^{k-1} p_j, & k \geq s+1, \end{cases} \quad (2.1)$$

with

$$\pi_0^{-1} = \sum_{k=0}^s \frac{\lambda^k}{k!} + \frac{\lambda^s}{s!} F\left(\frac{\lambda}{s}\right) \quad (2.2)$$

and

$$F(x) = \sum_{n=0}^{\infty} p_s \cdots p_{s+n} x^{n+1}. \quad (2.3)$$

From (2.2) it can be seen that the stationary distribution exists when $F(\frac{\lambda}{s}) < \infty$. Since $p_{s+n} \in [0, 1]$, we have that $F(\frac{\lambda}{s}) < \infty$ when $0 \leq \lambda < s$. When $\lambda \geq s$ we need to be more careful. The radius of convergence of the power series $F(x)$ is given by $1/P$ with

$$P := \limsup_{n \rightarrow \infty} (p_s \cdots p_{s+n})^{\frac{1}{n+1}} \in [0, 1]. \quad (2.4)$$

In a major portion of the main text, we assume the following condition:

$$P \in [0, 1), \quad F\left(\frac{1}{P} - 0\right) = \lim_{x \uparrow \frac{1}{P}} F(x) = \infty. \quad (2.5)$$

Under this condition it can be easily observed from (2.3) that the stability condition for our system becomes

$$\lambda \in [0, \lambda_P) \quad \text{with} \quad \lambda_P = \frac{s}{P}, \quad (2.6)$$

where $\lambda_P = \infty$ when $P = 0$. The condition (2.5) is certainly not as general as possible to develop the theory, but it excludes cases that need separate consideration, thereby distracting attention from the bottom line of the exposition. Stability is guaranteed when $\lambda < s$. Also, when $\lim_{k \rightarrow \infty} p_k = 0$, we have $P = 0$ and thus stability for all $\lambda \geq 0$. Condition (2.5) is also satisfied for the case $p_k = p \in (0, 1)$ for all $k \geq s$, where $F(x) = px/(1-px)$ so that $F(\frac{1}{P} - 0) = \infty$, with $P = p$, indeed. We exclude at this point the case $P = 1$, which would for example occur in the case $p_k = 1$, $k \geq s$, and the cases that $F(\frac{1}{P} - 0) < \infty$. However, in Section 6 and Appendix A, the results are proved under general conditions.

Let

$$B(s, \lambda) = \frac{\lambda^s / s!}{\sum_{k=0}^s \lambda^k / k!} \quad (2.7)$$

denote the Erlang B formula representing the stationary blocking probability in an $M/M/s/s$ system. A crucial performance measure is the stationary probability $D_F(s, \lambda) = \sum_{k=s}^{\infty} \pi_k$ that an arriving customer finds all servers occupied, given by

$$D_F^{-1}(s, \lambda) = \frac{B^{-1}(s, \lambda) + F(\frac{\lambda}{s})}{1 + F(\frac{\lambda}{s})}, \quad (2.8)$$

where D_F^{-1} is short-hand notation for $(D_F)^{-1}$. Note that for $p_k = 0$, $k \geq s$, the term $F(\lambda/s)$ vanishes, and the probability $D_F(s, \lambda)$ reduces to $B(s, \lambda)$. Also, for $p_k = 1$, $k \geq s$, the probability $D_F(s, \lambda)$ reduces to the Erlang C formula given by

$$C(s, \lambda) = \frac{\frac{\lambda^s}{(s-1)!(s-\lambda)}}{\sum_{k=0}^{s-1} \lambda^k / k! + \frac{\lambda^s}{(s-1)!(s-\lambda)}}, \quad (2.9)$$

representing the stationary delay probability in an $M/M/s$ system.

Another important performance measure is the stationary probability $D_F^R(s, \lambda) = \sum_{k=s}^{\infty} \pi_k p_k$ of being rejected, given by

$$D_F^{-R}(s, \lambda) = \frac{B^{-1}(s, \lambda) + F(\frac{\lambda}{s})}{1 + (1 - s/\lambda)F(\frac{\lambda}{s})}. \quad (2.10)$$

It follows from results in Section 6 that

$$\max\{0, 1 - \frac{s}{\lambda}\} \leq D_F^R(s, \lambda) \leq B(s, \lambda) \leq D_F(s, \lambda) \leq 1 \quad (2.11)$$

for $0 \leq \lambda < \lambda_P$. Moreover, as a consequence of condition (2.5) we have

$$D_F(s, \lambda_P - 0) = 1, \quad D_F^R(s, \lambda_P - 0) = 1 - P. \quad (2.12)$$

The birth-and-death process $(C(t))_{t \geq 0}$ relies on the assumption that rejected customers are considered lost. Alternatively, we could assume that rejected customers reattempt to enter the system after some time. In that case, rejected customers start producing reattempts until they are allowed to enter. Assume that periods between successive reattempts of a rejected customer are exponentially distributed with rate μ , independent of interarrival and service times. The system can then be described as a two-dimensional process $(C(t), N(t))_{t \geq 0}$ with $C(t)$ the number of customers in the system and $N(t)$ the number of rejected customers at time t . Under the above assumptions this process is a continuous-time Markov chain on the lattice infinite strip $\{0, 1, \dots, s\} \times \mathbb{Z}_+$.

Since the transition rates of this process clearly depend on the second coordinate, the process $\{(C(t), N(t)); t \geq 0\}$ is hard to analyze. In fact, even deriving the stationary distribution poses analytical difficulties, and no closed-form solution seems to be available. We therefore make the following assumption: Reattempts arrive to the system according to a Poisson process with rate Ω , independent of the Poisson process of customers that arrive to the system for the first time. This assumption is also known as the *retrials see time averages* (RTA) approximation. Under this

assumption, the total flow of customers arriving to the system is a Poisson process with rate $\lambda + \Omega$. The unknown rate Ω should then be the solution to

$$\Omega = (\lambda + \Omega)D_F^R(s, \lambda + \Omega). \quad (2.13)$$

Equation (2.13) is intuitively clear as it equates two expressions for the rate of reattempts. In some cases the RTA approximation can be theoretically justified. Cohen [4] showed that the system with $p_k = 0$ for $k \geq s$, in the limit as $\mu \downarrow 0$, behaves as an Erlang loss system, except with an increased arrival intensity. More specifically, as $\mu \downarrow 0$, the distribution of the number of busy servers converges to the corresponding distribution for the standard Erlang loss system $M/M/s/s$ (which is a truncated Poisson distribution), but with increased arrival rate $\lambda + \Omega$, where Ω is defined as the solution to (2.13). Indeed, in the case of infinitely long retrial times, it fits intuition that the flow of reattempts is independent from the flow of primary customers. For retrial queues with finite retrial times, the RTA approximation has proved useful and accurate for many retrial systems. We shall refer to (2.13) as the *generalized Cohen equation*.

Theorem 1 (Unique solution Cohen's equation). *Under condition (2.5), there is a unique solution $\Omega_{s,F}(\gamma)$ of equation (2.13) for any $\lambda \in (0, s)$.*

A proof of Theorem 1 can be distilled from [5, Section 3]. In Appendix A we shall present a self-contained proof of Theorem 1 under general conditions.

3 QED limits

The QED regime for many-server systems refers to scaling of the arrival rate λ and the number of servers s such that, while both λ and s increase toward infinity, the traffic intensity $\rho = \lambda/s$ approaches unity and

$$(1 - \rho)\sqrt{s} \rightarrow \gamma, \quad (3.1)$$

where γ is a fixed constant. The scaling combines large capacity with high utilization. For the Erlang loss and delay systems, this kind of scaling leads to the classical results (see e.g. [10, Section 5.2]), for $\gamma \in (-\infty, \sqrt{s})$ fixed,

$$\frac{1}{\sqrt{s}} B^{-1}(s, s - \gamma\sqrt{s}) = \frac{\Phi(\gamma)}{\varphi(\gamma)} + O\left(\frac{1}{\sqrt{s}}\right), \quad s \rightarrow \infty, \quad (3.2)$$

and for $\gamma \in (0, \sqrt{s})$ fixed,

$$\lim_{s \rightarrow \infty} C(s, s - \gamma\sqrt{s}) = \left(1 + \gamma \frac{\Phi(\gamma)}{\phi(\gamma)}\right)^{-1}, \quad (3.3)$$

where $\Phi(x)$ and $\varphi(x)$ denote the standard normal cumulative distribution function and density, respectively.

The following result will prove useful in establishing QED limiting results.

Lemma 2 (Two decompositions). *For $\lambda \in [0, \lambda_P)$,*

$$D_F^{-1}(s, \lambda) = (1 - q_\lambda)B^{-1}(s, \lambda) + q_\lambda C^{-1}(s, \lambda), \quad (3.4)$$

$$D_F^{-R}(s, \lambda) = B^{-1}(s, \lambda) + \frac{q_\lambda}{1 - q_\lambda} C^{-1}(s, \lambda), \quad (3.5)$$

where

$$q_\lambda = \frac{\frac{s}{\lambda} F\left(\frac{\lambda}{s}\right)}{1 + F\left(\frac{\lambda}{s}\right)}. \quad (3.6)$$

The proof of Lemma 2 is given in Appendix A. It can also be obtained from more general results in [13]. Note that the function $C(s, \lambda)$ is also defined for $\lambda > s$, while in the $M/M/s$ queue the stability condition is $\lambda < s$. It is further shown in Appendix A that $0 \leq q_\lambda \leq 1$, with $q_\lambda = 1$ if and only if $p_k = 1$, $k \geq s$. Thus, for instance, D_F^{-R} always exceeds B^{-1} and the excess is given by the second term of the right-hand side of (3.5), which is the product of a factor entirely determined by the admission policy and the Erlang C formula. Also, D_F^{-1} is a convex combination, with a γ -dependent convexity parameter $1 - q_\lambda$, of the Erlang B and C formulae. When an admission policy is mild, implying that q_λ is close to 1, we have that the Erlang C formula is dominant. When an admission policy is strict, the Erlang B formula is dominant. Aside from these general comments, the variety of weight functions q_λ that can occur, see (3.6), is rather substantial. In the QED regime, though, the Erlang B formula is always dominant. Indeed, in the QED limit, we have $\lambda/s \rightarrow 1$, and so $q_\lambda \rightarrow F(1)/(1 + F(1))$ which is a finite number $\in (0, 1)$ by our condition (2.5). Now (3.2) and (3.3) show that B^{-1} grows like \sqrt{s} while C^{-1} remains bounded as $s \rightarrow \infty$.

We now apply the scaling (3.1) to the system with admission policy. We henceforth keep working with the notation for the QED regime in (3.1), which is why we reformulate the stability condition (2.6) as

$$\gamma \in (\gamma_P, \sqrt{s}] \quad (3.7)$$

with

$$\gamma_P = -\frac{1-P}{P} \sqrt{s} \in (-\infty, 0). \quad (3.8)$$

Theorem 3 (QED limits without retrials). *Under condition (2.5), for $\lambda = s - \gamma\sqrt{s}$, with $\gamma \in (-\infty, \sqrt{s}]$ fixed,*

$$\lim_{s \rightarrow \infty} \sqrt{s} D_F(s, s - \gamma\sqrt{s}) = (1 + F(1)) \frac{\phi(\gamma)}{\Phi(\gamma)}, \quad (3.9)$$

$$\lim_{s \rightarrow \infty} \sqrt{s} D_F^R(s, s - \gamma\sqrt{s}) = \frac{\phi(\gamma)}{\Phi(\gamma)}. \quad (3.10)$$

Proof. We have by continuity of $F(x)$ at $x = 1$ that for fixed $\gamma \in (-\infty, \sqrt{s}]$

$$F\left(\frac{s - \gamma\sqrt{s}}{s}\right) = F\left(1 - \frac{\gamma}{\sqrt{s}}\right) = F(1) + o(1), \quad s \rightarrow \infty. \quad (3.11)$$

Therefore,

$$\frac{1}{\sqrt{s}} D_F^{-1}(s, s - \gamma\sqrt{s}) = \frac{1}{1 + F(1)} \frac{\Phi(\gamma)}{\varphi(\gamma)} + o(1), \quad s \rightarrow \infty, \quad (3.12)$$

and

$$\frac{1}{\sqrt{s}} D_F^{-R}(s, s - \gamma\sqrt{s}) = \frac{\Phi(\gamma)}{\varphi(\gamma)} + O\left(\frac{1}{\sqrt{s}}\right), \quad s \rightarrow \infty. \quad (3.13)$$

This implies (3.9) and (3.10). \square

A first observation is that the limiting expressions (3.9) and (3.10) are similar as for the Erlang B formula in (3.2), and the only difference between the limits of D_F and B is the factor $1 + F(1)$, which incorporates all information about the admission policy.

Let $\bar{D}_F(s, \lambda) = D_F(s, \lambda + \Omega)$ and $\bar{D}_F^R(s, \lambda) = D_F^R(s, \lambda + \Omega)$ with Ω as in (2.13). Hence, $\bar{D}_F(s, \lambda)$ and $\bar{D}_F^R(s, \lambda)$ are the stationary probability that an arriving customer finds all servers occupied, and the stationary probability that an arriving customer is rejected, respectively, in the system with retrials using the RTA approximation.

Theorem 4 (QED limits with retrials). *Under condition (2.5), for $\lambda = s - \gamma\sqrt{s}$, with $\gamma \in (-\infty, \sqrt{s}]$ fixed, and with Ω defined as in (2.13),*

$$\lim_{s \rightarrow \infty} \frac{\Omega}{\sqrt{s}} = a \quad (3.14)$$

with a the unique positive solution of the equation

$$a = \frac{\varphi(\gamma - a)}{\Phi(\gamma - a)}. \quad (3.15)$$

Furthermore,

$$\lim_{s \rightarrow \infty} \sqrt{s} \bar{D}_F(s, s - \gamma\sqrt{s}) = (1 + F(1)) \frac{\phi(\gamma - a)}{\Phi(\gamma - a)}, \quad (3.16)$$

$$\lim_{s \rightarrow \infty} \sqrt{s} \bar{D}_F^R(s, s - \gamma\sqrt{s}) = \frac{\phi(\gamma - a)}{\Phi(\gamma - a)}. \quad (3.17)$$

The proof of Theorem 4 is presented in Section 7. Theorem 4 shows that the additional load due to retrials Ω , for a system with many servers, is of the order \sqrt{s} . In particular, as the number of servers grows large, Ω is well approximated by $a\sqrt{s}$, where a is a constant that no longer depends on s . This also means that for the overall retrial system the arrival rate $\lambda + \Omega$ is approximately $s - (\gamma - a)\sqrt{s}$.

4 Dimensioning problems

First consider the situation without retrials, and the problem of finding the arrival rate λ such that the probability $D_F(s, \lambda)$ to find all servers occupied or the probability $D_F^R(s, \lambda)$ that service is denied altogether has a prescribed value. Here the number of servers and the admission policy, embodied by F , are assumed to be given.

Problem 1 For fixed s, ε , find γ such that

$$\sqrt{s} D(s, s - \gamma\sqrt{s}) = \varepsilon \quad \text{with} \quad D = D_F \text{ or } D_F^R. \quad (4.1)$$

In Section 6 it will be shown that, under condition (2.5), $D_F(s, s - \gamma\sqrt{s})$ decreases strictly from 1 at $\gamma = \gamma_P$ to 0 at $\gamma = \sqrt{s}$. Unfortunately, such a result does not hold for D_F^R : There are policies F satisfying condition (2.5) such that $D_F^R(s, s - \gamma\sqrt{s})$ is not monotonic as a function of γ , see Appendix A. These policies are, however, rather rare. Relevant policies for which $D_F^R(s, s - \gamma\sqrt{s})$ is monotonic include

$$p_k = p \in (0, 1), \quad k \geq s, \quad (4.2)$$

and, for some $N \geq s$,

$$p_k = \begin{cases} 1, & s \leq k \leq N, \\ 0, & k \geq N + 1. \end{cases} \quad (4.3)$$

When $D_F^R(s, s - \gamma\sqrt{s})$ is monotonic, it decreases from $1 - P$ at $\gamma = \gamma_P$ to 0 at $\gamma = \sqrt{s}$.

We have the following result:

Theorem 5 (Unique solutions). *Under condition (2.5),*

(i) *Equation (4.1) with $D = D_F$ has a unique solution $\gamma = \gamma_{s,F}(\varepsilon)$ for any $\varepsilon \in (0, \sqrt{s})$.*

Assuming further that $D_F^R(s, s - \gamma\sqrt{s})$ is monotonic in γ ,

(ii) *Equation (4.1) with $D = D_F^R$ has a unique solution $\gamma = \gamma_{s,F}^R(\varepsilon)$ for any $\varepsilon \in (0, (1 - P)\sqrt{s})$.*

We next consider Problem 1 in the QED regime, and first introduce some definitions. Let for $\gamma \in (\gamma_P, \sqrt{s})$

$$g_{s,F}(\gamma) := \sqrt{s}D_F(s, s - \gamma\sqrt{s}), \quad g_{s,F}^R(\gamma) := \sqrt{s}D_F^R(s, s - \gamma\sqrt{s}). \quad (4.4)$$

Furthermore, define for $\gamma \in \mathbb{R}$,

$$g_\infty(\gamma) = \frac{\phi(\gamma)}{\Phi(\gamma)}, \quad (4.5)$$

and define $\gamma_{\infty,F}(\varepsilon)$ and $\gamma_{\infty,F}^R(\varepsilon)$ as the solutions of

$$(1 + F(1))g_\infty(\gamma) = \varepsilon \quad \text{and} \quad g_\infty(\gamma) = \varepsilon, \quad (4.6)$$

respectively. It is well-known, see [2, Subsection 4.1] that $g_\infty(\gamma)$ strictly decreases from $+\infty$ at $\gamma = -\infty$ to 0 at $\gamma = \infty$, and so both equations in (4.6) have unique solutions when $\varepsilon > 0$.

In Theorem 7 below, we give a limit result for $\gamma_{s,F}(\varepsilon)$ and $\gamma_{s,F}^R(\varepsilon)$ as $s \rightarrow \infty$ that involves $\gamma_{\infty,F}(\varepsilon)$ and $\gamma_{\infty,F}^R(\varepsilon)$, respectively. For this result, the following observations are made. From (2.8) and (2.10) we get

$$g_{s,F}(\gamma) = \sqrt{s}B(s, s - \gamma\sqrt{s}) \frac{1 + F(1 - \gamma/\sqrt{s})}{1 + B(s, s - \gamma\sqrt{s})F(1 - \gamma/\sqrt{s})}, \quad (4.7)$$

and

$$g_{s,F}^R(\gamma) = \sqrt{s}B(s, s - \gamma\sqrt{s}) \frac{1 - \frac{\gamma}{\sqrt{s}}(1 - \gamma/\sqrt{s})^{-1}F(1 - \gamma/\sqrt{s})}{1 + B(s, s - \gamma\sqrt{s})F(1 - \gamma/\sqrt{s})}. \quad (4.8)$$

Now we have from [9, Theorem 14],

$$\sqrt{s}B(s, s - \gamma\sqrt{s}) = g_\infty(\gamma) + \frac{1}{\sqrt{s}}h_\infty(\gamma) + O(s^{-1}), \quad (4.9)$$

where

$$h_\infty(\gamma) = -\frac{1}{3}(\gamma^3 + (\gamma^2 + 2)g_\infty(\gamma))g_\infty(\gamma), \quad (4.10)$$

and the O in (4.9) holds uniformly in any bounded set of γ 's. Using (4.9), together with

$$F(1 - \frac{\gamma}{\sqrt{s}}) = F(1) - \frac{\gamma}{\sqrt{s}}F'(1) + O(s^{-1}), \quad (4.11)$$

which holds because of assumption (2.5), the following result is established upon computation.

Lemma 6. Under condition (2.5),

$$g_{s,F}(\gamma) = (1 + F(1))g_\infty(\gamma) + \frac{1}{\sqrt{s}}h_{\infty,F}(\gamma) + O(s^{-1}), \quad (4.12)$$

$$g_{s,F}^R(\gamma) = g_\infty(\gamma) + \frac{1}{\sqrt{s}}h_{\infty,F}^R(\gamma) + O(s^{-1}), \quad (4.13)$$

where

$$h_{\infty,F}(\gamma) = (1 + F(1))h_\infty(\gamma) - (\gamma F'(1) + (1 + F(1))F(1))g_\infty(\gamma), \quad (4.14)$$

$$h_{\infty,F}^R(\gamma) = h_\infty(\gamma) - (\gamma + g_\infty(\gamma))g_\infty(\gamma)F(1). \quad (4.15)$$

The O in (4.12) and (4.13) holds uniformly in any bounded set of γ 's.

Theorem 7 (Asymptotic dimensioning without retrials). Under condition (2.5),

$$\gamma_{s,F}(\varepsilon) = \gamma_{\infty,F}(\varepsilon) + \frac{1}{\sqrt{s}}\eta_{\infty,F}(\varepsilon) + O(s^{-1}), \quad (4.16)$$

$$\gamma_{s,F}^R(\varepsilon) = \gamma_\infty^R(\varepsilon) + \frac{1}{\sqrt{s}}\eta_{\infty,F}^R(\varepsilon) + O(s^{-1}), \quad (4.17)$$

with

$$\eta_{\infty,F}(\varepsilon) = -\frac{h_{\infty,F}(\gamma_{\infty,F}(\varepsilon))}{(1 + F(1))g'_\infty(\gamma_{\infty,F}(\varepsilon))}, \quad (4.18)$$

$$\eta_{\infty,F}^R(\varepsilon) = -\frac{h_{\infty,F}^R(\gamma_\infty^R(\varepsilon))}{g'_\infty(\gamma_\infty^R(\varepsilon))}. \quad (4.19)$$

Theorem 7 gives the limits of $\gamma_{s,F}(\varepsilon)$ and $\gamma_{s,F}^R(\varepsilon)$ as $s \rightarrow \infty$, together with first-order corrections $\eta_{\infty,F}(\varepsilon)$ and $\eta_{\infty,F}^R(\varepsilon)$ that are simple functions of the limits $\gamma_{\infty,F}(\varepsilon)$ and $\gamma_\infty^R(\varepsilon)$. It is here instrumental to note that

$$g'_\infty(\gamma) = -g_\infty(\gamma)(\gamma + g_\infty(\gamma)). \quad (4.20)$$

The proof of Theorem 7 will be given in Section 8.

We next discuss the numerical experiments that we conducted to illustrate the analytical results. Remember that the objective of Problem 1 is to determine maximal sustainable load λ such that the rejection probability $D_F^R(s, \lambda)$ is below a threshold $\varepsilon_s := \varepsilon/\sqrt{s}$. In Section 1 we have denoted the true maximal load by λ_{opt} , and we have explained the concepts of square-root staffing and asymptotic dimensioning, in order to obtain accurate estimates of λ_{opt} that are asymptotically sharp in the QED regime.

The conventional square-root staffing rule is to use the QED approximation $\sqrt{s}D_F^R(s, \lambda) \approx D_*(\gamma)$, obtain the solution to $D_*(\gamma) = \varepsilon_s$, say γ_* , and then prescribe the load as $\lambda_* = s - \gamma_*\sqrt{s}$. Theorem 7 allows for refined-square root staffing based on a better QED approximation $\sqrt{s}D_F^R(s, \lambda) \approx D_\bullet(\gamma)$ and the solution γ_\bullet to $D_\bullet(\gamma) = \varepsilon_s$.

For the asymptotic dimensioning sketched above and in Section 1, applied to Problem 1, we identify the following key functions and parameters:

$$D_*(\gamma) = \frac{\phi(\gamma)}{\Phi(\gamma)}, \quad (4.21)$$

$$D_\bullet(\gamma) = D_*(\gamma) + \frac{1}{\sqrt{s}} h_{\infty, F}^R(\gamma), \quad (4.22)$$

$$\gamma_* = \gamma_\infty^R(\varepsilon), \quad (4.23)$$

$$\gamma_\bullet = \gamma_* + \frac{1}{\sqrt{s}} \eta_{\infty, F}^R(\varepsilon), \quad (4.24)$$

and using the square-root rule $\lambda = s - \gamma\sqrt{s}$,

$$\lambda_* = s - \gamma_*\sqrt{s}, \quad (4.25)$$

$$\lambda_\bullet = s - \gamma_\bullet\sqrt{s} = \lambda_* + r_\bullet, \quad (4.26)$$

with

$$r_\bullet = \frac{h_{\infty, F}^R(\gamma_*)}{g'_{\infty}(\gamma_*)}. \quad (4.27)$$

Table 1 presents results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with $s = 100$ servers. Note that $|\lambda_{\text{opt}} - \lambda_\bullet|$ is always less than 0.1, and how the refinements r_\bullet lead to much sharper estimates of the true optimal values.

ε	λ_{opt}	λ_*	λ_\bullet	r_\bullet	$\sqrt{s}D_F^R(s, \lambda_*)$	$\sqrt{s}D_F^R(s, \lambda_\bullet)$
0.010	75.324	72.836	75.409	2.573	0.004	0.010
0.020	77.554	75.504	77.621	2.117	0.011	0.020
0.030	78.996	77.201	79.053	1.852	0.018	0.030
0.040	80.096	78.479	80.146	1.667	0.026	0.041
0.050	80.999	79.519	81.045	1.525	0.034	0.051
0.060	81.774	80.405	81.816	1.411	0.043	0.061
0.070	82.458	81.181	82.497	1.315	0.052	0.071
0.080	83.073	81.876	83.110	1.234	0.061	0.081
0.090	83.636	82.507	83.671	1.164	0.071	0.091
0.100	84.157	83.088	84.190	1.102	0.080	0.101

Table 1: Results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with $D = D_F^R$ and $s = 100$ servers.

Table 2 presents results for the same situation, except with $p_k = p = 0.5$, $k \geq s$. This control policy is thus more lenient, in the sense that when all servers are occupied, 50% (instead of 10%) of the arrivals is admitted. Note that despite the fact that the refinements r_\bullet are larger, and $|\lambda_{\text{opt}} - \lambda_\bullet|$ is always less than 1, the estimates are less accurate. Based on this example, and many other numerical experiments not reported here, we conclude that square-root staffing becomes less accurate for systems with admission control policies that allow more customers to enter. Such systems really benefit from the refined staffing rules.

ε	λ_{opt}	λ_*	λ_\bullet	r_\bullet	$\sqrt{s}D_F^R(s, \lambda_*)$	$\sqrt{s}D_F^R(s, \lambda_\bullet)$
0.010	75.910	72.836	76.298	3.462	0.003	0.011
0.020	78.162	75.504	78.510	3.006	0.009	0.022
0.030	79.619	77.201	79.942	2.741	0.015	0.033
0.040	80.730	78.479	81.035	2.556	0.022	0.043
0.050	81.642	79.519	81.933	2.414	0.029	0.054
0.060	82.425	80.405	82.705	2.300	0.037	0.064
0.070	83.116	81.181	83.386	2.204	0.045	0.074
0.080	83.738	81.876	83.999	2.123	0.053	0.084
0.090	84.307	82.507	84.560	2.053	0.061	0.095
0.100	84.832	83.088	85.078	1.991	0.070	0.105

Table 2: Results for the admission policy $p_k = p = 0.5$, $k \geq s$, and for Problem 1 with $s = 100$ servers.

Let us now turn to the same dimensioning problem, but then for the system *with* retrials.

Problem 2 For fixed s, ε , find λ such that

$$\sqrt{s}D_F^R(s, \lambda + \Omega) = \varepsilon \quad (4.28)$$

with Ω defined as in (2.13).

We only consider the case D_F^R in (4.28), but results for D_F can be obtained in a similar manner. Let us first bring the generalized Cohen equation (2.13) in a form that is amenable for analysis in the QED regime. Write $\Omega = a\sqrt{s}$ so that (2.13) becomes

$$a\sqrt{s} = (s - (\gamma - a)\sqrt{s}) D_F^R(s, s - (\gamma - a)\sqrt{s}). \quad (4.29)$$

With $f_{s,F}^R$ defined as

$$f_{s,F}^R(\gamma) := \left(1 - \frac{\gamma}{\sqrt{s}}\right) g_{s,F}^R(\gamma) = \sqrt{s} \left(1 - \frac{\gamma}{\sqrt{s}}\right) D_F^R(s, s - \gamma\sqrt{s}), \quad (4.30)$$

we can write (4.29) concisely as

$$a = f_{s,F}^R(\gamma - a) \quad (4.31)$$

in which γ is given and a is to be solved.

We thus have $D_F^R(s, \lambda + \Omega) = \varepsilon/\sqrt{s}$ if and only if

$$g_{s,F}^R(\gamma - a_{s,F}(\gamma)) = \varepsilon, \quad (4.32)$$

where $a = a_{s,F}(\gamma)$ solves for $\gamma > 0$ the equation (4.31).

Using $f_{s,F}^R(\delta) = (1 - \delta/\sqrt{s}) g_{s,F}^R(\delta)$, the equation in (4.32) takes the form

$$\frac{a}{1 - (\gamma - a)/\sqrt{s}} = \varepsilon, \quad (4.33)$$

where $a = a_{s,F}(\gamma)$. This gives

$$a = \frac{1 - \gamma/\sqrt{s}}{1 - \varepsilon/\sqrt{s}} \varepsilon, \quad (4.34)$$

and therefore

$$\gamma - a = \frac{\gamma - \varepsilon}{1 - \varepsilon/\sqrt{s}}. \quad (4.35)$$

Hence, we can solve Problem 2 by first solving $\delta = \delta_{s,F}^R(\varepsilon)$ from

$$g_{s,F}^R(\delta) = \varepsilon, \quad (4.36)$$

see (4.32), and then setting $(\gamma - \varepsilon)/(1 - \varepsilon/\sqrt{s}) = \delta$, i.e.,

$$\gamma = \gamma_{s,F}^R(\varepsilon) = \delta + \varepsilon - \delta\varepsilon/\sqrt{s}; \quad \delta = \delta_{s,F}^R(\varepsilon). \quad (4.37)$$

As to (4.36) we operate under the assumption of monotonicity of D_F^R as made in the beginning of this section.

Denote by $\delta = \delta_\infty(\varepsilon)$ the solution of

$$g_\infty(\delta) = \varepsilon. \quad (4.38)$$

Observe that $\delta_\infty(\varepsilon) = \gamma_\infty^R(\varepsilon)$, see (4.6). Then we have by Theorem 7, (4.34) and (4.35) that

$$\delta_{s,F}^R(\varepsilon) = \delta_\infty(\varepsilon) - \frac{1}{\sqrt{s}} \frac{h_{\infty,F}^R(\delta_\infty(\varepsilon))}{g'_{\infty}(\delta_\infty(\varepsilon))} + O(s^{-1}). \quad (4.39)$$

Using this in (4.37), we arrive at the following result.

Theorem 8 (Asymptotic dimensioning with retrials). *Under condition (2.5), and assuming that $D_F^R(s, s - \gamma\sqrt{s})$ is monotonic in γ ,*

$$\gamma_{s,F}^R(\varepsilon) = \varepsilon + \delta_\infty(\varepsilon) + \frac{1}{\sqrt{s}} \theta_{\infty,F}^R(\varepsilon) + O(s^{-1}), \quad (4.40)$$

with

$$\theta_{\infty,F}^R(\varepsilon) = -\delta_\infty(\varepsilon)\varepsilon - \frac{h_{\infty,F}^R(\delta_\infty(\varepsilon))}{g'_{\infty}(\delta_\infty(\varepsilon))}. \quad (4.41)$$

For the asymptotic dimensioning scheme, applied to Problem 1 *with retrials*, we identify the following key functions and parameters:

$$\gamma_* = \varepsilon + \delta_\infty(\varepsilon), \quad (4.42)$$

$$\gamma_\bullet = \gamma_* + \frac{1}{\sqrt{s}} \theta_{\infty,F}^R(\varepsilon), \quad (4.43)$$

and using the square-root rule $\lambda = s - \gamma\sqrt{s}$,

$$\lambda_* = s - \gamma_*\sqrt{s}, \quad (4.44)$$

$$\lambda_\bullet = s - \gamma_\bullet\sqrt{s} = \lambda_* + r_\bullet, \quad (4.45)$$

with

$$r_\bullet = (\gamma_* - \varepsilon)\varepsilon + \frac{h_{\infty,F}^R(\gamma_* - \varepsilon)}{g'_{\infty}(\gamma_* - \varepsilon)}. \quad (4.46)$$

Table 3 presents results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with retrials and $s = 100$ servers. Table 4 displays the results for $p = 0.5$.

ε	λ_{opt}	λ_*	λ_\bullet	r_\bullet	$\sqrt{s}D_F^R(s, \lambda_*)$	$\sqrt{s}D_F^R(s, \lambda_\bullet)$
0.010	75.249	72.736	75.336	2.600	0.004	0.010
0.020	77.399	75.304	77.470	2.166	0.010	0.020
0.030	78.759	76.901	78.822	1.921	0.018	0.031
0.040	79.775	78.079	79.832	1.753	0.025	0.041
0.050	80.594	79.019	80.647	1.628	0.034	0.051
0.060	81.283	79.805	81.333	1.528	0.042	0.061
0.070	81.880	80.481	81.929	1.447	0.051	0.071
0.080	82.409	81.076	82.455	1.379	0.059	0.081
0.090	82.884	81.607	82.929	1.321	0.068	0.091
0.100	83.315	82.088	83.359	1.271	0.077	0.101

Table 3: Results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 2 with retrials and $s = 100$ servers.

ε	λ_{opt}	λ_*	λ_\bullet	r_\bullet	$\sqrt{s}D_F^R(s, \lambda_*)$	$\sqrt{s}D_F^R(s, \lambda_\bullet)$
0.010	75.834	72.736	76.225	3.489	0.003	0.011
0.020	78.006	75.304	78.359	3.055	0.009	0.022
0.030	79.380	76.901	79.711	2.810	0.015	0.033
0.040	80.407	78.079	80.721	2.642	0.021	0.043
0.050	81.234	79.019	81.536	2.516	0.028	0.054
0.060	81.930	79.805	82.222	2.417	0.036	0.064
0.070	82.534	80.481	82.817	2.336	0.043	0.074
0.080	83.068	81.076	83.344	2.268	0.051	0.085
0.090	83.548	81.607	83.817	2.210	0.059	0.095
0.100	83.984	82.088	84.248	2.160	0.067	0.105

Table 4: Results for the admission policy $p_k = p = 0.5$, $k \geq s$, and for Problem 2 with retrials and $s = 100$ servers.

5 Bistability

In Section 4 we have considered dimensioning problems that were formulated directly in terms of the probabilities D_F and D_F^R , and we have seen that these problems admit generically one solution when the prescribed value for D_F or D_F^R is in the appropriate range. The situation is different when dimensioning problems of a more complicated nature are considered. In this section we consider carried traffic quantities $\lambda(1 - D(s, \lambda))$ with $D = D_F$ or D_F^R , without and with retrials, and we will see that the corresponding dimensioning problems may have two solutions. One could refer to this situation as *bistability*, which has also been observed in certain loss networks with alternative two-link routing [7]. The bistability stresses the fact that sample paths of the Markov process tend to be concentrated around two relatively stable points of the system.

Note that $\lambda(1 - D_F(s, \lambda))$ is the rate of arrivals that enter the system in one attempt, and $\lambda(1 - D_F^R(s, \lambda))$ is the rate of arrivals that pass the system, possibly after having to wait in the queue that arises when all servers are occupied.

We first consider the system without retrials.

Problem 3 For fixed s, ε , find λ such that

$$\lambda(1 - D(s, \lambda)) = \varepsilon \quad \text{with} \quad D = D_F \text{ or } D_F^R. \quad (5.1)$$

In Section 6 we prove the following result.

Theorem 9 ((Non-)uniqueness of solutions). *Under condition (2.5),*

(i) *Equation (5.1) with $D = D_F$ has at least two solutions $\lambda \in (0, \lambda_P)$ when $\varepsilon > 0$ is sufficiently small.*

(ii) *Equation (5.1) with $D = D_F^R$ has a unique solution $\lambda \in (0, \lambda_P)$ when $\varepsilon \in (0, s)$.*

Theorem 9(i) is intuitively clear since $\lambda(1 - D_F(s, \lambda))$ is positive for all $\lambda \in (0, \lambda_P)$ while it vanishes when $\lambda \downarrow 0$ or $\lambda \uparrow \lambda_P$. The result of Theorem 9(ii) is less obvious and depends on monotonicity of $\lambda(1 - D_F^R(s, \lambda))$ as a function of $\lambda \in (0, \lambda_P)$. For the case $p_k = p \in (0, 1)$, $k \geq s$, it follows from (6.22) and Proposition 15(iv) that $\lambda(1 - D_F(s, \lambda))$ is strictly concave, and so (5.1) with $D = D_F$ has *exactly* two solutions when $\varepsilon > 0$ is sufficiently small.

We next consider carried traffic quantities in the case of retrials. For convenience, we shall restrict ourselves to the choice $D = D_F^R$ in (5.1).

Problem 4 For fixed s, ε , find λ such that

$$\lambda(1 - D_F^R(s, \lambda + \Omega)) = \varepsilon \quad (5.2)$$

with Ω defined as in (2.13).

Observe that $\lambda(1 - D_F^R(s, \lambda + \Omega)) = \varepsilon$ if and only if

$$(s - \gamma\sqrt{s}) \left(1 - \frac{1}{\sqrt{s}} g_{s,F}^R(\gamma - a_{s,F}(\gamma)) \right) = \varepsilon. \quad (5.3)$$

Using $f_{s,F}^R(\delta) = (1 - \delta/\sqrt{s}) g_{s,F}^R(\delta)$ and (4.31) in (5.3), we can write equation (5.2) as

$$(s - \gamma\sqrt{s}) \left(1 - \frac{1}{\sqrt{s}} \frac{a}{1 - (\gamma - a)/\sqrt{s}} \right) = \sqrt{s} \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a} = \varepsilon, \quad (5.4)$$

where $a = a_{s,F}(\gamma)$. It is now not possible to simply eliminate a , and a study of the function

$$L_{s,F}(\gamma) = \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a_{s,F}(\gamma)} \quad (5.5)$$

is required.

We now give a detailed presentation of what can be achieved analytically for the case that $F \equiv 0$ ($p_k = 0$ for all $k \geq s$). It is a challenging problem to explore in what respect this detailed analytic result can be extended to more general admission policies. We thus consider the carried-traffic problem in (5.3) for the case that $F \equiv 0$ and study the function

$$L_s(\gamma) = L_{s,F}(\gamma) = \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a_s(\gamma)}, \quad 0 < \gamma < \sqrt{s}, \quad (5.6)$$

where $a_s(\gamma)$ is the solution of the Cohen equation $a = f_s(\gamma - a)$ with $f_s = f_{s,F \equiv 0}$. Both f_s and a_s have been studied in great detail in [2]. Using the results of [2] and extensions thereof, the following is shown in Appendix B.

Theorem 10. *There holds*

$$L_s(\gamma) \in (0, \sqrt{s} - \gamma), \quad \gamma \in (0, \sqrt{s}), \quad (5.7)$$

$$L_s(\gamma) = \gamma s(1 + O(\gamma\sqrt{s})), \quad \gamma \downarrow 0, \quad (5.8)$$

$$L_s(\gamma) = (\sqrt{s} - \gamma) \left(1 + O\left(\frac{e^s}{\sqrt{s}} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^s\right) \right), \quad \gamma \uparrow \sqrt{s}. \quad (5.9)$$

Furthermore, $L_s(\gamma)$ is unimodal on $(0, \sqrt{s})$, and the maximum of $L_s(\gamma)$ is assumed at the unique solution $\gamma = \hat{\gamma}_s$ of the equation

$$\gamma a_s(\gamma) = \frac{1}{2} \left(1 - \frac{\gamma}{\sqrt{s}} \right), \quad (5.10)$$

and

$$L_s(\hat{\gamma}_s) = \frac{\hat{\gamma}_s}{\hat{\gamma}_s + \frac{1}{2\sqrt{s}}} (\sqrt{s} - \hat{\gamma}_s). \quad (5.11)$$

Finally,

$$\hat{\gamma}_s > f_s(0) = \gamma_s^* > \left(\frac{1}{2} + \frac{1}{16s} \right)^{1/2} - \frac{1}{4\sqrt{s}}, \quad s \geq 2, \quad (5.12)$$

and $\hat{\gamma}_s$ increases in s from $1/2$ at $s = 1$ to 1.034113461 at $s = \infty$, with $\hat{\gamma}_s \approx 1$ when $s = 550$.

From the detailed information provided by Theorem 10, it is seen that the equation

$$\frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a} = \frac{\varepsilon}{\sqrt{s}}, \quad (5.13)$$

see (5.4), has two, one or zero solutions according as $\varepsilon <, =$ or $> \sqrt{s} L_s(\hat{\gamma}_s)$ with $L_s(\hat{\gamma}_s)$ given in (5.11). Furthermore, it is seen that $\sqrt{s} L_s(\hat{\gamma}_s) = s + O(\sqrt{s})$, and that the highest carried-traffic numbers occur in a γ -region somewhat away from, but relatively close to, $\gamma = 0$.

Theorem 11 (Bistability). *Problem 4 has two, one or zero solutions according as $\varepsilon <, =$ or $> \sqrt{s} L_s(\hat{\gamma}_s)$ with $L_s(\hat{\gamma}_s)$ given in (5.11).*

In Figure 1 we display $\frac{1}{\sqrt{s}} L_s(\delta\sqrt{s})$, $0 < \delta < 1$, for $s = 1, 5, 10, 50, 100$. It is observed that the graphs approximate the graph of the function $1 - \delta$, $0 < \delta < 1$, when s gets large.

6 Proof of results on dimensioning

In this section we drop condition (2.5) so that both $P = 1$ and $F(\lambda_P - 0) < \infty$ are allowed.

The equations in (4.1) and in (5.1) take the form

$$g(\gamma) = \varepsilon, \quad g = g_{s,F} \text{ or } g_{s,F}^R, \quad (6.1)$$

and

$$\gamma + f(\gamma) = \frac{s - \varepsilon}{\sqrt{s}}, \quad f = f_{s,F} \text{ or } f_{s,F}^R, \quad (6.2)$$

respectively, where $g_{s,F}^R$ and $f_{s,F}^R$ as in (4.4) and (4.30), and we define for $\gamma \in (\gamma_P, \sqrt{s}]$,

$$f_{s,F}(\gamma) := \left(1 - \frac{\gamma}{\sqrt{s}} \right) g_{s,F}(\gamma) = \sqrt{s} \left(1 - \frac{\gamma}{\sqrt{s}} \right) D_F(s, s - \gamma\sqrt{s}). \quad (6.3)$$

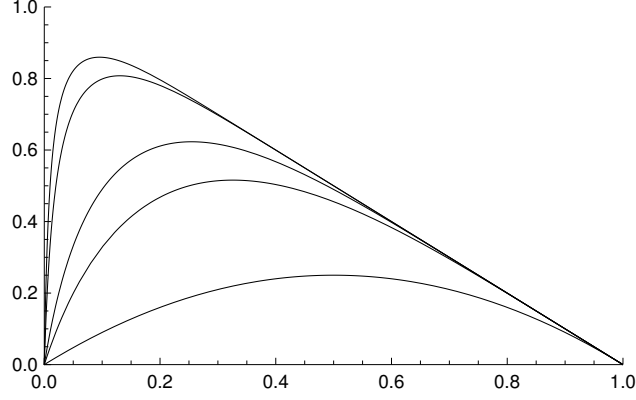


Figure 1: Plot of $\frac{1}{\sqrt{s}} L_s(\delta\sqrt{s})$, $0 < \delta < 1$, for the cases $s = 1, 5, 10, 50, 100$, with maximizer $\hat{\gamma}_s/\sqrt{s}$ decreasing with s and maximum value $\frac{1}{\sqrt{s}} L_s(\hat{\gamma}_s)$ increasing with s .

To get insight and pertinent results about when the equations in (6.1) and (6.2) have (unique) solutions and how to find these solutions, we relate the functions $f_{s,F}$ and $f_{s,F}^R$ to the functions f_s and g_s given by

$$f_s(\gamma) = \left(1 - \frac{\gamma}{\sqrt{s}}\right) g_s(\gamma) = \sqrt{s} \left(1 - \frac{\gamma}{\sqrt{s}}\right) B(s, s - \gamma\sqrt{s}), \quad \gamma \leq \sqrt{s}. \quad (6.4)$$

The latter functions have been studied in considerable detail in [2], in particular with respect to monotonicity properties. Some of these properties are collected in the beginning of Appendix A.

Lemma 12. For $\gamma \in (\gamma_P, \sqrt{s}]$,

$$f_{s,F}(\gamma) = f_s(\gamma) \frac{1 + \left(1 - \frac{\gamma}{\sqrt{s}}\right) H_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)}, \quad (6.5)$$

and

$$f_{s,F}^R(\gamma) = f_s(\gamma) \frac{1 - \frac{\gamma}{\sqrt{s}} H_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)}, \quad (6.6)$$

with

$$H_s(\gamma) = \sum_{n=0}^{\infty} p_s \cdots p_{s+n} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n. \quad (6.7)$$

Proof. This follows in a straightforward manner from (2.8) and (2.10) and the definition of H_s , where we note that $F(1 - \frac{\gamma}{\sqrt{s}}) = (1 - \frac{\gamma}{\sqrt{s}}) H_s(\gamma)$. \square

Lemma 12 shows that $f_{s,F}$ and $f_{s,F}^R$ factorize into an admission policy independent part $f_s(\gamma)$ and a part comprising the admission policy via H_s . By dividing either side of (6.5) and (6.6) by

$(1 - \gamma/\sqrt{s})$, it is seen that a similar result as Lemma 12 holds for $g_{s,F}$ and $g_{s,F}^R$, with the same factors comprising the admission policy as in (6.5) and (6.6).

The following result gives a global picture for $f_{s,F}$ and $f_{s,F}^R$ in terms of inequalities. We observe that $F(\lambda_P - 0) = \infty \Leftrightarrow H_s(\gamma_P + 0) = \infty$.

Proposition 13. (i) For $\gamma_P < \gamma < \sqrt{s}$,

$$\max\{0, -\gamma\} \leq f_{s,F}^R(\gamma) \leq f_s(\gamma) \leq f_{s,F}(\gamma) \leq \min\left\{\sqrt{s} - \gamma, \frac{\sqrt{s} f_s(\gamma)}{\gamma + f_s(\gamma)}\right\}. \quad (6.8)$$

(ii) There is equality in the first inequality in (6.8) if and only if $\gamma \in (0, \sqrt{s})$ and $p_k = 1$ for all $k \geq s$, and in that case

$$H(\gamma) = \frac{\sqrt{s}}{\gamma}, \quad \gamma_P = 0 < \gamma < \sqrt{s}. \quad (6.9)$$

(iii) There is equality in the second inequality in (6.8) for any $\gamma \in (\gamma_P, \sqrt{s})$ if and only if $p_k = 0$ for all $k \geq s$. There is equality in the third inequality in (6.8) for any $\gamma \in (\gamma_P, \sqrt{s})$ if and only if $p_k = 0$ for all $k \geq s$.

(iv) There is equality in the fourth inequality in (6.8) if and only if $\gamma \in (0, \sqrt{s})$ and $p_k = 1$ for all $k \geq s$.

(v) For $\gamma = \sqrt{s}$,

$$f_{s,F}^R(\gamma) = f_s(\gamma) = f_{s,F}(\gamma) = 0. \quad (6.10)$$

(vi) $f_{s,F}^R(\gamma_P + 0) = -\gamma_P$ if and only if $H_s(\gamma_P + 0) = \infty$.

(vii) $f_{s,F}(\gamma_P + 0) = \sqrt{s} - \gamma_P$ if and only if $H_s(\gamma_P + 0) = \infty$.

The proof of Proposition 13 is given in Appendix A and uses the representations (6.5) and (6.6).

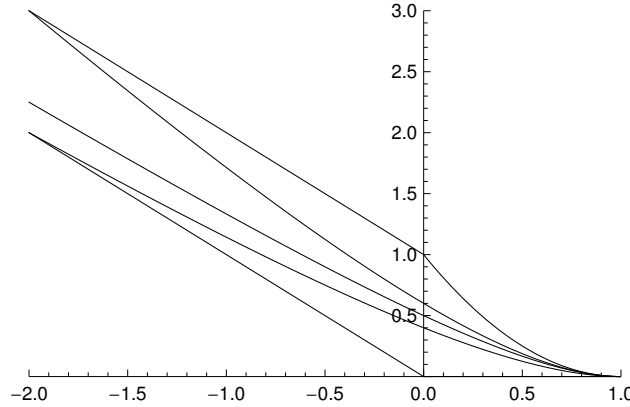


Figure 2: Picture illustrating $f_{s,F}^R$, $f_{s,F}$ and f_s with bounds as in Proposition 13 for the case that $0 < P < 1$ and $H_s(\gamma_P + 0) = \infty$ (actual choice: $s = 1$, $p_k = p = 1/3$, $k \geq s$)

In Figure 2 we show $f_{s,F}^R$, f_s and $f_{s,F}$, together with the bounds in (6.8) for the case that $-\infty < \gamma_P < 0$ and $H(\gamma_P + 0) = \infty$ (actual choice: $s = 1$ and $p_k = p = 1/3$, $k \geq s$).

In general, monotonicity properties for the functions $f_{s,F}$, $g_{s,F}^R$ and of the functions $\gamma + f_{s,F}(\gamma)$ and $\gamma + f_{s,F}^R(\gamma)$, see (6.1) and (6.2) are not easy to establish or manifestly not true. We now present some results, positive and negative, on monotonicity and convexity of the functions in (6.3) and (4.30).

Proposition 14. (i) $f_{s,F}$ and $g_{s,F}$ are strictly decreasing.

(ii) $\gamma + f_{s,F}^R(\gamma)$, $\gamma_P < \gamma \leq \sqrt{s}$, is strictly increasing.

The proof of Proposition 14 is given in Appendix A. An interesting example is obtained when we choose

$$p_k = p \in (0, 1), \quad k \geq s. \quad (6.11)$$

Then $P = p$, $\gamma_P = -\frac{1-p}{p} \sqrt{s} \in (-\infty, 0)$, and

$$H_s(\gamma) = \sum_{n=0}^{\infty} p^{n+1} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n = \frac{p}{1 - p(1 - \gamma/\sqrt{s})}, \quad \gamma_P < \gamma \leq \sqrt{s}. \quad (6.12)$$

Furthermore, $H_s(\gamma_P + 0) = \infty$, and we compute from (6.5), (6.6) and (6.12)

$$f_{s,F}(\gamma) = \frac{f_s(\gamma)}{1 - p + \frac{p}{\sqrt{s}}(\gamma + f_s(\gamma))}, \quad f_{s,F}^R(\gamma) = \frac{(1-p)f_s(\gamma)}{1 - p + \frac{p}{\sqrt{s}}(\gamma + f_s(\gamma))}. \quad (6.13)$$

There is for this case the following result.

Proposition 15. Assume $p_k = p \in (0, 1)$, $k \geq s$, so that $f_{s,F}$ and $f_{s,F}^R$ are given by (6.13), and let $g_{s,F}$ and $g_{s,F}^R$ be the functions associated to $f_{s,F}$ and $f_{s,F}^R$ according to (6.3) and (4.30), respectively. Then,

(i) $f_{s,F}$ and $f_{s,F}^R$ are strictly decreasing and convex in $\gamma_P < \gamma \leq \sqrt{s}$, with

$$f_{s,F}(\gamma_P + 0) = \sqrt{s} - \gamma_P = \frac{\sqrt{s}}{p}, \quad f_{s,F}^R(\gamma_P + 0) = -\gamma_P = \frac{1-p}{p} \sqrt{s}, \quad (6.14)$$

and $f_{s,F}(\sqrt{s}) = f_{s,F}^R(\sqrt{s}) = 0$.

(ii) $g_{s,F}$ and $g_{s,F}^R$ are strictly decreasing in $\gamma_P < \gamma \leq \sqrt{s}$, with

$$g_{s,F}(\gamma_P + 0) = \sqrt{s}, \quad g_{s,F}^R(\gamma_P + 0) = (1-p)\sqrt{s}, \quad (6.15)$$

and $g_{s,F}(\sqrt{s}) = g_{s,F}^R(\sqrt{s}) = 0$.

(iii) $\gamma + f_{s,F}^R(\gamma)$ is strictly increasing and convex in $\gamma_P < \gamma \leq \sqrt{s}$, with

$$\gamma + f_{s,F}^R(\gamma) \Big|_{\gamma=\gamma_P+0} = 0, \quad \gamma + f_{s,F}^R(\gamma) \Big|_{\gamma=\sqrt{s}} = \sqrt{s}. \quad (6.16)$$

(iv) $\gamma + f_{s,F}(\gamma)$ is non-monotonic and strictly convex in $\gamma_P < \gamma \leq \sqrt{s}$, with

$$\gamma + f_{s,F}(\gamma) \Big|_{\gamma=\gamma_P+0} = \gamma + f_{s,F}(\gamma) \Big|_{\gamma=\sqrt{s}} = \sqrt{s}. \quad (6.17)$$

The proof of Proposition 15 is given in Appendix A.

We finally have the following result.

Proposition 16. *Assume that p_k is chosen as in (4.3). Then $g_{s,F}^R(\gamma)$ is strictly decreasing in $\gamma \in (\gamma_P, \sqrt{s})$.*

The proof of Proposition 16 is given in the Appendix A.

We now show how the various results given in this section can be used to prove Theorem 5 and Theorem 9.

6.1 Proof of Theorem 5

Under condition (2.5) we have $F(\lambda_P - 0) = H_s(\gamma_P + 0) = \infty$. It follows from (6.3) and (6.5) that

$$g_{s,F}(\gamma_P + 0) = \frac{f_{s,F}(\gamma_P + 0)}{1 - \gamma_P/\sqrt{s}} = \frac{f_s(\gamma_P)}{1 - \gamma_P/\sqrt{s}} \frac{1 - \gamma_P/\sqrt{s}}{f_s(\gamma_P)/\sqrt{s}} = \sqrt{s}. \quad (6.18)$$

In a similar fashion it follows from (4.4), (4.30) and (6.6), using $\gamma_P = -(1 - P)P^{-1}\sqrt{s}$, that

$$g_{s,F}^R(\gamma_P + 0) = (1 - P)\sqrt{s}. \quad (6.19)$$

Furthermore,

$$D_F(s, \lambda) = \sqrt{s}g_{s,F}(\gamma), \quad D_F^R(s, \lambda) = \sqrt{s}g_{s,F}^R(\gamma) \quad (6.20)$$

when $\lambda = s - \gamma\sqrt{s}$ while $g_{s,F}(\sqrt{s}) = g_{s,F}^R(\sqrt{s}) = 0$ as $B(s, 0) = 0$. Then Theorem 5(i) follows from Proposition 14(i) while Theorem 5(ii) holds by the monotonicity assumption made in the discussion preceding Theorem 5.

6.2 Proof of Theorem 9

The function $\lambda(1 - D_F(s, \lambda))$ depends continuously on $\lambda \in (0, \lambda_P)$, is positive for $\lambda \in (0, \lambda_P)$, and satisfies

$$\lim_{\lambda \downarrow 0} \lambda(1 - D_F(s, \lambda)) = 0 = \lim_{\lambda \uparrow \lambda_P} \lambda(1 - D_F(s, \lambda)), \quad (6.21)$$

see (6.18) and (6.20). This yields assertion (i).

A computation, using (4.30) and (6.3), shows that

$$\lambda(1 - D_F(s, \lambda)) = s - \sqrt{s}(\gamma + f_{s,F}(\gamma)), \quad (6.22)$$

when $\lambda = s - \gamma\sqrt{s}$. Hence, by Proposition 14(ii), we have that $\lambda(1 - D_F^R(s, \lambda))$ is strictly increasing in $\lambda \in (0, \lambda_P)$. Furthermore,

$$\lim_{\lambda \downarrow 0} \lambda(1 - D_F^R(s, \lambda)) = 0, \quad \lim_{\lambda \uparrow \lambda_P} \lambda(1 - D_F^R(s, \lambda)) = s \quad (6.23)$$

by (6.18) and (6.20). This yields assertion (ii).

Let us conclude this section with the following two observations. Concavity of $\lambda(1 - D(s, \lambda))$ when $p_k = p \in (0, 1)$, $k \geq s$, follows from (6.22) and Proposition 15(ii). Monotonicity of $D_F^R(s, \lambda)$ when p_k is given by (6.3) follows from (6.20) and Proposition 16.

7 Proof of Theorem 4

We prove Theorem 4 under condition (2.5), where we assume $P \in (0, 1)$ (the case $P = 0$ requires only minor modification of the analysis below). By Theorem 1, the generalized Cohen equation (2.13), written in QED coordinates as in (4.31), has a unique solution $a_{s,F}(\gamma)$ for any $\gamma \in (0, \sqrt{s})$.

Theorem 17. *For any $\gamma > 0$,*

$$a_\infty(\gamma) \geq a_{s,F}(\gamma) = a_\infty(\gamma) + O\left(\frac{1}{\sqrt{s}}\right), \quad s \rightarrow \infty, \quad (7.1)$$

where the O holds uniformly in any compact set of $\gamma \in (0, \infty)$, and where $a_\infty(\gamma)$ is the unique solution of

$$a = f_\infty(\gamma - a) = \frac{\varphi(\gamma - a)}{\Phi(\gamma - a)}, \quad (7.2)$$

see [2, Section 3.3].

Proof. We have $f_{s,F}^R(\delta) \leq f_s(\delta) < f_\infty(\delta)$ and $f_{s,F}^R(\delta) = f_\infty(\delta) + O(\frac{1}{\sqrt{s}})$ uniformly in any compact set of $\delta \in \mathbb{R}$, see (6.6), Proposition 13(i) and [2, Propositions 5 and 6]. Let $\gamma > 0$. We prove below that $a_{s,F}(\gamma) \leq a_\infty(\gamma)$. Therefore, there is an $M > 0$ such that

$$f_\infty(\gamma - a) - \frac{M}{\sqrt{s}} \leq f_{s,F}^R(\gamma - a) \leq f_\infty(\gamma - a) \quad (7.3)$$

holds for $a \in [0, a_\infty(\gamma)]$ and all $s \geq 1$. Since $a + f_\infty(\gamma - a)$ is convex in a and $f'_\infty(\delta) > -1$, $\delta \in \mathbb{R}$, see [2, (21)], it follows from the mean value theorem that there is an $\eta > 0$ such that

$$f_\infty(\gamma - a) - a \geq \eta(a_\infty(\gamma) - a), \quad a \in [0, a_\infty(\gamma)], \quad (7.4)$$

where it is observed that $f_\infty(\gamma - a) - a = 0$ at $a = a_\infty(\gamma)$. Then

$$\begin{aligned} 0 = f_{s,F}^R(\gamma - a_{s,F}(\gamma)) - a_{s,F}(\gamma) &\geq f_\infty(\gamma - a_{s,F}(\gamma)) - a_{s,F}(\gamma) - \frac{M}{\sqrt{s}} \\ &\geq \eta(a_\infty(\gamma) - a_{s,F}(\gamma)) - \frac{M}{\sqrt{s}}, \end{aligned} \quad (7.5)$$

and so

$$a_{s,F}(\gamma) \geq a_\infty(\gamma) - \frac{M}{\eta\sqrt{s}}. \quad (7.6)$$

In the above argument, M and η can be chosen independently of γ in any compact subset of $(0, \infty)$, and so the result follows.

We still have to show that $a_{s,F}(\gamma) \leq a_\infty(\gamma)$. We have from $f_{s,F}^R(\delta) \leq f_s(\delta)$ that

$$f_\infty(\gamma - a_{s,F}(\gamma)) \geq f_{s,F}^R(\gamma - a_{s,F}(\gamma)) = a_{s,F}(\gamma). \quad (7.7)$$

Therefore,

$$\gamma - a_\infty(\gamma) + f_\infty(\gamma - a_\infty(\gamma)) = \gamma \geq \gamma - a_{s,F}(\gamma) + f_\infty(\gamma - a_{s,F}(\gamma)). \quad (7.8)$$

Now $\delta + f_\infty(\delta)$ is increasing, see [2, (21)], and so it follows from (7.8) that $\gamma - a_\infty(\gamma) \geq \gamma - a_{s,F}(\gamma)$, i.e., $a_{s,F}(\gamma) \leq a_\infty(\gamma)$. \square

The limiting behavior of $D_F(s, \lambda + \Omega)$ and $D_F^R(s, \lambda + \Omega)$ is now easily found. We have, see (6.3) and (4.12), for $\gamma > 0$

$$\begin{aligned} \frac{1}{\sqrt{s}} D_F^{-1}(s, s - (\gamma - a_{s,F}(\gamma))) &= \frac{1 - (\gamma - a_{s,F}(\gamma))/\sqrt{s}}{f_{s,F}(\gamma - a_{s,F}(\gamma))} \\ \rightarrow \left(1 + H(0) \frac{\phi(\gamma - a_\infty(\gamma))}{\Phi(\gamma - a_\infty(\gamma))}\right)^{-1} &= ((1 + H(0)) a_\infty(\gamma))^{-1}, \quad s \rightarrow \infty. \end{aligned} \quad (7.9)$$

Similarly,

$$\begin{aligned} \frac{1}{\sqrt{s}} D_F^{-R}(s, s - (\gamma - a_{s,F}(\gamma))) &= \frac{1 - (\gamma - a_{s,F}(\gamma))/\sqrt{s}}{f_{s,F}^R(\gamma - a_{s,F}(\gamma))} \\ &= \frac{1}{a_{s,F}(\gamma)} (1 - (\gamma - a_{s,F}(\gamma))/\sqrt{s}) \rightarrow a_\infty^{-1}(\gamma), \quad s \rightarrow \infty, \end{aligned} \quad (7.10)$$

where it also has been used that $a = a_{s,F}(\gamma)$ satisfies (4.31).

8 Proof of Theorem 7

We give the proof of (4.16), (4.18) in detail; the proof of (4.17) and (4.19) being quite similar. Take any $\gamma_1, \gamma_2 \in \mathbb{R}$ with

$$\gamma_1 < \gamma_{\infty,F}(\varepsilon) < \gamma_2. \quad (8.1)$$

From strict decreasingness of $g_\infty(\gamma)$ and the definition of $\gamma_{\infty,F}(\varepsilon)$, we have

$$(1 + F(1))g_\infty(\gamma_2) < (1 + F(1))g_\infty(\gamma_{\infty,F}(\varepsilon)) = \varepsilon < (1 + F(1))g_\infty(\gamma_1). \quad (8.2)$$

Because $g_{s,F}(\gamma) \rightarrow (1 + F(1))g_\infty(\gamma)$ as $s \rightarrow \infty$, we have from (8.2) that

$$g_{s,F}(\gamma_2) < \varepsilon < g_{s,F}(\gamma_1) \quad (8.3)$$

when s is large. By monotonicity of $g_{s,F}$, this implies that $\gamma_{s,F}(\varepsilon) \in [\gamma_1, \gamma_2]$ for large s . Since γ_1 and γ_2 in (8.1) are arbitrary, it follows that $\gamma_{s,F}(\varepsilon) \rightarrow \gamma_{\infty,F}(\varepsilon)$ as $s \rightarrow \infty$. In particular, $\gamma_{s,F}(\varepsilon)$ is bounded in $s \geq 1$.

We next show from the weakened form

$$g_{s,F}(\gamma) = (1 + F(1))g_\infty(\gamma) + O(s^{-1/2}) \quad (8.4)$$

of (4.12) that

$$\gamma_{s,F}(\varepsilon) = \gamma_{\infty,F}(\varepsilon) + O(s^{-1/2}). \quad (8.5)$$

To that end, we consider (8.4) with $\gamma = \gamma_{s,F}(\varepsilon)$ and write

$$g_\infty(\gamma_{s,F}(\varepsilon)) = g_\infty(\gamma_{\infty,F}(\varepsilon)) + (\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_\infty(\gamma_{\infty,F}(\varepsilon)) + O((\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))^2). \quad (8.6)$$

Using this, together with

$$g_{s,F}(\gamma_{s,F}(\varepsilon)) = \varepsilon = (1 + F(1))g_\infty(\gamma_{\infty,F}(\varepsilon)), \quad (8.7)$$

in (8.4) we get

$$\varepsilon = \varepsilon + (1 + F(1))(\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_{\infty}(\gamma_{\infty,F}(\varepsilon)) + O((\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))^2) + O(s^{-1/2}). \quad (8.8)$$

It is known, see [2, (21)], that $g'_{\infty}(\gamma)$ is negative and bounded away from 0 when γ is in a bounded set. From (8.8) we therefore get (8.5).

We finally show (4.16), and for this we repeat the argument for showing (8.5), but now using the full strength of (4.12) with $\gamma = \gamma_{s,F}(\varepsilon)$. Using (8.5) in (8.6) yields

$$g_{\infty}(\gamma_{s,F}(\varepsilon)) = g_{\infty}(\gamma_{\infty,F}(\varepsilon)) + (\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_{\infty}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1}). \quad (8.9)$$

Furthermore, again by (8.5),

$$h_{\infty,F}(\gamma_{s,F}(\varepsilon)) = h_{\infty,F}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1/2}). \quad (8.10)$$

When we use this in (4.12) with $\gamma = \gamma_{s,F}(\varepsilon)$ together with (8.7), we obtain

$$\varepsilon = \varepsilon + (1 + F(1))(\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_{\infty}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1}) + \frac{1}{\sqrt{s}}h_{\infty,F}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1}). \quad (8.11)$$

From this (4.16) and (4.18) follow at once.

References

- [1] J.R. Artalejo and A. Gómez-Corral. *Retrial Queueing Systems*. Springer-Verlag, 2008.
- [2] F. Avram, A.J.E.M. Janssen, and J.S.H. van Leeuwen. Loss systems with slow retrials in the Halfin-Whitt regime. *Adv. in Appl. Probab.*, 45:274-294, 2013.
- [3] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Oper. Res.*, 52:17-34, 2004.
- [4] J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18:49-100, 1957.
- [5] G.I. Falin and J.R. Artalejo. Approximations for multiserver queues with balking/retrial discipline. *OR Spektrum*, 17:239-244, 1995.
- [6] G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman & Hall, London, 1997.
- [7] R.J. Gibbens, P.J. Hunt, and F.P. Kelly. Bistability in communication networks. In *Disorder in Physical Systems: a Volume in Honour of John M. Hammersley*, pages 113-127. Oxford University Press, 1990.
- [8] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29:567-588, 1981.
- [9] D. Jagerman. Some properties of the Erlang loss function. *Bell System Technical Journal*, 53:525-551, 1974.

- [10] A.J.E.M. Janssen, J.S.H. van Leeuwen, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. in Appl. Probab.*, 40:122-143, 2008.
- [11] A.J.E.M. Janssen, J.S.H. van Leeuwen, and B. Zwart. Refining square root safety staffing by expanding Erlang C. *Operations Research*, 59:1512–1522, 2011.
- [12] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.
- [13] M. Nesenbergs. A hybrid of erlang B and C formulas and its applications. *IEEE Transactions on Communications*, 27:59–68, 1979.
- [14] B. Zhang, J.S.H. van Leeuwen, and B. Zwart. Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Oper. Res.*, 60:461-474, 2012.

A Remaining proofs, except proof of Theorem 10

We start by recalling some basic properties, shown in [2], of the functions f_s and g_s given in (6.4). We have

- a. $f_s(\gamma)$ is strictly convex and decreases strictly in $-\infty < \gamma \leq \sqrt{s}$ from $+\infty$ to 0, and $\gamma + f_s(\gamma)$ increases strictly in $-\infty < \gamma \leq \sqrt{s}$ from 0 to \sqrt{s} . Furthermore, $\gamma + f_s(\gamma) = O(\frac{1}{\gamma})$, $\gamma \rightarrow -\infty$.
- b. $g_s(\gamma)$ decreases strictly in $-\infty < \gamma \leq \sqrt{s}$ from \sqrt{s} to 0.
- c. $f_s(\gamma)$ and $g_s(\gamma)$ increase in $s \geq 1$ to $f_\infty(\gamma) = g_\infty(\gamma) = \varphi(\gamma)/\Phi(\gamma)$ uniformly in any compact set of $\gamma \in \mathbb{R}$.

Proof of Proposition 13. We have for $0 < \gamma < \sqrt{s}$ that

$$H_s(\gamma) \leq \sum_{n=0}^{\infty} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n = \frac{\sqrt{s}}{\gamma}, \quad (\text{A.1})$$

with equality if and only if $p_{s+n} = 1$, $n = 0, 1, \dots$. Hence, $1 - \gamma H_s(\gamma)/\sqrt{s} \geq 0$ for $0 < \gamma < \sqrt{s}$, and evidently $1 - \gamma H_s(\gamma)/\sqrt{s} \geq 0$ for $\gamma_P < \gamma \leq 0$. Therefore, from (6.6), $f_{s,F}^R(\gamma) \geq 0$, $\gamma_P < \gamma < \sqrt{s}$, with equality for any γ if and only if $p_{s+n} = 1$, $n = 0, 1, \dots$. Next, we write (6.6) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}^R(\gamma) = -\gamma + \frac{\gamma + f_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)}, \quad (\text{A.2})$$

and then it follows from $\gamma + f_s(\gamma) > 0$ that $f_{s,F}^R(\gamma) > -\gamma$. This proves the first inequality in (6.8).

Next, we write (6.6) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}^R(\gamma) = f_s(\gamma) - \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma) \frac{\gamma + f_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)}, \quad (\text{A.3})$$

and it follows from $\gamma + f_s(\gamma) > 0$ that $f_{s,F}^R(\gamma) \leq f_s(\gamma)$, with equality if and only if $H_s(\gamma) = 0$ if and only if $p_{s+n} = 0$, $n = 0, 1, \dots$. This proves the second inequality in (6.8).

Next, we write (6.5) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}(\gamma) = f_s(\gamma) + f_s(\gamma) H_s(\gamma) \frac{1 - (\gamma + f_s(\gamma))/\sqrt{s}}{1 + f_s(\gamma) H_s(\gamma)/\sqrt{s}}, \quad (\text{A.4})$$

and it follows from $\gamma + f_s(\gamma) < \sqrt{s}$ that $f_{s,F}(\gamma) \geq f_s(\gamma)$, with equality if and only if $H_s(\gamma) = 0$ if and only if $p_{s+n} = 0$, $n = 0, 1, \dots$. This proves the third inequality in (6.8).

Next, we write (6.5) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}(\gamma) = \sqrt{s} - \gamma - \sqrt{s} \frac{\sqrt{s} - \gamma - f_s(\gamma)}{\sqrt{s} + f_s(\gamma) H_s(\gamma)}, \quad (\text{A.5})$$

and it follows from $\gamma + f_s(\gamma) < \sqrt{s}$ that $f_{s,F}(\gamma) < \sqrt{s} - \gamma$. Since $\sqrt{s} - \gamma > f_s(\gamma)$ for $\gamma < \sqrt{s}$, we have that the function $x \geq 0 \mapsto (1 + x(1 - \gamma/\sqrt{s}))/ (1 + x f_s(\gamma)/\sqrt{s})$ is strictly increasing. From (6.5) and (A.1) it then follows (with $x = H_s(\gamma) \leq \sqrt{s}/\gamma$) that $f_{s,F}(\gamma) \leq \sqrt{s} f_s(\gamma)/(\gamma + f_s(\gamma))$ for $0 < \gamma < \sqrt{s}$, with equality if and only if $p_{s+n} = 0$, $n = 0, 1, \dots$. Furthermore, for $\gamma \leq 0$, we have

$$\sqrt{s} - \gamma - \frac{\sqrt{s} f_s(\gamma)}{\gamma + f_s(\gamma)} = \frac{\gamma(\sqrt{s} - \gamma - f_s(\gamma))}{\gamma + f_s(\gamma)} \leq 0, \quad (\text{A.6})$$

with equality if and only if $\gamma = 0$. This proves the fourth inequality in (6.8).

The cases of equality in the inequalities in (6.8) have been indicated already along with their proofs, and this settles Proposition 13, (ii)–(iv).

Proposition 13(v) follows from 1(i) and the fact that $f_s(\sqrt{s}) = 0$.

Proposition 13(vi)–(vii) follow from the representations (6.6) and (6.5) and the fact that $H_s(\gamma)$ increases to $H_s(\gamma_P + 0)$ as γ decreases to γ_P by non-negativity of all p_k . This completes the proof of Proposition 13.

Note. We have the following consequences of (A.3), (A.4) and $\gamma + f_s(\gamma) = O(\frac{1}{\gamma})$, $\gamma \rightarrow -\infty$.

- a. $f_{s,F}^R(\gamma) = f_s(\gamma) + O(\frac{1}{\gamma})$, $\gamma_P < \gamma < 0$,
- b. $f_{s,F}(\gamma) = f_s(\gamma) + \sqrt{s} + O(\frac{1}{\gamma})$, $\gamma_P < \gamma < 0$, when $F \neq 0$,
- c. $g_{s,F}^R(\gamma), g_{s,F}(\gamma) = \sqrt{s} + O(\frac{1}{\gamma})$, $\gamma_P < \gamma < 0$.

These results are in particular relevant when $P = 0$ so that $\gamma_P = -\infty$.

Proof of Proposition 14. (i) Since $f_{s,F}(\gamma) = (1 - \gamma/\sqrt{s}) g_{s,F}(\gamma)$, it is sufficient to show that $g_{s,F}(\gamma)$ is strictly decreasing. We have from (A.5) that

$$g_{s,F}(\gamma) = \frac{f_{s,F}(\gamma)}{1 - \gamma/\sqrt{s}} = \sqrt{s} \left(1 - \frac{1 - \frac{1}{\sqrt{s}} g_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)} \right). \quad (\text{A.7})$$

Now $g_s(\gamma)$ strictly decreases in $-\infty < \gamma \leq \sqrt{s}$ from \sqrt{s} to 0, and $f_s(\gamma)H_s(\gamma)$ is non-negative and decreasing in $\gamma_P < \gamma \leq \sqrt{s}$. It follows that

$$\frac{1 - \frac{1}{\sqrt{s}}g_s(\gamma)}{1 + \frac{1}{\sqrt{s}}f_s(\gamma)H_s(\gamma)} \quad (\text{A.8})$$

is non-negative and strictly increasing in $\gamma_P < \gamma \leq \sqrt{s}$, and the proof is complete.

(ii) We have that $\gamma + f_s(\gamma)$ is positive and strictly increasing in $-\infty < \gamma \leq \sqrt{s}$, and $f_s(\gamma)H_s(\gamma)$ is non-negative and decreasing in $\gamma_P < \gamma \leq \sqrt{s}$. It follows that

$$\frac{\gamma + f_s(\gamma)}{1 + \frac{1}{\sqrt{s}}f_s(\gamma)H_s(\gamma)} \quad (\text{A.9})$$

is strictly increasing in $\gamma_P < \gamma \leq \sqrt{s}$. Then it follows from the representation (A.2) of $f_{s,F}^R$ that $\gamma + f_{s,F}^R(\gamma)$ is strictly increasing in $\gamma_P < \gamma \leq \sqrt{s}$.

Proof of Proposition 15. With $p_k = p \in (0, 1)$ for $k \geq s$, we have $f_{s,F}$ and $f_{s,F}^R$ as in (6.13) so that, in particular, $f_{s,F}^R = (1 - p)f_{s,F}$.

(i) We have that $\gamma + f_s(\gamma)$ is positive and strictly increasing in $-\infty < \gamma \leq \sqrt{s}$ and that $f_s(\gamma)$ is positive and strictly decreasing in $-\infty < \gamma \leq \sqrt{s}$. Hence, from (6.13), both $f_{s,F}(\gamma)$ and $f_{s,F}^R(\gamma)$ are strictly decreasing in $\gamma_P < \gamma \leq \sqrt{s}$. We next show (strict) convexity of $f_{s,F}^R$. We write (6.13) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}^R(\gamma) = \frac{f_s(\gamma)}{1 + b(\gamma + f_s(\gamma))}; \quad b = \frac{p}{(1-p)\sqrt{s}} = \frac{-1}{\gamma_P} > 0, \quad (\text{A.10})$$

and we compute for $\gamma_P < \gamma \leq \sqrt{s}$

$$\begin{aligned} (f_{s,F}^R)''(\gamma) &= \left(\frac{f_s'(\gamma)(1 + b\gamma) - bf_s(\gamma)}{(1 + b(\gamma + f_s(\gamma)))^2} \right)' \\ &= \frac{f_s''(\gamma)(1 + b\gamma)(1 + b(\gamma + f_s(\gamma))) - 2b(f_s'(\gamma)(1 + b\gamma) - bf_s(\gamma))(1 + f_s'(\gamma))}{(1 + b(\gamma + f_s(\gamma)))^3}. \end{aligned} \quad (\text{A.11})$$

From

$$f_s''(\gamma), \quad 1 + b\gamma > 0, \quad f_s(\gamma > \max\{0, -\gamma\}), \quad -1 < f_s'(\gamma) < 0 \quad (\text{A.12})$$

for $\gamma_P < \gamma < \sqrt{s}$, it follows that $(f_{s,F}^R)''(\gamma) > 0$ for $\gamma_P < \gamma < \sqrt{s}$. Hence, $f_{s,F}^R(\gamma)$ is strictly convex in $\gamma_P < \gamma < \sqrt{s}$, and so is $f_{s,F}(\gamma)$, see (6.13).

The values of $f_{s,F}$, $f_{s,F}^R$ at $\gamma = \gamma_P + 0$ and $\gamma = \sqrt{s}$, as given in and below (6.14), follow from Proposition 1(v)–(vii) and the fact that $H_s(\gamma_P + 0) = \infty$, see (6.12), in this case.

(ii) We have for $\gamma_P < \gamma < \sqrt{s}$ from (6.3), (4.30), (6.4) and (6.13) that

$$g_{s,F}(\gamma) = \frac{g_s(\gamma)}{1 - p + \frac{1}{\sqrt{s}}p(\gamma + f_s(\gamma))}. \quad (\text{A.13})$$

Now $g_s(\gamma)$ is strictly decreasing and positive in $-\infty < \gamma < \sqrt{s}$ and $\gamma + f_s(\gamma)$ is strictly increasing and positive in $-\infty < \gamma < \sqrt{s}$. Hence, $g_{s,F}(\gamma)$ is positive and strictly decreasing in $\gamma_P < \gamma < \sqrt{s}$, and so is $g_{s,F}^R(\gamma) = (1-p)g_{s,F}(\gamma)$. The values assumed by $g_{s,F}$ and $g_{s,F}^R$ at $\gamma = \gamma_P + 0$ and $\gamma = \sqrt{s}$ follow from (i).

(iii) We have from (i) that $\gamma + f_{s,F}^R(\gamma)$ is strictly convex in $\gamma_P < \gamma < \sqrt{s}$, and $(f_{s,F}^R)'(\gamma) > -1$ for $\gamma_P < \gamma \leq \sqrt{s}$ by Proposition 14(ii). The values assumed by $\gamma + f_{s,F}^R(\gamma)$ at $\gamma = \gamma_P + 0$ and $\gamma = \sqrt{s}$ follow from (i).

(iv) Strict convexity of $\gamma + f_{s,F}(\gamma)$ in $\gamma_P < \gamma < \sqrt{s}$ follows from (i). The values assumed by $\gamma + f_{s,F}(\gamma)$ at $\gamma = \gamma_P + 0$ and $\gamma = \sqrt{s}$ also follow from (i). From these values it is seen that $\gamma + f_{s,F}(\gamma)$ is non-monotonic in $\gamma_P < \gamma < \sqrt{s}$.

This completes the proof of Proposition 15.

Note. It follows from the expression for $(f_{s,F}^R)'(\gamma)$ in (A.11) that

$$(f_{s,F}^R)'(\gamma_P + 0) = \frac{\gamma_P}{f_s(\gamma_P)} > -1, \quad (\text{A.14})$$

and that

$$(f_{s,F})'(\gamma_P + 0) = \frac{1}{1-p} \frac{\gamma_P}{f_s(\gamma_P)} < -1, \quad (\text{A.15})$$

where the two inequalities follow from $0 < \delta + f_s(\delta) < \sqrt{s}$ and the definition of γ_P , see (A.10).

Proof of Proposition 16. With $p_k = 1$, $s \leq k \leq N$, $p_k = 0$, $k > N$, we have $P = 0$ and $H_s(\gamma) = \gamma^{-1}\sqrt{s}(1 - (1 - \gamma/\sqrt{s})^{N+1})$. For $\gamma \leq \sqrt{s}$ it then follows from (6.6) and (4.30) that

$$g_{s,F}^R(\gamma) = \frac{(1 - \gamma/\sqrt{s})^N f_s(\gamma)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma) \sum_{n=0}^N (1 - \gamma/\sqrt{s})^n}. \quad (\text{A.16})$$

A computation shows that

$$\begin{aligned} & \sqrt{s}(g_{s,F}^R)'(\gamma) \frac{(\sqrt{s} + f_s(\gamma) H_s(\gamma))^2}{(1 - \gamma/\sqrt{s})^{N-1}} \\ &= -N f_s(\gamma) + \left(1 - \frac{\gamma}{\sqrt{s}}\right) \sqrt{s} f_s'(\gamma) - \frac{1}{\sqrt{s}} f_s^2(\gamma) \sum_{n=0}^N (N-n) \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n, \end{aligned} \quad (\text{A.17})$$

and all terms on the second line of (A.17) are negative. Hence, $g_{s,F}^R$ is strictly decreasing, as required.

Some examples. Take $H_s(\gamma)$ in (6.7) of the form

$$H_s(\gamma) = \sum_{n=0}^{\infty} q_n \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n, \quad q_n = \frac{\varepsilon P^n}{(n+1)^\alpha}, \quad (\text{A.18})$$

with $0 < \varepsilon < 1$, $\alpha > 0$, $0 < P < 1$. Then $0 < q_n < 1$, q_n is decreasing in n and $\limsup q_n^{1/(n+1)} = P$. When $\alpha > 1$, we have that $H_s(\gamma_P + 0) < \infty$. Consequently, in this case

$$1 - P < D_F^R(s, \lambda_P - 0) < D_F(s, \lambda_P - 0) < 1. \quad (\text{A.19})$$

When $1 < \alpha < 2$ we have that $H'_s(\gamma_P + 0) = -\infty$, and in this case $(D_F^R)'(s, \lambda_P - 0) = -\infty$, showing that $D_F^R(s, \lambda)$ is increasing for λ close to λ_P . An example where $F(\frac{1}{P} - 0) = H_s(\gamma_P + 0) = \infty$ and $g_{s,F}^R(\gamma_P + 0) = \infty = -(D_F^R)'(s, \lambda_P - 0)$ is provided by the choice $s = 1$ and $q_n = \frac{1}{10} \binom{-1/2}{n} (-\frac{1}{2})^n$, $n = 0, 1, \dots$, yielding $H_1(\gamma) = \frac{1}{10}(1 + \gamma)^{-1/2}$ with $P = 2$, $\gamma_P = -1$.

Proof of Lemma 2. We have from (6.4) and (6.5) that

$$\begin{aligned}
\frac{1}{\sqrt{s}} D_F^{-1}(s, s - \gamma\sqrt{s}) &= \frac{1 - \gamma/\sqrt{s}}{f_{s,F}(\gamma)} \\
&= \frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} \frac{1 + \frac{1}{\sqrt{s}} f_s(\gamma) H_s(\gamma)}{1 + (1 - \gamma/\sqrt{s}) H_s(\gamma)} \\
&= \frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} \frac{1 - \frac{1}{\sqrt{s}} \gamma H_s(\gamma) + \frac{1}{\sqrt{s}} H_s(\gamma)(f_s(\gamma) + \gamma)}{1 + (1 - \gamma/\sqrt{s}) H_s(\gamma)} \\
&= (1 - p(\gamma)) \frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} + p(\gamma) \frac{1}{\sqrt{s}} \frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} (f_s(\gamma) + \gamma), \tag{A.20}
\end{aligned}$$

with $p(\gamma) = q_\lambda$ where q_λ is given in (3.6). Now, by (6.4),

$$\frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} = \frac{1}{\sqrt{s}} B^{-1}(s, s - \gamma\sqrt{s}), \tag{A.21}$$

and

$$\frac{1 - \gamma/\sqrt{s}}{f_s(\gamma)} (f_s(\gamma) + \gamma) = 1 - \frac{\gamma}{\sqrt{s}} + \frac{\gamma}{\sqrt{s}} B^{-1}(s, s - \gamma\sqrt{s}) = C^{-1}(s, s - \gamma\sqrt{s}). \tag{A.22}$$

Then (3.4) follows from (A.20)–(A.22). The proof of (3.5) is similar.

We observe that $0 \leq p(\gamma) \leq 1$, which follows from (A.1).

Proof of Theorem 1. From [5, Section 3] one can extract a proof of Theorem 1. This proof depends on basic properties of birth-death processes. The proof that we give here is presented in QED coordinates and uses the analytic properties of $f_{s,F}^R$, such as given in Propositions 13 and 14. The proof is given under general conditions, so that also $P = 1$ and $F(\frac{1}{P} - 0), H_s(\gamma_P + 0) < \infty$ is allowed.

We distinguish the following cases:

Case a. $P = 0$. Then $\gamma_P = -\infty$, and both $H_s(\gamma)$ and $f_{s,F}^R(\gamma)$ are well-defined and analytic in $-\infty < \gamma \leq \sqrt{s}$. Set $\gamma_{P,F} = 0$.

Case b. $P \in (0, 1]$. Then $\gamma_P = -(1 - P)\sqrt{s}/P \in (-\infty, 0]$, and both $H_s(\gamma)$ and $f_{s,F}^R(\gamma)$ are well-defined and analytic in $\gamma_P < \gamma \leq \sqrt{s}$. In this case b, we distinguish the subcases

b1. $H_s(\gamma_P + 0) < \infty$. Then by Abel's theorem, $H_s(\gamma)$ is continuous in $\gamma_P \leq \gamma \leq \sqrt{s}$, and so is $f_{s,F}^R(\gamma)$.

b2. $H_s(\gamma_P + 0) = \infty$.

In these cases we have from (6.6) and (A.2)

$$f_{s,F}^R(\gamma_P + 0) = -\gamma_P + \frac{\gamma_P + f_s(\gamma_P)}{1 + \frac{1}{\sqrt{s}} f_s(\gamma_P) H_s(\gamma_P + 0)}. \quad (\text{A.23})$$

Hence $f_{s,F}^R(\gamma_P + 0) > -\gamma_P$ in case b1 while $f_{s,F}^R(\gamma_P + 0) = -\gamma_P$ in case b2. Set $\gamma_{P,F} = f_{s,F}^R(\gamma_P + 0) + \gamma_P$.

In case $P = 0$, we have by Proposition 13 and 14 that

$$f_{s,F}^R(\gamma) > -\gamma, \quad (f_{s,F}^R)'(\gamma) > -1, \quad (\text{A.24})$$

for $\gamma < \sqrt{s}$, and it follows as in the proof of [2], Theorem 10 in Section 4.3 that for any $\gamma \in (0, \sqrt{s})$ there is a unique solution a of the equation $a = f_{s,F}^R(\gamma - a)$. This solution, $a_{s,F}(\gamma)$, satisfies $a_{s,F}(\gamma) \rightarrow +\infty$ as $\gamma \downarrow 0$. For if $b := \liminf_{\gamma \downarrow 0} a_{s,F}(\gamma) < \infty$, we would have $b = \liminf_{\gamma \downarrow 0} f(\gamma - a_{s,F}(\gamma)) = f(-b)$, contradicting the first item in (A.24). Similarly, it can be shown, compare the beginning of the proof of [2], Theorem 8 in Section 4.8, by considering $c := \limsup_{\gamma \uparrow \sqrt{s}} a_{s,F}(\gamma)$, that $a_{s,F}(\gamma) \rightarrow 0$ as $\gamma \uparrow \sqrt{s}$. Assume now that $P \in (0, 1]$, in which we exclude the case that $p_k = 1$, $k \geq s$. Now (A.24) holds for $\gamma_P < \gamma < \sqrt{s}$. Let $\gamma \in (\gamma_{P,F}, \sqrt{s})$. We have

$$f_{s,F}^R(\gamma - 0) > 0, \quad (\text{A.25})$$

while

$$f_{s,F}^R(\gamma - (\gamma - \gamma_P) + 0) = \gamma_{P,F} - \gamma_P < \gamma - \gamma_P, \quad (\text{A.26})$$

and so

$$f_{s,F}^R(\gamma - \delta) < \delta \quad (\text{A.27})$$

when δ is less than but close to $\gamma - \gamma_P > 0$. By continuity, it follows from (A.25) and (A.27) that the equation $a = f_{s,F}^R(\gamma - a)$ has a solution a , and this solution is unique by (A.24). To show that this solution, $a_{s,F}(\gamma)$, satisfies $a_{s,F}(\gamma) \rightarrow f_{s,F}^R(\gamma_P + 0)$ as $\gamma \downarrow \gamma_{P,F}$, we need the following lemma.

Lemma 18. *Let $c < 0 < d$ and let $h : (c, d) \rightarrow \mathbb{R}$ be smooth and such that $h(a) > 0$, $h'(a) < 1$ for $c < a < d$ while $h(a) \rightarrow d$ when $a \uparrow d$. Then for any ε , $0 < \varepsilon < -c$, there is a unique solution $a(\varepsilon)$ of the equation $a = h(a - \varepsilon)$, and $a(\varepsilon) \rightarrow d$ as $\varepsilon \downarrow 0$.*

Proof. Let $\varepsilon \in (0, -c)$. Now

$$h(0 - \varepsilon) - 0 = h(-\varepsilon) > 0; \quad h(a - \varepsilon) - a \rightarrow -\varepsilon < 0, \quad a \uparrow d + \varepsilon. \quad (\text{A.28})$$

Hence, by continuity, there is an $a \in (0, d + \varepsilon)$ such that $a = h(a - \varepsilon)$, and this a is unique by the assumption that $h'(b) < 1$ for $b \in (c, d)$. To proceed, we first show that $h(a) > a$ when $c < a < d$. Indeed, when $h(a) < a$ for some $a \in (c, d)$, we would have

$$h(a_1) - h(a) = h(a_1) - a_1 - (h(a) - a) + a - a_1 > a_1 - a \quad (\text{A.29})$$

when a_1 is sufficiently close to d , contradicting $h'(b) < 1$ for all $b \in (c, d)$. Therefore, $h(a) \geq a$ for $a \in (c, d)$, and $h(a) = a$ for some $a \in (c, d)$ cannot occur either. For otherwise, we would have for $a_1 > a$ that

$$h(a_1) - h(a) \geq a_1 - a, \quad (\text{A.30})$$

contradicting $h'(b) < 1$ for all $b \in (c, d)$.

We now show that $a(\varepsilon) \rightarrow d$ as $\varepsilon \downarrow 0$. We have from $a(\varepsilon) < d + \varepsilon$, see (A.28), that $\limsup_{\varepsilon \downarrow 0} a(\varepsilon) \leq d$. Now suppose that $b := \liminf_{\varepsilon \downarrow 0} a(\varepsilon) < d$. Take $\varepsilon_n > 0$, $\varepsilon_n \rightarrow 0$ such that $a(\varepsilon_n) \rightarrow b < d$. Then

$$0 = h(a(\varepsilon_n) - \varepsilon_n) - a(\varepsilon_n) \rightarrow h(b) - b, \quad (\text{A.31})$$

contradicting $h(b) > b$ for all $b \in (c, d)$. This completes the proof of the lemma.

Taking in the lemma

$$h(a) = f_{s,F}^R(\gamma_{P,F} - a), \quad c := -\sqrt{s} + \gamma_{P,F} < a < f_{s,F}^R(\gamma_P + 0) =: d, \quad (\text{A.32})$$

it follows that $a_{s,F}(\gamma) \rightarrow f_{s,F}^R(\gamma_P + 0)$ as $\gamma \downarrow \gamma_{P,F}$. Finally, $a_{s,F}(\gamma) \rightarrow 0$ as $\gamma \uparrow \sqrt{s}$ can be shown by using the same argument as in [2], Theorem 8 in Section 4.8 for showing that $a(\gamma) \rightarrow 0$ as $\gamma \uparrow \sqrt{s}$. This completes the proof.

B Proof of Theorem 10

In this appendix we present the proof of Theorem 10 on the function L_s in (5.6), given in terms of the function $a_s(\gamma)$ that solves Cohen's equation $a = f_s(\gamma - a)$. The proofs rely heavily on (extensions of) the results in [2]. In particular, we use

$$\gamma a_s(\gamma) = 1 - \frac{2\gamma}{\sqrt{s}} - \left(1 - \frac{2}{s}\right)\gamma^2 + 4\left(1 - \frac{1}{s}\right)\frac{\gamma^3}{\sqrt{s}} + O(\gamma^4), \quad \gamma \downarrow 0, \quad (\text{B.1})$$

which is a sharpening of [2, Theorem 3]. This sharpening can be obtained by the method to prove [2, Theorem 3] where, as an intermediate step, [2, Proposition 2] should be sharpened to

$$f_s(\delta) = -\delta - \frac{1}{\delta} - \frac{2}{\delta^2\sqrt{s}} + \left(2 - \frac{6}{s}\right)\frac{1}{\delta^3} + \left(16 - \frac{24}{s}\right)\frac{1}{\delta^4\sqrt{s}} + O\left(\frac{1}{\delta^5}\right), \quad \delta \rightarrow -\infty, \quad (\text{B.2})$$

using the methods of [2, Section 4.1]. We shall also use and sharpen [2, Proposition 1],

$$1 - \frac{2}{\sqrt{s}}\gamma - \gamma^2 < \gamma a_s(\gamma) < 1 - \frac{1}{\sqrt{s}}\gamma, \quad 0 < \gamma < \sqrt{s}. \quad (\text{B.3})$$

Proof of $0 < L_s(\gamma) < \sqrt{s} - \gamma$, $0 < \gamma < \sqrt{s}$. This follows from the definition in (5.6) and $a_s(\gamma) > 0$.

Proof of $L_s(\gamma) = \gamma s(1 + O(\gamma\sqrt{s}))$, $\gamma \downarrow 0$. This follows from the definition in (5.6) and (B.3).

Proof of $L_s(\gamma) = (\sqrt{s} - \gamma)(1 + O(s^{-1/2}e^s(1 - \gamma/\sqrt{s})^s))$, $\gamma \uparrow \sqrt{s}$. This follows from the proof of [2, Theorem 4], and Stirling's formula.

We next show the results on unimodality.

Proposition 19. *We have for $0 < \gamma < \sqrt{s}$*

$$L'_s(\gamma) = 0 \Leftrightarrow \gamma a_s(\gamma) = \frac{1}{2}(1 - \gamma/\sqrt{s}). \quad (\text{B.4})$$

Proof. We have from the definition of L_s in (5.6)

$$L'_s(\gamma) = 0 \Leftrightarrow \sqrt{s} - \gamma + 2a_s(\gamma) + (\sqrt{s} - \gamma) a'_s(\gamma) = 0. \quad (\text{B.5})$$

By implicit differentiation in [2, (14)] and the expression in [2, Subsection 4.3] for f'_s in terms of f_s we have

$$a'_s(\gamma) = \frac{-a_s(\gamma)(\gamma + 1/\sqrt{s})}{1 - \gamma/\sqrt{s} - \gamma a_s(\gamma)}, \quad 0 < \gamma < \sqrt{s}. \quad (\text{B.6})$$

Using this in (B.5) with the facts that $\gamma > 0$ and $1 - \gamma/\sqrt{s} - \gamma a_s(\gamma) > 0$, we have for $0 < \gamma < \sqrt{s}$

$$L'_s(\gamma) = 0 \Leftrightarrow (\gamma a_s(\gamma))^2 + (1 - \gamma/\sqrt{s})(\gamma\sqrt{s} - \frac{1}{2})\gamma a_s(\gamma) - \frac{1}{2}(1 - \gamma/\sqrt{s})^2 \gamma\sqrt{s} = 0. \quad (\text{B.7})$$

The quadratic in $\gamma a_s(\gamma)$ occurring in the second proposition in (B.7) has the roots

$$\gamma a_s(\gamma) = -\frac{1}{2}(1 - \gamma/\sqrt{s})(\gamma\sqrt{s} - \frac{1}{2}) \pm \frac{1}{2}(1 - \gamma/\sqrt{s})(\gamma\sqrt{s} + \frac{1}{2}). \quad (\text{B.8})$$

Since $\gamma a_s(\gamma) > 0$, only the root in (B.8) with the + sign needs to be considered. The latter root equals $\frac{1}{2}(1 - \gamma/\sqrt{s})$, and this completes the proof.

To show unimodality of L_s , we should consider the function $\gamma a_s(\gamma)/(1 - \gamma/\sqrt{s})$, $0 < \gamma < \sqrt{s}$. This function assumes the values 1 and 0 at $\gamma = 0+$ and $\gamma = \sqrt{s} - 0$, and so, by Proposition 19, it is sufficient to show that this function is strictly decreasing in $0 < \gamma < \sqrt{s}$. The result we show below is somewhat stronger and will also be used in the proof of Proposition 21.

Proposition 20. *$\gamma a_s(\gamma)/(1 - \gamma/\sqrt{s})^2$ decreases strictly in $0 < \gamma < \sqrt{s}$ when $s > 1$.*

Proof. We compute

$$((1 - \gamma/\sqrt{s})^{-2} \gamma a_s(\gamma))' = (1 - \gamma/\sqrt{s})^{-2} \left[\frac{2}{\sqrt{s}} (1 - \gamma/\sqrt{s})^{-1} \gamma a_s(\gamma) + a_s(\gamma) - \gamma a'_s(\gamma) \right]. \quad (\text{B.9})$$

Using (B.6), we thus see that for $0 < \gamma < \sqrt{s}$

$$\begin{aligned} & ((1 - \gamma/\sqrt{s})^{-2} \gamma a_s(\gamma))' < 0 \\ \Leftrightarrow & (1 - \gamma a_s(\gamma) - \gamma/\sqrt{s}) \left(\frac{2\gamma}{\sqrt{s}} + 1 - \gamma/\sqrt{s} \right) - \gamma(\gamma + 1/\sqrt{s})(1 - \gamma/\sqrt{s}) < 0 \\ \Leftrightarrow & (1 - \gamma/\sqrt{s})(1 - \gamma^2) - \gamma a_s(\gamma)(1 + \gamma/\sqrt{s}) < 0. \end{aligned} \quad (\text{B.10})$$

It is therefore sufficient to show that for $0 < \gamma < \sqrt{s}$

$$\gamma a_s(\gamma) > (1 - \gamma^2) \frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}. \quad (\text{B.11})$$

We compute

$$\frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}(1 - \gamma^2) = 1 - \frac{2\gamma}{\sqrt{s}} - \left(1 - \frac{2}{s}\right)\gamma^2 + 2\left(1 - \frac{1}{s}\right)\frac{\gamma^3}{\sqrt{s}} + O(\gamma^4), \quad \gamma \downarrow 0, \quad (\text{B.12})$$

and so, by (B.1), we see that (B.11) holds for small positive γ . Now suppose that γ , $0 < \gamma < \sqrt{s}$, is such that

$$\gamma a_s(\gamma) = (1 - \gamma^2) \frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}. \quad (\text{B.13})$$

At such a γ we compute, using (B.6) and (B.12) twice,

$$\begin{aligned} (\gamma a_s(\gamma))' &= a_s(\gamma) - \frac{\gamma a_s(\gamma)(\gamma + 1/\sqrt{s})}{1 - \gamma/\sqrt{s} - \gamma a_s(\gamma)} \\ &= a_s(\gamma) - \frac{1 + \gamma/\sqrt{s}}{1 - \gamma/\sqrt{s}} a_s(\gamma) = -\frac{2}{\sqrt{s}} \frac{1 - \gamma^2}{1 + \gamma/\sqrt{s}}. \end{aligned} \quad (\text{B.14})$$

At the same time, we compute

$$\left((1 - \gamma^2) \frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}\right)' = -\frac{2}{\sqrt{s}} \frac{1 + \gamma\sqrt{s} - \gamma^2 - \gamma^3/\sqrt{s}}{(1 + \gamma/\sqrt{s})^2}. \quad (\text{B.15})$$

Since $s > 1$, we have for $0 < \gamma < \sqrt{s}$

$$1 + \gamma\sqrt{s} - \gamma^2 - \gamma^3/\sqrt{s} > (1 + \gamma/\sqrt{s})(1 - \gamma^2) = 1 + \gamma/\sqrt{s} - \gamma^2 - \gamma^3/\sqrt{s}. \quad (\text{B.16})$$

Hence, at a $\gamma \in (0, \sqrt{s})$ where (B.13) holds, we have

$$(\gamma a_s(\gamma))' > \left((1 - \gamma^2) \frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}\right)'. \quad (\text{B.17})$$

This is in particular so for

$$\gamma_0 := \inf \{0 < \gamma < \sqrt{s} \mid (\text{B.13}) \text{ holds}\}. \quad (\text{B.18})$$

This $\gamma_0 \in (0, \sqrt{s})$ since (B.11) holds for small positive γ and since we have assumed that there is a $\gamma \in (0, \sqrt{s})$ such that (B.13) holds. However, validity of (B.13) and (B.17) for $\gamma = \gamma_0$ implies that

$$\gamma a_s(\gamma) < (1 - \gamma^2) \frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}} \quad (\text{B.19})$$

holds for γ 's close to but less than γ_0 . However, (B.11) holds for γ 's close to 0, and so there is, by continuity, a $\gamma_1 < \gamma_0$ such that (B.13) holds. Contradiction, see (B.18). This proves that (B.11) holds for all $\gamma \in (0, \sqrt{s})$, and the proof is complete.

Note. We have $\gamma a_{s=1}(\gamma) = (1 - \gamma)^2$, $0 < \gamma \leq 1$.

We have from Proposition 20 that there is a unique root $\gamma = \hat{\gamma}_s \in (0, \sqrt{s})$ of

$$\gamma a_s(\gamma) = \frac{1}{2}(1 - \gamma/\sqrt{s}). \quad (\text{B.20})$$

From the behavior of $L_s(\gamma)$ near $\gamma = 0$ and $\gamma = \sqrt{s}$, it thus follows that L_s is unimodal, with unique maximum at $\gamma = \hat{\gamma}_s$. The value of $L_s(\gamma)$ at $\gamma = \hat{\gamma}_s$, see (5.11), is easily obtained by inserting (B.20) into the definition of L_s in (5.6).

We proceed by showing some properties of $\hat{\gamma}_s$.

Proposition 21. For $s > 1$,

$$\hat{\gamma}_s > f_s(0) > \left(\frac{1}{2} + \frac{1}{16s}\right)^{1/2} - \frac{1}{4\sqrt{s}}. \quad (\text{B.21})$$

Proof. It follows from $a_s(\gamma) = f_s(\gamma - a_s(\gamma))$ that $a_s(f_s(0)) = f_s(0)$. Using the inequality in (B.11) with $\gamma = f_s(0)$, we get

$$f_s^2(0) = (1 - f_s^2(0)) \frac{1 - f_s(0)/\sqrt{s}}{1 + f_s(0)/\sqrt{s}}. \quad (\text{B.22})$$

Working this out (noting that the cubic terms $f_s^3(0)$ cancel), we find

$$f_s^2(0) > \frac{1}{2} \left(1 - \frac{f_s(0)}{\sqrt{s}}\right). \quad (\text{B.23})$$

That is, the value assumed by $\gamma a_s(\gamma)/(1 - \gamma/\sqrt{s})$ at $\gamma = f_s(0)$ exceeds $1/2$. Since $\gamma a_s(\gamma)/(1 - \gamma/\sqrt{s})$ is strictly decreasing in $0 < \gamma < \sqrt{s}$, the first inequality in (B.21) follows. Finally, from (B.23) we also get

$$\left(f_s(0) + \frac{1}{4\sqrt{s}}\right)^2 - \frac{1}{2} - \frac{1}{16s} > 0, \quad (\text{B.24})$$

and this gives the second inequality in (B.21). \square

Note. We have $\hat{\gamma}_{s=1} = 1/2$.

Theorem 22. $\hat{\gamma}_s$ increases strictly in $s \geq 1$ from $1/2$ at $s = 1$ to $1.034113461\dots$ at $s = \infty$.

Proof. Let $s > t \geq 1$. From (B.3) we have for $u \geq 1$ that

$$p_u(\gamma) := \frac{\gamma a_u(\gamma)}{1 - \gamma/\sqrt{u}} = 1 - \gamma/\sqrt{u} + O(\gamma^2), \quad \gamma \downarrow 0, \quad (\text{B.25})$$

and so

$$\frac{\gamma a_s(\gamma)}{1 - \gamma/\sqrt{s}} > \frac{\gamma a_t(\gamma)}{1 - \gamma/\sqrt{t}} \quad (\text{B.26})$$

holds for small positive γ . Now suppose we have a $\gamma \in (0, \sqrt{t})$ such that

$$p := \frac{\gamma a_s(\gamma)}{1 - \gamma/\sqrt{s}} = \frac{\gamma a_t(\gamma)}{1 - \gamma/\sqrt{t}}. \quad (\text{B.27})$$

We have $p \in (0, 1)$. We compute, using (B.6) and the definition of p_u in (B.25)

$$\begin{aligned} \left(\frac{a_u(\gamma)}{1 - \gamma/\sqrt{u}}\right)' &= \frac{-a_u(\gamma)}{1 - \gamma/\sqrt{u}} \frac{\gamma + 1/\sqrt{u}}{1 - \gamma/\sqrt{u} - \gamma a_u(\gamma)} + \frac{1}{\sqrt{u}} \frac{a_u(\gamma)}{(1 - \gamma/\sqrt{u})^2} \\ &= -p_u(\gamma) \frac{\gamma + 1/\sqrt{u}}{\gamma(1 - \gamma/\sqrt{u})(1 - p_u(\gamma))} + \frac{p_u(\gamma)}{\sqrt{u}} \frac{1}{\gamma(1 - \gamma/\sqrt{u})}. \end{aligned} \quad (\text{B.28})$$

Hence, at a γ where (B.27) holds, we have

$$\left(\frac{a_u(\gamma)}{1 - \gamma/\sqrt{u}}\right)' = \frac{-p}{\gamma(1-p)} \frac{\gamma + 1/\sqrt{u}}{1 - \gamma/\sqrt{u}} + \frac{p}{\gamma\sqrt{u}} \frac{1}{1 - \gamma/\sqrt{u}}, \quad u = s, t. \quad (\text{B.29})$$

Therefore, at such a γ we have

$$\begin{aligned}
& \left(\frac{a_s(\gamma)}{1 - \gamma/\sqrt{s}} \right)' - \left(\frac{a_t(\gamma)}{1 - \gamma/\sqrt{t}} \right)' \\
&= \frac{-p}{\gamma(1-p)} \left(\frac{\gamma + 1/\sqrt{s}}{1 - \gamma/\sqrt{s}} - \frac{\gamma + 1/\sqrt{t}}{1 - \gamma/\sqrt{t}} \right) + \frac{p}{\gamma} \left(\frac{1}{\sqrt{s} - \gamma} - \frac{1}{\sqrt{t} - \gamma} \right) \\
&= \frac{p(1/\sqrt{s} - 1/\sqrt{t})}{\gamma(1 - \gamma/\sqrt{s})(1 - \gamma/\sqrt{t})} \left(-\frac{1 + \gamma^2}{1-p} + 1 \right) > 0
\end{aligned} \tag{B.30}$$

since $s > t$, $p \in (0, 1)$ and $\gamma \in (0, \sqrt{t})$. Letting $\gamma_0 = \inf \{0 < \gamma < \sqrt{t} \mid \text{(B.27) holds}\} \in (0, \sqrt{t})$, we arrive at a contradiction in the same way as in the proof of (B.11). Hence, (B.26) holds for all $\gamma \in (0, \sqrt{t})$. Then from decreasingness of $\gamma a_s(\gamma)/(1 - \gamma/\sqrt{s})$ as a function of γ , it is then seen that $\hat{\gamma}_s > \hat{\gamma}_t$.

The limit value of $\hat{\gamma}_s$ as $s \rightarrow \infty$ follows on considering the equation $\gamma a_\infty(\gamma) = 1/2$ in which $a_\infty(\gamma)$ is the unique solution of the Cohen equation in (7.2). We still need a lemma.

Lemma 23. $\gamma a_\infty(\gamma)$ strictly decreases in $0 < \gamma < \infty$ from 1 to 0.

Proof. Using

$$a'_\infty(\gamma) = \frac{-\gamma a_\infty(\gamma)}{1 - \gamma a_\infty(\gamma)}, \tag{B.31}$$

compare (B.6), it follows for $\gamma > 0$ that

$$(\gamma a_\infty(\gamma))' < 0 \Leftrightarrow \gamma a_\infty(\gamma) > 1 - \gamma^2. \tag{B.32}$$

From [2, (20)], it is seen that $\gamma a_\infty(\gamma) > 1 - \gamma^2$ holds for small positive γ . At a $\gamma > 0$ where $\gamma a_\infty(\gamma) = 1 - \gamma^2$, we compute $(\gamma a_\infty(\gamma))' = 0 > -2\gamma = (1 - \gamma^2)'$ and so, by the method of the proofs of (B.11) and (B.26), such a γ does not exist. Hence, $\gamma a_\infty(\gamma)$ is strictly decreasing in $\gamma > 0$. Finally, $\gamma a_\infty(\gamma) \rightarrow 1$ as $\gamma \downarrow 0$ follows from [2], Theorem 14, and $\gamma a_\infty(\gamma) \rightarrow 0$ as $\gamma \rightarrow \infty$ follows from

$$\gamma a_\infty(\gamma) = \gamma f_\infty(\gamma - a_\infty(\gamma)) < \gamma f_\infty(\gamma - a_\infty(1)) \rightarrow 0, \quad \gamma \rightarrow \infty. \tag{B.33}$$

This completes the proof of the lemma.

We conclude from Lemma 23 that $\gamma a_\infty(\gamma) = 1/2$ has a unique solution $\hat{\gamma}_\infty$, which can be determined numerically by a two-stage Newton method, using conveniently

$$f'_\infty(\delta) = -f_\infty(\delta)(\delta + f_\infty(\delta)) \quad \text{and} \quad \text{(B.31)} \tag{B.34}$$

for solving $a = f_\infty(\gamma - a)$ for $a = a_\infty(\gamma)$ and $\gamma a_\infty(\gamma) = 1/2$ for γ , respectively.

The value of $\hat{\gamma}_s$ for $s = 1$ follows from the Note before Theorem 22. \square