Optimality gaps in asymptotic dimensioning of many-server systems

Jaron Sanders*, S.C. Borst†, A.J.E.M. Janssen‡, and J.S.H. van Leeuwaarden§

Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

The Quality-and-Efficiency-Driven (QED) regime provides a basis for solving asymptotic dimensioning problems that trade off revenue, costs and service quality. We derive bounds for the *optimality gaps* that capture the differences between the true optimum and the asymptotic optimum based on the QED approximations. Our bounds generalize earlier results for classical many-server systems. We also apply our bounds to a many-server system with threshold control.

Keywords: QED regime, Halfin–Whitt regime, queues in heavy traffic, asymptotic analysis, asymptotic dimensioning, optimality gap

1 Introduction

The theory of square-root staffing in many-server systems ranks among the most celebrated principles in applied probability. The general idea behind square-root staffing is as follows: a finite server system is modeled as a system in heavy traffic, where the number of servers s is large, whereas at the same time, the system is critically loaded. Under Markovian assumptions, and denoting the load on the system by λ, this can be achieved by setting $s = \lambda + \beta\sqrt{\lambda}$ and letting $\lambda \to \infty$ while keeping $\beta > 0$ fixed, or alternatively setting $\lambda = s - \gamma\sqrt{s}$ and letting $s \to \infty$ while keeping $\gamma > 0$ fixed. In both cases, the system reaches the desirable Quality-and-Efficiency-Driven (QED) regime.

*jaron.sanders@tue.nl
†s.c.borst@tue.nl
‡a.j.e.m.janssen@tue.nl
§j.s.h.v.leeuwaarden@tue.nl

The QED regime refers to mathematically defined conditions in which both customers and system operators benefit from the advantages that come with systems that operate efficiently at large scale, which is particularly relevant for systems in e.g. health care, cloud computing, and customer services. Such conditions manifest themselves in a low delay probability and negligible mean delay, despite the fact that the system utilization is high. Properties of this sort can be proven rigorously for systems such as the $M/M/s$ queue by establishing stochastic-process limits under the aforementioned QED scalings [2]. The QED regime also creates a natural environment for solving dimensioning problems that achieve an acceptable trade-off between service quality and capacity. Quality is usually formulated in terms of some target service level. Take for instance the probability that an arriving customer experiences delay. The target could be to keep the delay probability below some value $\epsilon \in (0,1)$. The smaller ϵ, the better the offered quality of service. Once the target service level is set, the objective from the operator's perspective is to determine the highest load λ such that the target ϵ is still met.

For the $M/M/s$ queue, it was shown by Borst et al. [1] that such dimensioning procedures combined with QED approximations have certain asymptotic optimality properties. To illustrate this, consider the case of linear costs, i.e. waiting cost are b per customer per unit time, and staffing cost are c per server per unit time. Denoting the total cost by $K_\lambda(s)$, it can be shown that when $s = \lambda + \beta\sqrt{\lambda}$ and $\beta > 0$,

$$K_\lambda(s) = b\lambda\frac{C_\lambda(s)}{s-\lambda} + cs = c\lambda + \sqrt{\lambda}\Big(c\beta + \frac{b}{\beta}C_\lambda(s)\Big) \tag{1}$$

with $C_\lambda(s)$ the delay probability in the $M/M/s$ queue. The first term $c\lambda$ represents the cost of the minimally required capacity λ, while the second term gathers the cost factors that are all $O(\sqrt{\lambda})$. Halfin and Whitt [2] showed that in the QED regime $C_\lambda(s)$ converges to a nondegenerate limit $C_0(\beta) \in (0,1)$, so that in the QED regime one only needs to determine $\beta_0 = \arg\min_\beta\{c\beta + bC_0(\beta)/\beta\}$, and then set $s_0 = [\lambda + \beta_0\sqrt{\lambda}]$ as an approximation for the optimal number of servers $s^{\rm opt} = \arg\min_s\{K_\lambda(s)\}$. Borst et al. [1] called this procedure *asymptotic dimensioning*.

Based on the QED limiting regime, one expects that such approximate solutions are accurate for large relative loads λ. For the *optimality gaps* $|s_0 - s^{\rm opt}|$ and $|K_\lambda(s_0) - K_\lambda(s^{\rm opt})|$, i.e. inaccuracies that arise from the fact that the actual system is of finite size, Borst et al. [1] showed through numerical experiments that the approximation s_0 performs exceptionally well in almost all circumstances, even when systems are only moderately sized. A rigorous underpinning for these observations was provided by Janssen et al. [5], who used *refined* QED approximations to quantify the optimality gaps. The delay probability, for instance, was shown to behave as $C_0(\beta) + C_1(\beta)/\sqrt{\lambda} + O(\lambda^{-1})$, which in turn was used to estimate the optimality gaps for the dimensioning problem in (1). Zhang et al. [8] obtained similar results for optimality gaps in the context of the $M/M/s+M$ queue, in which customers may abandon before receiving service.

Motivated by the results in [5, 8], Randhawa [6] took a more abstract approach to quantify optimality gaps of asymptotic dimensioning problems. He showed under general assumptions that when the approximation to the objective function is accurate up to $O(1)$, the prescriptions that are derived from this approximation are $o(1)$-optimal. The optimality gap thus becomes zero asymptotically. This general setup was shown in [6] to apply to the $M/M/s$ queues in the QED regime, which confirmed and sharpened the results on the optimality gaps in [5, 8] by implying that $|K_\lambda(s_0) - K_\lambda(s^{\text{opt}})| = o(1)$. The abstract framework in [6], however, can only be applied if refined approximations as in [5, 8] are available.

Such refined approximations were recently developed in [4, 7] for a broad class of many-server systems operating in the QED regime with $\lambda = s - \gamma\sqrt{s}$, and equipped with an admission control policy and a revenue structure. For a wide range of performance metrics, $M_s(\gamma)$ say, these refinements are of the form $M_s(\gamma) = M_0(\gamma) + M_1(\gamma)/\sqrt{s} + \cdots$. The method in [4, 7] can deliver as many higher-order terms as needed, and generate all the refinements obtained in [5, 8, 6].

In the present paper, we demonstrate how the results in [4, 7] can be leveraged to determine the optimality gaps of novel asymptotic dimensioning problems for a large class of many-server systems. Our main result (Theorem 1) provides generic bounds for the optimality gaps that become sharper when more terms in the QED expansion for $M_s(\gamma)$ are included.

2 Model description

2.1 Service systems with admission control and revenues

We consider many-server systems with s parallel servers, to which customers arrive according to a Poisson process with rate λ. Every customer requires an exponentially distributed service time with mean one. If a customer arrives and finds $k - s \geq 0$ customers waiting, the customer is allowed to join the queue with probability $p_s(k-s)$ and is rejected with probability $1 - p_s(k-s)$. The total number of customers in the system evolves as a birth–death process $\{X_s(t)\}_{t\geq 0}$ and has a stationary distribution

$$\pi_s(k) = \begin{cases} Z^{-1}, & k = 0, \\ Z^{-1}\frac{(s\rho)^k}{k!}, & k = 1, 2, \ldots, s, \\ Z^{-1}\frac{s^s\rho^k}{s!}\prod_{i=0}^{k-s-1} p_s(i), & k = s+1, s+2, \ldots, \end{cases} \tag{2}$$

where $\rho = \lambda/s$, $Z = \sum_{k=0}^{s}(s\rho)^k/k! + ((s\rho)^s/s!)F_s(\rho)$, and $F_s(\rho) = \sum_{n=0}^{\infty} p_s(0)\cdot \ldots \cdot p_s(n)\rho^{n+1}$. The stationary distribution in (2) exists if and only if the relative load ρ and the *admission control policy* $\{p_s(k)\}_{k\in\mathbb{N}_0}$ are such that $F_s(\rho) < \infty$.

Next, we assume that the system generates revenue at rate $r_s(k) \in \mathbb{R}$ when there are k customers in the system. The sequence $\{r_s(k)\}_{k\in\mathbb{N}_0}$ will be called the *revenue structure*. The stationary rate at which the system generates revenue is

then given by

$$R_s(\gamma) = \sum_{k=0}^{\infty} r_s(k)\pi_s(k), \tag{3}$$

which depends via the equilibrium distribution on the admission control policy. Ref. [7] discusses the problem of maximizing the revenue rate over the set of all admission control policies.

One advantage of considering general admission control policies and revenue structures is that one can study different service systems and steady-state performance measures through one unifying framework. For example, the equilibrium behavior of the canonical $M/M/s/s$, $M/M/s$, and $M/M/s+M$ systems can be recovered by choosing $p_s(k-s) = 0$, $p_s(k-s) = 1$, and $p_s(k-s) = 1/(1+(k-s+1)\theta/s)$, respectively. Here, θ corresponds to the rate at which waiting customers abandon from the $M/M/s+M$ system. Similarly, the delay probability $D_s(\gamma) = \sum_{k=s}^{\infty} \pi_s(k)$ can be recovered by setting $r_s(k) = \mathbb{1}[k \geq s]$, the mean queue length $Q_s(\gamma) = \sum_{k=s}^{\infty}(k-s)\pi_s(k)$ is recovered when considering $r_s(k) = (k-s)\mathbb{1}[k \geq s]$, and the average number of idle servers $I_s(\gamma) = \sum_{k=0}^{s-1}(s-k)\pi_s(k)$ follows from $r_s(k) = (s-k)\mathbb{1}[k < s]$.

As a primary example we will consider a scenario where besides the waiting cost $b > 0$ incurred per customer per unit time, a fee $a > 0$ is received for every served customer, and a penalty $d \geq 0$ is imposed for rejecting a customer. The latter cost accounts for the degree of revenue loss from the admission control policy. Denoting by $D_s^R(\gamma) = \sum_{k=s}^{\infty}(1-p_s(k-s))\pi_s(k)$ the probability that an arriving customer is rejected, and by $W_s(\gamma) = \sum_{k=s}^{\infty}((k-s+1)/s)p_s(k-s)\pi_s(k)$ the expected waiting time of an arriving customer, the total system revenue rate is given by

$$R_s(\gamma) = a\lambda(1 - D_s^R(\gamma)) - b\lambda W_s(\gamma) - d\lambda D_s^R(\gamma). \tag{4}$$

By virtue of Little's law $\lambda W_s(\gamma) = Q_s(\gamma)$ and $\lambda(1 - D_s^R(\gamma)) = s - I_s(\gamma)$, and since $\lambda = s - \gamma\sqrt{s}$, the revenue rate can equivalently be expressed as

$$R_s(\gamma) = as + d\gamma\sqrt{s} - (a+d)I_s(\gamma) - bQ_s(\gamma). \tag{5}$$

This scenario therefore corresponds to the revenue structure

$$r_s(k) = \begin{cases} ak + d\gamma\sqrt{s} - d(s-k) & k < s, \\ as + d\gamma\sqrt{s} - b(k-s), & k \geq s. \end{cases} \tag{6}$$

2.2 QED scaling and refinements

We now discuss how to apply the QED scaling to obtain an asymptotic expansion for $R_s(\gamma)$ for general revenue structures $\{r_s(k)\}_{k\in\mathbb{N}_0}$, which we will exploit in §3 to characterize the asymptotic optimality gap. We impose the following three conditions throughout this paper:

(i) The arrival rate and system size are coupled via the scaling $\lambda = s - \gamma\sqrt{s}$;

(ii) $p_s(0) \cdot \ldots \cdot p_s(n) \to f((n+1)/\sqrt{s})$ where $f(x)$ is either a continuous, nonincreasing function, or $f(x) = \mathbb{1}[x \le \eta]$;

(iii) There exist sequences $\{n_s\}_{s\in\mathbb{N}_+}$, $\{q_s\}_{s\in\mathbb{N}_+}$ with $q_s > 0$, and a continuous function $r(x)$ that satisfy the scaling condition $(r_s(k) - n_s)/q_s \to r((k-s)/\sqrt{s})$.

It is proven in [4, 7] that $\lim_{s\to\infty}(R_s(\gamma) - n_s)/q_s = R_0(\gamma)$ under conditions (i)–(iii), with

$$R_0(\gamma) = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{-\frac{1}{2}x^2-\gamma x}\,\mathrm{d}x + \int_0^\infty r(x)f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\frac{\Phi(\gamma)}{\phi(\gamma)} + \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}. \tag{7}$$

Here, Φ and ϕ denote the cumulative distribution function and probability density function of the standard normal distribution. This asymptotic characterization of the revenue is leveraged in [7] to prove that for many revenue structures there exists an optimal admission control policy with a threshold structure.

Moreover, the method used to obtain (7) in [4, 7] can be extended to derive an asymptotic expansion of the form

$$R_s(\gamma) = n_s + q_s\Big(\sum_{i=0}^{j}\frac{R_i(\gamma)}{s^{i/2}} + O\Big(\frac{1}{s^{(j+1)/2}}\Big)\Big), \tag{8}$$

which is shown in A. We also provide closed-form expressions for the first two terms $R_0(\gamma)$ and $R_1(\gamma)$ for arbitrary $f(x)$ and $r(x)$. The asymptotic expansion in (8) is a crucial ingredient for determining the optimality gaps.

Let us discuss the asymptotic expansion in the context of (5). Denoting $n_s = as$ and extracting a term $q_s = \sqrt{s}$ yields $R_s(\gamma) = n_s + \sqrt{s}\hat{R}_s(\gamma)$ with

$$\hat{R}_s(\gamma) = d\gamma - (a+d)\frac{I_s(\gamma)}{\sqrt{s}} - b\frac{Q_s(\gamma)}{\sqrt{s}}. \tag{9}$$

Since our goal is to maximize $R_s(\gamma)$ over γ, and because the term n_s is constant and independent of γ, we only need to focus on the maximization of $\hat{R}_s(\gamma)$.

Recall (6) and note that the limiting revenue structure for the objective function in (9) is given by $r(x) = (a+d)x + d\gamma$ for $x < 0$, and $r(x) = -bx + d\gamma$ for $x \ge 0$. With an admission control policy $f(x) = \mathbb{1}[x \le \eta]$ where $\eta \ge 0$ denotes the admission threshold, it follows from (7) that

$$\lim_{s\to\infty}\hat{R}_s(\gamma) = \hat{R}_0(\gamma) = d\gamma - \frac{(a+d)\big(1+\gamma\frac{\Phi(\gamma)}{\phi(\gamma)}\big) + b\frac{1-(1+\gamma\eta)\mathrm{e}^{-\gamma\eta}}{\gamma^2}}{\frac{\Phi(\gamma)}{\phi(\gamma)} + \frac{1-\mathrm{e}^{-\gamma\eta}}{\gamma}}. \tag{10}$$

We prove in A that $\lim_{s\to\infty}\sqrt{s}(\hat{R}_s(\gamma) - \hat{R}_0(\gamma)) = \hat{R}_1(\gamma)$ and provide an explicit expression for $\hat{R}_1(\gamma)$. We then have both a first-order approximation $\hat{R}_0(\gamma)$ and a second-order approximation $\hat{R}_0(\gamma) + \hat{R}_1(\gamma)/\sqrt{s}$ for the objective function $\hat{R}_s(\gamma)$.

3 General revenue maximization

For general objective functions (3), we now aim for solving the dimensioning problem

$$\max_{\gamma\in\Gamma}\{R_s(\gamma)\}, \tag{11}$$

where we assume that $\Gamma = [\gamma_l, \gamma_r]$ is a compact interval contained in $(\gamma_s^{\min}, \infty)$, with $\gamma_s^{\min} = \inf\{\gamma \in \mathbb{R} | F_s(\rho) < \infty\}$. Denote the exact solution by

$$\gamma_s^{\text{opt}} = \arg\max_{\gamma\in\Gamma}\{R_s(\gamma)\}. \tag{12}$$

We assume that an asymptotic expansion of the form (8) is available for $R_s(\gamma)$ and its derivative $R_s{}'(\gamma)$, which we can then use to approximate the objective function. Hence we will consider

$$\gamma_{j,s} = \arg\max_{\gamma\in\Gamma}\Big\{R_0(\gamma) + \ldots + \frac{R_j(\gamma)}{s^{j/2}}\Big\} \tag{13}$$

as approximations for the exact solution γ_s^{opt}, which should be increasingly better for larger j and/or s. Note that $\gamma_{0,s} = \gamma_0$ is independent of s.

Denoting the ith derivative of a function $g(x)$ by $g^{(i)}(x)$, we assume also that $R_0{}^{(k)}(\gamma), \ldots, R_j{}^{(k)}(\gamma)$ are bounded on Γ for $k = 0, 1, 2$, and that $R_{j+1}{}^{(l)}(\gamma)$ is bounded on Γ for $l = 0, 1$. We furthermore assume that both the first-order optimizer γ_0 and the exact optimizer γ_s^{opt} exist, are unique and lie in the interior of Γ, and that $R_0(\gamma)$ is strictly concave on Γ and has a continuous derivative $R_0{}''(\gamma)$ on Γ. Finally, we assume that $f(x)$ is such that

$$\lim_{\gamma\downarrow\gamma^{\min}}\int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x = \infty, \tag{14}$$

where $\gamma^{\min} = \inf\{\gamma \in \mathbb{R} | \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x < \infty\}$. Ref. [4] discusses under which conditions assumption (14) is satisfied, and it is satisfied for instance for any admission control policy $f(x) = \mathbb{1}[x \le \eta]$ with admission threshold $\eta > 0$, since $\int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x = (1 - \mathrm{e}^{-\gamma\eta})/\gamma$ and $\gamma^{\min} = -\infty$.

3.1 Optimality gaps

We now derive the optimality gaps for the general optimization problem in (11). Theorem 1 formalizes that the approximating solutions $\gamma_{j,s}$ are asymptotically optimal through bounds for the optimality gaps, and that an approximation of order j yields a gap decay of order $j+1$. With minor modifications to the proof, the result also applies to minimization problems of the form $\min_{\gamma\in\Gamma}\{R_s(\gamma)\}$.

Theorem 1. *For $j = 0, 1, \ldots$, there exist constants M_j and K_j independent of s and $s_j \in \mathbb{N}_+$ such that for all $s \ge s_j$,*

$$|R_s(\gamma_s^{\text{opt}}) - R_s(\gamma_{j,s})| \le \frac{q_s M_j}{s^{(j+1)/2}}, \quad |\gamma_{j,s} - \gamma_s^{\text{opt}}| \le \frac{K_j}{s^{(j+1)/2}}. \tag{15}$$

Proof First, we will prove a monotonicity result, as well as the existence of optimizers.

Lemma 1. *There is an $s_0 \in \mathbb{N}_+$ such that for all $s \geq s_0$, the function $R_0(\gamma) + \sum_{i=1}^{j} R_i(\gamma)/s^{i/2}$ has a unique optimizer $\gamma_{j,s} \in \Gamma$ and a strictly decreasing derivative $R_0{}'(\gamma) + \sum_{i=1}^{j} R_i{}'(\gamma)/s^{i/2}$.*

Proof. Recall that γ_0 lies in the interior of Γ and that $R_0{}'(\gamma)$ is strictly decreasing on Γ by assumption, which implies that $R_0{}'(\gamma_l) > 0$. We seek $s_1 \in \mathbb{N}_+$ such that for all $s \geq s_1$, $R_0{}'(\gamma_l) + \sum_{i=1}^{j} R_i{}'(\gamma_l)/s^{i/2} > 0$. Note next that $(1/\sqrt{s}) \sum_{i=1}^{j} |R_i{}'(\gamma_l)| \geq -\sum_{i=1}^{j} R_i{}'(\gamma_l)/s^{i/2}$ for $s \in \mathbb{N}_+$. For all $s \in \mathbb{N}_+$ for which $R_0{}'(\gamma_l) > (1/\sqrt{s}) \sum_{i=1}^{j} |R_i{}'(\gamma_l)|$, we have thus consequently that $R_0{}'(\gamma_l) > -\sum_{i=1}^{j} R_i{}'(\gamma_l)/s^{i/2}$. We therefore pick $s_1 = \lceil (\sum_{i=1}^{j} |R_i{}'(\gamma_l)|/R_0{}'(\gamma_l))^2 \rceil$ to ensure that for all $s \geq s_1$, $R_0{}'(\gamma_l) + \sum_{i=1}^{j} R_i{}'(\gamma_l)/s^{i/2} > 0$. A similar result holds at $\gamma = \gamma_r$, i.e. we have $R_0{}'(\gamma_r) < 0$, and thus $s_2 = \lceil (\sum_{i=1}^{j} |R_i{}'(\gamma_r)|/R_0{}'(\gamma_r))^2 \rceil$ is such that for all $s \geq s_2$, $R_0{}'(\gamma_r) + \sum_{i=1}^{j} R_i{}'(\gamma_r)/s^{i/2} < 0$. The function $R_0(\gamma) + \sum_{i=1}^{j} R_i(\gamma)/s^{i/2}$ thus has a unique optimizer $\gamma_{j,s} \in \Gamma$.

Finally we turn to proving the monotonicity property of $R_0{}'(\gamma) + \sum_{i=1}^{j} R_i{}'(\gamma)/s^{i/2}$. By assumption, $R_0{}''(\gamma) < 0$ for all $\gamma \in [\gamma_l, \gamma_r]$. Similar to before, set $s_3 = \max_{\gamma \in [\gamma_l, \gamma_r]} \lceil (\sum_{i=1}^{j} |R_i{}''(\gamma)|/R_0{}''(\gamma))^2 \rceil$, and conclude that $R_0{}''(\gamma) + \sum_{i=1}^{j} R_i{}''(\gamma)/s^{i/2} < 0$ for all $s \geq s_3$ and all $\gamma \in [\gamma_l, \gamma_r]$. Finish the proof by setting $s_0 = \max\{s_1, s_2, s_3\}$, and by noting that s_0 is bounded. □

Recall that the unique optimizer γ_s^{opt} exists, and lies in the interior of Γ. Because γ_s^{opt} maximizes $R_s(\gamma)$, we have therefore by suboptimality that

$$0 \leq R_s(\gamma_s^{\mathrm{opt}}) - R_s(\gamma_{j,s}) = \Big[q_s \sum_{i=0}^{j} \frac{R_i(\gamma_{j,s})}{s^{i/2}} - R_s(\gamma_{j,s}) \Big] \tag{16}$$
$$- \Big[q_s \sum_{i=0}^{j} \frac{R_i(\gamma_s^{\mathrm{opt}})}{s^{i/2}} - R_s(\gamma_s^{\mathrm{opt}}) \Big] + q_s \Big[\sum_{i=0}^{j} \frac{R_i(\gamma_s^{\mathrm{opt}})}{s^{i/2}} - \sum_{i=0}^{j} \frac{R_i(\gamma_{j,s})}{s^{i/2}} \Big],$$

and subsequently by expansion (8) that

$$0 \leq R_s(\gamma_s^{\mathrm{opt}}) - R_s(\gamma_{j,s}) = q_s \Big[\sum_{i=0}^{j} \frac{R_i(\gamma_s^{\mathrm{opt}})}{s^{i/2}} - \sum_{i=0}^{j} \frac{R_i(\gamma_{j,s})}{s^{i/2}} \Big]$$
$$+ q_s \frac{R_{j+1}(\gamma_s^{\mathrm{opt}}) - R_{j+1}(\gamma_{j,s})}{s^{(j+1)/2}} + O\Big(\frac{q_s}{s^{(j+2)/2}} \Big). \tag{17}$$

Since $\gamma_{j,s}$ maximizes $\sum_{i=0}^{j} R_i(\gamma)/s^{i/2}$, we have by suboptimality that the term within square brackets is negative, i.e.

$$\sum_{i=0}^{j} \frac{R_i(\gamma_s^{\mathrm{opt}})}{s^{i/2}} - \sum_{i=0}^{j} \frac{R_i(\gamma_{j,s})}{s^{i/2}} \leq 0 \tag{18}$$

for all $s \in \mathbb{N}_0$. Therefore, since $q_s > 0$,

$$0 \leq R_s(\gamma_s^{\text{opt}}) - R_s(\gamma_{j,s}) \leq q_s \frac{R_{j+1}(\gamma_s^{\text{opt}}) - R_{j+1}(\gamma_{j,s})}{s^{(j+1)/2}} + O\Big(\frac{q_s}{s^{(j+2)/2}}\Big). \quad (19)$$

Since $R_{j+1}(\gamma)$ is bounded on Γ by assumption, the first claim in (15) follows.

Corollary 1. *There exist constants $M_j{}' > 0$ independent of s and $s_j{}' \in \mathbb{N}_+$ such that for all $s \geq s_j{}'$,*

$$\Big|\sum_{i=0}^{j} \frac{R_i{}'(\gamma_s^{\text{opt}})}{s^{i/2}} - \sum_{i=0}^{j} \frac{R_i{}'(\gamma_{j,s})}{s^{i/2}}\Big| = \Big|\sum_{i=0}^{j} \frac{R_i{}'(\gamma_s^{\text{opt}})}{s^{i/2}}\Big| \leq \frac{M_j{}'}{s^{(j+1)/2}}. \quad (20)$$

Proof. Note that $\gamma_{j,s}$ is the optimizer of $\sum_{i=0}^{j} R_i(\gamma)/s^{i/2}$, and that therefore $\sum_{i=0}^{j} R_i{}'(\gamma_{j,s})/s^{i/2} = 0$, which proves the leftmost equality.

Next, we examine the asymptotic expansion of the derivative of $R_s(\gamma)$, which we have assumed is available and of the form (8). It follows that

$$\begin{aligned} q_s \sum_{i=0}^{j} \frac{R_i{}'(\gamma_s^{\text{opt}})}{s^{i/2}} &= R_s'(\gamma_s^{\text{opt}}) - q_s \frac{R_{j+1}{}'(\gamma_s^{\text{opt}})}{s^{(j+1)/2}} + O\Big(\frac{q_s}{s^{(j+2)/2}}\Big) \\ &\overset{\text{(i)}}{=} -q_s \frac{R_{j+1}{}'(\gamma_s^{\text{opt}})}{s^{(j+1)/2}} + O\Big(\frac{q_s}{s^{(j+2)/2}}\Big), \end{aligned} \quad (21)$$

since (i) γ_s^{opt} optimizes $R_s(\gamma)$. Now (20) follows since $R_{j+1}{}'(\gamma)$ is bounded on Γ. □

We are now ready to establish the second claim in (15). Recall that for all $j \in \mathbb{N}_+$, and sufficiently large s, the function $\sum_{i=0}^{j} R_i{}'(\gamma)/s^{(i/2)}$ is strictly decreasing in $[\gamma_l, \gamma_r]$, see Lemma 1. Note also that $\gamma_0, \gamma_s^{\text{opt}}, \gamma_{j,s} \in [\gamma_l, \gamma_r]$. The mean value theorem implies then that

$$\Big|\sum_{i=0}^{j} \frac{R_i{}'(\gamma_s^{\text{opt}})}{s^{(i/2)}} - \sum_{i=0}^{j} \frac{R_i{}'(\gamma_{j,s})}{s^{(i/2)}}\Big| \geq m_{j,s} |\gamma_s^{\text{opt}} - \gamma_{j,s}| \quad (22)$$

with

$$m_{j,s} = -\max_{\gamma \in [\gamma_l, \gamma_r]} \Big\{ R_0{}''(\gamma) + \sum_{i=1}^{j} \frac{R_i{}''(\gamma)}{s^{i/2}} \Big\}, \quad (23)$$

where we have also used that the function $\sum_{i=0}^{j} R_i(\gamma)/s^{(i/2)}$ is optimized by $\gamma_{j,s}$. Combining with Corollary 1, it follows that

$$\frac{M_j{}'}{s^{(j+1)/2}} \geq \Big|\sum_{i=0}^{j} \frac{R_i{}'(\gamma_s^{\text{opt}})}{s^{i/2}} - \sum_{i=0}^{j} \frac{R_i{}'(\gamma_{j,s})}{s^{i/2}}\Big| \geq m_{j,s} |\gamma_s^{\text{opt}} - \gamma_{j,s}|, \quad (24)$$

which is almost the second claim in (15). What remains is to remove the dependency on s of $m_{j,s}$.

To that end, remark that $\max_{\gamma\in[\gamma_l,\gamma_r]} {R_0}''(\gamma) < 0$ by continuity and pointwise negativity of ${R_0}''(\gamma)$. Then, bound

$$\begin{aligned} m_{j,s} &= \min_{\gamma\in[\gamma_l,\gamma_r]} \Big\{ -{R_0}''(\gamma) - \sum_{i=1}^{j} \frac{{R_i}''(\gamma)}{s^{i/2}} \Big\} \\ &\geq \min_{\gamma\in[\gamma_l,\gamma_r]} \{-{R_0}''(\gamma)\} + \frac{1}{\sqrt{s}} \min_{\gamma\in[\gamma_l,\gamma_r]} \Big\{ -\sum_{i=1}^{j} \frac{{R_i}''(\gamma)}{s^{(i-1)/2}} \Big\} \qquad (25) \\ &\geq \min_{\gamma\in[\gamma_l,\gamma_r]} \{-{R_0}''(\gamma)\} - \frac{1}{\sqrt{{s_j}'}} \max_{\gamma\in[\gamma_l,\gamma_r]} \Big\{ \sum_{i=1}^{j} \frac{{R_i}''(\gamma)}{({s_j}')^{(i-1)/2}} \Big\} =: m_j, \end{aligned}$$

and increase the value of ${s_j}'$ if necessary to ensure that $m_j > 0$.

Summarizing, we now have that ${M_j}'/s^{(j+1)/2} \geq m_j|\gamma_s^{\mathrm{opt}} - \gamma_{j,s}|$ for all $s \geq {s_j}'$. Setting $K_j = {M_j}'/m_j$ completes the proof. □

3.2 Dimensioning under a delay constraint

The approach of §3.1 can also be used to solve delay constrained dimensioning problems. As an example, consider finding

$$\gamma_s^{\mathrm{opt}} = \arg_{\gamma\in\Gamma}\{D_s(\gamma) = \epsilon\}, \qquad (26)$$

where $\epsilon \in (0,1)$. Since we have an asymptotic expansion of the form (8) for the delay probability, see A, instead of solving for γ_s^{opt} directly we can obtain approximations $\gamma_{j,s} = \arg_{\gamma\in\Gamma_j}\{\sum_{i=0}^{j} D_i(\gamma)/s^{i/2} = \varepsilon\}$. The optimality gaps can be calculated using a similar proof technique.

Corollary 2. *For $j = 0, 1, \ldots$ there exist finite constants M_j, $K_j > 0$ independent of s and $s_j \in \mathbb{N}_+$ such that for all $s \geq s_j$,*

$$|D_s(\gamma_{j,s}) - \varepsilon| \leq \frac{M_j}{s^{(j+1)/2}}, \quad |\gamma_{j,s} - \gamma_s^{\mathrm{opt}}| \leq \frac{K_j}{s^{(j+1)/2}}. \qquad (27)$$

4 Approaches to asymptotic dimensioning

4.1 Asymptotic revenue maximization with a threshold

We will now consider new approaches to asymptotic dimensioning in the context of optimizing the objective function (4). We start with considering the revenue maximization problem described in §2.1 while assuming that the threshold is fixed. This concretely requires us to maximize $\hat{R}_0(\gamma)$ in (9) over γ given a fixed $\eta < \infty$.

The accuracy of our asymptotic expansion as an approximation to the objective function $\hat{R}_s(\gamma)$ is examined in Figure 1, which shows the function $\hat{R}_s(\gamma)$ with its first- and second-order approximations for a system of size $s = 10$.

We conclude that both approximations are remarkably accurate for this relatively small system. Near the optimizer γ_s^{opt}, the second-order approximation is almost indistinguishable from the objective function. The maximizer of the second-order approximation $\hat{R}_0(\gamma)+\hat{R}_1(\gamma)/\sqrt{s}$ is also closer to the maximizer of $\hat{R}_s(\gamma)$ than the maximizer of the first-order approximation $\hat{R}_0(\gamma)$ is. This illustrates that including higher-order correction terms in the asymptotic expansion indeed reduces the optimality gap.

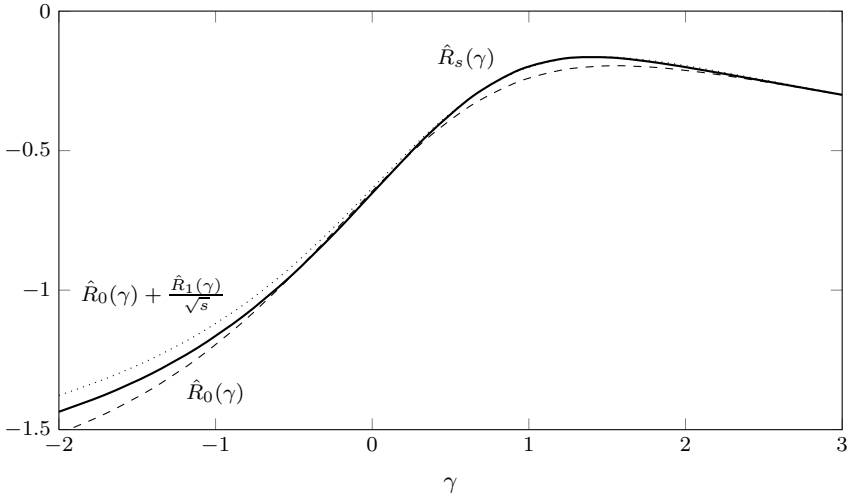

Figure 1: The function $\hat{R}_s(\gamma)$ as function of γ for $a = 0.1$, $b = 1$, and $\eta = 2$, for a system of size $s = 10$ with its first-order approximation $\hat{R}_0(\gamma)$ (dashed curve) and its second-order approximation $\hat{R}_0(\gamma) + \hat{R}_1(\gamma)/\sqrt{s}$ (dotted curve).

The absolute error $|\hat{R}_s(\gamma) - \hat{R}_0(\gamma) - \hat{R}_1(\gamma)/\sqrt{s}|$ is plotted in Figure 2 as function of s for $\gamma = 2$. A fit is provided which confirms that the asymptotic expansion is indeed accurate up to $O(1/s)$, as suggested by the asymptotic expansion in (8). The jumps in the data points are caused by rounding in the admission control, since $p_s(k - s) = \mathbb{1}[k - s \leq \lfloor \eta\sqrt{s} \rfloor]$.

We also examine optimality gaps in Figure 2, which shows first- and second-order optimality gaps. Again notice that jumps occur because of the rounding in the control policy. Furthermore, we have provided fits that confirm that the optimality gap is of order $O(1/\sqrt{s})$ when the asymptotic approximation is of order $O(1)$, and that the optimality is of order $O(1/s)$ when the asymptotic approximation is of order $O(1/\sqrt{s})$.

4.2 Joint dimensioning and admission control

We now introduce joint dimensioning and admission control, which has not been studied in the QED literature. In [7], it is proven that the admission control policy that maximizes the system's revenue rate has a threshold structure, and that for fixed $\gamma < \infty$ there exists an optimal threshold level η^{opt}. In §4.1, we

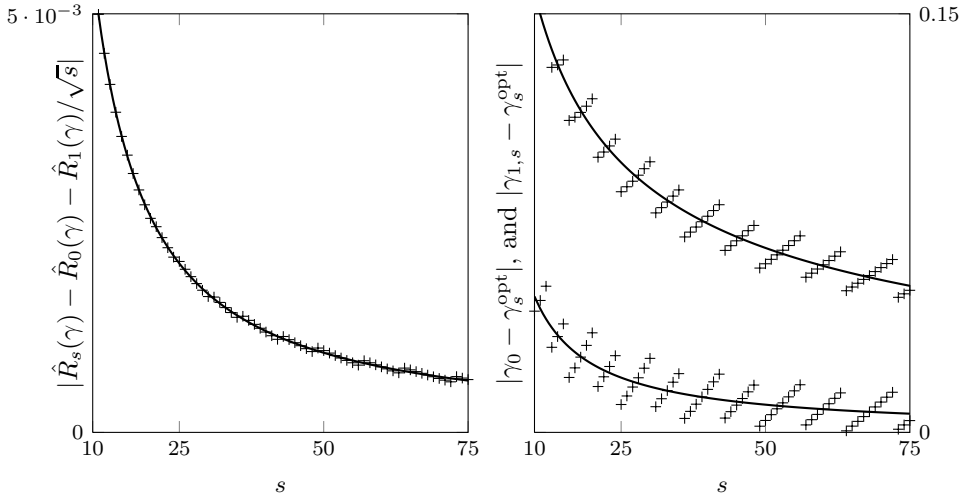

Figure 2: (left) The absolute error $|\hat{R}_s(\gamma) - \hat{R}_0(\gamma) - \hat{R}_1(\gamma)/\sqrt{s}|$ as function of s for $a = 0.1$, $b = 1$, and $\eta = 2$, for a critically scaled system with $\gamma = 2$. Plotted also is the curve $e(s) = c_1 + c_2/s^{c_3}$ with fit parameters $c_1 = 4.0 \cdot 10^{-5}$, $c_2 = 7.2 \cdot 10^{-2}$, and $c_3 = 1.1$ (continuous). (right) The top data points give the optimality gap $|\gamma_0 - \gamma_s^{\mathrm{opt}}|$, and the bottom data points $|\gamma_{1,s} - \gamma_s^{\mathrm{opt}}|$. The top fit is for $e(s) = c_1 + c_2/\sqrt{s}$ with $c_1 = 2.3 \cdot 10^{-4}$ and $c_2 = 4.9 \cdot 10^{-1}$, the bottom $e(s) = c_1 + c_2/s$ with $c_1 = -1.0 \cdot 10^{-5}$ and $c_2 = 1.3 \cdot 10^{-3}$.

fixed $\eta < \infty$ and searched for γ^{opt}. Now consider their joint optimization, that is to find

$$(\gamma^{\mathrm{opt}}, \eta^{\mathrm{opt}}) = \arg \max_{\gamma \in \mathbb{R},\, \eta \geq 0} \{\hat{R}_0(\gamma, \eta)\}. \tag{28}$$

We now illustrate numerically that joint dimensioning and admission control provides important improvements compared to optimizing over γ only as in §4.1. Table 1 displays solution pairs $(\gamma^{\mathrm{opt}}, \eta^{\mathrm{opt}})$ to the maximization problem in (28) for various ratios $r_1 = a/(a+d)$ and $r_2 = (a+d)/b$. Note that these two ratios are sufficient to describe all possible optimization problems, i.e. that occur for different a, b, and d, which can for example be seen by dividing the objective function in (10) by d and then noting that $(a+d)/d = r_2/(r_2 - r_1 r_2)$ and $b/d = 1/(r_2 - r_1 r_2)$. From Table 1, we see that the optimization problem is well-posed, since nondegenerate optimal pairs exist.

Table 2 contains the percentage improvements that can be achieved by solving the joint dimensioning and admission control problem, compared to the classical approach of finding $\gamma_\infty^{\mathrm{opt}} = \arg\max_{\gamma \geq 0} \hat{R}_s(\gamma, \eta = \infty)$. We should note that this concerns the maximization of the second-order term in (4), the leading order cannot be influenced through optimization. Note also that γ^{opt} can become negative.

4.3 Refined dimensioning

The QED refinements can also be used to derive refined dimensioning levels. The idea is that a higher-order asymptotic solution $\gamma_{1,s}$ can be expressed as a function of the lower-order asymptotic solution γ_0. To see this, consider the following representation of $\gamma_{1,s}$ in the context of dimensioning under a delay constraint as discussed in §3.2.

Theorem 2. *For sufficiently large s, the first-order solution is given as $\gamma_{1,s} = \gamma_0 + \bar{\gamma}_0$, where*

$$\bar{\gamma}_0 = \sum_{n=1}^{\infty} \frac{(-1)^n}{(n-1)!} \Big(\frac{\mathrm{d}}{\mathrm{d}\gamma}\Big)^{n-1} \Big[(A'(\gamma) + E'(\gamma)) \Big(\frac{\gamma}{A(\gamma)}\Big)^{n+1} (E(\gamma))^n\Big]_{\gamma=0}, \tag{29}$$

and for small γ the auxiliary functions are defined as $A(\gamma) = D_0(\gamma + D_0^{\leftarrow}(\varepsilon)) - \varepsilon$, and $E(\gamma) = D_1(\gamma + D_0^{\leftarrow}(\varepsilon))/\sqrt{s}$.

Proof. We mimick the standard proof of Lagrange's inversion theorem. Both A and E are analytic in the neighborhood of $\gamma = 0$ by analyticity of $B_s(\gamma)$ around $\gamma = 0$ and analyticity of $F_1(\gamma)$ in $\mathrm{Re}\{\gamma\} > \gamma^{\min}$, with $\gamma^{\min} < 0$ by assumption. Also, $A(0) = 0$, $A'(0) \neq 0$.

By taking s sufficiently large, say $s \geq s_0$, we can arrange that there is an $r > 0$ such that $A'(0) + G'(0) \neq 0$ and $|E(\gamma)/A(\gamma) \leq 1/2$ for $|\gamma| = r$ and $s \geq s_0$, while $\gamma = 0$ is the only zero of $F(\gamma)$ in $|\gamma| \leq r$. By Rouché's theorem, for any $s \geq s_0$ there is a unique $\bar{\gamma}_0$ in $|\gamma| \leq r$ such that $A(\bar{\gamma}_0) + E(\bar{\gamma}_0) = 0$. Thus the solution $\gamma_{1,s}$ near γ_0 of the equation $D_0(\gamma) + D_1(\gamma)/\sqrt{s} = \varepsilon$ is given as

Table 1: The optimal threshold, hedge pair $(\gamma^{\text{opt}}, \eta^{\text{opt}})$ for different ratios of $a/(a+d)$ and $(a+d)/b$.

$a/(a+d)$	$(a+d)/b$								
	1/5	1/4	1/3	1/2	1	2	3	4	5
0.1	(1.9, 0.4)	(1.9, 0.5)	(1.9, 0.6)	(1.8, 0.9)	(1.6, 1.6)	(1.4, 2.9)	(1.2, 3.9)	(1.1, 4.7)	(1.1, 5.5)
0.2	(1.5, 0.3)	(1.5, 0.4)	(1.5, 0.5)	(1.4, 0.7)	(1.3, 1.3)	(1.1, 2.3)	(1.0, 3.1)	(0.9, 3.8)	(0.8, 4.4)
0.3	(1.2, 0.3)	(1.2, 0.3)	(1.1, 0.4)	(1.1, 0.6)	(1.0, 1.1)	(0.8, 2.0)	(0.7, 2.6)	(0.7, 3.2)	(0.6, 3.7)
0.4	(0.9, 0.2)	(0.8, 0.3)	(0.8, 0.4)	(0.8, 0.5)	(0.7, 1.0)	(0.6, 1.7)	(0.6, 2.3)	(0.5, 2.8)	(0.5, 3.2)
0.5	(0.5, 0.2)	(0.5, 0.2)	(0.5, 0.3)	(0.5, 0.5)	(0.5, 0.8)	(0.4, 1.5)	(0.4, 2.0)	(0.3, 2.4)	(0.3, 2.8)
0.6	(0.2, 0.2)	(0.2, 0.2)	(0.2, 0.3)	(0.2, 0.4)	(0.2, 0.7)	(0.2, 1.2)	(0.2, 1.7)	(0.1, 2.0)	(0.1, 2.4)
0.7	(-0.3, 0.1)	(-0.3, 0.2)	(-0.3, 0.2)	(-0.2, 0.3)	(-0.2, 0.6)	(-0.2, 1.0)	(-0.1, 1.4)	(-0.1, 1.7)	(-0.1, 2.0)
0.8	(-0.9, 0.1)	(-0.9, 0.1)	(-0.9, 0.2)	(-0.8, 0.2)	(-0.7, 0.4)	(-0.6, 0.8)	(-0.5, 1.1)	(-0.5, 1.3)	(-0.5, 1.6)
0.9	(-2.1, 0.1)	(-2.1, 0.1)	(-2.1, 0.1)	(-2.0, 0.2)	(-1.8, 0.3)	(-1.5, 0.5)	(-1.4, 0.7)	(-1.3, 0.9)	(-1.2, 1.0)

Table 2: Pairs $(\gamma^{\text{opt}}/\gamma_\infty^{\text{opt}}, 100 \cdot (R^{\text{opt}} - R_\infty^{\text{opt}})/|R_\infty^{\text{opt}}|)$ for ratios $a/(a+d)$ and $(a+d)/b$.

$a/(a+d)$	$(a+d)/b$								
	1/5	1/4	1/3	1/2	1	2	3	4	5
0.1	(0.9, 7)	(0.9, 6)	(0.9, 5)	(0.9, 3)	(1.0, 1)	(1.0, 0)	(1.0, 0)	(1.0, 0)	(1.0, 0)
0.2	(0.8, 13)	(0.8, 11)	(0.8, 9)	(0.8, 7)	(0.9, 4)	(0.9, 2)	(0.9, 1)	(1.0, 1)	(1.0, 1)
0.3	(0.6, 18)	(0.7, 16)	(0.7, 14)	(0.7, 11)	(0.8, 7)	(0.8, 4)	(0.9, 3)	(0.9, 3)	(0.9, 2)
0.4	(0.5, 23)	(0.5, 22)	(0.5, 19)	(0.6, 16)	(0.6, 11)	(0.7, 8)	(0.7, 7)	(0.7, 6)	(0.7, 5)
0.5	(0.3, 29)	(0.3, 28)	(0.4, 25)	(0.4, 22)	(0.4, 17)	(0.5, 13)	(0.5, 11)	(0.5, 10)	(0.5, 9)
0.6	(0.1, 36)	(0.1, 34)	(0.1, 32)	(0.1, 29)	(0.2, 23)	(0.2, 19)	(0.2, 17)	(0.2, 16)	(0.2, 15)
0.7	(-0.2, 44)	(-0.2, 42)	(-0.2, 40)	(-0.2, 37)	(-0.2, 31)	(-0.2, 27)	(-0.2, 25)	(-0.2, 23)	(-0.2, 22)
0.8	(-0.6, 54)	(-0.6, 52)	(-0.7, 50)	(-0.7, 47)	(-0.8, 42)	(-0.9, 38)	(-0.9, 36)	(-0.9, 34)	(-1.0, 33)
0.9	(-1.5, 67)	(-1.5, 65)	(-1.6, 64)	(-1.8, 62)	(-2.0, 58)	(-2.3, 54)	(-2.5, 52)	(-2.6, 51)	(-2.6, 50)

$\gamma_{1,s} = \gamma_0 + \bar{\gamma}_0$. By Cauchy's theorem, there is for $\bar{\gamma}_0$ the integral representation

$$\bar{\gamma}_0 = \frac{1}{2\pi \mathrm{i}} \int_{|\gamma|=r} \frac{\gamma(A'(\gamma) + E'(\gamma))}{A(\gamma) + E(\gamma)} \, \mathrm{d}\gamma. \tag{30}$$

Since we have $0 \neq |A(\gamma)| \geq 2|E(\gamma)|$ on $|\gamma| = r$, it follows that

$$\bar{\gamma}_0 = \sum_{n=0}^{\infty} \frac{(-1)^n}{2\pi \mathrm{i}} \int_{|\gamma|=r} \gamma(A'(\gamma) + E'(\gamma)) \frac{(E(\gamma))^n}{(A(\gamma))^{n+1}} \, \mathrm{d}\gamma. \tag{31}$$

Due to analyticity of $\gamma/A(\gamma)$, the term with $n = 0$ vanishes. For $n = 1, 2, \ldots$, we furthermore have that

$$\begin{aligned} &\frac{1}{2\pi \mathrm{i}} \int_{|\gamma|=r} \gamma(A'(\gamma) + E'(\gamma)) \frac{(E(\gamma))^n}{(A(\gamma))^{n+1}} \, \mathrm{d}\gamma \\ &= \frac{1}{2\pi \mathrm{i}} \int_{|\gamma|=r} \frac{1}{\gamma^n} (A'(\gamma) + E'(\gamma)) \Big(\frac{\gamma}{A(\gamma)} \Big)^{n+1} (E(\gamma))^n \, \mathrm{d}\gamma \\ &= \frac{1}{(n-1)!} \Big(\frac{\mathrm{d}}{\mathrm{d}\gamma} \Big)^{n-1} \Big[(A'(\gamma) + E'(\gamma)) \Big(\frac{\gamma}{A(\gamma)} \Big)^{n+1} (E(\gamma))^n \Big]_{\gamma=0}. \end{aligned} \tag{32}$$

This is the result in (29), and concludes the proof. □

When using the first two terms in (29), we get for $\bar{\gamma}_0$ the approximation

$$-\frac{E}{A'} + \frac{EE'}{(A')^2} - \frac{A''E^2}{2(A')^3} - \frac{3A''E'E^2}{(A')^4} + \frac{E''E^2}{(A')^3} + \frac{2E(E')^2}{(A')^3}, \tag{33}$$

with all functions evaluated at $\gamma = 0$. The first term gives an $s^{-1/2}$-correction of γ_0, the second term gives an s^{-1}-contribution, and other terms give contributions of $O(s^{-3/2})$ or smaller. Summarizing, we thus have that

$$\gamma_{1,s} = \gamma_0 + \bar{\gamma}_0 = \gamma_0 - \frac{1}{\sqrt{s}} \frac{D_1(\gamma_0)}{{D_0}'(\gamma_0)} + O\Big(\frac{1}{s}\Big). \tag{34}$$

References

[1] S. C. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, Feb. 2004.

[2] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, June 1981.

[3] D. L. Jagerman. Some properties of the Erlang loss function. *Bell System Technical Journal*, 53(3):525–551, Mar. 1974.

[4] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and J. Sanders. Scaled control in the QED regime. *Performance Evaluation*, 70(10):750–769, Oct. 2013.

[5] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and B. Zwart. Refining square-root safety staffing by expanding Erlang C. *Operations Research*, 59(6):1512–1522, Dec. 2011.

[6] R. S. Randhawa. The optimality gap of asymptotically-derived prescriptions with applications to queueing systems. *arXiv:1210.2706 [math]*, Oct. 2012.

[7] J. Sanders, S. C. Borst, A. J. E. M. Janssen, and J. van Leeuwaarden. Optimal admission control for many-server systems with QED-driven revenues. *ArXiv e-prints*, nov 2014. Submitted, preprint available on arXiv.

[8] B. Zhang, J. S. H. van Leeuwaarden, and B. Zwart. Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Operations Research*, 60(2):461–474, 2012.

Acknowledgments

This research was financially supported by The Netherlands Organization for Scientific Research (NWO) in the framework of the TOP-GO program and by an ERC Starting Grant.

A Higher-order terms in asymptotic expansions

We now provide closed-form expressions for the asymptotic expansion in (8). We drop the dependence on γ for notational convenience, and prove the following result.

Theorem 3. *As* $s \to \infty$, $(R_s - n_s)/q_s = R_0 + R_1/\sqrt{s} + O(1/s)$, *where*

$$R_0 = \frac{W_0^{\mathrm{L}} + W_0^{\mathrm{R}}}{B_0 + F_0}, \quad R_1 = \frac{W_1^{\mathrm{L}} + W_1^{\mathrm{R}}}{B_0 + F_0} - \frac{(W_0^{\mathrm{L}} + W_0^{\mathrm{R}})(B_1 + F_1)}{(B_0 + F_0)^2}, \tag{35}$$

and

$$\begin{aligned} W_0^{\mathrm{L}} &= \int_{-\infty}^{0} r(x) \mathrm{e}^{-\frac{1}{2}x^2 - \gamma x} \, \mathrm{d}x, \\ W_1^{\mathrm{L}} &= \tfrac{1}{2} \int_{-\infty}^{0} (\tfrac{1}{3}x^3 - (1+\gamma^2)x) r(x) \mathrm{e}^{-\frac{1}{2}x^2 - \gamma x} \, \mathrm{d}x + r(0), \\ W_0^{\mathrm{R}} &= \int_{0}^{\infty} r(x) f(x) \mathrm{e}^{-\gamma x} \, \mathrm{d}x, \\ W_1^{\mathrm{R}} &= -\tfrac{1}{2}\gamma^2 \int_{0}^{\infty} x r(x) f(x) \mathrm{e}^{-\gamma x} \, \mathrm{d}x - \tfrac{1}{2} r(0) f(0), \end{aligned} \tag{36}$$

as well as

$$B_0 = \frac{\Phi(\gamma)}{\phi(\gamma)}, \;\; B_1 = \tfrac{1}{3}\Big(2 + \gamma^2 + \gamma^3 \frac{\Phi(\gamma)}{\phi(\gamma)}\Big), \tag{37}$$
$$F_0 = \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x, \;\; F_1 = -\tfrac{1}{2}\gamma^2 \int_0^\infty x f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x - \tfrac{1}{2}f(0).$$

Proof. Note after substituting (2) into $R_s = \sum_{k=0}^\infty r_s(k)\pi_s(k)$, that asymptotically

$$\frac{R_s - n_s}{q_s} = \frac{\sum_{k=0}^s r\big(\frac{k-s}{\sqrt{s}}\big)\frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!}\sum_{k=s+1}^\infty r\big(\frac{k-s}{\sqrt{s}}\big)\rho^{k-s} f\big(\frac{k-s}{\sqrt{s}}\big)}{\sum_{k=0}^s \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!}\sum_{k=s+1}^\infty \rho^{k-s} f\big(\frac{k-s}{\sqrt{s}}\big)}. \tag{38}$$

Dividing by the factor $(s\rho)^s/s!$, we obtain the form

$$\frac{R_s - n_s}{q_s} = \frac{W_s^{\mathrm{L}} + W_s^{\mathrm{R}}}{B_s^{-1} + F_s}, \tag{39}$$

where we have introduced notation for the Erlang B formula, $B_s(\rho) = ((s\rho)^s/s!)/(\sum_{k=0}^s (s\rho)^k/k!)$, and we have defined $F_s = \sum_{n=0}^\infty \rho^{n+1} f((n+1)/\sqrt{s})$, $W_s^{\mathrm{L}} = \sum_{k=0}^s r((k-s)/\sqrt{s})(s!(s\rho)^{k-s})/k!$, and $W_s^{\mathrm{R}} = \sum_{n=0}^\infty r((n+1)/\sqrt{s})\rho^{n+1} f((n+1)/\sqrt{s})$. In [4, 7], it is proven using Jagerman's asymptotic expansions [3] for Erlang B's formula that asymptotically,

$$W_s^{\mathrm{L}} = \sqrt{s}W_0^{\mathrm{L}} + W_1^{\mathrm{L}} + O\Big(\frac{1}{\sqrt{s}}\Big), \;\; W_s^{\mathrm{R}} = \sqrt{s}W_0^{\mathrm{R}} + W_1^{\mathrm{R}} + O\Big(\frac{1}{\sqrt{s}}\Big),$$
$$B_s^{-1} = \sqrt{s}B_0 + B_1 + O\Big(\frac{1}{\sqrt{s}}\Big), \;\; F_s = \sqrt{s}F_0 + F_1 + O\Big(\frac{1}{\sqrt{s}}\Big),$$

with the coefficients as given in (36) and (37). After substituting these asymptotic expansions into (39), we obtain

$$\frac{R_s - n_s}{q_s} = \frac{W_0^{\mathrm{L}} + W_0^{\mathrm{R}}}{B_0 + F_0} \cdot \frac{1 + \frac{1}{\sqrt{s}}(W_1^{\mathrm{L}} + W_1^{\mathrm{R}})/(W_0^{\mathrm{L}} + W_0^{\mathrm{R}}) + O(\frac{1}{s})}{1 + \frac{1}{\sqrt{s}}(B_1 + F_1)/(B_0 + F_0) + O(\frac{1}{s})}.$$

By then utilizing the Taylor expansion $1/(1+x) = 1 - x + O(x^2)$, we obtain the result. ∎

Delay probability The delay probability $D_s = \sum_{k=s}^\infty \pi_s(k)$ can be represented by $r_s(k) = \mathbb{1}[k \geq s]$, recall (3). This corresponds asymptotically to $r(x) = \mathbb{1}[x \geq 0]$. introduce for convenience $\mathcal{L} = \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x$. Then, $W_0^{\mathrm{L}} = W_1^{\mathrm{L}} = 0$, $W_0^{\mathrm{R}} = \mathcal{L}$, $W_1^{\mathrm{R}} = \frac{1}{2}\gamma^2\mathcal{L}'$, $F_0 = \mathcal{L}$, and $F_1 = \frac{1}{2}\gamma^2\mathcal{L}' - \frac{1}{2}$. It follows that $D_0 = \mathcal{L}/(\Phi(\gamma)/\phi(\gamma) + \mathcal{L})$ and

$$D_1 = \frac{\frac{1}{2}\gamma^2\mathcal{L}'}{\frac{\Phi(\gamma)}{\phi(\gamma)} + \mathcal{L}} - \frac{\mathcal{L}(\frac{1}{3}\big(2 + \gamma^2 + \gamma^3\frac{\Phi(\gamma)}{\phi(\gamma)}\big) + \frac{1}{2}\gamma^2\mathcal{L}' - \frac{1}{2})}{(\frac{\Phi(\gamma)}{\phi(\gamma)} + \mathcal{L})^2}. \tag{40}$$

Queue length The mean queue length $Q_s = \sum_{k=s}^{\infty}(k-s)\pi_s(k)$ can be represented by $r_s(k) = (k-s)\mathbb{1}[k \geq s]$. Scaling so that $Q_s/\sqrt{s} = \sum_{k=s}^{\infty}((k-s)/\sqrt{s})\pi_s(k)$, we see that the revenue structure can asymptotically be related to the revenue profile $r(x) = x\mathbb{1}[x \geq 0]$. Therefore, $W_0^{\mathrm{L}} = W_1^{\mathrm{L}} = 0$, $W_0^{\mathrm{R}} = -\mathcal{L}'$, $W_1^{\mathrm{R}} = -\frac{1}{2}\gamma^2\mathcal{L}''$, $F_0 = \mathcal{L}$, and $F_1 = \frac{1}{2}\gamma^2\mathcal{L}' - \frac{1}{2}$. Thus $Q_s/\sqrt{s} = Q_0 + Q_1/\sqrt{s} + O(1/s)$ with $Q_0 = -\mathcal{L}'/(\Phi(\gamma)/\phi(\gamma) + \mathcal{L})$ and

$$Q_1 = -\frac{\frac{1}{2}\gamma^2\mathcal{L}'}{\frac{\Phi(\gamma)}{\phi(\gamma)} + \mathcal{L}} + \frac{\mathcal{L}'(\frac{1}{3}\big(2 + \gamma^2 + \gamma^3\frac{\Phi(\gamma)}{\phi(\gamma)}\big) + \frac{1}{2}\gamma^2\mathcal{L}' - \frac{1}{2})}{(\frac{\Phi(\gamma)}{\phi(\gamma)} + \mathcal{L})^2}. \tag{41}$$