http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 7, No. 2, December 2017, pp. 315–341
ISSN 1946-5238 (print), ISSN 1946-5238 (online)

Optimal Admission Control for Many-Server Systems with QED-Driven Revenues

Jaron Sanders,[a] S. C. Borst,[a] A. J. E. M. Janssen,[a] J. S. H. van Leeuwaarden[a]

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

Contact: mail@jaronsanders.nl (JS); s.c.borst@tue.nl (SCB); a.j.e.m.janssen@tue.nl (AJEMJ); j.s.h.v.leeuwaarden@tue.nl (JSHvL)

Received: September 2015
Accepted: September 2017

MSC2010 Subject Classification: 90B22; 60K25; 93E03; 65K10; 34E05

https://doi.org/10.1287/stsy.2017.0004

Abstract. We consider Markovian many-server systems with admission control operating in a Quality-and-Efficiency-Driven (QED) regime, where the relative utilization approaches unity while the number of servers grows large, providing natural Economies-of-Scale. In order to determine the optimal admission control policy, we adopt a revenue maximization framework, and suppose that the revenue rate attains a maximum when no customers are waiting and no servers are idling. When the revenue function scales properly with the system size, we show that a nondegenerate optimization problem arises in the limit. Detailed analysis demonstrates that the revenue is maximized by nontrivial policies that bar customers from entering when the queue length exceeds a certain threshold of the order of the typical square-root level variation in the system occupancy. We identify a fundamental equation characterizing the optimal threshold, which we extensively leverage to provide broadly applicable upper/lower bounds for the optimal threshold, establish its monotonicity, and examine its asymptotic behavior, all for general revenue structures. For linear and exponential revenue structures, we present explicit expressions for the optimal threshold.

History: Former designation of this paper was SSY-2015-202.

Funding: 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.

Keywords: queueing • Quality-and-Efficiency-Driven (QED) regime • asymptotic analysis • optimal control • threshold control • variational calculus

1. Introduction

Large-scale systems that operate in the Quality-and-Efficiency Driven (QED) regime dwarf the usual trade-off between high system utilization and short waiting times. In order to achieve these dual goals, the system is scaled so as to approach full utilization, while the number of servers grows simultaneously large, rendering crucial Economies-of-Scale. Specifically, for a Markovian many-server system with Poisson arrival rate λ, exponential unit-mean service times and s servers, the load $\rho = \lambda/s$ is driven to unity in the QED regime in accordance with

$$(1-\rho)\sqrt{s} \to \gamma, \quad s \to \infty, \tag{1}$$

for some fixed parameter $\gamma \in \mathbb{R}_+$. As s grows large, the stationary probability of delay then tends to a limit, say $g(\gamma)$, which may take any value in $(0,1)$, depending on the parameter γ. Since the *conditional* queue length distribution is geometric with mean $\rho/(1-\rho) \approx \sqrt{s}/\gamma$, it follows that the stationary mean number of waiting customers scales as $g(\gamma)\sqrt{s}/\gamma$. Little's law then in turn implies that the mean stationary waiting time of a customer falls off as $g(\gamma)/(\gamma\sqrt{s})$.

The QED scaling behavior also manifests itself in process level limits, where the evolution of the system occupancy, properly centered around s and normalized by $\sqrt{s}$, converges to a diffusion process as $s \to \infty$, which again is fully characterized by the single parameter γ. This reflects that the system state typically hovers around the full-occupancy level s, with natural fluctuations of the order $\sqrt{s}$.

The QED scaling laws provide a powerful framework for system dimensioning, i.e., matching the service capacity and traffic demand so as to achieve a certain target performance level or optimize a certain cost metric. Suppose, for instance, that the objective is to find the number of servers s for a given arrival rate λ (or

equivalently, determine what arrival rate λ can be supported with a given number s of servers) such that a target delay probability $\epsilon \in (0,1)$ is attained. The above-mentioned convergence results for the delay probability then provide the natural guideline to match the service capacity and traffic volume in accordance with $\lambda = s - \gamma_\epsilon \sqrt{s}$, where the value of γ_ϵ is such that $g(\gamma_\epsilon) = \epsilon$.

As an alternative objective, imagine we aim to strike a balance between the expenses incurred for staffing servers and the dissatisfaction experienced by waiting customers. Specifically, suppose a (salary) cost c is associated with each server per unit of time and a (possibly fictitious) holding charge h is imposed for every waiting customer per unit of time. Writing $\lambda = s - \gamma\sqrt{s}$ in accordance with (1), and recalling that the mean number of waiting customers scales as $g(\gamma)\sqrt{s}/\gamma$, we find that the total operating cost per time unit scales as

$$cs + h\frac{g(\gamma)\sqrt{s}}{\gamma} = \lambda c + c\gamma\sqrt{s} + h\frac{g(\gamma)\sqrt{s}}{\gamma} = \lambda c + \left(c\gamma + h\frac{g(\gamma)}{\gamma}\right)\sqrt{s}.$$

This then suggests to set the number of servers in accordance with $s = \lambda + \gamma_{c,h}\sqrt{s}$, where $\gamma_{c,h} = \arg\min_{\gamma>0}(c\gamma + hg(\gamma)/\gamma)$ in order to minimize the total operating cost per time unit. Exploiting the powerful QED limit theorems, such convenient capacity sizing rules can in fact be shown to achieve optimality in some suitable asymptotic sense.

As illustrated by the above two paragraphs, the QED scaling laws can be leveraged for the purpose of dimensioning, with the objective to balance the service capacity and traffic demand so as to achieve a certain target performance standard or optimize a certain cost criterion. A critical assumption, however, is that all customers are admitted into the system and eventually served, which may in fact not be consistent with the relevant objective functions in the dimensioning, let alone be optimal in any sense.

Motivated by the latter observation, we focus in the present paper on the optimal admission control problem for a given performance or cost criterion. Admission control acts on an operational time scale, with decisions occurring continuously whenever customers arrive, as opposed to capacity planning decisions which tend to involve longer time scales. Indeed, we assume that the service capacity and traffic volume are given, and balanced in accordance with (1), but do allow for the value of γ to be negative, since admission control provides a mechanism to deal with overload conditions. While a negative value of γ may not be a plausible outcome of a deliberate optimization process, in practice an overload of that order might well result from typical forecast errors.

We formulate the admission control problem in a revenue maximization framework, and suppose that revenue is generated at rate $r_s(k)$ when the system occupancy is k. As noted above, both from a customer satisfaction perspective and a system efficiency viewpoint, the ideal operating condition for the system is around the full occupancy level s, where no customers are waiting and no servers are idling. Hence we assume that the function $r_s(k)$ is unimodal, increasing in k for $k \le s$ and decreasing in k for $k \ge s$, thus attaining its maximum at $k = s$.

We consider probabilistic control policies, which admit arriving customers with probability $p_s(k-s)$ when the system occupancy is k, independent of any prior admission decisions. It is obviously advantageous to admit customers as long as free servers are available, since it will not lead to any wait and drive the system closer to the ideal operating point s, boosting the instantaneous revenue rate. Thus we stipulate that $p_s(k-s) = 1$ for all $k < s$.

For $k \ge s$, it is far less evident whether to admit customers or not. Admitting a customer will then result in a wait and move the system away from the ideal operating point, reducing the instantaneous revenue rate. On the other hand, admitting a customer may prevent the system occupancy from falling below the ideal operating point in the future. The potential long-term gain may outweigh the adverse near-term effect, so there may be a net benefit, but the incentive weakens as the queue grows. The fundamental challenge in the design of admission control policies is to find exactly the point where the marginal utility reaches zero, so as to strike the optimal balance between the conflicting near-term and longer-term considerations.

Since the service capacity and traffic volume are governed by (1), the QED scaling laws imply that, at least for $\gamma > 0$ and without any admission control, the system occupancy varies around the ideal operating point s, with typical deviations of the order $\sqrt{s}$. It is therefore natural to suppose that the revenue rates and admission probabilities scale in a consistent manner, and in the limit behave as functions of the properly centered and normalized state variable $(k-s)/\sqrt{s}$. Specifically, we assume that the revenue rates satisfy the scaling condition

$$\frac{r_s(k) - n_s}{q_s} \to r\left(\frac{k-s}{\sqrt{s}}\right), \quad s \to \infty, \tag{2}$$

with n_s a nominal revenue rate attained at the ideal operating point, q_s a scaling coefficient, and r a unimodal function, which represents the scaled reduction in revenue rate associated with deviations from the optimal operating point s. For example, with $[x]^+ = \max\{0, x\}$, any revenue structure of the form

$$r_s(k) = n_s - \alpha^-([s-k]^+)^{\beta^-} - \alpha^+([k-s]^+)^{\beta^+}$$

satisfies (2) when $q_s = s^{\max\{\beta^-,\beta^+\}/2}$, in which case

$$r(x) = -\alpha^-([-x]^+)^{\beta^-}\mathbb{1}[\beta^- \geq \beta^+] - \alpha^+([x]^+)^{\beta^+}\mathbb{1}[\beta^- \leq \beta^+].$$

Note that these revenue structures impose polynomial penalties on deviations from the ideal operating point. Similar to (2), we assume that the admission probabilities satisfy a scaling condition, namely

$$p_s(0)\cdots p_s(k-s) = f\left(\frac{k-s}{\sqrt{s}}\right), \quad k \geq s, \tag{3}$$

with f a non-increasing function and $f(0) = 1$. In particular, we allow for $f(x) = \mathbb{1}[0 \leq x < \eta]$, which corresponds to an admission threshold control $p_s(k-s) = \mathbb{1}[k-s \leq \lfloor \eta\sqrt{s} \rfloor]$.

In Section 2 we discuss the fact that the optimal admission policy is indeed such a threshold control, with the value of η asymptotically being determined by the function r, which we later prove in Section 4. The optimality of a threshold policy may not come as a surprise, and can in fact be established in the pre-limit ($s < \infty$). However, the pre-limit optimality proof only yields the structural property, and does not furnish any characterization of how the optimal threshold depends on the system characteristics or provide any computational procedure for actually obtaining the optimal value. In contrast, our asymptotic framework (as $s \to \infty$) produces a specific equation characterizing the optimal threshold value, which does offer explicit insight in the dependence on the key system parameters and can serve as a basis for an efficient numerical computation or even a closed-form expression in certain cases. This is particularly valuable for large-scale systems where a brute-force enumerative search procedure may prove prohibitive.

Let us finally discuss the precise form of the revenue rates that serve as the objective function that needs to be maximized by the optimal threshold. We will mostly focus on the average *system-governed* revenue rate defined as

$$R_s(\{p_s(k)\}_{k\geq 0}) = \sum_{k=0}^{\infty} r_s(k)\pi_s(k). \tag{4}$$

From the system's perspective, this means that the revenue is simply governed by the state-dependent revenue rate $r_s(k)$ weighed according to the stationary distribution, with $\pi_s(k)$ denoting the stationary probability of state k.

An alternative would be to consider the *customer reward* rate

$$\hat{R}_s(\{p_s(k)\}_{k\geq 0}) = \lambda \sum_{k=0}^{\infty} \hat{r}_s(k)p_s(k-s)\pi_s(k). \tag{5}$$

Here, $\hat{r}_s(k)$ can be interpreted as the state-dependent reward when admitting a customer in state k, and since this happens with probability $p_s(k)$ at intensity λ, we obtain (5). While this paper primarily focuses on (4), we show in Section 2.3 that there is an intimate connection with (5); a system-governed reward structure $\{r_s(k)\}_{k\in\mathbb{N}_0}$ can be translated into a customer reward structure $\{\hat{r}_s(k)\}_{k\in\mathbb{N}_0}$, and vice versa.

1.1. Contributions and Related Literature

A diverse range of control problems have been considered in the queueing literature, and we refer the reader to Lippman (1975), Stidham (1985), Kushner and Dupuis (2001), Meyn (2008), Çil et al. (2009) for background. Threshold control policies are found to be optimal in a variety of contexts in such as De Waal (1990), Chen and Frank (2001), Bekker and Borst (2006), and many (implicit) characterizations of optimal threshold values have been obtained in Naor (1969), Yildirim and Hasenbein (2010), and Borgs et al. (2014). For (single-server) queues in a conventional heavy-traffic regime, optimality of threshold control policies has been established by studying limiting diffusion control problems in Ghosh and Weerasinghe (2007), Ward and Kumar (2008), and Ghosh and Weerasinghe (2010).

The analysis of control problems in the QED regime has mostly focused on routing and scheduling, see Atar et al. (2004), Atar (2005a, b), Atar et al. (2006), and Gurvich and Whitt (2009). Threshold policies in the context of many-server systems in the QED regime have been considered in Armony and Maglaras (2004), Massey and

Wallace (2005), Whitt (2005), and Whitt (2004). General admission control, however, has only received limited attention in the QED regime, see for instance Koçağa and Ward (2010), Weerasinghe and Mandelbaum (2013). These studies specifically account for abandonments, which create a trade-off between the rejection of a new arrival and the risk of that arrival later abandoning without receiving service, with the associated costly increase of server idleness.

In the present paper we address the optimal admission control problem from a revenue maximization perspective. Diverse considerations of revenue maximization problems as a function of pricing, capacity constraints and service differentiation have shown that the QED regime and other heavy-traffic regimes emerge naturally as the optimal regimes to operate in Maglaras and Zeevi (2003, 2005), Kumar and Randhawa (2010), and Maglaras et al. (2017). This implies in particular that the QED regime provides an appropriate framework to address further questions of revenue maximization. We therefore start from the premise that the system is operated in the QED regime, and then diverge from Maglaras and Zeevi (2003, 2005), Kumar and Randhawa (2010), and Maglaras et al. (2017) by assuming that the revenue function $r_s(k)$ satisfies the scaling condition (2). Building on the recent work in Janssen et al. (2013) allows us to show that additional nondegenerate optimization problems arise in the QED limit. This suggests that the QED regime is the optimal operating regime in a broader range of scenarios when underlying pricing and revenue structures are scaled appropriately, complementing the conclusions in Maglaras and Zeevi (2003, 2005), Kumar and Randhawa (2010), and Maglaras et al. (2017). Lastly, our analysis shows that nontrivial threshold control policies are optimal when $r_s(k)$ peaks around the ideal operating point $k \approx s$.

In Section 2 we present a fundamental equation which implicitly determines the asymptotically optimal threshold. The subsequent analysis of this equation in Section 3 yields valuable insight into the dependence of the optimal threshold on the revenue structure, and provides a basis for an efficient numerical scheme. Closed-form expressions for the optimal threshold can only be derived when considering specific revenue structures.

We will, for example, show that for *linearly decreasing* revenue rates, the optimal threshold can be (explicitly) expressed in terms of the Lambert W function (Corless et al. 1996). We note that a linearly decreasing revenue structure has also been considered in Borgs et al. (2014) for determining the optimal threshold k^{opt} in an $M/M/s/k$ system, and there also, k^{opt} is expressed in terms of the Lambert W function. Besides assuming a static revenue and finite threshold k, a crucial difference between (Borgs et al. 2014) and this paper is that our revenue structure scales as in (2), so that the threshold k is suitable for the QED regime. Our work thus extends (Borgs et al. 2014), both in terms of scalable and more general revenue structures.

In terms of mathematical techniques, we use Euler–Maclaurin (EM) summation (Olver 2010) and asymptotic results by Jagerman (1974) to analyze the asymptotic behavior of (4) as $s \to \infty$. This approach was used recently for many-server systems with admission control in the QED regime (Janssen et al. 2013), and is now extended by incorporating suitably scaled revenue structures in Section 2. These ingredients then pave the way to determine the optimal admission control policy in the QED regime in Section 3. In Section 4, we use Hilbert-space theory from analysis, and techniques from variational calculus, to prove the existence of optimal control policies, and to establish that control policies with an admission threshold which scales with the natural $\sqrt{s}$ order of variation are optimal in the QED regime.

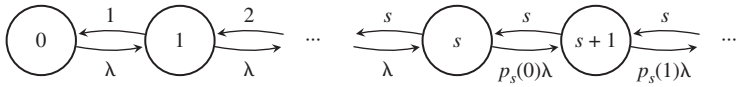
2. Revenue Maximization Framework

We now develop an asymptotic framework for determining an optimal admission control policy for a given performance or cost criterion. In Section 2.1 we describe the basic model for the system dynamics, which is an extension of the classical $M/M/s$ system. Specifically, the model incorporates admission control and is augmented with a revenue structure, which describes the revenue rate as a function of the system occupancy. Adopting this flexible apparatus, the problem of finding an optimal admission control policy is formulated in terms of a revenue maximization objective.

2.1. Markovian Many-Server Systems with Admission Control

Consider a system with s parallel servers where customers arrive according to a Poisson process with rate λ. Customers require exponentially distributed service times with unit mean. A customer that finds upon arrival k customers in the system is taken into service immediately if $k < s$, or may join a queue of waiting customers if $k \geq s$. If all servers are occupied, a newly arriving customer is admitted into the system with probability $p_s(k-s)$, and denied access otherwise. We refer to the probabilities $\{p_s(k)\}_{k\geq 0}$ as the *admission control policy*. If we denote the number of customers in the system at time t by $Q_s(t)$, and make the usual independence assumptions, then

Figure 1. Transition diagram of the process $\{Q_s(t)\}_{t\geq 0}$.



$\{Q_s(t)\}_{t\geq 0}$ constitutes a Markov process (see Figure 1 for its transition diagram). The stationary distribution $\pi_s(k) = \lim_{t\to\infty} \mathbb{P}[Q_s(t) = k]$ is given by

$$\pi_s(k) = \begin{cases} \pi_s(0)\dfrac{(s\rho)^k}{k!}, & k = 1, 2, \ldots, s, \\ \pi_s(0)\dfrac{s^s\rho^k}{s!}\displaystyle\prod_{i=0}^{k-s-1} p_s(i), & k = s+1, s+2, \ldots, \end{cases} \tag{6}$$

with

$$\rho = \frac{\lambda}{s}, \qquad \pi_s(0) = \left(\sum_{k=0}^{s}\frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!}F_s(\rho)\right)^{-1} \tag{7}$$

and

$$F_s(\rho) = \sum_{n=0}^{\infty} p_s(0)\cdots p_s(n)\rho^{n+1}. \tag{8}$$

From (6)–(8), we see that the stationary distribution 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 $F_s(\rho) < \infty$ (Janssen et al. 2013), which always holds in case $\rho < 1$.

With k customers in the system, we assume that the system generates revenue at rate $r_s(k) \in \mathbb{R}$. We call $\{r_s(k)\}_{k\geq 0}$ the *revenue structure*. Our objective is to find an admission control policy in terms of the probabilities $\{p_s(k)\}_{k\geq 0}$ that maximizes the average stationary revenue rate, i.e.,

$$\begin{aligned} \text{to maximize} \quad & R_s(\{p_s(k)\}_{k\geq 0}), \quad \text{over} \quad \{p_s(k)\}_{k\geq 0}, \\ \text{subject to} \quad & 0 \leq p_s(k) \leq 1, \quad k \in \mathbb{N}_0, \text{ and } F_s(\rho) < \infty. \end{aligned} \tag{9}$$

2.2. QED-Driven Asymptotic Optimization Framework

We now construct an asymptotic optimization framework where the limit laws of the Quality-and-Efficiency-Driven (QED) regime can be leveraged by imposing suitable assumptions on the admission control policy and revenue structure. In order for the system to operate in the QED regime, we couple the arrival rate to the number of servers as

$$\lambda = s - \gamma\sqrt{s}, \quad \gamma \in \mathbb{R}. \tag{10}$$

For the admission control policy we assume the form in (3), with f either a nonincreasing, bounded, and twice differentiable continuous function, or a step function, which we will refer to as the *asymptotic admission control profile*. We also assume the revenue structure has the scaling property (2), with r a piecewise bounded, twice differentiable continuous function with bounded derivatives. We will refer to r as the *asymptotic revenue profile*. These assumptions allow us to establish Proposition 1 by considering the stationary average revenue rate $R_s(\{p_s(k)\}_{k\geq 0})$ as a Riemann sum and using Euler–Maclaurin (EM) summation to identify its limiting integral expression, the proof of which can be found in Appendix A. Let $\phi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$ and $\Phi(x) = \int_{-\infty}^{x}\phi(u)\,\mathrm{d}u$ denote the probability density function and cumulative distribution function of the standard normal distribution, respectively.

Proposition 1. *If $r^{(i)}$ is continuous and bounded for $i = 0, 1, 2$, and either* (i) *f is smooth, and $(f(x)\exp(-\gamma x))^{(i)}$ is exponentially small as $x \to \infty$ for $i = 0, 1, 2$, or* (ii) *$f(x) = \mathbb{1}[0 \leq x < \eta]$ with a fixed, finite $\eta > 0$, then*

$$\lim_{s\to\infty}\frac{R_s(\{p_s(k)\}_{k\geq 0}) - n_s}{q_s} = R(f),$$

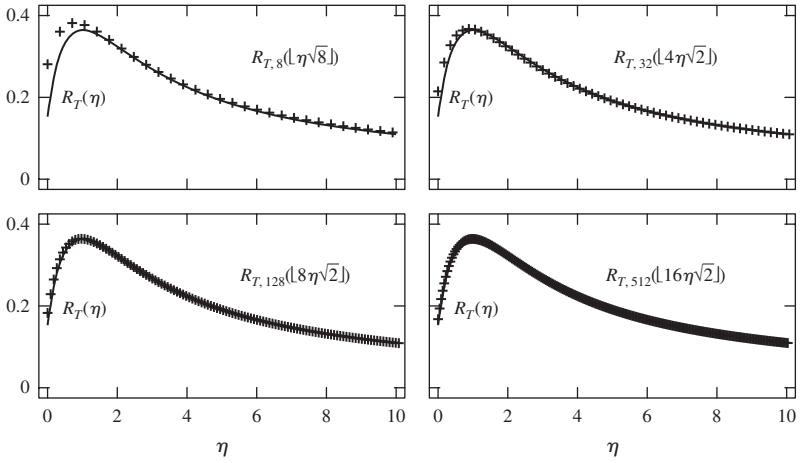
where in case (i)

$$R(f) = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{(1/2)x^2-\gamma x}\,\mathrm{d}x + \int_{0}^{\infty} r(x)f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\Phi(\gamma)/\phi(\gamma) + \int_{0}^{\infty} f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}, \tag{11}$$

and in case (ii)

$$R(\mathbb{1}[0 \leq x < \eta]) = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{-(1/2)x^2-\gamma x}\,\mathrm{d}x + \int_{0}^{\eta} r(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\Phi(\gamma)/\phi(\gamma) + (1-\mathrm{e}^{-\gamma\eta})/\gamma}, \quad \eta \geq 0. \tag{12}$$

Figure 2. $R_{\mathrm{T},s}(\lfloor \eta\sqrt{s} \rfloor)$ and $R_{\mathrm{T}}(\eta)$ for $s = 8, 32, 128, 256$ servers.



Because of the importance of the threshold policy, we will henceforth use the short-hand notations $R_{\mathrm{T},s}(\tau) = R_s(\{\mathbb{1}[k \le s + \tau]\}_{k\ge 0})$ and $R_{\mathrm{T}}(\eta) = R(\mathbb{1}[0 \le x < \eta])$ to indicate threshold policies.

Example 1 (Exponential Revenue)**.** Consider a revenue structure $r_s(k) = \exp(b(k-s)/\sqrt{s})$ for $k < s$ and $r_s(k) = \exp(-d(k-s)/\sqrt{s})$ for $k \ge s$, and with $b, d > 0$. Taking $n_s = 0$ and $q_s = 1$, the asymptotic revenue profile is $r(x) = \exp(bx)$ for $x < 0$ and $r(x) = \exp(-dx)$ for $x \ge 0$, so that according to Proposition 1 for threshold policies,

$$\lim_{s\to\infty} R_{\mathrm{T},s}(\lfloor \eta\sqrt{s} \rfloor) = R_{\mathrm{T}}(\eta) = \frac{\Phi(\gamma - b)/(\phi(\gamma - b)) + (1 - \mathrm{e}^{-(d+\gamma)\eta})/(d+\gamma)}{\Phi(\gamma)/\phi(\gamma) + (1 - \mathrm{e}^{-\gamma\eta})/\gamma}.$$

Figure 2 plots $R_{\mathrm{T},s}(\lfloor \eta\sqrt{s} \rfloor)$ for a finite system with $s = 8, 32, 128, 256$ servers, respectively, together with its limit $R_{\mathrm{T}}(\eta)$. Here, we set $b = 5$, $d = 1$, and $\gamma = 0.01$. Note that the approximation $R_{\mathrm{T},s}(\lfloor \eta\sqrt{s} \rfloor) \approx R_{\mathrm{T}}(\eta)$ is remarkably accurate, even for relatively small systems, an observation which in fact seems to hold for most revenue structures and parameter choices. For this particular revenue structure, we see that the average revenue rate peaks around $\eta^{\mathrm{opt}} \approx 1.0$. In Example 2, we confirm this observation by determining η^{opt} numerically.

An alternative way of establishing that the limit of $R_s(\{p_s(k)\}_{k\ge 0})$ is $R(f)$, is by exploiting the stochastic-process limit for $\{Q_s(t)\}_{t\ge 0}$. It was shown in Janssen et al. (2013) that under condition (i) in Proposition 1, together with (3) and (10), the normalized process $\{\hat{Q}_s(t)\}_{t\ge 0}$ with $\hat{Q}_s(t) = (Q_s(t) - s)/\sqrt{s}$ converges weakly to a stochastic-process limit $\{D(t)\}_{t\ge 0}$ with stationary density

$$w(x) = \begin{cases} Z^{-1}\mathrm{e}^{-(1/2)x^2 - \gamma x}, & x < 0, \\ Z^{-1} f(x)\mathrm{e}^{-\gamma x}, & x \ge 0, \end{cases}$$

where $Z = \Phi(\gamma)/\phi(\gamma) + \int_0^\infty f(x)\exp(-\gamma x)\,\mathrm{d}x$. This diffusion process essentially behaves as a Brownian motion when $x \ge 0$, and as an Ornstein–Uhlenbeck process when $x < 0$. When additionally assuming (2), the limiting system revenue can be written as

$$R(f) = \int_{-\infty}^{\infty} r(x)w(x)\,\mathrm{d}x,$$

the stationary revenue rate generated by the stochastic-process limit. So an alternative method to prove Proposition 1 would be to first formally establish weak convergence at the process level, then prove that limits with respect to space and time can be interchanged, and finally use the stationary behavior of the stochastic-process limit. This is a common approach in the QED literature (Halfin and Whitt 1981, Garnett et al. 2002). Instead, we construct a direct, purely analytic proof, that additionally gives insight into the error that is made when approximating $R_s(\{p_s(k)\}_{k\ge 0})$ by $R(f)$ for finite s. These error estimates are available in Appendix A for future reference.

With Proposition 1 at hand, we are naturally led to consider the asymptotic optimization problem, namely,

$$\begin{aligned} &\text{to maximize} \quad R(f) \text{ over } f, \\ &\text{subject to} \quad 0 \le f(x) \le 1, \quad x \in [0,\infty), \quad \text{and} \quad \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x < \infty. \end{aligned} \tag{13}$$

The condition $\int_0^\infty f(x)e^{-\gamma x}\,dx < \infty$ is the limiting form of the stability condition $F_s(\rho) < \infty$, see Janssen et al. (2013). Also note that we do not restrict f to be monotone. We prove for the optimization problem in (13) the following in Section 4.

Proposition 2. *If r is nonincreasing for $x \geq 0$, then there exist optimal asymptotic admission controls that solve (13). Moreover, the optimal asymptotic admission control profiles have a threshold structure of the form*

$$f(x) = \mathbb{1}[0 \leq x < \eta^{\text{opt}}],$$

where η^{opt} is any solution of

$$r(\eta) = R_{\text{T}}(\eta) \tag{14}$$

if $r(0) > R_{\text{T}}(0)$, and $\eta^{\text{opt}} = 0$ if $r(0) \leq R_{\text{T}}(0)$. If r is strictly decreasing in $x \geq 0$, then η^{opt} is unique.

Recall that the optimality of a threshold policy should not come as a surprise, and could in fact be shown in the pre-limit and within a far wider class of policies than those satisfying (3). The strength of Proposition 2 lies in the characterization (14) of η^{opt}. We refer to (14) as the *threshold equation*: it is a powerful basis on which to obtain numerical solutions, closed-form expressions, bounds, and asymptotic expansions for η^{opt}. Results for η^{opt} of this nature are presented in Section 3.

Example 2 (Exponential Revenue Revisited)**.** Let us revisit Example 1, where $r(x) = \exp(bx)$ for $x < 0$ and $r(x) = \exp(-dx)$ for $x \geq 0$. The threshold equation, (14), takes the form

$$e^{-d\eta}\left(\frac{\Phi(\gamma)}{\phi(\gamma)} + \frac{1 - e^{-\gamma\eta}}{\gamma}\right) = \frac{\Phi(\gamma - b)}{\phi(\gamma - b)} + \frac{1 - e^{-(d+\gamma)\eta}}{d + \gamma}, \tag{15}$$

which we study in depth in Section 3.4.2. When $b = 5$, $d = 1$, and $\gamma = 0.01$, solving (15) numerically yields $\eta^{\text{opt}} \approx 1.00985$, which supports our earlier observation that $\eta^{\text{opt}} \approx 1$ in Example 1.

The true optimal admission threshold $\tau^{\text{opt}} = \arg\max_{\tau \in \mathbb{N}_0} R_{\text{T},s}(\tau)$ is plotted in Figure 3 as a function of s, along with the asymptotic QED approximation $\tau^{\text{opt}} \approx \lfloor \eta^{\text{opt}}\sqrt{s} \rfloor$. We observe that the QED approximation is accurate, even for a relatively small number of servers, and exact in the majority of cases. This is reflected in Figure 4, which plots the relative optimality gap as a function of s. The relative optimality gap is zero for the vast majority of s values, and as low as 10^{-2} for systems with as few as 10 servers.

We remark that when utilizing the asymptotic optimal threshold provided by Proposition 2 in a finite system, the proof of Proposition 1 in Appendix A guarantees that $R_{s,\text{T}}(\lfloor \eta^{\text{opt}}\sqrt{s} \rfloor) - R_{\text{T}}(\eta^{\text{opt}}) = O(1/\sqrt{s})$. In other words, a finite system that utilizes the asymptotic optimal threshold achieves a revenue within $O(1/\sqrt{s})$ of the solution to (13).

2.3. Customer Reward Maximization

In Section 1 we have discussed the difference between revenues seen from the system's perspective and from the customer's perspective. Although the emphasis lies on the system's perspective, as in Section 2.2, we now show how results for the customer's perspective can be obtained.

Figure 3. The true optimal admission threshold τ^{opt} as a function of s, together with the (almost indistinguishable) QED approximation $\lfloor \eta^{\text{opt}}\sqrt{s} \rfloor$.

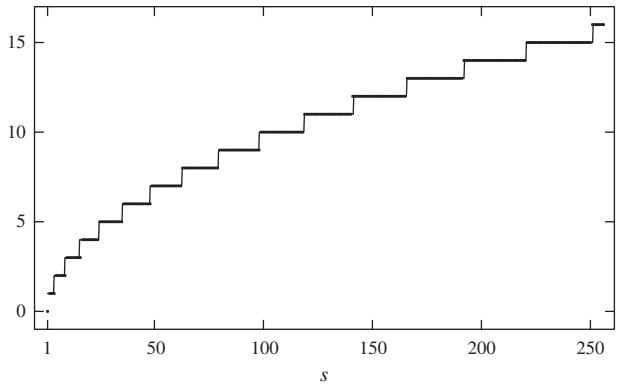
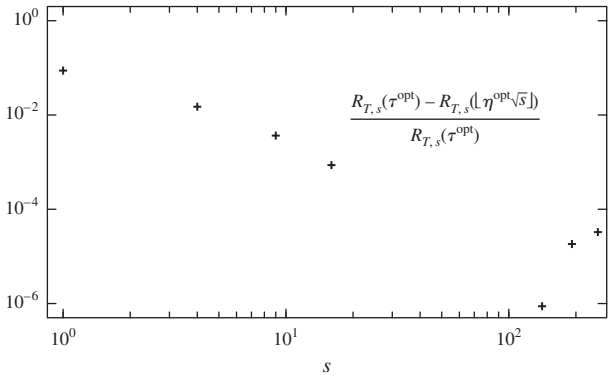


Figure 4. The relative error $(R_{\mathrm{T},s}(\tau^{\mathrm{opt}}) - R_{\mathrm{T},s}(\lfloor \eta^{\mathrm{opt}}\sqrt{s}\rfloor))/R_{\mathrm{T},s}(\tau^{\mathrm{opt}})$ as a function of s. The missing points indicate an error that is strictly zero. The errors that are non-zero arise due to the QED approximation for the optimal admission threshold being off by just one state.



2.3.1. Linear Revenue Structure. If revenue is generated at rate $a > 0$ for each customer that is being served, and cost is incurred at rate $b > 0$ for each customer that is waiting for service, the revenue structure is given by

$$r_s(k) = \begin{cases} ak, & k \le s, \\ as - b(k-s), & k \ge s. \end{cases} \tag{16}$$

When $n_s = as$ and $q_s = \sqrt{s}$, the revenue structure in (16) satisfies the scaling condition in (2), with

$$r(x) = \begin{cases} ax, & x \le 0, \\ -bx, & x \ge 0. \end{cases} \tag{17}$$

Consequently, Proposition 1 implies that

$$\lim_{s\to\infty} \frac{R_s(\{p_s(k)\}_{k\ge 0}) - as}{\sqrt{s}} = \frac{a(1+\gamma(\Phi(\gamma)/\phi(\gamma))) - b\int_0^\infty x f(x) \mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\Phi(\gamma)/\phi(\gamma) + \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x},$$

for any profile f, and Proposition 2 reveals that the optimal control is $f(x) = \mathbb{1}[0 \le x < \eta^{\mathrm{opt}}]$ with η^{opt} the unique solution of the threshold equation (14), which with $c = a/b$ becomes

$$\eta\left(\frac{\Phi(\gamma)}{\phi(\gamma)} + \frac{1-\mathrm{e}^{-\gamma\eta}}{\gamma}\right) = c\left(1 + \gamma\frac{\Phi(\gamma)}{\phi(\gamma)}\right) + \frac{1-(1+\gamma\eta)\mathrm{e}^{-\gamma\eta}}{\gamma^2}. \tag{18}$$

We see that η^{opt} depends only on c. The threshold equation (18) is studied extensively in Section 3.4.1. A minor variation of the arguments used there to prove Proposition 8, shows that

$$\eta^{\mathrm{opt}} = r_0 + \frac{1}{\gamma} W\left(\frac{\gamma \mathrm{e}^{-\gamma r_0}}{a_0}\right), \tag{19}$$

where W denotes the Lambert W function, i.e., the solution to the equation $W(z)\exp(W(z)) = z$, and

$$a_0 = -\gamma - \gamma^2\frac{\Phi(\gamma)}{\phi(\gamma)}, \qquad r_0 = c\gamma + \frac{1}{\gamma + \gamma^2(\Phi(\gamma)/\phi(\gamma))}.$$

2.3.2. Relating System-Governed Revenue to Customer Rewards. From (19), it can be deduced that for large values of γ, $\eta^{\mathrm{opt}} \approx c\gamma$ (see the proof of Proposition 10, and for a discussion on the asymptotic behavior of the threshold equation for general revenue structures, we refer to Section 3.3). Thus, asymptotically, the optimal threshold value is approximately equal to the product of the staffing slack γ and the ratio of the service revenue a and the waiting cost b.

The asymptotic behavior $\eta^{\mathrm{opt}} \approx c\gamma$ may be explained as follows. For each arriving customer, we must balance the expected revenue a when that customer is admitted and eventually served against the expected waiting

cost incurred for that customer as well as the additional waiting cost for customers arriving after that customer. When the arriving customer finds τ customers waiting, the overall waiting cost incurred by admitting that customer may be shown to behave roughly as $b\tau/(\gamma\sqrt{s})$ for large values of γ. Equating a with $b\tau/(\gamma\sqrt{s})$ then yields that the optimal threshold value should approximately be $\tau^{\text{opt}} \approx c\gamma\sqrt{s}$.

The stationary average revenue rate $R_s(\{p_s(k)\}_{k\geq 0})$ under revenue structure (16) is therefore the same as when a reward $a > 0$ is received for each admitted customer and a penalty $b\mathbb{E}[W]$ is charged when the expected waiting time of that customer is $\mathbb{E}[W]$, with $b > 0$. In the latter case the stationary average reward earned may be expressed as in (5), where now

$$\hat{r}_s(k) = a - b\max\left\{0, \frac{k-s+1}{s}\right\} \tag{20}$$

denotes a *customer reward*.

The *system-governed* revenue rate and the *customer* reward rate are in this case equivalent. To see this, write

$$\hat{R}_s(\{p_s(k)\}_{k\geq 0}) = a\lambda\left(\sum_{k=0}^{s-1}\pi_s(k) + \sum_{k=s}^{\infty}p_s(k-s)\pi_s(k)\right) - b\lambda\sum_{k=s}^{\infty}\frac{k-s+1}{s}p_s(k-s)\pi_s(k).$$

Then note that because the arrival rate multiplied by the probability that an arriving customer is admitted must equal the expected number of busy servers, and by local balance $\lambda\pi_s(k)p_s(k-s) = s\pi_s(k+1)$ for $k = s, s+1, \ldots$, we have

$$\begin{aligned}\hat{R}_s(\{p_s(k)\}_{k\geq 0}) &= a\left(\sum_{k=0}^{s-1}k\pi_s(k) + \sum_{k=s}^{\infty}s\pi_s(k)\right) - b\sum_{k=s+1}^{\infty}(k-s)\pi_s(k)\\ &= \sum_{k=0}^{s-1}ak\pi_s(k) + \sum_{k=s}^{\infty}(as - b(k-s))\pi_s(k) \overset{(16)}{=} \sum_{k=0}^{\infty}r_s(k)\pi_s(k) = R_s(\{p_s(k)\}_{k\geq 0}).\end{aligned}$$

The optimal threshold in (19) thus maximizes the customer reward rate $\hat{R}_s$ asymptotically as well, i.e., in this example by Proposition 2,

$$\lim_{s\to\infty}\frac{\hat{R}_s(\{p_s(k)\}_{k\geq 0}) - a}{\sqrt{s}} = \lim_{s\to\infty}\frac{R_s(\{p_s(k)\}_{k\geq 0}) - a}{\sqrt{s}} = R(f).$$

In fact, for *any* system-governed revenue rate $r_s(k)$, the related customer reward structure $\hat{r}_s(k)$ is given by

$$\hat{r}_s(k) = \frac{r_s(k+1)}{\min\{k+1, s\}}, \quad k \in \mathbb{N}_0, \tag{21}$$

because then

$$\begin{aligned}\hat{R}_s(\{p_s(k)\}_{k\geq 0}) &= \sum_{k=0}^{s-1}\hat{r}_s(k)\lambda\pi_s(k) + \sum_{k=s}^{\infty}\hat{r}_s(k)\lambda p_s(k-s)\pi_s(k) = \sum_{k=0}^{s-1}\hat{r}_s(k)(k+1)\pi_s(k+1) + \sum_{k=s}^{\infty}\hat{r}_s(k)s\pi_s(k+1)\\ &= \sum_{k=0}^{s}\hat{r}_s(k-1)k\pi_s(k) + \sum_{k=s+1}^{\infty}\hat{r}_s(k-1)s\pi_s(k) = R_s(\{p_s(k)\}_{k\geq 0}),\end{aligned}$$

using local balance, $\lambda\pi_s(k) = (k+1)\pi_s(k+1)$ for $k = 0, 1, \ldots, s-1$ and $\lambda p_s(k-s)\pi_s(k) = s\pi_s(k+1)$ for $k = s, s+1, \ldots$.

Proposition 3. *For any system-governed revenue rate $r_s(k)$, the customer reward structure $\hat{r}_s(k)$ in (21) guarantees that the average system-governed revenue rate $R_s(\{p_s(k)\}_{k\geq 0})$ equals the customer reward rate $\hat{R}_s(\{p_s(k)\}_{k\geq 0})$.*

In particular, Proposition 3 implies that counterparts to Propositions 1 and 2 hold for the customer reward rate $\hat{R}_s(\{p_s(k)\}_{k\geq 0})$, assuming the customer reward structure $\hat{r}_s(k)$ is appropriately scaled.

3. Properties of the Optimal Threshold

We focus throughout this paper on maximization of the average system-governed revenue rate. In Section 2 we have established threshold optimality and derived the threshold equation that defines the optimal threshold η^{opt}. In this section we obtain a series of results about η^{opt}. In Section 3.1, we present a procedure (for general revenue functions) to obtain an upper bound $\eta^{\max}$ and a lower bound $\eta^{\min}$ on the optimal threshold η^{opt}. Section 3.2 discusses our monotonicity results. Specifically, we prove that η^{opt} increases with $\gamma \in \mathbb{R}$, and that $R_{\text{T}}(0)$, $R_{\text{T}}(\eta^{\text{opt}})$

both decrease with $\gamma \in \mathbb{R}$. In Section 3.3, we derive asymptotic descriptions of the optimal threshold for general revenue structures, even if the revenue structures would not allow for an explicit characterization. We prove that $\eta^{\text{opt}} \approx r^{\leftarrow}(r(-\gamma))$ as $\gamma \to \infty$, and that $\eta^{\text{opt}} \approx -(1/\gamma)\ln(1 - r'(0-)/r'(0+))$ as $\gamma \to -\infty$. In Section 3.4, we derive explicit characterizations of η^{opt} for linear and exponential revenue structures.

From here on, we assume that $r(x)$ is piecewise smooth and bounded on $(-\infty, 0)$ and $(0, \infty)$, and continuous at 0 with $r(\pm 0) = 1 = r(0)$. We also assume that $r(x)$ is increasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$, with $0 \le r(x) \le r(0) = 1$. Revenue functions for which $r(0) > 0$ and $r(0) \ne 1$ can be considered through analysis of the scaled revenue function $\bar{r}(x) = r(x)/r(0)$. For notational convenience, we also define $r_{\text{L}}(x)$ and $r_{\text{R}}(x)$ as

$$r(x) = \begin{cases} r_{\text{L}}(x), & x < 0, \\ r_{\text{R}}(x), & x \ge 0, \end{cases}$$

and introduce $A = \int_{-\infty}^{0} r_{\text{L}}(x) \mathrm{e}^{-(1/2)x^2 - \gamma x}\, \mathrm{d}x$, and $B = \Phi(\gamma)/\phi(\gamma)$. Note that $R_{\text{T}}(0) = A/B$.

Corollary 1. *Under these assumptions, there exists a solution $\eta^{\text{opt}} > 0$ of the threshold equation. This solution is positive, unless $r_{\text{L}}(x) = 1$, and unique when $r'_{\text{R}}(x) < 0$ for all $x \ge 0$ such that $r_{\text{R}}(x) > 0$.*

Proof. Note that these assumptions on r are slightly stronger than in Proposition 2. This corollary is directly implied by our proof of Proposition 2; see the explanation between (48) and (49). □

3.1. General Bounds

Denote the inverse function of r by $r^{\leftarrow}$. The following bounds hold for general revenue structures, are readily calculated, and provide insight into the optimal thresholds. Later in Sections 3.4.1 and 3.4.2, we illustrate Proposition 4 for a linear revenue structure, and an exponential revenue structure, respectively.

Proposition 4. *When r is strictly decreasing for $x \ge 0$, $\eta^{\max} = r_{\text{R}}^{\leftarrow}(R_{\text{lower}}) \ge \eta^{\text{opt}}$, and $\eta^{\min} = r_{\text{R}}^{\leftarrow}(R_{\text{upper}}) < \eta^{\text{opt}}$. Here,*

$$R_{\text{lower}} = R_{\text{T}}(0) = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{-(1/2)x^2 - \gamma x}\, \mathrm{d}x}{\Phi(\gamma)/\phi(\gamma)}, \qquad R_{\text{upper}} = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{-(1/2)x^2 - \gamma x}\, \mathrm{d}x + \int_{0}^{\eta^{\max}} \mathrm{e}^{-\gamma x}\, \mathrm{d}x}{\Phi(\gamma)/\phi(\gamma) + \int_{0}^{\eta^{\max}} \mathrm{e}^{-\gamma x}\, \mathrm{d}x}.$$

Proof. The assumptions on $r_{\text{R}}(x)$ imply that its inverse function $r_{\text{R}}^{\leftarrow}(y)$ exists, and that it is also strictly decreasing. It is therefore sufficient to provide upper and lower bounds on $R_{\text{T}}(\eta)$ that are independent of η.

For threshold control policies, the system revenue is given by (12). Also recall that the optimal threshold η^{opt} solves the threshold equation, i.e., $r_{\text{R}}(\eta^{\text{opt}}) = R_{\text{T}}(\eta^{\text{opt}})$. By suboptimality, we immediately obtain $R_{\text{lower}} \le R_{\text{T}}(\eta^{\text{opt}})$, and so $\eta^{\text{opt}} \le \eta^{\max}$ by monotonicity.

We will first derive an alternative forms of the threshold equation. For instance, rewriting (14) into

$$\left(B + \int_{0}^{\eta} \mathrm{e}^{-\gamma x}\, \mathrm{d}x\right) r(\eta) = A + \int_{0}^{\eta} r(x)\mathrm{e}^{-\gamma x}\, \mathrm{d}x, \tag{22}$$

dividing by B, and using $R_{\text{T}}(0) = A/B$ gives

$$r(\eta) - R_{\text{T}}(0) = -\frac{r(\eta)}{B}\int_{0}^{\eta} \mathrm{e}^{-\gamma x}\, \mathrm{d}x + \int_{0}^{\eta} r(x)\mathrm{e}^{-\gamma x}\, \mathrm{d}x.$$

We then identify the right-hand member as being a result of an integration by parts, to arrive at the alternative form

$$r(\eta) - R_{\text{T}}(0) = -\left[\frac{r(x)}{B}\int_{0}^{x} \mathrm{e}^{-\gamma u}\, \mathrm{d}u\right]_{0}^{\eta} + \int_{0}^{\eta} \mathrm{e}^{-\gamma x}\, \mathrm{d}x = -\int_{0}^{\eta} \frac{r'(x)}{B}\int_{0}^{x} \mathrm{e}^{-\gamma u}\, \mathrm{d}u\, \mathrm{d}x. \tag{23}$$

Let $c(\eta) = \int_{0}^{\eta} \mathrm{e}^{-\gamma x}\, \mathrm{d}x = (1 - \mathrm{e}^{-\gamma\eta})/\gamma$ if $\gamma \ne 0$ and $c(\eta) = \eta$ if $\gamma = 0$. Since $c(\eta)$ is increasing in η and $-r'(x) \ge 0$ for $x \ge 0$, we have for $\eta \ge 0$

$$-\frac{1}{B}\int_{0}^{\eta} r'(x)c(x)\, \mathrm{d}x < -\frac{1}{B}c(\eta)\int_{0}^{\eta} r'(x)\, \mathrm{d}x = \frac{1}{B}c(\eta)(1 - r(\eta)).$$

Let $\eta = \hat{\eta}$ be the (unique) solution of the equation

$$r(\eta) - R_{\text{T}}(0) = \frac{1}{B}c(\eta)(1 - r(\eta)). \tag{24}$$

Then

$$r(\hat{\eta}) - R_{\mathrm{T}}(0) = \frac{1}{B}c(\hat{\eta})(1 - r(\hat{\eta})) > -\frac{1}{B}\int_0^{\hat{\eta}} r'(x)c(x)\,\mathrm{d}x,$$

and so $0 < \hat{\eta} < \eta^{\mathrm{opt}}$. We have from (24) that

$$r(\hat{\eta}) = \frac{R_{\mathrm{T}}(0) + (1/B)c(\hat{\eta})}{1 + (1/B)c(\hat{\eta})} = \frac{A + c(\hat{\eta})}{B + c(\hat{\eta})}. \tag{25}$$

From $\hat{\eta} < \eta^{\mathrm{opt}} < \eta^{\max} = r^{\leftarrow}(R_{\mathrm{T}}(0))$, we then find

$$c(\hat{\eta}) < c(\eta^{\max}), \qquad \frac{A + c(\hat{\eta})}{B + c(\hat{\eta})} < \frac{A + c(\eta^{\max})}{B + c(\eta^{\max})} = R_{\mathrm{upper}},$$

since $0 < A < B$, i.e., $R_{\mathrm{T}}(0) < 1$. This completes the proof. □

3.2. Monotonicity

We next investigate the influence of the slack γ on the optimal threshold.

Proposition 5. *The revenue $R_{\mathrm{T}}(0)$ decreases in $\gamma \in \mathbb{R}$.*

Proof. Write $r(-x) = u(x)$ so that $0 \le u(x) \le 1 = u(0)$ and

$$R_{\mathrm{T}}(0) = \frac{\int_0^\infty u(x)\mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x}{\int_0^\infty \mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x}, \quad \gamma \in \mathbb{R},$$

and calculate

$$\frac{\mathrm{d}R_{\mathrm{T}}(0)}{\mathrm{d}\gamma} = \frac{\int_0^\infty \mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x \int_0^\infty x u(x)\mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x}{\left(\int_0^\infty \mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x\right)^2} - \frac{\int_0^\infty u(x)\mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x \int_0^\infty x\mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x}{\left(\int_0^\infty \mathrm{e}^{-(1/2)x^2+\gamma x}\,\mathrm{d}x\right)^2}.$$

The numerator can be written as

$$N = \int_0^\infty \int_0^\infty (x-y)u(x)\mathrm{e}^{-(1/2)x^2+\gamma x}\mathrm{e}^{-(1/2)y^2+\gamma y}\,\mathrm{d}x\mathrm{d}y.$$

Suppose that $u(x) = \mathbb{1}[0 \le x < a]$ for some $a > 0$. Then

$$N = \int_0^a \int_0^\infty (x-y)\mathrm{e}^{-(1/2)x^2+\gamma x}\mathrm{e}^{-(1/2)y^2+\gamma y}\,\mathrm{d}x\mathrm{d}y \le \int_0^a \int_0^a (x-y)\mathrm{e}^{-(1/2)x^2+\gamma x}\mathrm{e}^{-(1/2)y^2+\gamma y}\,\mathrm{d}x\mathrm{d}y = 0.$$

In general, we can write $u(x) = -\int_0^\infty \mathbb{1}[0 \le x < a]u'(a)\,\mathrm{d}a = -\int_x^\infty u'(a)\,\mathrm{d}a$ with $u'(a) < 0$, to arrive at

$$N = \int_0^\infty -u'(a)\left(\int_0^\infty \int_0^\infty (x-y)\mathbb{1}[0 \le x < a]\mathrm{e}^{-(1/2)x^2+\gamma x}\mathrm{e}^{-(1/2)y^2+\gamma y}\,\mathrm{d}x\mathrm{d}y\right)\mathrm{d}a \le 0.$$

This concludes the proof. □

Proposition 6. *The optimal threshold η^{opt} increases in $\gamma \in \mathbb{R}$, and $R_{\mathrm{T}}(\eta^{\mathrm{opt}})$ decreases in $\gamma \in \mathbb{R}$.*

Proof. By Proposition 5, we have that $R_{\mathrm{T}}(0)$ decreases in $\gamma \in \mathbb{R}$. Furthermore, for any $\eta > 0$, we have that $\int_0^\eta \mathrm{e}^{-\gamma x}\,\mathrm{d}x/B$ decreases in $\gamma \in \mathbb{R}$. Consider the alternative form (23) of the threshold equation. For fixed η, the left member thus increases in γ, while the right member decreases in γ, since $r'(x) < 0$. The solution η^{opt} of the threshold equation therefore increases in $\gamma \in \mathbb{R}$.

To prove the second part of the claim, we recall that η^{opt} solves the threshold equation, so $R_{\mathrm{T}}(\eta^{\mathrm{opt}}) = r_{\mathrm{R}}(\eta^{\mathrm{opt}})$. Since $\eta^{\mathrm{opt}} \ge 0$ is increasing in $\gamma \in \mathbb{R}$, our assumptions on r imply that $r_{\mathrm{R}}(\eta^{\mathrm{opt}})$ is decreasing in $\gamma \in \mathbb{R}$. Hence, $R_{\mathrm{T}}(\eta^{\mathrm{opt}})$ is decreasing in $\gamma \in \mathbb{R}$ as well. □

Proposition 6 can be interpreted as follows. First note that an increase in γ means that fewer customers are served by the system, apart from the impact of a possible admission control policy. Then, for threshold control, an increased γ implies that the optimal threshold should increase, in order to serve more customers. This of course is a direct consequence of our revenue structure, which is designed to let the system operate close to the ideal operating point. A large γ drifts the process away from this ideal operating point, and this can be compensated for by a large threshold η^{opt}. Hence, although the slack γ and the threshold η^{opt} have quite different impacts on the system behavior, at a high level their monotonic relation can be understood, and underlines that the revenue structure introduced in this paper has the right properties for the QED regime.

3.3. Asymptotic Solutions

We now present asymptotic results for the optimal threshold in the regimes where the slack γ becomes extremely large or extremely small.

Proposition 7. *When $\gamma \to -\infty$, and if the revenue function has a cusp at $x = 0$, i.e., $r'_{\mathrm{R}}(0+) < 0 < r'_{\mathrm{L}}(0-)$, the optimal threshold is given by*

$$\eta^{\mathrm{opt}} = -\frac{1}{\gamma} \ln\left(1 - \frac{r'_{\mathrm{L}}(0-)}{r'_{\mathrm{R}}(0+)}\right) + O\left(\frac{1}{\gamma^2}\right). \tag{26}$$

Proof. We consider $\gamma \to -\infty$. From steepest descent analysis, we have for a smooth and bounded f on $(-\infty, 0]$,

$$\int_{-\infty}^{0} f(x) \mathrm{e}^{-\gamma x} \, \mathrm{d}x = -\frac{f(0)}{\gamma} - \frac{f'(0)}{\gamma^2} + O\left(\frac{1}{\gamma^3}\right), \quad \gamma \to -\infty.$$

Hence, it follows that

$$R_{\mathrm{T}}(0) = \frac{A}{B} = \frac{-1/\gamma - r'_{\mathrm{L}}(0-)/\gamma^2 + O(1/\gamma^3)}{-1/\gamma + O(1/\gamma^3)} = 1 + \frac{r'_{\mathrm{L}}(0-)}{\gamma} + O\left(\frac{1}{\gamma^2}\right), \quad \gamma \to -\infty. \tag{27}$$

From the upper bound $\eta^{\mathrm{opt}} < r^{\leftarrow}(R_{\mathrm{T}}(0))$ and $r(0) = 1$, $r'_{\mathrm{R}}(0+) < 0$, we thus see that $\eta^{\mathrm{opt}} = O(1/|\gamma|)$, $\gamma \to -\infty$, and so in the threshold equation, see (23), we only need to consider η's of $O(1/|\gamma|)$. In (23), we have $\int_0^x \exp(-\gamma u) \, \mathrm{d}u = (1 - \exp(-\gamma x))/\gamma$. Using that $1/(\gamma B) = 1 + O(1/\gamma^2)$, see (27), we get for the right-hand side of (23),

$$-\frac{1}{B} \int_0^{\eta} r'_{\mathrm{R}}(x) \int_0^x \mathrm{e}^{-\gamma u} \, \mathrm{d}u \, \mathrm{d}x = \int_0^{\eta} r'_{\mathrm{R}}(x)(1 - \mathrm{e}^{-\gamma x} \, \mathrm{d}x) \mathrm{d}x \left(1 + O\left(\frac{1}{\gamma^2}\right)\right).$$

Next,

$$r'_{\mathrm{R}}(x) = r'_{\mathrm{R}}(0+) + O\left(\frac{1}{\gamma}\right), \qquad 1 - \mathrm{e}^{-\gamma x} = O(1), \quad 0 \le x \le \eta = O\left(\frac{1}{\gamma}\right),$$

and so

$$-\frac{1}{B} \int_0^{\eta} r'_{\mathrm{R}}(x) \int_0^x \mathrm{e}^{-\gamma u} \, \mathrm{d}u \, \mathrm{d}x = -r'_{\mathrm{R}}(0+) \frac{1 - \mathrm{e}^{-\gamma\eta} - \gamma\eta}{\gamma} + O\left(\frac{1}{\gamma^2}\right). \tag{28}$$

Furthermore, for the left-hand side of (23), we have

$$r_{\mathrm{R}}(\eta) - R_{\mathrm{T}}(0) = 1 + r'_{\mathrm{R}}(0+)\eta + O\left(\frac{1}{\gamma^2}\right) - \left(1 + \frac{r'_{\mathrm{L}}(0-)}{\gamma} + O\left(\frac{1}{\gamma^2}\right)\right) = r'_{\mathrm{R}}(0+)\eta - \frac{r'_{\mathrm{L}}(0-)}{\gamma} + O\left(\frac{1}{\gamma^2}\right). \tag{29}$$

Equating (28) and (29) and simplifying, we find

$$r'_{\mathrm{R}}(0+)(1 - \mathrm{e}^{-\gamma\eta}) = r'_{\mathrm{L}}(0-) + O\left(\frac{1}{\gamma}\right),$$

and this gives (26). □

If $r_{\mathrm{L}}(x)$ is slowly varying, the optimal threshold is approximately given by

$$\eta^{\mathrm{opt}} \approx r_{\mathrm{R}}^{\leftarrow}(r_{\mathrm{L}}(-\gamma)) \tag{30}$$

as $\gamma \to \infty$. To see this, note that as $\gamma \to \infty$,

$$R_{\mathrm{T}}(0) = \frac{\int_{-\infty}^{0} r_{\mathrm{L}}(x) \mathrm{e}^{-(1/2)x^2 - \gamma x} \, \mathrm{d}x}{\int_{-\infty}^{0} \mathrm{e}^{-(1/2)x^2 - \gamma x} \, \mathrm{d}x} = \frac{\mathrm{e}^{(1/2)\gamma^2} \int_0^{\infty} r_{\mathrm{L}}(-x) \mathrm{e}^{-(1/2)(x-\gamma)^2} \, \mathrm{d}x}{\mathrm{e}^{(1/2)\gamma^2} \int_0^{\infty} \mathrm{e}^{-(1/2)(x-\gamma)^2} \, \mathrm{d}x} \approx r_{\mathrm{L}}(-\gamma).$$

A full analysis goes be beyond the scope of this paper, and would overly complicating our exposition. Instead, consider as an example $r_{\mathrm{L}}(x) = \exp(bx)$ with $b > 0$ small. We have as $\gamma \to \infty$ with exponentially small error

$$R_{\mathrm{T}}(0) = \frac{\Phi(\gamma - b)}{\phi(\gamma - b)} \cdot \frac{\phi(\gamma)}{\Phi(\gamma)} = \mathrm{e}^{-b\gamma} \mathrm{e}^{b^2/2} \left(1 - b \frac{\phi(\gamma)}{\Phi(\gamma)} + O(b^2)\right) = r_{\mathrm{L}}(-\gamma)(1 + O(b^2)). \tag{31}$$

When for instance $r_{\mathrm{R}}(x) = \exp(dx)$ with $d > 0$, we get that $\eta^{\mathrm{opt}} \approx r_{\mathrm{R}}^{\leftarrow}(\exp(-b\gamma + b^2/2)) = (b/d)\gamma - b^2/(2d)$ with exponentially small error, as $\gamma \to \infty$. Furthermore, the right-hand side in (23) is exponentially small as $\gamma \to \infty$, so that in good approximation the solution to the threshold equation is indeed given by (30).

3.4. Explicit Results for Two Special Cases

We now study the two special cases of linear and exponential revenue structures. For these cases we are able to find precise results for the η^{opt}. We demonstrate these results for some example systems, and also include the bounds and asymptotic results obtained in Sections 3.1 and 3.3, respectively.

3.4.1. Linear Revenue. We first present an explicit expression for the optimal threshold for the case of a linear revenue function,

$$r_{\text{R}}(x) = \left(1 - \frac{x}{d}\right)\mathbb{1}[0 \le x \le d], \quad x \ge 0,$$

and arbitrary $r_{\text{L}}(x)$. We distinguish between $\gamma \neq 0$ and $\gamma = 0$ in Propositions 8 and 9 below.

Proposition 8. *Assume $\gamma \neq 0$. Then*

$$\eta^{\text{opt}} = r_0 + \frac{1}{\gamma} W\left(\frac{\gamma e^{-\gamma r_0}}{a_0}\right), \tag{32}$$

where W denotes Lambert's W *function, see below* (19)*, and*

$$a_0 = -\gamma^2\left(B + \frac{1}{\gamma}\right), \qquad r_0 = \frac{d(B-A) + 1/\gamma^2}{B + 1/\gamma}. \tag{33}$$

Proof. It follows from Avram et al. (2013, Section 4) that $B + 1/\gamma \neq 0$ when $\gamma \neq 0$ so that a_0, r_0 in (33) are well-defined with $a_0 \neq 0$. From the threshold equation in (23), and $r_{\text{R}}(\eta) = 1 - \eta/d$ when $0 \le \eta \le d$, we see that $\eta = \eta^{\text{opt}}$ satisfies

$$1 - \frac{\eta}{d} - \frac{A}{B} = \frac{1}{d}\int_0^{\eta}\int_0^{x} e^{-\gamma u}\,du\,dx.$$

Now

$$\int_0^{\eta}\int_0^{x} e^{-\gamma u}\,du\,dx = \frac{1}{\gamma}\left(\eta - \frac{1 - e^{-\gamma\eta}}{\gamma}\right),$$

and this yields for $\eta = \eta^{\text{opt}}$ the equation

$$\gamma(\eta - r_0)e^{\gamma(\eta - r_0)} = \frac{\gamma}{a_0}e^{-\gamma r_0} \tag{34}$$

with a_0 and r_0 given in (33). Note that Equation (34) is of the form $W(z) \times \exp(W(z)) = z$, which is the defining equation for Lambert's W function, and this yields the result.

Proposition 8 provides a connection with the developments in Borgs et al. (2014). Furthermore, the optimal threshold η^{opt} is readily computed from it, taking care that the branch choice for W is such that the resulting η^{opt} is positive, continuous, and increasing as a function of γ. For this matter, the following result is relevant.

Proposition 9. *For $r_{\text{R}}(x) = 1 - x/d$ with $d > 0$, and arbitrary $r_{\text{L}}(x)$, as $\gamma \to 0$,*

$$\eta^{\text{opt}} = \sqrt{\frac{\pi}{2} + 2d\left(\sqrt{\frac{\pi}{2}} - \int_{-\infty}^{0} r_{\text{L}}(x)e^{-(1/2)x}\,dx\right)} - \sqrt{\frac{\pi}{2}} + O(\gamma). \tag{35}$$

Proof. In the threshold equation in (23), we set $\varepsilon = 1 - A/B$ and use $r_{\text{R}}(x) = 1 - x/d$, $r'_{\text{R}}(x) = -1/d$, to arrive at

$$d\varepsilon - \eta = \frac{1}{B}\int_0^{\eta}\int_0^{x} e^{-\gamma u}\,du\,dx.$$

Since

$$\int_0^{\eta}\int_0^{x} e^{-\gamma u}\,du\,dx = \int_0^{\eta}(x + O(\gamma x^2))\,dx = \frac{1}{2}\eta^2 + O(\gamma\eta^3),$$

we obtain the equation

$$d\varepsilon - \eta = \frac{1}{2B}\eta^2 + O(\gamma\eta^3). \tag{36}$$

Using that $\eta^{\text{opt}} < r^{\leftarrow}(A/B) = O(1)$ as $\gamma \to 0$, we find from (36) that as $\gamma \to 0$,

$$\eta^{\text{opt}} = \sqrt{B^2 + 2Bd\varepsilon} - B + O(\gamma) = \sqrt{B^2 + 2d(B-A)} - B + O(\gamma).$$

Finally (35) follows from the expansions

$$B = \sqrt{\frac{\pi}{2}} + O(\gamma), \qquad A = \int_{-\infty}^{0} r_{\mathrm{L}}(x) \mathrm{e}^{-(1/2)x^2}\, \mathrm{d}x + O(\gamma),$$

as $\gamma \to 0$. □

We may also study the regime $\gamma \to \infty$. Note that the following result coincides with the asymptotic behavior of $\eta^{\min}$ and $\eta^{\max}$ in Proposition 4.

Proposition 10. *For $r_{\mathrm{L}}(x) = \mathrm{e}^{bx}$, and $r_{\mathrm{R}}(x) = (d-x)/d$, as $\gamma \to \infty$,*

$$\eta^{\mathrm{opt}} = d\left(1 - \frac{\Phi(\gamma - b)\phi(\gamma)}{\phi(\gamma - b)\Phi(\gamma)}\right) + O\left(\frac{1}{\gamma}\mathrm{e}^{-(1/2)\gamma^2}\right).$$

Proof. The revenue structure implies that

$$A = \frac{\Phi(\gamma - b)}{\phi(\gamma - b)}, \qquad B = \frac{\Phi(\gamma)}{\phi(\gamma)}, \qquad \int_0^x \mathrm{e}^{-\gamma u}\, \mathrm{d}u = \frac{1 - \mathrm{e}^{-\gamma x}}{\gamma}.$$

Therefore, as $\gamma \to \infty$,

$$\frac{A}{B} = \frac{\Phi(\gamma - b)\phi(\gamma)}{\phi(\gamma - b)\Phi(\gamma)}, \qquad \frac{1}{B} = O(\mathrm{e}^{-(1/2)\gamma^2}), \qquad \int_0^x \mathrm{e}^{-\gamma u}\, \mathrm{d}u = O\left(\frac{1}{\gamma}\right).$$

Substituting in the threshold equation (23), we find that as $\gamma \to \infty$,

$$1 - \frac{\eta}{d} = \frac{\Phi(\gamma - b)\phi(\gamma)}{\phi(\gamma - b)\Phi(\gamma)} + O\left(\frac{1}{\gamma}\mathrm{e}^{-(1/2)\gamma^2}\right),$$

which completes the proof. □

Figure 5 displays η^{opt} given in Proposition 8 as a function of γ, together with the bounds given by Proposition 4,

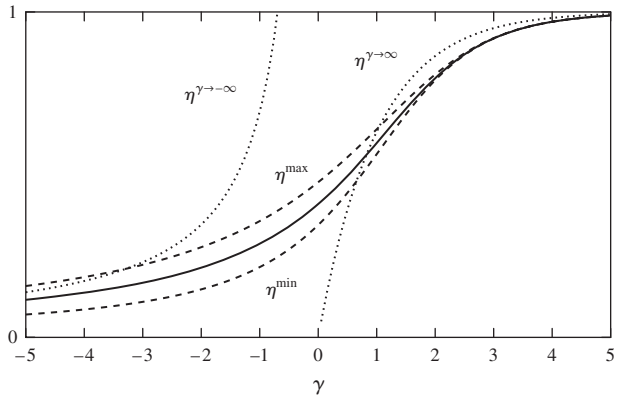
$$\eta^{\max} = d\left(1 - \frac{\Phi(\gamma - b)\phi(\gamma)}{\phi(\gamma - b)\Phi(\gamma)}\right), \qquad \eta^{\min} = d\left(1 - \frac{\Phi(\gamma - b)/\phi(\gamma - b) + \int_0^{\eta^{\max}} \mathrm{e}^{-\gamma x}\, \mathrm{d}x}{\Phi(\gamma)/\phi(\gamma) + \int_0^{\eta^{\max}} \mathrm{e}^{-\gamma x}\, \mathrm{d}x}\right),$$

and asymptotic solutions of Proposition 7,

$$\eta^{\gamma \to -\infty} = -\frac{1}{\gamma}\ln(1 + bd), \qquad \eta^{\gamma \to \infty} = d(1 - \mathrm{e}^{-b\gamma}).$$

Figure 5 also confirms the monotonicity of η^{opt} in γ established in Proposition 6. Note also the different regimes in which our approximations are valid, and that the bounds of Proposition 4 are tight as $\gamma \to \pm\infty$.

Figure 5. The optimal threshold η^{opt}, its bounds $\eta^{\min}$, $\eta^{\max}$, and its approximations $\eta^{\gamma \to \pm\infty}$, all as a function of γ, when $r_{\mathrm{L}}(x) = \exp(bx)$, $r_{\mathrm{R}}(x) = (d-x)/d$, and $b = d = 1$. The curve for the optimal threshold has been produced with (32).



3.4.2. Exponential Revenue. Consider $r_{\mathrm{L}}(x)$ arbitrary, and let $r_{\mathrm{R}}(x) = \exp(-\delta x)$ for $x \geq 0$, with $\delta > 0$. First, we will consider what happens asymptotically as $\delta \downarrow 0$ in the case $\gamma = 0$, which should be comparable to the case in Proposition 9. Then, we consider the case $\gamma = -\delta$, which like the linear revenue structure has a Lambert W solution. Finally, we consider what happens asymptotically when $\varepsilon = 1 - R_{\mathrm{T}}(x) > 0$ is small, and we check our results in the specific cases $\gamma = -2\delta$, $-\delta/2$ and $\gamma = \delta$, which have explicit solutions.

Proposition 11. *For $\gamma = 0$, as $\delta \downarrow 0$,*

$$\eta^{\mathrm{opt}} = \sqrt{\frac{2(B-A)}{\delta}} - \frac{2A+B}{3} + O(\sqrt{\delta}) = \sqrt{\frac{2}{\delta}\left(\sqrt{\frac{\pi}{2}} - \int_{-\infty}^{0} r(x)\mathrm{e}^{-(1/2)x^2}\,\mathrm{d}x\right)} + O(1). \tag{37}$$

Proof. When $\gamma = 0$, the threshold equation reads

$$\mathrm{e}^{\delta\eta} = 1 + \frac{\delta(B-A)}{1+A\delta} + \frac{\delta\eta}{1+A\delta}, \tag{38}$$

which follows from (22) with $\gamma = 0$. With $\delta > 0$, $\eta > 0$, the left-hand side of (38) exceeds $1 + \delta\eta + \delta^2\eta^2/2$, while the right-hand side is exceeded by $1 + \delta\eta + \delta(B-A)$. Therefore, the left-hand side of (38) exceeds the right-hand side if $\eta > \eta_*$, where $\eta_* = \sqrt{2(B-A)/\delta}$. This implies that $\eta^{\mathrm{opt}} \leq \eta^*$, and so we restrict attention to $0 \leq \eta \leq \eta_* = O(1/\sqrt{\delta})$ when considering (38). Expanding both sides of (38) gives

$$1 + \delta\eta + \frac{1}{2}\delta^2\eta^2 + \frac{1}{6}\delta^3\eta^3 + O(\delta^4\eta^4) = 1 + \delta(B-A) - \delta^2 A(B-A) + O(\delta^3) + \delta\eta - \delta^2 A\eta + O(\delta^3\eta). \tag{39}$$

Cancelling the terms $1 + \delta\eta$ at both sides of (39), and dividing by $\delta^2/2$ while remembering that $\eta = O(1/\sqrt{\delta})$, we get

$$\eta^2 = \frac{2(B-A)}{\delta} - 2\eta A - \frac{1}{3}\eta^3\delta + O(1).$$

Therefore,

$$\eta = \eta_*\left(1 - \frac{A\delta}{B-A}\eta - \frac{\delta^2}{6(B-A)}\eta^3 + O(\delta)\right)^{1/2} = \eta_*(1 + O(\sqrt{\delta})). \tag{40}$$

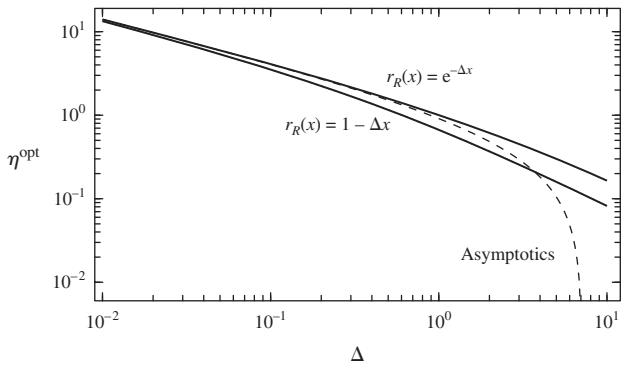
Thus $\eta = \eta_* + O(1)$, and inserting this in the right-hand side of the middle member in (40) yields

$$\eta = \eta_*\left(1 - \frac{A\delta}{B-A}\eta_* - \frac{\delta^2}{6(B-A)}\eta_*^3 + O(\delta)\right)^{1/2} = \eta_* - \frac{A\delta}{2(B-A)}\eta_*^2 - \frac{\delta^2}{12(B-A)}\eta_*^4 + O(\sqrt{\delta}) = \eta_* - \frac{2A+B}{3} + O(\sqrt{\delta}),$$

which is the result (37). □

Figure 6 draws for $\gamma = 0$ a comparison between $r_{\mathrm{R}}(x) = \exp(-\Delta x)$ and $r_{\mathrm{R}}(x) = 1 - \Delta x$. As expected, we see agreement when $\Delta \downarrow 0$, and for larger Δ the exponential revenue leads to slightly larger η^{opt} compared with linear revenues.

Figure 6. The optimal threshold η^{opt} in the exponential revenue case $r_{\mathrm{R}}(x) = \exp(-\Delta x)$, and in the linear revenue case $r_{\mathrm{R}}(x) = 1 - \Delta x$, as $\Delta \downarrow 0$. In both cases, $r'_{\mathrm{R}}(0+) = -\Delta$. The leading-order behavior established in Proposition 11 is also included.



When $\gamma = -\delta$, the threshold equation becomes

$$e^{-\delta\eta} - R_T(0) = \frac{\phi(-\delta)}{\Phi(-\delta)}\left(\eta - \frac{1-e^{-\delta\eta}}{\delta}\right),$$

or equivalently,

$$e^{-\delta\eta} = \frac{1}{\Phi(-\delta)/\phi(-\delta) - 1/\delta}\left(\eta - \frac{1}{\delta} + \frac{\Phi(-\delta)}{\phi(-\delta)}R_T(0)\right),$$

and the solution may again be expressed in terms the Lambert W function.

Proposition 12. *When* $\gamma = -\delta$, $\eta^{\text{opt}} = r_0 + (1/\delta)W(\delta e^{-\delta r_0}/a_0)$, *with*

$$a_0 = \frac{1}{\Phi(\gamma)/\phi(\gamma) - 1/\delta}, \qquad r_0 = \frac{1}{\delta} - \frac{\Phi(\gamma)}{\phi(\gamma)}R_T(0).$$

Proof. Immediate, since the standard form is $e^{-\delta\eta} = a_0(\eta - r_0)$. □

In case $\alpha = (\gamma+\delta)/\delta \neq 0, 1$, the threshold equation is given by, see (23),

$$e^{-\delta\eta} - R_T(0) = \frac{\delta}{B}\int_0^\eta e^{-\delta x}\frac{1-e^{-\gamma x}}{\gamma}\,dx = \frac{\delta}{B\gamma}\left(\frac{1-e^{-\delta\eta}}{\delta} - \frac{1-e^{-(\gamma+\delta)\eta}}{\gamma+\delta}\right), \tag{41}$$

After setting $z = e^{-\delta\eta} \in (0, 1]$, (41) takes the form

$$z - R_T(0) = \frac{1}{\gamma B}\left(1 - z - \frac{1-z^\alpha}{\alpha}\right). \tag{42}$$

Observe that the factor $1/(\gamma B)$ is positive when $\alpha > 1$, and negative when $\alpha < 1$. For values $\alpha = -1$, $1/2$, and 2, an explicit solution can be found in terms of the square-root function, see Proposition 19 in Appendix B. In all other cases, the solution is more involved. In certain regimes, however, a solution in terms of an infinite power series can be obtained, see Proposition 18 in Appendix B.

For illustrative purposes, we again plot the optimal threshold η^{opt} as a function of γ. It has been determined by numerically solving the threshold equation, and is plotted together with the bounds given by Proposition 4,

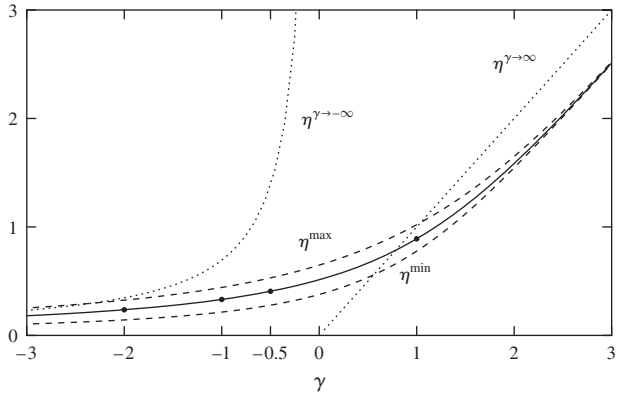
$$\eta^{\max} = -\frac{1}{d}\ln\left(\frac{g(\gamma-b)}{g(\gamma)}\right), \qquad \eta^{\min} = -\frac{1}{d}\ln\left(\frac{g(\gamma-b) + \int_0^{\eta^{\max}} e^{-\gamma x}\,dx}{g(\gamma) + \int_0^{\eta^{\max}} e^{-\gamma x}\,dx}\right),$$

and asymptotic solutions of Proposition 7,

$$\eta^{\gamma\to-\infty} = -\frac{1}{\gamma}\ln\left(1 + \frac{b}{d}\right), \qquad \eta^{\gamma\to\infty} = \frac{b\gamma}{d},$$

in Figure 7. Similar to Figure 5, Figure 7 also illustrates the monotonicity of η^{opt} in $\gamma \in \mathbb{R}$, the different regimes our approximations and bounds are valid, and how our bounds are tight as $\gamma \to \infty$. We have also indicated

Figure 7. The optimal threshold η^{opt}, its bounds $\eta^{\min}$, $\eta^{\max}$, and its approximations $\eta^{\gamma\to\pm\infty}$, all as a function of γ, when $r_L(x) = \exp(bx)$, $r_R(x) = \exp(-dx)$, and $b = d = 1$. The analytical solutions for $\alpha = -1$, 0, 1/2, and 2 provided by Propositions 12 and 19 are also indicated. The curve for the optimal threshold has been produced by numerically solving the threshold equation.



the analytical solutions for $\alpha = -1$, 0, 1/2, and 2, as provided by Propositions 12 and 19 in Appendix B. The asymptotic width 1/2 of the gap between the graphs of η^{opt} and $\eta^{\gamma\to\infty}$ is consistent with the refined asymptotics of η^{opt} as given below (31), case $b = d = 1$.

4. Optimality of Threshold Policies

We now present a proof of Proposition 2, the cornerstone for this paper that says that threshold policies are optimal, and that the optimal threshold satisfies the threshold equation. We first present in Section 4.1 a variational argument that gives an insightful way to derive Proposition 2 heuristically. Next, we present the formal proof of Proposition 2 in Section 4.2 using Hilbert-space theory.

4.1. Heuristic Based on a Variational Argument

For threshold controls $f(x) = \mathbb{1}[0 \le x < \eta]$ with $\eta \in [0,\infty)$, the QED limit of the long-term revenue (11) becomes (12). The optimal threshold η^{opt} can be found by equating

$$\frac{\mathrm{d}R}{\mathrm{d}\eta} = \frac{(B + (1-\mathrm{e}^{-\gamma\eta})/\gamma) r(\eta)\mathrm{e}^{-\gamma\eta} - (A + \int_0^\eta r(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x)\mathrm{e}^{-\gamma\eta}}{(B + (1-\mathrm{e}^{-\gamma\eta})/\gamma)^2} = \frac{r(\eta) - R_{\mathrm{T}}(\eta)}{\mathrm{e}^{\gamma\eta}(B + (1-\mathrm{e}^{-\gamma\eta})/\gamma)} \tag{43}$$

to zero, which shows that the optimal threshold η^{opt} solves the threshold equation (14), i.e., $r(\eta) = R_{\mathrm{T}}(\eta)$.

For any piecewise continuous function g on $[0,\infty)$ that is *admissible*, i.e., such that $0 \le f + \varepsilon g \le 1$ and $\int_0^\infty (f+\varepsilon g)\mathrm{e}^{-\gamma x}\,\mathrm{d}x < \infty$ for sufficiently small ε, define

$$\delta R(f;g) = \lim_{\varepsilon \downarrow 0} \frac{R(f+\varepsilon g) - R(f)}{\varepsilon}. \tag{44}$$

We call (44) the functional derivative of f with increment g, which can loosely be interpreted as a derivative of f in the direction of g, see Luenberger (1969) for background. Substituting (11) into (44) yields

$$\delta R(f;g) = \frac{\left(B + \int_0^\infty f\mathrm{e}^{-\gamma x}\,\mathrm{d}x\right)\int_0^\infty r g\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\left(B + \int_0^\infty f\mathrm{e}^{-\gamma x}\,\mathrm{d}x\right)^2} - \frac{\left(A + \int_0^\infty r f\mathrm{e}^{-\gamma x}\,\mathrm{d}x\right)\int_0^\infty g\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{\left(B + \int_0^\infty f\mathrm{e}^{-\gamma x}\,\mathrm{d}x\right)^2}. \tag{45}$$

Rewriting (45) gives

$$\delta R(f;g) = \frac{\int_0^\infty g(x)\mathrm{e}^{-\gamma x}[r(x) - R(f)]\,\mathrm{d}x}{B + \int_0^\infty f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}. \tag{46}$$

We can now examine the effect of small perturbations εg towards (or away from) policies f by studying the sign of (46). Specifically, it can be shown that for every perturbation g applied to the optimal threshold policy of Proposition 2, $\delta R(f^{\mathrm{opt}};g) \le 0$, indicating that these threshold policies are locally optimal. Moreover, it can be shown that for any other control f, a perturbation exists so that $\delta R(f;g) > 0$. Such other controls are therefore not locally optimal. Assuming the existence of an optimizer, these observations thus indeed indicate that the threshold control in Proposition 2 is optimal. We note that these observations crucially depend on the sign of $r(x) - R(f)$, as can be seen from (46). It is in fact the threshold equation (14) that specifies the point where a sign change occurs.

Note that while these arguments support Proposition 2, this section does not constitute a complete proof. In particular the existence of optimizers still needs to be established.

4.2. Formal Proof of Proposition 2

In the formal proof of Proposition 2 that now follows, we start by proving that there exist maximizers in Section 4.2.1. This ensures that our maximization problem is well-defined. In Section 4.2.2, we then derive necessary conditions for maximizers by perturbing the control towards (or away from) a threshold policy, as alluded to before, and in a formal manner using measure theory. Finally, we characterize in Section 4.2.3 the maximizers, by formally discarding pathological candidates.

With $r\colon \mathbb{R} \to [0,\infty)$ a smooth function, nonincreasing to 0 as $x \to \pm\infty$, and $\gamma \in \mathbb{R}$, recall that we are considering the maximization of the functional (11) with $f\colon [0,\infty) \to [0,1]$ is measurable and with $g(x) = f(x)\mathrm{e}^{-\gamma x} \in L^1([0,\infty))$. We do not assume f to be nonincreasing. Recall that $A = \int_{-\infty}^0 r(x)\exp(-\frac{1}{2}x^2 - \gamma x)\,\mathrm{d}x > 0$, $B = \Phi(\gamma)/\phi(\gamma) > 0$, and let $b(x) = \mathrm{e}^{-\gamma x}$ for $x \ge 0$. Then write $R(f)$ as

$$R(f) = \frac{A + \int_0^\infty r(x)g(x)\,\mathrm{d}x}{B + \int_0^\infty g(x)\,\mathrm{d}x} = L(g),$$

which is considered for all $g \in L^1([0,\infty))$ such that $0 \le g(x) \le b(x)$ for $0 \le x < \infty$. The objective is to maximize $L(g)$ over all such allowed g.

For notational convenience, write

$$L(g) = \frac{A}{B}\left(1 + \frac{\int_0^\infty s(x)g(x)\,dx}{1 + \int_0^\infty Sg(x)\,dx}\right), \tag{47}$$

where

$$s(x) = \frac{r(x)}{A} - \frac{1}{B}, \qquad S = \frac{1}{B}.$$

Recall that $r(x)$ is nonincreasing, implying that $s(x) \le s(0)$ for all $x \ge 0$. When $s(0) \le 0$, the maximum of (47) thus equals A/B, and is assumed by all allowed g that vanish outside the interval $[0, \sup\{x \in [0,\infty) \mid s(x) = 0\}]$. When $s(0) > 0$, define

$$x_0 = \inf\{x \in [0,\infty) \mid s(x) = 0\}, \tag{48}$$

which is positive and finite by smoothness of s and $r(x) \to 0$ as $x \to \infty$. Note that the set $\{x \in [0,\infty) \mid s(x) = 0\}$ consists of a single point when $r(x)$ is *strictly* decreasing as long as $r(x) > 0$. But even if $r(x)$ is not strictly decreasing, we have $s(x) \le 0$ for $x \ge x_0$. Because $g(x) \ge 0$ implies that

$$\int_{x_0}^\infty s(x)g(x)\,dx \le 0 \le \int_{x_0}^\infty Sg(x)\,dx, \tag{49}$$

we have

$$L(g\mathbb{I}[x \in [0,x_0)]) \ge L(g)$$

for all g. We may therefore restrict attention to allowed g supported on $[0,x_0]$. Such a g can be extended to any allowed function supported on $[0, \sup\{x \in [0,\infty) \mid s(x) = 0\}$ without changing the value $L(g)$. Therefore, we shall instead maximize

$$J(g) = \frac{\int_0^{x_0} s(x)g(x)\,dx}{1 + \int_0^{x_0} Sg(x)\,dx} \tag{50}$$

over all $g \in L^1([0,x_0])$ satisfying $0 \le g(x) \le b(x)$ for $0 \le x \le x_0$, in which $s(x)$ is a smooth function that is positive on $[0,x_0)$ and decreases to $s(x_0) = 0$.

4.2.1. Existence of Allowed Maximizers.

Proposition 13. *There exist maximizers $f^{\text{opt}} \in \mathcal{F}$ that maximize $R(f)$.*

Proof. We will use notions from the theory of Hilbert spaces and Lebesgue integration on the line. We consider maximization of $J(g)$ in (50) over all measurable g with $0 \le g(x) \le b(x)$ for a.e. $x \in [0,x_0]$.

For any $g \in L^1([0,x_0])$, the Lebesgue points of g, i.e., all $x_1 \in (0,x_0)$ such that

$$\lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} g(x_1 + x)\,dx \tag{51}$$

exists, is a subset of $[0,x_0]$ whose complement is a null set, and the limit in (51) agrees with $g(x_1)$ for a.e. $x_1 \in [0,x_0]$, Teschl (2014).

The set of allowed functions g is a closed and bounded set of the separable Hilbert space $L^2([0,x_0])$, and the functional $J(g)$ is bounded on this set. Hence, we can find a sequence of candidates $\{g_n\}_{n\in\mathbb{N}_0}$ of allowed g_n, such that

$$\lim_{n\to\infty} J(g_n) = \sup_{\text{allowed } g} \{J(g)\} < \infty.$$

We can subsequently find a subsequence $\{h_k\}_{k\in\mathbb{N}_0} = \{g_{n_k}\}_{k\in\mathbb{N}_0}$ such that h_k converges weakly to an $h \in L^2([0,x_0])$, Rudin (1987). Then

$$\sup_{\text{allowed } g} \{J(g)\} = \lim_{k\to\infty} J(h_k) = \lim_{k\to\infty} \frac{\int_0^\infty (r(x)/A - 1/B)h_k(x)\,dx}{1 + (1/B)\int_0^\infty h_k(x)\,dx} \overset{\text{(i)}}{=} \frac{\int_0^\infty (r(x)/A - 1/B)h(x)\,dx}{1 + (1/B)\int_0^\infty h(x)\,dx} = J(h), \tag{52}$$

where (i) follows from weak convergence. We now only need to show that h is allowed. We have for any $\varepsilon > 0$ and any $x_1 \in (0,x_0)$ by weak convergence that

$$\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} h(x_1 + x)\,dx = \lim_{k\to\infty} \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} h_k(x_1 + x)\,dx \in [0, b(x_1)],$$

since all h_k are allowed. Hence for all Lebesgue points $x_1 \in (0, x_0)$ of h we have

$$\lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} h(x_1 + x)\, \mathrm{d}x \in [0, b(x_1)],$$

and so $0 \le h(x_1) \le b(x_1)$ for a.e. $x_1 \in [0, x_0]$. This, together with (52) shows that h is an allowed maximizer. □

4.2.2. Necessary Condition for Maximizers.

Proposition 14. *For any maximizer $f^{\mathrm{opt}} \in \mathcal{F}$, $f(x) = 1$ if $r(x) > R(x)$, and $f(x) = 0$ if $r(x) < R(x)$.*

Proof. Let g be an allowed maximizer of $J(g)$. We shall equivalently show that for any Lebesgue point $x_1 \in (0, x_0)$ of g,

$$\frac{s(x_1)\big(1 + \int_0^{x_0} Sg(x)\, \mathrm{d}x\big)}{S \int_0^{x_0} s(x) g(x)\, \mathrm{d}x} > 1 \quad \Rightarrow \quad g(x_1) = b(x_1), \tag{53}$$

$$\frac{s(x_1)\big(1 + \int_0^{x_0} Sg(x)\, \mathrm{d}x\big)}{S \int_0^{x_0} s(x) g(x)\, \mathrm{d}x} < 1 \quad \Rightarrow \quad g(x_1) = 0. \tag{54}$$

Let $x_1 \in (0, x_0)$ be any Lebesgue point of g and assume that

$$\frac{s(x_1)\big(1 + \int_0^{x_0} Sg(x)\, \mathrm{d}x\big)}{S \int_0^{x_0} s(x) g(x)\, \mathrm{d}x} > 1. \tag{55}$$

Suppose that $g(x_1) < b(x_1)$. We shall derive a contradiction. Let $\varepsilon_0 > 0$ be small enough so that

$$\tfrac{1}{2}(g(x_1) + b(x_1)) \le \min_{x_1 - \varepsilon_0 \le x \le x_1 + \varepsilon_0} \{b(x)\}. \tag{56}$$

Along with g, consider for $0 < \varepsilon \le \varepsilon_0$ the function

$$g_\varepsilon(x) = \begin{cases} g(x), & x \notin [x_1 - \varepsilon, x_1 + \varepsilon], \\ \tfrac{1}{2}(g(x_1) + b(x_1)), & x \in [x_1 - \varepsilon, x_1 + \varepsilon]. \end{cases}$$

This g_ε is allowed by (56). Write $J(g)$ as

$$J(g) = \frac{C(\varepsilon) + I_s(\varepsilon; g)}{D(\varepsilon) + I_S(\varepsilon; g)},$$

where

$$C(\varepsilon) = \left(\int_0^{y-\varepsilon} + \int_{y+\varepsilon}^{x_0} \right) s(x) g(x)\, \mathrm{d}x, \qquad D(\varepsilon) = 1 + \left(\int_0^{y-\varepsilon} + \int_{y+\varepsilon}^{x_0} \right) Sg(x)\, \mathrm{d}x,$$

and

$$I_s(\varepsilon; g) = \int_{x_1 - \varepsilon}^{x_1 + \varepsilon} s(x) g(x)\, \mathrm{d}x, \qquad I_S(\varepsilon; g) = \int_{x_1 - \varepsilon}^{x_1 + \varepsilon} Sg(x)\, \mathrm{d}x. \tag{57}$$

We can do a similar thing with $J(g_\varepsilon)$, using the same numbers $C(\varepsilon)$ and $D(\varepsilon)$ and g replaced by g_ε in (57). We compute

$$J(g_\varepsilon) - J(g) = \frac{(C(\varepsilon) + I_s(\varepsilon; g_\varepsilon))(D(\varepsilon) + I_S(\varepsilon; g)) - (C(\varepsilon) + I_s(\varepsilon; g))(D(\varepsilon) + I_S(\varepsilon; g_\varepsilon))}{(D(\varepsilon) + I_S(\varepsilon; g))(D(\varepsilon) + I_S(\varepsilon; g_\varepsilon))}, \tag{58}$$

in which the numerator $N(g_\varepsilon; g)$ of the fraction at the right-hand side of (58) can be written as

$$N(g_\varepsilon; g) = (I_s(\varepsilon; g_\varepsilon) - I_s(\varepsilon; g)) D(\varepsilon) - (I_S(\varepsilon; g_\varepsilon) - I_S(\varepsilon; g)) C(\varepsilon) + I_s(\varepsilon; g_\varepsilon) I_S(\varepsilon; g) - I_s(\varepsilon; g) I_S(\varepsilon; g_\varepsilon).$$

Since x_1 is a Lebesgue point of g, we have as $\varepsilon \downarrow 0$

$$\frac{1}{2\varepsilon} I_s(\varepsilon; g_\varepsilon) \to \frac{1}{2} s(x_1)(g(x_1) + b(x_1)), \qquad \frac{1}{2\varepsilon} I_s(\varepsilon; g) \to s(x_1) g(x_1), \tag{59}$$

$$\frac{1}{2\varepsilon} I_S(\varepsilon; g_\varepsilon) \to \frac{1}{2} S(g(x_1) + b(x_1)), \qquad \frac{1}{2\varepsilon} I_S(\varepsilon; g) \to Sg(x_1), \tag{60}$$

while also

$$C(\varepsilon) \to \int_0^{x_0} s(x)g(x)\,\mathrm{d}x, \qquad D(\varepsilon) \to 1 + \int_0^{x_0} Sg(x)\,\mathrm{d}x.$$

Therefore,

$$\lim_{\varepsilon \downarrow 0} N(g_\varepsilon, g) = \frac{1}{2}(b(x_1) - g(x_1))\left(s(x_1)\left(1 + \int_0^{x_0} Sg(x)\,\mathrm{d}x\right) - S\int_0^{x_0} s(x)g(x)\,\mathrm{d}x\right) > 0$$

by assumption (55). Then $J(g_\varepsilon) - J(g) > 0$ when ε is sufficiently small, contradicting maximality of $J(g)$. Hence, we have proven the first relation in (54). The proof of the second relation is similar. □

4.2.3. Characterization of Maximizers. Proposition 14 does not exclude the possibility that a maximizer alternates between 0 and 1. Proposition 15 solves this problem by excluding the pathological candidates.

Proposition 15. *The quantity*

$$R(f;\eta) = \frac{A + \int_0^\eta r(x)f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}{B + \int_0^\eta f(x)\mathrm{e}^{-\gamma x}\,\mathrm{d}x}.$$

is uniquely maximized by

$$f(x) = \mathbb{1}[0 \le x \le \eta^{\mathrm{opt}}],$$

with η^{opt} a solution of the equation $r(\eta) = R_{\mathrm{T}}(\eta)$, apart from null functions and its value at any solution of $r(\eta) = R_{\mathrm{T}}(\eta)$.

Proof. Assume that g is a maximizer, and consider the continuous, decreasing function

$$t_g(x_1) = s(x_1)\left(1 + \int_0^{x_0} Sg(x)\,\mathrm{d}x\right) - S\int_0^{x_0} s(x)g(x)\,\mathrm{d}x,$$

which is positive at $x_1 = 0$ and negative (because $g \neq 0$) at $x_1 = x_0$ since s is decreasing with $s(0) > 0 = s(x_0)$. Let $x_{2,g}$, $x_{3,g}$ be such that $0 < x_{2,g} \le x_{3,g} < x_0$ and

$$t_g(x_1) = \begin{cases} > 0, & 0 \le x_1 < x_{2,g}, \\ = 0, & x_{2,g} \le x_1 \le x_{3,g}, \\ < 0, & x_{3,g} < x_1 \le x_0. \end{cases}$$

Note that $x_{2,g} = x_{3,g}$ when s is strictly decreasing on $[0, x_0]$, and that $s'(x) = 0$ for $x \in [x_{2,g}, x_{3,g}]$ when $x_{2,g} < x_{3,g}$. According to Janssen et al. (2013), we have

$$g(x_1) = b(x_1), \quad \text{a.e. } x_1 \in [0, x_{2,g}], \qquad \text{and} \qquad g(x_1) = 0, \quad \text{a.e. } x_1 \in [x_{3,g}, x_0]. \tag{61}$$

For an allowed $h \neq 0$, consider the continuous function

$$J(h;x_1) = \frac{\int_0^{x_1} s(x)h(x)\,\mathrm{d}x}{1 + \int_0^{x_1} Sh(x)\,\mathrm{d}x}, \quad 0 \le x_1 \le x_0. \tag{62}$$

We differentiate $J(h;x_1)$ with respect to x_1, where we use the fact that for any $k \in L^1([0,x_0])$,

$$\frac{\mathrm{d}}{\mathrm{d}x_1}\left[\int_0^{x_1} k(x)\,\mathrm{d}x\right] = k(x_1), \quad \text{a.e. } x_1 \in [0, x_0].$$

Thus we get for a.e. x_1 that

$$\frac{\mathrm{d}}{\mathrm{d}x_1}[J(h;x_1)] = \frac{N_h(x_1)}{D_h(x_1)}, \tag{63}$$

where $D_h(x_1) = (1 + \int_0^{x_0} Sh(x)\,\mathrm{d}x)^2$, and

$$N_h(x_1) = h(x_1)M_h(x_1) \tag{64}$$

with

$$M_h(x_1) = s(x_1)\left(1 + \int_0^{x_1} Sh(x)\,\mathrm{d}x\right) - S\int_0^{x_1} s(x)h(x)\,\mathrm{d}x.$$

Now $M_h(x_1)$ is a continuous function of $x_1 \in [0, x_0]$ with $M_h(x_0) < 0 < M_h(0)$ since $s(x_0) = 0 < s(0)$ and $h \neq 0$. Furthermore, $M_h(x_1)$ is differentiable at a.e. x_1, and one computes for a.e. x_1,

$$\frac{\mathrm{d}}{\mathrm{d}x_1}[M_h(x_1)] = s'(x_1)\left(1 + \int_0^{x_1} Sh(x)\,\mathrm{d}x\right). \tag{65}$$

Since s is decreasing, the right-hand side of (65) is nonpositive for all x_1 and negative for all x_1 with $s'(x_1) < 0$.

Now let g be a maximizer, and consider first the case that $s(x)$ is strictly decreasing. Then $x_{2,g} = x_{3,g}$ in (61). Next consider $h = b$ in (62) and further. It follows from (65) that M_b is strictly decreasing on $[0, x_0]$, and so M_b has a unique zero $\hat{x}$ on $[0, x_0]$. Therefore, by (63) and (64), $J(b; x_1)$ has a unique maximum at $x_1 = \hat{x}$. Then, from (61) and maximality of g, $x_{2,g} = \hat{x} = x_{3,g}$. Hence, J is uniquely maximized by

$$g(x_1) = b(x_1)\mathbb{1}[x_1 \in [0, \hat{x}]], \tag{66}$$

apart from null functions, with $\hat{x}$ the unique solution y of the equation

$$s(y)\left(1 + \int_0^{y} Sb(x)\,\mathrm{d}x\right) - S\int_0^{y} s(x)b(x)\,\mathrm{d}x = 0. \tag{67}$$

This handles the case that s is strictly decreasing.

When s' may vanish, we have to argue more carefully. In the case that $x_{2,g} = x_{3,g}$, we can proceed as earlier, with (66) emerging as maximizer and $x_{2,g} = y = x_{3,g}$. So assume we have a maximizer g with $x_{2,g} < x_{3,g}$, and consider $h = g$ in (62) and further. We have that $J(h = g; x_1)$ is constant in $x_1 \in [x_{3,g}, x_0]$. Furthermore, from $s'(x_1) = 0$ for $x_1 \in [x_{2,g}, x_{3,g}]$ and (64), we see that $J(h = g; x_1)$ is constant in $x_1 \in [x_{2,g}, x_{3,g}]$ as well. This constant value equals $J(g)$, and is equal to $J(b\mathbb{1}[x_1 \in [0, x_{2,g}]])$ since, due to (64), we have $J(g; \cdot) = J(\bar{g}; \cdot)$ when $g = \bar{g}$ a.e. outside $[x_{2,g}, x_{3,g}]$. We are then again in the previous situation, and the solutions y of (67) form now a whole interval $[y_2, y_3]$. The maximizers are again unique, apart from their values for $x_1 \in [y_2, y_3]$ that can be chosen arbitrarily between 0 and $b(x_1)$. □

5. Conclusions and Future Perspectives

The QED regime has gained tremendous popularity in the operations management literature, because it describes how large-scale service operations can achieve high system utilization while simultaneously maintaining short delays. Operating a system in the QED regime typically entails hiring a number of servers according to the square-root staffing rule $s = \lambda/\mu + \gamma\sqrt{\lambda/\mu}$, and has the added benefit that limiting performance measures can be described by elementary functions of just the one parameter γ. Through the square-root staffing rule, γ determines a hedge against variability or overcapacity, which is of the order of the natural fluctuations of the demand per time unit when the system operates in the QED regime. Classical problems of dimensioning large-scale systems in the QED regime can then be solved by optimizing objective functions solely dependent on γ.

Our paper adds a revenue maximization framework that complies with the classical dimensioning of QED systems by constructing scalable admission controls and revenue structures that remain meaningful in the QED regime (Proposition 1). As we have proven, our revenue framework naturally leads to an optimal control that bars customers from entering when the queue length of delayed customers exceeds the threshold $\eta^{\mathrm{opt}}\sqrt{s}$, provided that η^{opt} satisfies a fundamental *threshold equation* (Proposition 2). A detailed study of this threshold equation made it possible to characterize η^{opt} in terms of exact expressions, bounds, and asymptotic expansions. The weak assumptions made throughout this paper allow for application to a rich class of revenue structures, and an interesting direction for future work would therefore be the construction of realistic revenue structures based on specific case studies, expert opinions, or calibration to financial data.

Let us finally discuss the fascinating interplay between the parameters γ and η, which suggest that they act as communicating yet incomparable vessels. The optimal threshold η^{opt} increases with the overcapacity γ. Since more overcapacity roughly means fewer customers per server, and a larger threshold means more customers per server, we see that the optimization of revenues over the pair (γ, η) gives rise to an intricate two-dimensional framework in which the two parameters have radically different yet persistent effects in the QED regime. At the process level, the γ acts as a negative drift in the entire state space, while the η only interferes at the upper limit of the state space. Hence, while in this paper we have treated γ as given, and mostly focused on the behavior of the new parameter η, our framework paves the way for two-dimensional joint staffing and admission control problems. Gaining a deeper understanding of this interplay, and in relation to specific revenue structures, is a promising direction for future research.

Appendix A. Limiting Behavior of Long-Term QED Revenue

With $r_s(k) = r((k-s)/\sqrt{s})$ as in (2) and $\pi_s(k) = \lim_{t\to\infty} \mathbb{P}[Q_s(t) = k]$, (6), where p_s and f are related as in (3), we compute for $\rho = 1 - \gamma/\sqrt{s} > 0$,

$$\sum_{k=0}^{\infty} r_s(k)\pi_s(k) = \frac{\sum_{k=0}^{s} r((k-s)/\sqrt{s})((s\rho)^k/k!) + ((s\rho)^s/s!)\sum_{k=s+1}^{\infty} r((k-s)/\sqrt{s})\rho^{k-s} f((k-s)/\sqrt{s})}{\sum_{k=0}^{s}(s\rho)^k/k! + (s\rho)^s/s!\sum_{k=s+1}^{\infty}\rho^{k-s} f((k-s)/\sqrt{s})}.$$

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

$$\sum_{k=0}^{\infty} r_s(k)\pi_s(k) = \frac{W_s^{\mathrm{L}}(\rho) + W_s^{\mathrm{R}}(\rho)}{B_s^{-1}(\rho) + F_s(\rho)}.$$

Here,

$$B_s(\rho) = \frac{(s\rho)^s/s!}{\sum_{k=0}^{s}(s\rho)^k/k!}$$

is the Erlang B formula,

$$F_s(\rho) = \sum_{n=0}^{\infty} \rho^{n+1} f\left(\frac{n+1}{\sqrt{s}}\right)$$

as in (8), and

$$W_s^{\mathrm{L}}(\rho) = \sum_{k=0}^{s} r\left(\frac{k-s}{\sqrt{s}}\right)\frac{s!(s\rho)^{k-s}}{k!}, \tag{A.1}$$

$$W_s^{\mathrm{R}}(\rho) = \sum_{n=0}^{\infty} r\left(\frac{n+1}{\sqrt{s}}\right)\rho^{n+1} f\left(\frac{n+1}{\sqrt{s}}\right). \tag{A.2}$$

with superscripts L and R referring to the left-hand part $k = 0, 1, \ldots, s$ and right-hand part $k = s+1, s+2, \ldots$ of the summation range, respectively.

From Jagerman's asymptotic results for Erlang B, there is the approximation (Jagerman 1974, Theorem 14)

$$B_s^{-1}(\rho) = \sqrt{s}\psi(\gamma) + \chi(\gamma) + O\left(\frac{1}{\sqrt{s}}\right) \tag{A.3}$$

with $\psi(\gamma) = \Phi(\gamma)/\phi(\gamma)$ and $\chi(\gamma)$ expressible in terms of ϕ and Φ as well. For $F_s(\rho)$ there is the approximation (Janssen et al. 2013, Theorem 4.2),

$$F_s(\rho) = \sqrt{s}\mathscr{L}(\gamma) + \mathscr{M}(\gamma) + O\left(\frac{1}{\sqrt{s}}\right), \tag{A.4}$$

with $\mathscr{L}(\gamma) = \int_0^{\infty} f(x)\exp(-\gamma x)\,\mathrm{d}x$ and $\mathscr{M}(\gamma)$ expressible in terms of $\mathscr{L}'(\gamma)$. We aim at similar approximations for $W_s^{\mathrm{L}}(\rho)$ and $W_s^{\mathrm{R}}(\rho)$ in (A.1), (A.2).

We start by considering $W_s^{\mathrm{R}}(\rho)$ for the case that r and its first two derivatives are continuous and bounded in the two following situations:

(i) f is smooth; $f(y)\exp(-\gamma y)$ and its first two derivatives are exponentially small as $y \to \infty$;
(ii) $f = \mathbb{1}[x \in [0, \eta]]$ with $\eta > 0$.

A.1. Asymptotics of $W_s^{\mathrm{R}}(\rho)$

In the series expression for $W_s^{\mathrm{R}}(\rho)$, we have

$$\rho^{n+1} = \left(1 - \frac{\gamma}{\sqrt{s}}\right)^{n+1} = \mathrm{e}^{-(n+1)\gamma_s/\sqrt{s}}$$

with

$$\gamma_s = -\sqrt{s}\ln\left(1 - \frac{\gamma}{\sqrt{s}}\right) = \gamma + \frac{\gamma^2}{2\sqrt{s}} + \cdots > \gamma. \tag{A.5}$$

Hence, the conditions in case (i) are also valid when using γ_s instead of γ.

We obtain the following result.

Lemma 1. *For case* (i) *it holds that*

$$W_s^{\mathrm{R}}(\rho) = \sqrt{s}\int_0^{\infty} \mathrm{e}^{-\gamma_s y} r(y) f(y)\,\mathrm{d}y - \frac{1}{2} r(0) f(0) + O\left(\frac{1}{\sqrt{s}}\right). \tag{A.6}$$

For case (ii) *it holds that*

$$W_s^{\mathrm{R}}(\rho) = \sqrt{s}\int_0^{\eta} \mathrm{e}^{-\gamma_s y} r(y)\,\mathrm{d}y - \frac{1}{2} r(0) + \left(\lfloor \eta\sqrt{s}\rfloor - \left(\eta\sqrt{s} - \frac{1}{2}\right)\right)\mathrm{e}^{-\gamma_s \eta} r(\eta) + O\left(\frac{1}{\sqrt{s}}\right). \tag{A.7}$$

Proof. We use EM-summation as in Janssen et al. (2013, Appendix C), first instance in Janssen et al. (2013, (C.1)), case $m=1$, with the function

$$h(x)=g\left(\frac{x+1/2}{\sqrt{s}}\right),\quad x\geq 0,\qquad \text{and}\qquad g(y)=\mathrm{e}^{-\gamma_s y}r(y)f(y),\quad y\geq 0,$$

using a finite summation range $n=0,1,\ldots,N$, where we take $N=s$ in case (i) and $N=\lfloor\eta\sqrt{s}-3/2\rfloor$ in case (ii) In both cases, we have by smoothness of h on the range $[0,N+1]$ that

$$\sum_{n=0}^{N}h\left(n+\frac{1}{2}\right)=\int_0^{N+1}h(x)\,\mathrm{d}x+\frac{1}{2}B_2\left(\frac{1}{2}\right)(h^{(1)}(N+1)-h^{(1)}(0))+R, \tag{A.8}$$

where $|R|\leq\frac{1}{2}B_2\int_0^{N+1}|h^{(2)}(x)|\,\mathrm{d}x$. Due to our assumptions, it holds in both cases that

$$\frac{1}{2}B_2\left(\frac{1}{2}\right)(h^{(1)}(N+1)-h^{(1)}(0))+R=O\left(\frac{1}{\sqrt{s}}\right).$$

In case (i), the left-hand side of (A.8) equals $W_s^{\mathrm{R}}(\rho)$, apart from an error that is exponentially small as $s\to\infty$. In case (ii), the left-hand side of (A.8) and $W_s^{\mathrm{R}}(\rho)$ are related according to

$$W_s^{\mathrm{R}}(\rho)=\sum_{n=0}^{N}h\left(n+\frac{1}{2}\right)+g\left(\frac{\lfloor\eta\sqrt{s}\rfloor}{\sqrt{s}}\right)\left(\lfloor\eta\sqrt{s}\rfloor-\left\lfloor\eta\sqrt{s}-\frac{1}{2}\right\rfloor\right). \tag{A.9}$$

The second term at the right-hand side of (A.9) equals 0 or $g(\lfloor\eta\sqrt{s}\rfloor/\sqrt{s})$ accordingly as $\eta\sqrt{s}-\lfloor\eta\sqrt{s}\rfloor\geq$ or $<\frac{1}{2}$, i.e., accordingly as $N+1=\lfloor\eta\sqrt{s}\rfloor$ or $\lfloor\eta\sqrt{s}\rfloor-1$. Next, by smoothness of h and g on the relevant ranges, we have

$$\int_0^{N+1}h(x)\,\mathrm{d}x=\sqrt{s}\int_{1/(2\sqrt{s})}^{(N+3/2)/\sqrt{s}}g(y)\,\mathrm{d}y=\sqrt{s}\int_0^{(N+3/2)/\sqrt{s}}g(y)\,\mathrm{d}y-\frac{1}{2}g(0)+O\left(\frac{1}{\sqrt{s}}\right).$$

In case (i), we have that $\int_{(N+3/2)/\sqrt{s}}^{\infty}g(y)\,\mathrm{d}y$ is exponentially small as $s\to\infty$, since $N=s$, and this yields (A.6). In case (ii), we have

$$\begin{aligned}\int_0^{(N+3/2)/\sqrt{s}}g(y)\,\mathrm{d}y-\int_0^{\eta}g(y)\,\mathrm{d}y=\int_{\eta}^{(N+3/2)/\sqrt{s}}g(y)\,\mathrm{d}y&=\left(\frac{N+3/2}{\sqrt{s}}-\eta\right)g\left(\frac{\lfloor\eta\sqrt{s}\rfloor}{\sqrt{s}}\right)+O\left(\frac{1}{s}\right)\\&=\frac{1}{\sqrt{s}}\left(\left\lfloor\eta\sqrt{s}-\frac{1}{2}\right\rfloor-\left(\eta\sqrt{s}-\frac{1}{2}\right)\right)g\left(\frac{\lfloor\eta\sqrt{s}\rfloor}{\sqrt{s}}\right)+O\left(\frac{1}{s}\right),\end{aligned}$$

and with (A.9), this yields (A.7). This completes the proof. □

We denote for both case (i) and (ii)

$$\mathscr{L}_{rf}(\delta)=\int_0^{\infty}\mathrm{e}^{-\delta y}r(y)f(y)\,\mathrm{d}y \tag{A.10}$$

with $\delta\in\mathbb{R}$ such that the integral of the right-hand side of (A.10) converges absolutely. From (A.5) it is seen that, with the prime $'$ denoting differentiation,

$$\mathscr{L}_{rf}(\gamma_s)=\mathscr{L}_{rf}(\gamma)+\frac{\gamma^2}{2\sqrt{s}}\mathscr{L}'_{rf}(\gamma)+O\left(\frac{1}{s}\right).$$

Thus we get from Lemma 1 the following result.

Proposition 16. *For case* (i) *it holds that*

$$W_s^{\mathrm{R}}(\rho)=\sqrt{s}\mathscr{L}_{rf}(\gamma)+\frac{1}{2}\gamma^2\mathscr{L}'_{rf}(\gamma)-\frac{1}{2}r(0)f(0)+O\left(\frac{1}{\sqrt{s}}\right).$$

For case (ii) *it holds that*

$$W_s^{\mathrm{R}}(\rho)=\sqrt{s}\mathscr{L}_{rf}(\gamma)+\frac{1}{2}\gamma^2\mathscr{L}'_{rf}(\gamma)-\frac{1}{2}r(0)+\left(\lfloor\eta\sqrt{s}\rfloor-\left(\eta\sqrt{s}-\frac{1}{2}\right)\right)\mathrm{e}^{-\gamma\eta}r(\eta)+O\left(\frac{1}{\sqrt{s}}\right).$$

A.2. Asymptotics of $W_s^{\mathrm{L}}(\rho)$

We next consider $W_s^{\mathrm{L}}(\rho)$ for the case that $r\colon(-\infty,0]\to\mathbb{R}$ has bounded and continuous derivatives up to order 2. Using a change of variables, we write

$$W_s^{\mathrm{L}}(\rho)=r(0)+\sum_{k=1}^{s}r\left(\frac{-k}{\sqrt{s}}\right)\frac{s!s^{-k}}{(s-k)!}\rho^{-k}, \tag{A.11}$$

and we again intend to apply EM-summation to the series at the right-hand side of (A.11). We first present a bound and an approximation.

Lemma 2. *We have for* $|\gamma|/\sqrt{s} \le \frac{1}{2}$ *and* $\rho = 1-\gamma/\sqrt{s}$,

$$\frac{s!s^{-k}}{(s-k)!}\rho^{-k} \le \exp\left(-\frac{k(k-1)}{2s}+\frac{\gamma k}{\sqrt{s}}+\frac{\gamma^2 k}{s}\right), \quad k=1,2,\dots,s, \tag{A.12}$$

and

$$\frac{s!s^{-k}}{(s-k)!}\rho^{-k} = G_s\left(\frac{k}{\sqrt{s}}\right)\left(1+O\left(\frac{1}{s}P_6\left(\frac{k}{\sqrt{s}}\right)\right)\right), \quad k \le s^{2/3}, \tag{A.13}$$

where

$$G_s(y) = \mathrm{e}^{-(1/2)y^2+\gamma y}\left(1-\frac{1}{6\sqrt{s}}y^3+\frac{1}{2\sqrt{s}}(1+\gamma^2)y\right), \tag{A.14}$$

and $P_6(y)$ *is a polynomial in* y *of degree* 6 *with coefficients bounded by* 1 *(the constant implied by* $O(\cdot)$ *depends on* γ*).*

Proof. We have for $k=1,2,\dots,s$ and $|\gamma|/\sqrt{s} \le 1/2$, $\rho = 1-\gamma/\sqrt{s}$,

$$\begin{aligned}\frac{s!s^{-k}}{(s-k)!}\rho^{-k} &= \rho^{-k}\prod_{j=0}^{k-1}\left(1-\frac{j}{s}\right) = \exp\left(\sum_{j=0}^{k-1}\ln\left(1-\frac{j}{s}\right)-k\ln\left(1-\frac{\gamma}{\sqrt{s}}\right)\right)\\ &\le \exp\left(-\sum_{j=0}^{k-1}\frac{j}{s}+\frac{\gamma k}{\sqrt{s}}+\frac{\gamma^2 k}{s}\right) = \exp\left(-\frac{k(k-1)}{2s}+\frac{\gamma k}{\sqrt{s}}+\frac{\gamma^2 k}{s}\right),\end{aligned}$$

where it has been used that $-\ln(1-x) \le x+x^2$, $|x| \le 1/2$.

On the range $k \le s^{2/3}$, we further expand

$$\begin{aligned}\frac{s!s^{-k}}{(s-k)!}\rho^{-k} &= \exp\left(-\sum_{j=0}^{k-1}\left(\frac{j}{s}+\frac{j^2}{2s^2}+O\left(\frac{j^3}{s^3}\right)\right)+\frac{\gamma k}{\sqrt{s}}+\frac{\gamma^2 k}{s}+O\left(\frac{k}{s^{3/2}}\right)\right)\\ &= \exp\left(-\frac{k(k-1)}{2s}-\frac{k(k-1)(2k-1)}{12s^2}+O\left(\frac{k^4}{s^3}\right)+\frac{\gamma k}{\sqrt{s}}+\frac{\gamma^2 k}{s}+O\left(\frac{k}{s^{3/2}}\right)\right)\\ &= \exp\left(-\frac{k^2}{2s}+\frac{\gamma k}{\sqrt{s}}-\frac{k^3}{6s^2}+\frac{1}{2}(1+\gamma^2)\frac{k}{s}+O\left(\frac{k}{s^{3/2}}+\frac{k^2}{s^2}+\frac{k^4}{s^3}\right)\right)\end{aligned}$$

On the range $0 \le k \le s^{2/3}$ we have

$$\frac{k^3}{s^2}, \frac{k}{s}, \frac{k}{s^{3/2}}, \frac{k^2}{s^2}, \frac{k^4}{s^3} = O(1).$$

Hence, on the range $0 \le k \le s^{2/3}$,

$$\frac{s!s^{-k}}{(s-k)!}\rho^{-k} = \exp\left(-\frac{k^2}{2s}+\frac{\gamma k}{\sqrt{s}}\right)\left(1-\frac{k^3}{6s^2}+\frac{1}{2}(1+\gamma^2)\frac{k}{s}+O\left(\frac{k^6}{s^4}+\frac{k^4}{s^3}+\frac{k^2}{s^2}+\frac{k}{s^{3/2}}\right)\right) = G\left(\frac{k}{\sqrt{s}}\right)\left(1+O\left(\frac{1}{s}P_6\left(\frac{k}{\sqrt{s}}\right)\right)\right),$$

where $P_6(y) = y^6+y^4+y^2+y$. □

Proposition 17. *It holds that*

$$W_s^{\mathrm{L}}(\rho) = \sqrt{s}\int_{-\infty}^{0}\mathrm{e}^{-(1/2)y^2-\gamma y}r(y)\,\mathrm{d}y+\frac{1}{2}r(0)+\int_{-\infty}^{0}\mathrm{e}^{-(1/2)y^2-\gamma y}\left(\frac{1}{6}y^3-\frac{1}{2}(1+\gamma^2)y\right)r(y)\,\mathrm{d}y+O\left(\frac{1}{\sqrt{s}}\right).$$

Proof. With $v(y) = r(-y)$, we write

$$W_s^{\mathrm{L}}(\rho) = r(0)+\sum_{n=0}^{s-1}v\left(\frac{n+1}{\sqrt{s}}\right)\frac{s!s^{-n-1}}{(s-n-1)!}\rho^{-n-1}. \tag{A.15}$$

By the assumptions on r and the bound in (A.12), the contribution of the terms in the series in (A.15) with $n > s^{2/3}$ is $O(\exp(-Cs^{1/3}))$, $s\to\infty$, for any C with $0<C<1/2$. On the range $n=0,1,\dots,\lfloor s^{2/3}\rfloor-1=:N$, we can apply (A.13), and so, with exponentially small error,

$$W_s^{\mathrm{L}}(\rho) = r(0)+\sum_{n=0}^{N}v\left(\frac{n+1}{\sqrt{s}}\right)G_s\left(\frac{n+1}{\sqrt{s}}\right)\left(1+O\left(\frac{1}{s}P_6\left(\frac{n+1}{\sqrt{s}}\right)\right)\right).$$

By EM-summation, as used in the proof of Lemma 1 for the case (i) as considered there, we have

$$\sum_{n=0}^{N}v\left(\frac{n+1}{\sqrt{s}}\right)G_s\left(\frac{n+1}{\sqrt{s}}\right) = \sqrt{s}\int_0^{\infty}v(y)G_s(y)\,\mathrm{d}y-\frac{1}{2}v(0)G_s(0)+O\left(\frac{1}{\sqrt{s}}\right),$$

where we have extended the integration range $[0,(N+3/2)/\sqrt{s}]$ to $[0,\infty)$ at the expense of exponentially small error. Then the result follows on a change of the integration variable, noting that $v(y)=r(-y)$ and the definition of G_s in (A.14), implying $G_s(0)=1$. □

The result of Proposition 1 in the main text follows now from (A.3), (A.4), Propositions 16 and 17, by considering leading terms only.

Appendix B. Explicit Solutions for Exponential Revenue

Proposition 18. *When $\varepsilon = 1 - R_{\mathrm{T}}(0) > 0$ is sufficiently small,*

$$\eta^{\mathrm{opt}} = -\frac{1}{\delta}\ln\left(1 - \sum_{l=1}^{\infty} a_l \varepsilon^l\right), \tag{B.1}$$

where

$$a_1 = 1, \qquad a_2 = \frac{1}{2}(\alpha + \beta - 1), \tag{B.2}$$

$$a_{l+1} = \frac{1}{l+1}\left((l\alpha + (l+1)\beta - 1)a_l + \beta\sum_{i=2}^{l-1} i a_i a_{l+1-i}\right), \quad l = 2, 3, \ldots, \tag{B.3}$$

with $\beta = (1-\alpha)(1 + 1/(\gamma B))$ and the convention that $\sum_{i=2}^{l-1} = 0$ for $l = 2$.

Proof. With $\varepsilon = 1 - R_{\mathrm{T}}(0)$ and $w = 1 - z$, we can write (42) as

$$H(w) = w + \frac{1}{\gamma B}\left(w - \frac{1}{\alpha}(1 - (1-w)^{\alpha})\right) = \varepsilon. \tag{B.4}$$

Note that

$$H(w) = w + \frac{1}{\gamma B}\left(\frac{1}{2}(\alpha-1)w^2 - \frac{1}{6}(\alpha-1)(\alpha-2)w^3 + \cdots\right), \quad |w| < 1,$$

and so there is indeed a (unique) solution

$$w(\varepsilon) = \varepsilon + \sum_{l=2}^{\infty} a_l \varepsilon^l \tag{B.5}$$

of (B.4) when $|\varepsilon|$ is sufficiently small. To find the a_l we let $D = 1/(\gamma B)$, and we write (B.4) as

$$(1 + D)w - \frac{1}{\alpha}D + \frac{1}{\alpha}(1-w)^{\alpha} = \varepsilon, \quad w = w(\varepsilon). \tag{B.6}$$

Differentiating (B.6) with respect to ε, multiplying by $1 - w(\varepsilon)$, and eliminating $(1 - w(\varepsilon))^{\alpha}$ using (B.6) yields the equation

$$(1 - \alpha\varepsilon - \beta w(\varepsilon))w'(\varepsilon) = 1 - w(\varepsilon). \tag{B.7}$$

Inserting the power series (B.5) for $w(\varepsilon)$ and $1 + \sum_{l=1}^{\infty}(l+1)a_{l+1}\varepsilon^l$ for $w'(\varepsilon)$ into (B.7) gives

$$1 - \alpha\varepsilon + \sum_{l=1}^{\infty}(l+1)a_{l+1}\varepsilon^l - \alpha\sum_{l=2}^{\infty} l a_l \varepsilon^l - \beta\varepsilon - \beta\sum_{l=2}^{\infty} l a_l \varepsilon^l - \beta\sum_{l=2}^{\infty} a_l \varepsilon^l - \beta\sum_{l=2}^{\infty} a_l \varepsilon^l \sum_{l=1}^{\infty}(l+1)a_{l+1}\varepsilon^l = 1 - \varepsilon - \sum_{l=2}^{\infty} a_l \varepsilon^l. \tag{B.8}$$

Using that

$$\sum_{l=2}^{\infty} a_l \varepsilon^l \sum_{l=1}^{\infty}(l+1)a_{l+1}\varepsilon^l = \sum_{l=3}^{\infty}\left(\sum_{i=2}^{l-1} i a_i a_{l+1-i}\right)\varepsilon^l,$$

it follows that $a_1, a_2, a_3, \ldots$ can be found recursively as in (B.2)–(B.3), by equating coefficients in (B.8). The result (B.1) then follows from $\eta^{\mathrm{opt}} = -(1/\delta)\ln z = (1/\delta)\ln(1-w)$. The inequality $\beta < 0$ follows from the inequality $\gamma + \phi(\gamma)/\Phi(\gamma) > 0$, $\gamma \in \mathbb{R}$, given in Avram et al. (2013, Section 4). □

We consider next the cases $\alpha = -1$, $1/2$, and 2 that allow for solving the threshold equation explicitly, and that illustrate Proposition 18.

Proposition 19. *Let $t = -\gamma B/(1+\gamma B) > 0$, and $\varepsilon = 1 - R_{\mathrm{T}}(0)$, for the cases* (i) *and* (ii) *below. The optimal threshold η^{opt} is given as*

$$\eta^{\mathrm{opt}} = -\frac{1}{\delta}\ln(1 - w(\varepsilon)),$$

where $w(\varepsilon)$ is given by:

(i) $\alpha = -1$,

$$w(\varepsilon) = \frac{1}{2}t\left(\sqrt{(1+\varepsilon)^2 + \frac{4\varepsilon}{t}} - 1 - \varepsilon\right) = \varepsilon - \frac{1}{2}t\sum_{k=2}^{\infty}(-1)^k \frac{P_k(1+2/t) - P_{k-2}(1+2/t)}{2k-1}\varepsilon^k \tag{B.9}$$

for $|\varepsilon| < 1 + 2/t - \sqrt{(1+2/t)^2 - 1}$, and where P_k is the Legendre polynomial of degree k,

(ii) $\alpha = 1/2$,

$$w(\varepsilon) = \frac{2t}{1+\gamma B}\left(\sqrt{1 + \frac{\varepsilon}{t}} - 1\right) - t\varepsilon = \varepsilon + \frac{2t}{1+\gamma B}\sum_{k=2}^{\infty}\binom{1/2}{k}\left(\frac{\varepsilon}{t}\right)^k \tag{B.10}$$

for $|\varepsilon| < t$,

(iii) $\alpha = 2$,

$$w(\varepsilon) = -\gamma B + \sqrt{(\gamma B)^2 + 2\gamma B\varepsilon} = \varepsilon + \gamma B \sum_{k=2}^{\infty} \binom{1/2}{k} \left(\frac{2\varepsilon}{\gamma B}\right)^k \tag{B.11}$$

for $|\varepsilon| < \frac{1}{2}\gamma B$.

Proof. *Case* (i). When $\alpha = -1$, we can write the threshold equation as

$$w^2 + t(1+\varepsilon)w = t\varepsilon. \tag{B.12}$$

From the two solutions

$$w = -\frac{1}{2}t(1+\varepsilon) \pm \sqrt{\left(\frac{1}{2}t(1+\varepsilon)\right)^2 + t\varepsilon}$$

of (B.12), we take the one with the + sign so as to get w small and positive when ε is small and positive. This gives $w(\varepsilon)$ as in the first line of (B.9), the solution being analytic in the ε-range given in the second line of (B.9). To get the explicit series expression in (B.9), we integrate the generating function

$$\sum_{k=0}^{\infty} P_k(x)\varepsilon^k = (1 - 2x\varepsilon + \varepsilon^2)^{-1/2}$$

of the Legendre polynomials over x from -1 to $-1-2/t$, and we use for $k = 1, 2, \ldots$ that

$$P'_{k+1}(x) - P'_{k-1}(x) = (2k+1)P_k(x), \qquad P_{k+1}(-1) - P_{k-1}(-1) = 0,$$

see Szegö (1939, (4.7.29), (4.7.3-4)) for the case $\lambda = 1/2$.

Case (ii). When $\alpha = 1/2$, we can write the threshold equation as

$$2(1-w)^{1/2} = 2 + \gamma B\varepsilon - (1+\gamma B)w.$$

After squaring, we get the equation

$$w^2 + 2\frac{2 - (2+\gamma B\varepsilon)(1+\gamma B)}{(1+\gamma B)^2}w = \frac{4 - (2+\gamma B\varepsilon)^2}{(1+\gamma B)^2}.$$

After a lengthy calculation, this yields the two solutions

$$w = \frac{2\gamma B}{(1+\gamma B)^2}\left(1 + \frac{1}{2}(1+\gamma B)\varepsilon \pm \sqrt{1 - \frac{1+\gamma B}{\gamma B}\varepsilon}\right). \tag{B.13}$$

Noting that $-1 < \gamma B < 0$ in this case, and that w is small positive when ε is small positive, we take the $-$ sign in (B.13), and arrive at the square-root expression in (B.10), with t given earlier. The series expansion given in (B.10) and its validity range follow directly from this.

Case (iii). When $\alpha = 2$, we have $\gamma B > 0$, and the threshold equation can be written as

$$w^2 + 2\gamma B w = 2\gamma B\varepsilon.$$

Using again that w is small positive when ε is small positive, the result in (B.11) readily follows. □

References

Armony M, Maglaras C (2004) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.

Atar R (2005a) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15(4):2606–2650. MR2187306.

Atar R (2005b) A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* 15(1B):820–852.

Atar R, Mandelbaum A, Reiman MI (2004) A Brownian control problem for a simple queueing system in the Halfin–Whitt regime. *Systems and Control Lett.* 51(3–4):269–275.

Atar R, Mandelbaum A, Shaikhet G (2006) Queueing systems with many servers: Null controllability in heavy traffic. *Ann. Appl. Probab.* 16(4):1764–1804. MR2288704.

Avram F, Janssen AJEM, Leeuwaarden JSHV (2013) Loss systems with slow retrials in the Halfin–Whitt regime. *Adv. Appl. Probab.* 45(1): 274–294. MR3077549.

Bekker R, Borst SC (2006) Optimal admission control in queues with workload-dependent service rates. *Probab. Engrg. Inform. Sci.* 20(4): 543–570.

Borgs C, Chayes JT, Doroudi S, Harchol-Balter M, Xu K (2014) The optimal admission threshold in observable queues with state dependent pricing. *Probab. Engrg. Inform. Sci.* 28(1):101–119.

Çil EB, Örmeci EL, Karaesmen F (2009) Effects of system parameters on the optimal policy structure in a class of queueing control problems. *Queueing Systems* 61(4):273–304.

Chen H, Frank MZ (2001) State dependent pricing with a queue. *IIE Trans.* 33(10):847–860.

Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the Lambert *W* function. *Adv. Comput. Math.* 5(1):329–359.
De Waal P (1990) Overload control of telephone exchanges. Ph.D. thesis, Katholieke Universiteit Brabant, Tilburg.
Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.
Ghosh AP, Weerasinghe AP (2007) Optimal buffer size for a stochastic processing network in heavy traffic. *Queueing Systems* 55(3):147–159.
Ghosh AP, Weerasinghe AP (2010) Optimal buffer size and dynamic rate control for a queueing system with impatient customers in heavy traffic. *Stochastic Processes and Their Appl.* 120(11):2103–2141.
Gurvich I, Whitt W (2009) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.
Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
Jagerman DL (1974) Some properties of the Erlang loss function. *Bell System Tech. J.* 53(3):525–551.
Janssen AJEM, van Leeuwaarden JSH, Sanders J (2013) Scaled control in the QED regime. *Performance Evaluation* 70(10):750–769.
Koçağa YL, Ward AR (2010) Admission control for a multi-server queue with abandonment. *Queueing Systems* 65(3):275–323.
Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing Service Oper. Management* 12(3):511–526.
Kushner HJ, Dupuis P (2001) *Numerical Methods for Stochastic Control Problems in Continuous Time* (Springer, New York).
Lippman SA (1975) Applying a new device in the optimization of exponential queuing systems. *Oper. Res.* 23(4):687–710.
Luenberger DG (1969) *Optimization by Vector Space Methods* (John Wiley & Sons, New York).
Maglaras C, Zeevi A (2003) Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* 49(8):1018–1038.
Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53(2): 242–262.
Maglaras C, Yao J, Zeevi A (2017) Optimal price and delay differentiation in large-scale queueing systems. *Management Sci*, ePub ahead of print February 22, 2017, https://doi.org/10.1287/mnsc.2016.2713.
Massey W, Wallace R (2005) An asymptotically optimal design of $M/M/c/k$ queue. Technical report, Princeton University.
Meyn SP (2008) *Control Techniques for Complex Networks* (Cambridge University Press, New York).
Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.
Olver FWJ (2010) *NIST Handbook of Mathematical Functions* (Cambridge University Press, New York).
Rudin W (1987) *Real and Complex Analysis* (McGraw-Hill, New York).
Stidham S (1985) Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control* 30(8):705–713.
Szegö G (1939) *Orthogonal Polynomials* (American Mathematical Society, Providence, RI).
Teschl G (2014) *Topics in Real and Functional Analysis.* Unpublished, available at http://www.mat.univie.ac.at/~gerald/.
Ward AR, Kumar S (2008) Asymptotically optimal admission control of a queue with impatient customers. *Math. Oper. Res.* 33(1):167–202.
Weerasinghe A, Mandelbaum A (2013) Abandonment versus blocking in many-server queues: Asymptotic optimality in the QED regime. *Queueing Systems* 75(2–4):279–337.
Whitt W (2004) A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* 52(6):922–941.
Whitt W (2005) Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* 30(1):1–27.
Yildirim U, Hasenbein JJ (2010) Admission control and pricing in a queue with batch arrivals. *Oper. Res. Lett.* 38(5):427–431.